IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 August 2023, accepted 1 October 2023, date of publication 10 October 2023, date of current version 18 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323575

RESEARCH ARTICLE

EST-YOLOv5s: SAR Image Aircraft Target Detection Model Based on Improved YOLOv5s

MIN HUANG[1], WEIHAO YAN[1], WENHUI DAI[1], AND JINGYANG WANG[1,2]

[1]Hebei University of Science and Technology, Shijiazhuang 050018, China
[2]Hebei Intelligent Internet of Things Technology Innovation Center, Shijiazhuang, Hebei 050018, China

Corresponding author: Jingyang Wang (jingyangw@hebust.edu.cn)

This work was supported in part by the Defense Industrial Technology Development Program under Grant JCKYS2022DC10, and in part by the Foundation of Hebei Intelligent Internet of Things Technology Innovation Center under Grant KFZD2201.

ABSTRACT Due to the diversity of aircraft target scale and interference of background strong scattering in synthetic aperture radar (SAR) images, it is a challenge for target detection tasks. In response to these problems, this paper proposes a new SAR image aircraft target detection model named EST-YOLOv5s. The proposed model integrates the Efficient Channel Attention (ECA) mechanism into the C3 module of the backbone network, which enhances the scattering features of aircraft targets and suppresses irrelevant background information without increasing the number of parameters. Secondly, replace the bottleneck module in the last C3 module in the backbone network with the Swin Transformer Block. By using the shifted window partitioning approach to obtain the global perception ability, the problem of missed detection of small objects is improved. Finally, the Task-Specific Context Decoupling (TSCODE) head is used to balance the relationship between classification and regression so that different contextual details can be better utilized. In this paper, the SAR Aircraft Detection Dataset (SADD) is used as the experimental data set to compare with the baseline model YOLOv5s. The experimental outcomes indicate that the recall of the EST-YOLOv5s model reached 94.2%, the precision reached 97.3%, and the mAP@50 reached 97.8%, which were 2.3%, 1.7%, and 1.7% higher than YOLOv5s respectively. Furthermore, our model also meets the real-time requirements in terms of speed and exhibits strong anti-interference ability.

INDEX TERMS Aircraft target detection, anti-interference, SAR, EST-YOLOv5s, ECA, swin transformer, TSCODE head.

I. INTRODUCTION

Synthetic Aperture Radar (SAR) is an important means of obtaining ground object information using radar technology. Fig. 1 illustrates the process of creating a SAR image. The radar equipment transmits a sequence of pulse signals initially. These signals move at a high speed and are bounced back by the ground. Finally, the returned echo signals are analyzed to produce a radar image. Compared with optical remote sensing technology, SAR has unique advantages, such as getting high-quality images at night or under cloud cover regardless of the type of ground objects, and has high penetration capabilities. So, it is widely used in military security, environmental monitoring, resource exploration, and other fields [1], [2]. Target detection is a hot research topic in high-resolution synthetic aperture radar, receiving attention from numerous scholars. However, aircraft targets differ from ship and vehicle targets as they are high-value and time-sensitive [3]. The rapid and accurate detection of aircraft targets plays a crucial role in acquiring real-time military intelligence, such as assessing enemy combat effectiveness. It holds significance in operational decision-making [4] and civil aviation schedules [5].

In the past, SAR aircraft target detection algorithms mainly relied on extracting features from images. Based on the number of extracted feature selections, these algorithms can be classified into three types: algorithms with a single feature, algorithms with multiple features, and algorithms focused on expert systems [6]. Among them, the most common detection algorithm based on a single feature is the constant false alarm

The associate editor coordinating the review of this manuscript and approving it for publication was Fabrizio Santi.

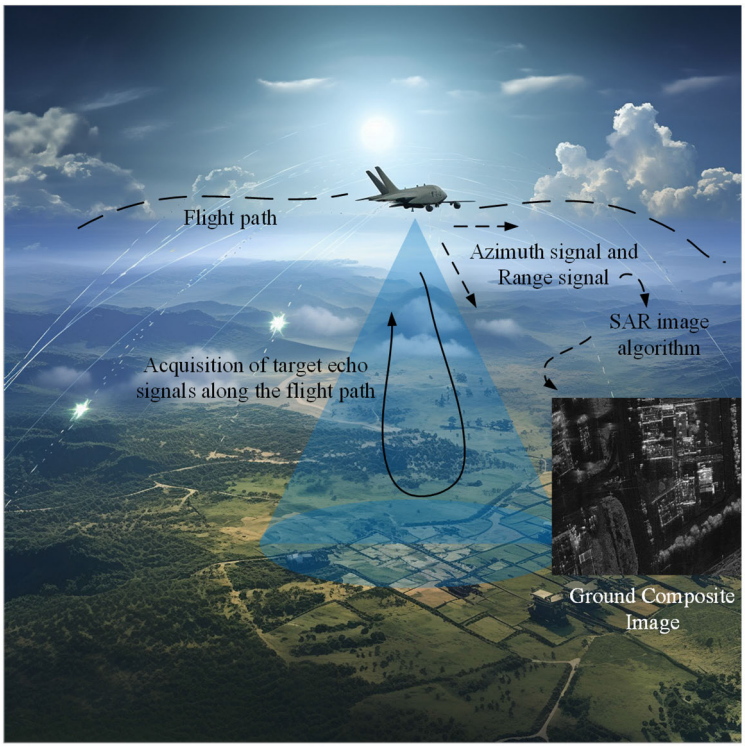

FIGURE 1. SAR imaging process.

rate algorithm (CFAR) [7] based on clutter statistics and threshold extraction. On this basis, researchers have delved into in-depth investigations concerning statistical features and non-uniform backgrounds. Many SAR target detection algorithms based on CFAR have been proposed and developed rapidly, such as superpixel-based CFAR (S-CFAR) [8], iterative-censoring CFAR (IC-CFAR) [9], intensity-space domain CFAR (ISD-CFAR) [10], outliers-robust CFAR (OR-CFAR) [11], segmentation and nonparametric CFAR (SnP-CFAR) [12], etc. In addition to CFAR-based methods, non-CFAR is mainly a target detection algorithm that takes the region as the basic unit and extracts the regional features using global and local regional saliency and local regional entropy change. The detection algorithm based on a single feature only relies on the strong scattering characteristics of the target for decision-making. But the multi-feature-based detection algorithm uses two or more features for detection, which is more suitable for scenarios with high target density or a significant presence of strong scattering clutters [13], [14], [15], [16], [17]. Expert system-based methods are mainly classified into template-based and model-based recognition methods. Template-based methods build a template library by extracting features from known types of targets and then similarly matching these features with those of the target under test. Still, this method relies on the similarity of the database to the test set. The model-based approach builds a model for each target category and stores it in the database to find the hypothesis prediction with the highest similarity to the image to be tested [18]. Nevertheless, due to inherent limitations in robustness and feature representation capabilities, traditional techniques cannot achieve desirable results for SAR target detection effects in complex background scattering and discrete scattering distribution scenarios.

In recent times, the rapid progress of artificial intelligence technology has led to the widespread utilization of advanced deep-learning models in object detection. Convolutional neural network (CNN), as the mainstream deep learning algorithm in target detection, has strong end-to-end feature extraction capabilities and is applied extensively across different computer vision [19], [20]. Cui et al. [21] proposed a pyramid network based on dense attention, which combines prominent features with global non-fuzzy features to enhance the accuracy of target detection in SAR images effectively. Wei et al. [22] proposed a high-resolution feature pyramid structure, which connects sub-networks from high-resolution to low-resolution in parallel, thereby enhancing the salient information of the target in the network and improving the detection effect of the algorithm. Aiming at the problem of redundant feature maps, Lin et al. [23] proposed a faster R-CNN algorithm based on compression and incentive mechanism for ship target detection in SAR images, encoding and screening the feature vectors extracted by the neural network, reversely correcting the sub-feature maps, ultimately enhancing the network's detection capabilities.

However, various interferences and noises often affect SAR images, such as atmospheric disturbances, surface clutter, electronic interference, etc., which may lead to image

quality degradation and affect subsequent analysis and applications. When dealing with SAR images that contain aircraft targets of varying sizes and numerous tiny targets, achieving satisfactory results becomes challenging for the algorithm above. In addition, during the aircraft detection process, the strong scattering points of the background objects distributed near the aircraft will weaken the scattering points of the aircraft targets, causing the scattering points of aircraft components to be confused with those of ground objects. This makes it challenging to identify and pinpoint the aircraft.

