TYPE Original Research
PUBLISHED 02 June 2023
DOI 10.3389/fnbot.2023.1170163

OPEN ACCESS

EDITED BY
Feihu Zhang,
Northwestern Polytechnical University, China

REVIEWED BY
Zhaoying Liu,
Beijing University of Technology, China
Yang Yang,
Yunnan Normal University, China
Wen Xin,
Northwestern Polytechnical University, China

*CORRESPONDENCE
Ruitao Lu
✉ lrt19880220@163.com

RECEIVED 20 February 2023
ACCEPTED 18 May 2023
PUBLISHED 02 June 2023

CITATION
Zhao K, Lu R, Wang S, Yang X, Li Q and Fan J (2023) ST-YOLOA: a Swin-transformer-based YOLO model with an attention mechanism for SAR ship detection under complex background. *Front. Neurorobot.* 17:1170163. doi: 10.3389/fnbot.2023.1170163

ST-YOLOA: a Swin-transformer-based YOLO model with an attention mechanism for SAR ship detection under complex background

Kai Zhao, Ruitao Lu*, Siyu Wang, Xiaogang Yang, Qingge Li and Jiwei Fan

Department of Automation, Rocket Force University of Engineering, Xi'an, China

A synthetic aperture radar (SAR) image is crucial for ship detection in computer vision. Due to the background clutter, pose variations, and scale changes, it is a challenge to construct a SAR ship detection model with low false-alarm rates and high accuracy. Therefore, this paper proposes a novel SAR ship detection model called ST-YOLOA. First, the Swin Transformer network architecture and coordinate attention (CA) model are embedded in the STCNet backbone network to enhance the feature extraction performance and capture global information. Second, we used the PANet path aggregation network with a residual structure to construct the feature pyramid to increase global feature extraction capability. Next, to cope with the local interference and semantic information loss problems, a novel up/down-sampling method is proposed. Finally, the decoupled detection head is used to achieve the predicted output of the target position and the boundary box to improve convergence speed and detection accuracy. To demonstrate the efficiency of the proposed method, we have constructed three SAR ship detection datasets: a norm test set (NTS), a complex test set (CTS), and a merged test set (MTS). The experimental results show that our ST-YOLOA achieved an accuracy of 97.37%, 75.69%, and 88.50% on the three datasets, respectively, superior to the effects of other state-of-the-art methods. Our ST-YOLOA performs favorably in complex scenarios, and the accuracy is 4.83% higher than YOLOX on the CTS. Moreover, ST-YOLOA achieves real-time detection with a speed of 21.4 FPS.

KEYWORDS

synthetic aperture radar (SAR) image, ship detection, Swin Transformer, YOLO, attention mechanism

1. Introduction

SAR imaging has a high resolution, a long detection range, and a strong anti-interference ability. It has a wide range of applications and development possibilities and is employed extensively in both civil and military fields, including surveying, mapping, catastrophe monitoring, marine observation, and military reconnaissance (Cumming and Wong, 2005; Moreira et al., 2013). Research on ship detection technology, such as precise terminal guidance, operational effectiveness assessment, and identification of ship targets in strategic port areas, is crucial for both military and civilian applications. Although significant progress has been made in the past few decades, ship detection in SAR images remains a challenge due to shape deformation, pose variation, and background clutter.

Traditional ship detection methods employ machine learning models to distinguish the ship target from the background in SAR images. These methods usually include two primary processes: object detection and target identification (Lu et al., 2020a,b). The most popular models employed in the classic detection approach are the constant false-alarm rate (CFAR) (Robey et al., 1992) and its variant algorithms. These methods detect ships by creating a statistical distribution model of the background clutter. By employing a decomposition strategy, Gao et al. (2018) suggested a CFAR algorithm based on a generalized gamma distribution to enhance the signal-to-noise ratio of SAR images. Wang et al. (2017) pro-posed a constant false-alarm detector in the intensity space domain, which uses data correlation to detect targets and wake pixels. To best suit the information provided by the ship distribution map (Schwegmann et al., 2015) proposed a method for transforming a scalar threshold into a threshold manifold using the simulated annealing (SA) process. However, the traditional SAR ship detection model relies on artificial design feature selection, which leads to poor detection robustness and generalization ability. In addition, this kind of algorithm requires a high contrast between the target image and the background image and is not suitable for detecting ship targets in complex environments.

With the development of deep learning theory, convolutional neural networks have made significant advancements in the field of target recognition and display advanced performance in target detection. The deep learning detection algorithms can be roughly classified into two categories: one-stage methods and two-stage methods (Girshick et al., 2014). Two-stage object detection methods perform region generation to obtain pre-selected boxes and then use sample classification and regression with border positioning through the convolutional neural network. Representative methods include R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), and Faster R-CNN (Ren et al., 2015). These methods have high detection accuracy but low detection efficiency. One-stage object detection methods use the backbone feature extraction network to directly locate and classify the target. Typical detection methods are YOLO (Redmon et al., 2016), SSD (Liu et al., 2016), and CenterNet (Duan et al., 2019). Although these methods are fast, they are prone to false detection and missing detection compared with the two-stage detection methods.

Based on the current deep-learning-based target detection, in recent years, numerous researchers have developed advanced algorithms for SAR ship detection. To address the challenges of ship detection in maritime environments, Gao et al. (2022) proposed an enhanced YOLOv5 SAR ship detection method based on target augmentation. By constructing the feature enhancement Swin converter (FESwin) module and the adjacent feature fusion (AFF) module (Li et al., 2022), proposed a detection model suitable for the strong scattering, multi-scale, and complicated background of ship objects in SAR images. To provide a visual converter system based on context-federated representation learning appropriate for SAR ship detection (Xia et al., 2022), creatively combined CNNs with transformers. Although the aforementioned methods address the issues of multi-scale targets, huge noise clutter, and complicated backgrounds, the detection accuracy and the calculation speed still restrict application in the real world. Small islands and nearby sea structures are the reasons for false detection in com-plex backgrounds. In addition, the dense distribution of ships in the dock and the sea causes multiple targets to overlap, leading to the low accuracy of models in detecting targets.

Based on the aforementioned analysis, in this paper, we have proposed a novel ship detection model called ST-YOLOA, which is more suitable for the actual complex environment in SAR images. We chose Swin Transformer (Liu et al., 2021) and YOLOX (Ge et al., 2021) as our basic models. The main contributions of this paper are as follows: (1) Together data on ship target features, we have proposed the STCNet backbone network. This network effectively solves the problem of insufficient feature extraction caused by strong scattering in SAR images. It enhances the processing ability of feature information by obtaining more significant feature information in different environments. It also has excellent global information modeling capabilities of Swin Transformer. (2) We have built a novel feature pyramid network based on an enhanced PANet for profoundly fusing high-level and low-level features. This network solves the issues of local information interference and attention diversion by using semantic and localization information. To improve the detection accuracy, we have adopted binary trilinear interpolation up-sampling for maintaining the original data of the feature map. (3) To effectively reduce the impact of noise in the feature map on the detection accuracy, classification and regression are handled separately. We have used EIOU as the localization loss function to cope with the sample imbalance and enhance the model generalization ability.

The rest of the paper is organized as follows. In Section 2, we review the prior work related to the proposed ST-YOLOA. In Section 3, we provide details of the main components and methodology of the proposed ST-YOLOA. Section 4 introduces the experimental settings, results, and analysis. The conclusions of the paper are drawn in Section 6.

