



RMRN-DETR: regression-optimized remote sensing image detection network based on multi-dimensional real-time detection and domain adaptation

Muzi Chen, Hanrui Zhang, Baohua Cheng, Kun Li, Yinuo Li, Chengzhi Guo, Jiurong Liu, Youbing Li, Lewei Jing & Xinchang Fang

To cite this article: Muzi Chen, Hanrui Zhang, Baohua Cheng, Kun Li, Yinuo Li, Chengzhi Guo, Jiurong Liu, Youbing Li, Lewei Jing & Xinchang Fang (2025) RMRN-DETR: regression-optimized remote sensing image detection network based on multi-dimensional real-time detection and domain adaptation, International Journal of Remote Sensing, 46:22, 8411-8439, DOI: [10.1080/01431161.2025.2564908](https://doi.org/10.1080/01431161.2025.2564908)

To link to this article: <https://doi.org/10.1080/01431161.2025.2564908>



Published online: 29 Oct 2025.



Submit your article to this journal [↗](#)



Article views: 81



View related articles [↗](#)



View Crossmark data [↗](#)



RMRN-DETR: regression-optimized remote sensing image detection network based on multi-dimensional real-time detection and domain adaptation

Muzi Chen^a, Hanrui Zhang^a, Baohua Cheng^a, Kun Li^a, Yinuo Li^a, Chengzhi Guo^a,
Jiurong Liu^a, Youbing Li^b, Lewei Jing^b and Xinchang Fang^b

^aSchool of Control and Mechanical, Tianjin chengjian University Tianjin, China; ^bSTECOL Corporation, Power Construction Corporation of China Tianjin, China

ABSTRACT

With the advancement of real-time object detection technology, maintaining high detection accuracy for small objects across multiple scales remains challenging. Conventional convolutional neural networks (CNNs) struggle to effectively capture multi-scale features, often failing to meet detection requirements. This study proposes RMRN-DETR, an optimized remote sensing image detection network based on multi-dimensional real-time detection and domain adaptation. First, we introduce a Multi-dimensional Real-time detection module (MR) to achieve efficient end-to-end accuracy improvement. Second, a Multi-dimensional Domain Adaptation module is proposed to address feature fusion across different scales, effectively capturing both low-level and high-level semantic information in a multi-scale hierarchy. Finally, a novel loss boundary regression module is introduced to enhance bounding box regression accuracy, precisely reflecting the discrepancy between predicted and ground-truth boxes. Experimental results demonstrate a 1.8% accuracy improvement over the baseline on the ROSD dataset and a 2.9% gain on the DIOR dataset. The proposed method significantly enhances the detection accuracy and efficiency of small objects in remote sensing images, demonstrating strong adaptability to complex multi-scale scenarios.

ARTICLE HISTORY

Received 11 April 2025

Accepted 17 September 2025

KEYWORDS

Small object detection;
Multi-dimensional real-time
detection; Multi-dimensional
domain adaptation;
Regression optimization;
Remote sensing detection

1. Introduction

Remote sensing object detection is a research hotspot in computer vision. It faces challenges like calibration and data uniformity, an overly broad monitoring range of information, and single-target detection data. Traditional object detection algorithms have poor generalization abilities and are generally less accurate than deep learning-based algorithms.

For small object detection, CNNs have high computational complexity, especially with large-scale models and datasets. The pooling layers in CNNs reduce information further. For example, a 24×24 target may shrink to about 1 pixel after four pooling layers, and it's

hard to distinguish because of low dimensionality. Moreover, remote sensing images have varying shooting scales and uncertain orientations, so object detectors need to be robust to orientation changes. Due to these challenges, the single feature map from standard convolutional networks can't meet the requirements of remote sensing object detection anymore. To address these issues, Ci, Jinlong et al. (Bokhovkin and Burnaev 2019) utilized RevVit as the backbone network, leveraging its multi-scale feature fusion mechanism to effectively handle targets of different sizes and shapes. Yang et al. (Chang et al. 2024) designed a GSD deep classification network, combining GSD deep features with an attention framework for multi-class object detection. Shen et al. (Ci et al. 2024) proposed a novel dynamic sensing and correlation loss detector (DCDet) for multi-scale object detection. Han et al. (Dong et al. 2021) adopted a self-adaptive pseudo-label assigner (SPA) to adapt to various scenarios. These methods work well in certain cases but require improvements in speed, small object detection, and unsupervised learning.

Traditional DETR uses Transformer as the core for end-to-end object detection by outputting target boxes via bipartite graph matching and discarding components like anchors and NMS. But it has drawbacks: slow training convergence, poor performance on small and dense targets, high computational cost and inefficiency of the attention mechanism on high-resolution images, and lack of domain adaptability.

In machine vision, DETR-based models have innovations in multiple aspects and can be divided into four categories. Structural optimization models like Deformable DETR and Conditional DETR improve encoders and decoders to enhance feature processing and target prediction. Training strategy optimization models such as DAB-DETR and UP-DETR accelerate convergence. Task expansion models like Oscar and Mask-DETR broaden application boundaries by enabling multi-modal and multi-task processing. Lightweight design models including TinyDETR and Mobile-DETR make them suitable for resource-constrained environments. Altogether, they break through traditional DETR's limitations and provide efficient and robust solutions for object detection and segmentation tasks.

Furthermore, balancing detection accuracy, information scale, and monitoring range has always been a critical issue in remote sensing detection technology. Traditional methods often fail to fully capture complex spatial and spectral information. Remote sensing images typically exhibit inconsistent data quality, high computational demands, and complex data processing, which may lead to excessive computational complexity and significant loss of detection accuracy. Luo et al. (Du, Zhang, and Zhang 2021) proposed a new model, DETR (M-DETR), which significantly improved image recognition rates. Liu (Guo et al. 2023) introduced a detection framework, DST-DETR, incorporating a convolutional module, PfConv, to enhance the AOD-Net model's ability to detect small objects in low-quality images. Liu (Han et al. 2024) optimized a deep learning model, Bearing-DETR, using the real-time detection transformer (RT-DETR) architecture, providing significant improvements in defect detection. Kong et al. (He et al. 2021) developed a Drone-DETR model based on RT-DETR, reducing redundant computations caused by complex backgrounds. Huang (Y. Huang and Yuan 2023) proposed a detection framework, AD-DETR, based on DETR (detection transformer), along with an asymmetric relationship fusion mechanism and a decoupled cross-attention head to focus more on visible and contributing regions. These methods trade off speed for accuracy, struggle with tiny objects, and rely on heavy computations.

Besides dealing with the limitations of single-scale feature extraction, the complexity of high-dimensional data in remote sensing detection makes it hard for traditional CNNs to extract and integrate key information effectively. Objects in remote sensing images may be at different scales, and traditional convolutional operations may lack flexibility in handling multi-scale features. So, multi-scale feature extraction methods are necessary. To enhance the superiority of multi-scale feature extraction methods in remote sensing object detection, Zhou et al. (Z. Huang and You 2023) proposed a multi-scale guided module (MSGM), fusing deep and shallow feature maps at multiple scales to reduce the loss of feature information in small objects. Wang et al. (Kong, Shang, and Jia 2024) developed an EfficientNetB4 model based on a Siamese network and a multi-scale gated fusion module, integrating dual-temporal multi-scale features to preserve boundary details and detect fully changed targets. Huang et al. (Li et al. 2024) introduced a multi-scale feature subtraction fusion network, reducing redundant pseudo-change features and improving training efficiency. Chang et al. (M. Liu et al. 2024) proposed a multi-scale attention network, utilizing multi-scale channel and spatial attention mechanisms to extract multi-scale building feature information. Zhang et al. (Z. Liu, Sun, and Wang 2024) presented a YOLO-MFD model, significantly improving target localization accuracy in remote sensing images. Wen et al. (Luo et al. 2024) proposed an automatic registration algorithm for remote sensing images with different spatial resolutions, enabling the alignment of images with varying sizes and resolutions. Xie et al. (Qu et al. 2023) employed a deep CNN with residual structures as the backbone network, achieving multi-scale object detection. He et al. (Shang et al. 2020) proposed an unsupervised change detection (CD) analysis framework based on multi-scale visual saliency coarse-to-fine fusion (MVSF), fusing multi-scale saliency at the pixel level. Dong et al. (Shen et al. 2024) introduced a multi-scale spatial attention region proposal network (MSA-RPN) for high-resolution optical remote sensing images, achieving high recall rates in small target region proposal generation. Shang et al. (Wang et al. 2024) proposed an end-to-end multi-scale adaptive feature fusion network (MANet) for effective fusion. These methods improve accuracy at the cost of higher complexity and remain vulnerable to extreme scales or background noise.

Besides multi-scale fusion, detection accuracy needs attention mechanisms. The core of such mechanisms is to focus on key info, reduce overload and improve efficiency. But traditional ones often have quadratic computational complexity and may overemphasize local info while ignoring the global context, affecting task performance. For remote sensing detection, developing a high-precision attention mechanism is crucial to solve the problem of low detection accuracy. Qu et al. (Wen et al. 2023) proposed a remote sensing small object detection network based on attention mechanisms and multi-scale feature fusion, enhancing target feature extraction capabilities at different scales. Li et al. (Xie et al. 2024) introduced a rotation-equivariant detector with enhanced feature fusion and attention modules, effectively extracting target information from remote sensing images and improving detection accuracy and stability. Guo et al. (Yang et al. 2023) proposed a hyperspectral anomaly detection algorithm based on channel attention mechanisms and LRX, better capturing features of different remote sensing images while reducing noise. Zhao et al. (Yuan and Xu 2021) designed a novel attention mechanism to enhance important information without losing weak information. In addition, traditional algorithms exhibit significant gaps between predicted values and ground

truth. Existing bounding box regression loss functions yield the same value for different prediction results, which reduces the convergence speed and accuracy of bounding box regression. Du et al. (Zhang and Zhu 2024) proposed a Scale-Sensitive IoU (SIOU) loss algorithm for remote sensing image object detection. By introducing an area adjustment factor γ into the loss function, the loss value of bounding boxes can be adjusted, quantitatively distinguishing different bounding boxes. Bokhovkin A et al. (Zhao et al. 2024) proposed a differentiable proxy function for boundary detection metric computation, which can be used as a loss function for binary segmentation in any neural network. Yuan et al. (Zhou et al. 2024) introduced a NeighborLoss function, which improves MIoU, Precision, Recall, and Accuracy compared to the cross-entropy loss function. While improving accuracy, these methods often trade off speed/efficiency and remain vulnerable to extreme imaging conditions (low resolution, noise, clutter). The loss function innovations help but require careful tuning and may not generalize across diverse datasets.