Since YOLOv5s is a model with a lighter depth and width in the YOLOv5 series, it performs well in terms of target detection speed. Therefore, in response to these problems, this paper proposes an EST-YOLOv5s model with anti-interference ability based on YOLOv5s to detect aircraft targets in SAR images. The main contributions of this paper are as follows:

(1) Combining Efficient Channel Attention with the C3 module in the backbone network of the YOLOv5s model forms a new module named ECAC3. This module enhances useful features in the network, suppresses irrelevant and redundant feature channels, and reduces the processing and computation of useless information by adaptively adjusting the correlation weights between channels. The network's perception of critical features is improved, making the network more sensitive to capturing the details of the target object.
(2) As the network structure becomes more and more complex, the feature information of most targets in SAR images is gradually lost. Therefore, the Swin Transformer Block and the C3 module are fused into the C3STR structure. It can better capture the global context information and establish long-distance feature dependencies. At the same time, the features of different positions are weighted and fused so that the global features can better participate in the target detection task, thereby improving the accuracy of target detection.
(3) This paper abandons the traditional coupled head in YOLOv5s and integrates the task-specific context decoupling head. By further decoupling the object detection task into two sub-tasks of object classification and bounding box regression, the network can be more focused on learning the specific features of their respective tasks. Introducing task-specific contextual information can enhance the correlation between object classification and bounding box regression, helping the model better understand the semantic information of targets. Thus, a high-resolution feature map with enhanced edge details is obtained to improve the detection accuracy of small objects.
(4) After introducing the ECAC3 and C3STR modules into the backbone network of YOLOv5s, it can more accurately capture the important features of aircraft targets in SAR images, while suppressing irrelevant information such as background clutter. In addition, introducing the TSCODE header can improve the model's ability to combine contextual semantic features, thereby enhancing its ability to understand the target and reducing its sensitivity to noise interference. In this paper, by simulating different levels of noise, the anti-interference experiment is carried out on the model, which verifies that EST-YOLOv5s has a strong anti-interference ability.

The subsequent sections of this paper are structured as follows. First, the paper simply introduces related work on object detection based on deep learning and object detection based on SAR images in Section II. Subsequently, this paper introduces the various components and functions of the EST-YOLOv5s model in Section III. Then, Section IV presents the basic situation of the dataset and experimental environment, along with an analysis and discussion of the detection performance and anti-interference ability of EST-YOLOv5s. Finally, we summarize the main work of this paper and future work to be done.

II. RELATED WORK

A. APPROACHES TO OBJECT DETECTION USING DEEP LEARNING ALGORITHMS

Current object detection algorithms based on convolutional neural networks can be categorized into two types: two-stage algorithms, which rely on candidate regions, and single-stage algorithms based on classification and regression. In 2014, Girshick et al. [24] proposed a two-stage target detection algorithm R-CNN, which uses a selective search algorithm to obtain candidate regions, CNN extraction features, image classification, and border regression for target detection. However, due to problems such as redundant calculation of candidate boxes, the efficiency of this approach is low. In 2015, Li et al. [25] introduced a spatial pyramid pooling layer to resolve this problem and proposed SPP-net, which greatly shortened the training time. On this basis, Girshick [26] combined the benefits of R-CNN and SPP-net, introducing Fast R-CNN, which samples fixed-size convolutional feature maps from candidate boxes of various sizes and only employs one scale for grid divide and merge. The calculation speed has been further improved, but there is still the problem of a large amount of calculation. On the basis of Fast R-CNN, Ren et al. [27] proposed Faster R-CNN, added the Region Proposal Network (RPN) to the backbone network, and extracted candidate frames in the convolutional feature layer of RPN by setting anchor points of different scales to achieve end-to-end training of networks. Nonetheless, relying solely on high-level features for object prediction while disregarding low-level features poses difficulty in detecting small objects. To increase the network's sensitivity to translation changes and improve target positioning accuracy, Dai et al. [28] shared the calculation of the region of interest and proposed the R-FCN structure. In addition, they further offered a Feature Pyramid Network (FPN) [29], which effectively

improves the network's performance in small object detection by utilizing multi-scale features and a top-down structure. This approach enables the network to handle objects of different scales better and improves accuracy on small objects.

The single-stage detection method eliminates the candidate box generation and screening process used in the traditional two-stage detection method. It integrates the detection task directly into a single neural network. Redmon et al. [30] introduced the You Only Look Once (YOLO) algorithm, which predicts targets classifying and regressing the input image. On this basis, Liu et al. [31] combined the anchor point method of Faster R-CNN with the regression idea of YOLO and then proposed the Single Shot MultiBox Detector (SSD). This method attains comparable accuracy to the two-stage approach while also operating at a swift pace. YOLOv2 [32], uses an anchoring mechanism based on k-means clustering to improve detection accuracy. Lin et al. [33] proposed the single-stage detection algorithm RetinaNet by using Focal loss to focus more on difficult-to-classify samples; the weight of easy-to-classify pieces is reduced, and its impact on training is reduced. It surpassed the existing two-stage algorithm in terms of accuracy. Yolov3 [34], proposed by Redmon et al., achieves significant improvements in accuracy and speed by introducing a more powerful backbone network, multi-scale detection, and improved training techniques. Although the accuracy of the YOLO algorithm is continuously improving, parameters and complexity are increasing sharply, and in turn, slows down the detection speed of the network. Wang et al. [35] propose a new backbone network, Cross Stage Partial Network (CSPNet), which effectively facilitates feature propagation and reuse by introducing cross-stage partial connections. YOLOv4 [36] proposed using CSPDarknet53 as the backbone feature extraction network, further improving the detection speed. Subsequently, YOLOv5 [37] emerges and garners increased attention from researchers due to its advantages in terms of speed and accuracy.

B. SAR TARGET DETECTION METHOD IN DEEP LEARNING

With the continuous progress of SAR imaging methods, using deep learning to process SAR images has become a new research hotspot. In 2017, Dou et al. [38] proposed a reconstruction method for aircraft targets using shape priors to accurately extract contour shape features, providing effective prior information for target recognition. In 2018, He et al. [39] introduced a technique that leverages depth shape priors to reconstruct aircraft in high-resolution SAR images. Used deep learning to model aircraft shape features and super-resolution reconstruction technology to achieve more accurate and realistic aircraft reconstruction results. In 2019, An et al. [40] proposed an improved object detector DRBox-v2, which improves the performance and effect of object detection in SAR images by introducing rotatable bounding boxes and Rotated Region of Interest (RROI) pooling operations. At the same time, they proposed a method that combined focus loss and difficult sample mining to improve the imbalance between positive and negative samples. Zhang et al. [41] proposed a new high-speed SAR ship detection method based on a deep separable convolutional neural network (DS-CNN). They adopted a lightweight network architecture and utilized D-Conv2D and P-Conv2D instead of the traditional C-CNN, significantly increasing the ship detection speed. In 2020, Zhang et al. [42] proposed a cascaded three-view network to fully characterize aircraft targets by introducing multi-view and an end-to-end training method. On this basis, the target is detected and identified by slice and image processing methods. In order to better capture the feature and context information of the target, Zhao et al. [43] introduced a pyramid attention mechanism and dilated convolution, which improved the accuracy and robustness of aircraft detection. And the accuracy on the Gaofen-3 airport dataset reached 85.68%. Guo et al. [44] introduced scatter transformation before feature extraction and designed an attention pyramid module to adaptively select and focus on important aircraft features, significantly improving the aircraft detection task. In 2022, aiming at the multi-scale problem of SAR ship targets in complex scenes, Guo et al. [45] proposed an improved YOLOv5 detection method named YOLOv5s_CBAM_BiFPN, using a Convolutional Block Attention Module (CBAM) and Bidirectional Feature Pyramid Network (BiFPN), which solved the problem of missed detection of multi-scale objects. Ge et al. [46] introduced a new spatial orientation attention module based on the YOLOX framework. And fused it with the path aggregation feature pyramid to capture feature transformations in different directions to highlight the features of aircraft targets in SAR images. Xu et al. [47] proposed the SBN-3D-SD model to improve shadow detection and tracking accuracy in Video-Synthetic Aperture Radar (Video-SAR) images. By taking advantage of the sparse properties of shadows, the low-rank properties of the background, and the Gaussian properties of shadows, the shadows are enhanced by decomposition in 3D space, which improves the accuracy of various shadow detection and tracking algorithms. Meanwhile, they also proposed a network model GWFEF-Net [48] for improving the performance of dual-polarization ship detection in synthetic aperture radar (SAR) images. This model improved ship detection performance by leveraging the dual-polarization property through feature enrichment, enhancement, fusion, and channel attention mechanisms. In 2023, Li et al. [49] proposed the model YOLOv5–L+ BiFPN + Swin Transformer + GAM. The proposed model solved the problem of detecting small-sized aircraft targets in SAR images by utilizing its powerful feature extraction capability.

