Article

Not peer-reviewed version

SO-RTDETR for Small Object Detection in Aerial Images

Jing Liu , Yanyan Cao * , Ying Wang * , Chaoping Guo , Haijing Zhang , Chunyu Dong

Posted Date: 25 October 2024

doi: 10.20944/preprints202410.1980.v1

Keywords: Small Object Detection; Convolutional Neural Network; Feature Extraction; RT-DETR; Aerial Images

Article

SO-RTDETR for Small Object Detection in Aerial Images

Jing Liu [1], Yanyan Cao [1,*], Ying Wang[2,*], Chaoping Guo [1], Haijing Zhang[1],Chunyu Dong [1]

[1] Xi'an Key Laboratory of Human-Machine Integration and Control Technology for Intelligent Rehabilitation, School of Computer Science, Xijing University, Xi'an 710123, China; 20180075@xijing.edu.cn (J.L.); 20190222@xijing.edu.cn (C.G.); 20200060@xijing.edu.cn (H.Z.); 20230037@xijing.edu.cn (C.D.);

[2] School of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430072, China; wangying@wsyu.edu.cn

* Correspondence: 20070077@xijing.edu.cn (Y.C.), wangying@wsyu.edu.cn(Y.W.)

Abstract: In aerial image object detection, small targets present significant challenges due to limited pixel information, complex backgrounds, and sensitivity to bounding box perturbations. To tackle these issues, we propose SO-RTDETR for small object detection. The model introduces a Cross-Scale Feature Fusion with S2 (S2-CCFF) module, a Parallelized Patch-Aware attention (PPA) module, and the Normalized Wasserstein Distance (NWD) loss function, leading to significant performance improvements. Specifically, the S2-CCFF module enhances small object information by incorporating an additional S2 layer, while SPDConv downsampling maintains key details and reduces computational cost. The CSPOK-Fusion mechanism integrates global, local, and large branch features, capturing multi-scale representations and effectively mitigating interference from complex backgrounds and occlusions, thereby enhancing the spatial representation of features across scales. The PPA module, embedded in the Backbone network, leverages multi-level feature fusion and attention mechanisms to retain and strengthen small object features, addressing the issue of information loss. The NWD loss function, by focusing on the relative positioning and shape differences of bounding boxes, increases robustness to minor perturbations, enhancing detection accuracy. Experimental results on the VisDrone and NWPU VHR-10 aerial datasets demonstrate that our approach outperforms state-of-the-art detectors.

Keywords: Small Object Detection; Convolutional Neural Network; Feature Extraction; RT-DETR; Aerial Images

1. Introduction

In aerial image processing, object detection plays a crucial role and is widely applied in geographic information systems, urban planning, environmental monitoring, disaster assessment, and other fields. However, due to the wide field of view in aerial images, the size of the resulting images is often larger than typical images, leading to a significant imbalance between foreground and background information. Many objects appear as small targets, characterized by limited pixel information, small object area, indistinct texture features, and low contrast with the background. Traditional detection methods struggle with these challenges, often facing limitations such as insufficient detection accuracy and loss of fine details.

As deep learning and convolutional neural networks(CNN) continue to evolve rapidly in recent years, object detection models utilizing the Transformer [1] architecture, such as DETR (Detection Transformer) [2], have made remarkable progress in detection tasks. By transforming object detection into an end-to-end sequence modeling problem, DETR eliminates the need for region proposals used in traditional detectors. Compared to R-CNN [3] and YOLO [4–9] methods, DETR avoids the complexity of post-processing and hyperparameter tuning by using query vectors as soft anchors instead of predefined anchor boxes for target localization. However, this design results in slow convergence and requires extended training time. To address this, researchers have proposed various improvements,

Citation: Liu, J.; Cao,Y.; Wang, Y.; Guo,C.; Zhang, H.; Dong, C.;. SO-RTDETR for Small Object Detection in Aerial Images. *Entropy* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

such as Deformable DETR [10], and the introduction of algorithms like RT-DETR [11] marks a maturation of DETR-based methods. These approaches have improved small object detection by incorporating multi-scale feature extraction modules, contextual information fusion strategies, and advanced bounding box regression mechanisms. Nevertheless, challenges such as complex backgrounds with occlusions, loss of object information, and the low tolerance to bounding box perturbations in aerial images still need to be addressed for small object detection [12].

Complex Backgrounds and Occlusions:Aerial images have a wide perspective and a large number of small targets present. For example, Figure 1 (a) shows a square scene from the VisDrone UAV dataset, containing numerous tiny objects. Additionally, substantial background information, including vegetation and buildings, is present. Figure 1 (b) illustrates small cars obscured by trees. The features of small objects are easily affected by background information or other disturbances, introducing noise into the learned feature representations. This noise weakens the depiction of small object features and hinders the model's learning and accurate prediction.

Loss of Details Regarding Small Objects: In deep learning algorithms, CNN networks typically construct multi-scale feature pyramids, which progressively decrease the spatial resolution of feature maps. While this helps retain most of the critical information, some object information is inevitably lost. For medium and large objects, this loss generally does not significantly affect detection performance, as their features remain prominent. However, for small objects, this loss severely impacts detection [13]. The features of small objects become weak on highly compressed feature maps, and their proximity to each other, as well as potential confusion with the background or other objects, increases detection complexity. As shown in Figure 1 (c), it difficult to make accurate predictions from these sparse and incomplete representations.

Low Tolerance to Bounding Box Perturbations: Localization is a fundamental aspect of object detection, typically achieved through bounding box regression. Intersection over Union (IoU) is a common metric for evaluating regression performance. Compared to normal objects, small objects are highly sensitive to slight shifts in their bounding boxes. As illustrated in Figure 1 (d) and Figure 1 (e), small objects (6x6) and normal objects (36x36) are shown. Box A indicates the ground truth (GT), whereas boxes B and C illustrate predicted boxes with slight diagonal shifts of 1 pixel and 2 pixels, respectively. The IoU assesses the overlap between the GT box and the predicted boxes. A 1-pixel shift for the small object reduces the IoU to 0.53, while a 2-pixel shift further decreases it to 0.14. In contrast, the IoU for the normal object remains relatively stable, changing from 0.90 to 0.72 under similar perturbations. This demonstrates that small objects exhibit significantly lower tolerance to bounding box regression errors compared to normal objects, complicating the model's learning in the regression branch.

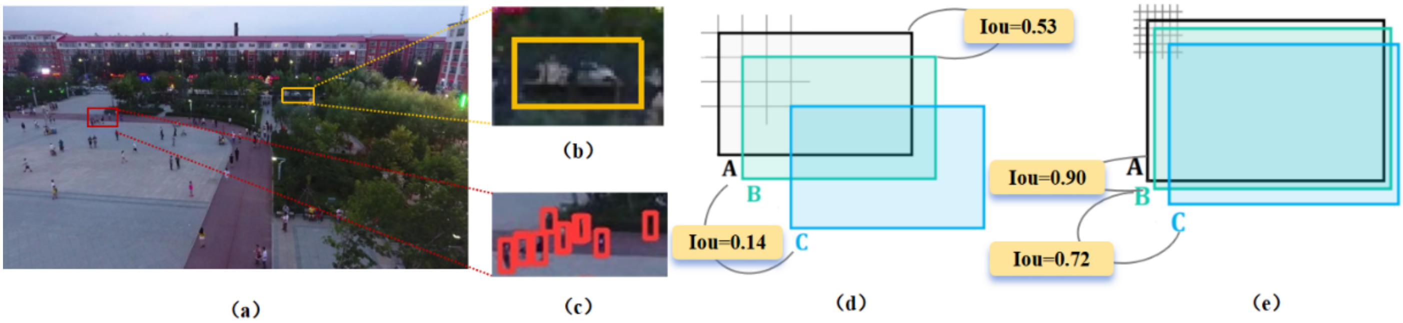


Figure 1. Challenges in small object detection. (a) Image from the VisDrone dataset; (b) Small object affected by noise; (c) Feature loss during small object detection; (d) Low tolerance to small object bounding box perturbations; (e) Minimal impact of bounding box perturbations on normal objects.

This paper presents SO-RTDETR, a model specifically designed for small object detection to tackle these challenges. The model integrates a Cross-Scale Feature Fusion with S2 (S2-CCFF) module, a Parallelized Patch-Aware attention (PPA) module, and the Normalized Wasserstein Distance (NWD) loss function. Unlike traditional methods, SO-RTDETR effectively captures small object features, minimizes background noise, enhances cross-scale feature fusion, and lessens the impact of bounding box perturbations, ultimately leading to improved detection accuracy and efficiency. The key contributions of this study include:

- The S2-CCFF module is proposed, where an S2 layer is added during cross-scale feature fusion to enrich small object information. To mitigate information loss caused by conventional downsampling, spatial downsampling is performed on the S2 layer using SPDConv, preserving key details while reducing computational complexity. Additionally, the CSPOK-Fusion module is designed to integrate multi-scale features across global, local, and large branches, effectively suppressing noise, capturing feature representations from global to local scales, and addressing complex background interference and occlusion issues, thereby improving detection accuracy.
- The PPA module is incorporated into the Backbone network, employing multi-level feature fusion and attention mechanisms to preserve and enhance small object representations. This ensures key information is retained across multiple downsampling stages, effectively mitigating small object information loss and improving subsequent detection accuracy.
- To address the low tolerance of bounding box perturbations, we introduce the NWD [14] loss function, which better captures differences in the relative position, shape, and size of bounding boxes. It focuses on the relative positional relationships between boxes rather than merely relying on overlap. This approach offers greater tolerance to minor bounding box perturbations.

The paper is structured as follows: Section 2 provides an in-depth overview of existing research on small object detection in aerial imagery. Section 3 introduces the architecture of the proposed model. In Section 4, we present ablation studies to evaluate the contribution of each module, alongside comparative experiments and visual performance analyses. Lastly, Section 5 concludes the work and discusses potential directions for future research.

2. Related work

The detection of small objects holds a pivotal position in the field of computer vision, particularly in high-resolution images with complex backgrounds. Small objects occupy fewer pixels, making them more susceptible to background noise and challenging to extract features from. To effectively address these challenges, researchers have proposed various approaches, including improved feature extraction mechanisms, multi-scale feature fusion, and the introduction of novel loss functions. This section reviews mainstream object detection algorithms and the latest advances in small object detection.

2.1. Convolutional Neural Network-Based Detection Methods

Conventional CNN-based detection approaches are typically divided into two-stage and one-stage detectors. Two-stage detectors, like Faster RCNN [15], Mask RCNN [16], and Cascade RCNN [17], initially generate region proposals, followed by a refinement process to improve classification accuracy and object localization. Libra R-CNN [18] addresses the imbalance issue in small object detection by refining raw features through non-local blocks to obtain balanced interaction features. Cascade R-CNN [17] improves small object detection accuracy by progressively refining predictions through multi-stage regression. This method optimizes the boundary boxes and class information of small objects in a stepwise manner. While two-stage algorithms are well-known for their superior detection accuracy, they also face challenges related to speed, training complexity, and optimization.

