Preprints.org

Article

Not peer-reviewed version

MA-YOLO: Multi-Scale Attention-Enhanced YOLO for Object Detection in Remote Sensing Images

TingSong Sun , JianMin Wang * , Jianyu Sun *

Posted Date: 2 May 2025

doi: 10.20944/preprints202505.0025.v1

Keywords: multi-scale object detection; remote sensing images; attention mechanism; multiangle pooling

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

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

Article

MA-YOLO: Multi-Scale Attention-Enhanced YOLO for Object Detection in Remote Sensing Images

Tingsong Sun, Jianmin Wang * and Jianyu Sun *

School of Surveying and Geospatial Science, Liaoning Technical University, Liaoning 12306, China;
* Correspondence: wjminlntu@163.com (J.W.); sjy170331@163.com (J.S.); Tel.: +86-176-1682-5884

Abstract: Object detection plays a crucial role in remote sensing by enabling automated information extraction and supporting downstream decision-making. However, remote sensing images often exhibit complex backgrounds and significant scale variations, leading to degraded detection performance. To address these challenges, we propose MA-YOLO, an enhanced variant of YOLOv7 designed for robust object detection in remote sensing imagery. First, we introduce the dilated convolution layer aggregation network (DELAN-1), which integrates MobileViTv3 and dilated convolution to effectively balance global and local feature extraction while reducing computational overhead. This improves semantic representation in complex backgrounds and across diverse object scales. Second, we propose the cross-layer feature fusion module (CFFM), which enhances information flow between the backbone and neck networks by fusing shallow positional information with deep semantic features, mitigating contextual information loss. Finally, we incorporate the multi-angle pooling attention module (MAPA) in the neck network, leveraging multi-angle pooling and Transformer-based attention to capture target features from multiple directions, improving feature extraction robustness and multi-scale detection performance. Extensive experiments on NWPU VHR-10, VisDrone2019, and RSOD datasets demonstrate the effectiveness and robustness of MA-YOLO. Specifically, on NWPU VHR-10, MA-YOLO achieves a 1.1% improvement in $mAP_{0.5}$ and a 0.9% increase in $mAP_{0.5:0.95}$ over YOLOv7, highlighting its superior capability in handling complex backgrounds and multi-scale object detection.

Keywords: multi-scale object detection; remote sensing images; attention mechanism; multiangle pooling

1. Introduction

With the rapid advancement of remote sensing technology, remote sensing imagery [1–5] has become essential in military reconnaissance, disaster detection, ecological monitoring, and various other applications . Accurate object detection in remote sensing images allows effective identification and localization of critical targets such as vehicles, ships, and individuals, significantly enhancing decision-making processes [6–8]. Despite substantial progress, object detection in remote sensing imagery still faces numerous challenges, including complex backgrounds, significant scale variations, limited resolution, and unclear object features, all of which reduce detection accuracy [9–11].

In recent years, deep learning methods, particularly convolutional neural networks (CNN), have demonstrated significant potential in remote sensing object detection [12–14]. CNN-based object detection algorithms are generally categorized into two-stage and one-stage methods. Representative two-stage methods include Faster R-CNN [15–17] and Mask R-CNN-based approaches [18,19]. These algorithms first generate candidate regions and then perform detailed feature extraction and classification. While these methods offer high accuracy, their dependence on two-step processes results in increased computational costs, limiting their application in real-time detection scenarios [20]. Moreover, the anchor-based design of two-stage detectors often requires careful hyperparameter tuning, which may not generalize well across different remote sensing datasets. In addition, their performance degrades when detecting small or densely packed objects, which are prevalent in complex aerial or satellite imagery.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

In contrast, one-stage detection methods, such as You Only Look Once (YOLO) [21], Single Shot MultiBox Detector (SSD) [22], CenterNet [23], and EfficientDet [24], directly predict bounding boxes and categories from images in a single forward pass, significantly improving computational efficiency [25]. Among these, the YOLO series has achieved widespread adoption due to its superior balance between speed and accuracy [26,27]. Recent advancements have further improved YOLO's detection capabilities, particularly in handling dense small objects and complex backgrounds [28–30]. However, there remains room for improvement regarding multi-scale object detection performance and contextual information preservation. Specifically, remote sensing images often contain objects with extreme scale variations and background clutter, posing significant challenges to existing one-stage models. Therefore, designing more effective feature fusion mechanisms and multi-scale representation strategies remains an active research direction in the field.

To address the above issues, this paper proposes an enhanced YOLOv7-based detection framework called MA-YOLO, specifically designed for robust object detection in complex remote sensing scenarios. We introduce three novel modules into YOLOv7: the Dilated Efficient Layer Aggregation Network (DELAN-1), the Cross-layer Feature Fusion Module (CFFM), and the Multi-Angle Pooling Attention Module (MAPA). The DELAN-1 module integrates the Dilated Mobile Vision Transformer version 3 (DMobileViTv3) to effectively capture both global and local semantic features. The CFFM module further bridges shallow positional features with deeper semantic information, enhancing contextual preservation and cross-layer representation capability. The MAPA module, incorporating multi-angle pooling with transformer attention mechanisms, ensures robust feature extraction from multiple spatial directions, improving multi-scale detection performance.

In summary, our main contributions are as follows:

- We propose the DELAN-1 module, integrating DMobileViTv3, which effectively captures multi-scale semantic information through a combination of MobileViTv3 and dilated convolution operations.
- We develop the CFFM module, facilitating the fusion of shallow positional information and deep semantic features, enhancing cross-layer information flow and context preservation.
- We design the MAPA module, utilizing multi-angle pooling combined with a lightweight transformer structure to robustly extract object features from diverse directions, significantly improving detection accuracy.
- Extensive experiments conducted on benchmark datasets NWPU VHR-10, VisDrone, and RSOD demonstrate that our proposed MA-YOLO method achieves superior detection accuracy, effectively addressing challenges posed by complex backgrounds, scale variations, and multi-object scenarios.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

Figure 1. Sample examples from the three datasets.

2. Related Work

2.1. Object Detection Based on YOLO

Object detection in remote sensing imagery [3] remains a formidable challenge due to complex backgrounds, significant scale variations, and densely distributed targets. To overcome these obstacles, researchers have developed a range of strategies, including multi-scale feature fusion, attention mechanisms, advanced data augmentation techniques, and optimized loss functions [4]. These approaches have significantly enhanced models' capability to distinguish targets in diverse environments and improved detection performance across multiple scales in complex remote sensing scenarios [5].

Among one-stage detection frameworks, the YOLO (You Only Look Once) series has gained widespread recognition for its superior balance between detection accuracy and computational efficiency. In particular, YOLOv7 emerges as one of the most advanced models in this series. It incorporates MaxPool (MP) and Efficient Layer Aggregation Network (ELAN) modules in its backbone, substantially improving feature extraction capabilities. Furthermore, its neck adopts a Path Aggregation Feature Pyramid Network (PAFPN) to facilitate effective multi-scale feature fusion, while the detection head integrates Re-parameterization (REP) blocks to accelerate bounding box predictions. However, despite these advancements, YOLOv7 and similar models still encounter challenges when processing complex backgrounds and detecting multi-scale targets commonly present in remote sensing imagery. The fixed receptive field and limited contextual modeling hinder the model's sensitivity to small or densely packed objects, particularly in cluttered scenes. Additionally, the high computational cost of deeper networks imposes constraints on their deployment in resource-limited environments.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

To address these limitations, recent studies in generative modeling offer valuable insights. Cross-modal conditioning and hierarchical feature fusion mechanisms have shown great potential in enhancing contextual understanding. For example, the IMAGPose framework [31] introduces cross-view attention and multi-level conditioning strategies to improve pose-guided person generation, providing inspiration for better contextual aggregation. Similarly, IMAGDressing-v1 [32] employs customizable conditioning strategies to adaptively control generation outputs, which motivates adaptive feature modulation in detection tasks. Moreover, progressive conditional diffusion models [33] demonstrate that iterative refinement of feature representations can significantly enhance fine-grained details in image generation.

Motivated by these advancements, our work integrates advanced attention mechanisms and hierarchical feature fusion modules to enhance remote sensing object detection. By drawing inspiration from generative frameworks, we aim to improve the model's ability to capture rich contextual information, strengthen multi-scale feature representations, and boost detection performance under complex scenarios.

2.2. Attention Mechanism

The attention mechanism has become a pivotal component in deep learning models, offering the ability to selectively focus on salient regions while suppressing irrelevant background noise [3]. By computing adaptive weights for input features, attention modules enhance feature representation, enabling models to capture more discriminative information, particularly beneficial for detecting small or occluded objects [5].