In this study, a remarkable regression-optimized remote sensing image detection network named RMRN-DETR is proposed, which is centred around multi-dimensional real-time detection and domain adaptation. Firstly, in contrast to the multi-scale fusion network of YOLO where the speed and accuracy are negatively impacted by the Non-Maximum Suppression (NMS), our model takes advantage of the end-to-end transformer-based detectors (DETR). Specifically, by thoroughly studying the real-time end-to-end object detector Transformer (RT-DETR) and harnessing the power of advanced DETR, we construct the RMRN-DETR model in two highly innovative steps. The first step involves the design of an efficient hybrid encoder. This encoder is unique as it utilizes cross-scale fusion and decoupled intra-scale interaction. Unlike traditional encoders that often struggle with effectively handling multi-scale features, our hybrid encoder can rapidly process multi-scale features, thereby significantly enhancing the model's ability to deal with complex remote sensing images that contain objects at various scales. Secondly, we build a decoder with uncertainty-minimized query selection. This decoder is an innovative component that not only provides high-quality initial queries but also plays a crucial role in improving the overall accuracy of the model. Moreover, it offers the flexibility to adjust the speed by varying the number of decoder layers, which enables the model to adapt to different application scenarios with diverse requirements for speed and accuracy. When it comes to the attention mechanism, we employ a local attention mechanism that draws inspiration from Convolutional Neural Networks (CNNs) to establish an effective vision transformer. To achieve a more optimal balance between computational complexity and the size of the receptive field, we take an innovative approach to extend the Vision Transformers (ViT) model. By meticulously analysing the patch interactions in the shallow layers of ViT, we extract two key characteristics, namely locality and sparsity. Based on these findings, we propose a novel multi-scale dilated attention (MSDA) mechanism. This mechanism is capable of simulating sparse patch and local interactions within sliding windows. Furthermore, by adopting a pyramid structure where we stack MSDA blocks at the lower layers and global multi-head self-attention blocks at the higher layers, we construct a multi-scale dilated Transformer (DilateFormer) model. This model architecture is a pioneering design in the field and provides a new way to handle visual information more effectively. In addition, considering the advantages and disadvantages of existing bounding box regression loss functions, we introduce a minimum point distance-based bounding box regression loss function, named MPDIoU. This function is inspired by the

geometric features of horizontal rectangles and serves as a brand-new metric for comparing the similarity between predicted and ground-truth bounding boxes. It offers a more precise and effective way to evaluate the performance of bounding box regression compared to traditional loss functions. Finally, extensive experiments have been carried out, and the results demonstrate that our RMRN-DETR model achieves performance that is comparable to state-of-the-art models across a wide variety of vision tasks. This validates the effectiveness and innovation of each of the unique design elements incorporated into our proposed model.

The main contributions of this paper are summarized as follows:

- (1) A multi-dimensional real-time detection module(MR)is proposed, eliminating the negative impact of NMS post-processing on real-time object detection, establishing an end-to-end speed benchmark, and improving detection accuracy.
- (2) A multi-dimensional domain adaptation module(MA) is introduced, along with a loss boundary regression module(BR), highlighting its advantages in handling bounding box regression problems and providing a more precise loss measurement method.
- (3) Our method greatly enhances small object detection, fills technological gaps, and improves the overall performance and practicality of remote sensing detection, standing out in the field.It has strong adaptability and can handle various scenarios.

2. Methodology

The network primarily adopts an efficient hybrid encoder architecture, combining multi-scale feature extraction, attention mechanisms, and a Transformer decoder to achieve efficient and accurate object detection. The core components of the network include the Efficient Hybrid Encoder, which encodes input image features and integrates the strengths of Convolutional Neural Networks (CNNs) and Transformers to capture both local and global features. Multi-scale feature extraction is performed using feature maps from multiple levels, such as F5, S4, and S3, while the AFF1 module fuses feature maps from different levels to enhance feature representation.

The Efficient RepGFPN component serves as an efficient feature pyramid network, employing reparameterization techniques for multi-scale feature fusion. The MSDA (Multi-Scale Dilated Attention) module captures critical feature information through a multi-scale attention mechanism. The network includes common layers such as Conv2d, Conv3d, Batch Normalization (BN), and activation functions (Act) for feature extraction and normalization, as well as Position Embedding to provide positional information for the Transformer.

Finally, the Discreteness-minimal Query Selection & Transformer Decoder & Head component optimizes query vector selection and generates detection results through the Transformer decoder, significantly improving the accuracy and efficiency of object detection. The overall structure of the network is illustrated in [Figure 1](#).

Input serves as the original data processed by the model. The backbone network extracts multi-scale basic features from the original image through different hierarchical feature extraction modules such as 'S3, S4, S5'. Features of AFF1 is to receive the multi – scale features (such as different hierarchical features like S3, S4, S5) output by the backbone network (BACKBONE), and then conduct preliminary integration of these features.

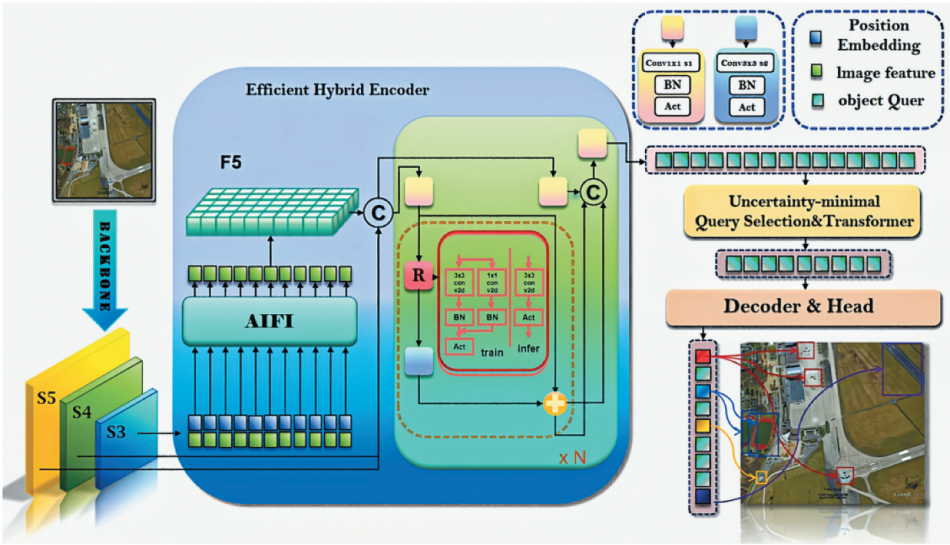


Figure 1. RMRN-DETR network structure.

The features after preliminary integration by the ‘AIFI’ module will be passed to the ‘F5’ module to further construct more abstract and expressive high – level features. The ‘AIFI’ module of the efficient hybrid encoder, the core feature processing module of the model, receives the multi-scale features output by the backbone network for preliminary integration. Then, the integrated features are input into the ‘F5’ module to further construct more abstract and expressive high-level features. The cyclic and repetitive modules (‘R’ and ‘xN’) gradually enhance the complexity and discriminability of the features. After fusion through ‘C’, the uncertainty-minimal query selection & transformer module searches for target-related information in the features, similar to the query mechanism in the Transformer. Position embedding injects positional information into the query vectors or features, enabling the model to perceive the spatial layout. Combined with ‘image features’, operations such as convolution (Conv), batch normalization (BN), and activation (Act) are used to select more reliable queries. Then, leveraging the attention characteristics of the Transformer, the queries interact deeply with the image features, focusing on key target areas and enhancing the representation ability of the features for the targets. In the decoder and head part, the decoder receives the features and queries processed by the Transformer and further decodes information such as the position and category of the targets. Finally, the head, the output module, outputs the target detection results through the classification head, regression head, etc.

2.1. A. Multi-dimensional real-time detection Module(MR)

YOLO requires Non-Maximum Suppression (NMS) for post-processing, which not only slows down inference speed but also introduces hyperparameters, leading to instability in both speed and accuracy. Additionally, the high computational cost limits its practicality and prevents it from fully leveraging the advantages of excluding NMS. To address the issues of reduced inference speed, hyperparameter dependency, and accuracy instability

caused by NMS post-processing in YOLO series models, we propose an efficient hybrid encoder architecture based on RT-DETR [32], combined with a reparameterization module to construct the Multi-dimensional Real-time Detection Module (MS-RTNet) named MR. In the domain dimension, implicit domain alignment and explicit attention are co-optimized. In the task dimension, the same feature pyramid is well handled to optimize detection and segmentation loss at the same time. It solves multi-dimensional problems such as scale dimension, space dimension, Domain dimension, Task dimension, cross-domain generalization (Domain dimension), multi-task compatibility (Task dimension), and computational efficiency adaptation (Dynamic dimension). This module significantly improves the efficiency of multi-scale feature processing through decoupled intra-scale interaction and cross-scale fusion, while supporting flexible speed adjustment by varying the number of decoder layers to adapt to different scenarios without retraining. The feature extraction process is optimized using reparameterization techniques, enabling the network to handle multi-scale features more flexibly and reduce computational complexity.

We design a multi-dimensional real-time detection module, integrating a reparameterization module to build an efficient hybrid encoder. By flattening, reshaping, and combining features, the module enhances the network's performance in tasks such as object detection and image segmentation. Specifically, the 'AIFI' module is part of the Efficient Hybrid Encoder in the object detection deep learning model architecture. Based on the description you provided, its main function is to receive the multi-scale features (such as S3, S4, S5) output by the backbone network (BACKBONE) and perform preliminary integration on these features. It performs intra-scale interaction on high-level features, avoiding semantic confusion and computational redundancy in low-level feature interactions, further reducing computational costs.

The feature fusion module in deep learning optimizes the network in multiple ways. Firstly, it utilizes 1×1 convolutional layers to adjust the number of channels in feature maps. Then, by leveraging 'Queen fusion', it enhances feature interaction, which promotes the dense information flow among feature maps of different scales and enables better fusion of semantic and spatial features at different scales. Moreover, the module adopts CSPNet to replace the traditional 3×3 convolution-based feature fusion and omits the upsampling process in traditional networks, so as to balance the real-time detection accuracy. Finally, the CSPNet is further upgraded by integrating reparameterization mechanisms and Efficient Layer Aggregation Network (ELAN) connections to improve the performance of the network.

