Parallel Space and Channel Attention for Stronger Remote Sensing Object Detection

Yuhui Zhao, Ruifeng Yang, Chenxia Guo, and Xiaole Chen

***Abstract*—The object detection of natural images tends to obtain advanced semantic information through multiple convolutions and pooling, ignoring the detailed information in the feature map. Pixel-level images may be the target we are looking for in remote sensing images. This article designs a new attention mechanism that can fully utilize the spatial and channel information of the image, strengthen the region of interest, and try to protect the image's original information. Find the most influential location information in the spatial dimension and the most influential feature map in the channel dimension. Strengthen important channels and positions in the feature map to make vital information stronger and weak information not lost. Combine the designed attention mechanism with existing modules to enhance YOLO-V7 detection capability. We have merged two publicly available remote sensing image datasets, increasing object types, and richer appearance features, which can better detect model performance. Experimental results on an improved dataset have shown that the enhanced model in this article can improve the detection ability of small- and medium-sized targets in complex backgrounds, with a 1% increase in mean average precision (mAP) value and a maximum improvement of 8.2% for single-class targets. Medium targets such as airports, dams, and soccer ball fields also increase by about 5%. We also conducted experiments on the DOTA1.0 dataset to demonstrate that mAP improved by 1.1%, with 13 target categories having higher APs. The improved model reduces computational complexity by 2.7%, which is very user-friendly for embedded devices.**

***Index Terms*—Attention mechanism, convolutional neural networks (CNNs), object detection, optical remote sensing images.**

I. Introduction

IN RECENT years, the rapid development of remote sensing technology has dramatically improved the quality and quantity of remote sensing images. Representative landmarks such as airports, ports, and stations and small objects such as planes, ships, and cars can all be captured, making remote sensing image object detection increasingly widely used in military, commercial, agricultural, livelihood, and other fields. Compared with natural images, object detection in remote sensing images still faces difficulties such as scale diversity, particular viewing angles, multiple small targets, multidirectional shooting, high background complexity, and rich background information. Changes in weather, scale, perspective, mutual occlusion of buildings, and similarity between objects and backgrounds can all affect the effectiveness of detection performance.

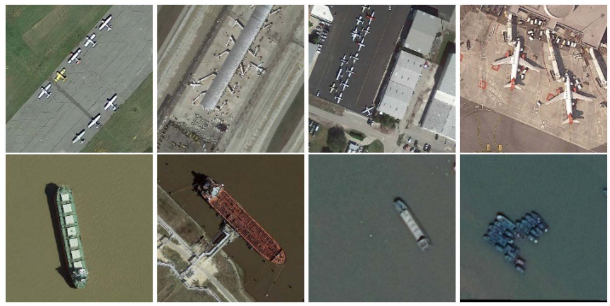

Fig. 1. Airplanes and ships captured by different satellites.

High-quality datasets and efficient network structures are crucial to obtaining good detection results, which are beneficial for model training and validation based on data-driven deep learning. The images of publicly available datasets come from various sensors with different temporal, spatial, spectral, and polarization characteristics. The same object exhibits multiple appearance characteristics. Imaging fusion technology can provide more information but is more complex and requires specialized datasets. We used a larger dataset to train the network model, improving remote sensing objects' detection ability. The widely used publicly available datasets include DOTA1.0 [1], DIOR [2], NWPU VHR-10 [3], LEVIR [4], and VEDAI [5]. As shown in Fig. 1, these images come from different satellites and exhibit various features. By merging datasets, the diversity and richness of the data are enhanced, the range of target types to be evaluated is expanded, and the number of instances available for analysis is increased. Merging datasets can help investigate the impact of multisource datasets on the accuracy and effectiveness of remote sensing target recognition and verify the generalization performance of the network.

Researchers have made progress in target detection in remote sensing images. Sun et al. [6] proposed an anchor-free method for ship target detection in HR SAR images. Zhang et al. [7] proposed a task-collaborated detector for oriented object detection in remote sensing image. We hope to achieve better results at higher baselines. The YOLO [8], [9], [10], [11] series has powerful object detection capabilities, with YOLO-V7 [12] having the best detection performance. We combine the parallel spatial attention and channel attention (P-SACA) module with YOLO-V7 to enhance the image's region of interest in both

Manuscript received 22 October 2023; revised 12 December 2023 and 21 December 2023; accepted 22 December 2023. Date of publication 26 December 2023; date of current version 10 January 2024. This work was supported by the Central Guidance on Local S&T Development Fund of Shanxi Province under Grant YDZJSX2022A027. *(Corresponding author: Ruifeng Yang.)*

Yuhui Zhao, Ruifeng Yang, and Chenxia Guo are with the School of Instrument and Electronics, the North University of China, Taiyuan 030051, China (e-mail: b20220627@nuc.edu.cn; yangruifeng@nuc.edu.cn; guochenxia@nuc.edu.cn).

Xiaole Chen is with China North Vehicle Research Institute, Beijing 100072, China (e-mail: 1529156348@qq.com).

Digital Object Identifier 10.1109/JSTARS.2023.3347235

channel and spatial dimensions when outputting anchor boxes of different scales in YOLO-V7. To address challenges associated with small objects and poor image quality encountered in remote sensing image object detection, we have introduced a space-to-depth (SPD) module [13]. The SPD module reduces the model's computational complexity and parameter size while ensuring that detection accuracy is unaffected. The P-SACA module explores the most influential channel and spatial information and multiplies it with the original feature map, emphasizing critical information but not wholly ignoring unimportant areas. Experimental results have shown that embedding the P-SACA module designed in this article and the SPD module into the YOLO-V7 model is more effective. In multiple object detection, the P-SACA module outperforms attention mechanism modules such as CBAM [14] and SE [40]. Our main contributions are summarized as follows.

1) We propose a new parallel spatial and channel attention module that fully utilizes input, spatial, and channel information. It strengthens key features while retaining weak information, making it more conducive to remote sensing target detection tasks with complex backgrounds and large-scale spans.
2) Integrate the designed attention mechanism into the YOLO-V7 model. The improved network model has a more vital exploration ability for large and complex datasets, reducing computational complexity by 2.7%. Experimental results on the DOTA1.0 dataset have shown that our model can improve mAP by 1.2%. In total, 13 types of targets can achieve higher AP.
3) Experiments on a larger dataset have shown that the improved model has more stronger object detection capabilities. It improves the detection performance of medium to large targets with fewer instances and weakens the impact of sample imbalance on detection performance. The detection accuracy is higher among the 15 target types. The mean average precision (mAP) increased by 1%. The maximum growth rate for single-category objects is 8.2%. Multiple medium-sized objects have increased by about 5%.

II. Related Work

Improving the quality of datasets and building more efficient network structures are the directions that researchers are constantly striving for.

