IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 March 2023, accepted 11 April 2023, date of publication 17 April 2023, date of current version 24 April 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3267435*

RESEARCH ARTICLE

# Remote Sensing Object Detection Based on Convolution and Swin Transformer

**XUZHAO JIANG AND YONGHONG WU**
Department of Statistics, Wuhan University of Technology, Wuhan 430070, China
Corresponding author: Yonghong Wu (whyflying2008@163.com)

This work was supported in part by the Natural Science Foundation of Hubei Province under Grant 2020CFB546, in part by the National Natural Science Foundation of China under Grant 12001411 and Grant 12201479, and in part by the Fundamental Research Funds for the Central Universities under Grant WUT: 2021IVB024 and Grant 2020-IB-003.

**ABSTRACT** Remote sensing object detection is an essential task for surveying the earth. It is challenging for the target detection algorithm in natural scenes to obtain satisfactory detection results in remote sensing images. In this paper, the RAST-YOLO (You only look once with Regin Attention and Swin Transformer) algorithm is proposed to address the problems of remote sensing object detection, such as significant differences in target scales, complex backgrounds, and tightly arranged small-size targets. To increase the information interaction range of the feature map, make full use of the background information of the object, and improve the detection accuracy of the object with a complex background, the Regin Attention (RA) mechanism combined with Swin Transformer as the backbone is proposed to extract features. To improve the detection accuracy of small objects, the C3D module is used to fuse deep and shallow semantic information and optimize the multi-scale problem of remote sensing targets. To evaluate the performance of RAST-YOLO, extensive experiments are performed on DIOR and TGRS-HRRSD datasets. The experimental results show that RAST achieves state-of-the-art detection accuracy with high efficiency and robustness. Specifically, compared with the baseline network, the mean average precision (mAP) of detection results is improved by 5% and 2.3% on DIOR and TGRS-HRRSD datasets, respectively, which demonstrates RAST-YOLO is effective and superior. Moreover, the lightweight structure of RAST-YOLO can ensure the real-time detection speed and obtain excellent detection results.

**INDEX TERMS** Remote sensing images, object detection, attention mechanism, swin transformer, multi-scale features.

## I. INTRODUCTION

Object detection in remote sensing images is crucial in interpreting aerial and satellite images, which is widely used in many fields, such as resource exploration [1], intelligent navigation [2], environmental monitoring [3] and target tracking [4]. The main task of remote sensing target detection is to determine whether there are targets of interest in remote sensing images and provide their spatial location. In recent years, with rapid development in aerospace and UAVs, numerous high-resolution and high-quality datasets have been created for remote sensing image processing. Compared with natural scene images, remote sensing object detection faces the following challenges: small data scale; similar appearance of objects in different categories; significant disparity appearance of objects in the same category; uneven distribution of small, medium, and large targets; sometimes dense and sometimes sparse target distribution; complex background and extreme imbalance in the number between classes, etc., as shown in Fig. 1.

It is seen from Fig.1 that, in (a) and (b), the object categories are aircrafts, but the backgrounds of the aircrafts are ocean and land, respectively. Moreover, the size difference of the aircrafts in (b) is significant, which is a common challenge for remote sensing object detection. The targets in (c) are sparse, while the targets in (d) are very dense and small. The target in (e) is a bridge, while the target in (f) is a dam. They are of different categories but have highly similar

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

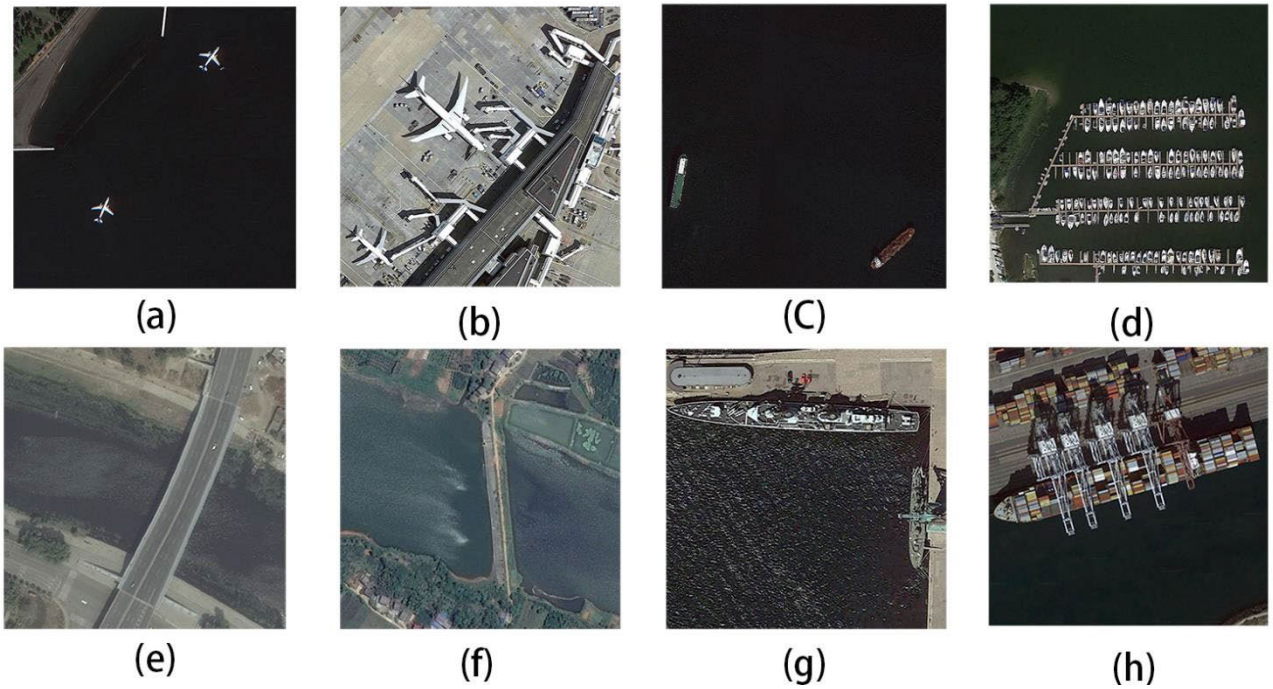


**FIGURE 1.** Common difficulties in remote sensing target detection.

appearances. The targets in (g) and (h) are both ships, but the former is a warship while the latter is a cargo ship. They attribute to the same category with different appearances. Therefore, it is difficult to obtain satisfactory results if the target detection methods for natural scene images are directly used to detect remote sensing targets.

Traditional object detection algorithms consist of several steps: feature extraction, feature transformation, and classifier prediction. The feature extraction stage mainly extracts target features, such as color, texture, shape and angle. The feature extraction methods include scale-invariant feature transform (SIFT) [5], histogram of oriented gradient (HOG) [6], and deformable part model [7]. Classifiers are used to identify specific classes of targets, including support vector machine (SVM) [8], random forest [9], and naive Bayesian algorithm [10], etc. which are based on manual empirical selection and inefficient because of their low target feature extraction capability, especially for deep semantic information. Thus, their robustness and generalization capability are poorer than those of deep learning methods.

Since Alex Krizhevsky et al. proposed Alexnet [11], deep learning has been developed rapidly, and convolutional neural networks have been used for various tasks in computer vision. The deep learning-based target detection algorithms are mainly divided into two categories:

(i) One-Stage Object Detection Algorithm, which include YOLO [12], SSD [13], Retinanet [14], CornerNet [15], etc. They do not need the Region Proposal. The class probability and coordinate location values of objects are directly generated in one stage.

(ii) Two-Stage Target Detection Algorithms, which include R-CNN [16], Fast R-CNN [17], Faster RCNN [18], Mask RCNN [19], etc. Their detection process consists of two stages. The first stage generates Region Proposals, which contain the approximate location information of targets and the second performs classification and location refinement of the Region Proposals.

Transformer [20] is the most advanced machine translation method for many natural language processing tasks (BERT [21], GPT [22]). Inspired by the successful application of Transformer in NLP, Alexey Dosovitskiy et al proposed ViT [23] (Vision Transformer), which is the first model that only uses Transformer to classify images without using convolution neural network. When pre-training the ImageNet-21 k data set or the JFT-300 M data set, the Vit surpassed the most advanced image recognition models at that time. From then on, Transformer began to shine in the visual space.

Transformer-based object detection algorithms can be divided into two categories according to the network structure: Transformer as Backbone uses CNN to extract features and realize prediction with Transformer, and Transformer as Neck uses Transformer as the backbone network and realizes prediction with CNN. Mainstream object detection algorithms based on Transformer neck include DETR [24], deformable DETR [25], ACT [26], UP-DETR [27], etc. Mainstream object detection algorithms based on Transformer backbone include FPT [28], Swin Transformer [29], DeMT [30], etc.

The object detection algorithm based on Transformer outperforms the traditional convolution neural network

algorithm in detection accuracy. But it still has many shortcomings, such as a large number of model parameters, slow training time and reasoning speed. The training process depends on large data sets. Because its calculation cost increases in square times with the increase of resolution, it is not suitable for processing high-resolution images.

The main performance indexes of the target detection model are the accuracy and speed of detection. The accuracy of target location and classification and the speed of algorithm detection are mainly considered. Object detection algorithm based on Transformer has achieved satisfactory results in accuracy, but its detection speed is not satisfactory.