The module contains multiple standard 2D convolutional layers (Conv2d), batch normalization layers (BN), and activation function layers (Act) for feature extraction and enhancement. The Conv+BN+Act combination further processes features, and it can be repeated N times to fit different network depths or complexity demands, aiming to boost model performance in object detection or image segmentation via efficient feature fusion and processing (Figure 2). Multi-scale Transformer encoders have computational redundancy as high-level features with rich semantics are extracted from low-level ones, causing ineffective interactions among cascaded multi-scale features. To solve this, we combine DINO-Deformable-R50 (evolving from several DETR variants with specific optimization strategies, using ResNet-50 as the backbone) and the smaller-scale RepGFPN from RT-DETR, applying reparameterization to optimize feature extraction. This enables more flexible multi-scale feature

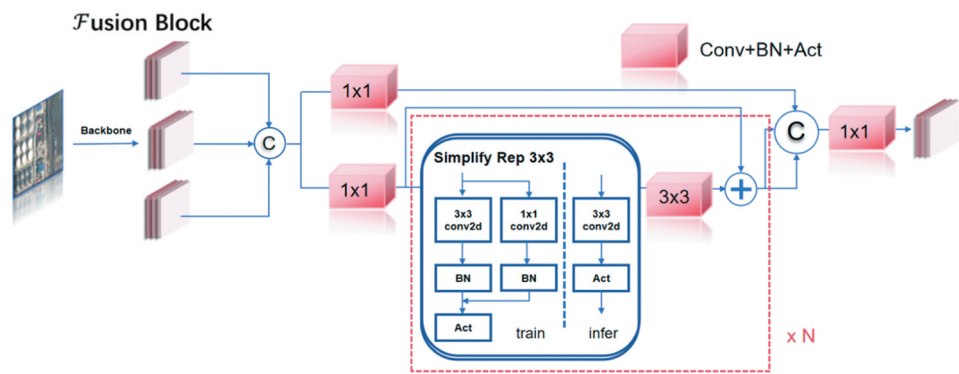


Figure 2. Taking efficient RepGFPN as the neck, high-level semantic and low-level spatial features were extracted and fused.

processing and reduces complexity, mainly for efficient multi-scale feature fusion tasks. The module integrates cross-scale and intra-scale feature interactions. The Single-Scale Encoder (SSE) processes single-scale features, and the Multi-Scale Encoder (MSE) deals with multi-scale ones. The Cross-Scale Fusion (CSF) module combines features across scales. Incorporating Attention-based Intra-scale Feature Interaction (AIFI) and Global Feature Pyramid Network (GFPN) (Figure 4), it enhances feature representation. Through feature concatenation (Concat) and enhancement (Enhanced), the module flexibly adapts to various configurations (variants A/B/C/D/E), significantly improving model performance in multi-scale object detection and image segmentation tasks (Figure 3).

In traditional Feature Pyramid Network (FPN), the top-down path is solely top-down, and the horizontal connection pattern has limitations in handling complex scenes. We derive the improved formulas for GFPN (Generalized Feature Pyramid Network), such as the weighted fusion formula for cross-level fusion that considers features from different levels.

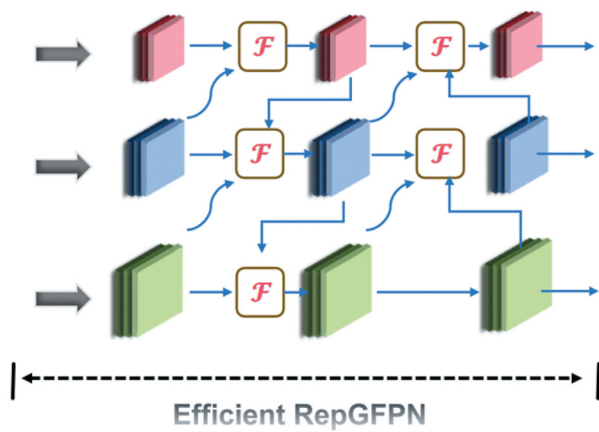


Figure 3. The encoder structure for each variant is included. SSE stands for single-scale Transformer encoder, MSE stands for multi-scale Transformer encoder, and CSF stands for cross-scale fusion. The two modules designed in the hybrid encoder are AIFI and GFPN.

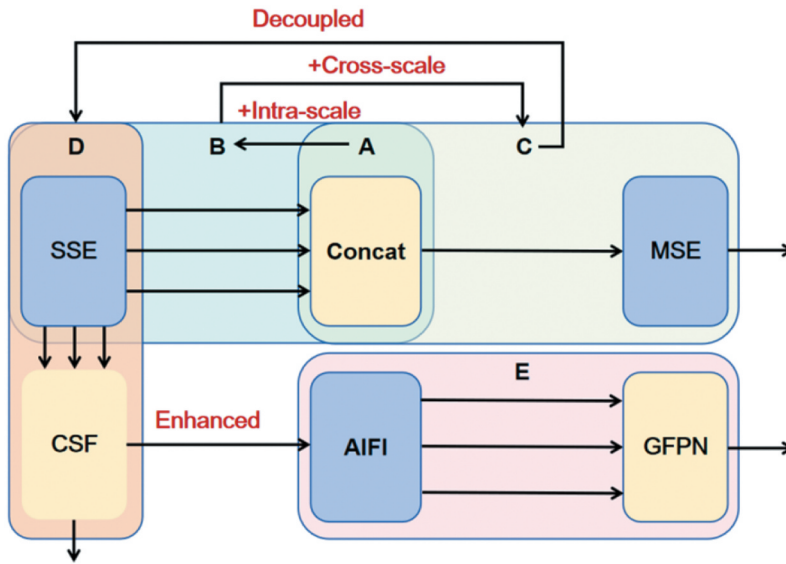


Figure 4. Attention-based feature pyramid structure and queen fusion mechanism (GFPN).

$$F_l = \sum_{i=1}^n w_i * \text{Conv}(C_{l_i}) \quad (1)$$

Let F_l denote the fused feature of the l -th layer in GFPN, C_{l_i} represent the features from different layers participating in the fusion, and w_i signify the corresponding weights.

This novel feature fusion strategy effectively integrates features from different scales, enhancing both richness and accuracy of feature representation. The module demonstrates compatibility with various backbone architectures (ResNet, EfficientNet, etc.) and delivers strong performance across diverse application scenarios.

We conduct experiments with data readers and lighter decoders, initially removing the multi-scale Transformer encoder from DINO-Deformable-R50. A single-scale Transformer encoder is inserted into variant A using variant B, implementing a multi-scale feature-sharing encoder through a single Transformer block for intra-scale feature interaction, then cascading it as output. Variant C utilizes cross-scale feature fusion achieved by B, feeding cascaded features into a multi-scale Transformer encoder to enable simultaneous intra-scale and cross-scale feature interactions. Variant D combines the single-scale Transformer encoder from the former with a PANet-style structure for the latter to realize intra-scale interaction and cross-scale fusion. Variant E enhances D by adding intra-scale interactions and inter-scale fusion.

Since low-level features lack semantic concepts and risk redundant or confusing interactions with high-level features, intra-scale interaction for low-level features becomes unnecessary. To apply self-attention operations to high-level features with richer semantic concepts, capturing relationships between conceptual entities for subsequent object localization and recognition modules, we employ the AIFI single-scale

Transformer encoder to perform intra-scale interaction on S5. This approach further reduces computational costs compared to variant D.

By introducing an improved feature pyramid structure and queen fusion mechanism, MS-RTNet demonstrates outstanding performance across multiple benchmark datasets, particularly in complex backgrounds and multi-scale object detection. Additionally, MS-RTNet adaptively adjusts feature importance, enhancing the network's overall adaptability to various tasks.

2.2. B.Multidimensional domain adaptation Module (MA)

Standard Vision Transformers (ViTs) [34] typically demand substantial computational resources and heavily rely on large-scale datasets, requiring longer training times for convergence which impacts experimental efficiency. These models primarily depend on global context for information processing, where the global attention's receptive field leads to quadratic computational costs. Another branch of Vision Transformers employs CNN-inspired local attention, modelling interactions between patches within small neighbourhoods. While this approach reduces computational expenses, it inherently suffers from performance limitations due to the constrained receptive field of local attention.

To address these challenges, we leverage the sparsity of block-level self-attention mechanisms across different scales and combine it with a multi-scale semantic information extraction method called Multi-Scale Dilated Attention (MSDA) to capture multi-scale semantic information. Given a feature map X , we obtain corresponding queries, keys, and values through linear projections. The feature map channels are then divided into n different heads, each applying different dilation rates for multi-scale SWDA to achieve multi-scale representation learning capability. The formula of SWDA is as follows:

$$X = SWDA(Q, K, V, r) \quad (2)$$

Here, Q , K , and V represent the query, key, and value matrices respectively. Each row of the three matrices indicates a single query/key/value feature vector. For a query at position (i, j) in the original feature map, SWDA sparsely selects keys and values for self-attention within a sliding window with a size of $w \times w$ centred at (i, j) .

Multi-Scale Dilated Attention (MSDA) is a feature enhancement module designed for complex scenarios. Its core objective is to enable the model to simultaneously capture semantic dependencies across different scales through a differentiated dilation rate design. Traditional attention mechanisms typically process features at a single scale, making it challenging to balance local details and global semantics. In contrast, MSDA achieves parallel modelling of multi-scale contextual information without increasing computational complexity by assigning distinct dilation rates to different branches. MSDA decomposes the input feature map into S parallel branches, each corresponding to a unique dilation rate (e.g. $d_1 = 1, d_2 = 3, d_3 = 5$). Through dilated convolution operations, branches with different dilation rates capture features with varying receptive fields: Small dilation rates (e.g. $d = 1$) focus on local details. Large dilation rates (e.g. $d > 1$) capture global semantics. The feature representation of the S -th branch is denoted as:

$$X_s = DilatedConv(X, d_s) \quad (3)$$

Following innovative vision modules like PVT and Swin, we adopt a pyramid architecture to develop a new efficient Multi-Scale Dilated Transformer (DilateFormer) model. MSDA blocks are stacked in shallow layers to capture low-level information, while global multi-head self-attention is stacked in deeper stages to simulate high-level interactions. The model extracts multi-scale semantic information through self-attention mechanisms across different heads, capturing various abstraction levels of images. By exploiting the sparsity of self-attention at different scales, it reduces redundancy without complex operations or additional computational overhead while maintaining performance. The feature map channels are split into multiple heads, each processing different feature subsets in parallel, enhancing the model's learning capacity and efficiency. Different dilation rates across heads enable MSDA to focus on various feature scales, comprehensively capturing image information. Outputs are merged through concatenation and aggregated via linear layers, yielding richer feature representations without extra computational cost.

This improvement effectively aggregates semantic information at different scales by setting varying dilation rates across heads. In remote sensing object detection, it significantly enhances accuracy and robustness through multi-scale analysis and dynamic feature extraction, adapting to diverse complex remote sensing scenarios. Figure 5 demonstrates the combined functionality and outstanding performance of modules A and B. Sliding Window Dilated Attention captures multi-scale contextual information through different dilation rates ($r = 1$, $r = 2$, $r = 3$) and fuses multi-scale features via concatenation. The Efficient Reparameterized Feature Pyramid Network optimizes feature extraction through reparameterization techniques, further enhanced by MSDA to strengthen feature representation. Their combination efficiently captures both local and multi-scale features while significantly improving feature fusion precision and efficiency.