Researchers have made many efforts to improve the quality of remote sensing images. They establish new algorithms and networks to fully utilize the multimodal information of remote sensing images, hoping to make the features of remote sensing images more prominent. Wang et al. [15] proposed a multistage self-guided separation network (MGSNet) and a self-guided network, fully utilizing the differences between the background and the target, as well as the similarity between the targets, to make the features of remote sensing images apparent. Wang et al. [16] proposed a representation enhanced status replay network (RSRNet) for multisource remote sensing image classification, which mainly solves the problem of representation bias and classifier bias accumulation and the problem of insufficient interaction of multisource information. Zhang et al. [17] proposed a spatial logical aggregation network (SLA-NET), focusing on the significance of spatial morphological differences and validating their effects through experiments. Hong et al. [18] provide a baseline solution by developing a general multimodal deep learning (MDL) framework, which dramatically inspires multimodal remote sensing image tasks. In [19], a new set of multimodal remote sensing benchmark datasets (called C2Seg dataset, including hyperspectral, multispectral, SAR) was built, facilitating the cross city semantic segmentation task research. The low-rank representation net (LRR-Net) was presented to hyperspectral anomaly detection, which did not rely on manual parameter settings to achieve better generalization performance [20]. The SpectralGPT [21] was created, which is purpose-built to handle spectral RS images using a novel 3-D generative pretrained transformer (GPT). The SpectralGPT fills the gap in applying spectral data in remote sensing. To address the issue of spectral variability in hyperspectral images, the augmented linear mixing model (LMM) is proposed [22], which applies a data-driven learning strategy in inverse problems of hyperspectral unmixing to solve spectral variability. For multimodal data of remote sensing images, Yao et al. [23] proposed a novel MDL framework by extending conventional ViT with minimal modifications, which propelled the development of land use and land cover (LULC) classification. Zhang et al. [24] proposed the interleaving perception convolutional neural network (IP-CNN) to integrate heterogeneous information and improve the joint classification performance of hyperspectral image (HSI) and light detection and ranging (LiDAR) data. Yao et al. [25] propose a novel coupled unmixing network with a cross-attention mechanism, CUCaNet for short, to enhance the spatial resolution of HSI using higher spatial-resolution multispectral image (MSI).

Insufficient data and imbalanced samples are important factors that constrain the development of deep learning. To solve the problem of scarce training data, Tai et al. [26] proposed the connection-free attention module, which can transmit the sharing features of electro-optical and SAR images from the source network to the target network for information supplement. Because of the limited effective offshore ship training samples obtained and the severe imbalance between positive and negative examples, Zhuang et al. [27] proposed the structured sparse representation model to realize more effective and robust offshore ship detection under the condition of a small sample set.

The YOLO algorithm has been improved by many researchers and can adapt to various object detection tasks. Shao et al. [28] constructed a night navigation ship dataset and proposed an improved algorithm called TASFF-YOLOV5, which achieved good ship detection results. Aiming at bridge detection in aerial images, Guo et al. [29] proposed a directional bridge detection model with water body segmentation as an auxiliary task to guide bridge positioning. It combines the advantages of semantic-segmentation-based supplementary supervision, water constraints, and instance switching-based data augmentation to improve detection results. Zhu et al. [30] proposed TPH-YOLOV5 to solve the problems of target density and motion ambiguity in UAV target recognition. It replaced the original prediction heads with transformer prediction heads to achieve better results. R. S. et al. [13] proposed a new CNN building block,

called SPD-Conv, for the problem that the performance of CNNs would decline rapidly under challenging tasks with low image resolution or small objects. Lin et al. [31] proposed a dynamic object detection framework named Dynamic-det for YOLO-V7. Through adaptive reasoning, the dynamic model can obtain significant accuracy and computational efficiency. Zheng et al. [32] improved YOLO-V7 by adding an attention mechanism and replacing the loss function, which can effectively realize the detection of small objects under complex backgrounds. Hussain et al. [33] proposed a framework for autonomous rack inspection based on computer vision around the YOLOv7 architecture. Zhu et al. [30] proposed an improved object detection algorithm for YOLO-V5 UAV capture scenes, which can solve the problem of multiple small and dense objects and complex backgrounds in high-altitude photography. A feature enhancement block (FE-Block) is first presented to generate adaptive weights for different receptive field features by convolution, assigning significant weights to shallow feature maps to improve small object feature extraction ability [34]. Sun et al. [35] proposed a YOLO-based arbitrary-oriented SAR ship detector using bidirectional feature fusion and angular classification (BiFA-YOLO).

Researchers hope that well-designed attention mechanisms can better utilize images. The CBAM module infers the attention map along two independent dimensions (channel and space), and then, multiplies the attention map with the input feature map for adaptive feature optimization. The authors demonstrated through experiments that the CBAM module is practical in classification and object detection networks. The SE module aims to assign different weights to different image positions from the perspective of the channel domain through a weight matrix to obtain more critical feature information. BotNet [41] is a conceptually simple yet powerful backbone architecture incorporating self-attention for multiple computer vision tasks. It replaces spatial convolution with global self-attention in ResNet's last three bottleneck blocks without making any other changes. It also reduces parameters and minimizes latency overhead, achieving good results in instance segmentation and object detection. The application of the C3TR module has achieved good results in object detection tasks with a multihead self-attention module and position encoding. The coordinate attention factorizes (CA) [42] channel attention into two 1-D feature encoding processes that aggregate features along the two spatial directions, which performs well in tasks such as image classification, object detection, and semantic segmentation.

In summary, current research on object detection in remote sensing images mainly focuses on small targets. Improving the quality of remote sensing images and fully utilizing multimodal information effectively enhance remote sensing image tasks. Many publicly available datasets with small object instances exhibit the characteristic of imbalanced samples. These have greatly improved the detection effect of small targets. However, the number of medium-sized targets is relatively small, and the background is complex. But medium-sized targets are landmark buildings with fixed positions, essential in navigation and positioning. CNNs obtain advanced semantic information about images through continuous convolution and pooling, but they may lose detailed information. A well-designed network structure can enhance feature extraction capabilities. The attention mechanism designed in this article can search for more influential feature maps in the channel dimension and essential positions in the spatial dimension. Enhance the input feature map in parallel from both channel and spatial dimensions, amplify important information, and retain detailed information. We embedded the designed attention mechanism into YOLO-V7, effectively improving the detection performance of small- and medium-sized targets under imbalanced sample conditions.

III. Dataset-Making and Object-Detection Network

This article uses two remote sensing datasets, DIOR and DOTA1.0, to form a more complex dataset, DIOR&DOTA. The new dataset contains 24 target categories and 380 754 annotated instances. The same target has more appearance features. In addition, we performed image augmentation to simulate natural scenes, including conventional rotation and scaling, weather conditions like rainy and foggy days, and motion blur.

This article employs the YOLO-V7 model as the baseline. YOLO-V7 has good detection performance for small objects such as airplanes, cars, and ships, with an accuracy of up to 90%. Due to imbalanced samples and significant differences in object size, many types of detection perform poorly. The method proposed in this article enhances the detection ability of complex background targets and effectively reduces the computational complexity. Uneven data distribution will guide the network to learn more features of multisample objects, weakening the detection ability of objects with fewer samples. The improved model has more vital learning ability and performs better on larger datasets under the same training conditions. Because it fully utilizes the information in channels and space, it can weaken the negative impact caused by imbalanced samples.

A. Datasets Preparation

The indispensability of datasets in the realm of deep learning is self-evident. Remote sensing images predominantly originate from satellites, encompassing various sensors and platforms, such as Google Earth, JL-1 satellite, and GF-2 satellite. Therefore, remote sensing images exhibit significant differences, exacerbating the challenges associated with target detection. Image fusion is employed to augment the detected attributes of the tested object and enhance its target features. However, this method requires high technical requirements. Designing a more comprehensive dataset and a network with stronger feature extraction capabilities can learn more features so that the model can effectively capture objects with different appearance features at multiple scales.

The DOTA1.0 dataset encompasses 2 806 aerial images and 188 282 instances. These images exhibit varying pixel dimensions, spanning from 800∗800 to 4000∗4000, accommodating objects of diverse sizes, orientations, and shapes. It has 15 target categories. The DIOR datasets comprise 23 463 images, categorized into 20 distinct target classes. The images are all 800∗800 in size and collected under different imaging conditions, including changes in weather, season, and image quality.