To solve the difficulties of remote sensing target detection and improve its accuracy and detection speed, the high-precision advantages of the target detection algorithm based on Transformer are used to obtain the main contributions as follows:

a) A feature extraction backbone network integrated with convolutional neural network and Transformer is proposed. This network can better extract rich information features from input images, increase the interaction range of feature information, make full use of the global background information and local details of the target, and improve the detection accuracy of remote sensing targets in complex backgrounds. C3D module is designed to generate a feature pyramid at the Neck stage, which enhances the fusion of deep semantic information and shallow location information. It not only perfectly identifies the same class of objects with different sizes and scales, but also improves the detection accuracy of small objects.
b) The C3D module is designed to generate feature pyramids in the Neck stage to enhance the fusion of deep semantic information and shallow location information, which perfectly identifies the same class of objects with different sizes and scales, and improves the detection accuracy of small objects.
c) The attention mechanism module RA (regional attention) is designed and used in parallel with the Swin Transformer [29], which extends information interaction within the window to the global level. The RA mechanism is superior to the existing attention mechanisms.
d) ACmix Plus Detector, a detector combining convolution and self-attention mechanisms, is designed, which improves network detection accuracy and recall for each type of objects.

The rest of this paper is as follows: In Section II, the applications of target detection in natural images and remote images are reviewed; In Section III, the target detection framework proposed in this paper is introduced in detail; Experimental results of RAST-YOLO proposed in this paper are shown in Section IV; Conclusions are drawn in Section V.

## II. APPLICATION OF OBJECT DETECTION

Object detection is one of the most critical and challenging branches in the field of computer vision, which has been widely used in many areas and attracted the attention of many researchers. In this section, the research progress of target detection in natural scenes and remote sensing images and its application in various industries are introduced in detail.

### A. OBJECT DETECTION IN NATRUAL SCENES

In past decades, deep learning-based target detection algorithms have been successfully applied to natural scene images. Convolutional neural network-based target detection algorithms are classified into two-stage detection algorithms with Region Proposal Network (RPN), and single-stage detection algorithms without RPN.

Two-stage detection algorithms divide the detection process into two stages. In the first stage, Region Proposal is generated, and in the second stage, candidate regions are classified (generally the position needs to be refined). RCNN [16] is the first object detection algorithm based on deep learning, which implements the whole detection process through deep neural networks. Compared with traditional object detection, it achieves significant improvement in accuracy and speed. However, the fully connected layer of RCNN requires a fixed-size image input, and cropping or stretching the original image affect the detection accuracy. In Fast RCNN [17], the structure of ROI Pooling between the convolutional layer and the fully connected layer effectively reduces the impact of cropping and stretching on detection accuracy. However, it takes Selective Search much time to find candidate frames. The Faster RCNN [18] algorithm replaces Selective Search in Fast RCNN [13] with RPN and uses the Anchor mechanism to link region generation with convolutional networks. Faster RCNN improves the detection speed to 17 FPS and achieves 70.4% detection accuracy in the test set of PASCL VOC [31].

One-stage detection algorithms, such as YOLO [12], SSD [13], RetinaNet [14], YOLOv3 [32], etc. do not need to generate the Region Proposal, and directly generate the class probability and coordinate position values of the object. The final detection result is directly obtained after a single detection. Thus, the speed of one-stage detection algorithms is faster. To address the slow detection speed of the two-stage detection algorithm, YOLO [12] applies a single neural network to the whole image, which predicts the bounding box and probability of each region simultaneously. The algorithm efficiently improves the speed of target detection, and its accuracy is better than that of two-stage algorithms, especially for detecting small targets. To improve the detection of multi-scale objects, SSD [13] detection algorithm introduces multi-reference and multi-resolution detection techniques and uses multi-scale feature maps for prediction, which significantly improves the accuracy of the one-stage detection algorithm, especially for small target objects. Retinanet [14] addresses the reason why the detection accuracy of one-stage algorithms is inferior to two-stage algorithms is the positive and negative sample class imbalance. The algorithm proposes Focal loss based on cross-entropy loss and improves the detection accuracy of hard-to-classify samples by increasing

its' weight in the loss function, thus improving the detection accuracy of the one-stage algorithm. YOLOv3 [32] designs Darknet53, a feature extraction network with excellent performance for multi-scale prediction with faster speed and higher accuracy.

### B. APPLICATION OF OBJECT DETECTION IN VARIOUS INDUSTRIES

In recent years, object detection has become a research hotspot in the field of computer vision. As a branch of image processing and computer vision, it is widely used in robot navigation, intelligent video surveillance, industrial detection, traffic monitoring and many other fields. In the summary of this section, successful application of target detection in various industries will be listed. They play an important role in the development of remote sensing. To solve the problems in infrared image target detection, such as poor texture information, low resolution and high noise, Zhang et al. [33] proposed the Deep-IRTarget backbone network, which consists of a frequency feature extractor, a spatial feature extractor and a dual-domain feature resource allocation model. And a resource allocation model RAF is designed to superimpose the features of frequency domain and spatial domain to construct the dual domain features. Thus, an ideal infrared image target detection effect is achieved. To better solve the problem of small target detection in infrared images, Wu et al. [34] proposed a simple and effective "U-Net in U-Net" framework, which embeds a tiny U-Net into a larger U-Net backbone, and realizes multi-level and multi-scale representation learning of objects. For live operation of distribution lines, Zhao et al. [35] designed an autonomous robot navigation system. They proposed an insulator and drop fuses target detection method based on the Larger Scale 'You Only Look Once' Version 4 (LS-YOLOv4) algorithm, which helps the robot grasp the power components of the manipulator to accurately identify the target. In the railway industry, the use of cameras to accurately and quickly detect targets is an important but challenging problem. Ye et al. [36] proposed the LFD algorithm, which improves the real-time detection accuracy of targets of different scales (especially small targets) without additional storage space and processing time, and is used for collision warning in the train safety system. Wang et al. [37] introduced Region-of-Interest into the YOLOv4 network to enhance train detection of pedestrians and signal lights in highly complex and harsh environments. The application of the above object detection algorithm based on deep learning in various industries provides certain ideas and insights for the development of remote sensing image object detection, and plays an indispensable role in its development.

### C. OBJECT DETECTION IN REMOTE SENSING

Due to its' excellent detection accuracy, two-stage target detection algorithms are used as baselines and appropriately improved to obtain superior remote sensing algorithms. In [38], RPN and local contextual feature fusion network

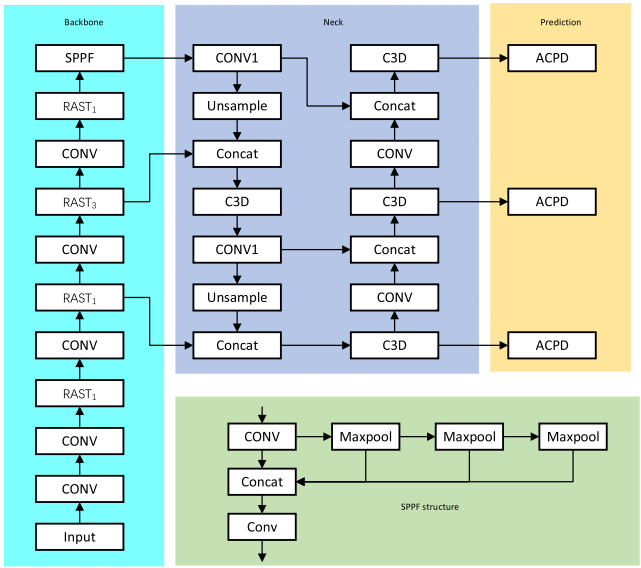


**FIGURE 2.** Network structure of RAST-YOLO.

based on Faster RCNN are redesigned for remote sensing images with multi-scale and multi-angle features of targets, such that multi-angle and multi-scale features of geospatial objects can be extracted. In [39], a structure-guided feature transform hybrid residual (SGFTHR) network based on FCOS [40] is proposed, which effectively improves the detection performance for a large number of small and dense objects in remote sensing images. Lv et al. proposed a multi-scale feature adaptive fusion (MFAF) method [41] for many multiscale objects and complex backgrounds of targets in remote sensing images. They used it in YOLOv4 [42] to improve its detection performance for multiscale targets successfully. Yang et al. [43] replaced feature extraction trunk CSPDarknet53 with ConvNeXt-S on the basis of YOLOv4, used EIoU loss function, and added CA attention mechanism into the network, thus improving the detection ability of remote sensing targets. In [44], the cross-scale feature fusion pyramid network (CF2PN) is proposed to solve the multi-scale problem of remote sensing images. In [45], the bi-directional contextual enhancement (CBD-E) method is proposed, which filters useless background information and collects useful background information to enhance its detection performance.

YOLOv5 is a one-stage target detection algorithm with real-time detectability and better detection accuracy, especially for small target objects. Moreover, the network structure of YOLOv5 is flexible and easy to change. In this paper, the existing target detection methods are comprehensively compared and YOLOv5 is used as the baseline to improve the performance of remote sensing object detection.