In contrast, one-stage detectors bypass the need for region proposals by employing a single neural network to simultaneously predict object classes and bounding box coordinates. Prominent one-stage detectors include SSD [19], RetinaNet [20], and the YOLO series [4–9]. These networks are well-suited for tasks where speed is critical but often trade off some accuracy. Despite their faster performance, detecting small objects remains a challenge due to detail loss caused by downsampling.

Recognizing the advantages of multi-scale features in object detection, researchers have introduced the FPN [21], which enhances the performance of small object detection. FPN constructs a bottom-up feature pyramid, enabling the model to fuse multi-scale features at different resolutions, thus enhancing its ability to detect small objects. The Path Aggregation Network (PANet) [22] enhances the FPN by employing fewer convolutional layers in its path-enhancement module, thereby retaining more information from lower layers. Additionally, PANet incorporates an adaptive feature pooling module, allowing regions of interest to contain multi-layer features, further boosting small object detection performance. NAS-FPN [23] employed reinforcement learning to train a controller that identifies the optimal model architecture within a predefined search space. BiFPN [24] improves the balance between accuracy and efficiency by eliminating nodes that have only one input edge and introducing an additional edge connecting the original input directly to the output node when their levels align. This method treats each bidirectional pathway as a unique feature network layer.

Despite the improvements made by the above methods in detecting objects at different scales, they have not fully considered the extraction of low-level positional information and the subtle interactions within the context, leading to insufficient feature representation when handling small objects. CF2PN [25] utilized multi-level feature fusion techniques to address the challenges of low efficiency in multi-scale object detection in aerial images. AugFPN [26] addressed the inconsistency between detailed and semantic information in feature maps by adopting a one-time supervision approach during feature fusion to bridge the information gap and introduced ASF to dynamically integrate features across various scales.

AFPN [27] boosted detection performance by merging adjacent low-level features while progressively incorporating high-level features, minimizing semantic gaps between distant levels, and ensuring a more consistent scale distribution. Gong [28] tackled the poor performance of FPN in small object detection by employing fusion factor techniques, which optimized fusion weights through statistical analysis of object counts in the dataset, significantly improving detection performance. Gao [29] rotated FPN's high-level semantic features to four different degrees, concatenating them along the channel dimension before processing through convolution layers. This strengthened the interaction between different perspectives of high-level semantic features, further enhancing global semantic information. Hu [30] introduced adaptive hierarchical upsampling to generate FPN features, providing substantial semantic compensation for low-level features and preventing dilution noise caused by FPN fusion. DN-FPN [31] addressed the problem of noisy features arising from the lack of regularization between features of different scales during fusion by using contrastive learning to suppress noise in each layer of FPN's top-down pathway, thereby improving small object detection accuracy. CFPT [32] introduced a feature pyramid network designed without upsampling, tailored for small object detection in aerial imagery. By enhancing feature interaction and global information utilization through cross-layer channel and spatial attention mechanisms, this approach effectively prevented information loss and improved detection performance. To overcome the limitations of existing FPNs, which often neglect low-level positional information and fine-grained context interaction, LR-FPN [33] proposed a location-refined feature pyramid network to enhance shallow positional information extraction and promote fine-grained context interaction.

2.2. Transformer-Based Detection Methods

Recently, Transformer-based detection methods have gained significant attention. The Vision Transformer (ViT) [34] has shown notable potential across diverse visual tasks. However, due to the computational complexity of traditional ViTs concerning image resolution, research has shifted towards lightweight alternatives. The Swin Transformer [35] improves efficiency by restricting self-attention to non-overlapping local windows while incorporating a shifted window strategy, which facilitates connections between adjacent windows. Its hierarchical structure offers multi-scale modeling flexibility and ensures linear computational complexity with image size.

The Detection Transformer (DETR) [2] introduced an end-to-end detection model leveraging self-attention mechanisms. By eliminating anchor box generation, DETR simplifies the detection process, directly locating targets via query vectors. However, the limited quantity and quality of positive samples may hinder performance on exceedingly small objects. In response, several improvements, such as Deformable DETR [10], have been proposed, incorporating deformable convolutions and sparse attention mechanisms to enhance detection efficiency, particularly in small-object detection tasks. RT-DETR [11], an optimized variant of DETR, reduces the complexity of the Transformer structure and refines the feature extraction module, enabling real-time object detection. O2DETR [36] substitutes attention modules with local convolutions and integrates multi-scale feature maps, demonstrating improved detection performance in scenarios involving rotated objects. Huang [37] employed advanced sampling techniques, including Sample Points Refinement (SPR), Scale-aligned Target (ST), and Task-decoupled Sample Reweighting (SR) mechanisms, to optimize positioning and attention distribution during the detection process. These methods enhance detection performance for small targets by constraining positioning, integrating scale information, and directing attention toward challenging samples.

Nevertheless, RT-DETR still faces challenges in effectively addressing small targets, primarily due to inadequate feature representation and suboptimal accuracy, particularly in aerial images where the low contrast between complex backgrounds and small objects exacerbates the issue.

2.3. Detection Methods for Small Objects in Aerial Images

ClusDet [38] adopts a coarse-to-fine strategy, starting with a Cluster Candidate Network (CPNet) to extract clustered regions, followed by a Scale Estimation Network (ScaleNet) to adjust region sizes. A detection network (DetecNet) then performs precise detection, improving small object recognition. DMNet [39] streamlines ClusDet's training by using a density map generation network to produce density maps for clustering predictions.

UFPMP-Det [40] employs a feature pyramid with multi-path aggregation and anchor optimization to predict sub-regions, which are combined for efficient inference, enhancing both detection accuracy and efficiency. CEASC [41] substitutes sparsely sampled feature statistics with global context features and implements an adaptive multi-layer masking strategy. This approach optimizes mask ratios across various scales, resulting in enhanced foreground coverage and improved precision and efficiency. DTSSNet [42] incorporates a manually designed block between the backbone and neck to enhance sensitivity to multi-scale features, along with a sample selection method tailored for small objects.

Drone-YOLO [43] adopts a three-layer PAFPN structure combined with a detection head tailored for small targets, significantly improving the algorithm's capability to detect small objects. ESOD [44] combines feature-level target searching with image block slicing, reusing the detector's Backbone for feature extraction and employing sparse detection heads to reduce computational waste in background areas, enabling efficient and accurate small object detection in high-resolution images.

FFCA-YOLO [45] enhances the network's sensitivity to local regions, multi-scale feature fusion, and global interrelations across channels and spatial dimensions through techniques like Feature Enhancement Modules (FEM), Feature Fusion Modules (FFM), and Spatial Context Awareness Modules (SCAM). This method achieves higher detection accuracy on small object detection datasets of aerial images and demonstrates good robustness under various simulated degradation conditions.

Liu [46] introduced supervision for micro-object regions during training, enabling the extraction of these small areas and creating a clear attention map. By employing this explicit attention map to modulate the feature map semantically, they suppress the background while enhancing regions containing micro-objects, thereby proving the method's effectiveness for small object detection.

To address the challenges posed by large-scale images and uneven object distributions, YOLC [47] introduces a Local Scale Module (LSM) that adaptively searches clustered regions to zoom in for precise detection. The regression loss is modified using Gaussian Wasserstein distance to obtain high-quality bounding boxes, while the detection head employs deformable convolutions and refinement techniques to enhance detection capabilities for small objects.

3. Methodologies

This section presents a comprehensive overview of the proposed small object detection framework for aerial images, termed SO-RTDETR. First, we summarize the methodology and present the overall network architecture. Subsequently, we elaborate on the composition of each module.

3.1. Overall Architecture

We select RT-DETR as the baseline framework, which is an optimized version of DETR, designed to deliver faster inference speeds and efficient real-time detection capabilities. After extracting multi-scale feature maps using a backbone network, RT-DETR utilizes an Efficient Hybrid Encoder to convert multi-scale features into a sequence of image features through attention-driven intra-scale feature interactions (AIFI) and CNN-based cross-scale feature fusion (CCFF). It then employs uncertainty-minimizing query selection to choose a fixed number of encoder features, which serve as initial object queries for the decoder. The decoder progressively refines these queries with the assistance of auxiliary prediction heads to produce object classifications and bounding boxes.

This study introduces the SO-RTDETR framework for small object detection in aerial imagery, as illustrated in Figure 2. The framework is specifically designed to enhance the detection of small targets and consists of three core components: the Backbone, Encoder, and Decoder. The Backbone constructs a five-layer feature pyramid using a CNN to extract features at different scales and information levels, represented as S1, S2, S3, S4, S5. To improve the ability to capture information about small objects, we introduce a Precision PPA module within the Backbone to retain fine detail information of small targets.

During the feature fusion stage, we introduce the S2-CCFF module, which incorporates the S2 layer for cross-scale feature fusion. This module effectively merges features rich in small object information with the S3, S4, and S5 layers, facilitating the learning of feature representations from global to local, and significantly reducing the impact of noise.After processing through uncertainty-minimizing queries and the Decoder, we utilize the Normalized Wasserstein Distance (NWD) metric in the detection head to assess the loss between the ground truth and the predicted bounding boxes, effectively addressing small object bounding box perturbations and enhancing detection performance for small targets.

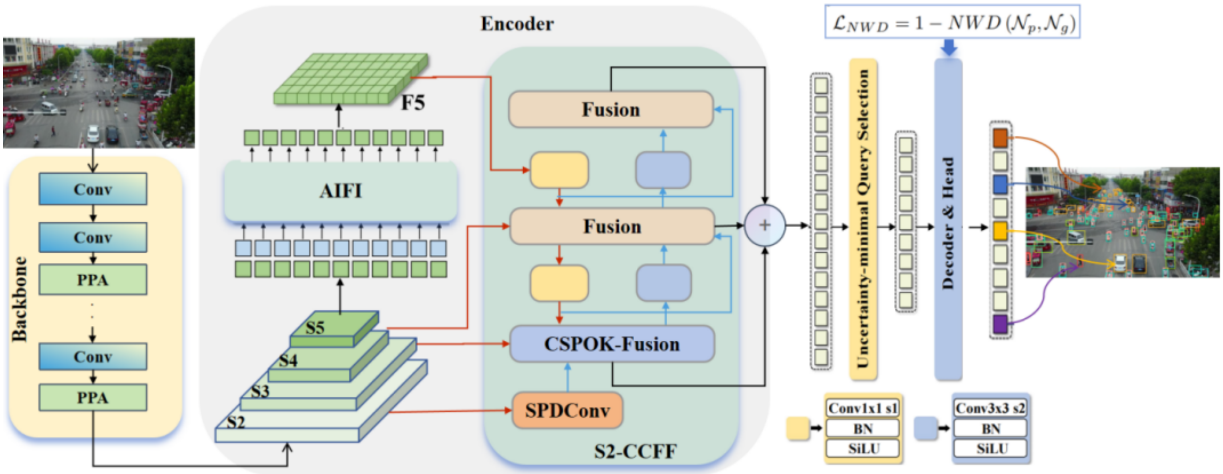


Figure 2. Overall Architecture of SO-RTDETR: The Backbone integrates the PPA module to retain essential information regarding small targets. S2-CCFF module incorporates the S2 layer,Enriched small target information. During detection, the Normalized Wasserstein Distance (NWD) loss is employed to address challenges associated with bounding box perturbations.