Within the DOTA1.0 dataset, 98% of the targets exhibit dimensions below 300 pixels, and 57% of the marks are less than

TABLE I
OPTICAL REMOTE SENSING IMAGE DATASET

Datasets	Categories	Images	Instances	Image width	Year
TAS	1	30	1319	792	2008
SZTAKI-INRIA	1	9	665	~800	2012
NWPU VHR-10	10	800	3775	~1000	2014
VEDAI	9	1210	3640	1024	2015
UCAS-AOD	2	910	6029	1280	2015
DLR 3K Vehicle	2	20	14235	5616	2015
HRSC2016	1	1070	2976	~1000	2016
RSOD	4	976	6950	~1000	2017
DOTA1.0	15	2806	188282	800-4000	2017
DIOR	20	23463	192472	800	2018
DIOR&DOTA1.0	24	44510	380754	800&1024	2023

TABLE II
DIOR&DOTA CATEGORIES

C1	airplane	C2	airport	C3	Baseball field	C4	Basketball court
C5	bridge	C6	chimney	C7	dam	C8	Expressway-Service-area
C9	Expressway-toll-station	C10	Golf field	C11	Ground-track-field	C12	harbor
C13	overpass	C14	ship	C15	stadium	C16	Storage tank
C17	Tennis court	C18	Train station	C19	vehicle	C20	windmill
C21	helicopter	C22	roundabout	C23	soccer-ball-field	C24	swimming-pool

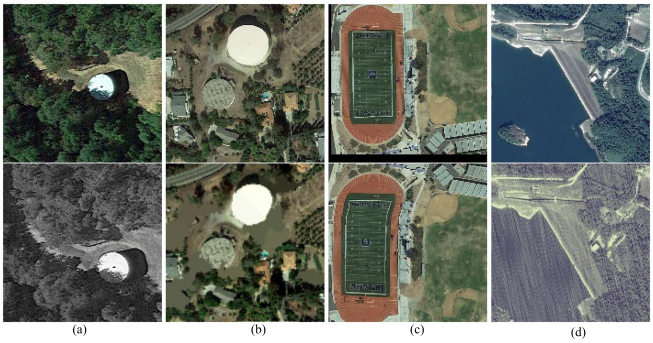

Fig. 2. Image enhancement results. (a) Turns the color image into a gray image. (b) Reduces image quality. (c) Distorts the image. (d) Simulates rainy days.

50 pixels. Similarly, in the DIOR dataset, approximately three-quarters of the target types possess dimensions below 300 pixels, and a majority of the eight targets are less than 100 pixels in size. The significant presence of a large number of small targets in annotated datasets makes networks inclined toward small target detection, which helps obtain relevant features unique to small targets. However, the uneven distribution of samples increases the difficulty of detecting large-sized objects.

The comparison between DIOR&DOTA and other popular datasets is shown in Table I. The newly established dataset has made significant progress in target categories and annotated instances. Table II shows the target types.

We randomly enhance the image to simulate various real-world scenarios, including rainy, foggy, snowy, and motion blur, and perform random rotation, translation, brightness changes, and other conventional image enhancements. As shown in Fig. 2, our dataset's richness and authenticity have been enhanced through image enhancement, enabling a more accurate representation of the real-world scenes depicted in the test images.

B. YOLO-V7

Compared to previous algorithms within the YOLO series, the primary innovation of YOLO-V7 lies in integrating model reparameterization into the network architecture. The concept of reparameterization was initially introduced within the REPVGG [36] framework, serving as a foundation for the novel approach of YOLO-V7. Furthermore, YOLO-V7 presents a novel network architecture that exhibits high efficiency, thereby enhancing the extraction of image features. The significant highlight of YOLO-V7 is the ELAN network architecture design, characterized by its simplicity and efficiency. As an efficient aggregation network, ELAN follows a similar concept to Resnet [37]. It divides the input feature map into two paths: one path undergoes multiple iterations of the CBS module to extract advanced semantic features. In contrast, the other path passes through the CBS module once to better preserve positional features. The SPPCPS module combined with SPP [38] and CBS [39] effectively expands the receptive field, enabling the algorithm to adapt to images of different resolutions. Finally, the feature maps from both approaches are combined through superposition. YOLO-V7 proposes a training method involving auxiliary heads to improve accuracy without affecting inference time. These auxiliary heads operate solely during training, optimizing the balance between inference time and training costs. Fig. 3 shows the network structure of YOLO-V7.

IV. IMPROVEMENTS TO YOLO-V7

As shown in Fig. 4, we embedded the P-SACA module in the last layer of the YOLO-V7 backbone and the SPD module in front of the head. This structural design can enhance the feature extraction ability of the model without increasing computational complexity.

After multiple convolution operations, the size of the feature map gradually decreases, and the number of channels increases. Using the P-SACA module in the last layer of the feature extraction network will increase the minimum computational complexity but significantly impact the subsequent network. The P-SACA calculates spatial attention and channel attention separately. When extracting channel features, the most influential feature channel is selected. Spatial features will select the feature

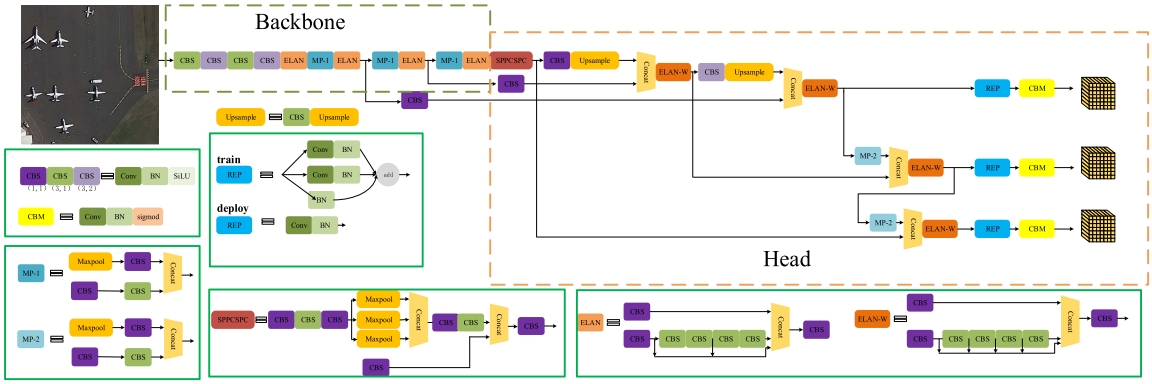

Fig. 3. YOLO-V7 network structure.

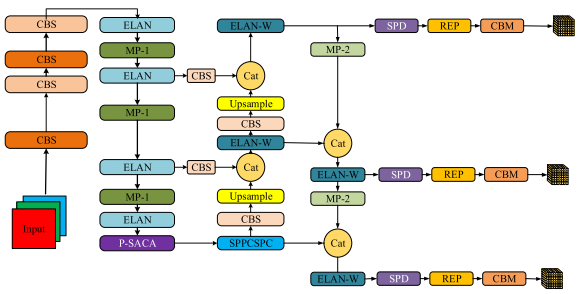

Fig. 4. SPD-P-SACA with YOLO-V7.

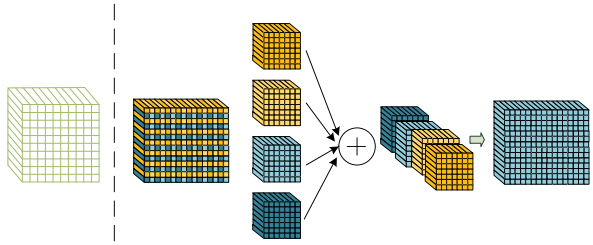

Fig. 5. SPD module.

area of greatest interest. Finally, the feature maps obtained from the two attention modules are overlaid with corresponding pixels to output the feature maps.

The SPD module carries out spatial information recombination and channel splicing for the feature map to reduce the size of the feature map and the amount of calculation. The SPD module performs interval sampling on the pixel points of the feature map, and the larger feature map yields four smaller feature maps. The feature map size decreases, and the dimension increases, but the information is not lost. It also plays a vital role in reducing computational complexity.