2. Related work

In this section, we briefly review the relevant literature regarding YOLO, Transformer, and the attention mechanism.

2.1. YOLO

The YOLO series is a typical network for one-stage detection. It handles object detection as a regression issue, where the bounding box coordinates and class probability are derived from picture pixels. YOLOX, as a typical representative of the YOLO series, has significant advantages in terms of speed and accuracy. Its essential modules include the Focus, the CSP bottleneck, SPP, PANet (Liu et al., 2018), and the decoupled detection head (Tian et al., 2019). The backbone of YOLOX is the CSP Darknet-53 (Bochkovskiy et al., 2020), which consists of several residual modules stacked one on top of the other. YOLOX uses PANet as the neck of the model, and the input is the three feature output layers output by the backbone network. It obtains features with richer semantic and localization information by feature fusion and sends them to the head for detection. At the head, YOLOX replaces the coupled detection head with the decoupled detection head using different branches for the classification and regression tasks, which significantly increases the convergence speed of the network. YOLOX eliminates the constraints of the original anchor (Zhang et al., 2020) of the YOLO series. The anchor-free mechanism substantially reduces the number of design parameters, maintaining effectiveness while significantly reducing time costs.

When the YOLO architecture is used for ship detection, it mainly has the following two disadvantages: (1) It has poor recognition where small target objects are concerned, and the positioning is inaccurate. (2) It lacks the ability to obtain global information on the image that can benefit the network in terms of accuracy and efficiency.

2.2. Transformer

Transformer (Vaswani et al., 2017) was initially applied in the field of natural language processing (NLP) and has proved to have many advantages. Transformer is not only powerful in modeling global contexts, but also excellent in establishing long-distance dependencies. With its rapid development in the field of NLP, Transformer has attracted widespread attention in the field of the computer vision field. Swin Transformer (ST) is considered the first successful attempt to bring it into computer vision. It enables the Transformer model to process images at different scales flexibly by applying a hierarchical structure similar to that of CNN. ST performs local self-attention calculations in the area of non-overlapping windows. It lowers the computational complexity of the number from a squared relationship to a linear relationship. Then it uses shifted window multi-head self-attention (SW-MSA) to achieve information interaction between non-overlapping windows.

As a general visual network, ST exhibits state-of-the-art performance in semantic segmentation, object detection, and image classification. However, ST has two clear drawbacks: (1) ST has a limited ability to encode contextual information and needs further improvement. (2) Because ST has more parameters than CNN, its training usually relies on a large number of training data.

2.3. Attention mechanism

Usually, attention mechanisms in the vision domain (Guo M. et al., 2022) include two types: spatial and channel. They extract better target features by assigning different weights to the feature points on the image. The spatial attention mechanism adds weights to the feature points containing object features in a single-channel feature map. On the other hand, the channel attention mechanism assigns more importance to feature channels containing component semantic information. Hu et al. (2018) proposed SENet, which analyzes the correlation between different feature channels and generates channel descriptions by fusing features across spatial dimensions, thus achieving selective emphasis on feature information and suppressing irrelevant feature information; Woo et al. integrated the feature channels and feature space between correlation proposed CBAM (Wang et al., 2018), which can focus on more profound feature semantic information; Wang et al. proposed CANet (Hou et al., 2021), which considers inter-channel relationships as well as location information over long distances, based on the spatial selectivity of the channel attention mechanism. In this paper, the different characteristics of SE, CBAM, and CA, are introduced into other model modules to improve the performance model further.

3. Methods

In this section, we first give a general overview of the ST-YOLOA target detection model and then discuss in detail the design ideas and network architecture of the ST-YOLOA model in three parts: feature extraction (Backbone), feature fusion (Neck), and target detection (Head), respectively. Figure 1 shows the ST-YOLOA network structure.

3.1. Overview

3.1.1. Backbone

In ST-YOLOA, we propose a backbone network called STCNet. It integrates the advantages of the Swin Transformer and the CA attention mechanism. Compared with the traditional CNN-based backbone feature extraction network, which only utilizes the information provided by regions in target localization, STCNet has good performance with dynamic attention and global modeling capability considering remote dependencies. The STCNet network adopts a layered architecture consisting of the Patch Embedding layer, the Swin Transformer Block, and the CA- Patch Merging layer composed of three parts.

3.1.2. Neck

In the neck of ST-YOLOA, we still use PANet to construct feature pyramids (Lin et al., 2017a) for feature depth fusion. In addition, we also introduce SE and CBMA attention mechanisms in the neck to enhance the focus on the target information and further improve the model performance.

3.1.3. Loss

The purpose of the loss function is mainly to make the model localization more accurate and recognition accuracy higher. Therefore, more advanced EIOU Loss is used in ST-YOLOA to accelerate the convergence and improve the model performance.

3.2. Backbone

3.2.1. Patch embedding layer

The patch embedding module first chunks the image at the front end of the feature extraction network, dividing the image into 4 × 4 non-overlapping blocks so that the feature dimension of each block is 4 × 4 × 3. Then, the original 2D image is converted into a series of 1D embedding vectors by projecting the feature dimensions to arbitrary dimensions through linear transformation, and the transformed embedding vectors are input

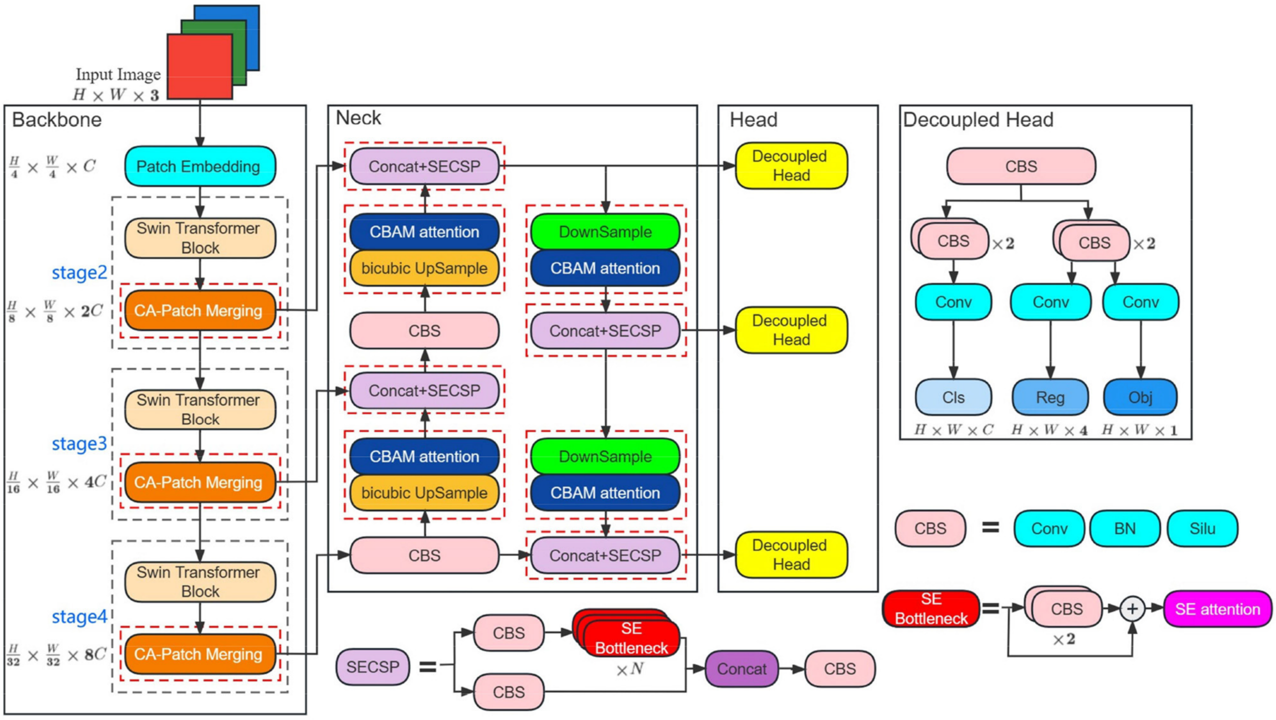

FIGURE 1
ST-YOLOA network structure. The red dashed boxes are the attention module addition locations. Conv, BN, and Silu denote the convolution, batch normalization, and SILU activation functions, respectively. Concat indicates the fully connected operation. Cls, reg, and obj represent the classification, localization, and confidence scores. *H*, *W*, and *C* denote the feature map's width, height, and number of channels.