## III. METHODOLOGY

The performance of existing object detection algorithms is compared in detection accuracy and speed and the framework

of YOLOv5-6.1 is used as the baseline of RAST-YOLO. The structure of RAST-YOLO contains a Backbone to extract features, a Neck to fuse features, and a Detector to get the result. Since information around targets is needed to identify remote sensing targets, the feature extraction module RASTn is designed, which integrates attention mechanism and convolutional neural network in feature extraction backbone. This module RASTn can effectively extract contextual information of feature maps and supplement information extraction around remote sensing targets. The CONV module is a combination of convolution, Batch Normalization, and activation functions. Its role is to reduce the size and expand the number of channels in the feature maps. SPPF module can realize the pyramid pool of adaptive size output and enhance the feature expression ability of feature extraction backbone. In Neck, based on dense linking from DenseNet [46] the C3 module in YOLOv5, called the C3D (C3DF) module, is redesigned. C3D(C3DF) significantly fuses the semantic information of deep and shallow layers and improves the detection accuracy of multi-scale targets via splicing shallow and deep feature maps. The improved ACmix [47] module combined with YOLO Detector forms the ACmix Plus Detector, which obtains global information across spatial dimensions and channel dimensions and enhances the recall and accuracy of the network. The structure of RAST-YOLO is shown in Fig.2, and the modules in the figure are described detailly in subsequent chapters.

The following is a description of the procedure for evaluating variables in RAST-YOLO. After data preprocessing, the resolution of the feature map is 640 × 640. In Backbone, the CONV module can reduce the feature map size and expand the number of channels of the feature map, while RASTn does not change the feature map size and the number of channels. The function of RASTn is to extract various feature information in the feature map. After the input image is processed by Backbone, it outputs feature images with sizes of 80 × 80, 40 × 40 and 20 × 20 to the Neck. In Neck, the size of convolution kernels in CONV1 module is 1 × 1, which does not change the size and number of channels of feature map, but only plays the role of feature fusion. The feature maps sent to Neck by Backbone have different sizes. The small-size feature maps contain deep semantic information, while the large-size feature maps contain shallow semantic information. Deep semantic information and shallow semantic information will be fused in Neck to construct a feature pyramid structure. The C3D module in Neck can fully integrate deep and shallow semantic information to improve the detection accuracy of multiscale remote sensing targets. After the feature maps are processed by the Neck of RAST-YOLO, the feature maps with sizes of 80 × 80, 40 × 40 and 20 × 20 are output to ACPD (ACmix Plus Detector) to predict the final results. ACPD predicts small-size remote sensing targets on large-size feature maps and large-size remote sensing targets on small-size feature maps, thus mentioning the detection accuracy of multi-scale remote sensing targets.

### A. REGION ATTENTION MECHANISM

RA mechanism module is embedded into the backbone of the RAST-YOLO, which gives more weight to interested remote sensing detection targets, effectively reduces the influence of complex backgrounds and improves the detection performance of remote sensing targets. RA mechanism module enables the RAST-YOLO algorithm to focus on the interested region and allocate more computational resources for it. The structure of the RA mechanism module is shown in Fig.3.

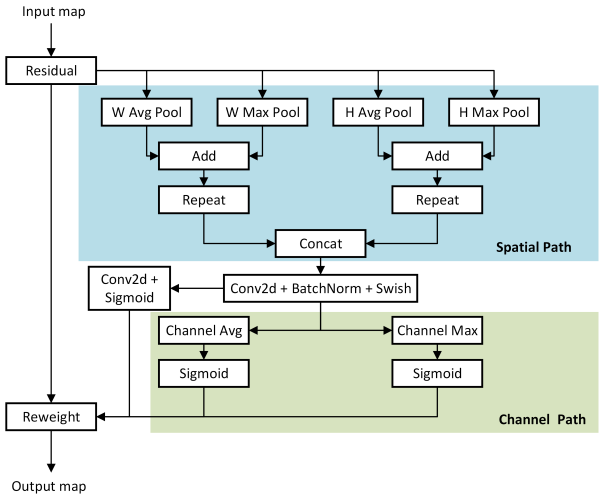


**FIGURE 3.** The structure of RA.

The CBAM [48] performs average pooling and maximum pooling on the input feature map along the channel direction, and compresses the channel dimension of the feature map into one dimension. Then it is stitched along the channel direction and compressed into one dimension with a convolution with kernel size of 7 × 7; The spatial attention features are obtained via the nonlinear activation function Sigmoid. It is difficult to capture detailed location information by pooling the input feature map along the channel direction directly. To capture more precise location information, an RA with pooling method similar to CA [49] is adopted. As shown in Figure 3, RA attention mechanism module enhances global feature extraction through channel and spatial dimensions, and integrates the attention feature information of the two dimensions to obtain global feature information.

Step 1. Let the feature map of the input RA mechanism module be

$$X = [x_1, x_2, x_3, \cdots, x_c] \in R^{C\times H\times W}, \tag{1}$$

which is calculated as follows.

$$x_c^{h-Avg}(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h, i); \tag{2}$$

$$x_c^{h-Max}(h) = \underset{0\le i<w}{Max}\{x_c(h, i)\}; \tag{3}$$

$$x_c^{w-Avg}(w) = \frac{1}{H}\sum_{0\le i<H} x_c(i, w); \tag{4}$$

$$x_c^{w-Max}(w) = \underset{0\le i<H}{Max}\{x_c(i, w)\}, \tag{5}$$

Step 2. The pooling results obtained in Step 1 are summed and copied along the pooling direction. $x_c^h(h, w)$ and $x_c^w(h, w)$ with the same size as the input feature map are obtained, which is calculated as follows:

$$x_c^h(h) = x_c^{h-Avg}(h) + x_c^{h-Max}(h); \quad (6)$$

$$x_c^w(w) = x_c^{w-Avg}(w) + x_c^{w-Max}(w); \quad (7)$$

$$x_c^h(h, w) = x_c^h(h); \quad (8)$$

$$x_c^w(h, w) = x_c^w(w), \quad (9)$$

Step 3. $x_c^h(h, w)$ and $x_c^w(h, w)$ obtained in Step 2 are stitched along the channel direction, and the number of channels is reduced to a convolution with kernel size of $1 \times 1$. After the Batch normalization is performed on the result, the nonlinear activation function Swish is used to calculate $f_1$. And then, the channels are expanded to the same dimension as the input feature map via a convolution with kernel size of $1 \times 1$ and the regional attention weight parameter $f_2$ is obtained, which is calculated as follows:

$$g_1 = F_1\left(\left[x_1^h, x_2^h, \cdots, x_c^h, x_1^w, x_2^w, \cdots, x_c^w\right]\right); \quad (10)$$

$$g_2 = BN(g_1); \quad (11)$$

$$f_1 = g_2 \times \sigma(g_2) \in R^{\left(\frac{C}{r}\right)\times H\times W}; \quad (12)$$

$$f_2 = \sigma[F_2(f_2)] \in R^{C\times H\times W}, \quad (13)$$

Step 4. For the feature fin Step 3, the mean and maximum values of its channel dimensions are found, respectively. After the nonlinear activation function Sigmoid calculation, the regional attention bias term parameters are obtained as in (14) and (15):

$$b_{Avg}(h, w) = \sigma\left[\frac{1}{\frac{C}{r}} \sum_{0\le i<\frac{C}{r}} f_1(i, h, w)\right] \in R^{1\times H\times W}; \quad (14)$$

$$b_{Max}(h, w) = \sigma\left[\underset{0\le i<\frac{C}{r}}{Max}\{f_1(i, h, w)\}\right] \in R^{1\times H\times W}, \quad (15)$$

Step 5. The final feature map with RA weights is calculated as in (16):

$$Output = \left(X + b_{Avg} + b_{Max}\right) * f_2. \quad (16)$$

The results calculated via the RA mechanism module are output.

### B. BACKBONE BASED ON RAST

The feature extraction network proposed in this paper consists of two modules alternatively. One is the Conv module, which has three steps in series: a two-dimensional convolution with kernel size of $1 \times 1$ and strides size of 2, Batch Normalization, and the Silu activation function. The Conv module is to reduce the length and width of the feature map and expand the number of dimensions. The module RASTn consists of 2D convolution, RA (Regin Attention) module, and Swin Transformer blocks in parallel, where RASTn denotes the STR module in its structure with n Swin Transformer blocks. The function of the RASTn module is to extract features and obtains the global background information and local details of the feature map. Its structure is shown in Fig.4:

The computational complexity of the multi-headed attention mechanism is proportional to the square of the size of the feature map. To reduce the computational complexity of multiple attention mechanisms and extend the range of information interactions. In Swin Transformer [29], the feature maps is divided into each window, and the attention mechanism is calculated for the pixels in each window and shifted window. However, the recognition and localization of objects in remote sensing images depends on the feature information of the global background. Information interactions in Swin Transformer exist only in individual windows and shifted windows, which can only capture local details of the target, but global background information is difficult to obtain. To achieve a wider range of information interactions and simultaneously obtain global background information and local details, the RAST feature extraction backbone is designed to combine Swin Transformer and RA modules.