The experimental results show that our improved network has good detection performance for small- and medium-sized objects.

A. *Space-to-Depth (SPD)*

There is a large amount of redundant information when detecting small objects in high-resolution images. However, each pixel is essential if the image's resolution is low or the object's size is small. As the depth of the network continues to increase, convolution and pooling increase the receptive field of each pixel in the feature map, enhancing the ability to represent abstract features while gradually losing shallow spatial information. The performance of the detector deteriorates. Consequently, the multilayer feature maps fail to provide high-level semantic features and fine-grained spatial information for accurate target localization. During the down-sampling process, the semantic information of small targets gradually diminishes. Large-scale targets with rich detail features require more robust semantic information for accurate classification, while detecting small-scale targets necessitates more precise spatial information for precise localization. These challenges underscore the limitations of the CNN architecture, where convolutional layers and pooling layers contribute to the loss of fine-grained details.

To address these limitations, we introduce the SPD module. Fig. 5 illustrates the image interval sampling approach employed by the SPD module to decrease the image size. The SPD module is similar to pooling in that it extracts local information from the image but does not directly discard the remaining information. The SPD concatenates the reconstructed feature maps in order based on channel dimensions without losing any information. This approach effectively integrates spatial and channel information, preserving all relevant data while minimizing computational demands.

B. *Attention Module With Parallel Spatial Attention and Channel Attention (P-SACA Model)*

Incorporating an appropriate attention mechanism into the network can significantly enhance its feature extraction capability. We conducted experiments using recognized and influential attention mechanisms such as CBAM, SE, Botnet, and other widely acknowledged effective methods to improve detection efficiency. However, these attempts did not yield satisfactory results. We define channel attention and spatial attention. Spatial attention can make the neural network pay more attention to the pixel region, which plays a decisive role in image classification, while ignoring the unimportant part. Channel attention can deal with the distribution relationship of channels in feature mapping. The superposition of two attention allocation forms enhances attention mechanisms' impact on model performance.

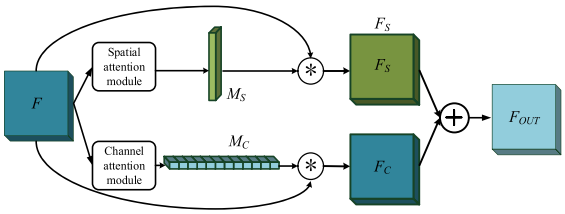

Fig. 6. P-SACA module.

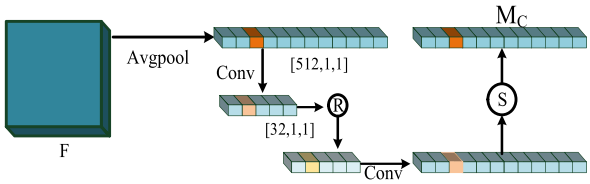

Fig. 7. Channel attention module.

Using the last layer feature map $F^{(C*H*W)}$ of the backbone as input, P-SACA generates a 1-D channel attention feature map $M_C{}^{(C*1*1)}$ and a 2-D spatial attention feature map $M_S{}^{(1*H*W)}$ in parallel. The entire process of extracting regions of interest can be summarized as follows:

$$F_c = M_c(F) \otimes F \tag{1}$$

$$F_s = M_s(F) \otimes F \tag{2}$$

$$F_{\text{out}} = F_c \oplus F_s \tag{3}$$

where $\otimes$ represents element by element multiplication and $\oplus$ represents element by element addition. Multiply the obtained M_c and M_s with the original image to obtain F_c and F_s, highlighting the region of interest without losing detailed information. Finally, the two feature maps obtained are overlaid element by element to get the output feature map F_{out}. The structure of the P-SACA module is shown in Fig. 6.

1) Channel Attention Module: Each channel of the multi-channel feature map is a detector for different features. The attention mechanism we designed explores the relationships between channels and identifies the most influential feature maps. To effectively calculate channel attention, we used the average pooling method to compress the spatial dimension of the input feature map. Then, we obtain the channel attention vector from the adaptive pooling feature map through convolution and activation functions, scale it using the sigmoid function, and apply it to the input feature map. Fig. 7 shows the execution process of the module in channel attention.

The channel attention is computed as

$$M_c = \sigma\left(f^{1\times1}\left(r\left(f^{1\times1}\left(\text{AvgPool}\left(F\right)\right)\right)\right)\right) \tag{4}$$

where σ represents the sigmoid function, and r represents the RELU activation function. The $f^{1\times1}$ represents a convolution operation with a filter size of 7×7.

2) Spatial Attention Module: Channel attention focuses on which feature detector is the most important, while spatial attention is more interested in which region to focus on in the 2-D plane. When calculating spatial attention, to extract as much important information as possible, we first perform maximum pooling and average pooling on the channel dimension of the feature map; superposition the obtained two 2-D feature maps into a new feature map with a channel number of 2; use 7×7 convolution check to convolution the input feature map, and use the sigmoid function to scale it to obtain M_s. The calculation process can be summarized as follows:

$$M_s = \sigma\left(f^{7\times7}\left[\text{AvgPool}\left(F\right);\text{MaxPool}\left(F\right)\right]\right) \tag{5}$$

where $[\text{AvgPool}(F);\text{MaxPool}(F)]$ represents a 3-D feature map formed by stacking two 2-D feature maps according to channel dimensions. The spatial attention module is shown in Fig. 8.

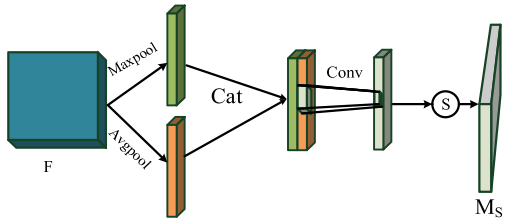

Fig. 8. Spatial attention module.

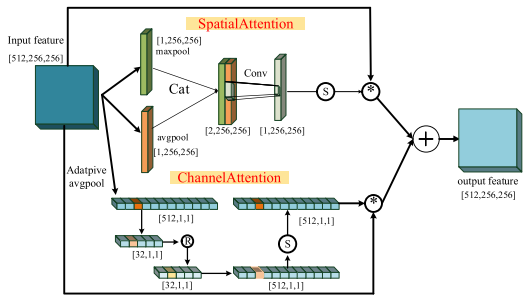

Fig. 9. P-SACA mode.

Finally, we will multiply the calculated M_s and M_c pixel by pixel with the original feature map to obtain F_s and F_c, respectively. The *H*, *W*, and *C* of F_s and F_c are consistent with F. Add F_c and F_s pixel by pixel to obtain the desired output feature map F_{out}. Fig. 9 illustrates the specific change process of using the P-SACA module feature map in the last layer of the backbone.

Compared to the CBAM module of concatenating spatial attention and channel attention, this article preserves more image information through parallel connection. Multiply the feature maps with important details on the channel and the feature maps with important information in the space with the original input feature maps, and then, cover them. By doing so, the image's original data will be saved to a greater extent while highlighting important information, which is meaningful for information-rich remote sensing images.

V. Experiments and Results

A. Experimental Environment

The experiment was conducted on a Windows 10 system with an Intel(R) Core (TM) i9-9940X CPU at 3.30 GHz and a GeForce RTX 3090 GPU with 24-GB memory. The entire framework was implemented in python. The implementation heavily relied on various libraries such as torch, SciPy, imgaug, matplotlib, opencv-python, and NumPy.