III. METHODOLOGY

A. MODEL OF YOLOv5s

YOLOv5 is a fast and accurate object detection model developed by Ultralytics and released in 2020. In 2022, Xu et al. [50] proposed a lightweight SAR shipborne

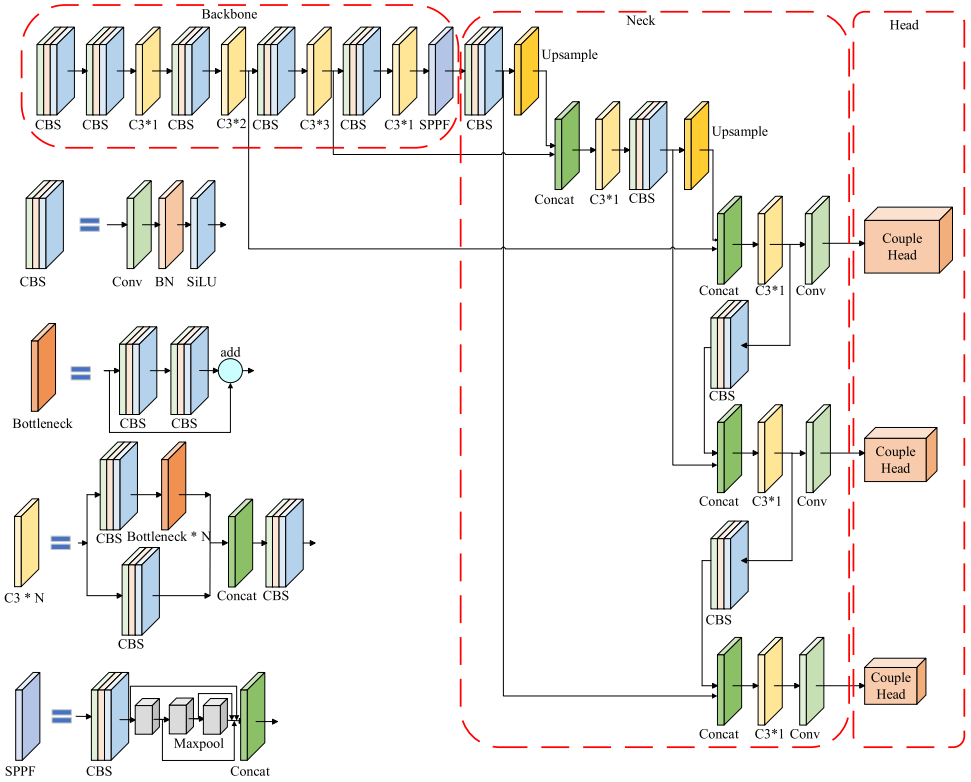

FIGURE 2. YOLOv5s network structure.

detector, Lite-YOLOv5, for detecting ships in maritime environments using synthetic aperture radar satellites. This method achieved high detection accuracy while reducing the model's size and computational requirements. Inspired by this, the present study selects YOLOv5s as the baseline model for detecting aircraft targets in SAR images.

As shown in Fig. 2, the basic structure of YOLOv5s is mainly composed of a backbone, neck, and detection head. The backbone network part specifically includes the CBS module, C3 module, and SPPF module. The CBS module comprises a convolutional layer, a batch normalization layer, and a non-linear activation function named SiLU. The C3 module plays a crucial role in YOLOv5s as it primarily focuses on feature fusion, reducing the model's dimension, and enhancing the representation ability of the feature map. It contains three CBS modules and N bottleneck structures. The bottleneck structure is similar to the residual structure of ResNet, which can better perform feature functions. The SPPF structure widens the perception area, and extracts and fuses advanced features by using a CBS module and three serial maximum pooling structures.

The structure of the neck part adopts an FPN structure and a PAN structure so that advanced strong semantic features are passed down to compensate for and enhance positioning information.

As the last part of YOLOv5s, the head uses anchor boxes and grid generation strategies to detect targets of different sizes at different scales to generate prediction boxes and corresponding category probabilities.

B. IMPROVED YOLOv5s NETWORK STRUCTURE

Based on the structure of YOLOv5s, this paper designs a new EST-YOLOv5s model for detecting aircraft targets in SAR images. The overall structure of the model is shown in Fig. 3. The lightweight channel attention structure is integrated into the first three C3 modules of the backbone network to constitute a new ECAC3 structure, which suppresses complex background information and enhances the ability to extract small objects while keeping the number of parameters unchanged. Because most aircraft targets are in the airport area, the targets are relatively dense, and other obstacles near the airport will disturb the scattering signature of the aircraft targets, which is prone to missed or false detections. Therefore, this paper proposes to replace the bottleneck structure in the last C3 structure in the backbone network with a Swin Transformer Block to enhance the perception of local geometric features. Relying solely on the learning ability of coupled-head networks to provide task-specific contextual information from shared feature maps often shows an imbalance between classification and regression tasks. Traditional methods generate more computational overhead, require longer training time, and reduce the efficiency of reasoning. Therefore, this paper proposes using the TSCODE head to handle complex context conflicts and improve network performance.

1) ECAC3 MODULE

As a computational model, the attention mechanism simulates the selectivity of human visual and cognitive processes

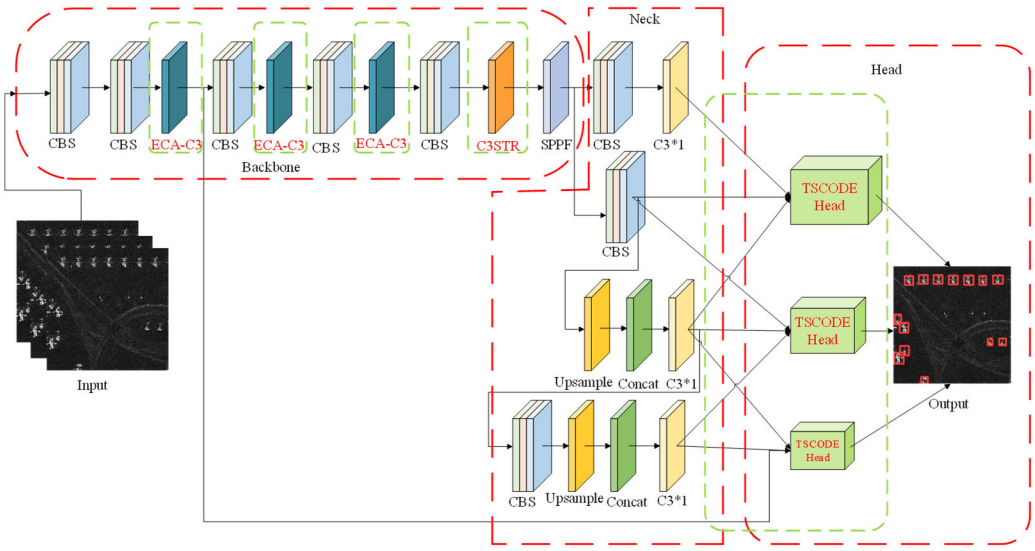

FIGURE 3. EST-YOLOv5s network model structure.

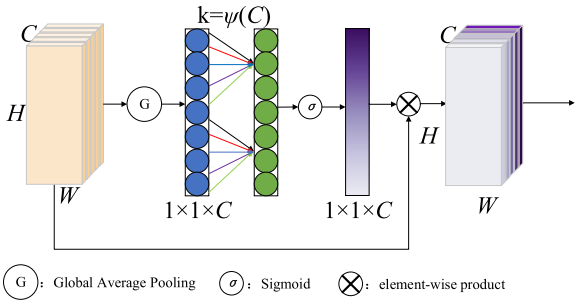

FIGURE 4. Structure of the ECA module.

and is widely used in natural language processing, computer vision, and other fields. Squeeze Excitation (SE) attention [51] and CBAM [52] are the most representative. Due to the high computational complexity of the former two, and ignoring that all spatial locations share the same channel attention weight. The ECA attention mechanism uses simple one-dimensional convolution operations, which does not significantly increase the computational overhead. It helps improve the feature extraction capabilities of the C3 module, enabling the network to learn more discriminative features and adaptively allocate attention to relevant channels. So, this paper chooses the ECA [53] mechanism that is both lightweight and retains information between different dimensions. Fig. 4 shows the composition of the ECA attention module.

First, the aggregated features of the feature map are obtained using global average pooling on feature maps of input size $C \times H \times W$, and channel weights are generated by performing a fast one-dimensional convolution of size $k = \psi(C)$. Among them, the hyperparameter $k = \psi(C)$ calculation method involved in one-dimensional convolution is shown in the following formula (1).

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{1}$$

The parameter k in this formula represents the size of the convolution kernel, which is adaptively determined by the mapping of the channel dimension C. Then, the sigmoid operation is performed on the obtained feature map to obtain a new feature map with a size of $1 \times 1 \times C$. Finally, the obtained channel attention weights are multiplied with the original feature map channel by channel to get a feature map with an output feature size of $C \times H \times W$. Among them, the meaning of C is the channel dimension, $|t|_{\text{odd}}$ means that the upward value is the odd number closest to t, and γ and b are 2 and 1, respectively, coefficients of a linear relationship. This paper integrates the ECA module into the C3 module in the backbone network of YOLOv5s. That is, after the feature map output by the convolutional layer is subjected to the global average pooling operation, the information on the channel dimension is obtained, and the local cross-channel interaction is realized through one-dimensional convolution, which avoids the loss of information caused by the reduction of the dimension. After adaptively weighting each channel, the greater the weight of the obtained channel, the more critical the feature represented by the channel. Conversely, subsequent feature extraction layers can discard smaller channel weights to enhance the representation of important features and suppress unimportant features. Thereby improving the detection accuracy of the model. As shown in Fig. 5, this figure shows the specific structure of the improved C3 module.