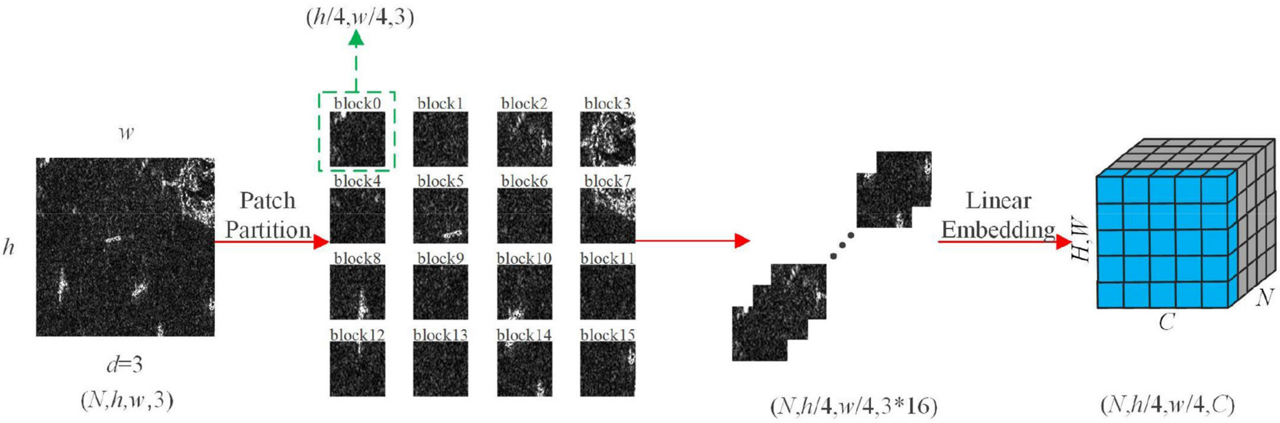

FIGURE 2
Patch Embedding structure. w and h are the length and width of the input feature map, d is the number of channel dimensions, and N is the batch size.

to three-stage feature extraction layers to generate a hierarchical feature representation. The structure is shown in Figure 2.

3.2.2. Swin transformer block

The Swin Transformer Block uses moving windows to calculate the attention between pixels, which helps to connect the front layer windows and reduce the complexity of the original attention calculation while overcoming the drawback of a lack of global effects, significantly enhancing the modeling effect.

In Figure 3, it can be seen that the multiheaded self-attention (MSA) mechanism in the Swin Transformer Blocks is constructed based on the shift window. There are two consecutive Swin Transformer Blocks. Each Swin Transformer Block consists of a LayerNorm (LN) layer, an MSA module, a residual connection, and a multilayer perceptron (MLP) that contains two fully connected layers using the GELU non-linear

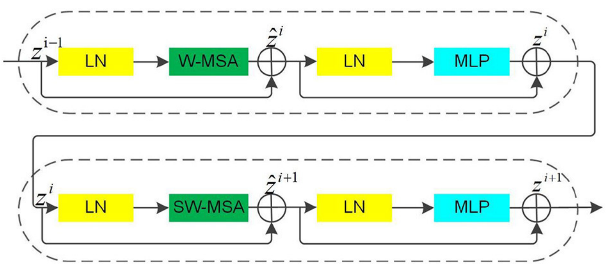

FIGURE 3
The Swin transformer blocks.

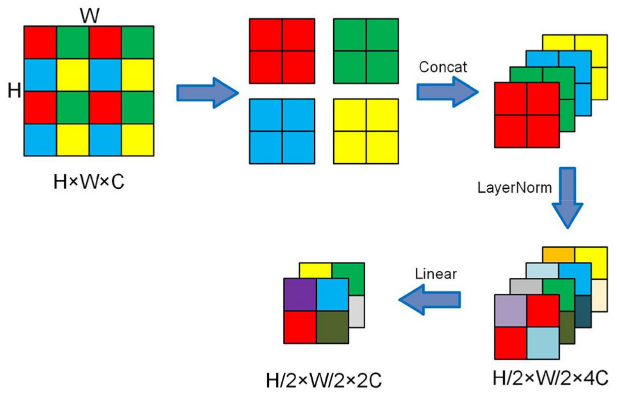

FIGURE 4
Schematic diagram of the Patch Merging layer.

activation function (Wang et al., 2021). The two consecutive Swin Transformer Blocks adopt the window multi-head self-attention (W-MSA) module and the shifted window multi-head self-attention (SW-MSA) module, respectively, which enables different windows to exchange information while reducing computational effort. Based on this window division mechanism, the continuous Swin Transformer Blocks are calculated as follows:

$$\hat{z}^i = W - MSA\left(LN\left(z^{i-1}\right)\right) + z^{i-1} \quad (1)$$

$$z^i = MLP\left(LN\left(\hat{z}^i\right)\right) + \hat{z}^i \quad (2)$$

$$\hat{z}^{i+1} = SW - MSA\left(LN\left(z^i\right)\right) + z^i \quad (3)$$

$$z^{i+1} = MLP\left(LN\left(\hat{z}^{i+1}\right)\right) + \hat{z}^{i+1} \quad (4)$$

Where $\hat{z}^i$ denotes the output of the (S)W-MSA module and z^i denotes the output of the MLP module of the i^{th} Block.

3.2.3. CA-patch merging

The Patch Merging layer is used to perform a down-sample operation before the feature output of the backbone network to reduce the feature map resolution and adjust the number of channels, thus forming a layered design and also saving some computational effort. Figure 4 presents the working process.

Considering the limited context encoding capability of the Swin Transformer, we add the CA attention mechanism after the Patch Merging layer. CA attention decomposes the channel attention work process into two one-dimensional feature encoding processes and then performs feature aggregation along two directions in space. Figure 5A illustrates the structure of the CA attention mechanism. It first pools the feature maps globally averaged in two dimensions, height, and width, using convolution kernels of dimensions (H, 1) and (1, W), respectively:

$$\begin{cases} z_c^h(h) = \frac{1}{W} \sum\limits_{0 \le i < W} x_c(h, i) \\ z_c^h(w) = \frac{1}{H} \sum\limits_{0 \le j < H} x_c(j, w) \end{cases} \quad (5)$$

The above transformations obtain the feature maps in the space in the width and height directions, respectively. Then CA attention performs the stitching operation on the feature maps and performs the F_1 transformation to obtain the feature map f. The formula is shown below:

$$f = \delta(F_1([z^h, z^w])) \quad (6)$$

Where F_1 is the 1×1 convolutional transform function, [,] denotes the splicing operation, and δ is the nonlinear activation function.