Various attention-based methods have been proposed to improve detection accuracy. For instance, the Recurrent Attention Model (RAM) [34] incorporates attention into recurrent networks, dynamically selecting important regions to boost detection performance. The Squeeze-and-Excitation Network (SENet) [35] introduces channel-wise attention, adaptively recalibrating feature responses to improve representation learning. Efficient Channel Attention (ECANet) [36] further simplifies channel attention computation by employing 1D convolutions, striking a balance between accuracy and efficiency. Additionally, the Cross Transformer Attention Module (CTAM) [37] combines convolutional and transformer-based attention to better capture both local details and global context. Transformer architectures, built upon self-attention mechanisms, excel at modeling long-range dependencies and capturing complex feature interactions. However, the quadratic computational complexity of self-attention with respect to input size poses scalability challenges in high-resolution remote sensing images.

To overcome these limitations, we propose the Multi-Angle Pooling Attention (MAPA) module, which integrates multi-angle pooling operations with a lightweight transformer design. This module enables the model to extract complementary features from multiple perspectives, effectively capturing both local textures and global semantic cues. By enhancing directional feature learning and reducing redundant computation, MAPA significantly improves the model's ability to understand complex scenes and detect objects more accurately.

3. Materials and Methods

To address the challenges of cluttered backgrounds and target variations in remote sensing images, we propose the MA-YOLO network model, as shown in Figure 2.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

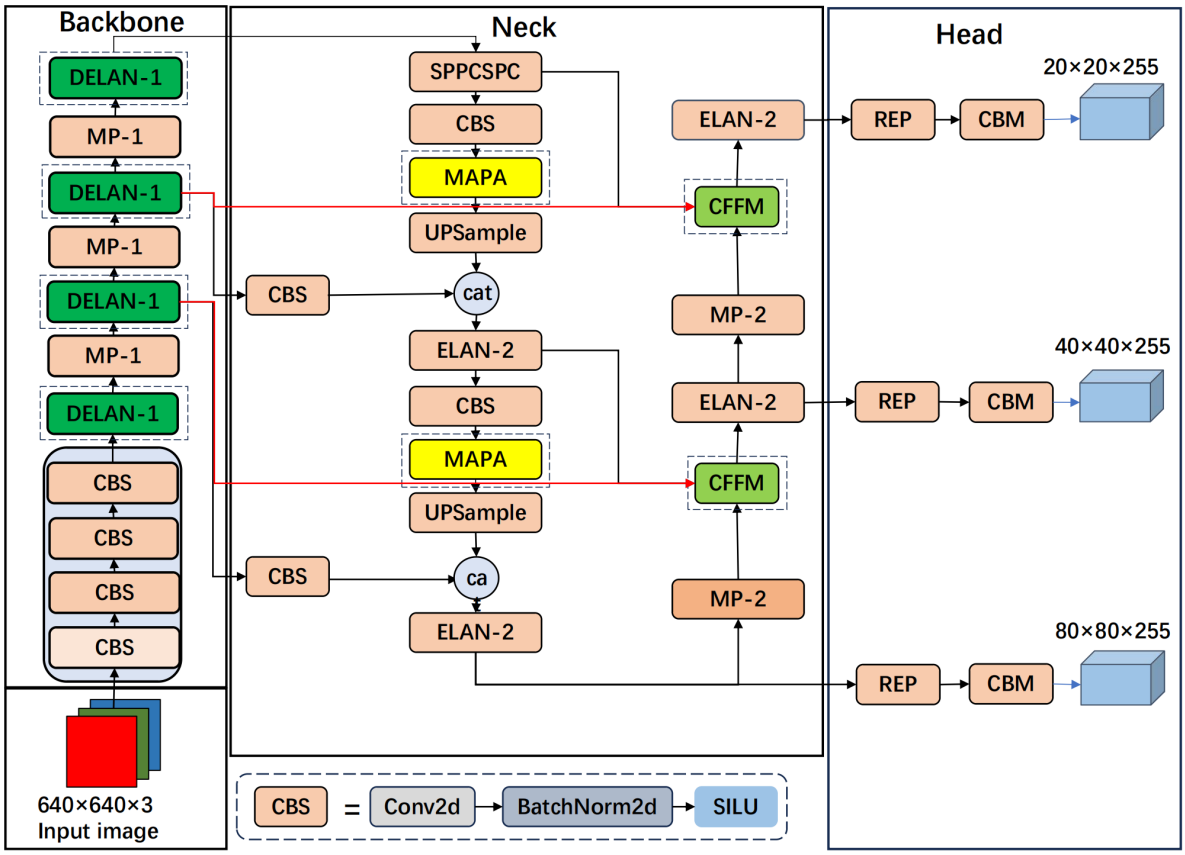

Figure 2. MA-YOLO block.

The MA-YOLO model is mainly composed of MP, SPPCSPC, as well as the newly introduced DEAN-1, CFFM, and MAPA modules. The DELAN-1 module is constituted by the DMobileViTv3 module and standard convolution. It can effectively extract fine-grained and coarse-grained information, thereby enhancing the learning ability of the network. The CFFM mechanism strengthens the model's cross-layer learning ability by integrating the features of the backbone network into the deep feature fusion path of the network neck. In addition, the MAPA module combines multi-angle pooling and the Transformer, extracting network features from multiple angles and improving the feature representation ability of the targets.

3.1. DELAN Block

In YOLOv7, the ELAN-1 module captures rich gradient information by controlling the long and short paths. However, in this module, local spatial context information is captured through 3×3 convolutions, which fails to effectively capture global features, resulting in a decline in the network's ability to learn complex features. To solve this problem, we attempt to replace the 3×3 convolution kernel with that of MobileViTv3[38] and redesign it to propose the DMobileViTv3 module. The Transformer module in MobileViTv3 uses the self-attention mechanism, which has high requirements for computing resources. Moreover, this module relies on global context information, which may lead to the loss of fine-grained information. To address these issues, we use Dilated Convolution to replace the Transformer module and substitute the depthwise separable convolution with a 3×3 convolution. As shown in Figure 3, the DMobileViTv3 module can capture global information with low computing resources, improving the network's ability to extract complex features.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

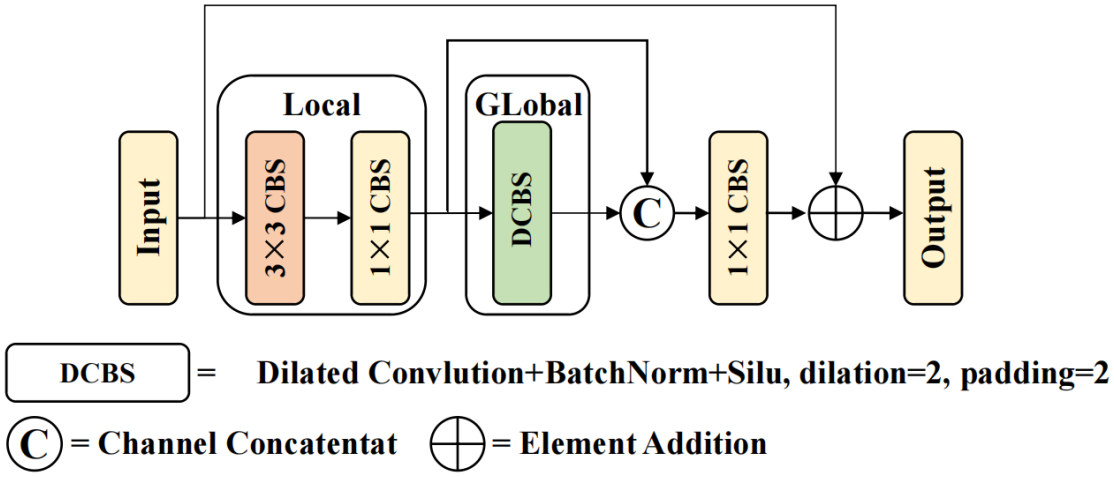

Figure 3. DMobileViTv3 block.

In the DMobileViTv3 module, first, 3×3 convolutions and 1×1 convolutions are used to capture local feature information. Subsequently, Dilated Convolution is employed to capture global feature information. Second, the feature map after capturing global feature information and the feature map after capturing local feature information are concatenated channel-wise, and the channel dimension is reduced through a 1×1 convolution. Finally, the concatenated feature map is summed with the input feature map through a residual structure to obtain the output feature map. As shown in Figure 4, replacing the first 3×3 convolution in the DELAN-1 module with the DMobileViTv3 is helpful for the extraction and fusion of global and local information features, improving the model's feature extraction ability in complex environments.