3.2. S2-CCFF Module

3.2.1. Module Structure

In small object detection tasks, challenges such as high background noise and information loss can significantly hinder the model's ability to capture target features. In deep learning algorithms, after constructing a multi-scale feature space through the backbone, multi-scale feature fusion is typically performed to capture features at different scales, enhancing detection accuracy. In RT-DETR, the interaction and fusion of cross-scale features continue to rely on the FPN as the optimal choice from a real-time perspective. The CCFF performs PAFPN operations on the S3 to S5 layers, as illustrated in Figure 3 (a). The Fusion module within this context is designed in the style of a CSPBlock. The multi-scale pyramid comprises the S1 to S5 layers, corresponding to high, medium, and low-resolution feature maps. Since the resolution of a feature map is inversely related to its receptive field, higher-resolution feature maps (e.g., S1) generally contain more fine-grained spatial information, whereas lower-resolution maps (e.g., S5) possess richer semantic information but less spatial detail. While higher-resolution feature maps like S1 retain more spatial information, their smaller receptive fields result in a lack of contextual information, making it challenging to distinguish small targets from the background or adjacent objects [48]. Conversely, relying solely on the S3, S4, and S5 layers may fail to fully utilize the fine-grained information provided by higher-resolution feature maps (e.g., S2), which is crucial for accurately detecting small targets. The common approach is to incorporate higher resolution feature layers (S2) to better preserve the spatial details of small targets and improve detection accuracy. However, this can introduce several issues, including increased computational load and longer post-processing times.

To address these challenges in aerial imagery, we propose the S2-CCFF module based on the CCFF module, as shown in Figure 3 (b). This module integrates the S2 layer into cross-scale feature fusion to improve the retention and enhancement of small target feature representations, with S2, S3, S4, and S5 corresponding to resolutions of 1/4, 1/8, 1/16, and 1/32 of the original image, respectively. To mitigate the increased computational burden and time consumption associated with adding the S2 layer, we first process the S2 feature layer using SPDConv [49], applying spatial downsampling to reduce the resolution of the S2 layer. This involves rearranging pixels to adjust the dimensions and structure of the feature map, allowing us to retain essential detail without directly handling high-resolution feature maps. This effectively resolves issues related to excessive computational demands and extended post-processing times after adding the S2 detection layer. Subsequently, the feature rich in small object information is fused with the S3 layer using the CSPOK-Fusion module. The CSPOK-Fusion module integrates concepts from CSP and Omni-Kernel [50], consisting of global, large, and local branches to effectively learn feature representations from global to local scales. This approach significantly suppresses background noise, enhances the detection performance of small targets, and reduces computational complexity.

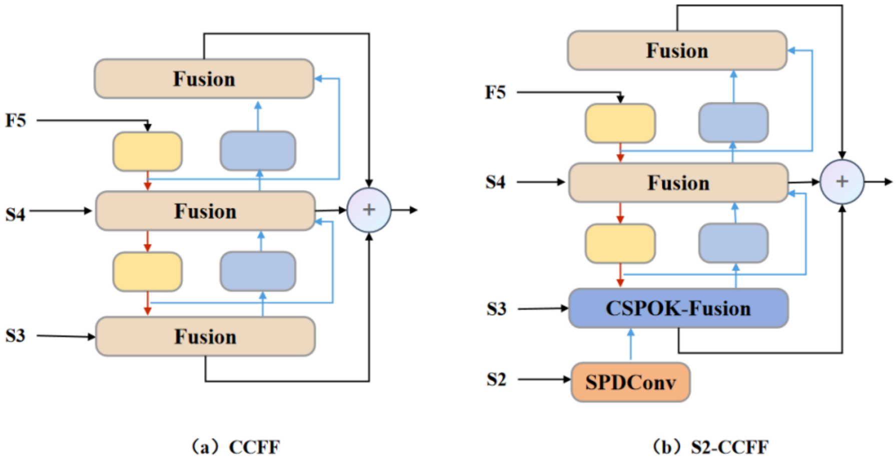
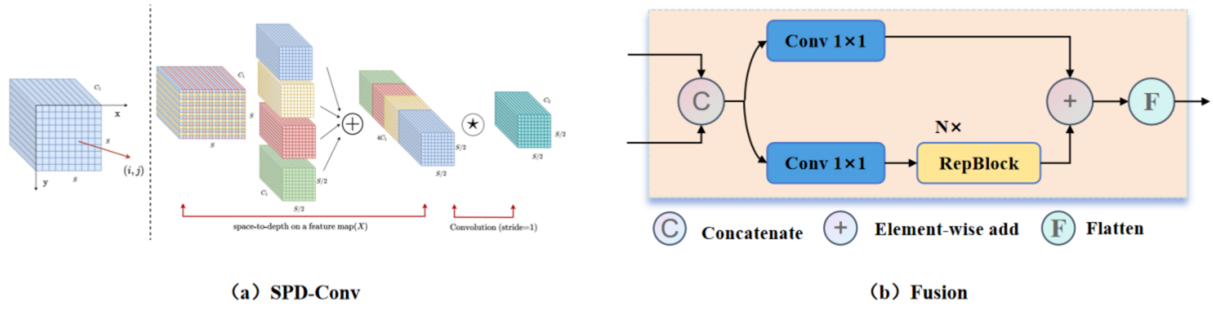


Figure 3. Comparison between CCFF and S2-CCFF.(a) represents the CCFF structure utilized in RT-DETR, while (b) illustrates the proposed S2-CCFF in this work.

The S2-CCFF mainly comprises the SPDConv, Fusion, and CSPOK-Fusion modules. In Figure 3 (b) , red arrows indicate downsampling, and blue arrows indicate upsampling. The SPDConv structure is depicted in Figure 4 (a). Its core concept is to introduce a novel CNN building block that replaces the conventional row-wise convolution and pooling layers found in CNN architectures. By combining spatial-to-depth (SPD) layers with non-row-wise convolution layers, SPDConv aims to enhance performance in detecting low-resolution images and small object targets. It effectively preserves fine-grained information, addressing the information loss problem associated with traditional row-wise convolution and pooling layers, thereby significantly improving performance in object detection and image classification tasks.

The Fusion Block, illustrated in Figure 4 (b), first receives features from different layers or scales and concatenates these features along the channel dimension using a concat operation. Subsequently, the features pass through two 1x1 convolutional layers. One branch enters the "RepBlock" module, where it undergoes stacking through N RepBlock units. The RepBlock is a block that incorporates various convolutional and activation operations, aimed at further enhancing the representational capability of the features. It provides more profound feature extraction across multiple levels of features. Next, the features processed by the RepBlock are added element-wise to another set of features that have not undergone RepBlock processing. This step functions as a skip connection, ensuring that features from different levels can be directly fused, thereby preserving the detailed information of shallow features while integrating the contextual information of deep features. Finally, the fused features are flattened in preparation for subsequent classification or regression tasks. This Fusion Block effectively merges information across multiple feature levels through concatenation, convolution operations, and residual connections, thereby strengthening the network’s ability to represent target features in small object detection tasks while reducing the interference from redundant features.



(a) SPD-Conv

(b) Fusion

Figure 4. Submodule Structures.(a) SPD-Conv, (b) Fusion Module.

The S2-CCFF module effectively addresses the challenges of significant noise interference and information loss in small object detection through its design, which incorporates adaptive convolution and multi-scale feature fusion. Compared to traditional feature fusion methods, the S2-CCFF module better retains and enhances the feature representation of small objects, enabling the detector to achieve higher accuracy and robustness in complex backgrounds and noisy environments.

3.2.2. CSPOK-Fusion Module

In the Fusion architecture, the cross-scale feature fusion is highly effective for integrating features from the S3 to S5 layers. However, when directly fusing S2 with S3 to S5, significant differences between shallow and deep features may lead to feature conflicts. Specifically, the low-level information from the S2 layer can obscure the high-level semantic features of the S3 to S5 layers, adversely affecting the overall performance of the network. Traditional fusion structures may not adequately address these disparities, which can diminish the effectiveness of the fusion process. Therefore, this study introduces the CSPOK-Fusion module, which incorporates the CSPOK module into the original Fusion architecture. Figure 5 illustrates the CSPOK-Fusion structure, which combines CSPOK with multi-branch convolution operations and RepBlock mechanisms through cross-channel fusion. The CSPOK-Fusion module enhances useful shallow detail information while suppressing redundant shallow information, thereby preventing shallow features from interfering with deep semantic features.

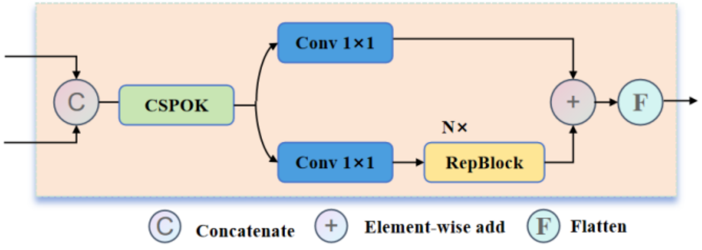


Figure 5. CSPOK-Fusion Structure. This structure is primarily used for the fusion of features from the S2 and S3 layers. Compared to the Fusion module, the CSPOK module has been added to enhance the integration process.

The CSPOK Block improves upon the Omni-Kernel Module (OKM) using the Cross Stage Partial (CSP) concept, as illustrated in Figure 6 (a). Initially, the input feature map X undergoes preliminary feature extraction through a convolutional layer. The CSP structure subsequently divides the input feature map into two branches, with the feature map split into two parts of $\frac{1}{4}$ and $\frac{3}{4}$. The left branch X_1, corresponding to the $\frac{1}{4}$ feature map, performs OKM operations, which involves channel compression via a 1×1 convolution, followed by feature extraction through three levels: local branch, large branch, and global branch. In the large branch, the convolution kernel size of 63 is replaced with 61. Meanwhile, the right branch X_2, representing the $\frac{3}{4}$ portion, directly concatenates with the feature map processed by OKM and fuses the two through a convolutional layer. This structure enhances the network's learning capacity without increasing computational overhead. It retains the robust contextual processing ability of the OKM module while ensuring that the other branch sufficiently preserves small target information, thereby improving the accuracy and speed of object detection.