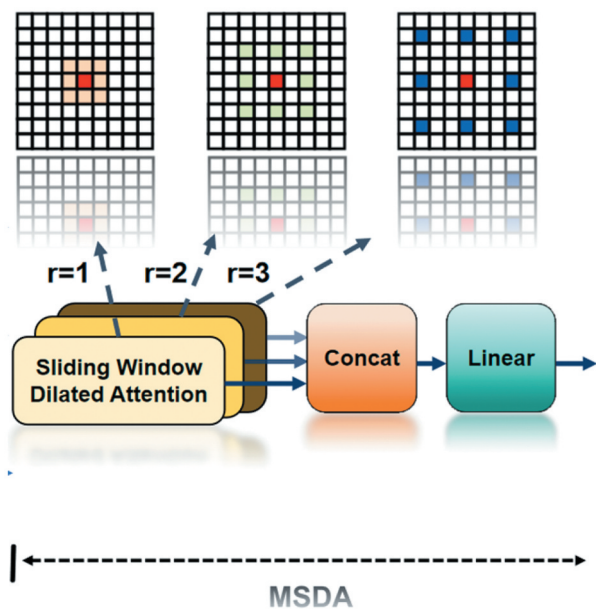


Figure 5. Multi-scale expansion of attention mechanisms (MSDA).

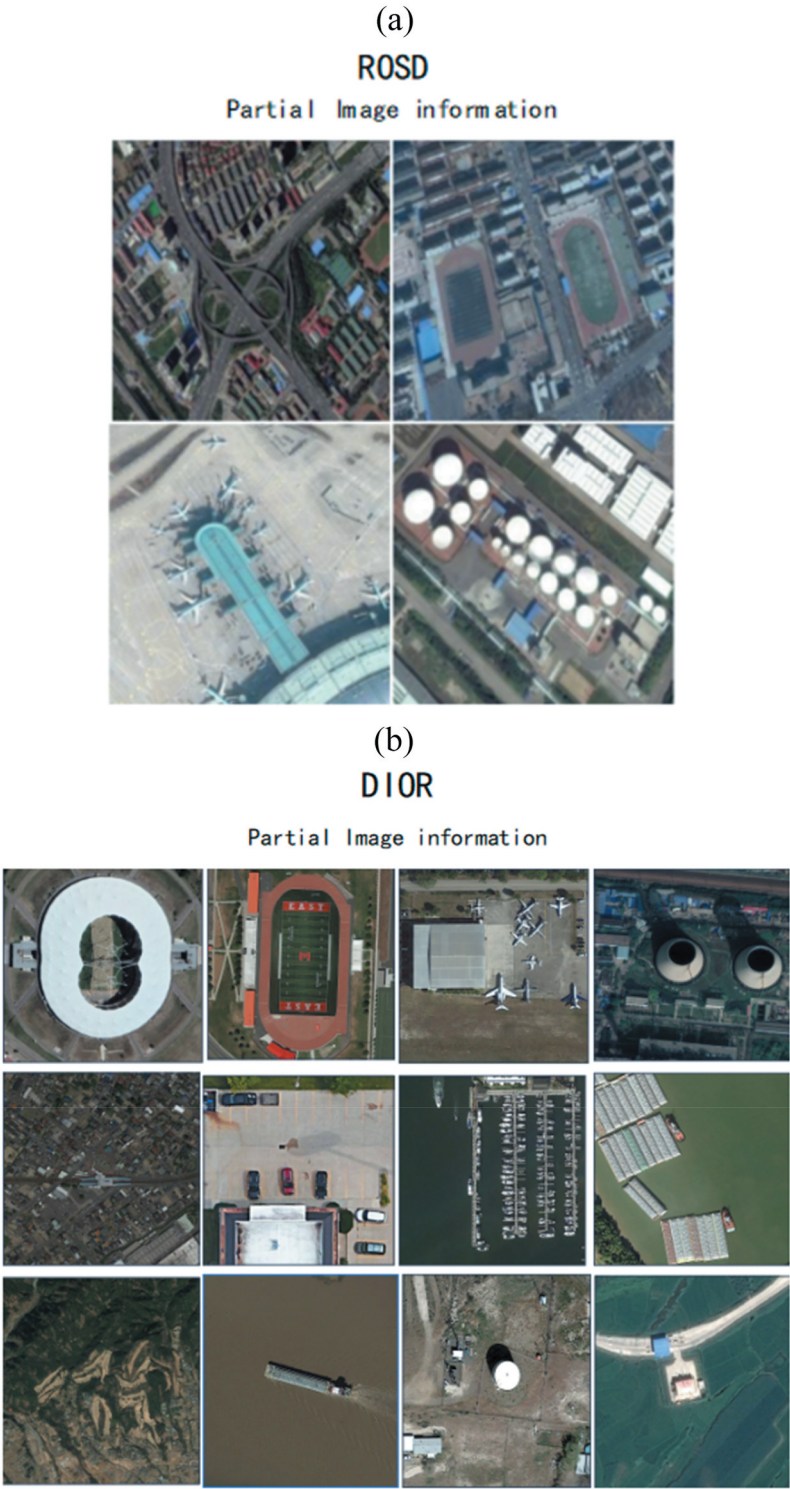


Figure 6. Includes partial image information from the ROSD (A) and DIOR (B) datasets.

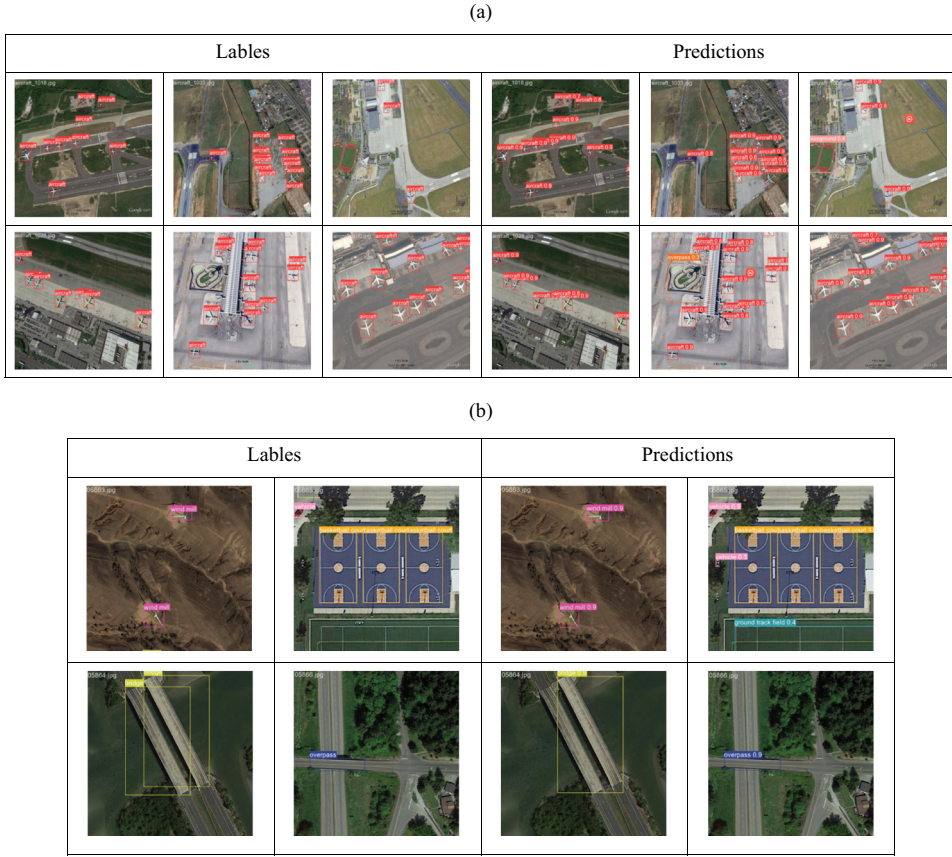


Figure 7. Detection results of RMRN-DETR on RSOD and DIOR datasets. (a) Detection results on RSOD dataset. (b) Detection results on DIOR dataset.

This approach provides robust support for multi-scale object detection in complex scenarios, suitable for tasks like object detection and image segmentation.

The rate $r \in \mathbb{N}^+$ is used to control the sparsity. Specifically, for the position (i, j) , the corresponding component x_{ij} of the output X from the SWDA operation is defined as follows:

$$x_{ij} = \text{Attention}(q_{ij}, K_r V_r) = \text{Softmax}\left(\frac{q_{ij} K_r^T}{\sqrt{d_k}}\right) V_r, 1 \leq i \leq W, 1 \leq j \leq H, \quad (4)$$

Here, H and W are the height and width of the feature map. K_r and V_r represent the keys and values selected from the feature maps K and V . Given a query located at (i, j) , the keys and values located at the following set of coordinates (i, j) .

2.3. C.Loss boundary regression Module (MR)

Existing IoU-based loss functions have significant limitations in object detection and localization, while the MPDIoU (Multi-Perspective Distance IoU) loss achieves notable optimizations through multi-dimensional innovations. Traditional IoU relies solely on

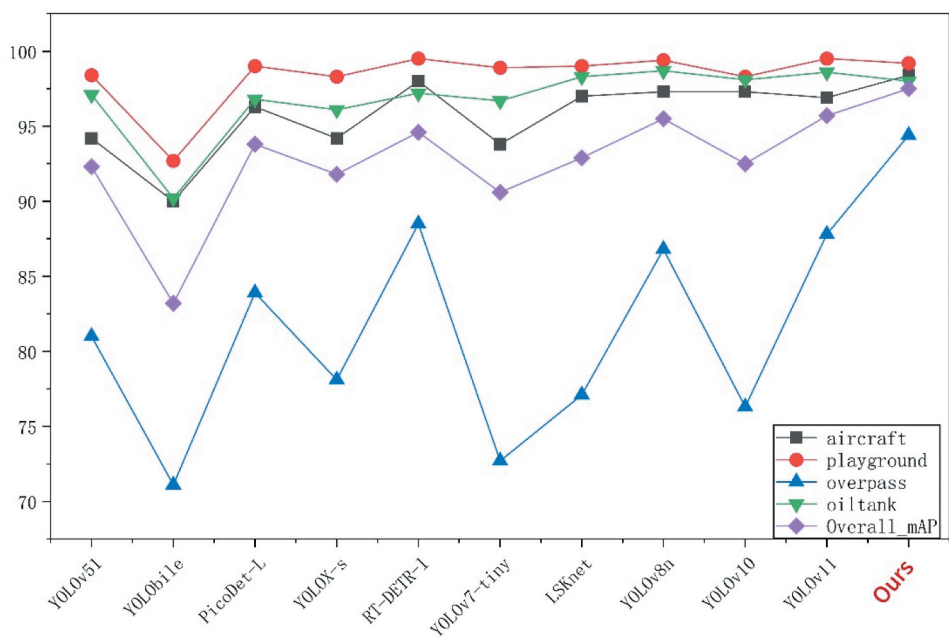


Figure 8. The data comparison chart of our average precision (MAP) score.