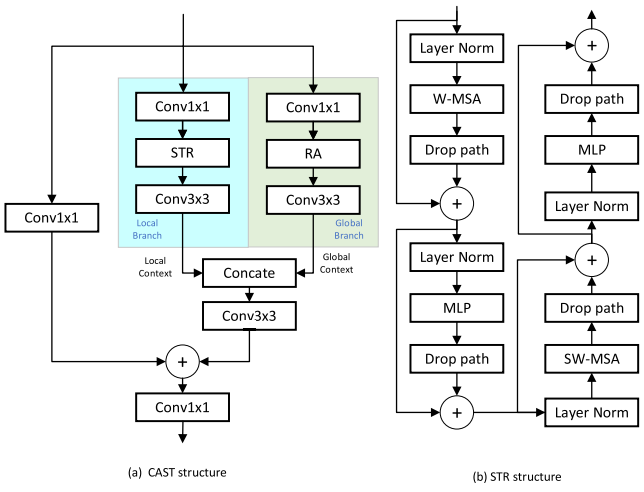


**FIGURE 4. The structures of RAST feature extractor and Swin Transformer.**

In RAST, the input feature map is expanded into three feature maps by using three convolutions with kernels of size $1 \times 1$. One is used as a residual link, and the others, computed by RA module and Swin Transformer Block module respectively, are stitched and the features are fused by using a 2D convolution with kernels of size $3 \times 3$. The background of remote sensing targets is extremely complex. However, the global background information is absolutely crucial for remote sensing object detection. And the local information containing rich spatial details is also indispensable. The RA module captures the feature information of the global background in both channel and spatial paths, while the Swin Transformer Block captures the local feature information with rich spatial details. By integrating the information captured by the two modules, the feature representation will contain both global and local information. This scheme effectively increases the receptive field and improves the detection

accuracy of remote sensing targets in complex backgrounds. Moreover, 2D convolution is used to disassemble the feature maps and send them to different modules for calculation and fusion, which improves the efficiency of processing procedure.

### C. DENSELY CONNECTED C3 (C3D)

C3 is a feature extraction module of yolov5, which has three convolutions with kernel size of $1 \times 1$ and several Bottlenecks composed of 2D convolutions. When the feature map is input into the C3, two convolutions are first calculated and two feature maps with half of the channel number are obtained, one as a residual link and the other into Bottlenecks. The bottleneck has two types of structures, one with residual links and the other without residual links. In the C3D structure proposed in this paper, dense links between Bottlenecks are added, which enhances the transmission of feature information between Bottlenecks and effectively uses the feature information. The comparison of C3 and C3D(F) structures is shown in Fig.5.

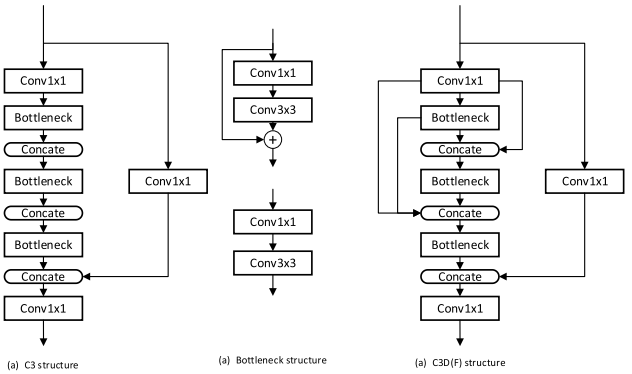


**FIGURE 5.** The structure of C3 and C3D.

Suppose the feature map of the input Bottleneck is $X \in R^{(C/2)\times H\times W}$. $F_i$ denotes the *ith* computation process of the Bottleneck. Then the computation process of C3 is as in (17):

$$f = F_3\left\{F_2\left[F_1\left(X\right)\right]\right\} \in R^{\left(\frac{C}{2}\right)\times H\times W}, \tag{17}$$

And the calculation process of C3D is as follows:

$$g_1 = F_1\left(X\right); \tag{18}$$

$$g_2 = F_2\left(\left[g_1, X\right]\right); \tag{19}$$

$$f = F_3\left(\left[g_2, g_1, X\right]\right) \in R^{\left(\frac{C}{2}\right)\times H\times W}. \tag{20}$$

C3D with dense links makes full use of the information of the input feature map compared with the residual network. It is more effective in overcoming the gradient disappearance of the deep neural network. In the Neck of RAST-YOLO, the input feature map is the combination of deep and shallow feature maps. The shallow feature map has less semantic information, but more location information, while the deep feature map has richer semantic information, but poorer location information. The fusion of the two effectively improves the detection accuracy of densely arranged small targets.

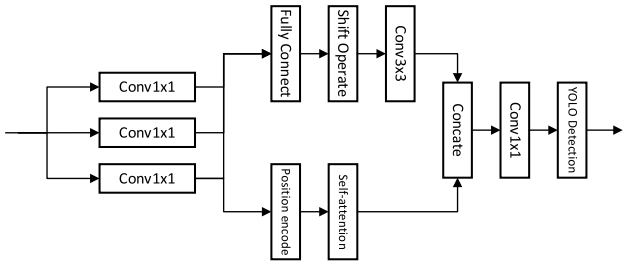


**FIGURE 6.** The structure of ACmix plus detector.

### D. ACMIX PLUS DETECTOR

Convolution and self-attentive mechanisms are two independent representational learning methods. A paper published in 2022 [47] demonstrates the potential connection between convolution and self-attentive mechanism because some computations of both methods are similar. For example, a 2D convolution with kernel size of k × k is decomposed into k × k convolutions with kernel size of 1 × 1. In computing the self-attentive mechanism, the encoding process to get the key, query, and value is regarded as a convolution with kernel size of 1 × 1. This process takes up more than 99% of the number of parameters and computation of the convolution. Meanwhile, the self-attentive mechanism inverts approximately 83% of the parameters into this process. Due to the potential strong connection between convolution and self-attention, the ACmix module is proposed to perfectly combine convolution and self-attention mechanism.

In the ACmix module, perform the convolution operation on the input feature map, and obtain three feature maps by using three 2D convolutions with kernel size of $1 \times 1$ respectively. These three feature maps are input to the convolution module and the self-attentive mechanism module to obtain the convolutional features $f_{Conv}$ and self-attentive features $f_{\mathrm{Att}}$. The results of output features are as in (21), where $\alpha$ and $\beta$ are learning parameters:

$$f = \alpha f_{Conv} + \beta f_{Att}, \tag{21}$$

The ACmix module is connected to each YOLO detector, which could improve the detection accuracy of YOLOv5. It is difficult to effectively fuse the features computed by the convolutional and self-attentive mechanisms through the learnable parameters $\alpha$ and $\beta$. In this paper, the convolutional and self-attentive features are stitched along the channel direction and fully fused through 2D convolution as in (22).

$$f = F\left(\left[f_{Conv}, f_{Att}\right]\right). \tag{22}$$

The structure of ACmix Plus Detector is shown in Fig.6, which is obtained by splicing the modified Acmix Plus and YOLO detectors in this paper.

## IV. EXPERIMENTS

To demonstrate the superiority of RAST-YOLO detection in remote sensing object detection, experimental procedures, including the data set, experimental parameter setting,

evaluation index, result comparison, analysis network interpretability, and visualization of detection effect, are shown in this section.

### A. EXPERIMENTAL CONFIGURATION

The experimental environment is windows 11 operating system, the computer running memory is 32GB, CPU is i9-12900K, GPU is RTX 3090, the deep learning framework is pytorch1.9, and the programming language is python3.9.

### B. DATASETS

DIOR [50] is a publicly available optical remote sensing image object detection dataset released by Northwestern Polytechnic University in 2019. DIOR consists of 23,463 high-quality remote sensing images and 192,472 instance objects, containing 20 common remote sensing category objects, such as airplanes, airports, baseball fields, basketball courts, bridges, chimneys, dams, expressway service areas, expressway toll stations, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. The remote sensing target detection dataset contains the most images and the most instances, with features as follows: Extensive range of object sizes; Rich images; High inter - class similarity and intra - class diversity; Uneven distribution of instances by category.

TGRS-HRRSD [51] is a dataset released by the University of Chinese Academy of Sciences in 2019, which consists of 21761 images and 55740 instance objects acquired from Google Earth and Baidu maps. TGRS-HRRSD contains 13 categories such as airplane, baseball diamond, basketball court, bridge, crossroad, ground track field, harbor, parking lot, ship, storage tank, T junction, tennis court, and vehicle. In the dataset, the number of samples in each category is balanced and each category contains about 4000 instances.

### C. EXPERIMENTAL PROTOCOL AND EVALUTATION INDICATIORS

Before training, the prior frame clustering algorithm of YOLOv5 is used to calculate the relative dataset. Dataset DIOR and GTRS-HRRSD are divided into training set, validation set, and test set in a ratio 1:1:2. The input image is 640 × 640 pixels, the training batch size is 16, the number of training epochs is 100, the optimizer is stochastic gradient descent, the initial learning rate size is set to 0.01, the momentum is 0.937 and the IOU threshold is 0.45.