The feature map f is then convolved in the original height and width direction and activated by the Sigmoid activation function to obtain the feature map attention weights g^h and g^w, which are given by the following equations:

$$\begin{cases} g^h = \sigma(F_h(f^h)) \\ g^w = \sigma(F_w(f^w)) \end{cases} \quad (7)$$

Where σ is the sigmoid activation function.

Finally, the CA attention mechanism is calculated by multiplicative weighting to obtain the output of the feature map with attention weights:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (8)$$

It encodes long-range dependencies by precise positional information, enabling our model to utilize global contextual details efficiently. At the same time, CA has both channel and spatial domain attention mechanisms. Its introduction can better capture direction-aware and location-sensitive information for more accurate localization to identify objects of interest and improve feature representation.

3.3. Neck

3.3.1. Improved CSPLayer—SECSP

CSPLayer (Wang et al., 2020) is mainly divided into two parts (Figure 1), a backbone part, which consists of shallow

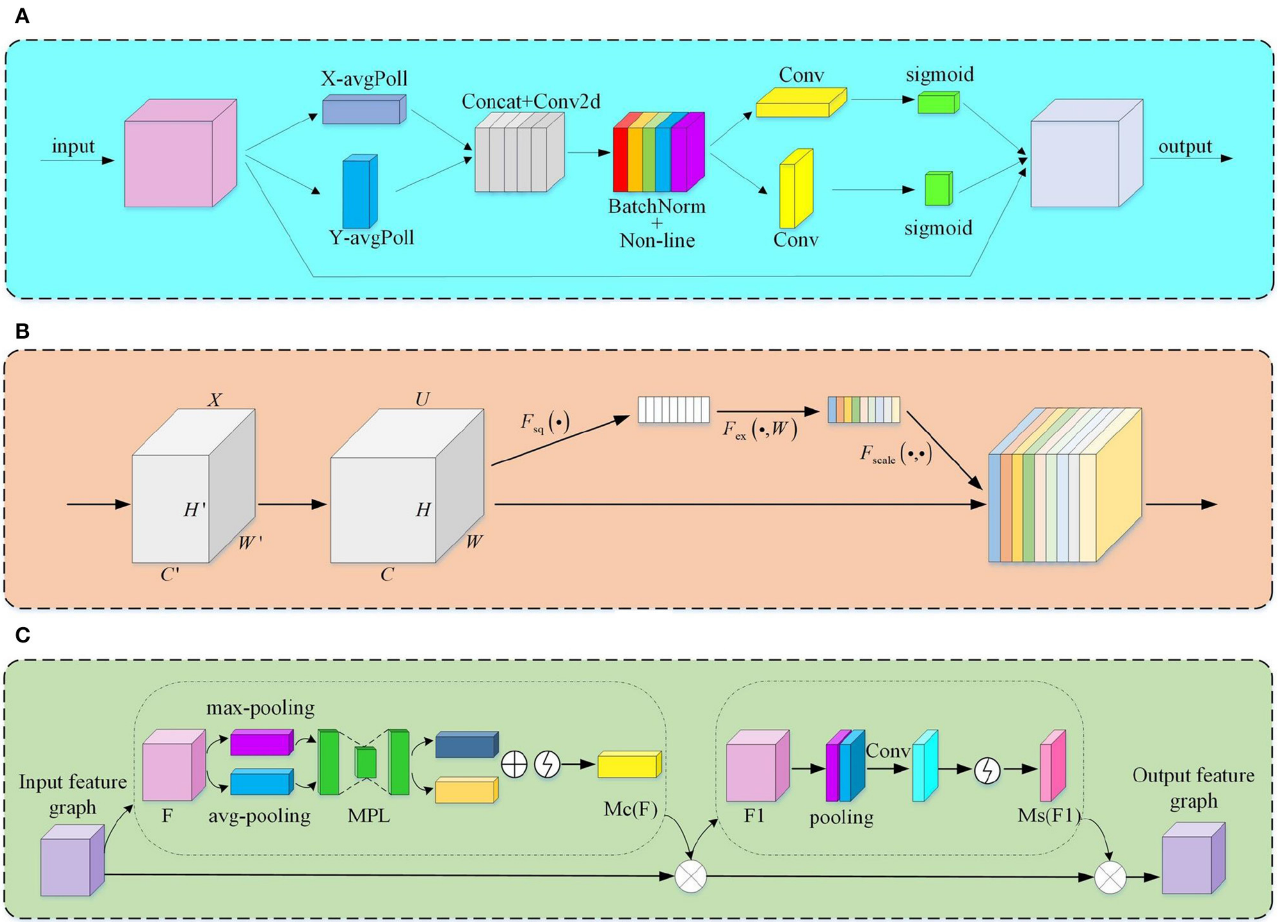

FIGURE 5
The module structure of the attention mechanism. **(A)** CA module; **(B)** SE module; **(C)** CA module.

convolutional branches and sub-residual branches, and a residual part, which is directly connected to the output part of the CSPLayer through a simply processed 1×1 convolutional layer. The sub-residual bottleneck structure is an essential component of the CSPNet. It has a 1×1 convolutional stacked layer and a 3×3 convolutional stacked layer. Additionally, shortcut connections are applied to directly add elements to the output of the convolutional layer. The feature extraction process of the CSP network is primarily carried out in the sub-residual bottleneck structure, and its application significantly alleviates the gradient disappearance problem.

The use of CSPLayer makes the model over-consider the surrounding contextual information (Liu et al., 2022), which causes local information interference. To solve this problem, we introduce the SE attention mechanism in the Bottleneck module to selectively emphasize the feature information, weaken the interference information, and further enhance the focus on the target features. Meanwhile, the 3 × 3 convolutional layer in Bottleneck needs to deal with a large number of parameter operations while causing a large number of parameter redundancies. SE performs feature compression on the feature map down the spatial dimension, squeezing the global spatial information into the channel description. The output feature map z_c of channel $\boldsymbol{c}$ after compression is:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{9}$$

Where x_c is the input, H and W represent the two directions of height and width in space, respectively. This process dramatically reduces the redundant parameters in the network. Figure 5B illustrates the structure of SE.

3.3.2. Improved up-sampling and down-sampling processes

The resolution of the feature maps at various sizes varies. Before feature fusion, down-sampling or up-sampling operations must shrink or enlarge the feature maps for feature fusion between feature maps of different scales. The process of compression and extension of feature maps brings about the problem of semantic information loss and the introduction of local interference. CBAM (as shown in Figure 5C) can focus on more profound feature semantic information by performing a hybrid pooling of both global average and global maximum over space and channels. Its introduction makes the model more robust. The specific working process is as follows:

$$\begin{cases} \mathrm{M_C(F)} = \sigma(\mathit{MLP}(\mathit{AvgPool}(\mathrm{F})) + \mathit{MLP}(\mathit{MaxPool}(\mathrm{F}))) \\ = \sigma(\mathrm{W_1}(\mathrm{W_0}(\mathrm{F^c_{avg}})) + \mathrm{W_1}(\mathrm{W_0}(\mathrm{F^c_{max}}))) \\ \mathrm{M_S(F)} = \sigma(f^{7\times7}([\mathit{AvgPool}(\mathrm{F}); \mathit{MaxPool}(\mathrm{F})])) \\ = \sigma(f^{7\times7}([\mathrm{F^s_{avg}};\mathrm{F^s_{max}}])) \end{cases} \tag{10}$$

Where $\mathrm{M_C(F)}$ and $\mathrm{M_S(F)}$ are one-dimensional and two-dimensional channel attention, respectively; σ is a sigmoid function; $W_0 \in R^{C/r\times C}$, $W_1 \in R^{C\times C/r}$; and $f^{7\times7}$ is a convolution kernel of 7×7 size.