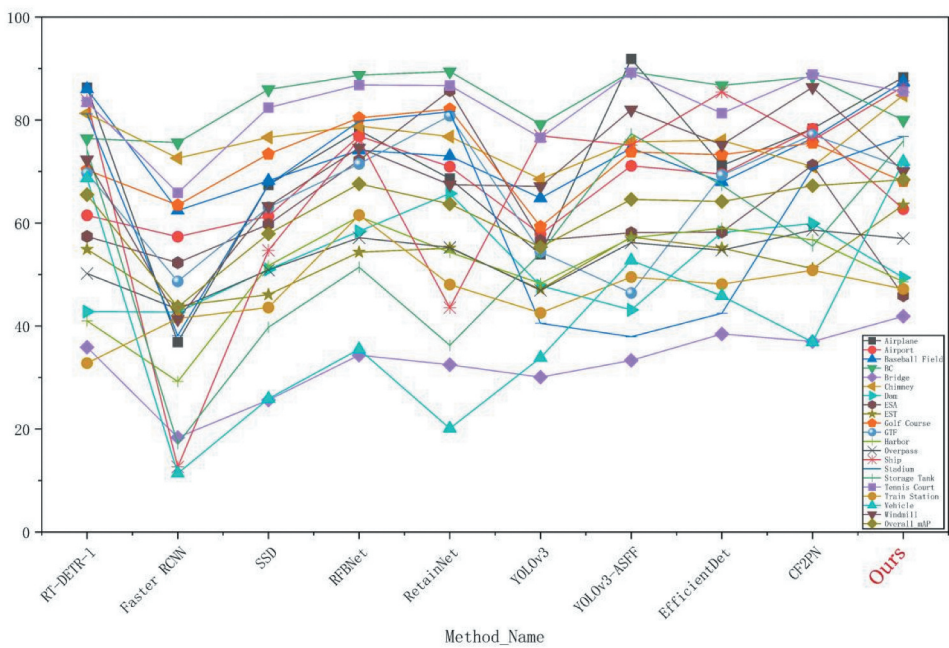


Figure 9. A data comparison chart showing the average score (MAP) of the average accuracy rates of different dior methods.



Figure 10. Comparison of the original method and our method.

the overlapping area, failing to optimize when there is no overlap and being insensitive to scale changes, which results in the same error penalty for both large and small objects. Although GloU introduces the difference in the area of the minimum enclosing convex hull to address the non-overlapping issue, it converges slowly for objects with extreme aspect ratios and lacks sufficient geometric constraints. DIoU incorporates the Euclidean distance between the centre points to enhance the convergence speed but lacks the modelling of multi-dimensional geometric relationships, leading to poor performance in scenarios with rotated objects. CloU further considers the difference in aspect ratios, yet its fixed rules render it ineffective when objects rotate or undergo drastic scale changes, and it has limited optimization for small objects. MPDIoU breaks through these bottlenecks. It replaces single-dimensional measurement with multi-perspective distances, such as horizontal, vertical, and diagonal distances, strengthening the geometric constraints on rotated objects. Meanwhile, by leveraging scale factors to impose greater distance penalties on small objects, it improves the robustness of small object detection. By integrating multi-dimensional geometric constraints and a scale-aware mechanism, MPDIoU can effectively handle non-overlapping and rotated objects, and optimize the localization of small objects in a targeted manner. Although its computational cost is slightly higher compared to DIoU and CloU, this can be improved through lightweight design, making its overall performance significantly superior to traditional methods.

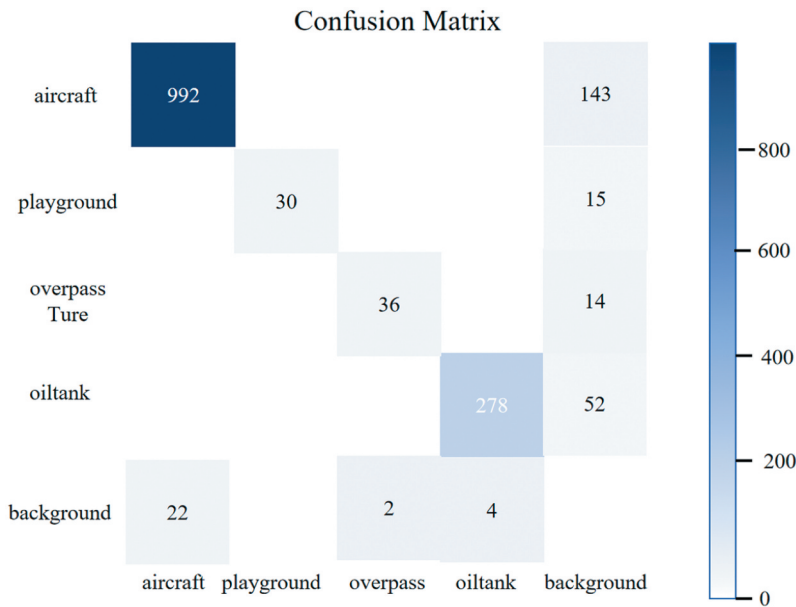


Figure 11. The ROSD dataset confusion matrix results are included.

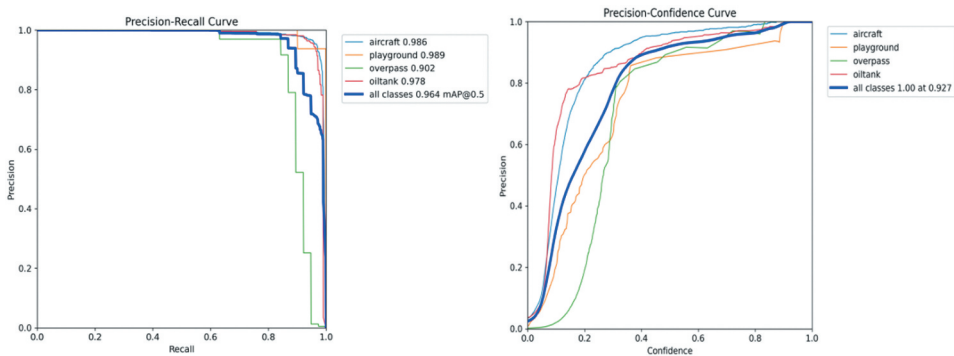


Figure 12. Precision and recall curves and precision confidence curves are only available for MR and MA.

The original IoU (Intersection over Union) serves as a fundamental metric for evaluating the overlap between predicted and ground-truth bounding boxes. It calculates the ratio between the intersection area and the union area of these two boxes. Specifically, the IoU formula can be expressed as:

$$IoU = \frac{A_{inter}}{A_{union}} \quad (5)$$

The intersection area refers to the overlapping region between predicted and ground-truth bounding boxes, while the union area represents their combined total coverage. In object detection tasks, IoU commonly evaluates prediction accuracy by comparing detected results against ground truth. A predefined

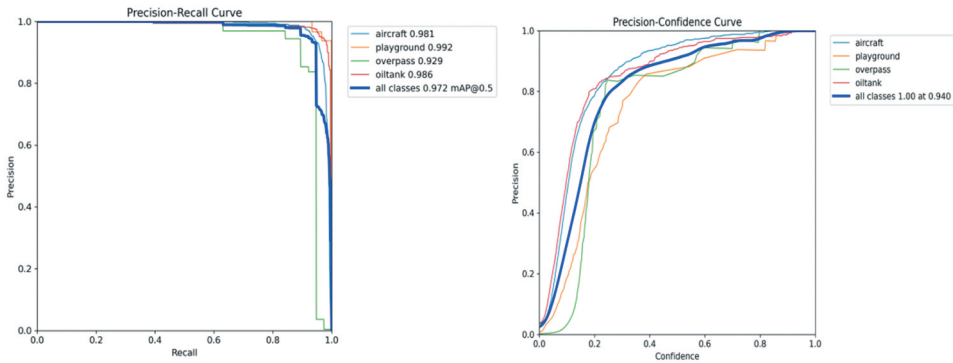


Figure 13. Precision and recall curves and precision confidence curves are only available for MR and BR.

threshold (e.g. 0.5) determines correct matches when IoU meets or exceeds this value, otherwise marking them as incorrect. During training, models optimize predictions by minimizing a loss function that drives predicted boxes towards their corresponding ground-truth annotations. Since norm-based loss functions misalign with the IoU evaluation metric, IoU-based loss functions like MLBR are introduced, formulated as:

$$LMPDI_{IoU} = 1 - MPDI_{IoU} \quad (6)$$

To mitigate the instability of traditional domain adaptation losses, we integrate MIoU into the domain alignment process: instead of relying solely on adversarial losses to minimize distribution differences, MIoU acts as a geometric constraint that aligns not just feature distributions but also the spatial consistency of target predictions across domains, such as when adapting from a source dataset (e.g. aerial images with clear weather) to a target dataset (e.g. foggy remote sensing scenes), where MIoU quantifies the overlap between source-style predictions and target ground truths, guiding the model to preserve spatial relationships (e.g. relative positions of buildings) that are invariant to domain shifts, thereby reducing the volatility of adversarial training by anchoring adaptation to concrete geometric metrics rather than abstract feature distributions.

Instead of relying solely on adversarial losses to align feature distributions, we integrate MIoU as a geometric anchor. MIoU quantifies the spatial overlap between predictions and ground truths across domains, providing a concrete, task-specific signal to supplement abstract distributional alignment. For example, when adapting from urban to rural remote sensing scenes, MIoU emphasizes preserving consistent overlap patterns between building footprints and their contextual features (e.g. roads, vegetation), reducing the volatility of adversarial training. This ensures losses remain tied to detection performance rather than arbitrary feature shifts, making them easier to balance.

3. Dataset used and detection results

3.1. A. Data sets and benchmarks used

The experiments first utilize the RSOD-Dataset released by Wuhan University, commonly employed in aerial remote sensing applications. This dataset contains four target categories: aircraft, playgrounds, overpasses, and oil tanks, comprising 976 images with 6950 annotations, where aircraft represent the most frequent category. RSOD dataset images typically feature high resolution (ranging from 720p [1280×720 pixels] to 4K [3840×2160 pixels]) to ensure detailed presentation and precise boundary annotations. The high-resolution characteristics facilitate accurate capture and labelling of object details and occlusions in complex scenarios.

Subsequently, we employ the large-scale public benchmark DIOR for optical remote sensing object detection. This dataset consists of 23,463 remote sensing images with 190,288 manually annotated axis-aligned bounding boxes (192,472 instances total). All images maintain 800×800 pixel dimensions with spatial resolutions varying from 0.5 m to 30 m. The dataset splits into training/validation sets (11,725 images) and test sets (11,738 images). It demonstrates large-scale diversity in object categories (20 classes including aircraft, airports, baseball fields, basketball courts, bridges, chimneys, dams, highway service areas, toll stations, harbours, golf courses, ground track fields, overpasses, ships, stadiums, storage tanks, tennis courts, train stations, vehicles, and windmills), instance counts, and image numbers, with significant variations in both spatial resolution and intra/inter-class object sizes.

Part of the two datasets are shown as in [Figure 6](#). All models undergo 150-epoch training on the training set, with input images uniformly resized to 640×640 pixels. Batch normalization and standard data augmentation techniques (random flipping, rotation) are applied during training. The experiments adopt SGD optimizer with momentum 0.9 and weight decay 0.0005, while initial learning rates follow cosine annealing schedules from 0.01 to 0.001.