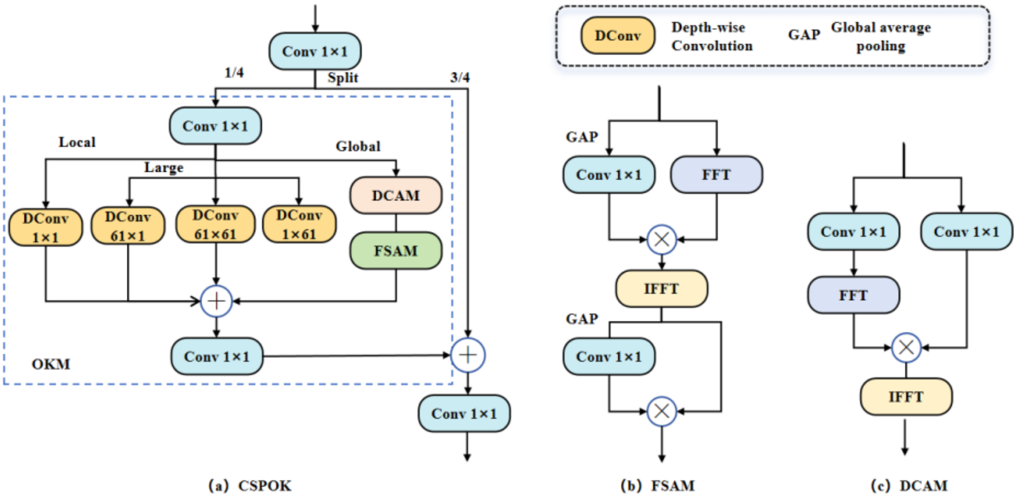


Figure 6. Structure of the CSPOK module. (a) Module structure; (b) FSAM structure; (c) DCAM structure.

(1) Local branch: performs 1x1 depth convolution to extract local features.

$$\begin{aligned} \mathrm{y} &= \mathrm{C}_{1\times 1}(\mathrm{X}_1) \\ \mathrm{O}_{\mathrm{locall}} &= \mathrm{DC}_{1\times 1}(\mathrm{y}) \end{aligned} \tag{1}$$

where O_{locall} represents the output of the local feature processing branch, $C_{1\times 1}$ denotes the 1×1 convolution, and$DC_{1\times 1}$ represents the 1×1 Depthwise convolution.

(2) Large branch: performs large-scale feature extraction through three depthwise convolutions (DC).

$$\mathrm{O}_{\mathrm{large}} = \mathrm{DC}_{1\times 61}(\mathrm{y}) + \mathrm{DC}_{61\times 61}(\mathrm{y}) + \mathrm{DC}_{61\times 1}(\mathrm{y}) \tag{2}$$

here, for the input y, after undergoing three depthwise convolutions 1×61, 61×61, 61×1 with large kernels, the results are summed to obtain the outcome of the large-scale feature processing O_{large} .

(3) Global branch: acquires global features through the DCAM and FSAM modules, represented as follows:

$$\mathrm{O}_{\mathrm{global}} = \mathrm{FSAM}(\mathrm{DCAM}(\mathrm{Z})) \tag{3}$$

The DCAM utilizes a dual-channel attention mechanism to enhance mutual information among feature channels, thereby improving their representation capability. Initially, input features undergo a 1×1 convolution, followed by transformation into the frequency domain via the Fast Fourier Transform (FFT). In this domain, dual-channel weighting is applied through two distinct convolution paths, processing feature information across different channels. The features are then transformed back to the spatial domain using the Inverse Fast Fourier Transform (IFFT) and subjected to another 1×1 convolution to reconstruct the feature map. This process is further refined by fusing the features to strengthen inter-channel correlation and enhance feature expressiveness.

The FSAM improves feature extraction efficacy by transforming the feature map into the frequency domain for processing. Initially, input features undergo a 1×1 convolution, followed by the application of the Fast Fourier Transform (FFT), which transforms the feature map from the spatial domain to the frequency domain. In this domain, different frequency components are emphasized using weighted attention to enhance critical frequency information. The feature map is then transformed back to the spatial domain through the Inverse Fast Fourier Transform (IFFT). Lastly, the attention-enhanced frequency domain features are fused with the original features, resulting in an enriched feature representation.

Ultimately, the local features, large-scale features, and global features are fused through element-wise addition. The $\frac{3}{4}$ feature map obtained from the initial convolution is added element-wise to the fused features and then processed through a convolution layer to generate the final feature map $\mathrm{F}_{\mathrm{out}}$.

$$F_{out} = \mathrm{Conv}(\mathrm{X}_2 + \mathrm{C}_{1\times1}(O_{\mathrm{local}} + O_{\mathrm{large}} + O_{\mathrm{global}})) \tag{4}$$

In summary, the CSPOK-Fusion structure is primarily designed for the fusion of feature maps from layers S2 and S3. Building upon the original fusion framework, the CSPOK module is incorporated to facilitate the gradual fusion of partial features and implement residual connections. This approach enhances the completeness and richness of feature representation while maintaining computational efficiency. For small object detection, the CSPOK module minimizes redundant feature transmission and ensures that critical information is not lost within the network hierarchy. This structure effectively integrates deep and shallow features while ensuring computational efficiency. Such a design is particularly well-suited for small object detection tasks, enabling the model to capture target details across different scales and enhancing its capability to detect small objects in complex backgrounds.

3.3. Backbone network structure

Feature extraction is critical for object detection in aerial images, commonly accomplished using a CNN-based architecture that constructs a five-layer feature pyramid to capture information at multiple scales and feature representations. However, many targets in aerial imagery appear as small objects, often occupying only a minor portion of the image. As the feature pyramid is constructed, these small objects are susceptible to gradual dilution or loss, while larger targets present increased complexity and diversity, which complicates the feature extraction process. To tackle these challenges, this study incorporates the PPA module [51] into the Backbone network, as illustrated in Figure 7. Layer 1 uses convolutional operations to extract low-level features from the input image. Layers 2 to 4 integrate Conv modules with the PPA module to generate multi-scale feature maps. The PPA utilizes hierarchical feature fusion and attention mechanisms to maintain and enhance small object representations, ensuring that vital information is preserved throughout multiple downsampling steps, ultimately improving detection accuracy in later stages.

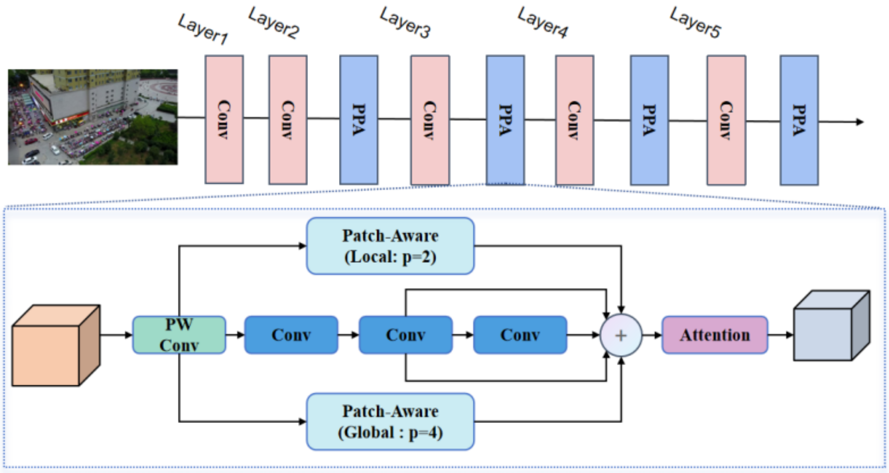


Figure 7. Backbone network structure. Layer 1 extracts low-level feature information through convolutional modules, Layers 2 to 5 comprise Conv modules integrated with the PPA module, which includes two components: multi-branch fusion and attention mechanisms. The multi-branch block consists of local and global branches.

The PPA employs a parallel multi-branch strategy, primarily comprising two components: multi-branch fusion and attention mechanisms. The fusion component integrates patch-aware and concatenated convolutions. By modifying the patch size parameter, local and global branches are distinguished, with the parameter 'p' set to 2 for local and 4 for global branches. Each branch is responsible for extracting features at varying scales and levels. This multi-branch approach enhances the capture of multi-scale features of targets, thereby improving the accuracy of small object detection.

Given the input feature tensor $\mathbf{F} \in \mathbb{R}^{H' \times W' \times C}$, it is initially processed through point-wise convolution to produce the desired output $\mathbf{F}' \in \mathbb{R}^{H' \times W' \times C'}$. Subsequently, calculations are performed for $\mathbf{F}_{\text{local}} \in \mathbb{R}^{H' \times W' \times C'}$, $\mathbf{F}_{\text{global}} \in \mathbb{R}^{H' \times W' \times C'}$, and $\mathbf{F}_{\text{conv}} \in \mathbb{R}^{H' \times W' \times C'}$ through the three branches, respectively. Finally, the three results are aggregated to generate the final feature representation.

In the global branch, computationally efficient operations are used, including Unfold and Reshape, to divide F′ into a series of spatially continuous blocks $(p \times p, H'/p, W'/p, C)$. Subsequently, each channel undergoes average processing to obtain the feature representation, which is then processed using a feed forward network (FFN) [26] for linear computation. An activation function is applied to acquire the probability distribution of the linear computed features in the spatial dimension, allowing for corresponding weight adjustments. In the weighted results, feature selection [27] is utilized to choose task-relevant features from the tokens and channels.

The PPA module aligns and refines contextual information from multiple layers progressively through a hierarchical fusion approach. For small objects, this method helps retain fine details that are typically lost in traditional networks. This is crucial for detecting small objects in complex backgrounds, enabling effective execution of aerial image detection.

3.4. NWD loss function

For small targets in aerial images, even small deviations in bounding boxes can lead to a significant decrease in IoU, thereby affecting localization accuracy. This low tolerance makes accurate localization of small targets particularly difficult. Due to the low tolerance of bounding box perturbations, the detection of small targets is more susceptible to interference from noise and background debris, leading to an increase in false positives and false negatives. To address this issue, we employs a novel measure of NWD similarity in place of the IoU. The NWD distance presents an innovative solution by first modeling the bounding box as a two-dimensional Gaussian distribution and then assessing their similarity through the calculation of the Wasserstein distance between the two distributions. Due to its capacity to quantify the similarity between distributions, Wasserstein distance is particularly well-suited for evaluating the similarity of small objects, even in cases of negligible or non-existent overlap. Furthermore, NWD is less sensitive to variations in object scale, enhancing its applicability for small object detection.

Specifically, for horizontal bounding boxes $R = (cx, cy, w, h)$, where (cx, cy) represent the center coordinates, and and denote the width and height, respectively. The two-stage Wasserstein distance between two bounding boxes is defined as follows:

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^{\mathrm{T}}, \left[cx_b, cy_b, \frac{w_b}{2} . \frac{h_b}{2} \right]^{\mathrm{T}} \right) \right\|_2^2 \tag{5}$$

here, the distance measure $W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ cannot be directly utilized as a similarity measure (i.e., a value between 0 and 1 corresponds to IoU). Consequently, we applied an exponential normalization to derive a new metric known as Normalized Wasserstein Distance (NWD):

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C} \right) \tag{6}$$

here, C is a constant closely associated with the dataset. The loss function based on NWD is defined as follows:

$$\mathcal{L}_{NWD} = 1 - NWD(\mathcal{N}_p, \mathcal{N}_g) \tag{7}$$

where $\mathcal{N}_p$ is the Gaussian distribution model of the prediction box P, and $\mathcal{N}_g$ is the Gaussian distribution model of the GT box G.

For small targets, the IoU metric is particularly sensitive to minor perturbations in the bounding box; even slight positional deviations can result in a substantial decline in IoU, adversely impacting detection accuracy. In contrast, the NWD distance assesses the similarity between Gaussian distributions of bounding boxes. It can evaluate the similarity between predicted and actual boxes more accurately, even when small perturbations are present, thereby enhancing detection accuracy.