The deconvolution up-sampling approximation is considered the inverse operation of convolution. It can restore the feature map better by introducing training parameters for learning. However, this up-sampling method is prone to a tessellation effect, which causes pixel blocks to appear in the image. On the other hand, the interpolation method does not require any parameter learning. It performs predictive estimation of unknown points based on known pixel points, which can expand the size of the image and achieve the effect of up-sampling. Therefore, we use the bicubic interpolation algorithm for up-sampling instead of deconvolution, which reduces many parameter operations while preserving the original image information. Figure 6 illustrates the improved up-and-down sampling process.

3.4. Head

Considering the fact that SAR images of ships in complex environments require a lot of feature information to identify targets, the commonly used coupled detection head will influence the model's performance and cannot detect the ship targets in SAR images of a complex environment. Therefore, the ST-YOLOX network model separates the classification and regression tasks by using a decoupled detection head for target detection to achieve the predicted output of target location and bounding box, which significantly increases the convergence speed and improves the accuracy of the model.

As for the loss function, SAR image ship detection is a single-class detection task. Hence, the loss function has only two components: localization loss (Reg) and confidence loss (Obj) (Jiang et al., 2022). The mathematical equations for these two components are as follows:

$$Loss = \frac{\lambda L_{reg} + L_{obj}}{N_{pos}} \tag{11}$$

where λ is the balance coefficient; N_{pos} represents the Anchor Points quantity of positive samples; L_{obj} indicates the confidence loss; in our paper, the binary cross-entropy loss (BCE loss) is used as this loss function to promote numerical stability; L_{reg} represents the localization loss, and the Efficient-IOU (EIOU) (Zhang et al., 2022) loss function is used. Its mathematical expression is:

$$L = L_{IOU} + L_{dis} + L_{asp} = 1 - IOU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \frac{\rho^2\left(w, w^{gt}\right)}{c_\omega^2} + \frac{\rho^2\left(h, h^{gt}\right)}{c_h^2} \tag{12}$$

where L_{IOU} is the overlap loss, L_{dis} is the center distance loss, and L_{asp} is the wide height loss. EIOU loss integrates the overlapping area, the distance to the center point, and the aspect ratio of the bounding box regression. It splits the loss term of the aspect ratio into the difference between the widths and the heights of the predicted and the minimum outer bounding boxes, which effectively solves the sample imbalance problem in the bounding box regression task, accelerates the convergence, improves the regression accuracy, and further optimizes the model.

With the aforementioned analysis, the ST-YOLOA detection model proposed in this paper applies to the detection of ship objects in SAR images under complex environments. It has the advantages of strong feature extraction capability, high utilization of high-level and low-level feature information, full information fusion, and robust performance, which is more suitable for ship target detection under realistic conditions.

4. Experiment

4.1. Experimental data and environment

4.1.1. Dataset

In the paper, the experimental data are based on the publicly accessible SAR-Ship-Dataset (Wang et al., 2018) from the Key Laboratory of Digital Earth, Institute of Space and Astronomical Information, Chinese Academy of Sciences. The primary data sources of this dataset are Sentinel-1 SAR data and domestic Gaofen-3 SAR data, which use three polarization techniques: single-polarization, double-polarization, and full-polarization. It used 108-view Sentinel-1 and 102-view Gaofen-3 high-resolution SAR images to build a SAR ship target deep learning sample library containing 43,819 images of 256 × 256 pixels with 59,535 ship targets in total. The dataset contains a wide variety of ship types and backgrounds, including sea-surface scenes with noise interference from the ocean and ships of different scales and nearshore scenes influenced by complex backgrounds, such as islands, land constructions, and port terminals.

We used 4,000 photographs from the SAR-Ship-Dataset as the dataset for our experiment. The training set and the test set were randomly divided according to the ratio of 8:2, and 20% of the training set was randomly selected as the validation set. To increase the data diversity and ensure the model had a better training effect, we used two data enhancement methods, Mosaic (Tian et al., 2019) and Mixup (Zhang et al., 2017), to perform data enhancement operations on the dataset.

To test the ship detection capability of this model in complex environments, we selected 450 SAR ship images in complex environments, such as near-coastal ship targets affected by surrounding non-ship targets, ship targets with blurred or obscured imaging, ship targets with coherent speckle noise and complex background information, and multi-scale ship targets. We named the two SAR ship detection test sets constructed as norm test set

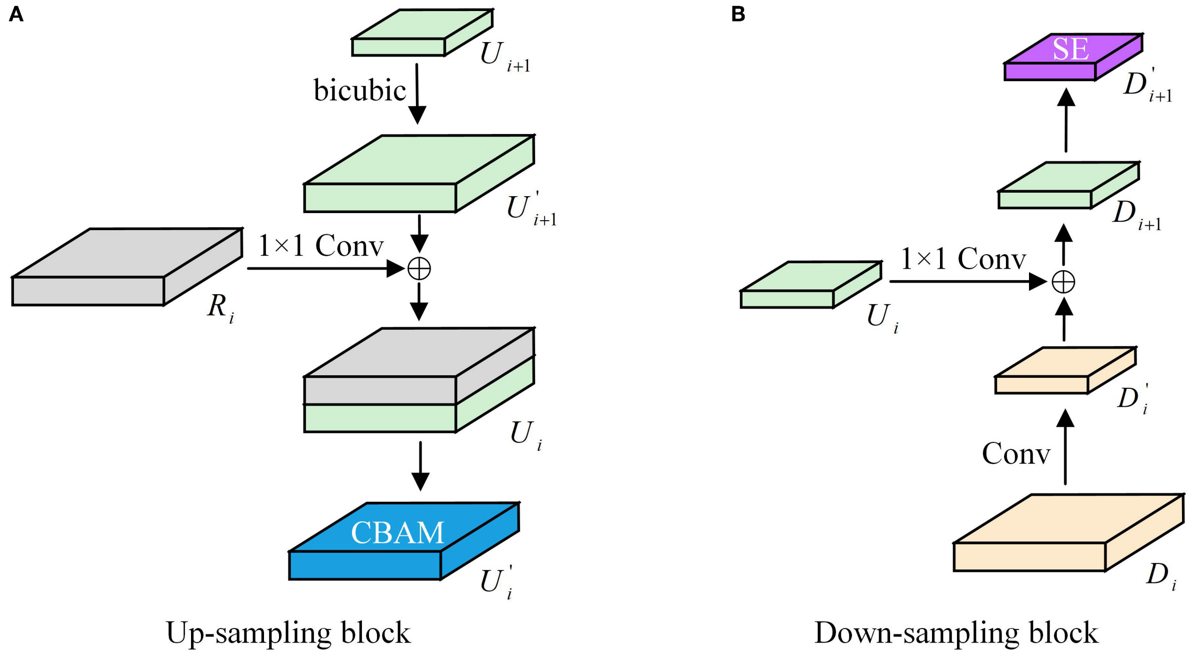

FIGURE 6
(A, B) Up/down-sampling structure diagram.

TABLE 1 Experimental data set details.