3.2. B. Experimental results data

The RSOD and DIOR datasets used for model validation in the paper are both public remote sensing datasets, and their division strictly follows official standards: the RSOD dataset is divided into a training set (70%), a validation set (15%), and a test set (15%). The training set is used for model parameter learning, the validation set for hyperparameter adjustment, and the test set for independent evaluation of final performance; the DIOR dataset is divided into a training set (80%), a validation set (10%), and a test set (10%). The division process maintains consistent category distribution of samples to ensure the reliability of evaluation results. The detection results of RMRN-DETR on the RSOD and DIOR datasets are shown in [Figure 7](#).

4. Analysis and evaluation of experimental results

This section evaluates and analyzes the target detection performance of our proposed method. The experimental framework employs Python programming language and PyTorch library (The version is 2021.1.1), running on a PC with Intel Core i9-13900HX

processor (2.20 GHz) and NVIDIA GeForce RTX 4060 Laptop GPU (8GB VRAM). We assess the detection performance of our model on two public remote sensing datasets and one self-built dataset, comparing the proposed method with existing classical target detection algorithms to validate its superiority. All experiments maintain consistent parameter settings and evaluation metrics to ensure fair comparisons. The implementation details include standard data preprocessing procedures and identical training protocols across all compared methods. Quantitative results demonstrate consistent performance improvements in both detection accuracy and computational efficiency.

4.1. A. Model accuracy performance evaluation

We train all models on the training set and evaluate them on the test set. The cache setting remains False to prevent image caching, forcing reloading during each access. All input images are resized to 640×640 pixels for model training, which significantly accelerates training and reduces memory consumption – particularly crucial under limited GPU memory conditions. The models undergo 150 training epochs, with each epoch representing a complete pass through the training data. The batch size maintains 1 image per iteration.

Mean Average Precision (mAP) quantifies detection accuracy by averaging precision scores across various confidence thresholds. This metric provides comprehensive evaluation of precision-recall trade-offs by averaging AP scores across all categories. The AP calculation follows this definition:

$$AP = \int_0^1 P(R) dR \quad (7)$$

mAP is defined as:

$$mAP = \frac{1}{N} \sum_{n=1}^N AP_n \quad (8)$$

We use mean Average Precision (mAP) as the evaluation metric for the proposed method. The architecture consists of three main components: a backbone network, a head, and a detection decoder. Based on RT-DETR (Real-Time Detection Transformer) framework, it employs Transformer mechanisms to enhance global context understanding while maintaining efficient feature extraction through CNN (Convolutional Neural Network) and Depthwise Separable Convolution (DWConv) designs, achieving efficient and accurate multi-scale object detection. The model outputs 80 object categories. At each stage, depthwise convolution downsampling reduces spatial resolution to extract more abstract features, ultimately decreasing feature map resolution to P4/32 for obtaining high-level global information. The head transforms features from the backbone into final detection outputs including object categories, bounding boxes and related information. The complete model contains 747 layers with a computational capacity of 97.4 GFLOPS (97.4 billion floating-point operations per second).

In comparative experiments, we retrain and test publicly available models of other algorithms using their default settings. As shown in [Table 1](#), compared with object detection algorithms including YOLOv51, YOLObile, PicoDet-L, YOLOX-s, lite-YOLOv5,

Class	YOLOW51	YOLObile	PicoDet-L	YOLOX-s	RT-DETR-l	YOLO v7 -tiny	LSK net	YOLOv8n	YOLOv10	YOLO v11	Ours
aircraft	94.20	90.00	96.30	94.20	98.00	93.80	97.00	97.3	97.3	96.9	98.40
play-ground	98.40	92.70	99.00	98.30	99.50	98.90	99.00	99.4	98.3	99.5	99.20
overpass	81.00	71.10	83.90	78.10	88.50	72.70	77.10	86.8	76.3	87.8	94.40
oiltank	97.10	90.20	96.80	96.10	97.20	96.69	98.30	98.7	98.1	98.6	98.00
map	92.30	83.20	93.80	91.80	94.60	90.60	92.90	95.5	92.5	95.7	97.50

Class	YOLOW51	YOLObile	PicoDet-L	YOLOX-s	RT-DETR-l	YOLO v7 -tiny	LSK net	YOLOv8n	YOLOv10	YOLO v11	Ours
aircraft	94.20	90.00	96.30	94.20	98.00	93.80	97.00	97.3	97.3	96.9	98.40
play-ground	98.40	92.70	99.00	98.30	99.50	98.90	99.00	99.4	98.3	99.5	99.20
overpass	81.00	71.10	83.90	78.10	88.50	72.70	77.10	86.8	76.3	87.8	94.40
oiltank	97.10	90.20	96.80	96.10	97.20	96.69	98.30	98.7	98.1	98.6	98.00
map	92.30	83.20	93.80	91.80	94.60	90.60	92.90	95.5	92.5	95.7	97.50

redetr-l, YOLOv7-tiny, YOLOv8n, YOLOv10, YOLOv11 and LSKnet, the proposed method demonstrates superior mAP performance across all datasets. Compared to YOLO-based algorithms, our method achieves mAP improvements ranging from 5.2% to 9.4%, with 2.9% accuracy gain over the original network. Table 2 shows comparison results with other methods including Faster RCNN, SSD, RFBNet, RetainNet, YOLOv3, YOLOv3-ASFF, EfficientDet, CSFF and CF2PN. Our method obtains the best mAP improvement across all datasets, showing performance gains from 0.4% to 24.69% compared to YOLO-based algorithms while maintaining the 2.9% accuracy improvement over the baseline network.

It is illustrated with experimental data of rosd dataset. Through experiments, we verify that the preprocessing time for a single image is 0.3 ms, the inference time is 33.7 ms, and the post-processing time is 0.2 ms, with the total processing delay as low as 34.2 ms. Calculated using the frames per second (FPS) formula, $FPS \approx 1/0.0342 \approx 29.24$, which meets the requirements for real-time applications. After verification, our experimental data includes key metrics. The model's floating – point operations (FLOPs) is 97.4 GFLOPs. The GPU used is the NVIDIA GeForce RTX 4060 Laptop GPU, and its video memory size is 8188 MiB. For FPS (frames per second), we have conducted meticulous experimental tests. After multiple runs on the ROSD and DIOR datasets and taking the average, the FPS value of our method is faster compared to the baseline methods. In terms of latency, our latency is lower than that of the baseline methods, which also reflects our advantage in improving real-time performance from another perspective. Regarding FLOPs (floating-point operations), the FLOPs of our method are such that when the model performs a complete forward calculation, it needs to execute approximately 97.4 billion floating-point operations. Compared with the baseline methods, we have certain advantages in computational efficiency. In terms of the number of parameters, we have carefully counted the number of parameters of the entire model. While our module realizes its functions, and in order to enable better comparison and ensure a certain level of accuracy, its number of parameters is larger than that of some baseline methods. However, this also ensures the flexibility of our model in subsequent expansion and comparison operations.

The chart comparing our average precision (MAP) scores is shown in Figure 8.

The average score (MAP) of the average accuracy rates for different disinfection methods is shown in Figure 9 and the comparison between the original method and our method shows that the accuracy advantage of our method is clearly demonstrated in Figure 10.

We select pictures from the ROSD dataset as illustrative examples. These pictures encompass typical complex scenarios, including small targets, crowded objects, and cross-domain migration. By comparing the results shown on the left and right sides of these pictures, our network exhibits remarkable advantages. It is quite evident that when handling crowded objects, our network can effectively segment the targets by leveraging its precise bounding box generation ability. This, in turn, enhances the recognition and localization accuracy and successfully averts false detections and confusions. When it comes to detecting small targets, our network has the capacity to strengthen the feature extraction and localization capabilities, thereby increasing both the detection success rate and accuracy. Specifically, each small target in every picture witnesses an average improvement of 0.2. In the face of complex situations like cross-domain migration, our network can stably detect targets under diverse image conditions, generate precise

Table 2. Mean average precision (MAP) scores for various Diior methods, where BC represents the basketball court, ESA represents the highway service area, ETS represents the highway toll station, and GTF represents the track venue. In addition, the best entry for each obstacle category is shown in bold.

Categories	RT-DETR		Faster RCNN	SSD	RFB		Retain Net	YOLOv3	YOLOv3-ASFF	EfficientDet	CF2PN	Ours
	-l				Net							
Airplane	86.3	36.9	67.42	77.97	68.63	53.92	91.86	71.18	78.32	88.3		
Airport	61.5	57.34	61.33	76.89	70.95	57.86	71.11	69.51	78.29	62.7		
Baseball Field	86.0	62.44	68.23	74.15	73	64.84	74.54	67.97	76.48	87.4		
BC	76.4	75.63	85.94	88.71	89.42	79.19	89.34	86.72	88.4	80.0		
Bridge	35.9	18.36	25.63	34.35	32.49	30.08	33.33	38.47	37	41.9		
Chimney	81.3	72.60	76.63	78.84	76.79	68.52	75.77	76.06	70.95	84.8		
Dam	42.8	42.69	51.00	58.39	65.74	47.83	43.13	58.18	59.9	49.4		
ESA	57.4	52.30	59.77	72.19	85.71	56.63	58.12	58.27	71.23	45.9		
EST	54.9	43.90	46.11	54.34	55.13	47.13	57.28	55.15	51.15	63.5		
Golf Course	70.3	63.49	73.39	80.41	82.11	59.30	73.81	73.26	75.55	68.1		
GTF	69.7	48.68	63.06	71.48	80.82	54.38	46.42	69.28	77.14	70.8		
Harbor	41.0	29.22	51.72	61.26	54.31	48.28	57.11	59.00	56.75	48.6		
Overpass	50.2	43.27	50.89	57.15	55.26	46.95	56.21	54.72	58.65	57.0		
Ship	83.8	12.62	54.69	76.3	43.58	76.95	75.11	85.42	76.06	86.5		
Stadium	81.2	38.15	67.45	79.75	81.62	40.54	37.92	42.49	70.61	76.8		
Storage Tank	74.2	17.16	39.76	51.42	36.26	54.18	77.33	67.34	55.52	75.9		
Tennis Court	83.5	65.86	82.41	86.81	86.70	76.58	89.21	81.32	88.84	85.5		
Train Station	32.8	41.47	43.60	61.57	48.07	42.56	49.51	48.14	50.83	47.2		
Vehicle	68.7	11.39	25.89	35.5	20.11	33.84	52.74	45.86	36.89	71.9		
Windmill	72.3	41.33	63.26	74.61	67.45	67.10	82.00	75.21	86.36	70.0		
mAP	65.5	43.71	57.91	67.61	63.71	55.33	64.59	64.18	67.25	68.4		