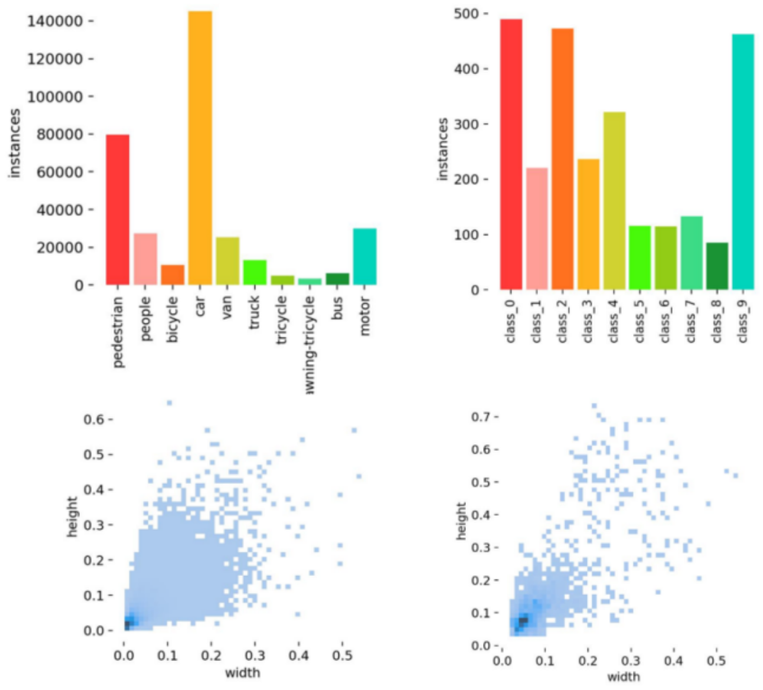
4. Experiment

4.1. DataSet

This study utilizes two widely used aerial image datasets for object detection: VisDrone 2019 and NWPU VHR-10. VisDrone 2019 is a large-scale dataset comprising 10,209 images captured by drones in urban and suburban areas across 14 Chinese cities. With a resolution of 2000 × 1500 pixels, it includes 542,000 instances spanning 10 common traffic scene categories such as pedestrians and vehicles. The dataset presents challenges like occlusion and varying viewpoints, with a high density of small objects, making it suitable for small object detection tasks.

The NWPU VHR-10 dataset, released by Northwestern Polytechnical University, contains 650 annotated high-resolution remote sensing images covering 10 object categories, such as airplanes, ships, and bridges. These images are sourced from Google Earth and the Vaihingen dataset, encompassing diverse scene types with varying object densities, ranging from dense to sparse distributions. The dataset is partitioned into training, validation, and test sets with a 7:2:1 split. Both datasets contain numerous small, densely packed objects, providing a robust environment for evaluating object detection models in aerial imagery.

Figure 8 shows the histogram of class instance distribution (first row) and the scatter plot of target box width and height distribution (second row) for two datasets. Figure 8 (a) Listed as VisDrone 2019 dataset, Figure 8 (b) as NWPU VHR-10 dataset. The number of categories in the first row shows that the number of categories in VisDrone and NWPU VHR-10 is not equal, with significant differences. The width and height of the target in the second dataset are relatively small, which poses certain challenges for object detection.



(a) VisDrone **(b) NWPU VHR-10**

Figure 8. Scatter plot of category distribution and width/height distribution for the VisDrone 2019 dataset and NWPU VHR-10 dataset, with the first row representing category distribution and the second-row representing width/height distribution.

4.2. Evaluation metrics and Environment

4.2.1. Evaluation metrics

In this experiment, the complexity of the algorithm is quantified in terms of floating point operations (FLOPs). The performance metrics used for comparative evaluation of the network include precision (P), recall (R), average precision (AP), and mean average precision (mAP) for each class. All predicted results are classified as positive samples. The evaluation framework defines true positive (TP) as the count of correctly detected positive samples, false positive (FP) as the number of incorrectly identified positive samples, and false negative (FN) as the actual targets that were missed during detection.

Precision is defined as the probability of correct predictions among all predictions, thereby assessing the accuracy of the algorithm's predictions. Recall represents the ratio of correctly predicted results to actual occurrences, measuring the algorithm's ability to identify all target objects. These metrics correspond to the probabilities of false detection and missed detection, respectively.

$$\begin{aligned}\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}}\end{aligned} \tag{8}$$

The area under the precision-recall (P-R) curve obtained from different numbers of positive samples represents the average precision (AP) for each class, while the mean average precision (mAP) is calculated as the average of the average precision across all classes. The formulas are as follows:

$$\begin{aligned}\text{AP} &= \int_0^1 \text{Precison (t)dt} \\ \text{mAP} &= \frac{\sum_{\text{n}=1}^{\text{N}} \text{AP}_\text{n}}{\text{N}}\end{aligned} \tag{9}$$

Additionally, we employ other evaluation metrics. P indicating the parameter size. This metric reflects the complexity of the model. GFLOPs refer to the number of floating-point operations performed by the model during execution, serving as a crucial indicator for assessing computational complexity.

4.2.2. Experimental Environment

The experimental environment and setup are as follows: the model was implemented on an A800 GPU cluster equipped with three 256GB GPUs, running on the Red Hat 4.8.5-28 operating system. Python version 3.11 and CUDA version 12.1 were used. Standard data augmentation techniques, such as random cropping, flipping, and scaling, were applied to enhance the model's generalization capability. The baseline model employed was RT-DETR, using ResNet50 as the feature extraction CNN. Other parameters used in this study are listed in the table below:

Table 1. Hyperparameter setting .

Parameter	Value
optimizer	AdamW
base_learning_rate	0.0001
weight_decay	0.0001
global_gradient_clip_norm	0.1
linear_warmup_steps	2000
minimum learning rate	0.00001

4.3. Ablation Study

Ablation experiments were performed on the VisDrone and NWPU VHR-10 datasets to assess the effectiveness of the proposed improvements outlined in this section. Utilizing RT-DETR as the baseline and ResNet-R18 as the backbone network, the results obtained from the VisDrone validation dataset are displayed in Table 2.

Table 2. Ablation experiments on the VisDrone$_{v}aldataset$.

PPA	S2-CCFF	NWD	P(M)	GFLOPS	mAP
×	×	×	38.61	56.8	44.8
✓	×	×	27.46	62.1	45.6
×	✓	×	39.8	66.9	46.8
×	×	✓	38.61	57.0	45.1
✓	✓	×	37.39	77.7	47.7
✓	✓	✓	37.39	76.0	47.9

The mPA of the baseline is 44.8%. Mainly because RT-DETR uses Transformer's self-attention mechanism, although it can capture global information, when facing extremely small targets, the global features may be too sparse to fully focus on the target itself. This can cause the features of small targets to be ignored in larger contexts, thereby affecting detection accuracy.

After adding the PPA module, Parameter reduced by 11.15 M approximately, GFLOPs increased, and mPA increased to 0.8%. Mainly because the PPA module adopts a multi branch fusion and attention mechanism. In multi-branch fusion, each branch uses different convolution operations (such as patch-aware and concatenated convolutions) to extract features, which are fused in subsequent stages. This fusion not only combines information from different scales but also enhances feature richness and representation ability through diverse convolution operations. The attention mechanism can further focus on key areas in the image, ensuring that key information is saved through multiple downsampling steps, thereby improving the detection accuracy of small targets. After only adding the S2-CCFF module, the number of parameters and calculations increased, and the accuracy also improved to 2%. This is because the S2-CCFF module adds an S2 layer that contains rich information about small targets during multi-scale feature fusion. At the same time, the CSPOK module adopts the CSP idea and global, large, and local branches to provide the model with a multi-granularity receptive field, effectively learning global to local feature representations. This improves the model's ability to extract and fuse features for small targets and effectively suppresses background noise. However due to the addition of the S2 layer, the parameters of the model and the computer have increased. After introducing the NWD loss function, GFLOP remained almost unchanged, but mAP increased by 0.3%. This is due to the ability of NWD to evaluate the similarity between Gaussian distributions of bounding boxes. Even when small target bounding boxes encounter small deviations, NWD can more accurately measure the similarity between predicted boxes and real boxes than traditional methods, thereby optimizing detection accuracy. When using both PPA and S2-CCFF modules simultaneously, the parameter count is between the case of using only PPA or only S2-CCFF, and GFLOPs reach the highest of 77.7 G, but the accuracy is also improved by 2.9%. Indicating that these modules may have a synergistic effect. They can better adapt to small object detection tasks by cooperating and optimizing various aspects of the network, jointly improving the performance of small object detection tasks in aerial images. After adding three modules simultaneously, there was a 3.1% increase compared to the baseline. Indicating that these modules may have a synergistic effect. They can better adapt to small object detection in aerial images and jointly improve detection performance by cooperating and optimizing various aspects of the network.

In the NWPU VHR-10 dataset, we can obtain similar experimental results, as shown in Table 3 compared to having only one module, networks with overlapping blocks achieve better performance. The addition of three modules enables the model to gradually align and refine contextual information from multiple layers, ensuring that key information is saved through multiple downsampling steps, effectively fusing deep and shallow features, and effectively suppressing the interference of complex backgrounds to achieve better regression and classification results. Meanwhile, the experiment also showed that the proposed module had no conflicts, and when all proposed methods were adopted, the model exhibited the best performance of 89.5 on both datasets.

Table 3. Ablation experiments on the NWPU VHR-10 dataset .

PPA	S2-CCFF	NWD	P(M)	GFLOPS	mAP
×	×	×	38.61	57.0	86.4
✓	×	×	35.05	60.2	87.2
×	✓	×	39.80	65.2	88.2
×	×	✓	38.61	57.0	87.3
✓	✓	×	39.2	76.0	89.3
✓	✓	✓	39.2	76.0	89.5

4.4. Comparative Experiment

4.4.1. VisDrone

To further assess the effectiveness of our method, comparative experiments were performed on the VisDrone dataset alongside other established methods, as presented in Table 4. Faster R-CNN and Cascade R-CNN are classified as two-stage methods, while the remaining methods are one-stage approaches. The results indicate that ATSS achieves the highest performance in mAP and mAP_l, while SO-RTDETR shows excellent performance in small target detection (mAP_s=0.157) and is highly competitive at mAP_50 (0.393), making it suitable for applications that require high-sensitivity small target detection. YOLOX has the weakest overall performance, while RT-DETR shows a balance in small object detection and mAP_50, but both perform poorly under stricter conditions.

Table 4. Comparative experiment based on VisDrone_test

.