Dataset		Number of images	Target background	Number of ships
Train		2,560	General background	3,440
Val		640	General background	843
Test	NTS	800	General background	1,020
	CTS	450	Complex background	943
	MTS	1,250	General Background + Complex background	1,963

(NTS) and complex test set (CTS), respectively, and combined the two sets into one merged test set (MTS). We used the combined performance of the model in these three test sets as the criterion to verify its detection ability. The details of the ship data set used for the experiments are shown in Table 1.

4.1.2. Evaluation indicators

This paper chooses the average precision (AP) (Everingham et al., 2009) as the main evaluation index to assess the effect of SAR image ship detection. It contains two parameter metrics, Precision and Recall. The calculation formula is:

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \\ AP = \int_0^1 P\,(r)dr \end{cases} \tag{13}$$

where *TP* (true positive) is the number of ships marked as ship targets, *FN* (false negative) is the number of ship targets marked as non-ships, *FP* (false positive) is the number of non-ships marked as ship targets, and *P(r)* is the area under the PR curve with precision and recall, which is AP.

Also, to better measure, the comprehensive performance of the model, Parameters, GFLOPs, and FPS are introduced as evaluation metrics in this paper.

4.1.3. Experimental environment and parameter setting

In this paper, the experimental environment was based on Linux system architecture, using the Ubuntu 18.04 operating system, equipped with an Intel(R) Core i9 10980 XE CPU and NVIDIA RTX 2080TI graphics card with 11 GB video memory. The deep learning framework used PyTorch, with accelerated training via CUDA 10.1 and cuDNN 7.6.

In this paper, the experimental hyperparameters are referred to the literature (An et al., 2019; Yuan and Zhang, 2021; Wu et al., 2022), and the main settings are as follows: setting the training period to 300 epochs, the maximum learning rate of the model to 0.01, and the minimum learning rate to 0.0001. The optimizer was stochastic gradient descent (SGD), and the weights decayed

TABLE 2 The ablation experiments results.

Serial number	Swin-T	Attention	EIOU loss	AP/%			FPS	GFLOPs	Parameters
				NTS	CTS	MTS			
1	-	-	-	96.23	70.86	85.52	**22.74**	**27.27**	**8.94**
2	√	-	-	96.30	73.81	86.72	22.24	89.60	32.72
3	-	√	-	96.94	71.64	86.37	20.24	27.28	9.00
4	-	-	√	97.23	73.61	88.27	*22.38*	**27.27**	**8.94**
5	√	√	-	96.71	74.19	87.74	21.60	89.61	32.77
6	√	-	√	97.28	*74.40*	88.16	22.00	89.60	32.72
7	-	√	√	**97.81**	72.40	86.93	20.49	27.28	9.00
8	√	√	√	*97.37*	**75.69**	**88.50**	21.40	89.61	32.77

Bold indicates the best result of each column, italic is the second best result; "-" is no module added, "√" is the module added.

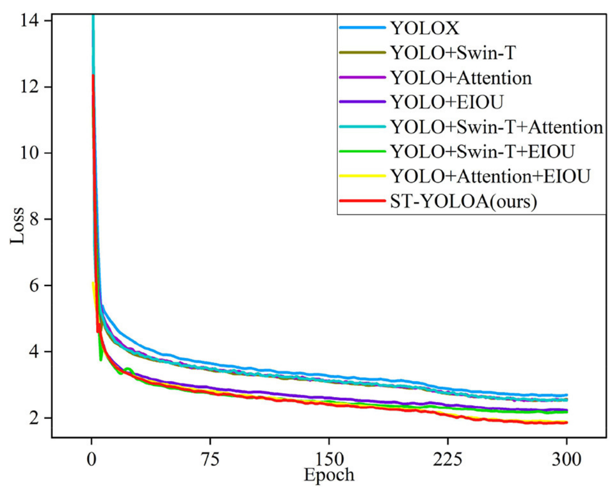

FIGURE 7
Loss function curves of the ablation experiments.

at a rate of 0.0005. To increase the speed of data reading, we employed multi-threaded data reading and used cosine (COS) as the learning rate descent method. The test was run with the following parameters: a non-maximum suppression threshold of 0.65, a confidence level of 0.001, and a prediction probability threshold of 0.5.

4.2. Ablation experiments

To validate the performance of each major module and loss function in the ST-YOLOA model, we performed ablation experiments. In these experiments, as the benchmark model, we used the YOLOX network, which through eight groups of networks with various structures were used to test the effects of various strategies on the detection effectiveness of the model. These strategies included changing the Swin Transformer backbone network; adding CA, SE, and CBAM attention mechanisms; modules; and using the EIOU loss function. We used the same experimental equipment and training parameters in each experiment to test and validate the detection effect in the NTS, CTS, and MTS. Table 2 shows the results of the ablation experiments.

The benchmark model YOLOX network is the least efficient, as seen in Table 2. Comparing the ordinal number 2 in the table, we can see that the improvement in the average accuracy AP on the NTS using the Swin Transformer backbone is insignificant. However, the improvement in the detection of CTS is significant. Swin Transformer contains a large number of parameters but does not have much impact on the detection speed. Its introduction gives the model an AP improvement of 2.95% with almost no loss in detection speed FPS. This demonstrates that Swin Transformer can focus on global information, particularly for the extraction of sophisticated features with a significantly enhanced effect. Serial number 3 adds CA, SE, and CBAM attention mechanism modules to the feature extraction and feature fusion sections, improving the AP of the NTS and the CTS by 0.71 and 0.78%, respectively, while the FPS decreases by 2.5 frames per second, showing that this method is capable of adaptively focusing on and using useful local feature information to lower the rate of missed detection but complicates the model computationally and structurally. Serial number 4 introduces the EIOU loss function and achieves good detection results on both test sets, with an AP improvement of 1.00 and 2.75%, respectively. Although the FPS decreases slightly, this still demonstrates that the introduction of the EIOU loss enhances convergence speed and prevents the degradation of model training caused by the uneven distribution of positive and negative samples, which is a successful addition strategy. The trials in serial numbers 5 to 8 are composite multi-strategy experiments. A comparison of serial numbers 2 to 4 shows that the use of a variety of tactics together achieves better results than the use of just one strategy. From serial numbers 5 and 6, it can be seen that Swin Transformer effectively addressess any speed reduction brought on by the addition of other modules. Serial number 7 uses both the attention module and EIOU loss to ensure that the model works optimally on the NTS. Serial number 8 is the ST-YOLOA network model proposed in this paper. Compared with YOLOX, its AP is improved by 1.14 and 4.83% in the NTS and the CTS, respectively. Although the effect is slightly reduced compared to serial number 7 on NTS, it still achieved second place in the comparison experiment. Its detection speed and detection effect in a complex environment are also substantially ahead.

TABLE 3 Ablation experiments of attentional mechanisms.

SE	CA	CBAM	Precision/%	Recall/%	AP/%	FPS	GFLOPs	Parameters
			79.87	83.49	86.72	22.24	89.60	32.72
√			84.83	81.46	87.46	21.94	89.60	32.72
√	√		85.71	82.53	87.64	21.87	89.60	32.75
√	√	√	**86.27**	**83.61**	**87.74**	21.60	89.61	32.77

Bold represents the maximum value in each column. √ represents the module that added one row of that column to the network.

TABLE 4 Performance comparison of the different algorithms.