bounding boxes, and make dependable judgements regarding confidence levels. Moreover, it can enhance its adaptability while maintaining high accuracy without any alterations. All in all, our network has significantly boosted its performance in complex situations and has correspondingly enhanced the detection accuracy and adaptability. Compared with RT – DETR – L, our method showcases significant advantages in the realm of object detection. In terms of detection accuracy, our method can precisely lock onto objects with a relatively high level of precision, substantially reducing the occurrence of errors and guaranteeing the reliability of detection results. Meanwhile, concerning the coverage of object recognition, our method demonstrates a stronger capability and can identify multiple types of objects that pose difficulties for RT – DETR – L to detect. Consequently, our method attains a rather obvious superiority in overall object detection capabilities, offering better and more comprehensive detection support for relevant applications. From the set of comparison pictures between RT – DETR – L and our method, it can be clearly perceived that the boundary regression module BR plays a crucial role in improving network accuracy. In the aspect of target localization, taking aircraft as an instance, the bounding boxes generated by the method incorporating the BR module can adhere to the actual contours of aircraft more closely compared to those of RT – DETR – L. It can determine the positions of aircraft more accurately, and the confidence annotations are also more consistent with the actual situation, which significantly decreases the probabilities of false detection and missed detection. In multi – target recognition scenarios, for pictures containing multiple aircraft targets, the BR module empowers our method not only to identify a greater number of aircraft targets but also to ensure that the bounding boxes corresponding to each target precisely define the scope of the aircraft. Additionally, the confidence annotations are more reasonable. However, in such scenarios, RT – DETR – L often encounters issues like inaccurate positioning of the bounding boxes for some aircraft targets and even fails to detect certain targets. Besides, for non – aircraft targets such as oil tanks (e.g. ‘oiltank’) and overpasses (e.g. ‘overpass’), the network equipped with the BR module also performs outstandingly. The generated bounding boxes can accurately enclose these targets, and the confidence scores assigned are also reasonable. This comprehensively reflects that the BR module effectively enhances the detection accuracy and generalization ability of the network for different types of targets. In contrast, RT – DETR – L is evidently inferior to our method with the BR module when it comes to recognizing these non – main detection targets.

4.2. B.Ablation experiments

We conduct ablation studies to evaluate the detection performance of our proposed RMRN-DETR algorithm, examining the impact of various algorithmic components for continuous optimization. Table 2 presents experimental results on the RSOD dataset with different algorithm enhancements, using evaluation metrics including mean Average Precision (mAP) and Recall as shown in Table 3.

In Method (1), we incorporate GFPN neck to optimize and fuse high-level semantic features with low-level spatial features. The Fusion Block implements a composite structure of Batch Normalization (BN) and Activation function (Act) through a ‘Simplified Rep 3x3’ design – using 3×3 convolution during training while simplifying to 1×1

Table 3. Results of ablation experiments on the RSOD dataset.

Method	MR	MA	BR	mAP/%	Recall/%	mAP50-95)%
RT-DETR-I				94.6	94.4	68.4
Methods(1)	✓			95.5(+0.9)	91(−3.4)	71.1(+2.7)
Methods(2)		✓		96.7(+2.1)	91.3(−3.1)	71.4(+3.0)
Methods(3)			✓	95.9(+1.3)	90.4(−4)	68(−0.4)
Methods(4)		✓	✓	97.2(+2.6)	94.8(+0.4)	71.4(+3)
Methods(5)	✓		✓	97.2(+2.6)	93.3(−1.1)	71(+2.6)
Methods(6)	✓	✓		96.4(+1.8)	94.3(−0.1)	73.3(+4.9)
Methods(7)	✓	✓	✓	97.5(+2.9)	95.9(+1.5)	71.6(+3.2)

convolution during inference for efficiency. This modification improves model accuracy by 0.9% but slightly reduces Recall.

Method (2) introduces the Multi-Scale Dilated Attention (MSDA) module, which obtains corresponding queries, keys and values through linear projection of feature map X. The feature channels are divided into n different heads, each performing multi-scale SWDA with varying dilation rates. All outputs are concatenated and fed into a linear layer for feature aggregation. This implementation further reduces parameter count compared to the previous network, achieving 0.9% mAP improvement and 3.3% Recall increase, though Recall remains below baseline values.

Method (3) implements a novel bounding box similarity metric MPDIoU, calculating minimum point distance between horizontal rectangles. This approach comprehensively considers overlap area, centre point distance, and width/height deviations, effectively distinguishing boxes with identical aspect ratios but different sizes/positions through direct keypoint distance computation. Compared to the previous step, it delivers 1.1% mAP improvement and 1.6% Recall increase while providing more precise loss measurement and significant speed enhancement. Overall, the complete solution achieves 1.8% mAP and 1.5% Recall improvements over the original network.

4.3. C.Confusion matrix

With the rapid development of machine learning and artificial intelligence, we require more detailed and systematic methods to understand model classification capabilities and their performance across different categories.

Confusion matrix analysis serves as a method in machine learning and statistics for evaluating classification model performance. It provides a comprehensive summary of classification results, applicable to supervised learning classification problems. The matrix presents sample classification outcomes in tabular form, comparing model predictions against actual labels.

In a confusion matrix, rows represent the true classes of samples, while columns indicate predicted classes. This structured representation enables precise assessment of true positives, false positives, true negatives, and false negatives, facilitating performance metric calculations such as accuracy, precision, recall, and F1-score. The analysis supports both binary and multi-class classification evaluation, with each cell containing counts of predictions matching specific true-predicted class pairs.

True Positives (TP) indicate the number of instances where the model correctly predicts positive cases as positive. False Negatives (FN) represent cases where the model

incorrectly predicts positive instances as negative. False Positives (FP) denote instances where the model wrongly predicts negative cases as positive. True Negatives (TN) count cases where the model accurately predicts negative instances as negative. These four fundamental components form the basis for calculating all subsequent classification performance metrics. The matrix structure allows direct visualization of correct classifications versus different types of errors across all target categories. Each cell value corresponds to absolute counts rather than percentages or normalized values. This tabular representation enables immediate identification of model strengths and weaknesses in class discrimination.

Precision is how many of the positive predictions are correct and is defined as follows:

$$\text{Precision} = TP / (TP + FP) \quad (9)$$

Recall (Recal) is defined as how many of the true positive samples are correctly predicted as positive and is defined as follows:

$$\text{Recall} = TP / (TP + FN) \quad (10)$$

The F1-Score is defined as a combination of precision and recall and is defined as follows:

$$F1 = 2 * (\text{Precision} * \text{Recal}) / (\text{Precision} + \text{Recal}) \quad (11)$$

Accuracy is defined as the proportion of all samples that the model correctly predicts and is defined as follows:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (12)$$

The four basic components of the confusion matrix are, from top to bottom and left to right: True Positives (TP), True Negatives (TN), False Positives (FP) 和 False Negatives (FN). In [Figure 11](#), we present the confusion matrix.

In the confusion matrix analysis using the ROSD dataset, aircraft detection achieves 97.8% precision, 87.4% recall, 92.3% F1-score, and 85.7% accuracy. Playground detection shows 100% precision, 66.7% recall, 80% F1-score, and 66.7% accuracy. Overpass detection demonstrates 94.7% precision, 72% recall, 82% F1-score, and 69.2% accuracy. Oil tank detection attains 98.6% precision, 84.2% recall, 90.8% F1-score, and 83.2% accuracy. The experimental results confirm our algorithm's strong performance in object detection accuracy for remote sensing images with complex backgrounds, particularly showing robust detection capability across different target categories with varying characteristics. The comprehensive metrics indicate balanced performance between precision and recall for most object classes, with particularly outstanding results in aircraft and oil tank detection scenarios. These quantitative measurements validate the method's effectiveness in practical remote sensing applications requiring high-precision target identification.

4.4. D.Curve analysis of experimental results

Next, we visually show the superiority of our method by comparing the precision recall curve and the precision curve by the ablation experiment.

When using only the MR and MA Module, Precision increases by 1.8 and recall decreases by 0.1. The precision and recall curves as well as the precision confidence curve are shown in Figure 12.

When using only the MR and BR Module, Precision increases by 2.6 and recall decreases by 1.1. The precision and recall curves as well as the precision confidence curve are shown in Figure 13.

When using only the MA and BR Module, Precision increases by 2.6 and recall increases by 0.4. The precision and recall curves as well as the precision confidence curve are shown in Figure 14.

Taking the ROSD dataset as an example, Figure 14 shows the Precision-Recall Curve commonly used to evaluate object detection model performance. The x-axis represents Recall, indicating the proportion of correctly detected positive samples among all actual positives. The y-axis shows Precision, denoting the ratio of correctly detected positives to all samples predicted as positive. The legend displays Precision-Recall curves for different categories: 'aircraft' (AP = 0.984), 'playground' (AP = 0.992), 'overpass' (AP = 0.944), 'oil-tank' (AP = 0.980), along with the mean Average Precision (mAP) curve for all categories. At a 0.5 IoU threshold, the mAP reaches 0.975 across all classes. Curves closer to the top-right corner indicate better model performance for that category. The 'playground' curve

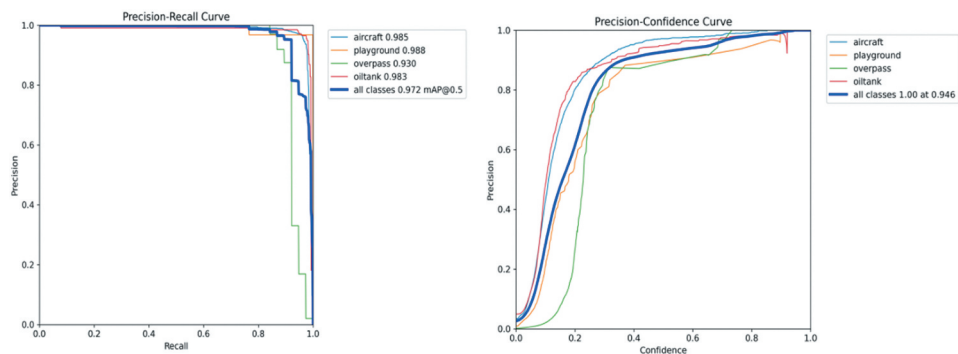


Figure 14. Precision and recall curves and precision confidence curves are only available for MA and BR.

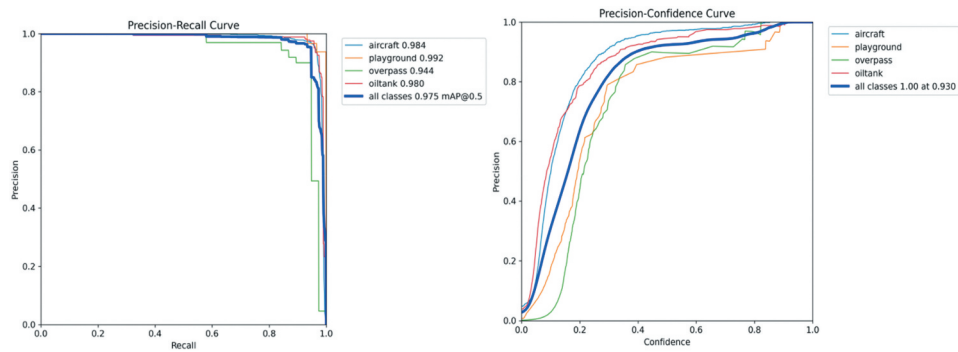


Figure 15. Precision-recall curve precision-confidence curve diagram.

appears nearest to the ideal corner, demonstrating superior balance between precision and recall, while the overall mAP curve confirms the model's high detection capability.