2) C3STR MODULE

Originally, Transformer was proposed by Google in the paper "Attention is All You Need" [54] for the field of natural

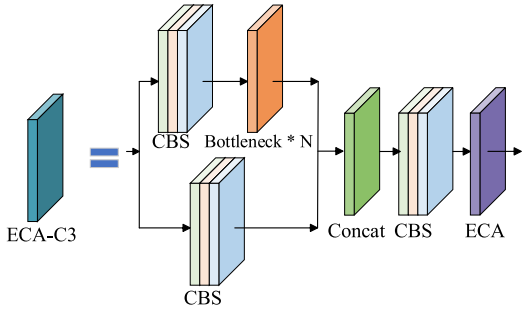

FIGURE 5. Structure of ECAC3 module.

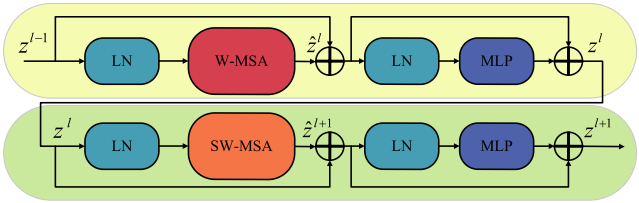

FIGURE 6. Structure of swin transformer block.

language processing. Vision Transformer [55] was introduced into computer vision in 2020 and achieved remarkable results. This shift has led to a series of networks and papers such as DeiT [56], Swin Transformer [57], DETR [58], SETR [59], and GANsformer [60], among others. These models exhibit excellent performance and application potential in computer vision tasks. Researchers have opened up new possibilities and achieved impressive achievements by applying the Transformer model to computer vision.

Since the original C3 module of YOLOv5s cannot obtain enough global context information, Transformer can compensate for this defect and improve the recognition effect of small targets in complex backgrounds. However, Transformer requires a lot of computing power, so this article chooses Swin Transformer Blocks to improve the C3 module. The Swin Transformer Block consists of two key self-attention modules: W-MSA and SW-MSA, which are integral components. A GELU nonlinear multi-layer perceptron (MLP) follows each module. In Fig. 6, each MSA module and MLP is connected to a LayerNorm (LN) layer, and a residual structure is added for connection.

As a self-attention mechanism, W-MSA often divides images without overlapping when processing images and computes self-attention in each window, respectively. If each window has M×M blocks, then the complexities of MSA and W-MSA are calculated by formulas (2) and (3).

$$\Omega(\text{MSA}) = 4\text{HWC}^2 + 2(\text{HW})^2\text{C} \tag{2}$$

$$\Omega(\text{W-MSA}) = 4\text{HWC}^2 + 2\text{M}^2\text{HWC} \tag{3}$$

In formulas (2) and (3), C represents a constant, and M generally takes the default value of 7. According to the formula, the computational complexity of W-MSA is linear concerning the size of the input image, while that of MSA is

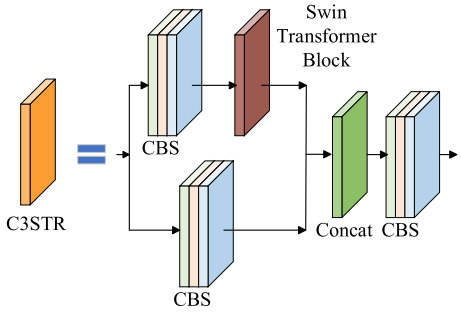

FIGURE 7. Structure of C3STR.

quadratic. In this way, W-MSA has reduced complexity but is short of information interaction across windows. By shifting the window division method, SW-MSA makes up for this shortcoming with a comprehensive perception of the overall information. Formulas (4), (5), (6), and (7) represent the calculation process of the Swin Transformer block.

$$\hat{\text{z}} = \text{W-MSA}\left(\text{LN}\left(\text{z}^{l-1}\right)\right) + \text{z}^{l-1} \tag{4}$$

$$\text{z}^{l} = \text{MLP}\left(\text{LN}\left(\hat{\text{z}}^{l}\right)\right) + \hat{\text{z}}^{l} \tag{5}$$

$$\hat{\text{z}}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(\text{z}^{l}\right)\right) + \text{z}^{l} \tag{6}$$

$$\text{z}^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{\text{z}}^{l+1}\right)\right) + \hat{\text{z}}^{l+1} \tag{7}$$

Among them, the four parameters $\hat{\text{z}}^{l}$, z^{l}, $\hat{\text{z}}^{l+1}$, and z^{l+1} represent the feature map information of the W-MSA, MLP, SW-MSA, and MLP output in sequence. By using sliding windows and cross-window connections, Swin Transformer can interact with features at different spatial locations and utilize multi-level feature extraction to obtain a more comprehensive image representation.

After referring to the studies of C3NRT [61] and C3-Trans [62], this article integrates the Swin Transformer Block into the C3 module to obtain a new C3STR module, as shown in Fig. 7. In this new structure, the original bottleneck block is substituted with the Swin Transformer Block. This module utilizes the correlation between different positions to globally model the pixels in the feature map through the selfattention mechanism and then propagates and integrates features of different scales in the feature map through the cross-window attention mechanism to increase the receptive field. This way, the association of feature information between different windows is realized, thereby improving the network's ability to detect multi-scale targets.

3) TSCODE HEAD

The detection head obtains prediction results by detecting objects of different sizes. The YOLOv5s model achieves this task by coupling detection heads. However, YOLOX [63] pointed out that the coupling detection head may damage the performance and affect the accuracy of network detection [64], [65]. It introduced the decoupling detection head into

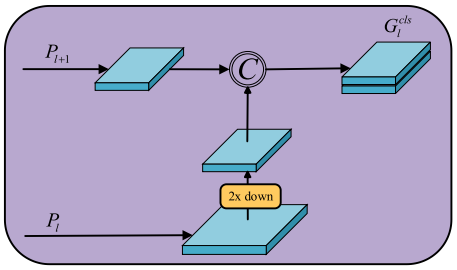

FIGURE 8. Semantic context encoding for classification (SCE).

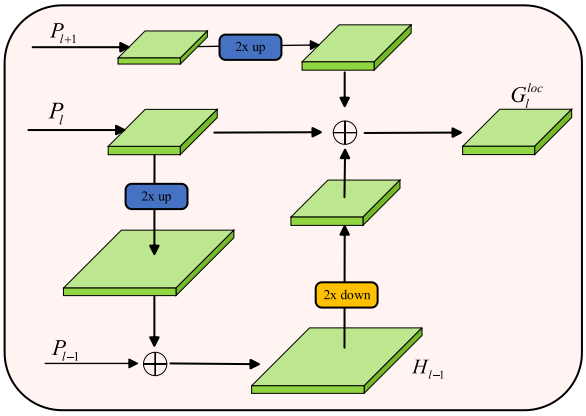

FIGURE 9. Detail-preserving encoding for localization (DPE).

the YOLO family for the first time, improving the network convergence speed and performance. The TSCODE head decouples the feature encoding of object classification and frame regression based on the original. It utilizes feature maps with different semantic contexts in the two branches. The classification branch generates spatially coarser but semantically richer feature maps. While the localization branch provides feature maps containing more detailed texture and boundary information. The encoding process used for classification and localization is shown in Fig. 8 and Fig. 9.

As shown in Fig. 8, for each pyramid level l, SCE utilizes feature maps from two levels, P_l and P_{l+1}, to generate semantically rich feature maps for classification. Specifically, after P_l is down sampled by 2 times, it is connected with P_{l+1} to generate the final G_l^{cls}. The formula is shown in formula (8).

$$G_l^{cls} = Concat(DConv(P_l), P_{l+1}) \tag{8}$$

where Concat and DConv denote concatenation and shared downsampling convolutional layers, respectively. In this way, not only can the sparsity of the salient features of the P_l layer be utilized, but it also benefits from the rich context semantics of the P_{l+1} layer. It is more helpful to infer those aircraft targets with weak scattering characteristics in SAR images.

Likewise, as shown in Fig. 9, DPE accepts feature maps from three pyramid levels, namely P_{l-1}, P_l, and P_{l+1}. P_{l-1} provides more detail and edge features, while P_{l+1} provides a more comprehensive perspective of the object. First, P_l is upsampled two times and then aggregated with P_{l-1}. After

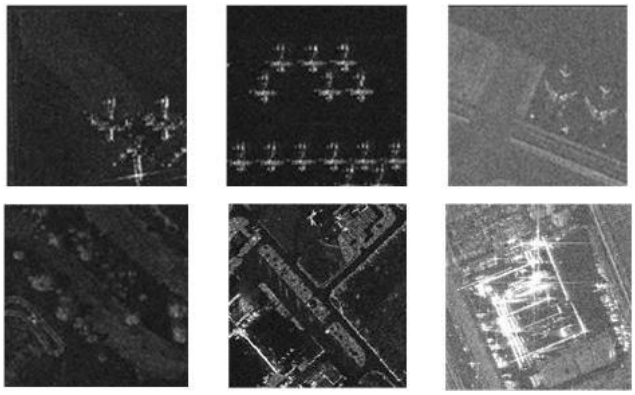

FIGURE 10. Examples of part of the dataset's positive sample and negative sample slice.

the obtained feature map H_{l-1} is downsampled by two times, it is aggregated with P_{l+1} and P_l after two times upsampling to generate the final G_l^{loc}. The formula is shown in formula (9).

$$G_l^{loc} = P_l + \mu(P_{l+1}) + DConv(\mu(P_l) + P_{l-1}) \tag{9}$$

where μ denotes upsampling, and DConv denotes another shared downsampling convolutional layer. This method can obtain more texture details and boundary information to accurately predict the SAR aircraft target's location.