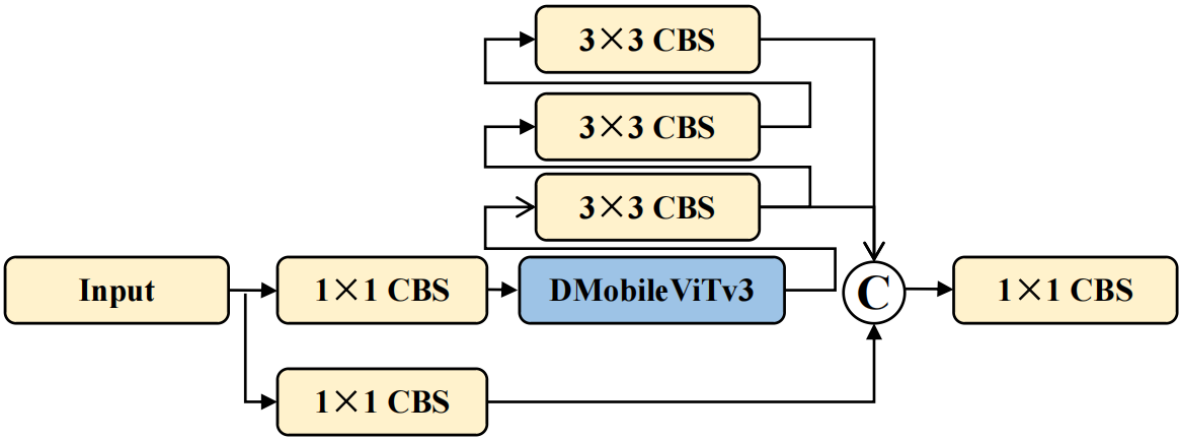

Figure 4. The DELAN-1 module is in the backbone network.

3.2. Semantic and Detail Enhanced Feature Pyramid

In the original YOLOv7 network neck module, the Path Aggregation Network (PAN) and the Feature Pyramid Network (FPN) are used for multi-scale feature fusion, effectively shortening the fusion path of deep semantic information and shallow location information. First, in the PAN-FPN network, a top-down upsampling method is adopted to achieve the fusion of features at different scales, so that the features at the bottom layer contain rich semantic information. Subsequently, a bottom-up downsampling method is used for feature fusion, further improving the feature representation ability of the network. However, fusing different features through upsampling and downsampling methods will cause the loss of some fine-grained information and reduce the accuracy of target detection. Therefore, in order to improve the network's target detection ability in complex backgrounds and at different scales and shorten the information transmission path, we propose the CFFM network structure.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

First, we introduce the CBS module between the backbone network and the neck. Through this module, the output feature map of the backbone network and the feature map of the neck are unified in size, facilitating subsequent feature fusion. Second, the feature map of the backbone network and the features of the neck are summed, and the summed feature map is passed through a global average pooling layer and a linear layer and then through a Sigmoid function to obtain the feature weight. The obtained weight feature is multiplied by the summed feature map. Subsequently, to prevent gradient vanishing, we add a residual structure and concatenate the multiplied feature map with the feature map of the neck. Finally, the features obtained by fusing the backbone network and the neck are concatenated to obtain the output feature map, and its structure is shown in Figure 5.

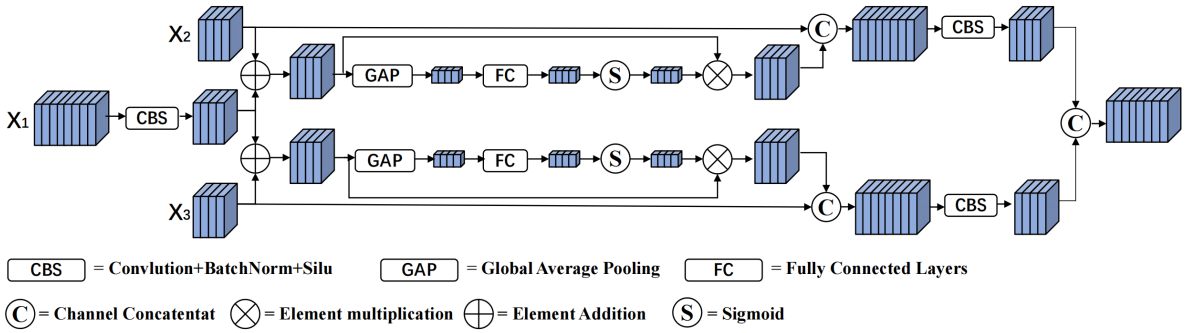

Figure 5. CFFM block.

In the figure, X_1, X_2, and X_3 respectively represent the input feature of the backbone network, the output feature map in the top-down process of the neck network, and the output feature of the previous layer network. Through this module, the fusion of the location information in the backbone network and the semantic information of the deep feature map is strengthened, the fusion path of the shallow features and the deep features is shortened, and the representation ability of the multi-scale features is improved.

3.3. MAPA Attention Mechanism

In remote sensing images, the small size of the targets and the complex background pose a rather significant challenge to the accurate recognition of the targets. To solve this problem, researchers have proposed many attention mechanisms. The attention mechanisms proposed by previous researchers have addressed the limitations of feature redundancy and the inability to focus on important areas that occur when only using convolutional model networks [39]. By introducing an attention mechanism, the model can better understand the importance of different areas for target detection. In order to improve the accuracy of target detection, optimize the number of parameters of the attention model, enhance the feature representation ability, and avoid excessive loss of information, we integrate the self-attention mechanism and multi-angle pooling and propose a MAPM attention mechanism. It aims to extract the feature information of the targets from different directions, avoiding the situation of insufficient information in a single direction, and introduce it into the neck network of YOLOv7, as shown in Figure 6.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

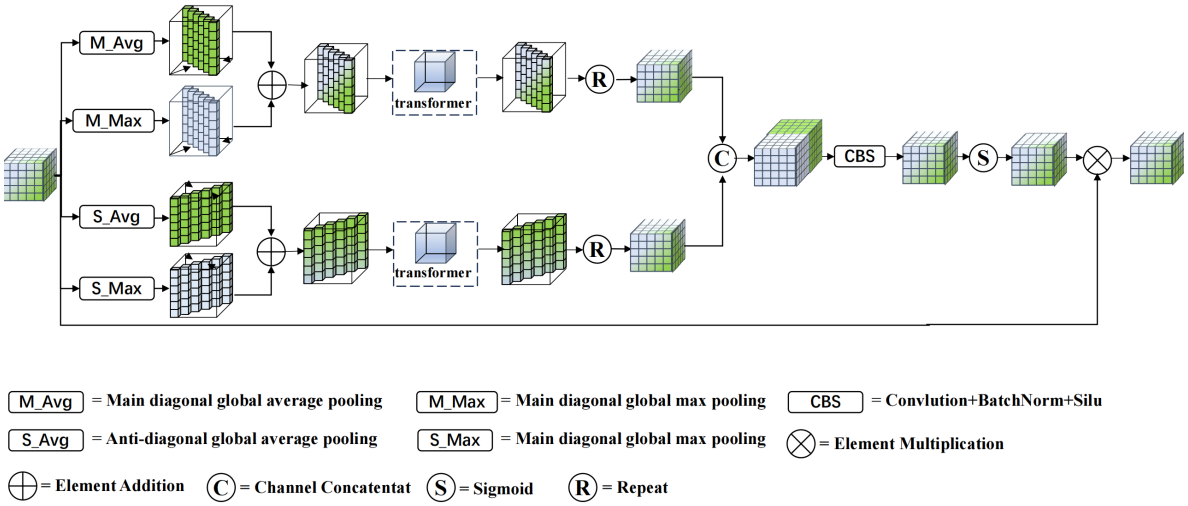

Figure 6. MAPA block.

The MAPM attention mechanism mainly consists of two modules. First, multi-angle pooling improves the feature extraction ability from different angles. Subsequently, the lightweight Transformer module is used to capture global long-range dependencies to obtain the weight matrix of the feature map, and the weight matrix is multiplied by the input feature map.

In the MAPM attention mechanism, multi-angle pooling is used, including four pooling methods: the main diagonal global average pooling, the main diagonal global maximum pooling, the secondary diagonal global average pooling, and the secondary diagonal global maximum pooling. Multi-angle pooling can extract key features from different angles, improve the learning ability of the model, and help capture global information. Suppose the input feature map is represented as$X_{in} = [X_1, X_2, X_3, \ldots, X_c] \in R^{H\times W\times C}$, where the X_C feature layer is represented as:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1w} \\ a_{21} & a_{22} & \cdots & a_{2w} \\ \cdots & \cdots & \cdots & \cdots \\ a_{h1} & a_{h2} & \cdots & a_{hw} \end{bmatrix}_{H\times W} . \tag{1}$$