The evaluation indexes in this paper are Precision (P), Recall, $mAP_{50}$, and $mAP_{50:95}$. Precision is the percentage of correctly predicted positive samples in all predicted positive samples. Recall is the percentage of correctly predicted positive samples in positive samples. Ap is an area under the P-R curve, and mAP is the mean value of all categories of AP. The calculation procedure is as follows:

$$P = \frac{TP}{TP + FP}; \tag{23}$$

**TABLE 1. Ablation experiments on DIOR dataset.**

| Experiments | P | R | $mAP_{50}$ | $mAP_{50:95}$ |
|---|---|---|---|---|
| (1) | 0.818 | 0.643 | 0.698 | 0.465 |
| (2) | 0.816 | 0.627 | 0.687 | 0.453 |
| (3) | 0.815 | 0.637 | 0.687 | 0.437 |
| (4) | 0.822 | 0.631 | 0.686 | 0.449 |
| (5) | 0.804 | 0.605 | 0.658 | 0.408 |
| (6) | 0.823 | 0.609 | 0.672 | 0.437 |
| (7) | 0.809 | 0.62 | 0.675 | 0.424 |
| (8) | 0.771 | 0.61 | 0.648 | 0.39 |

$$R = \frac{TP}{TP + FN}; \tag{24}$$

$$AP = \int_0^1 PdR; \tag{25}$$

$$mAP = \frac{\sum_i AP_i}{m}. \tag{26}$$

where $TP$, $FP$, $TN$ and $FN$ denote the number of true positive samples, false positive samples, true negative samples and false negative samples, respectively, m is the number of sample categories. Moreover, $mAP_{50}$ indicates that the sample is judged to be positive when the IOU of the predicted box and the real box are greater than 0.5. $mAP_{50:95}$ is the mean value of $mAP_{50}$, $mAP_{55}$, $mAP_{60}$, $mAP_{65}$, $mAP_{70}$, $mAP_{75}$, $mAP_{80}$, $mAP_{85}$, $mAP_{90}$ and $mAP_{95}$.

### D. ANALYSIS OF EXPERIMENTAL RESULTS

Five sets of experiments on the ablation of different improvement parts, the interpretability analysis of the network, the comparison between RA attention mechanism and other attention mechanisms, the comparison between RAST-YOLO and other algorithms, and the analysis of the visualization on detection results, are designed in this section.

#### 1) ABLATION EXPERIMENTS

In order to verify the positive effects of RAST, C3D, and ACPD (ACmix Plus Detector) on remote sensing target detection proposed in this paper. We designed a set of ablation experiments based on DIOR data set. (1) RAST-YOLO model; (2) The feature extraction backbone RAST was removed on the basis of RAST-YOLO; (3) Remove C3D on the basis of RAST-YOLO: (4) Remove ACmix Plus Detector on the basis of RAST-YOLO; (5) RAST and C3D were removed based on RAST-YOLO; (6) Remove feature extraction backbone RAST and ACmix Plus Detector on the basis of RAST-YOLO; (7) Remove C3D and ACmix Plus Detector on the basis of RAST-YOLO; (8) On the basis of RAST-YOLO, the feature extraction backbone RAST, C3D and ACmix Plus Detector were removed(namely YOLOv5). Experiments on the DIOR dataset are conducted under the same experimental conditions. The experimental results are shown in TABLE.1

It follows from TABLE.1 that the precision, recall and AP of RAST-YOLO significantly decrease after removing the proposed modules.

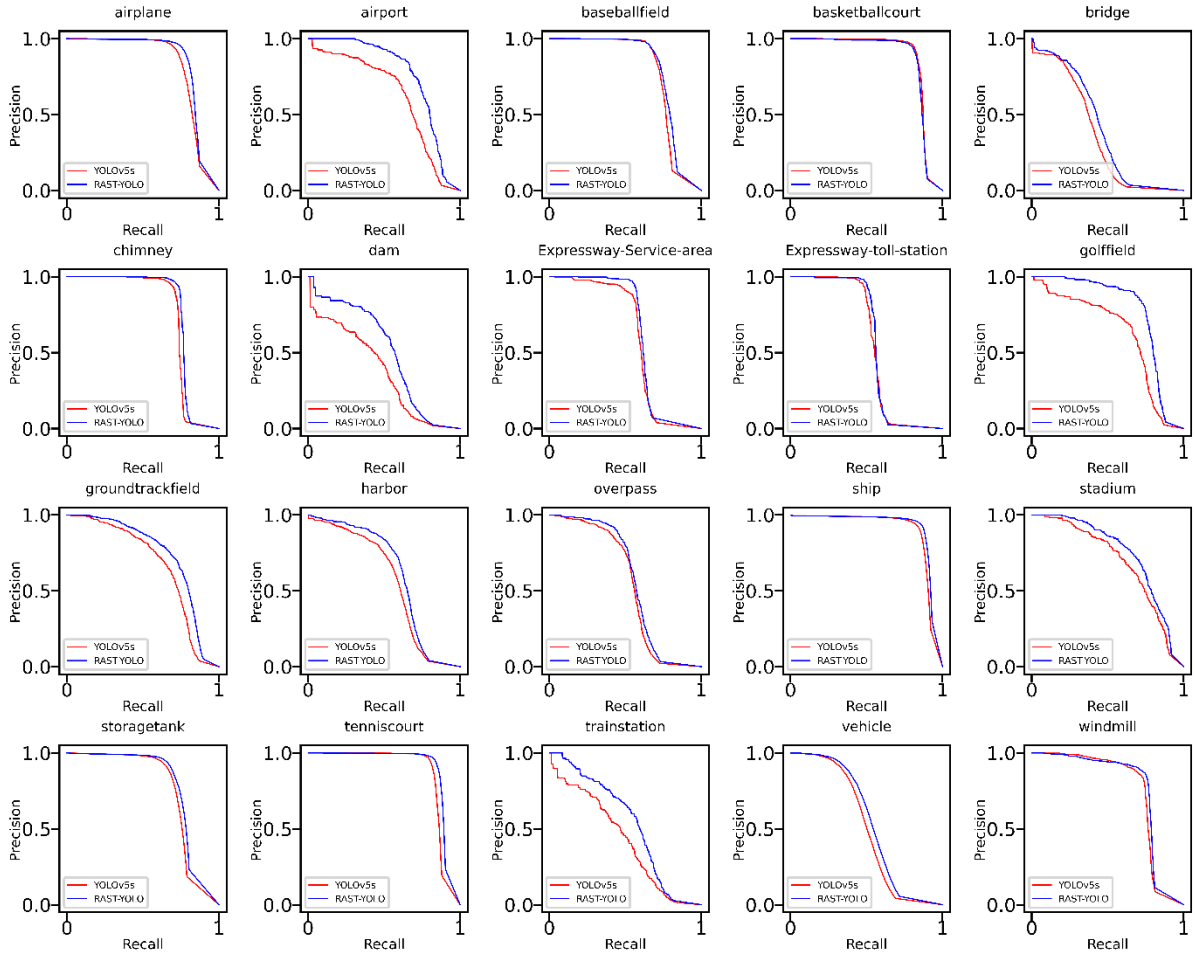


**FIGURE 7.** Comparison of P-R curves of YOLOv5 and RAST-YOLO.

i) When the feature extraction backbone RAST is removed, the recall of the RAST-YOLO is reduced by 1.6%; ii) When the C3D module of RAST-YOLO is removed, the mAP50:95 of the model decreases by 2.8%; iii) When the ACmix Plus Detector was removed and the detector of YOLOv5 was used, the recall of the model was reduced by 1.2%, but the precision was increased by 0.4%; iv) When the feature extraction backbone RAST and C3D module of RAST-YOLO were removed simultaneously, the precision of the model decreased by 1.4%, and the recall decreased by 3.8%; v) When the feature extraction backbone RAST of RAST-YOLO and ACmix Plus Detector were removed, the recall of the models decreased by 3.4% and mAP50 decreased by 2.6%, but the precision increased by 0.5%; vi) When removing C3D module and ACmix Plus Detector, model recall decreased by 2.3%, and mAP50:95 decreased by 4.1%; vii) Compared with YOLOv5, the precision of model detection of RAST-YOLO increased by 4.7%, recall increased by 3.3%, mAP50 increased by 5%, and mAP50:95 increased by 7.5%. Therefore, the ablation experiment can firmly verify the positive effect of RAST, C3D, and ACmix Plus Detector proposed in this paper on remote sensing object detection.

In the Fig.7, the P-R curves of YOLOv5 and RAST-YOLO are compared for each of the 20 categories in the DIOR dataset, and the area below the curve is the AP for each category. Specially, the detection effect of RAST-YOLO in airports, golf courses, dams and railway stations is obviously better than that of YOLOv5.

### 2) COMPARISON OF INTERPRETABLE ANALYSIS BETWEEN NETWORKS

The interpretably comparative analysis between RAST-YOLO and YOLOv5 is performed by using the Grad-CAM++ [52]. Three images containing common difficulties in remote sensing target detection are selected in each of the DIOR and TGRS-HRRSD datasets to test the performance of RAST-YOLO and YOLOv5. These images, which include common difficulties in remote sensing target detection, are used to test the performance of RAST-YOLO and YOLOv5.

The selected images, containing the heat map output by RAST-YOLO and YOLOv5, are shown in Fig.8. In the heat map, the model tends to be more sensitive and pays more attention to the redder regions.