TABLE III
DIOR EXPERIMENTS ON MULTIPLE OBJECT DETECTION ALGORITHMS ON DIOR DATASETS

Neural network /classes	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20	mAP
R-CNN	35.6	43.0	53.8	62.3	15.6	53.7	33.7	50.2	33.5	50.1	49.3	39.5	30.9	9.1	60.8	18.0	54.0	36.1	9.1	16.4	37.7
RICNN	39.1	61.0	60.1	66.3	25.5	63.3	41.1	51.7	36.6	55.9	58.9	43.5	39.0	9.1	61.1	19.1	63.5	46.1	11.4	31.5	44.2
RICAOD	42.2	69.7	62.0	79.0	27.7	68.9	50.1	60.5	49.3	64.4	65.3	42.3	46.8	11.7	53.5	24.5	70.3	53.3	20.4	56.2	50.9
CSFF	57.2	79.6	70.1	87.4	46.1	76.6	62.7	82.6	73.2	78.2	81.6	50.7	59.5	73.3	63.4	58.5	85.9	61.9	42.9	86.9	68.0
RIFD-CNN	56.6	53.2	79.9	69.0	29.0	71.5	63.1	69.0	56.0	68.9	62.4	51.2	51.1	31.7	73.6	41.5	79.5	40.1	28.5	46.9	56.1
Fast R-CNN	44.2	66.8	67.0	60.5	15.6	72.3	52.0	65.9	44.8	72.1	62.9	46.2	38.0	32.1	71.0	35.0	58.3	37.9	19.2	38.1	50.0
Faster R-CNN	53.6	49.3	78.8	66.2	28.0	70.9	62.3	69.0	55.2	68.0	56.9	50.2	50.1	27.7	73.0	39.8	75.2	38.6	23.6	45.4	54.1
SSD	59.5	72.7	72.4	75.7	29.7	65.8	56.6	63.5	53.1	65.3	68.6	49.4	48.1	59.2	61.0	46.6	76.3	55.1	27.4	65.7	58.6
YOLOv3	72.2	29.2	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1	61.1	44.9	49.7	87.4	70.6	68.7	87.3	29.4	48.3	78.7	57.1
Faster R-CNN with FPN-ResNet50	54.1	71.4	63.3	81.0	42.6	72.5	57.5	68.7	62.1	73.1	76.5	42.8	56.0	71.8	57.0	53.5	81.2	53.0	43.1	80.9	63.1
Faster R-CNN with FPN-ResNet101	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0	76.8	46.4	57.2	71.8	68.3	53.8	81.1	59.5	43.1	81.2	65.1
Mask RCNN with FPN-ResNet50	53.8	72.3	63.2	81.0	38.7	72.6	55.9	71.6	67.0	73.0	75.8	44.2	56.5	71.9	58.6	53.6	81.1	54.0	43.1	81.1	63.5
Mask RCNN with FPN-ResNet101	53.9	76.6	63.2	80.9	40.2	72.5	60.4	76.3	62.5	76.0	75.9	46.5	57.4	71.8	68.3	53.7	81.0	62.3	43.0	81.0	65.2
RetinaNet-ResNet50	53.7	77.3	69.0	81.3	44.1	72.3	62.5	76.2	66.0	77.7	74.2	50.7	59.6	71.2	69.3	44.8	81.3	54.2	45.1	83.4	65.7
RetinaNet-ResNet101	53.3	77.0	69.3	85.0	44.1	73.2	62.4	78.6	62.8	78.6	76.6	49.9	59.6	71.1	68.4	45.8	81.3	55.2	44.4	85.5	66.1
SCRDet++(RetinaNet) *-ResNet50	64.3	78.9	73.2	85.7	45.8	75.9	68.4	79.2	68.9	77.6	77.8	56.7	62.1	70.3	67.6	60.4	80.9	63.7	44.4	84.5	69.3
FPN*	66.5	83.0	71.8	83.0	50.4	75.7	70.2	81	74.8	79.0	77.7	55.2	62.0	72.2	72.1	68.6	81.2	66.0	54.5	89.0	71.7
SCRDET++(FPN*)-ResNet50	66.3	83.3	74.3	87.3	54.4	77.9	70.0	84.2	77.9	80.7	81.2	56.7	63.7	73.2	71.9	71.2	83.4	62.2	55.6	90.0	73.2
PANet-ResNet50	61.9	70.4	71.0	80.4	38.9	72.5	56.6	68.4	60.0	69.0	74.6	41.6	55.8	71.7	72.9	62.3	81.2	54.6	48.2	86.7	63.8
PANet-ResNet101	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0	73.4	45.3	56.9	71.7	70.4	62.0	80.9	57.0	47.2	84.5	66.1
Corner-net	58.8	84.2	72.0	80.8	46.4	75.3	64.3	81.6	76.3	79.5	79.5	26.1	60.6	37.6	70.7	45.2	84.0	57.1	43.0	75.9	64.9
deformable-detr-100	75.6	79.0	68.7	85.3	39.8	77.0	58.5	57.0	56.3	75.7	65.7	32.5	57.6	59.7	52.1	51.7	82.2	55.9	43.2	72.9	62.3
YOLO-V5	84.1	74.4	75.4	89.1	43.6	77.9	57.1	62.2	59.9	72.5	73.9	58.9	56.9	88.6	69.9	76.7	87.6	56.0	54.5	80.2	70.0
YOLOV7-tiny	92.2	73.0	82.9	88.8	39.7	76.0	52.6	65.7	61.6	69.1	76.5	61.5	55.2	89.4	79.6	75.8	90.4	55.2	51.0	78.5	70.7
YOLO-V8n	79.1	62.8	73.3	85.3	32.9	74.2	43.7	56.8	51.2	65.3	62.9	59.6	53.4	87.9	54.0	69.7	86.3	43.4	45.4	74.8	63.1

B. Implementation Details

In this study, we adopt the learning strategy of the YOLO-V7 model. The initial learning rate is set to 0.01, and a cyclic learning rate of 0.1 is employed. Set the batch size to 6. The experiment needs 150 epochs to converge the model. The input image size is 640 × 640. To fairly compare the models, all experiments did not use pretrained weights.

We retained the dataset allocation method and directly merged the original DOTA1.0 and DIOR datasets for training, validation, and testing. This approach ensures consistency in data partitioning with the original datasets, facilitating a more accurate comparison.

C. Evaluation Measures

The average precision is commonly used by object detection methods to evaluate the merits of the algorithms. This metric is also used in this article to compare with other methods.

The precision is the percentage of positive samples among all samples, and it is calculated as

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

where true positive (TP) is the number of positive samples detected correctly, and false positive (FP) is the number of samples incorrectly judged as positive by a negative sample.

The average precision (AP) is the area enclosed by the recall rate of the correctness curve and the x-axis. It is expressed as

$$\mathrm{AP} = \int_0^1 P(y)\, d_y \tag{7}$$

where y is the recall curve under different intersection ratio thresholds.

The mAP refers to the average of the AP of all categories contained in the detection target, which is calculated as

$$\mathrm{mAP} = \frac{\sum_n^N \mathrm{AP}_n}{N}. \tag{8}$$

D. Results and Analysis

This article demonstrates through experimental analysis on the DIOR dataset that the YOLO-V7 network has the best detection performance for remote sensing targets, as shown in Table III. Therefore, we chose YOLO-V7 as our baseline. Categories C1–C24 correspond to Table II. The mAP represents the average accuracy of all types, compiled in percentage format. The following tables all use the same method.

We combined the P-SACA module with the YOLO-V7 algorithm and introduced an SPD module in the detection head. As shown in Table IV, we compared the designed attention mechanism with some proven methods through experiments. Table V shows the parameter count and computational complexity of these methods. Despite our network design exhibiting slightly lower average accuracy, it requires less computational power. Notably, we have eight types of objects that have better detection performance.

As shown in Table VI, we conducted experimental verification on the DOTA1.0 dataset. Our model was based on YOLO-V7 and achieved an improvement in AP among 13 target categories, with an increase of 1.1% in mAP.