Then the number of elements output by the multi-angle global pooling of the feature layer is

$$O = H + W - 1. \tag{2}$$

The elements output by the multi-angle global pooling of the feature layer are:

$$\left[A_1, A_2, \cdots, A_m, \cdots, A_O\right] \in R^{O\times C}. \tag{3}$$

Main diagonal global average pooling and main diagonal global maximum pooling are to calculate the mean and maximum values of the elements on the secondary diagonal of the X_C feature layer. Secondary diagonal global average pooling and secondary diagonal global maximum pooling are to calculate the mean and maximum values of the main diagonal elements of the $X_C$feature layer. The computation methods are as follows:

$$A_M^{Avg} = \begin{cases} Q_1 \times \sum\limits_{\substack{i=1, \\ j=m}}^{m,1} a_{i,j}, & m \in \left[1, \dfrac{H+W}{2}\right] \\ Q_2 \times \sum\limits_{\substack{i=H, \\ j=m-W+1}}^{m-H+1,W} a_{i,j}, & m \in \left(\dfrac{H+W}{2}, O\right], \end{cases} \tag{4}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

$$A_M^{Max} = \begin{cases} \underset{\substack{i=1,\\ j=m}}{\overset{m,1}{\text{Max}}} a_{i,j}, & m \in \left[1, \frac{H+W}{2}\right] \\ \underset{\substack{i=H,\\ j=m-W+1}}{\overset{m-H+1,W}{\text{Max}}} a_{i,j}, & m \in \left(\frac{H+W}{2}, O\right], \end{cases} \tag{5}$$

$$A_S^{Avg} = \begin{cases} Q_1 \times \sum_{\substack{i=1,\\ j=W-m+1}}^{m,W} a_{i,j}, & m \in \left[1, \frac{H+W}{2}\right] \\ Q_2 \times \sum_{\substack{i=m-H+1,\\ j=1}}^{H,W+H-m} a_{i,j}, & m \in \left(\frac{H+W}{2}, O\right], \end{cases} \tag{6}$$

$$A_S^{Max} = \begin{cases} \underset{\substack{i=1,\\ j=W-m+1}}{\overset{m,W}{\text{Max}}} a_{i,j}, & m \in \left[1, \frac{H+W}{2}\right] \\ \underset{\substack{i=m-H+1,\\ j=1}}{\overset{H,W+H-m}{\text{Max}}} a_{i,j}, & m \in \left(\frac{H+W}{2}, O\right], \end{cases} \tag{7}$$

$$Q_1 = \frac{1}{m}, \tag{8}$$

$$Q_1 = \frac{1}{H+W-m}. \tag{9}$$

Here, A_M^{Avg} , A_M^{Max} , A_S^{Avg} and A_S^{Max} respectively represent the outputs of the X_C feature layer after main diagonal global average pooling, main diagonal global maximum pooling, secondary diagonal global average pooling, and secondary diagonal global maximum pooling.

For the multi-angle pooling module, first, the input feature X_{in} is subjected to main diagonal global average pooling and main diagonal global maximum pooling, and the feature maps are summed after obtaining them. Subsequently, the input feature X_{in} is again subjected to secondary diagonal global average pooling and secondary diagonal global maximum pooling, and the resulting feature maps are summed. Finally, the results of the two summations are fed into the Transformer module. The calculation process is expressed as:

$$X_{in}^M = M\_Avg(X_{in}) + M\_Max(X_{in}), \tag{10}$$

$$X_{in}^S = S\_Avg(X_{in}) + S\_Max(X_{in}). \tag{11}$$

Here, X_{in}^M and X_{in}^S respectively represent the outputs of main diagonal global pooling and secondary diagonal global pooling. $M_Avg(\cdot)$is the main diagonal global average pooling operation, $M\_Max(\cdot)$ is the main diagonal global maximum pooling operation, $S\_Avg(\cdot)$ is the secondary diagonal global average pooling operation, and $S\_Max(\cdot)$ is is the secondary diagonal global maximum pooling operation. The size of the feature output after multi-angle pooling is: $X_{in} \in R^{(H+W-1)\times C}$.

To capture the long-range dependencies between features, we introduce a lightweight Transformer module, which is mainly composed of two parts: the Neighborhood Attention (NA) module and the Multi-Layer Perceptron (MLP). Each part consists of LayerNorm and Dropout layers, which can improve the stability of model training and prevent overfitting. The volume of the feature map after multi-angle pooling becomes smaller, and feeding it into the lightweight Transformer module can accelerate the calculation speed. The calculation process of the NA module is:

$$X_{in\_NA}^M = Dropout(NA(LN(X_{in}^M)) + X_{in}^M, \tag{12}$$

$$X_{in\_NA}^S = Dropout(NA(LN(X_{in}^S)) + X_{in}^S. \tag{13}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

Here, $X^{M}_{in\_NA}$ and $X^{S}_{in\_NA}$ represent the outputs of the main diagonal pooling and secondary diagonal pooling after passing through the lightweight Transformer module, $Dropout(\cdot)$ is the dropout layer, $NA(\cdot)$ is the neighborhood attention operation, and $LN(\cdot)$ is the layernorm layer.

The computational process for the Multi-Layer Perceptron (MLP) module is as follows:

$$X^{M}_{MLP} = Dropout(MLP(LN(X^{M}_{in\_NA})) + X^{M}_{in\_NA}, \tag{14}$$

$$X^{S}_{MLP} = Dropout(MLP(LN(X^{S}_{in\_NA})) + X^{S}_{in\_NA}. \tag{15}$$

Here, X^{M}_{MLP} and X^{S}_{MLP}represent the outputs after passing through the forward feed layer and $MLP(\cdot)$ is the forward feed layer operation.

The feature map output by the Transformer is processed through repeat to obtain a feature map with the same size as Xin, and then they are concatenated. Subsequently, the dimension is reduced through a convolution operation. Finally, it passes through a Sigmoid function to obtain the weight feature, which is multiplied by X_{in} to obtain the feature map. The computational process is as follows:

$$\hat{X}^{MS}_{in} = Conv(Cat(repeat(X^{M}_{MLP}), repeat(X^{S}_{MLP}))), \tag{16}$$

$$X_{out} = mul(Sigmoid(\hat{X}^{MS}_{in}), X_{in}). \tag{17}$$

Here, $Cat(\cdot)$is the channel concatenation operation, $repeat(\cdot)$ is the feature element replication, and $Conv(\cdot)$ is the convolution operation.

4. Results

To validate the proposed MA-YOLO method's superiority, it is compared with multiple state-of-the-art MA-YOLO approaches on three large-scale datasets, namely, VisDrone2019, NWPU VHR-10 , and RSOD.

4.1. Datasets

The NWPU VHR-10 dataset comes from Northwestern Polytechnical University in China and is often used in high-resolution remote sensing target detection scenarios. This dataset includes 10 different categories of objects: tank, aircraft, basketball court, baseball field, ship, vehicle, bridge, port, sports field, and tennis court. We divide the NWPU VHR-10 dataset into a training set and a validation set according to the ratio of 8:2.

The VisDrone2019 dataset contains remote sensing images captured from the perspective of unmanned aerial vehicles (UAVs) and is widely used in the research of UAV target detection. The VisDrone2019 dataset consists of 10,209 static images, covering 10 different categories, namely: tricycle, van, bus, boxcar, truck, car, person, bicycle, pedestrian, and motor. According to the official website download method, the VisDrone2019 dataset is divided into a training set of 6,471 images and a validation set of 548 images.

The RSOD dataset is derived from the DIOR dataset and contains high-resolution satellite and aerial images, which are commonly used for small target detection in remote sensing. The RSOD dataset includes 20 different categories of targets, aiming to provide diverse sample targets from different scenes and geographical locations, including aircraft, playground, tank, overpass, etc. We divide the RSOD dataset into a training set and a validation set according to the ratio of 8:2.

4.2. Evaluation Metrics

Performance metrics in deep learning are indicators of model performance, aiding personnel in optimizing and improving models. To quantitatively analyze model performance, we study metrics for object detection including precision (P), recall (R), mean Average Precision (mAP), model parameters (Params), and Giga Floating Point Operations Per Second (GFLOPs). Recall is the proportion of actual positive samples correctly predicted as positive by the model. Precision is the proportion of correctly

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

predicted sample labels among all predicted samples. AP represents the area under the P-R curve, with AP values for each category being independent. mAP is the mean Average Precision for multi-class object detection, evaluating the accuracy of object detection. Parameters can be used to assess model complexity. We record the results as $mAP_{0.5}$ when the Intersection over Union (IOU) threshold is 0.5. By varying the threshold from 0.5 to 0.95 with an increment of 0.05, we calculate the average of these values to obtain $mAP_{0.5:0.95}$

$$Precision = \frac{TP}{TP + FP}, \tag{18}$$

$$\text{Re}call = \frac{TP}{TP + FN}, \tag{19}$$

$$AP = \int_0^1 Precision(\text{Re}call)d(\text{Re}call). \tag{20}$$

Here, TP represents True Positives,FP represents False Positives, andFNrepresents False Negatives.

4.3. Implementation Details

Our experiments are based on PyTorch, using 2 NVIDIA GeForce RTX 3090 (24G video memory) graphics cards. The experimental configuration environment is Python 3.8, Pytorch 2.0.0, and Cuda 11.8. The input sample size is 640×640. In the training stage, we use the SGD optimizer with an initial learning rate of 0.01 and a momentum of 0.937. We set the batch size to 32 and the maximum number of epochs to 300. The remaining hyperparameters are the same as the default settings of YOLOv7.