Algorithm	NTS			CTS			MTS			FPS	GFLOPs	Parames
	P/%	R%	AP/%	P/%	R/%	AP/%	P/%	R/%	AP/%			
CenterNet	*94.97*	85.20	96.22	**82.64**	57.05	70.95	**89.85**	71.68	85.54	25.13	109.34	32.67M
Faster R-CNN	61.84	**98.33**	95.98	38.88	**76.35**	65.03	50.04	**87.72**	83.18	25.47	401.91	136.69M
SSD	90.40	85.88	94.43	68.68	57.90	66.48	80.61	72.44	82.83	**50.40**	273.40	23.61M
RetinaNet	92.82	91.27	96.80	72.23	64.26	71.36	83.44	78.30	86.47	39.15	163.49	36.33M
EfficientDet	**95.60**	87.35	96.81	*81.33*	55.89	72.32	*89.75*	72.24	86.95	*45.64*	**7.40**	**3.83M**
YOLOv5	90.86	90.69	95.65	69.28	63.63	67.03	80.94	77.69	83.54	32.88	*16.38*	*7.06M*
YOLOX	88.79	94.71	96.23	72.77	67.44	70.86	82.66	82.32	85.52	22.74	27.27	8.94M
YOLOv7	92.06	95.49	*97.36*	75.70	68.40	*73.56*	84.76	82.48	87.50	34.60	105.11	37.20M
ST-YOLOA	91.82	*95.78*	**97.37**	74.24	*72.75*	**75.69**	84.04	*83.44*	**88.50**	21.40	89.61	32.77M

Bold font denotes the best outcome for each column, italics is the second best result.

The loss function curve in Figure 7 clearly illustrates our algorithm's superior performance. In conclusion, the model in this paper meets the real-time detection criteria, while displaying substantially improved target detection accuracy, especially showing excellent detection performance in the complicated environment of the dataset.

In this paper, we introduce three attention mechanisms to enhance the network performance according to the characteristics of different modules to enhance the focus on ship targets. We conducted ablation experiments on three attentional mechanisms, SE, CA, and CBAM, to validate the effectiveness of each attentional mechanism. Table 3 shows the experimental results. The results show that the combination of the three attention mechanisms works optimally.

4.3. Comparative experiments

To objectively evaluate the detection effectiveness of the ST-YOLOA model, we performed comparative experiments using our model and other existing target detection methods. The range of comparison algorithms covers a wide range, among which CenterNet, Faster R-CNN, SSD, RetinaNet (Lin et al., 2017b) and EfficientDet (Tan et al., 2020) are classical target detection models, YOLOv5 (Jocher, 2020) and YOLOX are newly published high-performance target detection models in recent years, and YOLOv7 (Wang et al., 2022) is one of the most advanced detection models at present. Table 4 displays the results of the comparison experiments.

CenterNet predicts the bounding box by learning the centroid and corner point pairs in the feature map without relying on the predetermined anchor box. It has the highest precision rate but a poor recall rate. As a representative algorithm of two-stage detection, the Faster R-CNN detection recall rate has improved significantly but the detection accuracy can be still improved. SSD has been experimentally shown to be unable to efficiently detect ships in complicated surroundings in a SAR picture, despite having a higher detection speed and a more condensed network model. RetinaNet surpasses the previous two-stage algorithm in terms of accuracy and the last detection single-stage algorithm in terms of speed, but there is still much room for improvement efficientDet, as a lightweight network model, has the lowest number of parameters and computations, but its accuracy still needs further improvement. YOLOv5 and YOLOX, as new single-stage detection models, have fewer parameters and computation and more concise network models, which perform better than traditional detection algorithms. However, there is still a certain degree of false detection and leakage, with serious false detection and leakage problems in complex conditions. YOLOv7 shows excellent performance in speed and accuracy, but the detection capability for complex backgrounds still needs further improvement. Compared with other models, ST-YOLOA has an average number of parameters and computation and has higher detection accuracy. It can also meet the basic requirements of real-time detection. Therefore, ST-YOLOA has good overall performance in terms of comprehensive detection, particularly when it comes to SAR image ship detection in complicated environments.

In this research, to demonstrate the detection performance of various models, we compared the PR (precision–recall) curves of each model on two separate test sets. The PR comparison curves are displayed in Figure 8.

To visually compare and analyze the detection effects of the ST-YOLOA model and other algorithms in different scenarios, we

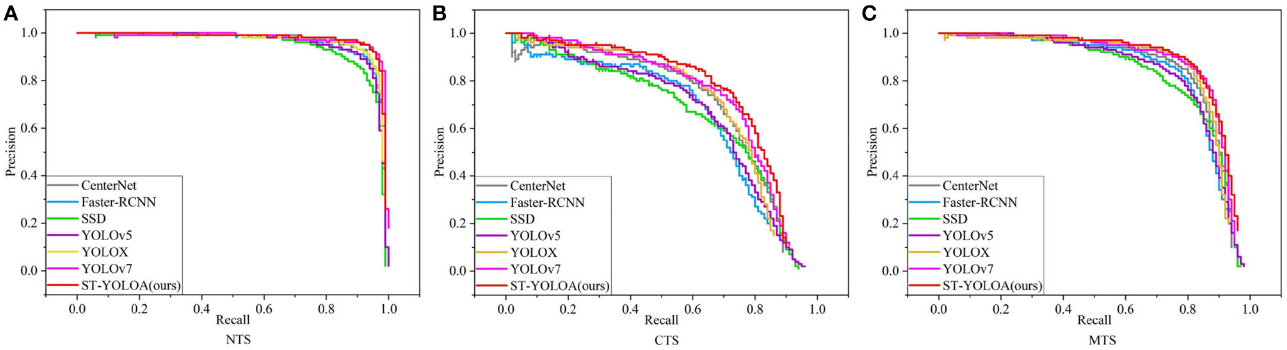

FIGURE 8
Comparison of the PR curves of different algorithm models. **(A)** NTS; **(B)** CTS; **(C)** MTS.