TABLE IV
COMPARISON OF NETWORK MODEL IMPROVEMENT METHODS ON DIOR DATASETS

model	YOLO V7	YOLOV7-CBAM	YOLOV7-SE	YOLOV7-BOTNET	YOLOV7-C3TR	YOLOV7-SEC3	YOLOV7-DynamicDer-main	YOLOV7-P_SACA	YOLOV7-SPD	SPD-P-SACA
mean	79.4	73.8	75.3	77.9	75.1	75.5	75.6	78.0	78.9	78.7
C1	94.7	92.7	94.1	94.6	93.3	93.6	94.5	94.9	94.3	93.9
C2	88.0	76.6	78.6	84.7	79.9	80.3	78.4	82.9	87.2	84.4
C3	85.9	79.5	84.3	85.9	84.0	85.0	83.1	85.1	84.4	85.8
C4	92.7	90.0	91.1	92.2	90.9	90.5	91.2	92.1	92.0	91.9
C5	51.9	43.9	46.4	49.0	46.4	46.3	47.6	50.7	52.3	50.5
C6	84.1	79.1	79.6	82.4	79.4	79.3	81.8	82.7	83.1	83.1
C7	72.8	52.4	61.3	67.7	57.6	60.1	59.5	66.7	72.0	69.0
C8	82.3	71.5	71.9	77.9	71.8	72.0	74.5	80.1	82.7	79.2
C9	75.7	68.2	67.8	76.6	68.5	68.4	69.4	75.1	74.9	76.6
C10	83.8	80.8	80.4	81.6	80.5	80.6	79.9	81.9	83.3	84.1
C11	82.9	76.9	78.4	81.7	79.0	80.4	77.9	81.7	82.8	83.1
C12	66.2	62.8	64.9	65.2	64.6	65.1	63.4	64.7	64.7	65.2
C13	64.9	60.0	61.4	63.5	60.8	60.6	61.8	64.6	65.3	65.4
C14	91.8	91.7	91.5	91.6	91.5	91.3	91.4	91.2	91.4	91.3
C15	79.3	82.9	80.9	83.0	81.5	82.2	79.3	80.5	80.3	81.9
C16	84.5	81.9	80.7	82.7	81.1	80.9	82.9	82.3	83.8	83.1
C17	92.4	91.5	91.8	92.3	91.9	91.4	91.5	91.7	92.2	92.7
C18	66.2	52.6	59.4	60.8	57.2	59.8	58.9	65.0	64.5	68.1
C19	64.4	59.5	60.2	61.6	59.7	59.6	62.5	62.4	63.1	62.4
C20	82.9	91.6	81.6	83.1	82.0	81.7	82.5	82.6	83.0	83.2

The red mark is just to emphasize that different methods have different effects on different objects.

TABLE V
COMPARISON OF THE PARAMETERS AND COMPUTATION OF THE IMPROVED MODEL

Model	YOLOV7	YOLOV7-CBAM	YOLOV7-SE	YOLOV7-BONNET	YOLOV7-C3TR	YOLOV7-SEC3	YOLOV7-DynamicDer-main	YOLOV7-P_SACA	YOLOV7-SPD	SPD-P_SACA
layers	314	426	320	494	335	337	846	323	318	329
parameters	36584258	40396626	56212290	48461666	42224962	60933594	77899012	37503908	37633346	38733187
GFLPS	105.4	112.8	139.2	114.7	108.1	143.0	183.2	105.2	100.8	102.6

TABLE VI
COMPARISON OF NETWORK MODEL IMPROVEMENT METHODS ON DOTA1.0

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Roi-Trans [43]	88.6	78.5	43.4	75.9	68.8	73.6	83.6	90.7	77.2	81.4	58.4	53.5	62.8	58.9	47.7	69.6
S^2A-Net [44]	89.1	82.8	48.4	71.1	78.1	78.4	87.3	90.8	84.9	85.6	60.4	62.6	65.3	69.1	57.9	74.1
PIoU [45]	80.9	69.7	24.1	60.2	38.3	64.4	64.8	90.9	77.2	70.4	46.5	37.1	57.1	61.9	64.0	60.5
DRN* [46]	89.5	83.2	49.0	62.2	70.6	74.3	84.0	90.7	84.6	85.4	55.8	60.8	71.6	68.8	63.9	72.9
CAF [47]	89.3	81.7	51.8	67.2	80.0	78.2	84.4	90.8	83.4	85.5	54.8	67.7	73.0	70.2	64.9	75.0
ProbIoU [48]	89.1	72.2	46.9	62.2	75.8	74.7	86.6	89.6	78.3	83.1	55.8	64.0	65.5	65.4	46.2	70.0
PolarDet [49]	89.7	87.0	45.3	63.3	78.4	76.6	87.1	90.8	80.6	85.9	61.0	67.9	68.2	74.6	68.7	75.0
YOLOV7	73.4	87.8	93.8	76.6	89.7	85.5	70.2	67.8	94.8	63.0	80.8	60.8	71.4	50.0	46.5	74.1
Ours	73.2	**88.7**	93.5	**76.9**	**89.9**	**85.8**	**70.7**	**68.8**	**94.9**	**63.9**	**81.2**	**61.9**	**73.5**	**52.3**	**52.4**	**75.2**

The bold values are only meant to emphasize in which categories our model performs better.

Table VII presents a performance comparison between YOLO-V7 and our model on the DIOR&DOTA datasets. The first two rows correspond to the results obtained from the validation datasets, while the last two rows compare a larger test dataset. The experimental results indicate that our improved network model has a more vital exploration ability for complex large datasets with reduced computational complexity. The mAP shows a 1% increase, with higher detection accuracy observed across 15 target types. The train station exhibits the most significant improvement, with a remarkable increase of 8.2%. Medium-sized targets such as airports, dams, and soccer ball fields have also increased by about 5%, significantly impacting navigation.

For the target detection task of remote sensing images, detecting objects with drastic scale changes in complex backgrounds and detecting objects with tiny targets are the biggest challenges. Figs. 10 and 11 provide more direct evidence that our detection performance is better.

TABLE VII
COMPARISON OF NETWORK MODEL IMPROVEMENT METHODS

model	mAP	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
YOLO-V7 (valid)	78.3	93.5	84.6	89.3	83.8	86.3	89.6	65.5	74.2	64.2	78.9	85.1	74.0
SPD-P_SACA (valid)	**79.3**	**93.6**	**87.5**	**89.5**	**83.8**	86.1	89.5	**71.0**	**74.5**	**66.6**	**80.0**	83.7	**76.2**
YOLO-V7 (test)	70.4	79.1	81.0	79.1	90.4	77.4	76.8	56.4	61.0	54.9	77.2	76.4	63.1
SPD-P_SACA (test)	**71.2**	**79.4**	**85.0**	77.9	89.8	**78.2**	76.7	**61.6**	**61.7**	**57.1**	**78.7**	73.2	**64.4**
model	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**	**C20**	**C21**	**C22**	**C23**	**C24**	
YOLO-V7 (valid)	68.9	95.1	96.6	80.1	96.5	61.3	82.3	82.5	49.9	84.1	55.9	57.5	
SPD-P_SACA (valid)	**69.0**	**95.2**	**96.7**	80.0	**96.5**	**69.5**	82.0	**82.7**	**50.5**	84.0	**58.9**	56.7	
YOLO-V7 (test)	58.3	86.3	75.3	87.3	94.3	63.5	66.4	78.7	48.5	82.0	35.3	41.2	
SPD-P_SACA (test)	**59.1**	**87.4**	73.0	87.2	**94.4**	**68.4**	66.2	**79.9**	**48.8**	81.4	**40.2**	39.1	

The bold values are only meant to emphasize in which categories our model performs better.