Method	mAP	mAP_0.5	mAP_0.75	mAP_s	mAP_m	mAP_l
Faster-RCNN [15]	0.205	0.342	0.219	0.100	0.295	0.433
Cascade-RCNN [17]	0.208	0.337	0.224	0.101	0.299	0.452
TOOD [52]	0.214	0.346	0.230	0.104	0.303	0.416
ATSS [53]	**0.216**	0.349	**0.231**	0.102	0.308	**0.458**
RetinaNet [20]	0.178	0.294	0.189	0.067	0.265	0.430
RTMDet [54]	0.184	0.312	0.213	0.077	0.288	0.445
YOLOX [55]	0.156	0.283	0.155	0.078	0.213	0.288
RT-DETR [11]	0.159	0.365	0.107	0.138	0.284	0.231
SO-RTDETR	0.176	**0.393**	0.123	**0.157**	**0.330**	0.227

This is because ATSS introduces the adaptive training sample selection (ATSS) mechanism, which adaptively selects positive and negative samples for each target and combines with the FPN to effectively integrate multi-scale features, especially with high robustness for detecting large targets. The adaptive positive sample selection mechanism helps to improve detection accuracy under strict IoU thresholds (such as mAP_l), and it can accurately select the most suitable anchor box. In addition, ATSS can effectively respond to targets of different scales through this mechanism, resulting in the best overall mAP performance.

Although YOLOX is a one-stage method, its performance in small object detection and high IoU threshold is poor. The anchor-free strategy adopted by YOLOX is more flexible in localization, but its ability to express features of small targets is insufficient. In addition, YOLOX's FPN design cannot fully capture the detailed information of small targets when processing high-density small target scenes, resulting in a lower overall mAP. In addition, under stricter IoU thresholds such as mAP_75 and mAP_l, YOLOX's robustness is not as good as ATSS and other methods because its positive and negative sample allocation mechanism has not been fully optimized for high IoU.

RT-DETR, as a DETR-based network, utilizes the Transformer architecture and its powerful self-attention mechanism helps capture global contextual information, resulting in a relatively balanced performance in small object detection and mAP_50. Transformer can effectively handle the relationships between objects and enhance feature learning, making it suitable for multi-scale object detection. The performance of RT-DETR is weak under strict conditions such as mAP_75 and mAP_l, mainly due to the Transformer model's strong dependence on training data and slow convergence. At high IoU thresholds, the regression accuracy of RT-DETR is insufficient, making it difficult to match models with adaptive mechanisms such as ATSS. In addition, the optimization of RT-DETR may not fully adapt to small deviations of small targets under stricter IoU conditions.

SO-RTDETR is a network optimized for small object detection. It introduces modules such as S2-CCFF and PPA basis on RT-DETR, which significantly enhance the model's perception of small targets by enhancing fine-grained feature extraction capabilities. Especially in the feature pyramid structure, the fine capture of small target information is more sensitive, avoiding the loss of small target features in the downsampling process of traditional detection networks. Therefore, SO-RTDETR performs well in small target detection (mAP_s=0.157) and mAP_50 (0.393), making it suitable for small target scenarios that require high-sensitivity detection.

The VisDrone dataset in Figure 9 displays city streets and roads at different times and locations. This includes various complex urban environments, including busy streets, intersecting roads, buildings, and bridges, which pose challenges to detection algorithms. From the detection effect diagram of the method in this article, it can be seen that the model has robustness for detecting small targets in different scenes, times, and types, especially for targets in complex scenes, as well as in varying lighting conditions, occlusions, and background interferences.

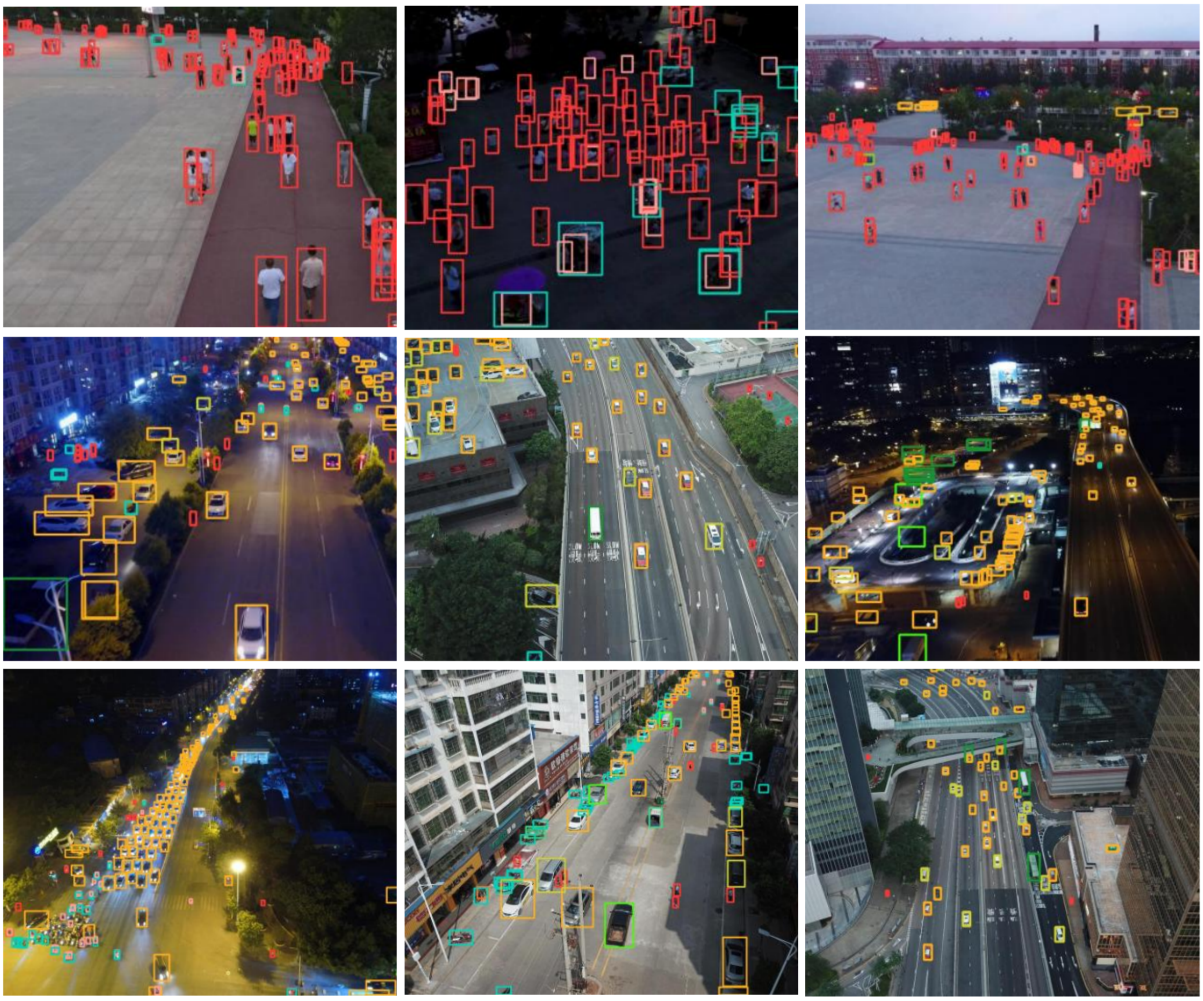


Figure 9. Figure 9: Detection results of SO-RTDETR on the VisDrone dataset.

To further assess the effectiveness of the proposed method, feature maps were generated with the Encoder on two VisDrone images. In these maps, colors represent different activation intensities, with green and yellow regions indicating higher levels of activation. These areas signify features that substantially contribute to the model's decision-making process **??**. As shown in Figure 10 (a) list as two original input images. The first image is a square, and the following image is a street with many vehicles and pedestrians. The second column shows the feature map generated by the baseline model, reflecting the spatial detail information extracted from the image. However, these feature maps appear noisy and scattered when focusing on specific areas of interest, such as pedestrians in squares or vehicles on streets. The activation areas are distributed throughout the feature maps, lacking concentrated attention to small targets. The third column shows the feature maps generated by the SO-RTDETR model, demonstrating a more focused attention mechanism, especially in areas where there may be targets such as pedestrians and vehicles. In contrast, the activation area is clearer and exhibits stronger concentration in the relevant areas of small object detection. This indicates that the method proposed in this article can better capture feature information related to small objects. Compared with the baseline, the feature map of SO-RTDETR significantly reduces irrelevant noise and focuses more on key areas in the image, indicating higher efficiency in small object detection.

The visual representation of the feature map indicates that the proposed method concentrates more on specific regions within the image, resulting in the extraction of more pronounced features. In contrast, the feature map generated by the Baseline method displays a more dispersed color distribution, suggesting that the activated feature areas are broader. This lack of focus on target regions adversely impacts detection accuracy, potentially leading to false positives or missed detections.

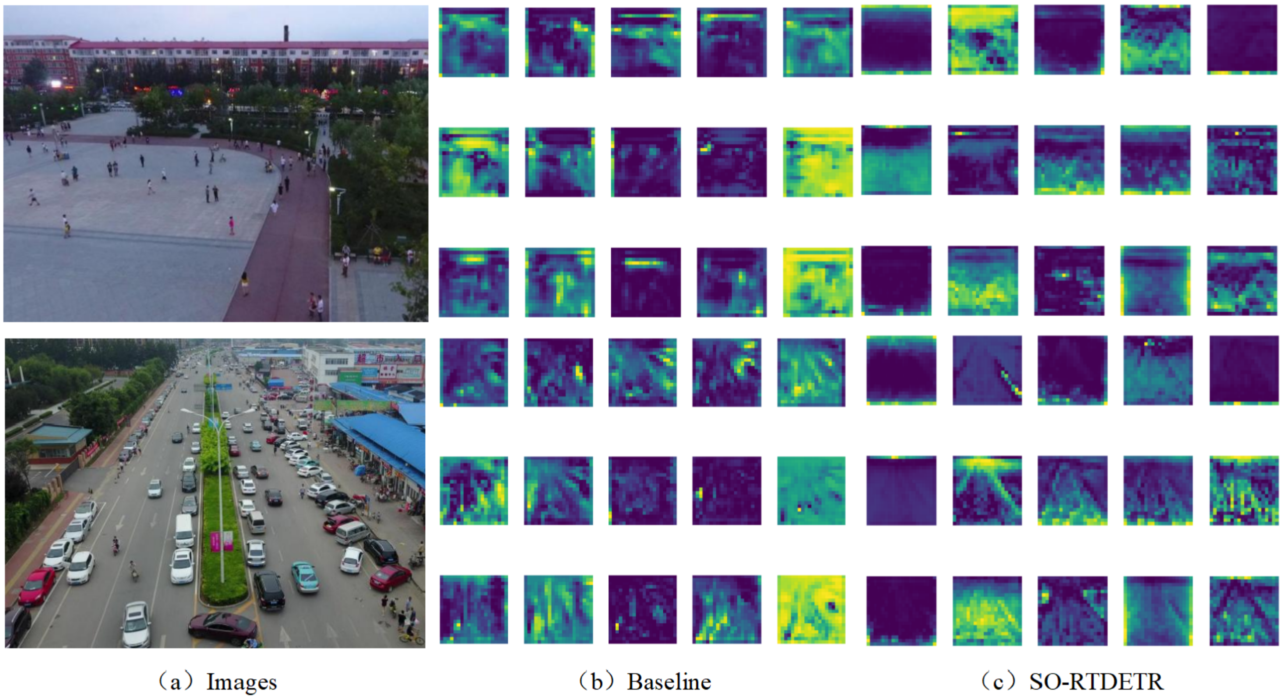


Figure 10. Feature maps created after Encoder. (a) shows two images from the VisDrone dataset, (b) illustrates the feature map produced by the Baseline method, while panel (c) depicts the feature map generated by the proposed method. The figure demonstrates that the feature maps derived from our method exhibit richer detailed information and a more pronounced hierarchical structure, effectively capturing a substantial number of small target features.