This section proposes solving the conflict between classification and regression tasks by replacing the coupling head in YOLOv5s with the TSCODE [66] head. Deep feature maps are richer in contextual semantic information. Classification tasks focus more on which class the extracted features are most similar to the existing category. Therefore, fusing each level with its deeper feature maps is more helpful in improving classification confidence. In the positioning task, the main focus lies in assessing the level of coincidence between the predicted and ground-truth box position coordinates. Combining the feature map of each layer with its adjacent deep and shallow features can obtain richer spatial detail information for bounding box parameter correction, making the bounding box more precisely snap to the target. Therefore, the TSCODE head can effectively improve the accuracy of SAR aircraft target detection.

IV. EXPERIMENT

A. DATASET

This paper uses the SAR Aircraft Detection Dataset (SADD) [67] to prove the credibility and effect of the EST-YOLOv5s model. This dataset was obtained from the TerraSAR-X satellite, and the image resolution ranges from 0.5 to 3 meters. Some positive and negative sample slices in the dataset are shown in Fig. 10.

The positive samples of this data set contain various complex target backgrounds, such as airport runways, parking lot airports, etc. The negative samples mainly include

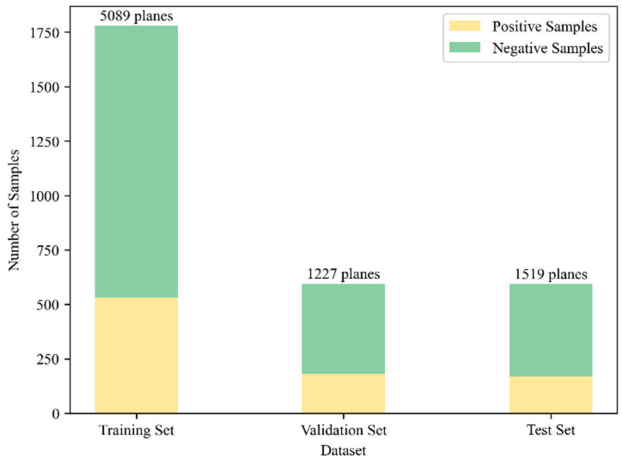

FIGURE 11. Dataset partition.

TABLE 1. Configuration parameters of experimental platform.

Parameter	Configuration
Operating System	Ubuntu 20.04 (64-bit)
Central Processing Unit	Intel(R) Xeon(R) Platinum 8255C CPU@2.50 GHz
Random Access Memory	40 GB
GPU accelerated environment	Cuda11.1, cuDNN8.0.5
Graphic Processing Unit	NVIDIA GeForce RTX 2080 Ti (11 GB)
Framework	PyTorch1.9.0
Programming language	Python 3.8

open spaces and forests around the airport. The dataset has 2966 slices of size 224 × 224, including 7835 aircraft targets.

The dataset is randomly divided into training set, verification set, and test set using a ratio of 6:2:2. The training set contains 533 positive samples and 1246 negative samples, with a total of 5089 aircraft. The validation set includes 181 positive samples and 413 negative samples, with a total of 1227 aircraft. The test set includes 169 positive samples and 424 negative samples, with a total of 1915 aircraft. Fig. 11 shows the division of the dataset.

B. EXPERIMENTAL ENVIRONMENT

All experiments in this paper are carried out in the hardware and software environment shown in Table 1.

In this paper, we set several key hyperparameters, including training steps, warmup epoch, warmup initial momentum, batch size, optimization algorithm, initial learning rate, momentum, and weight decay. The specific hyperparameter settings are shown in Table 2.

C. INDICATORS OF EVALUATION

This study evaluates the performance of EST-YOLOv5s by comparing the performance of EST-YOLOv5s with various other models in image detection. Specifically, this paper focuses on the localization accuracy of objects and the ratio of missed or false detections. This paper uses some common performance metrics to measure how good the improved model is. These metrics include precision, recall, mAP, F1, number of parameters, FPS, etc. In order to obtain the final experimental data, this experiment uses the previously divided training set and verification set to train the model and the test set to test the trained model.

TABLE 2. Hyperparameters of the model.

Hyperparameters	Value
training steps	500 epochs
warmup epoch	3
warmup initial momentum	0.8
batch size during training	16
batch size during testing	1
optimization algorithm	SGD
initial learning rate	0.01
momentum	0.937
weight decay	0.0005

The precision refers to the ratio between the number of positive samples correctly predicted by the model and the total number of positive samples predicted. Its specific calculation method is shown in the following formula (10).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

The recall is the ratio of the number of positive samples correctly predicted by the model to the number of positive samples in all targets. Formula (11) shows how the recall is calculated.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

In formulas (10) and (11), TP represents the number of correct recognitions by the model, FP represents the number of wrong recognitions by the model, and FN represents the number of SAR aircraft targets not detected by the model. In addition, the F1 score is used as a comprehensive evaluation index calculated based on accuracy and recall to comprehensively evaluate the performance of the model. Its calculation formula is shown in formula (12).

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Using the P-R curve can visualize the relationship between precision and recall, where precision is represented on the vertical axis and recall is represented on the horizontal axis.

The proportion of the region enclosed by the curve and the axes defines the mean precision. A good model is able to maintain high precision with gradually increasing recall. If the average precision value is higher, the model performs better. Use formula (13) to calculate the average precision.

$$\text{AP} = \int_0^1 \text{P(R)dR} \tag{13}$$

TABLE 3. Comparison between five different versions of YOLOv5.

Model	Precision	Recall	mAP@50	mAP@50:95	Parameters (M)	FPS (f/s)
YOLOv5n	0.938	0.88	0.933	0.734	1.76	98.03
YOLOv5s	0.956	0.919	0.961	0.797	7.01	102.04
YOLOv5m	0.972	0.907	0.96	0.829	20.85	57.8
YOLOv5l	0.981	0.901	0.963	0.839	46.1	59.52
YOLOv5x	0.974	0.914	0.966	0.844	86.17	51.28

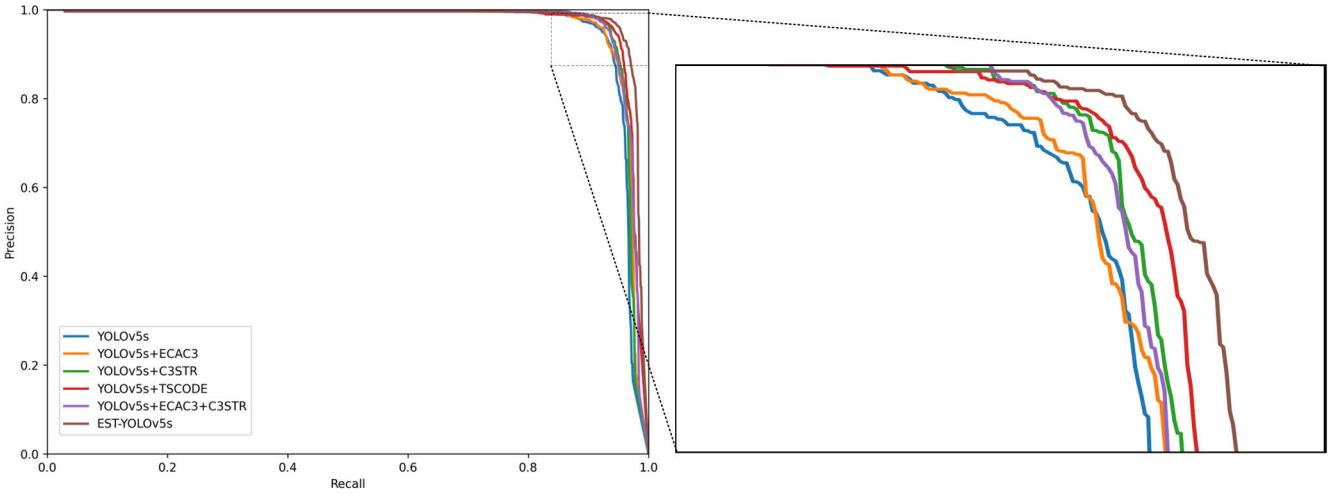

FIGURE 12. P-R curve Comparison of each model in the ablation experiment.

mAP is a metric used to evaluate the accuracy of object recognition. It represents the average accuracy achieved across different categories. A higher mAP indicates better overall detection performance and increased accuracy of the model. The calculation of average precision is expressed by the following formula (14).

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} \text{AP}_i \tag{14}$$

FPS represents the rate at which the algorithm detects and processes images. If it takes t seconds to process each image, the calculation can be expressed using the formula (15).

$$\text{FPS} = \frac{1}{t} \tag{15}$$

In the case of the target detection model, a higher FPS implies reduced latency, indicating superior real-time performance and faster computational speed. The model's efficiency improves as the FPS increases.