(a) (b) (c)

Fig. 10. Comparison of small target detection results in complex background. The first line is our model, and the second is the YOLO-V7 model. (a) Detected ground track fields with tiny pixels. (b) Detected multiple ground track fields and bridges. (c) Detected more than two vehicles.

(a) (b) (c) (d)

Fig. 11. Detection results of large objects in complex background. The first line is our model, and the second is the YOLO-V7 model. (a) Large receptive field. (b) Detected a dam. (c) Detected more than one tennis court. (d) Detected the bridge.

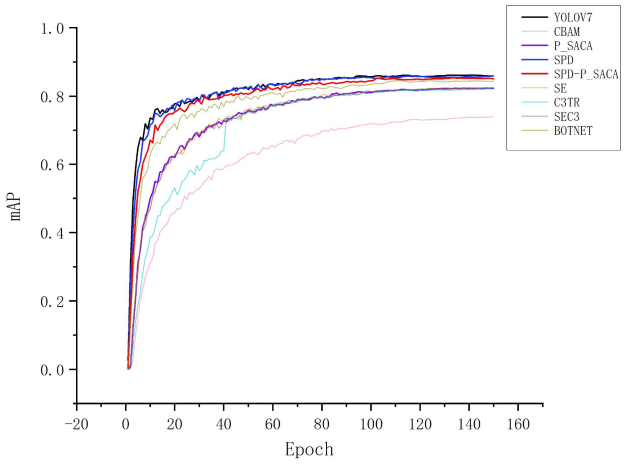

Fig. 12. Comparison of the mAP convergence speed.

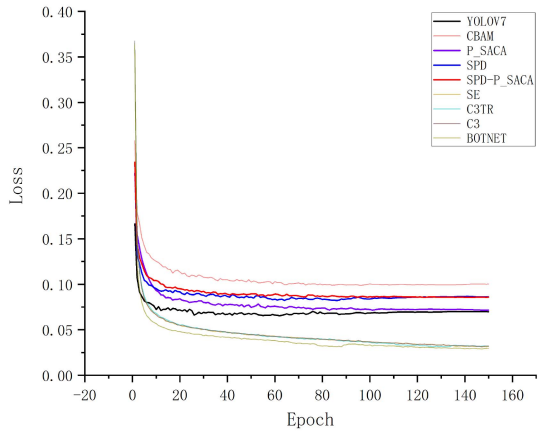

Fig. 13. Comparison of loss convergence speed.

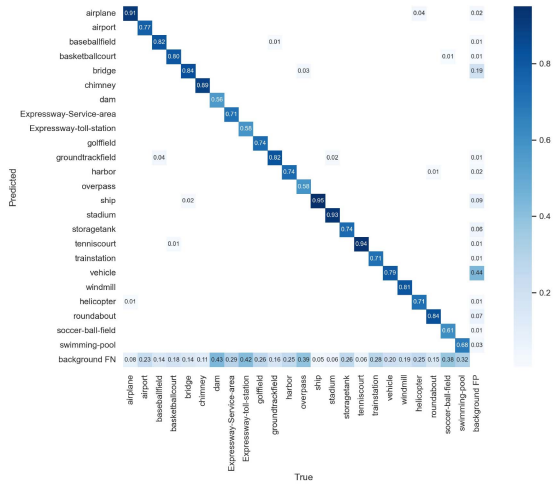

Fig. 14. Confusion matrix of detection results.

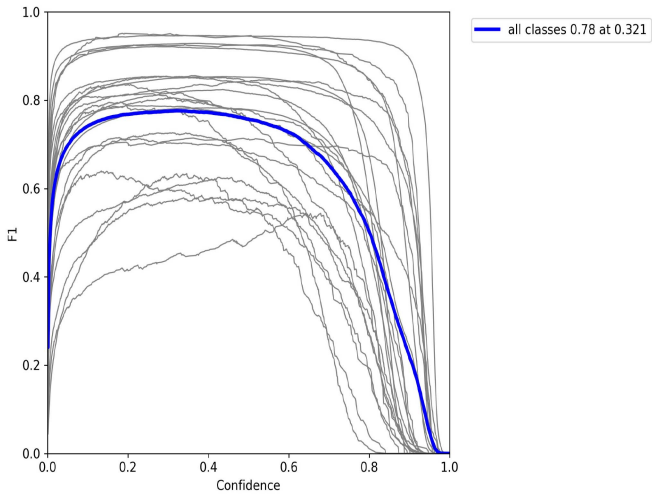

Fig. 15. F1-score.

In Fig. 10(a) and (b), our model can detect ground-track fields in extremely complex backgrounds, which are difficult for humans to recognize. The bridge in Fig. 10(b) and the car in Fig. 10(c) are more similar to the background color, making recognition more difficult. However, our model still completed the task. The aforementioned indicates that in complex backgrounds, our model has a stronger ability for feature extraction and will not ignore detailed information, which is very helpful for small object detection. In Fig. 11(a), YOLO-V7 divides an extensive golf course into three parts for recognition. The dam in Fig. 11(b) and the bridge in Fig. 11(c) only have slender edge information. These pieces of information are easily overlooked in convolutional networks. Our network has achieved target recognition with more robust information extraction capabilities. In Fig. 11(d), missed detections were avoided. Fig. 11 shows that our model has stronger feature extraction capabilities even for medium to large objects.

We compared the convergence speed of mAP of multiple methods combined with YOLO-V7 models, as shown in Fig. 12. The results indicate that the convergence speed of the P-SACA module is faster than other attention mechanisms, and the SPD model further improves the convergence speed of the model. Based on the analysis in Tables IV and VII, the SPD module can accelerate the convergence of the model, while the P-SACA module can have stronger feature extraction capabilities. The SPD-P-SACA module fully absorbs their respective advantages and performs better on large datasets than YOLO-V7. The loss curve in Fig. 13 again indicates that our model converges faster than other methods. Faster convergence speed represents lower training costs, which is significant for deep learning.

Fig. 14 shows the confusion matrix obtained by our model during training, which indicates that the complex background is a significant constraint on the poor detection performance of remote sensing image targets. The F1-score in Fig. 15 provides a balanced evaluation of model performance, with the gray curve representing 24 target categories and the blue curve representing the average value. The F1-Score curve indicates that our model can achieve good results when the confidence level is less than 0.8.

V. Conclusion

Remote sensing images have complex backgrounds and large object scale spans. The existing public datasets have relatively

few types of objects, and their distribution is highly uneven. The sources of images are also quite complex. To improve the target detection in remote sensing images, we designed the P-SACA module and combined it with existing modules to improve YOLO-V7. The improved network has better detection performance on the large dataset. First, we have designed a new attention mechanism for the parallel use of spatial and channel attention. It has a more robust feature extraction ability and stronger resistance to complex background interference. The module converges faster than other attention mechanisms and can reduce training costs. Second, we will combine the P-SACA module with the SPD module to improve the YOLO-V7 model. Experiments have shown that the improved model has a more vital detection ability for medium-sized targets with fewer samples and a more vital exploration ability on larger datasets. The mAP shows an increase of 1%. Higher detection accuracy was observed among 15 target types. The improvement of the railway station was the most significant, with a considerable increase of 8.2%. Midsize objects such as airports, dams, and catch courts have also increased by about 5%. Improving detection accuracy at airports and train stations will play a critical role in navigation. Third, we established the optical remote sensing target detection dataset DIOR&DOTA with more types of targets and instances through dataset merging. This dataset has a more vital ability to explore the effectiveness of remote sensing target detection. As far as we know, it has the most target types and corresponding instances, with more external features for the same target. It means that it will evaluate the target detection model more strictly. Finally, the experimental results show that the improved model in this article reduces the computational complexity by 2.7%, which is extremely friendly to embedded devices.