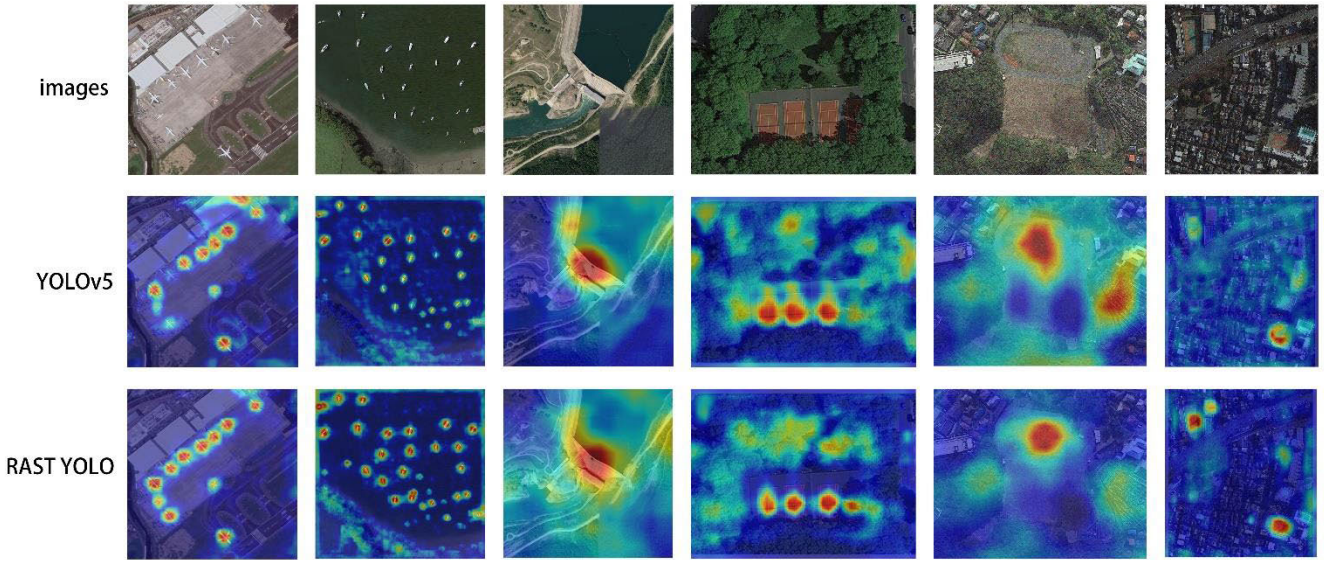


**FIGURE 8.** Interpretable comparison of YOLOv5 and RAST YOLO.

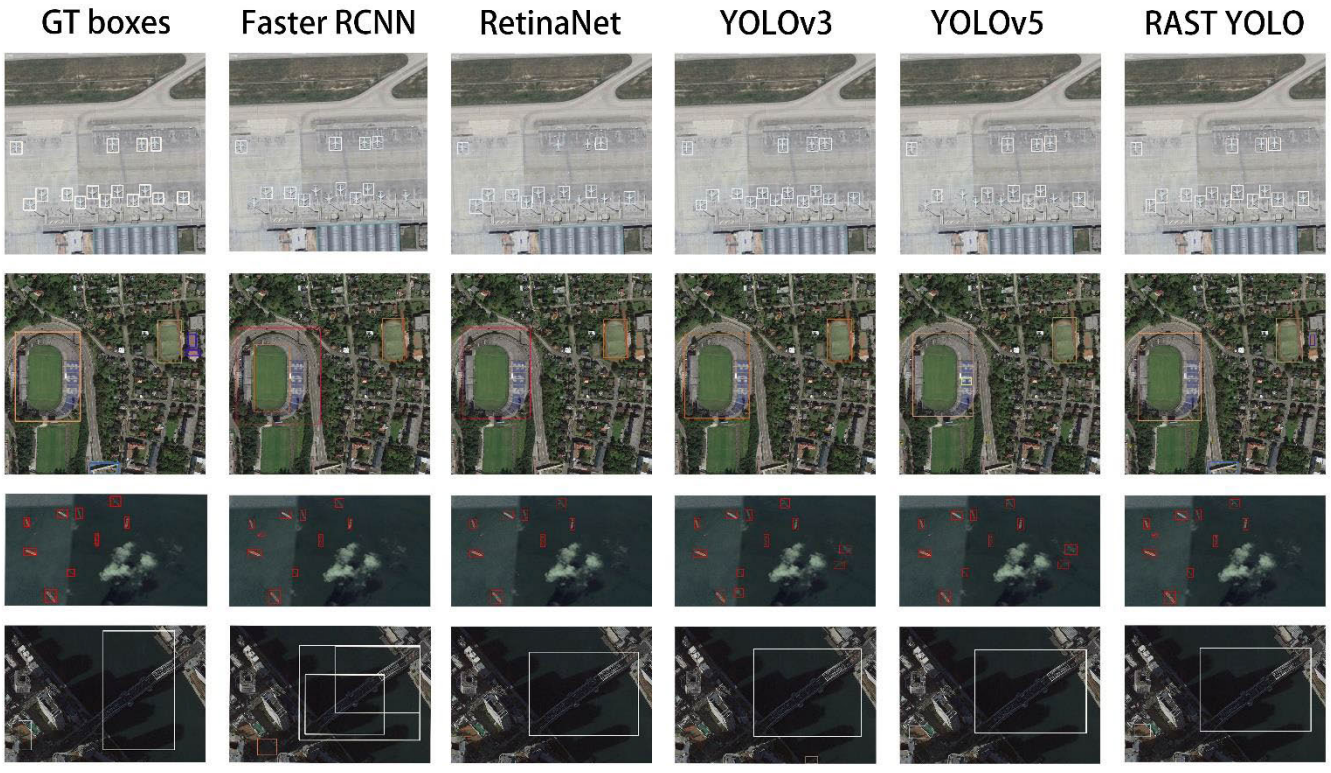


**FIGURE 9.** Comparison of visual detection results.

As can be seen from Fig. 8, the focus positions of the heat maps generated by RAST-YOLO are more accurate than those generated by YOLOv5. Specifically, aircrafts with different sizes are relatively orderly distributed in the first image. YOLOv5 ignores some small-sized aircrafts, compared with RAST-YOLO. Irregularly distributed ships with various sizes fill the second image. The focus of the heat map generated by RAST-YOLO covers all ships, while YOLOv5 ignores some small ships. In the third image, the dam needs to be detected.

The heat map generated by YOLOv5 focuses on the embankment with the same color and shape as the dam, while the heat map generated by RAST-YOLO precisely focuses on the dam. In the fourth image, four tennis courts need to be detected, one of which is partially obscured by branches and their shadows. RAST-YOLO pays more attention to the

**TABLE 2. Comparison of attention mechanisms.**

| ATTS | P | R | MAP50 | MAP50:95 |
|---|---|---|---|---|
| SE | **0.825** | 0.637 | 0.694 | 0.464 |
| GAM | 0.821 | 0.636 | 0.692 | 0.46 |
| CBAM | 0.807 | 0.635 | 0.688 | 0.657 |
| CA | 0.822 | 0.636 | 0.694 | 0.462 |
| RA | 0.818 | **0.643** | **0.698** | **0.465** |
| - | 0.824 | 0.612 | 0.676 | 0.447 |

obscured tennis court than YOLOv5. The fifth image contains an old ground track field that looks very similar to the background. The heat map generated by RAST-YOLO is precisely focused on the ground track field, but the heat map generated by YOLOv5 is partially focused on the open space around the area. In the sixth image, two ground track fields between urban buildings with complex backgrounds needs to be detected. The heat map generated by YOLOv5 ignores one of them, while the heat map generated by RAST-YOLO perfectly focuses on all fields.

Compared with the baseline, RAST-YOLO improves the detection accuracy of small sizes targets in complex backgrounds, partially obscured objects, and irregularly arranged remote sensing targets.

#### 3) PERFORMANCE COMPARISON BETWEEN ATTENTEION MECHANISMS

To verify the performance of RA(Region Attention) proposed in this paper, and prove the rationality of combining Swin Transformer Block with attention mechanism in RASTn module. The experimental comparison group in this section includes the performance comparison of the attention mechanism of RA with SE [53], CBAM [48], CA [49] and GAM [54], as well as the comparison of using attention mechanism and not using any attention mechanism. Experiments were conducted on the DIOR data set under the same experimental setup, and the experimental results are shown in TABLE.2:

As can be seen from TABLE.2, the recall rate, mAP50 and mAP50:95 of the detection results of RA mechanism are 0.6%, 0.4% and 0.1% higher than those of SE [53] attention mechanism, respectively. Recall, $mAP_{50}$ and $mAP_{50:95}$ of RA are 0.7%, 0.6%, and 0.5% higher than those of GAM [54], respectively. Precision, recall, $mAP_{50}$ and $mAP_{50:95}$ of RA are 1.1%, 0.8%, 1%, and 0.7% higher than those of CBAM [48], respectively. Recall, $mAP_{50}$ and $mAP_{50:95}$ of RA are 0.7%, 0.5%, and 0.3% higher than CA [49], respectively. However, the accuracy of detection results of SE, GAM and CA is all slightly higher than that of RA.

When no attention mechanism module is used, the global background information is lost, and only the local detail information obtained by the Swin Transformer Block is available. The detection result of the model is obviously worse than the combination of Swin Transformer Block and attention mechanism. Specifically, combining Swin Transformer Block with the attention mechanism effectively increases the recall of $mAP_{50}$ and $mAP_{50:95}$ by about 2%. It can be demonstrated that combining the local detail information captured by Swin Transformer Block with the global background feature captured by RA module can effectively improve the accuracy of remote sensing object detection.

In summary, combining with Swin Transformer Block and attention mechanism in this paper can effectively improve the detection accuracy of remote sensing targets, and the comprehensive performance of RA attention mechanism is better than the other four mainstream attention mechanisms. Therefore, the RA attention mechanism and Swin Transformer are used to synthesize the RAST feature extraction backbone RAST.