D. MODEL SELECTION EXPERIMENT

By comparing the five different versions of YOLOv5, we can conclude from Table 3 that YOLOv5n has the lowest model parameters, but its accuracy is also the lowest. In contrast, YOLOv5m, YOLOv5l, and YOLOv5x have slightly higher accuracy than YOLOv5s, but their number of parameters has dramatically increased. The FPS has also decreased significantly, meaning it will take up more storage space and consume more computing resources and detection time. The SADD dataset used in this paper has a relatively small scale, and employing a smaller model can help mitigate the risk of overfitting. Therefore, this paper uses YOLOv5s as the improved baseline model to find an appropriate balance between speed and accuracy.

E. ABLATION EXPERIMENT

In this paper, there are many improvements in the EST-YOLOv5s model. It mainly contains: integrating the ECA mechanism into the first three C3 modules in the backbone network, substituting the bottleneck block in the last C3 module in the backbone with the Swin Transformer Block, and swapping the coupling head with the TSCODE head. To prove the impact of these methods on EST-YOLOv5s, this paper conducts ablation experiments. Table 4 shows the final results. Where "$\surd$" represents the use of this modular approach. Fig. 12 compares P-R curves after the fusion of other modules and the original model.

As shown in Table 4 and Fig. 12, the performance of the YOLOv5s model after adding various improvement points has improved compared with the original model. The mAP@50 of EST-YOLOv5s is approximately 1.7% higher than that of the baseline network. The number of model parameters in this paper has increased, but the accuracy of

TABLE 4. Results of ablation experiments.

Experiment number	YOLOv5s			mAP@50	mAP@50:95	Precision	Recall	Parameters (M)	FPS (f/s)
	ECAC3	C3STR	TSCODE						
A	×	×	×	0.961	0.797	0.956	0.919	7.01	102
B	√	×	×	0.966	0.796	0.96	0.922	7.01	91
C	×	√	×	0.967	0.804	0.967	0.93	7.15	82
D	×	×	√	0.974	0.827	0.965	0.935	16.14	69
E	√	√	×	0.971	0.793	0.97	0.924	7.15	86
F	√	√	√	0.978	0.825	0.973	0.942	16.27	65

TABLE 5. Comparison of EST-YOLOv5s model with other models.

Model	Precision	Recall	mAP@50	F1	Parameters(M)	FLOPS(G)	FPS(f/s)
Faster R-CNN	0.851	0.957	0.888	0.901	41.12	33.18	36
RetinaNet	0.547	0.949	0.885	0.694	36.1	20.44	42
YOLOv3-SPP	0.973	0.912	0.967	0.942	9.56	23.6	102
YOLOv3-tiny	0.931	0.77	0.856	0.843	8.67	13.0	286
YOLOv5s	0.956	0.919	0.961	0.937	7.01	15.9	102
TPH-YOLOv5	0.962	0.703	0.81	0.812	60.34	138.6	35
YOLOv5s_CBAM_BiFPN	0.976	0.902	0.956	0.938	7.27	16.4	64
YOLOv5-L+BiFPN+SwinTransformer+GAM	0.962	0.905	0.953	0.818	85.36	1103.3	35
EST-YOLOv5s	0.973	0.942	0.978	0.957	16.27	73.3	65

TABLE 6. Comparison of models with different levels of noise interference.

Model	mAP@50 (5%)	mAP@50 (10%)	mAP@50 (15%)	mAP@50 (20%)	mAP@50 (25%)
Faster R-CNN	0.612	0.434	0.263	0.175	0.091
RetinaNet	0.678	0.424	0.320	0.169	0.091
YOLOv3-SPP	0.865	0.669	0.475	0.309	0.175
YOLOv3-tiny	0.553	0.317	0.152	0.094	0.054
YOLOv5s	0.917	0.818	0.687	0.541	0.435
TPH-YOLOv5	0.785	0.758	0.728	0.71	0.677
YOLOv5s_CBAM_BiFPN	0.916	0.818	0.657	0.531	0.442
YOLOv5-L+BiFPN+SwinTransformer+GAM	0.903	0.804	0.674	0.548	0.472
EST-YOLOv5s	0.961	0.924	0.887	0.81	0.747

the model in this paper is 97.8%, which is the highest compared with other models. The FPS can also meet real-time requirements. This paper adds the ECA mechanism to the C3 structure in scheme B. The experimental results show that mAP@50 has increased by 0.5%, precision and recall have increased by 0.4% and 0.3%, respectively, and the parameters remain at 7.01 M. This shows that the ECAC3 module does not increase the complexity of the model while improving the accuracy. In scheme C, this paper introduces Swin Transformer into the C3 module. SW-MSA uses a sliding window to segment the feature map. This method allows certain windows to receive information from multiple windows above, reducing the perception area loss and thereby improving the model's representation ability and nonlinear expression. The results show that mAP@50, mAP@50:95, precision, and recall increased by 0.6%, 0.7%, 1.1%, and 0.3%, respectively. Solution D is to introduce the TSCODE head into the YOLOv5s model to replace the original coupling head structure, separate the two tasks of classification and positioning, and introduce richer semantic information and more edge information features for positioning. This method obtained the highest recall rate of 82.7%. In scheme E, we combine schemes A and B. The mAP, precision, and recall values have been improved compared with schemes A and B, respectively. Scheme F is the result of EST-YOLOv5s, these experimental results show that mAP@50 has increased by 1.3%, mAP@50:95 has risen by 3%, the precision and recall values have also increased by 0.9% and 1.6%, respectively, and the number of parameters has increased significantly. Therefore, the decoupling head is also a part that should not be overlooked, and it is a significant approach to enhancing the performance of the object detection model.

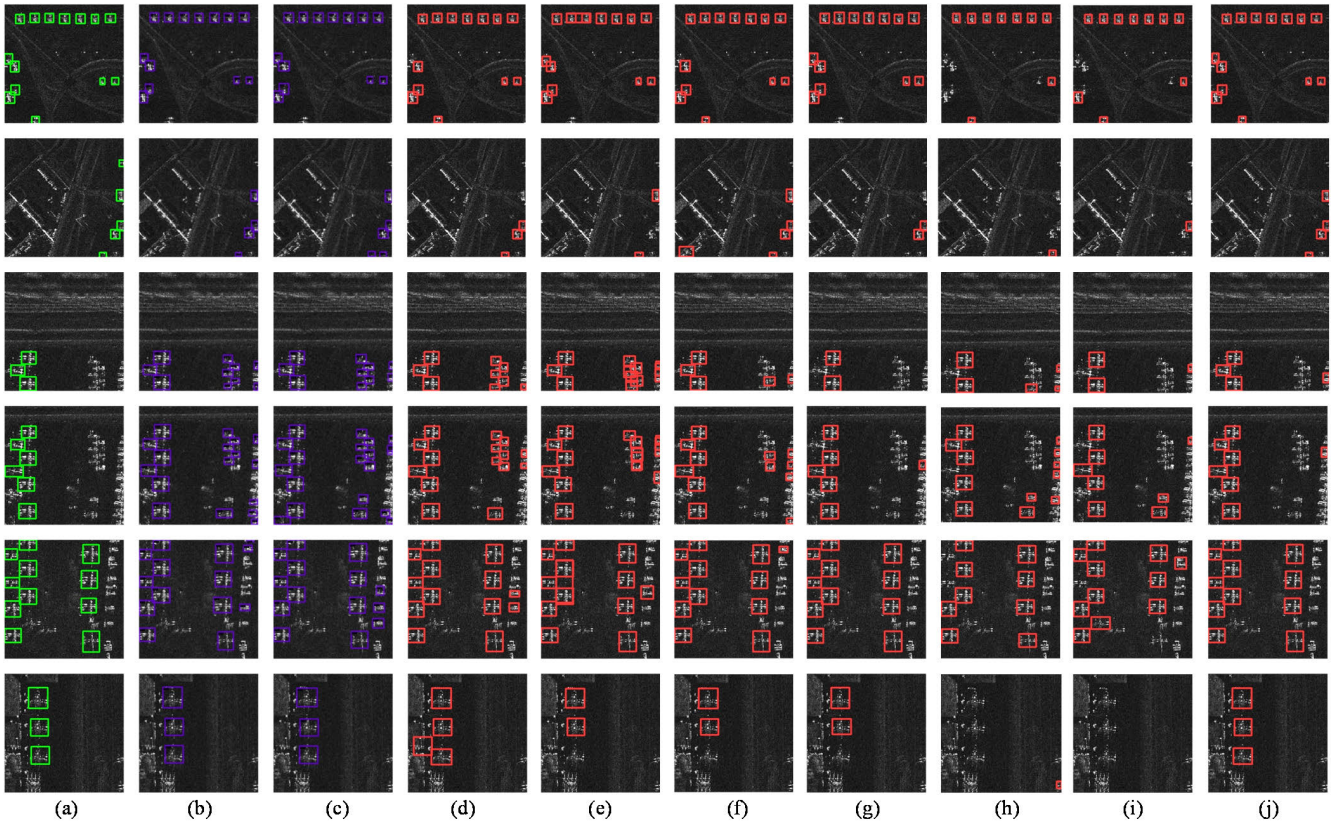

FIGURE 13. Comparison of actual test results of different models. (a) Ground Truth (b) Faster R-CNN (c) RetinaNet (d) YOLOv3-SPP (e) YOLOv3-tiny (f) YOLOv5s (g) TPH-YOLOv5 (h) YOLOv5s_CBAM_BiFPN (i) YOLOv5-L+BiFPN+SwinTransformer+GAM (j) EST-YOLOv5s.