Fig. 14 shows that the background significantly impacts target detection in remote sensing images. Imbalanced samples in the dataset result in significant differences in detection performance among different categories. We will improve the detection performance in the future by addressing complex backgrounds and imbalanced data samples. We hope to explore network models with smaller parameters and computational complexity, which can be more conveniently embedded into devices.

REFERENCES

[1] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 3974–3983, doi: 10.1109/CVPR.2018.00418.

[2] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022, doi: 10.1109/TPAMI.2021.3117983.

[3] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016, doi: 10.1109/TGRS.2016.2601622.

[4] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018, doi: 10.1109/TIP.2017.2773199.

[5] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery : A small target detection benchmark," *J. Vis. Commun. Image Representation*, vol. 34, pp. 187–203, 2016, doi: 10.1016/j.jvcir.2015.11.002.

[6] Z. Sun et al., "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, Jul. 2021, doi: 10.1109/JSTARS.2021.3099483.

[7] C. Zhang, B. Xiong, X. Li, and G. Kuang, "TCD: Task-collaborated detector for oriented objects in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 4700714, doi: 10.1109/TGRS.2023.3244953.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[11] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," Apr. 2020, *arXiv abs/2004.10934*.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 7464–7475, doi: 10.1109/CVPR52729.2023.00721.

[13] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," Aug. 2022, *arXiv:2208.03641*.

[14] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, "CBAM: Convolutional block attention module," Jul. 2018, *arXiv abs/1807.06521*.

[15] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote-sensing scene classification via multistage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5615312, doi: 10.1109/TGRS.2023.3295797.

[16] J. Wang, W. Li, Y. Wang, R. Tao, and Q. Du, "Representation-enhanced status replay network for multi-source remote-sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3286422.

[17] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5501212, doi: 10.1109/TGRS.2022.3233847.

[18] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021, doi: 10.1109/TGRS.2020.3016820.

[19] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856, doi: 10.1016/j.rse.2023.113856.

[20] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, May 2023, Art. no. 5513412, doi: 10.1109/TGRS.2023.3279834.

[21] D. Hong et al., "SpectralGPT: Spectral Foundation model," Nov. 2023, *arXiv abs/2311.07113*.

[22] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019, doi: 10.1109/TIP.2018.2878958.

[23] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 5514415, doi: 10.1109/TGRS.2023.3284671.

[24] H. Zhang, J. Yao, L. Ni, L. Gao, and M. Huang, "Multimodal attention-aware convolutional neural networks for classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3635–3644, Jul. 2023, doi: 10.1109/JSTARS.2022.3187730.

[25] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Computer Vision – ECCV 2020 (Lecture Notes in Computer Science)*, vol. 12374, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-58526-6_13.

[26] Y. Tai, Y. Tan, S. Xiong, Z. Sun, and J. Tian, "Few-shot transfer learning for SAR image classification without extra SAR samples," *IEEE J. Sel.*

Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 2240–2253, Mar. 2022, doi: 10.1109/JSTARS.2022.3155406.
[27] Y. Zhuang, L. Li, and H. Chen, "Small sample set inshore ship detection from VHR optical remote sensing images based on structured sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2145–2160, Apr. 2020, doi: 10.1109/JSTARS.2020.2987827.
[28] J. Shao, Q. Yang, C. Luo, R. Li, Y. Zhou, and F. Zhang, "Vessel detection from nighttime remote sensing imagery based on deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12536–12544, Nov. 2021, doi: 10.1109/JSTARS.2021.3125834.
[29] H. Guo, R. Zhang, Y. Wang, W. Yang, H.-C. Li, and G.-S. Xia, "Accurate bridge detection in aerial images with an auxiliary waterbody extraction task," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9651–9666, Sep. 2021, doi: 10.1109/JSTARS.2021.3112705.
[30] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Montreal, BC, Canada, 2021, pp. 2778–2788, doi: 10.1109/ICCVW54120.2021.00312.
[31] Z.-H. Lin, Y. Wang, J. Zhang, and X. Chu, "DynamicDet: A unified dynamic architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6282–6291.
[32] J. Zheng, H. Wu, H. Zhang, Z. Wang, and W. Xu, "Insulator-defect detection algorithm based on improved YOLOv7," *Sensors*, vol. 22, no. 22, 2022, Art. no. 8801, doi: 10.3390/s22228801.
[33] M. Hussain, H. Al-Aqrabi, M. Munawar, R. Hill, and T. Alsboui, "Domain feature mapping with YOLOv7 for automated edge-based pallet racking inspections," *Sensors*, vol. 22, no. 18, 2022, Art. no. 6927, doi: 10.3390/s22186927.
[34] Z. Liu, X. Gao, Y. Wan, J. Wang, and H. Lyu, "An improved YOLOv5 method for small object detection in UAV capture scenes," *IEEE Access*, vol. 11, pp. 14365–14374, 2023, doi: 10.1109/ACCESS.2023.3241005.
[35] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4209, doi: 10.3390/rs13214209.
[36] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 13728–13737, doi: 10.1109/CVPR46437.2021.01352.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.
[39] S. I.C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
[41] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 16514–16524, doi: 10.1109/CVPR46437.2021.01625.
[42] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 13708–13717, doi: 10.1109/CVPR46437.2021.01350.
[43] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 2844–2853, doi: 10.1109/CVPR.2019.00296.
[44] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2021, Art. no. 5602511, doi: 10.1109/TGRS.2021.3062048.
[45] Z. Chen et al., "PIoU loss: Towards accurate oriented object detection in complex environments," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–17.
[46] X. Pan et al., "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 11204–11213, doi: 10.1109/CVPR42600.2020.01122.
[47] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond Bounding-Box: Convex-hull feature adaptation for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 8788–8797, doi: 10.1109/CVPR46437.2021.00868.
[48] J. M. Llerena, L. F. Zeni, L. N. Kristen, and C. Jung, "Gaussian bounding boxes and probabilistic intersection-over-union for object detection," Jun. 2021, *arXiv:2106.06072*.
[49] P. Zhao, Z. Qu, Y. Bu, W. Tan, and Q. Guan, "PolarDet: A fast, more precise detector for rotated target in aerial images," *Int. J. Remote Sens.*, 2021, vol. 42, no. 15, pp. 5821–5851.

Yuhui Zhao received the B.S. degree in measurement and control technology and instrumentation, in 2020, from the North University of China, Taiyuan, China, where he is currently working toward the Ph.D. degree in instrument science and technology.

His research interests include image processing, machine learning (including deep learning), and its applications to computer vision tasks.

Ruifeng Yang received the B.S., M.S., and Ph.D. degrees in measurement technologies and instruments from the North University of China, Taiyuan, China, in 1992, 1999, and 2005, respectively.

From 2010 to 2012, he has studied with the Postdoctoral Center for Control Engineering, Beijing University of Aeronautics and Astronautics. He is currently a Professor with the School of Instrument and Electronics, North University of China, and the Director with the Automatic Test Equipment and System Engineering Research Center of Shanxi Province. His research interests include equipment test and system integration, automated testing and control, intelligent instruments, image processing, and machine vision.

Chenxia Guo received the B.S. degree in automation from the North University of China, Taiyuan, China, in 2001, the M.S. degree in test measurement technology and instrument from the North University of China, Taiyuan, in 2006, and the Ph.D. degree in instrument science and technology from the North University of China, in 2014.

She is currently an Associate Professor with the School of Instrument and Electronics, North University of China. Her research interests include visual measurement, automated testing and control, and complex electromechanical.

Xiaole Chen received the Ph.D. degree in instrument science and technology from the North University of China, Taiyuan, China, in 2022.

She is currently involved in object detection research with the China North Vehicle Research Institute, Beijing, China. Her research interests include image processing, deep learning, and its applications to computer vision tasks.