Figure 15 presents the Precision-Confidence Curve for evaluating object detection model performance. The x-axis represents Confidence, indicating the model's certainty in detection results (range 0–1), while the y-axis shows Precision, measuring the proportion of true positives among all predicted positives (range 0–1). The legend contains four coloured curves representing different categories: 'aircraft', 'playground', 'overpass', and 'oiltank', plus a bold blue curve labelled 'all classes 1.00 at 0.930' indicating perfect precision achieved at 0.930 confidence threshold for combined categories. Curves closer to the top-right corner demonstrate better model performance, maintaining high precision at elevated confidence levels. The 'all classes' curve maintains consistently high precision across most confidence values, confirming robust detection reliability. All category-specific curves show stable precision above 0.9 confidence thresholds, with 'playground' exhibiting the most optimal performance pattern. This visualization effectively demonstrates the model's capability to balance detection confidence with prediction accuracy across diverse object categories.

These 8 figures (focusing on Precision-Recall curves and Precision-Confidence curves) clearly present the differences and advantages in object detection performance through comparisons of different module combinations and verification of a specific model. At the module combination level, the curves for the MR & MA combination show no specific fluctuations in precision and recall. In contrast, the MR & BR combination can increase precision by 2.6 but causes a 1.1 decrease in recall; its Precision-Recall curve also shows a more obvious precision decline in the high-recall region, and the category distribution of its Precision-Confidence curve is more scattered. The MA & BR combination, however, demonstrates prominent advantages. It not only increases precision by 2.6 as well but also raises recall by 0.4. Moreover, its Precision-Recall curve maintains precision more stably as recall increases, and its Precision-Confidence curve shows a faster precision rise and a higher matching degree with confidence when confidence improves. At the model performance level, the curves related to the RMRN-DETR model based on the ROSD dataset have even more distinct advantages. All categories have relatively high AP values, among which the 'playground' category reaches an AP value of 0.992, and its Precision-Recall curve is closest to the upper-right corner. Meanwhile, the Precision-Confidence curve for 'all classes' achieves a perfect precision of 1.00 when the confidence is 0.930, and the curves for all categories maintain stable precision when the confidence is above 0.9. These facts fully reflect the model's excellent ability to balance detection confidence and prediction precision in multi-category object detection.

5. Conclusion

This paper proposes RMRN-DETR, a regression-optimized remote sensing image detection network based on multi-dimensional real-time detection and domain adaptation. The MR effectively improves bounding box regression accuracy and demonstrates superior performance in multi-scale object detection. The MA further enhances the model's adaptability to targets at different scales, particularly improving detection accuracy in complex scenarios. Additionally, the loss boundary regression module significantly refines bounding box localization precision through pixel-level optimization. The experimental results show that on the ROSD dataset, our method achieves an mAP improvement of 5.2%–9.4%

compared to other approaches, with a 1.8% accuracy gain over the original network. On the DIOR dataset, it demonstrates an mAP increase of 0.4%–24.69% against competing methods, outperforming the baseline by 2.9%. These results highlight its superior performance in object detection tasks, proving its effectiveness in handling complex backgrounds and multi-scale object detection. Future work will focus on model lightweighting to enhance practical deployment efficiency and adaptability. Through network structure optimization and efficient parameter compression techniques, the model will maintain high accuracy while significantly reducing computational requirements for diverse real-world applications.

Disclosure statement

The authors declare that there is no conflict of interests with respect to the research, authorship, and publication of this article.

Data availability statement

The DIOR dataset is available at <http://www.escience.cn/people/gongcheng/DIOR.html>. For related published papers on the DIOR dataset, refer to 'Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark'.

The RSOD dataset can be downloaded from <https://github.com/RSODataset>. Published papers associated with the RSOD dataset are accessible at <https://ieeexplore.ieee.org/abstract/document/7827088>.

References

- Bokhovkin, A., and E. Burnaev. 2019. "Boundary Loss for Remote Sensing Imagery Semantic Segmentation[J]." *Springer, Cham*, <https://doi.org/10.1007/978-3-030-22808-8-38>.
- Chang, J., X. He, P. Li, et al. 2024. "Multi-Scale Attention Network for Building Extraction from High-Resolution Remote Sensing Images." *Sensors* 24 (3): 1010. <https://doi.org/10.3390/s24031010>.
- Ci, J., H. Tan, H. Zhai, et al. 2024. "Radiation Anomaly Detection of Sub-Band Optical Remote Sensing Images Based on Multiscale Deep Dynamic Fusion and Adaptive Optimization." *Remote Sensing* 16 (16): 2953. <https://doi.org/10.3390/rs16162953>.
- Dong, R., L. Jiao, Y. Zhang, et al. 2021. "A Multi-Scale Spatial Attention Region Proposal Network for High-Resolution Optical Remote Sensing Imagery." *Remote Sensing* 13 (17): 3362. <https://doi.org/10.3390/rs13173362>.
- Du, S., B. Zhang, and P. Zhang. 2021. "Scale-Sensitive IOU Loss: An Improved Regression Loss Function in Remote Sensing Object Detection." *IEEE Access* 9:141258–141272. <https://doi.org/10.1109/ACCESS.2021.3119562>.
- Guo, H., H. Wang, X. Song, et al. 2023. "Anomaly Detection of Remote Sensing Images Based on the Channel Attention Mechanism and LRX." *Applied Sciences* 13 (12): 6988. <https://doi.org/10.3390/app13126988>.
- Han, J., W. Yang, Y. Wang, L. Chen, and Z. Luo. 2024. "Remote Sensing Teacher: Cross-Domain Detection Transformer with Learnable Frequency-Enhanced Feature Alignment in Remote Sensing Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–14. <https://doi.org/10.1109/TGRS.2024.3378284>.
- He, P., X. Zhao, Y. Shi, et al. 2021. "Unsupervised Change Detection from Remotely Sensed Images Based on Multi-Scale Visual Saliency Coarse-To-Fine Fusion." *Remote Sensing* 13 (4): 630. <https://doi.org/10.3390/rs13040630>.

- Huang, Y., and G. Yuan. 2023. "AD-DETR: DETR with asymmetrical relation and decoupled attention in crowded scenes[J]." *Mathematical Biosciences and Engineering* 20 (8). <https://doi.org/10.3934/mbe.2023633>.
- Huang, Z., and H. You. 2023. "MFSFNet: Multi-Scale Feature Subtraction Fusion Network for Remote Sensing Image Change Detection." *Remote Sensing* 15 (15): 3740. <https://doi.org/10.3390/rs15153740>.
- Kong, Y., X. Shang, and S. Jia. 2024. "Drone-DETR: Efficient Small Object Detection for Remote Sensing Image Using Enhanced rt-Detr Model." *Sensors* 24 (17): 5496. <https://doi.org/10.3390/s24175496>.
- Li, H., E. Tian, W. Zhang, et al. 2024. "Improving Remote Sensing Object Detection by Using Feature Extraction and Rotational Equivariant Attention." *International Journal of Remote Sensing* 45 (11): 3789–3806. <https://doi.org/10.1080/01431161.2024.2354132>.
- Liu, M., H. Wang, L. Du, et al. 2024. "Bearing-Detr: A Lightweight Deep Learning Model for Bearing Defect Detection Based on Rt-Detr." *Sensors* 24 (13): 4262. <https://doi.org/10.3390/s24134262>.
- Liu, Z., C. Sun, and X. Wang. 2024. "dst-Detr: Image Dehazing rt-Detr for Safety Helmet Detection in Foggy Weather." *Sensors* 24 (14): 4628. <https://doi.org/10.3390/s24144628>.
- Luo, F., Y. Dai, J. Fuentes, et al. 2024. "m-Detr: Multi-Scale DETR for Optical Music Recognition." *Expert Systems with Applications* 249:123664. <https://doi.org/10.1016/j.eswa.2024.123664>.
- Qu, J., Z. Tang, L. Zhang, et al. 2023. "Remote Sensing Small Object Detection Network Based on Attention Mechanism and Multi-Scale Feature Fusion." *Remote Sensing* 15 (11): 2728. <https://doi.org/10.3390/rs15112728>.
- Shang, R., J. Zhang, L. Jiao, et al. 2020. "Multi-Scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images." *Remote Sensing* 12 (5): 872. <https://doi.org/10.3390/rs12050872>.
- Shen, C., J. Qian, C. Wang, D. Yan, and C. Zhong. 2024. "Dynamic Sensing and Correlation Loss Detector for Small Object Detection in Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–12. <https://doi.org/10.1109/TGRS.2024.3407858>.
- Wang, Y., M. Wang, Z. Hao, et al. 2024. "MSGFNet: Multi-Scale Gated Fusion Network for Remote Sensing Image Change Detection." *Remote Sensing* 16 (3): 572. <https://doi.org/10.3390/rs16030572>.
- Wen, G., P. Cao, H. Wang, et al. 2023. "ms-Ssd: Multi-Scale Single Shot Detector for Ship Detection in Remote Sensing Images." *Applied Intelligence* 53 (2): 1586–1604. <https://doi.org/10.1007/s10489-022-03549-6>.
- Xie, Q., D. Zhou, R. Tang, and H. Feng. 2024. "A Deep CNN-Based Detection Method for Multi-Scale Fine-Grained Objects in Remote Sensing Images." *IEEE Access* 12:15622–15630. <https://doi.org/10.1109/ACCESS.2024.3356716>.
- Yang, Y., C. Wang, Z. Cai, P. Song, G. Huang, M. Cheng, and Y. Zang. 2023. "GSDDet: Ground Sample Distance Guided Object Detection for Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–12. <https://doi.org/10.1109/TGRS.2023.3309838>.
- Yuan, W., and W. Xu. 2021. "Neighborloss: A Loss Function Considering Spatial Correlation for Semantic Segmentation of Remote Sensing Image." *IEEE Access* 9:75641–75649. <https://doi.org/10.1109/ACCESS.2021.3082076>.
- Zhang, Z., and W. Zhu. 2024. "yolo-Mfd: Remote Sensing Image Object Detection with Multi-Scale Fusion Dynamic Head." *Computers, Materials & Continua* 79 (2). <https://doi.org/10.32604/cmc.2024.048755>.
- Zhao, Y., R. Yang, C. Guo, and X. Chen. 2024. "Parallel Space and Channel Attention for Stronger Remote Sensing Object Detection." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17:2610–2621. <https://doi.org/10.1109/JSTARS.2023.3347235>.
- Zhou, H., W. Liu, K. Sun, et al. 2024. "MSCANet: A multi-scale context-aware network for remote sensing object detection[J]." *Earth Science Informatics*, 17 (6). <https://doi.org/10.1007/s12145-024-01447-8>