#### 4) PERFORMANCE COMPARISON BETWEEN RAST YOLO AND OTHER ALGORITHMS

To verify its performance, RAST-YOLO is compared with mainstream algorithms such as Faster RCNN [18], YOLOv3 [32], YOLOv5, Retinanet [14], CF2PN [44] and CBD-E [45] on DIOR, Faster RCNN [18], YOLOv3 [32], YOLOv5, Retinanet [14], MFDF [41] and SGFTHR [39] on TGRS-HRRSD, where the backbone and feature fusion networks of Faster RCNN and Retinanet are Resnet50 [55] and FPN [56], respectively. Moreover, RAST-YOLOs is the lightweight network of RAST-YOLO. Its structure is the same as the RAST-YOLO, but its network width is three-quarters of that of RAST-YOLO, and its number of parameters is about 56% of that of RAST-YOLO. We compare it simultaneously with the mainstream state-of-the-art algorithms mentioned above. The experimental results are shown in TABLE.3 and TABLE.4.

It follows from TABLE.3 that $mAP_{50}$ obtained by RAST-YOLO is 0.698, which is superior to all compared algorithms and about 2% higher than CBD-E [45] and CF2PN [44] in DIOR dataset. In the detection results of each category, RAST-YOLO achieves more significant results than other algorithms in categories of both aircraft and ships which contain multi-scale targets, and in the small car target category, which outperforms CBD-E [45], CF2PN [44] by about 10%. Meanwhile, the best detection results for the complex background targets (categories of baseball fields and tennis courts) are the best. However, the RAST-YOLO is inferior to CBD-E [45] and CF2PN [44] in the detection of bridges and dams, and windmills, respectively.

On TGRS-HRRSD, RAST-YOLO achieves the $mAP_{50}$ of 0.907, which is superior to all other algorithms. Similar to the results on DIOR, RAST-YOLO achieves satisfactory detection accuracy on multi-scale categories. Its AP on aircraft and ship is 0.992 and 0.974, respectively, which is superior to all the compared algorithms. RAST-YOLO exceeds all the compared algorithms. In the detection results of baseball diamonds and baseball fields with complex backgrounds. However, RAST-YOLO is inferior to MFDF [41] and SGFTHR [39] in detecting tennis courts, harbors and vehicles, respectively.

Moreover, the lightweight network RAST-YOLOs also achieved excellent test results in this experiment. RAST-YOLOs only use approximately 56% of the parameters of RAST-YOLO. However, among the test results in the DIOR

**TABLE 3.** Object detection results on DIOR.

| Categories | Faster RCNN | YOLOv3 | RetinaNet | YOLOv5 | CBD-E | CF2PN | RAST-YOLO | RAST-YOLOs |
|---|---|---|---|---|---|---|---|---|
| Airplane | 0.638 | 0.748 | 0.626 | 0.820 | 0.542 | 0.783 | 0.843 | **0.844** |
| Airport | 0.616 | 0.730 | 0.722 | 0.622 | 0.77 | **0.783** | 0.764 | 0.737 |
| Baseball field | 0.669 | 0.690 | 0.682 | 0.767 | 0.715 | 0.765 | **0.787** | 0.75 |
| Basketball court | 0.846 | 0.875 | 0.848 | 0.868 | 0.871 | **0.884** | 0.859 | 0.857 |
| Bridge | 0.280 | 0.320 | 0.505 | 0.354 | **0.446** | 0.370 | 0.402 | 0.376 |
| Chimney | 0.730 | 0.751 | 0.767 | 0.742 | 0.754 | 0.701 | **0.768** | 0.767 |
| Dam | 0.445 | 0.482 | 0.545 | 0.374 | **0.635** | 0.599 | 0.502 | 0.518 |
| Expressway service area | 0.526 | 0.565 | 0.564 | 0.594 | **0.762** | 0.712 | 0.626 | 0.6 |
| Expressway toll station | 0.423 | 0.477 | 0.471 | 0.556 | **0.653** | 0.512 | 0.565 | 0.564 |
| Golf field | 0.712 | 0.752 | 0.749 | 0.619 | **0.793** | 0.756 | 0.771 | 0.742 |
| Ground track field | 0.659 | 0.656 | 0.677 | 0.668 | **0.795** | 0.771 | 0.737 | 0.689 |
| Harbor | 0.486 | 0.538 | 0.427 | 0.560 | 0.475 | 0.568 | 0.611 | **0.613** |
| Overpass | 0.501 | 0.520 | 0.526 | 0.537 | **0.593** | 0.587 | 0.566 | 0.552 |
| Ship | 0.716 | 0.732 | 0.680 | 0.893 | 0.691 | 0.761 | **0.911** | 0.903 |
| Stadium | 0.340 | 0.302 | 0.466 | 0.687 | 0.697 | 0.706 | **0.743** | 0.7 |
| Storage tank | 0.634 | 0.702 | 0.471 | 0.753 | 0.643 | 0.555 | **0.779** | 0.759 |
| Tennis court | 0.775 | 0.836 | 0.7743 | 0.864 | 0.845 | 0.888 | **0.893** | 0.87 |
| train station | 0.403 | 0.486 | 0.400 | 0.423 | **0.594** | 0.508 | 0.533 | 0.548 |
| Vehicle | 0.459 | 0.439 | 0.371 | 0.502 | 0.447 | 0.369 | **0.54** | 0.516 |
| Windmill | 0.698 | 0.723 | 0.719 | 0.748 | 0.831 | **0.864** | 0.762 | 0.759 |
| mAP50 | 0.578 | 0.616 | 0.590 | 0.648 | 0.678 | 0.673 | **0.698** | 0.683 |
| mAP50:95 | 0.358 | 0.427 | 0.396 | 0.390 | - | - | **0.465** | 0.447 |

**TABLE 4.** Object detection results on TGRS-HRRSD.

| Categories | Faster RCNN | YOLOv3 | RetinaNet | YOLOv5 | MFDF | SGFTHR | RAST-YOLO | RAST-YOLOs |
|---|---|---|---|---|---|---|---|---|
| Airplane | 0.978 | 0.985 | 0.976 | 0.991 | 0.984 | 0.973 | **0.992** | 0.989 |
| Baseball diamond | 0.837 | 0.870 | 0.856 | 0.865 | 0.882 | 0.896 | **0.914** | 0.87 |
| Basketball court | 0.635 | 0.670 | 0.603 | 0.722 | 0.709 | 0.578 | **0.725** | 0.691 |
| Bridge | 0.846 | 0.880 | 0.878 | 0.915 | 0.871 | 0.905 | **0.940** | 0.937 |
| Crossroad | 0.862 | 0.832 | 0.859 | 0.894 | 0.855 | 0.927 | **0.942** | 0.935 |
| Ground track field | 0.961 | 0.969 | 0.971 | 0.965 | 0.971 | 0.968 | **0.983** | 0.977 |
| Harbor | 0.927 | 0.935 | 0.914 | 0.947 | 0.925 | **0.962** | 0.948 | 0.943 |
| Parking lot | 0.558 | 0.673 | 0.549 | 0.653 | 0.652 | 0.604 | **0.700** | 0.664 |
| Ship | 0.881 | 0.912 | 0.891 | 0.917 | 0.911 | 0.912 | **0.933** | 0.924 |
| Storage tank | 0.929 | 0.945 | 0.947 | 0.973 | 0.953 | 0.950 | **0.974** | 0.972 |
| T junction | 0.646 | 0.715 | 0.632 | 0.779 | 0.708 | 0.742 | **0.849** | 0.846 |
| Tennis court | 0.908 | 0.911 | 0.905 | 0.912 | **0.936** | 0.849 | 0.929 | 0.917 |
| Vehicle | 0.935 | 0.924 | 0.922 | 0.956 | 0.940 | **0.964** | 0.959 | 0.94 |
| mAP50 | 0.839 | 0.863 | 0.839 | 0.884 | 0.869 | 0.864 | **0.907** | 0.893 |
| mAP50:95 | 0.543 | 0.582 | 0.553 | 0.563 | 0.484 | - | **0.616** | 0.589 |

dataset, $mAP_{50}$ and $mAP_{50:95}$ are only 1.5% and 1.8% worse than RAST-YOLO, respectively. Moreover, the comprehensive detection results are more excellent than CBD-E [44], CF2PN [43] and other excellent algorithms. However, in the test results on the TGRS-HRRSD dataset, the $mAP_{50}$ and $mAP_{50:95}$ RAST-YOLOs are only 1.4% and 2.7% worse than the RAST-YOLO, respectively. Furthermore, the comprehensive detection results of RAST-YOLOs exceed advanced algorithms such as SGFTHR [39] and MFDF [41]. RAST-YOLOs also achieved excellent results in the speed test. Furthermore, it takes RAST-YOLOs 19.6 to detect each remote sensing image and the detection speed reaches 51.02FPS when detecting remote sensing images with resolution of $640 \times 640$, which can ensure the real-time detection speed.

In conclusion, the results of the RAST-YOLO algorithm for remote sensing object detection are significantly superior to those obtained by Faster RCNN [18], Retinanet [14], YOLOv3 [32] and other algorithms for natural scenes. Compared with the SOTA remote sensing object detection algorithms, the detection results in most categories still have some advantages. Moreover, the lightweight structure of RAST-YOLO can ensure the real-time detection speed and obtain excellent detection results. From the above analysis of the experimental results, it can be shown that the RAST-YOLO proposed in this paper has significant advantage on remote sensing target detection.