F. COMPARATIVE EXPERIMENT

To prove the effectiveness of EST-YOLOv5s in SAR image aircraft target detection, a comparative experiment was carried out on the SADD dataset, and the proposed EST-YOLOv5s model was compared with several other target detection methods. Table 5 shows the performance comparison results between different models. Among them, Faster R-CNN, RetinaNet, YOLOv3-SPP, YOLOv3-tiny, and YOLOv5s are classic target detection methods, while the model YOLOv5s_CBAM_BiFPN and YOLOv5-L+BiFPN+SwinTransformer+GAM are specially designed for SAR ship and aircraft target detection.

In Table 5, compared with several other methods, EST-YOLOv5s has the highest mAP@50, which is 97.8%. Followed by YOLOv3-SPP, it realizes local fusion and global features by adding the SPP module based on YOLOv3. This addition is conducive to detecting situations with a significant difference in target sizes within the image. Although TPH-YOLOv5 performs well in solving small targets and multi-scale problems on UAV aerial photography datasets, its mAP@50 and F1 values on SAR datasets are low, only 81.0% and 81.2%. Moreover, the parameters of the network are relatively large, and the speed is relatively slow. Due to the complex network structure of Faster R-CNN and RetinaNet, a large number of parameters, and their anchor generation mechanism, their recall rate is generally high, but their precision is low, and it is easy to generate many wrong prediction frames. Compared with other networks, YOLOv3-tiny is more lightweight, with the least number of parameters, only 8.67 M, and a speed of 286 f/s, which is the fastest compared with other models. The Precision value and mAP@50 of the models YOLOv5s_CBAM_BiFPN and YOLOv5-L+BiFPN+SwinTransformer+GAM are relatively high, and they have certain applicability to SAR images. However, these models have lower recall values, and it is easy to miss the detection of aircraft targets. The parameter volume of EST-YOLOv5s is 16.27 M, the computing resources consumed are 73.3 GFLOPS, which is moderate compared to other networks, and the FPS is 65 f/s, which meets the requirements of real-time detection.

In order to verify that our model has anti-interference ability, this paper adds five different levels of salt and pepper noise to the dataset to simulate the interference of different degrees of complex environments on the model. Table 6 shows the mAP@50 of different models at different noise rates. Among them, mAP@50(5%), mAP@50(10%), mAP@50(15%), mAP@50(20%), and mAP@50(25%) represent the mAP@50 when the noise rate is 5%, 10%, 15%, 20%, and 25%, respectively. It is not difficult to conclude from Table 6 that as the noise level increases, the detection accuracy of all models gradually decreases with different trends. Among them, the model YOLOv5s_CBAM_BiFPN

is most similar to the downward trend of YOLOv5s. From the stability point of view, the curves of EST-YOLOv5s and TPH-YOLOv5s models decrease slowly in the whole range of salt-and-pepper-noise rates and remain relatively stable compared with other models.

The model EST-YOLOv5s maintains a high mAP value at different noise rates. As the noise rate increases, the detection accuracy gap between EST-YOLOv5s and other models grows, which means it has a stronger anti-interference ability. To sum up, the model can effectively filter out interference such as noise and strong scattering through reasonable feature representation and local information utilization and focus on detecting and identifying objects. At the same time, the model can better adapt to the changes in these disturbances and produce stable and accurate detection results in different situations.

G. VISUAL ANALYSIS

The detection results of the EST-YOLOv5s model and other traditional target detection models are displayed in Fig. 13 alongside the actual values of the dataset, where a green box represents the real value. Through the visualization of test results and the comparison with the real deal, you can see It is found that in a complex background environment, Faster R-CNN, RetinaNet, YOLOv3-SPP, YOLOv3-tiny, and YOLOv5s networks have produced more errors and redundant detection frames due to insufficient feature extraction. However, the multi-scale target detection capabilities of the models YOLOv5s_CBAM_BiFPN and YOLOv5-L+BiFPN+SwinTransformer+GAM are poor, and there are a large number of missed detections for medium-scale targets. At the same time, TPH-YOLOv5 is not easy to identify targets with fuzzy features, and some targets are not detected. Only EST-YOLOv5s successfully detected all the small targets for the first set of pictures. Based on the overall results, EST-YOLOv5s has the best overall performance, which significantly decreases the probability of missed detection and bounding box positioning errors in the SAR image detection process.

V. CONCLUSION

This paper proposes a new target detection model named EST-YOLOv5s to detect dense aircraft small targets and multi-scale aircraft targets in SAR images in complex environments. First, introducing a lightweight ECA mechanism can enhance the detailed features of the aircraft target while maintaining constant parameters. Then, the Swin Transformer is introduced in the last C3 module of the backbone network, and the global features are made more fully utilized. Finally, the TSCODE head replaces the traditional coupled detection head, enabling the model to have the ability to introduce task-specific contextual information while focusing on learning the respective specific features of classification and regression. These characteristics of the model also make it have stronger anti-interference ability and better robustness under complex conditions, but the model has the disadvantages of a complex model and long model training time.

In order to evaluate the ability of EST-YOLOv5s to detect aircraft targets in SAR images, this paper conducts comparative experiments on the SADD dataset. The experimental results show that the mAP@50 and F1 values of EST-YOLOv5s are higher than those of the original YOLOv5s, reaching 97.8% and 95.7%, respectively. In addition, this paper also conducts ablation experiments and anti-interference experiments to verify the effectiveness of the proposed module and the anti-interference performance of the model. EST-YOLOv5s can be widely used in essential fields such as military surveillance, natural disaster monitoring, and search and rescue. However, it is ineffective in automatic driving and video tracking. In the future, we will consider further developing the model in the direction of lightweight while ensuring the accuracy of the model.

REFERENCES

[1] B. van den Broek, E. den Breejen, R. Dekker, and A. Smith, "Change detection and maritime situation awareness in the channel area–feasibility of space borne SAR for maritime situation awareness," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 7436–7439, doi: 10.1109/IGARSS.2012.6351942.

[2] L. Cai, W. Shi, H. Zhang, and M. Hao, "Object-oriented change detection method based on adaptive multi-method combination for remote-sensing images," *Int. J. Remote Sens.*, vol. 37, no. 22, pp. 5457–5471, Nov. 2016, doi: 10.1080/01431161.2016.1232871.

[3] Z. Wang, N. Xu, J. Guo, C. Zhang, and B. Wang, "SCFNet: Semantic condition constraint guided feature aware network for aircraft detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5239420, doi: 10.1109/TGRS.2022.3224599.

[4] Y. Luo, H. Song, R. Wang, Y. Deng, F. Zhao, and Z. Xu, "Arc FMCW SAR and applications in ground monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5989–5998, Sep. 2014, doi: 10.1109/TGRS.2014.2325905.

[5] N. Liu, Z. Cui, Z. Cao, Y. Pi, and S. Dang, "Airport detection in large-scale SAR images via line segment grouping and saliency analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 434–438, Mar. 2018, doi: 10.1109/LGRS.2018.2792421.

[6] K. El-Darymli, P. McGuire, D. Power, and C. Moloney, "Target detection in synthetic aperture radar imagery: A state-of-the-art survey," *J. Appl. Remote Sens.*, vol. 7, no. 1, Mar. 2013, Art. no. 071598, doi: 10.1117/1.JRS.7.071598.

[7] G. Gao, L. Liu, L. Zhao, G. Shi, and G. Kuang, "An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 6, pp. 1685–1697, Jun. 2009, doi: 10.1109/TGRS.2008.2006504.

[8] W. Yu, Y. Wang, H. Liu, and J. He, "Superpixel-based CFAR target detection for high-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 730–734, May 2016, doi: 10.1109/LGRS.2016.2540809.

[9] W. An, C. Xie, and X. Yuan, "An improved iterative censoring scheme for CFAR ship detection with SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4585–4595, Aug. 2014, doi: 10.1109/TGRS.2013.2282820.

[10] C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain CFAR method for ship detection in HR SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 529–533, Apr. 2017, doi: 10.1109/LGRS.2017.2654450.

[11] J. Ai, Q. Luo, X. Yang, Z. Yin, and H. Xu, "Outliers-robust CFAR detector of Gaussian clutter based on the truncated-maximum-likelihood-estimator in SAR imagery," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2039–2049, May 2020, doi: 10.1109/TITS.2019.2911692.

[12] J. Karvonen, A. Gegiuc, T. Niskanen, A. Montonen, J. Buus-Hinkler, and E. Rinne, "Iceberg detection in dual-polarized C-band SAR imagery by segmentation and nonparametric CFAR (SnP-CFAR)," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4300812, doi: 10.1109/TGRS.2021.3070312.