Figure 11 shows the comparison of object detection results between our method and RT-DETR, Cascade RCNN, and YOLOX in different scenarios. By analyzing the detection performance of different methods in various scenarios, it can be seen from the figure that our method successfully detected more small targets, especially pedestrians and small vehicles, in multiple scenarios (such as squares, nighttime roads, rural intersections, etc.), demonstrating its sensitivity to small targets and crowded scenes. This indicates that the method has good feature extraction and precise localization capabilities. In complex scenarios, the method proposed in this article can maintain a high detection rate, especially in high-density traffic areas, demonstrating strong adaptability. The performance of RT-DETR is relatively balanced in various scenarios, especially in nighttime roads and high-density vehicle scenes, where it can detect most vehicles and pedestrians, demonstrating good ability in detecting medium-sized targets. However, RT-DETR has some shortcomings in detecting small targets such as distant pedestrians or small vehicles, especially in some long-distance or low-resolution scenes where there are relatively more missed detections. Cascade RCNN has shown good detection performance for large vehicles and targets close to cameras in multiple scenarios, demonstrating its strong detection ability for large and medium-distance targets. However, in some target-dense scenarios (such as urban streets), Cascade RCNN fails to detect pedestrians and small targets at long distances well, especially in scenes such as squares and rural intersections, where the phenomenon of missed detection is more obvious, indicating that it has certain limitations in detecting small or long-distance targets. YOLOX has low target detection accuracy in multiple scenarios, especially in urban roads and nighttime scenes, where many small targets have not been successfully detected. This indicates that YOLOX performs relatively poorly in handling complex scenes and small object detection. Although the overall performance is not as good as other methods, YOLOX still shows good localization ability when detecting large targets, such as large vehicles close to the camera.

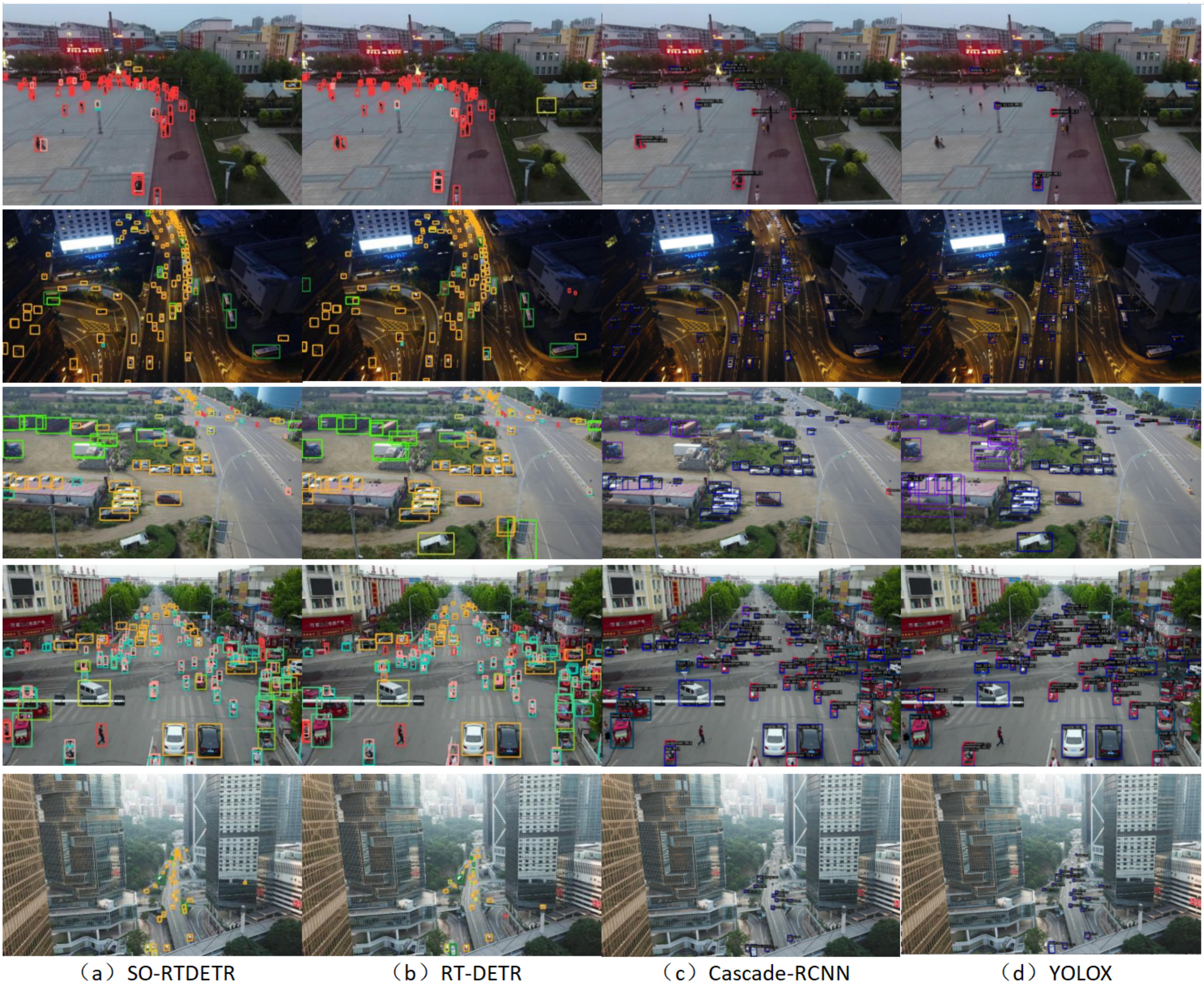


Figure 11. Comparison of detection performance between the SO-RTDETR and popular methods

SO-RTDETR achieved good detection results, mainly due to the difficulty of small object detection. A series of targeted optimization measures were proposed, such as the S2-CCFF module, PPA module, and NWD loss function. These modules help the model more effectively capture the detailed features of small targets, especially in distinguishing targets in complex backgrounds, enhancing the detection accuracy and recall rate of small targets.detection accuracy.

4.4.2. NWPU VHR-10

The proposed method was assessed against leading techniques on the NWPU VHR-10 dataset, with findings summarized in Table 5. The results indicate that two-stage methods, such as Faster R-CNN and Cascade R-CNN, surpass various one-stage approaches, especially in the NWPU dataset, which features diverse land types and often ambiguous boundaries that complicate classification [56]. Additionally, the dataset contains numerous small features and densely populated areas with complex backgrounds, including buildings and trees, which can hinder classification accuracy due to occlusions and shadows.

The two-stage approach first generates high-quality candidate regions (RoI) via a Region Proposal Network (RPN), followed by refined classification and localization of these regions. This two-step process improves target localization accuracy, particularly in cases with significant variations in target size, shape, and aspect ratio, allowing the two-stage method to capitalize on its strengths.

Overall, Faster RCNN and RetinaNet perform equally well and are suitable for scenarios with multiple target scales. YOLOX and Cascade RCNN perform well under stricter IoU conditions, but there is room for improvement on small targets. RTMDet performs well under medium to large targets and high IoU conditions, making it suitable for large target detection tasks in scenes. This is mainly due to multi-level feature fusion and efficient loss function optimization. SO-RTDETR enhances its sensitivity to small targets through innovative module design and loss function optimization, resulting in outstanding performance in small target detection tasks. Having high mAP and mAP50, it is suitable for high-sensitivity small target detection tasks.

Table 5. Comparative experiment based on VisDrone_test.

Method	mAP	mAP_0.5	mAP_0.75	mAP_s	mAP_m	mAP_l
Faster-RCNN [15]	0.512	0.878	0.55	**0.45**	0.48	0.544
Cascade-RCNN [17]	0.543	0.881	0.583	0.35	0.486	0.576
TOOD [52]	0.482	0.876	0.498	0.45	0.446	0.549
ATSS [53]	0.459	0.813	0.481	0.114	0.463	0.487
RetinaNet [20]	0.512	0.815	0.611	0.412	0.621	0.562
RTMDet [54]	0.562	0.878	0.641	0.419	**0.636**	0.571
YOLOX [55]	0.522	0.841	0.615	0.345	0.505	0.588
RT-DETR [11]	0.499	0.853	0.570	0.274	0.507	**0.679**
SO-RTDETR	**0.570**	**0.882**	**0.629**	0.403	0.546	0.678

Figure 12 compares the detection performance of SO-RTDETR with the baseline approach on the NWPU VHR-10 dataset. The first row shows results from our method, while the second row presents the baseline results. Green boxes indicate true positives (TP), blue boxes represent false positives (FP), and red boxes denote false negatives (FN). The results highlight that the baseline method has a considerable number of false positives and missed detections. In contrast, our method effectively reduces both FP and FN, leading to more accurate object identification and improved detection accuracy.

doi:10.20944/preprints202410.1980.v1

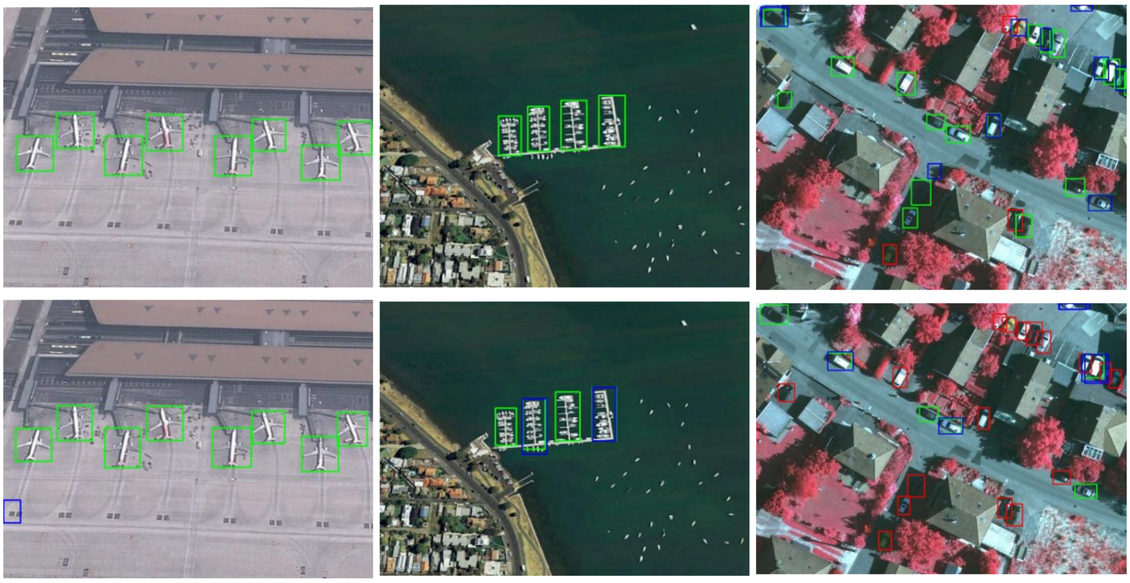


Figure 12. Comparison of detection performance, The first row displays the results from SO-RTDETR, while the second row shows the baseline method. The proposed approach significantly reduces false positives (FP) and false negatives (FN), thereby enhancing detection accuracy.