#### 5) COMPARISON OF THE VISUAL DETECTION RESULTS BETWEEN RAST-YOLO AND OTHER ALGORITHMS

In this paper, two images are selected from DIOR and TGRS-HRRSD datasets, respectively. The visual detection results are compared between Faster RCNN [18], Retinanet [14], YOLOv3 [32], YOLOv5, and RAST-YOLO, which are shown in Fig. 9. In the pictures of the detection results, the colors of the detection boxes are used to indicate the predicted categories.

The detection results between the algorithms are analyzed as follows. The first image required detecting densely packed aircraft at airfield. Faster RCNN [18], Retinanet [14], and YOLOv5 do not detect all aircraft. They are not as good as RAST-YOLO for small size aircraft. The second image needs to detect a stadium, a ground track field, two tennis courts, and an overpass. Faster RCNN [18], Retinanet [14], and YOLOv3 [32] incorrectly detect the ground track field as a stadium because of their similar appearance. Because the tennis courts and overpasses are very small and located

at the edge of the picture, Faster RCNN [18], Retinanet [14], YOLOv3 [32], and YOLOv5 do not detect them while RAST-YOLO successfully detected them. The third image needs to detect several ships at sea, and the clouds interfere with the detection of ships. YOLOv3 [32] and YOLOv5 incorrectly detect the clouds as ships, while the Retinanet [14] ignores a small-sized ship. However, Faster RCNN [18] and RAST-YOLO successfully detect all the ships. Compared with the GT boxes, the detection frame of RAST-YOLO is more accurately located. The fourth image needs to detect a bridge and a ground track field. The shadow of the tall buildings has some influences on the detection of the bridge. Faster RCNN [18] incorrectly detects this bridge as three bridges. Meanwhile, Faster RCNN [18], Retinanet [14] and YOLOv3 [32] ignored the athletic field. YOLOv5 and RAST-YOLO detect both the bridge and the athletic field, and RAST-YOLO more accurately locates the ground track field than YOLOv5. From the analysis above, it is obvious that RAST-YOLO is more effective than the mainstream target detection algorithms in detecting small-sized targets and complex background targets in remote sensing images. RAST-YOLO can more effectively deal with the interference of remote sensing images such as weather, climate, light, and shadow in the detection process. Thus, the superiority of RAST-YOLO in remote sensing target detection is verified.

## V. CONCLUSION

Complex background targets, small-scale targets, and multi-scale targets are the challenges in remote sensing object detection. Based on the framework of YOLOv5, RA mechanism is proposed and combined with Swin Transformer as the backbone to extract features. It can sufficiently extract global background information and local details of the target, effectively increasing the interaction range of feature information and reducing the impact of complex backgrounds on remote sensing object detection. The proposed C3D module fully integrates deep semantic information with shallow semantic information to build a more effective feature pyramid, and improve the detection accuracy of multi-scale targets and small targets. And the global and local information is fully used again with ACmix Plus Detector to output more accurate categories and target localization. Experimental comparison and analysis indicate that RAST-YOLO significantly outperforms the mainstream natural scene target detection algorithm and shows certain advantages compared with other excellent remote sensing target detection algorithms, which provides new ideas and insights for researchers to process remote sensing images.

## REFERENCES


[1] H. Lee, H. K. Jung, S. H. Cho, Y. Kim, H. Rim, and S. K. Lee, "Real-time localization for underwater moving object using precalculated DC electric field template," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5813–5823, Oct. 2018.

[2] I. Muhammad, K. Ying, M. Nithish, J. Xin, Z. Xinge, and C. C. Cheah, "Robot-assisted object detection for construction automation: Data and information-driven approach," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 6, pp. 2845–2856, Dec. 2021.

[3] M. Zurowietz and T. W. Nattkemper, "Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration," *IEEE Access*, vol. 8, pp. 143558–143568, 2020.

[4] B. Yan, E. Paolini, L. Xu, and H. Lu, "A target detection and tracking method for multiple radar systems," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5114721.

[5] W.-L. Zhao and C.-W. Ngo, "Flip-invariant SIFT for copy and object detection," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 980–991, Mar. 2013.

[6] F. Gao, C. M. Wang, and C. H. Li, "A combined object detection method with application to pedestrian detection," *IEEE Access*, vol. 8, pp. 194457–194465, 2020.

[7] Y. Tang, X. Wang, E. Dellandréa, and L. Chen, "Weakly supervised learning of deformable part-based models for object detection via region proposals," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 393–407, Feb. 2017.

[8] B. Yang, Z. Jia, J. Yang, and N. K. Kasabov, "Video snow removal based on self-adaptation snow detection and patch-based Gaussian mixture model," *IEEE Access*, vol. 8, pp. 160188–160201, 2020.

[9] B. V. Lad, M. F. Hashmi, and A. G. Keskar, "Boundary preserved salient object detection using guided filter based hybridization approach of transformation and spatial domain analysis," *IEEE Access*, vol. 10, pp. 67230–67246, 2022.

[10] A. K. Nsaif, S. H. M. Ali, K. N. Jassim, A. K. Nseaf, R. Sulaiman, A. Al-Qaraghuli, O. Wahdan, and N. A. Nayan, "FRCNN-GNB: Cascade faster R-CNN with Gabor filters and Naïve Bayes for enhanced eye detection," *IEEE Access*, vol. 9, pp. 15708–15719, 2021.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multiBox detector," *Computer Vision ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[15] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," 2018, *arXiv:1808.01244*.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[22] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, vol. 1, 2018, pp. 5–12.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K.: Springer, Aug. 2020, pp. 213–229.

[25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[26] M. Zheng, P. Gao, R. Zhang, K. Li, X. Wang, H. Li, and H. Dong, "End-to-end object detection with adaptive clustering transformer," 2020, *arXiv:2011.09315*.

[27] Z. Dai, B. Cai, Y. Lin, and J. Chen, "UP-DETR: Unsupervised pre-training for object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1601–1610.
[28] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Proc. 16th Eur. Conf.*, Glasgow, U.K.: Springer, Aug. 2020, pp. 323–339.
[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[30] Y. Xu, Y. Yang, and L. Zhang, "DeMT: Deformable mixer transformer for multi-task learning of dense prediction," 2023, *arXiv:2301.03461*.
[31] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.
[32] A. Farhadi and J. Redmon, "YOLOv3: An incremental improvement," in *Proc. Comput. Vis. pattern Recognit.*, vol. 1804. Berlin, Germany: Springer, 2018, pp. 1–6.
[33] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, "Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation," *IEEE Trans. Multimedia*, vol. 24, pp. 1735–1749, 2022.
[34] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.
[35] H. Zhao, C. Wang, R. Guo, X. Rong, J. Guo, Q. Yang, L. Yang, Y. Zhao, and Y. Li, "Autonomous live working robot navigation with real-time detection and motion planning system on distribution line," *High Voltage*, vol. 7, no. 6, pp. 1204–1216, Dec. 2022.
[36] T. Ye, C. Ren, X. Zhang, G. Zhai, and R. Wang, "Application of lightweight railway transit object detector," *IEEE Trans. Ind. Electron.*, vol. 68, no. 10, pp. 10269–10280, Oct. 2021.
[37] H. Wang, H. Pei, and J. Zhang, "Detection of locomotive signal lights and pedestrians on railway tracks using improved YOLOv4," *IEEE Access*, vol. 10, pp. 15495–15505, 2022.
[38] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.
[39] J. Li, H. Zhang, R. Song, W. Xie, Y. Li, and Q. Du, "Structure-guided feature transform hybrid residual network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610713.
[40] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
[41] H. Lv, W. Qian, T. Chen, H. Yang, and X. Zhou, "Multiscale feature adaptive fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[42] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[43] X. Yang, J. Zhao, H. Zhang, C. Dai, L. Zhao, Z. Ji, and I. Ganchev, "Remote sensing image detection based on YOLOv4 improvements," *IEEE Access*, vol. 10, pp. 95527–95538, 2022.
[44] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, "CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection," *Remote Sens.*, vol. 13, no. 5, p. 847, Feb. 2021.
[45] J. Zhang, C. Xie, X. Xu, Z. Shi, and B. Pan, "A contextual bidirectional enhancement method for remote sensing image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4518–4531, 2020.
[46] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.
[47] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," 2021, *arXiv:2111.14556*.
[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.
[49] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," 2021, *arXiv:2103.02907*.
[50] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.
[51] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
[52] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N Balasubramanian, "Grad-CAM++: Improved visual explanations for deep convolutional networks," 2017, *arXiv:1710.11063*.
[53] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*.
[54] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," 2021, *arXiv:2112.05561*.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[56] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

**XUZHAO JIANG** received the B.M. degree from Anhui Jianzhu University, Hefei, China, in 2021. He is currently pursuing the M.S. degree with the Department of Statistics, Wuhan University of Technology, Wuhan, China. His research interests include remote sensing object detection and semantic segmentation, machine learning, and deep learning.

**YONGHONG WU** received the degree from the Department of Mathematics, Huazhong University of Science and Technology, Wuhan, China, in 2005, and the Ph.D. degree from the Department of Control Science and Engineering, Huazhong University of Science and Technology, in 2011. He has been an Associate Professor with the School of Science, Wuhan University of Technology, Wuhan. His current research interests include coordinated control and optimization of multiagent networks, fault-tolerant control, complex dynamical networks, impulsive and hybrid systems, and dissipative systems.