4.4. Comparison with State-of-the-Art Methods

To verify the effectiveness and advantages of our improved algorithm based on YOLOv7, we compare the improved MA-YOLO algorithm with state-of-the-art methods. We conduct comparative analyses with advanced algorithms on the VisDrone2019, NWPU VHR-10, and RSOD datasets, using the same hyperparameters and training strategies in the experiments.

4.4.1. Comparisons on NWPU VHR-10 Dataset

Table 1 shows the comparison results on the NWPU VHR-10 dataset. Compared with other advanced algorithms, our MA-YOLO algorithm has the highest evaluation indicators. Its Precision (P), Recall (R), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ are 93.6%, 90.7%, 94.5%, and 57.4% respectively. Compared with the lightweight YOLOv10-n algorithm, our algorithm improves Precision (P), Recall (R), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ by 6.0%, 14.8%, 9.5%, and 5.7% respectively. Compared with the YOLOv11-X algorithm with a relatively large number of parameters, our algorithm improves Precision (P), Recall (R), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ by 1.5%, 8.1%, 5.0%, and 0.6% respectively. The experimental results demonstrate that our algorithm has a high detection accuracy.

Table 1. COMPARISON EXPERIMENTS ON NWPU VHR-10.

Method	P(%)	R(%)	mAP$_{.50}$(%)	mAP$_{.50:.95}$(%)	Params(M)	GFLOPs(G)
YOLOv6-n	92.2	84.9	90.2	54.3	4.2	11.8
YOLOv6-s	90.8	86.9	91.8	56.3	16.3	44.0
YOLOv7-tiny	90.4	86.5	91.2	53.2	6.0	13.3
YOLOv7	91.0	90.5	93.4	56.5	37.2	105.3
YOLOv8-s	90.2	87.0	90.9	56.6	11.1	28.5
YOLOv10-n	87.6	75.9	85.0	51.7	**2.7**	**8.2**
YOLOv10-m	85.1	82.2	88.7	55.4	16.5	63.5
YOLOv10-l	88.2	85.3	89.5	55.6	25.7	126.4
YOLOv11-l	90.6	79.3	88.3	54.8	25.3	86.6
YOLOv11-x	92.1	82.6	89.5	56.8	56.8	194.5
Ours	**93.6**	**90.7**	**94.5**	**57.4**	49.1	126.7

To more intuitively compare the experimental results, we draw the Precision- Recall (P-R) curves of our MA-YOLO algorithm and other advanced algorithms. The area under the P-R curve represents

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

the Mean Average Precision (mAP) of all categories at a threshold of 0.5. The larger the area under the P-R curve, the better the detection performance of the corresponding algorithm, as shown in Figure 7. The area under the P-R curve of our MA-YOLO algorithm is the largest, indicating that our MA-YOLO algorithm has a high target recognition accuracy. In the NWPU VHR-10 dataset, the target scales vary greatly and the background is complex. Our MA-YOLO algorithm, by adding DELAN-1 to the backbone network, helps with target recognition in complex backgrounds. The CFFM module enhances the model's adaptability to feature information extraction at different scales through cross-layer feature fusion, and the MAPA extracts information from multiple angles, avoiding the situation of insufficient information extraction in a single direction.

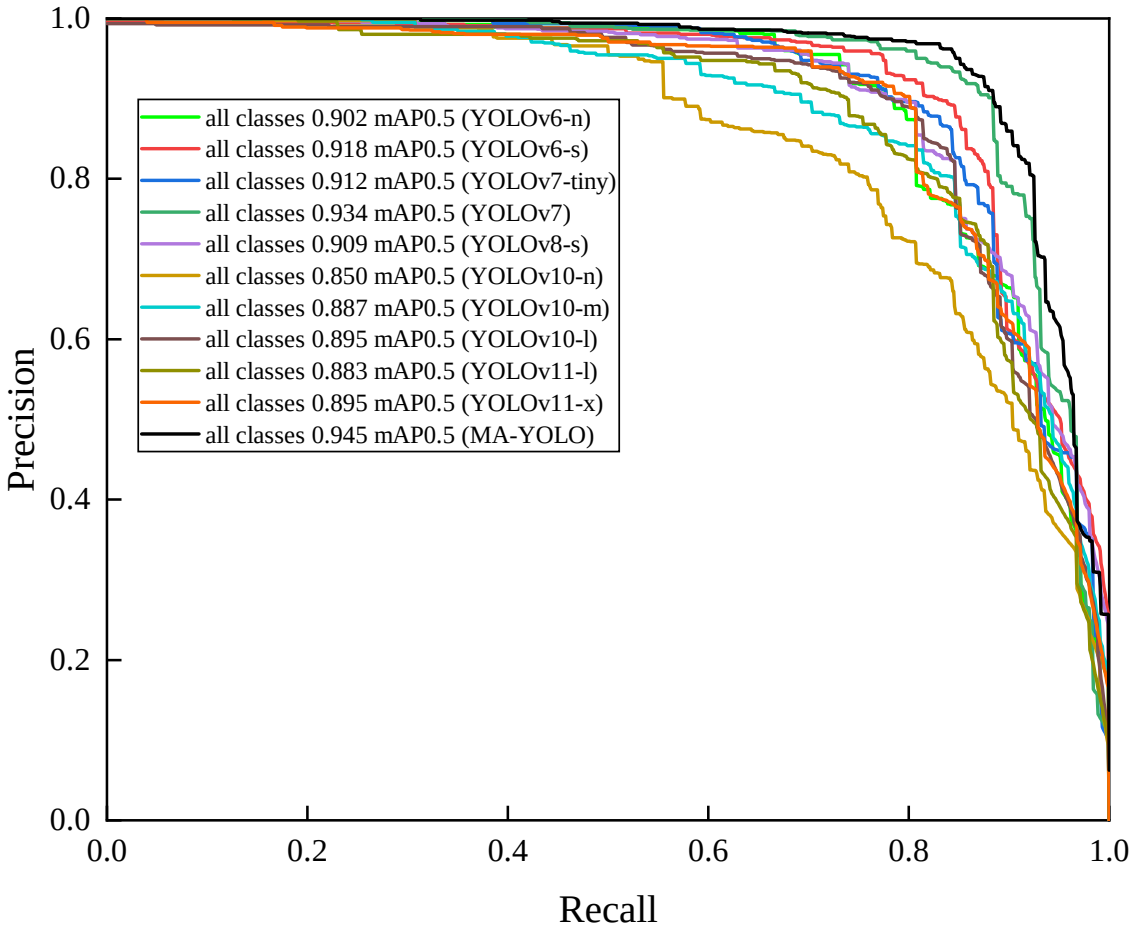

Figure 7. MA-YOLO algorithm and the P-R curves of other advanced algorithms.

4.4.2. Comparisons on VisDrone2019-Val Dataset

To further analyze the advantages of our model, we compare it with other advanced models on the VisDrone2019 dataset, as shown in Table 2. The MA-YOLO algorithm has the highest indicators in all aspects. Its Precision (P), Recall (R), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ are 58.4%, 50.4%, 50.0%, and 28.5% respectively. Compared with the lightweight TA-YOLO-tiny algorithm, our algorithm improves $mAP_{0.5}$ and $mAP_{0.5:0.95}$ by 13.5% and 6.7% respectively. Compared with the YOLOv8-x algorithm with more parameters, our algorithm improves $mAP_{0.5}$ and $mAP_{0.5:0.95}$ by 4.4% and 0.3% respectively. VisDrone2010 is a set of remote sensing target images captured by UAVs, and small-scale targets account for a relatively high proportion. The experimental results show that our MA-YOLO algorithm has great advantages in processing UAV aerial images and small targets.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

Table 2. COMPARISON EXPERIMENTS ON VisDrone2019-val.

Method	P(%)	R(%)	$mAP_{.50}$(%)	$mAP_{.50:.95}$(%)	Params(M)	GFLOPs(G)
YOLOv7-tiny	48.4	38.6	36.6	19.2	6.0	13.3
YOLOv7	58.3	49.3	49.1	28.0	37.2	105.3
YOLOv8-m[40]	55.6	41.6	43.3	26.5	25.9	79.3
YOLOv8-l[40]	57.2	43.2	45.3	27.7	43.6	165.7
YOLOv8-x[40]	56.0	44.0	45.6	28.2	68.2	258.5
TA-YOLO-tiny[41]	48.5	34.9	36.5	21.8	**2.3**	**7.9**
TA-YOLO-n[41]	50.2	38.9	40.1	24.1	3.8	14.1
TA-YOLO-s[41]	53.9	44.3	45.4	27.7	13.9	43.3
YOLO-GE-n[42]	49.3	41.2	40.7	23.7	3.5	-
YOLO-GE-s[42]	54.6	46.2	47.2	28.4	13.0	-
Ours	**58.4**	**50.4**	**50.0**	**28.5**	49.1	126.7

b

4.4.3. Comparisons on RSOD Dataset

To further compare with YOLO series algorithms, we conduct a comparison on the RSOD dataset, as shown in Table 3. Our MA-YOLO algorithm has the highest evaluation indicators in all aspects. The Precision (P), Recall (R), $mAP_{0.5}$, and$mAP_{0.5:0.95}$ of this algorithm are 96.9%, 91.1%, 95.6%, and 67.0% respectively. Compared with the most advanced YOLOv11-m, our algorithm improves$mAP_{0.5}$ and $mAP_{0.5:0.95}$ by 1.7% and 1.5% respectively. Compared with the lightweight YOLOv11-n, our algorithm improves $mAP_{0.5}$ and $mAP_{0.5:0.95}$ by 2.9% and 2.7% respectively. These comparative experiments further verify the advantages and feasibility of our algorithm in target detection with complex backgrounds and large-scale differences.