Figure 12 presents a bar chart comparing the performance of two datasets (VisDrone and NWPU VHR-10) in detection. The blue bars in the figure represent the baseline detection results, while the green bars represent the results of our method. The results depicted in the figure indicate that the SO-RTDETR enhances detection accuracy relative to the baseline approach, while also reducing the occurrence of false positives and missed detections. These findings demonstrate the effectiveness of the method in improving detection performance.

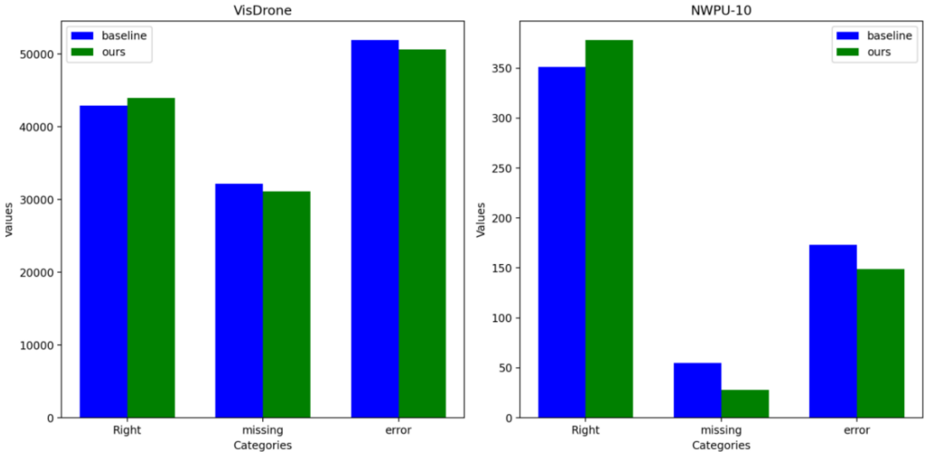


Figure 13. Detailed data comparison of detection results for VisDrone and NWPU VHR-10 datasets.

5. Conclusion

This article proposes an innovative small object detection model SO-RTDETR to address key issues in small object detection. This model combines the S2-CCFF module, PPA module, and NWD loss function, significantly improving the performance of small object detection. The main advantages of SO-RTDETR are reflected in the following aspects. Firstly, the S2-CCFF module enriches the information representation of small targets by adding S2 layers for cross-scale feature fusion, while using SPDConv to reduce computational complexity and preserve key details, successfully improving the accuracy and efficiency of detection. Secondly, the PPA module enhances the feature representation of small targets in the Backbone network through hierarchical feature fusion and attention mechanism, ensuring that key information is not lost during multiple downsampling processes, thereby improving detection performance. In addition, the introduced NWD loss function improves the tolerance of the model to boundary box disturbances by better measuring the relative position and shape differences of the boundary boxes, further enhancing the robustness of the model.

Although SO-RTDETR has achieved significant performance improvements in small object detection tasks, there are still some issues that require further research and optimization. Firstly, when dealing with extremely complex backgrounds or severe occlusions, although the model suppresses some noise, there is still a possibility of false positives or false negatives. Secondly, although SPDConv reduces computational complexity, the model still faces significant computational resource requirements when processing high-resolution images. Future research directions can focus on the following aspects: firstly, further optimizing feature fusion strategies and exploring more effective multi-scale feature extraction methods to improve adaptability to complex scenes; Secondly, by combining lightweight network structures, the computational overhead of the model can be reduced, making it more suitable for practical deployment; The third is to introduce more contextual and semantic information to enhance the model's ability to detect small targets in complex backgrounds. Through these improvements, SO-RTDETR is expected to demonstrate higher efficiency and wider applicability in practical applications.

Author Contributions: Conceptualization, Jing Liu and Yanyan Cao; Methodology, Jing Liu, Ying Wang and Chunyu Dong; Software, Chunyu Dong, Chaoping Guo and Haijing Zhang; Validation, Chunyu Dong; Visualization, Yanyan Cao; Writing – original draft, Jing Liu and Haijing Zhang; Writing – review & editing, Ying Wang and Chaoping Guo. All authors have read and agreed to the published version of this manuscript.

Funding: This research was funded by Science and Technology Research Project of Education Department of Hubei Province (B2023362) and Excellent Young and Middle aged Science and Technology Innovation Team Project for Higher Education Institutions of Hubei Province (T2023045).

Data Availability Statement: The VisDrone is available at https://github.com/VisDrone (accessed on 10 May 2024) ; The NWPU-10 is available at https://labelbox.com/datasets/nwpu-vhr-10/(accessed on 15 March 2024) ;

Acknowledgments: The author acknowledges the referees and the editor for carefully reading this paper and giving many helpful comments. The authors also express their gratitude to the reviewers for their insightful comments.

References

1. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.
2. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European conference on computer vision. Springer, 2020, pp. 213–229.
3. Girshick, R. Fast r-cnn. *arXiv preprint arXiv:1504.08083* **2015**.
4. Redmon, J. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
5. Redmon, J. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
6. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
7. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.
8. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616* **2024**.
9. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458* **2024**.
10. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* **2020**.

doi:10.20944/preprints202410.1980.v1

11. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16965–16974.
12. Miri Rekavandi, A.; Rashidi, S.; Boussaid, F.; Hoefs, S.; Akbas, E.; et al. Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art. *arXiv e-prints* **2023**, pp. arXiv–2309.
13. Xiao, J.; Wu, Y.; Chen, Y.; Wang, S.; Wang, Z.; Ma, J. LSTFE-Net: Long Short-Term Feature Enhancement Network for Video Small Object Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14613–14622.
14. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* **2022**, *190*, 79–93.
15. 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.
16. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
17. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence* **2019**, *43*, 1483–1498.
18. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 821–830.
19. 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.
20. Lin, T. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002* **2017**.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
22. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768.
23. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7036–7045.
24. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.
25. Huang, W.; Li, G.; Chen, Q.; Ju, M.; Qu, J. CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection. *Remote Sensing* **2021**, *13*, 847.
26. Chen, K.; Cao, Y.; Loy, C.C.; Lin, D.; Feichtenhofer, C. Feature pyramid grids. *arXiv preprint arXiv:2004.03580* **2020**.
27. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic feature pyramid network for object detection. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2023, pp. 2184–2189.
28. Gong, Y.; Yu, X.; Ding, Y.; Peng, X.; Zhao, J.; Han, Z. Effective fusion factor in FPN for tiny object detection. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 1160–1168.
29. Gao, T.; Niu, Q.; Zhang, J.; Chen, T.; Mei, S.; Jubair, A. Global to local: A scale-aware network for remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
30. Hu, X.; Xu, W.; Gan, Y.; Su, J.; Zhang, J. Towards disturbance rejection in feature pyramid network. *IEEE Transactions on Artificial Intelligence* **2022**, *4*, 946–958.
31. Liu, H.I.; Tseng, Y.W.; Chang, K.C.; Wang, P.J.; Shuai, H.H.; Cheng, W.H. A DeNoising FPN with Transformer R-CNN for Tiny Object Detection. *IEEE Transactions on Geoscience and Remote Sensing* **2024**.
32. Du, Z.; Hu, Z.; Zhao, G.; Jin, Y.; Ma, H. Cross-Layer Feature Pyramid Transformer for Small Object Detection in Aerial Images. *arXiv preprint arXiv:2407.19696* **2024**.
33. 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**.
34. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
36. Ma, T.; Mao, M.; Zheng, H.; Gao, P.; Wang, X.; Han, S.; Ding, E.; Zhang, B.; Doermann, D. Oriented object detection with transformer. *arXiv preprint arXiv:2106.03146* **2021**.
37. Huang, Z.; Zhang, C.; Jin, M.; Wu, F.; Liu, C.; Jin, X. Better Sampling, towards Better End-to-end Small Object Detection. *arXiv preprint arXiv:2407.06127* **2024**.
38. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered object detection in aerial images. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8311–8320.
39. Li, C.; Yang, T.; Zhu, S.; Chen, C.; Guan, S. Density map guided object detection in aerial images. In Proceedings of the proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 190–191.

40. Huang, Y.; Chen, J.; Huang, D. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2022, Vol. 36, pp. 1026–1033.
41. Du, B.; Huang, Y.; Chen, J.; Huang, D. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 13435–13444.
42. Chen, L.; Liu, C.; Li, W.; Xu, Q.; Deng, H. DTSSNet: Dynamic Training Sample Selection Network for UAV Object Detection. *IEEE Transactions on Geoscience and Remote Sensing* **2024**.
43. Zhang, Z. Drone-YOLO: an efficient neural network method for target detection in drone images. *Drones* **2023**, *7*, 526.
44. Liu, K.; Fu, Z.; Jin, S.; Chen, Z.; Zhou, F.; Jiang, R.; Chen, Y.; Ye, J. ESOD: Efficient Small Object Detection on High-Resolution Images. *arXiv preprint arXiv:2407.16424* **2024**.
45. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**.
46. Khalili, B.; Smyth, A.W. SOD-YOLOv8—Enhancing YOLOv8 for Small Object Detection in Aerial Imagery and Traffic Scenes. *Sensors* **2024**, *24*, 6209.
47. Liu, C.; Gao, G.; Huang, Z.; Hu, Z.; Liu, Q.; Wang, Y. YOLC: You Only Look Clusters for Tiny Object Detection in Aerial Images. *IEEE Transactions on Intelligent Transportation Systems* **2024**.
48. Xiao, J.; Guo, H.; Zhou, J.; Zhao, T.; Yu, Q.; Chen, Y. Tiny object detection with context enhancement and feature purification. *Expert Systems with Applications* **2023**, *211*, 118665.
49. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In Proceedings of the Joint European conference on machine learning and knowledge discovery in databases. Springer, 2022, pp. 443–459.
50. Cui, Y.; Ren, W.; Knoll, A. Omni-Kernel Network for Image Restoration. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 1426–1434.
51. Xu, S.; Zheng, S.; Xu, W.; Xu, R.; Wang, C.; Zhang, J.; Teng, X.; Li, A.; Guo, L. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. *arXiv preprint arXiv:2403.10778* **2024**.
52. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2021, pp. 3490–3499.
53. Biffi, L.J.; Mitishita, E.; Liesenberg, V.; Santos, A.A.d.; Gonçalves, D.N.; Estrabis, N.V.; Silva, J.d.A.; Osco, L.P.; Ramos, A.P.M.; Centeno, J.A.S.; et al. ATSS deep learning-based approach to detect apple fruits. *Remote Sensing* **2020**, *13*, 54.
54. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784* **2022**.
55. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* **2021**.
56. Liu, J.; Jing, D.; Zhang, H.; Dong, C. SRFAD-Net: Scale-Robust Feature Aggregation and Diffusion Network for Object Detection in Remote Sensing Images. *Electronics* **2024**, *13*, 2358.

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.