[13] X. Hou, W. Ao, Q. Song, J. Lai, H. Wang, and F. Xu, "FUSAR-ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition," *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 36–54, Apr. 2020, doi: 10.1007/s11432-019-2772-5.

[14] W. Ao, F. Xu, Y. Li, and H. Wang, "Detection and discrimination of ship targets in complex background from spaceborne ALOS-2 SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 2, pp. 536–550, Feb. 2018, doi: 10.1109/JSTARS.2017.2787573.

[15] X. Leng, K. Ji, X. Xing, S. Zhou, and H. Zou, "Area ratio invariant feature group for ship detection in SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2376–2388, Jul. 2018, doi: 10.1109/JSTARS.2018.2820078.

[16] C. He, M. Tu, D. Xiong, F. Tu, and M. Liao, "Adaptive component selection-based discriminative model for object detection in high-resolution SAR imagery," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 2, p. 72, Feb. 2018, doi: 10.3390/ijgi7020072.

[17] C. He, S. Li, Z. Liao, and M. Liao, "Texture classification of PolSAR data based on sparse coding of wavelet polarization textons," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 8, pp. 4576–4590, Aug. 2013, doi: 10.1109/TGRS.2012.2236338.

[18] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of image degradation and degradation removal to CNN-based image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1239–1253, Apr. 2021, doi: 10.1109/TPAMI.2019.2950923.

[19] H.-Y. Han, Y.-C. Chen, P.-Y. Hsiao, and L.-C. Fu, "Using channel-wise attention for deep CNN based real-time semantic segmentation with class-aware edge information," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1041–1051, Feb. 2021, doi: 10.1109/TITS.2019.2962094.

[20] 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, doi: 10.1109/ICCV.2019.00972.

[21] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019, doi: 10.1109/TGRS.2019.2923988.

[22] S. Wei, H. Su, J. Ming, C. Wang, M. Yan, D. Kumar, J. Shi, and X. Zhang, "Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet," *Remote Sens.*, vol. 12, no. 1, p. 167, Jan. 2020, doi: 10.3390/rs12010167.

[23] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019, doi: 10.1109/LGRS.2018.2882551.

[24] 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, doi: 10.1109/CVPR.2014.81.

[25] 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.

[26] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[27] 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, doi: 10.1109/TPAMI.2016.2577031.

[28] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. 30th Int. Conf. Neural Inf. Syst. (NIPS)*, Barcelona, Spain, Dec. 2016, pp. 379–387, doi: 10.5555/3157096.3157139.

[29] 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, doi: 10.1109/CVPR.2017.106.

[30] 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. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

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

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[34] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[35] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580, doi: 10.1109/CVPRW50498.2020.00203.

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

[37] G. Jocher, "YOLOv5 by ultralytics (version 7.0) [computer software]," Tech. Rep., 2020. [Online]. Available: https://zenodo.org/record/7347926/export/hx, doi: 10.5281/zenodo.7347926.

[38] F. Dou, W. Diao, X. Sun, Y. Zhang, and K. Fu, "Aircraft reconstruction in high-resolution SAR images using deep shape prior," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 11, p. 330, Oct. 2017, doi: 10.3390/ijgi6110330.

[39] C. He, M. Tu, D. Xiong, F. Tu, and M. Liao, "A component-based multi-layer parallel network for airplane detection in SAR imagery," *Remote Sens.*, vol. 10, no. 7, p. 1016, Jun. 2018, doi: 10.3390/rs10071016.

[40] Q. An, Z. Pan, L. Liu, and H. You, "DRBox-v2: An improved detector with rotatable boxes for target detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8333–8349, Nov. 2019, doi: 10.1109/TGRS.2019.2920534.

[41] T. Zhang, X. Zhang, J. Shi, and S. Wei, "Depthwise separable convolution neural network for high-speed SAR ship detection," *Remote Sens.*, vol. 11, no. 21, p. 2483, Oct. 2019, doi: 10.3390/rs11212483.

[42] L. Zhang, C. Li, L. Zhao, B. Xiong, S. Quan, and G. Kuang, "A cascaded three-look network for aircraft detection in SAR images," *Remote Sens. Lett.*, vol. 11, no. 1, pp. 57–65, Jan. 2020, doi: 10.1080/2150704X.2019.1681599.

[43] Y. Zhao, L. Zhao, C. Li, and G. Kuang, "Pyramid attention dilated network for aircraft detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 662–666, Apr. 2021, doi: 10.1109/LGRS.2020.2981255.

[44] Q. Guo, H. Wang, and F. Xu, "Scattering enhanced attention pyramid network for aircraft detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7570–7587, Sep. 2021, doi: 10.1109/TGRS.2020.3027762.

[45] Y. Guo, S. Chen, R. Zhan, W. Wang, and J. Zhang, "SAR ship detection based on YOLOv5 using CBAM and BiFPN," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 2147–2150, doi: 10.1109/IGARSS46834.2022.9884180.

[46] J. Ge, C. Wang, B. Zhang, C. Xu, and X. Wen, "Azimuth-sensitive object detection of high-resolution SAR images in complex scenes by using a spatial orientation attention enhancement network," *Remote Sens.*, vol. 14, no. 9, p. 2198, May 2022, doi: 10.3390/rs14092198.

[47] X. Xu, X. Zhang, T. Zhang, Z. Yang, J. Shi, and X. Zhan, "Shadow-background-noise 3D spatial decomposition using sparse low-rank Gaussian properties for video-SAR moving target shadow enhancement," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3223514.

[48] X. Xu, X. Zhang, Z. Shao, J. Shi, S. Wei, T. Zhang, and T. Zeng, "A group-wise feature enhancement-and-fusion network with dual-polarization feature enrichment for SAR ship detection," *Remote Sens.*, vol. 14, no. 20, p. 5276, Oct. 2022, doi: 10.3390/rs14205276.

[49] J. Li, W. Zhu, Y. Yang, L. Qiu, and B. Zhu, "Detection of aircraft targets in SAR images based on improved YOLOv5," *Electron. Opt. Contr.*, vol. 30, no. 8, pp. 61–67, Aug. 2023, doi: 10.3969/j.issn.1671-637X.2023.08.011.

[50] X. Xu, X. Zhang, and T. Zhang, "Lite-YOLOv5: A lightweight deep learning detector for on-board ship detection in large-scene Sentinel-1 SAR images," *Remote Sens.*, vol. 14, no. 4, p. 1018. Feb. 2022, doi: 10.3390/rs14041018.

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[52] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[53] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.
[54] A. Vaswani, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010, doi: 10.5555/3295222.3295349.
[55] 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*.
[56] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.
[57] 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, doi: 10.1109/ICCV48922.2021.00986.
[58] 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*.
[59] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886, doi: 10.1109/CVPR46437.2021.00681.
[60] D. A. Hudson and C. L. Zitnick, "Generative adversarial transformers," 2021, *arXiv:2103.01209*.
[61] Y. Liu, G. He, Z. Wang, W. Li, and H. Huang, "NRT-YOLO: Improved YOLOv5 based on nested residual transformer for tiny remote sensing object detection," *Sensors*, vol. 22, no. 13, p. 4953, Jun. 2022, doi: 10.3390/s22134953.
[62] 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 (ICCVW)*, Oct. 2021, pp. 2778–2788, doi: 10.1109/ICCVW54120.2021.00312.
[63] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
[64] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11560–11569, doi: 10.1109/CVPR42600.2020.01158.
[65] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, and Y. Fu, "Rethinking classification and localization for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10183–10192, doi: 10.1109/CVPR42600.2020.01020.
[66] J. Zhuang, Z. Qin, H. Yu, and X. Chen, "Task-specific context decoupling for object detection," 2023, *arXiv:2303.01047*.
[67] P. Zhang, H. Xu, T. Tian, P. Gao, L. Li, T. Zhao, N. Zhang, and J. Tian, "SEFEPNet: Scale expansion and feature enhancement pyramid network for SAR aircraft detection with small sample dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3365–3375, 2022, doi: 10.1109/JSTARS.2022.3169339.

MIN HUANG received the B.Eng. degree in electrical information engineering from the Hebei University of Science and Technology, Shijiazhuang, Hebei, China, in 2002, and the M.Eng. degree in software engineering from the Beijing Institute of Technology, Beijing, China, in 2004. He is currently pursuing the Ph.D. degree in electrical engineering with the National Key Laboratory on Electromagnetic Environment Effects, Shijiazhuang. He is an Associate Professor with the Hebei University of Science and Technology. His research interests include machine learning, natural language processing, big data processing, and artificial intelligence.

WEIHAO YAN is currently pursuing the master's degree with the Hebei University of Science and Technology. His research interests include image processing and machine vision.

WENHUI DAI is currently pursuing the master's degree with the Hebei University of Science and Technology. His research interests include computer vision and deep learning.

JINGYANG WANG received the B.Eng. degree in computer software from Lanzhou University, China, in 1995, and the M.Sc. degree in software engineering from the Beijing University of Technology, China, in 2007. He is currently a Professor with the Hebei University of Science and Technology, Shijiazhuang, Hebei, China. His research interests include machine learning, deep learning, natural language processing, and big data processing.