Table 3. COMPARISON EXPERIMENTS ON RSOD.

Method	P(%)	R(%)	$mAP_{.50}$(%)	$mAP_{.50:.95}$(%)	Params(M)	GFLOPs(G)
YOLOv7-tiny	92.5	89.1	91.1	62.0	6.0	13.3
YOLOv7	94.1	91.9	93.1	66.4	37.2	105.3
YOLOv8-n	90.3	87.2	89.2	63.1	3.0	8.2
YOLOv8-s	88.3	88.7	90.3	65.0	11.1	28.5
YOLOv10-n	88.5	86.8	89.5	62.5	2.7	8.2
YOLOv10-s	92.1	88.4	92.8	65.2	8.1	24.8
YOLOv10-m	92.9	88.9	94.0	64.6	16.5	63.5
YOLOv11-n	91.2	90.2	92.7	64.3	**2.6**	**6.4**
YOLOv11-s	91.4	89.4	92.8	65.4	9.4	21.6
YOLOv11-m	92.8	90.5	93.9	65.5	20.1	68.2
Ours	**96.9**	**91.1**	**95.6**	**67.0**	49.1	126.7

4.5. Ablation Studies and Analysis

The comparison results presented in Table 1, Table 2, and Table 3 indicate that the proposed MA-YOLO method outperforms several state-of-the-art methods, including YOLOv8, YOLOv10, YOLOv11, and others. Next, a comprehensive analysis of the proposed MA-YOLO method from an ablation perspective will be conducted to explore the logic behind its superior performance.

To evaluate the advantages of our MA-YOLO algorithm over the baseline model, we gradually add our improved modules based on the YOLOv7 baseline algorithm to verify the effectiveness of all our improved modules in the experiment, enabling us to better understand and optimize the model. Through this experiment, the feasibility of our MA-YOLO algorithm for target recognition in multi-scale complex backgrounds has been verified. The NWPU VHR-10 dataset contains targets of different scales and complex backgrounds, which poses a great challenge for remote sensing target recognition. Our ablation experiment is carried out on the NWPU VHR-10 dataset, and Table 4 shows the results of our ablation experiment.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

Table 4. Performance with different enhancements evaluated on the NWPU VHR-10 dataset.

Id	MAPA	DELAN-1	CFFM	P(%)	R(%)	mAP$_{.50}$(%)	mAP$_{.50:.95}$(%)	Params(M)	GFLOPs(G)
1	×	×	×	91.0	90.5	93.4	56.5	**37.2**	**105.3**
2	✓	×	×	92.6	89.9	93.8	56.6	39.4	108.6
3	×	✓	×	**95.2**	88.3	93.9	56.8	39.1	113.5
4	×	×	✓	94.5	89.7	94.0	56.9	45.1	115.0
5	✓	✓	×	94.4	89.9	94.1	57.1	41.2	116.9
6	✓	×	✓	94.4	88.1	94.3	57.0	47.3	118.4
7	×	✓	✓	93.6	90.1	94.2	57.2	46.9	123.3
8	✓	✓	✓	93.6	**90.7**	**94.5**	**57.4**	49.1	126.7

Our improved modules have improved the detection performance to varying degrees compared with the baseline model algorithm. Experiment 2, Experiment 3, and Experiment 4 are the evaluation results of a single improved module compared with the baseline model. In Experiment 2, when we added the MAPA module, the Precision (P), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ increased by 1.6%, 0.4%, and 0.1% respectively compared with the baseline model. This is because the improved MAPA module obtains more comprehensive background and detail information through multi-angle information extraction and fusion. In Experiment 3, when we added the DELAN-1 module, the Precision (P), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ increased by 4.2%, 0.5%, and 0.3% respectively compared with the baseline model. This is because by adding the DMobileViTv3 module to ELAN-1, this module can capture both global and local information simultaneously, improving the ability of target recognition in complex scenes. In Experiment 4, when we added the CFFM module, the Precision (P), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ increased by 3.5%, 0.6%, and 0.4% respectively compared with the baseline model. This is because this module effectively improves the cross-layer learning ability of the model by fusing the location information of the backbone network with the deep semantic information of the neck. It is worth noting that although the Recall (R) has decreased compared with the baseline model, higher detection accuracy is shown in other detection evaluation indicators.

To further verify the impact of our improved modules on the experiment, we conduct Experiment 5, Experiment 6, and Experiment 7, which are the evaluation results of adding two improved modules compared with the baseline model. In Experiment 5, when we added the MAPA and DELAN-1 modules, the Precision (P), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ increased by 3.4%, 0.7%, and 0.6% respectively compared with the baseline model. In Experiment 6, when we added the MAPA and CFFM modules, the Precision (P), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ increased by 3.4%, 0.9%, and 0.5% respectively compared with the baseline model. In Experiment 7, when we added the DELAN-1 and CFFM modules, the Precision (P), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ increased by 2.6%, 0.8%, and 0.7% respectively compared with the baseline model. It should be noted that our evaluation indicators $mAP_{0.5}$ and $mAP_{0.5:0.95}$ have also increased compared with a single improved module.

In Experiment 8, the MA-YOLO model integrating all modules shows higher detection accuracy. The Precision (P), Recall (R), $mAP_{0.5}$, and $mAP_{0.5:0.95}$ have been improved to 93.6%, 90.7%, 94.5%, and 57.4% respectively. It should be noted that although the recall rate of our MA-YOLO model is not the highest among the 7 improved experiments, it has increased by 0.2% compared with the baseline model. These experimental results show that adding different numbers of improved modules compared with the baseline model can effectively improve the target detection ability, verifying the advantages and feasibility of our MA-YOLO model in remote sensing target detection. This is because the MA-YOLO algorithm fuses shallow features and deep features through cross-layer feature fusion, enabling the model to distinguish the differences between the background and the target in complex backgrounds. And by extracting target feature information from different directions, it avoids the situation of insufficient extraction of target information and improves the detection accuracy.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

4.6. Visualization

4.6.1. Visual Results on NWPU VHR-10 Dataset

Figure 7 shows some visual detection results of the MA-YOLO model and YOLOv7 on the NWPU VHR-10 dataset. The results show that our MA-YOLO model has a better advantage in target detection under different scales and complex backgrounds. The confidence scores of target detection of the MA-YOLO model are generally higher than those of the YOLOv7 model. In the first group of visual results, the detection confidence of our MA-YOLO algorithm for the larger-scale ground track field and the moderately scaled baseball diamond is higher than that of the YOLOv7 algorithm. In the second group of visual results, in the comparison of vehicles in a complex background with partial shadow occlusion, the detection confidence of our MA-YOLO algorithm is higher than that of the YOLOv7 algorithm. In the third group of visual results, the YOLOv7 model made an error in detecting airplanes of different scales of the same object, identifying one airplane as two airplanes. However, our MA-YOLO model accurately detected the object, and in the detection of smaller-scale airplanes, the detection confidence of our algorithm is higher than that of the baseline model. The results show that our MA-YOLO model has a great advantage in target detection under complex backgrounds and different scales.

4.6.2. Visual Results on VisDrone2019 Dataset

To further verify the ability of our MA-YOLO algorithm to detect dense small targets in complex backgrounds, we compare some visual results of the MA-YOLO algorithm and the YOLOv7 algorithm on the VisDrone2019 dataset, as shown in Figure 8 . Since the targets in the VisDrone2019 dataset are relatively small, we have enlarged the visual results. In the first group of visual results, the YOLOv7 algorithm failed to detect the tricycle, while our MA-YOLO algorithm accurately detected it. In the second group of visual results, the YOLOv7 algorithm failed to detect the motor, while our MA-YOLO algorithm accurately detected the target. In the third group of visual results, due to the truck being blocked by leaves and the motor being blocked by a sign, the YOLOv7 algorithm failed to detect these two targets, while our MA-YOLO algorithm accurately detected the targets. This is because the MA-YOLO algorithm extracts the feature information of small targets from different directions, avoiding the situation of insufficient information extraction in a single direction, verifying that our MA-YOLO algorithm has a great advantage in detecting dense small targets in complex backgrounds.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

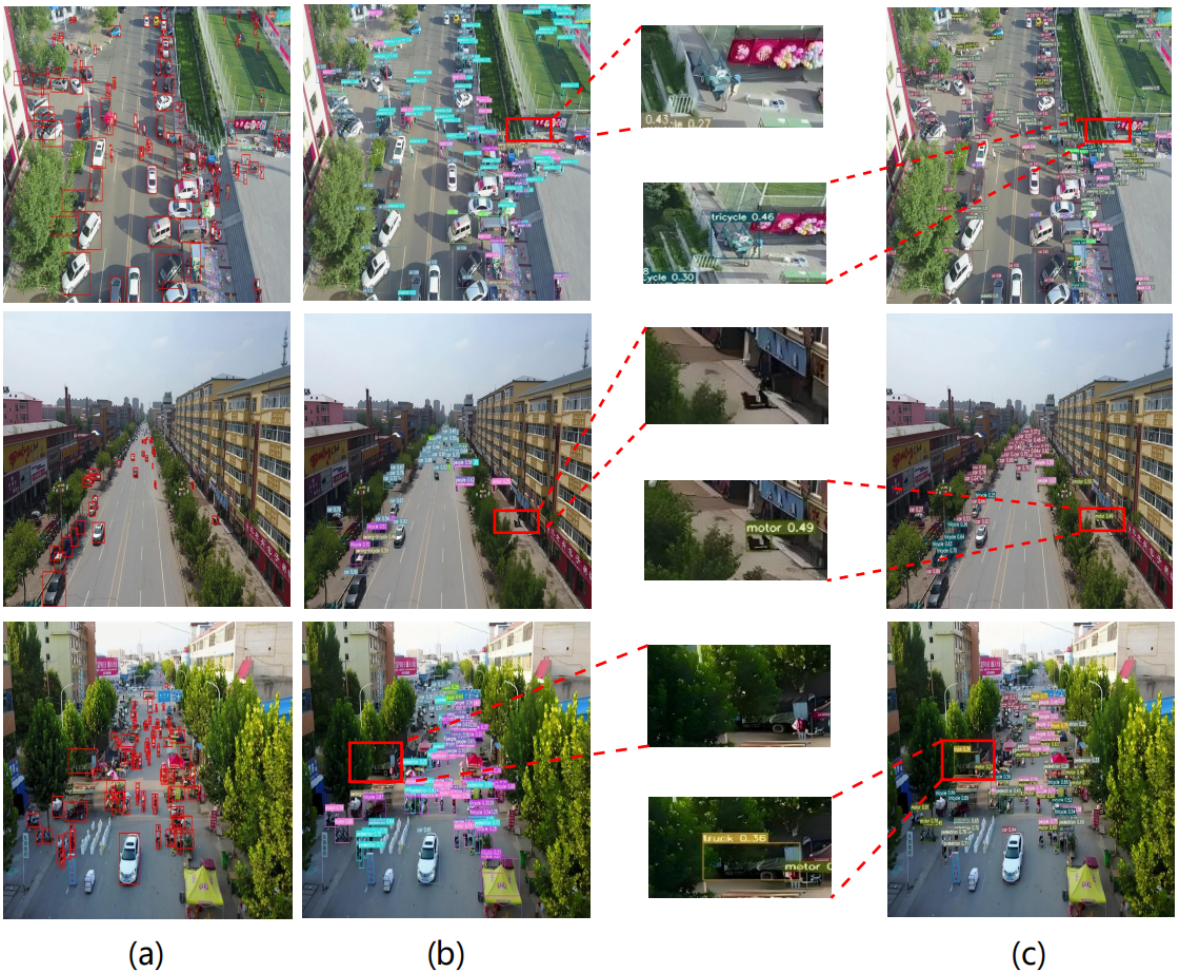

Figure 8. Comparison of some detection examples in the VisDrone2019 dataset: (a) Ground truth, (b) YOLOv7, (c) MA-YOLO.

4.6.3. Visual Results on RSOD Dataset

Figure 9 shows some visual detection results on the RSOD dataset. In the first and second groups of visual results, the YOLOv7 algorithm missed the detection of one aircraft, while our MA-YOLO algorithm accurately detected the aircraft. This is because the missed aircraft target is small and the background is complex, and our YMA-YOLO algorithm can extract feature information from different angles, avoiding the situation of insufficient extraction of feature information of small targets in complex backgrounds. In the third group of visual results, due to the large amount of image noise, the image becomes blurry, which has a certain impact on the detection results. The YOLOv7 algorithm only detected one oiltank, while our MA-YOLO algorithm detected three oiltanks, further indicating the advantage of the MA-YOLO algorithm in target detection in complex noisy environments.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

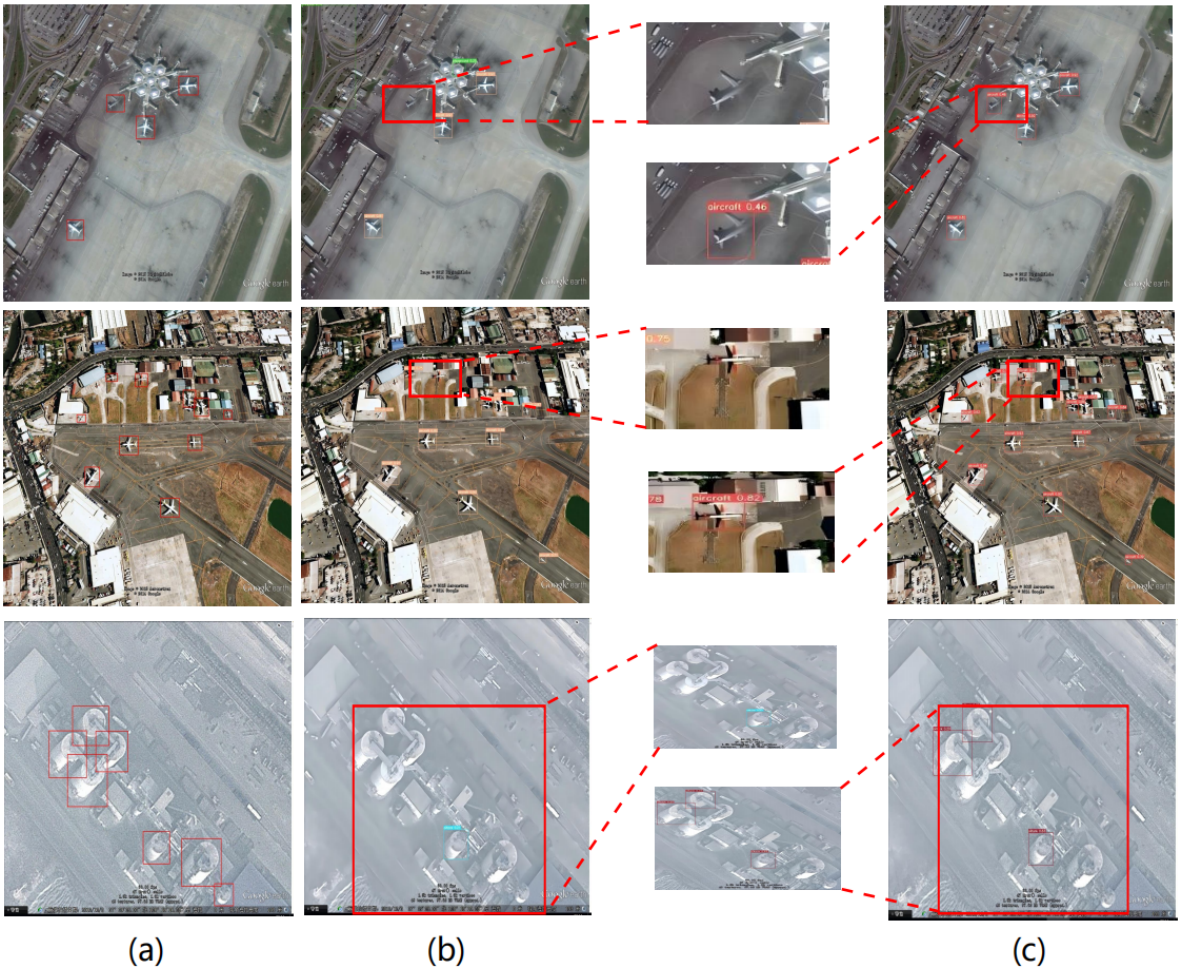

Figure 9. Comparison of some detection examples in the RSOD dataset: (a) Ground truth, (b) YOLOv7, (c) MA-YOLO.

5. Conclusion

This paper aims to solve the problems of complex backgrounds and large-scale variations in remote sensing target detection and further improve the detection accuracy. We proposed a target detection model MA-YOLO for this problem. Firstly, we used the DELAN-1 module in the backbone network to solve the problem of insufficient information extraction and balance the extraction of global and local information. Secondly, we introduced the CFFM module in the backbone network and the neck, which fused the shallow location information and the deep semantic information, helping the model learn context information and improving the cross-layer learning ability. Subsequently, we introduced the MAPA model in the neck network. This model combined multi-angle feature value extraction with the Transformer model, reducing the computational complexity of the self-attention mechanism in the Transformer, and improving the feature learning ability under different target scales through feature extraction from different angles.

Further experimental results have proven that our model can adapt to remote sensing target detection under different scales. The DELAN-1, CFFM, and MAPA modules, when added to MA-YOLO, simultaneously increase the model's params and GFLOPS. In future research, we plan to improve the lightweight and detection accuracy of the model through pruning and distillation techniques, making it more efficient in enhancing the target detection ability under different scales and improving the target detection speed. We also plan to apply this model to high-resolution images and use image compression technology to improve the detection accuracy of the network. This will further improve the model's ability to adapt to target detection under different tasks.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

Author Contributions: Conceptualization, T.S. and W.L.; methodology, T.S.; software, T.S. and J.W.; validation, T.S.and J.S.; formal analysis, T.S. and J.S.; writing—original draft preparation, T.S., J.W. and J.S.; writing—review and editing, T.S. and J.W.; visualization, T.S. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: Liaoning Provincial Department of Education Basic Research Project, grant number JYTQN2023202.

Data Availability Statement: The datasets presented in this study can be downloaded here: https://github.com/Gaoshuaikun/NWPU-VHR-10 (NWPU-VHR-10). https://github.com/VisDrone/VisDrone-Dataset (VisDrone2019). https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset- (RSOD).

Acknowledgments: We thank the editors and reviewers for their hard work and valuable advice.

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

References

1. Zhang, Z. A study on harbor target recognition in high resolution optical remote sensing image. *University of Science and Technology of China: Hefei, China* **2005**.
2. Liang, J. A review of the development of YOLO object detection algorithm. *Applied and Computational Engineering* **2024**, *71*, 39–46.
3. Weng, W.; Wei, M.; Ren, J.; Shen, F. Enhancing Aerial Object Detection with Selective Frequency Interaction Network. *IEEE Transactions on Artificial Intelligence* **2024**, *1*, 1–12.
4. Li, H.; Zhang, R.; Pan, Y.; Ren, J.; Shen, F. LR-FPN: Enhancing Remote Sensing Object Detection with Location Refined Feature Pyramid Network. *arXiv preprint arXiv:2404.01614* **2024**.
5. Qiao, C.; Shen, F.; Wang, X.; Wang, R.; Cao, F.; Zhao, S.; Li, C. A Novel Multi-Frequency Coordinated Module for SAR Ship Detection. In Proceedings of the 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2022, pp. 804–811.
6. Fan, X.; Hu, Z.; Zhao, Y.; Chen, J.; Wei, T.; Huang, Z. A small ship object detection method for satellite remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**.
7. Jiang, S.; Huang, H.; Yang, J.; Zhang, X.; Wang, S. Innovative Research on Small Object Detection and Recognition in Remote Sensing Images Using YOLOv5. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2024**, *48*, 77–83.
8. Shangguan, Y.; Li, J.; Chen, Z.; Ren, L.; Hua, Z. Multiscale attention fusion graph network for remote sensing building change detection. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–18.
9. Reddy, C.C.S.; Divya, R.; Roshini, S.; Shreya, B.; Divyasri, K. YOLO-Extract: Improved YOLOv5 for Aircraft Object Detection in Remote Sensing Images. *International Journal of Information Technology and Computer Engineering* **2024**, *12*, 460–470.
10. Li, Z.; Yuan, J.; Li, G.; Wang, H.; Li, X.; Li, D.; Wang, X. RSI-YOLO: Object detection method for remote sensing images based on improved YOLO. *Sensors* **2023**, *23*, 6414.
11. Liu, W.; Liu, J.; Su, X.; Nie, H.; Luo, B. Multi-level domain perturbation for source-free object detection in remote sensing images. *Geo-Spatial Information Science* **2024**, pp. 1–17.
12. Liu, C.; Zhang, S.; Hu, M.; Song, Q. Object detection in remote sensing images based on adaptive multi-scale feature fusion method. *Remote Sensing* **2024**, *16*, 907.
13. Zhang, Y.; Ma, M.; Wang, Z.; Li, J.; Sun, Y. POD-YOLO Object Detection Model Based on Bi-directional Dynamic Cross-level Pyramid Network. *Engineering Letters* **2024**, *32*.
14. Han, C.; Wu, C.; Hu, M.; Li, J.; Chen, H. C2F-SemiCD: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**.
15. Girshick, R. Fast r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* **2016**, *39*, 1137–1149.
17. Tang, K.; Xu, F.; Chen, X.; Dong, Q.; Yuan, Y.; Chen, J. The ClearSCD model: Comprehensively leveraging semantics and change relationships for semantic change detection in high spatial resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **2024**, *211*, 299–317.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

18. Huiming, Y.; Fuxin, X. A remote sensing image target recognition method based on improved Mask-RCNN model. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, 2021, pp. 436–439.
19. Yang, Y.; Tang, X.; Ma, J.; Zhang, X.; Pei, S.; Jiao, L. ECPS: Cross pseudo supervision based on ensemble learning for semi-supervised remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–17.
20. 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 Sensing* **2023**, *15*, 2429.
21. Liu, X.; Gong, W.; Shang, L.; Li, X.; Gong, Z. Remote sensing image target detection and recognition based on yolov5. *Remote Sensing* **2023**, *15*, 4459.
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 21–37.
23. Shi, T.; Gong, J.; Hu, J.; Zhi, X.; Zhang, W.; Zhang, Y.; Zhang, P.; Bao, G. Feature-enhanced CenterNet for small object detection in remote sensing images. *Remote Sensing* **2022**, *14*, 5488.
24. Li, R.; Wu, J.; Cao, L. Ship target detection of unmanned surface vehicle base on efficientdet. *Systems Science & Control Engineering* **2022**, *10*, 264–271.
25. Pérez, A.; Cruz, M.S.; Martín, D.S.; Gutiérrez, J.M. Transformer based super-resolution downscaling for regional reanalysis: Full domain vs tiling approaches. *arXiv preprint arXiv:2410.12728* **2024**.
26. Wang, K.; Zhou, H.; Wu, H.; Yuan, G. RN-YOLO: A Small Target Detection Model for Aerial Remote-Sensing Images. *Electronics* **2024**, *13*, 2383.
27. Liu, W.; Liu, J.; Su, X.; Nie, H.; Luo, B. Source-free domain adaptive object detection in remote sensing images. *arXiv preprint arXiv:2401.17916* **2024**.
28. Han, H.; Zhu, F.; Zhu, B.; Wu, H. Target detection of remote sensing image based on an improved YOLOv5. *IEEE Geoscience and Remote Sensing Letters* **2023**, *20*, 1–5.
29. Liu, R.; Chen, L.; Cai, G.; Liu, Y. Research on small target detection in remote sensing images based on improved YOLOv7. In Proceedings of the International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2023). SPIE, 2023, Vol. 12941, pp. 997–1004.
30. Nie, H.; Pang, H.; Ma, M.; Zheng, R. A Lightweight Remote Sensing Small Target Image Detection Algorithm Based on Improved YOLOv8. *Sensors* **2024**, *24*, 2952.
31. Shen, F.; Tang, J. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In Proceedings of the The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
32. Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; Tang, J. IMAGDressing-v1: Customizable Virtual Dressing. *arXiv preprint arXiv:2407.12705* **2024**.
33. Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313* **2023**.
34. Mnih, V.; Heess, N.; Graves, A.; et al. Recurrent models of visual attention. *Advances in neural information processing systems* **2014**, *27*.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
36. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.
37. Lang, K.; Cui, J.; Yang, M.; Wang, H.; Wang, Z.; Shen, H. A Convolution with Transformer Attention Module Integrating Local and Global Features for Object Detection in Remote Sensing Based on YOLOv8n. *Remote Sensing* **2024**, *16*, 906.
38. Wadekar, S.N.; Chaurasia, A. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv:2209.15159* **2022**.
39. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 815–825.
40. Jocher, G.; Chaurasia, A.; Qiu, J.Y. by Ultralytics (2023) https://github. com/ultralytics/ultralytics, 2023.
41. Li, M.; Chen, Y.; Zhang, T.; Huang, W. TA-YOLO: a lightweight small object detection model based on multi-dimensional trans-attention module for remote sensing images. *Complex & Intelligent Systems* **2024**, *10*, 5459–5473.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 May 2025 doi:10.20944/preprints202505.0025.v1

42. Yue, M.; Zhang, L.; Zhang, Y.; Zhang, H. An Improved YOLOv8 Detector for Multi-scale Target Detection in Remote Sensing Images. *IEEE Access* **2024**.

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