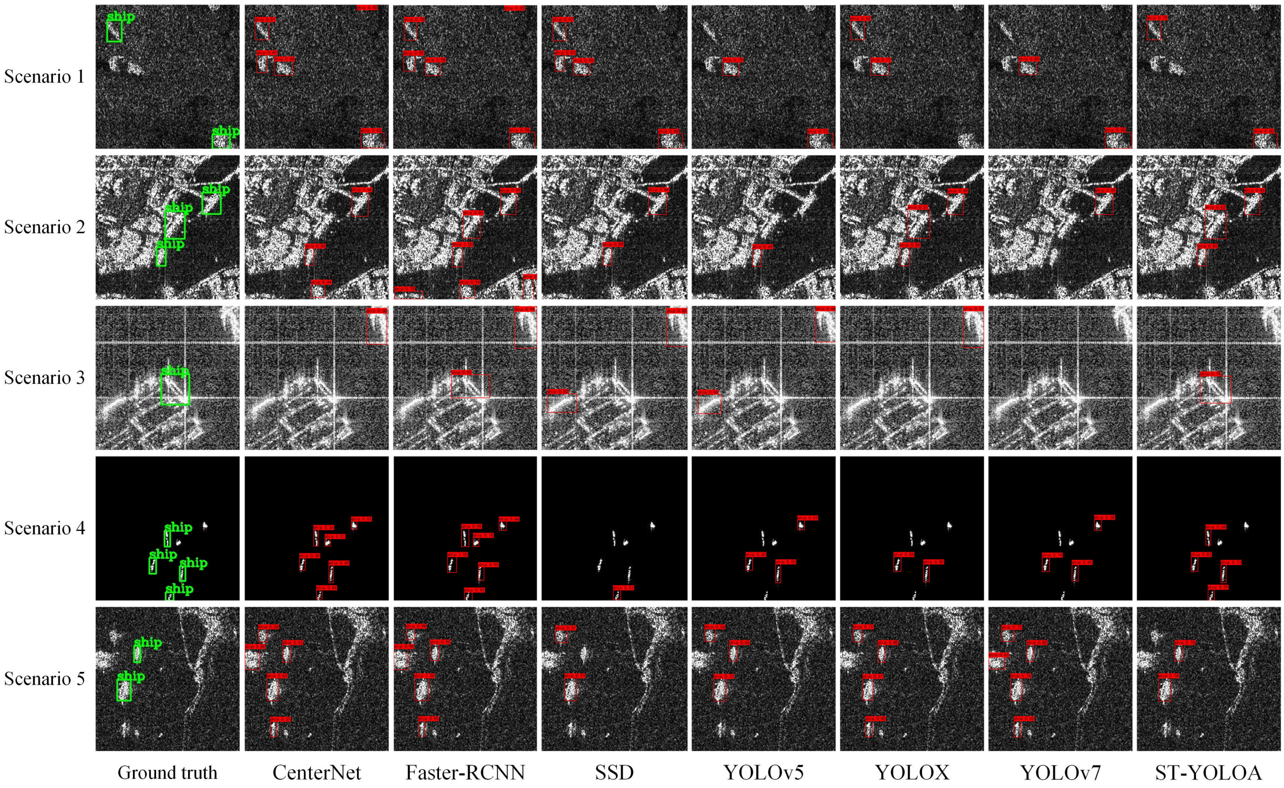

FIGURE 9
Comparison of the detection results.

selected SAR images containing near-coast and far-sea ship targets. Figure 8 shows the detection effect, where the first column of Figure 9 shows the actual labeling result, and other columns show the detection results of each algorithm.

4.4. Generalization ability test

In this paper, to illustrate the generalization capacity of ST-YOLOA, we used two distinct ways for partitioning data (Guo W. et al., 2022): (1) Partitioning the data at random into five ratios: {9:1, 8:2, 7:3, 6:4, 5:5}. (2) Partitioning the data multiple times at random into the ratio 8:2. The test results of the two methods of dataset division are provided in Tables 5, 6.

Table 5 shows that although the number of samples of ship targets in the test samples that are randomly divided by different ratios of the dataset varies significantly, the average accuracy of ST-YOLOA does not change much. However, even though the average accuracy of the samples divided in the ratio of 5:5 among them differed more than the others, it exhibits a good detection ability, which is analyzed because the detection effect degrades as a result of an insufficient number of training samples. The variance of AP for each sample in this experiment is 0.03085, and the variances of the precision and recall are 0.08452 and 0.2078, respectively. This

TABLE 5 Sample cutting in different proportions.

Proportioning	Precision %	Recall %	AP %
9:1	92.17	95.84	97.24
8:2	91.82	95.78	97.37
7:3	91.92	96.92	97.44
6:4	91.38	96.28	97.27
5:5	91.70	96.23	96.98
Mean	91.798	96.21	97.26
Variance	0.08452	0.2078	0.03085

TABLE 6 Multiple sample cuts in the same proportion.

Cutting times	Precision %	Recall %	AP %
1st	91.63	96.60	97.39
2nd	91.82	95.78	97.37
3rd	91.69	96.39	97.27
4th	92.00	96.51	97.24
5th	91.47	97.05	97.36
Mean	91.722	96.266	97.326
Variance	0.03997	0.11733	0.00443

indicates that the ST-YOLOA model proposed in this paper has a stable detection effect for test sets with different numbers of data samples and shows a strong generalization ability.

Due to the same number of samples and smaller variations in the number of ship targets, as shown in Table 6, the variance of the experiment's indicators is lower for samples divided multiple times at the same scale. The mean and variance of the SA-YOLOA model for the same proportion of samples divided multiple times were 91.722% and 0.03997 for precision, 96.266% and 0.11733 for recall, and 97.326% and 0.00443 for mean precision, respectively.

The information in Tables 5, 6 leads to the conclusion that ST-YOLOA performs well and has excellent generalization capacity, both in test samples with various ratios of randomly divided datasets and in test samples with the same proportion of multiple divided datasets.

4.5. Detection effect of the ST-YOLOA model in different scenarios

To visualize the detection effect of the ST-YOLOA model and further measure the model performance, this section first shows the schematic of the confusion matrix of the ST-YOLOA algorithm. As shown in Figure 10, the confusion matrix demonstrates that ST-YOLOA has good performance.

In this study, we demonstrate the effect of ship target detection under different scenarios and scales, including near-shore and far sea. Figure 11 presents the detection effect in each scenario. The first and second rows are near-shore ship targets near islands and near-shore buildings, respectively. Such targets have complex backgrounds and are susceptible to the influence of other non-ship targets around them. Multiple near-shore ship targets can easily be framed by a single detection box due to the dense docking of ships, which suppresses candidate boxes with high overlap and low prediction scores. The third row is a small, dense target in the distance that is easy to miss because it has a small ship scale. The ship target in the fourth row is prone to erroneous target localization since it has indistinct target borders and complicated background information. In all four aforementioned scenarios, the ST-YOLOA model significantly improved the detection rate and accuracy, as can be seen from the figure, and produced positive detection results.

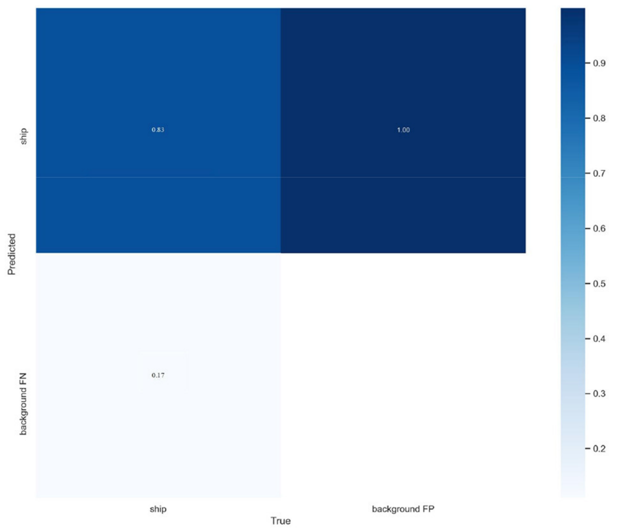

FIGURE 10
Confusion matrix.

4.6. Limitations and discussion

The results of previous experimental studies show that our model achieves sound visual effects in SAR ship detection in complex scenes. It is demonstrated that the ST-YOLOA model can learn global features and can be used to extract more powerful semantic features for ship target detection in harsh environments and complex scenes. However, our approach still suffers from some limitations.

The relatively high computational complexity and large number of parameters of the Swin Transformer module lead to more extended training and inference time. As seen from the experimental ablation results in Table 1, although we have used the Swin-Transformer network with a smaller model as much as possible, its use still introduces many parameters compared to the base model. The Swin Transformer network has a solid global modeling capability, capturing rich global feature information and integrating global data. This process requires a vast amount of support operations, resulting in more parameters and computations than other models. At the same time, the computational complexity of the Swin Transformer increases with the length of the input sequence. When dealing with very long input sequences, Swin Transformer may face problems such as high computational complexity and large memory consumption,

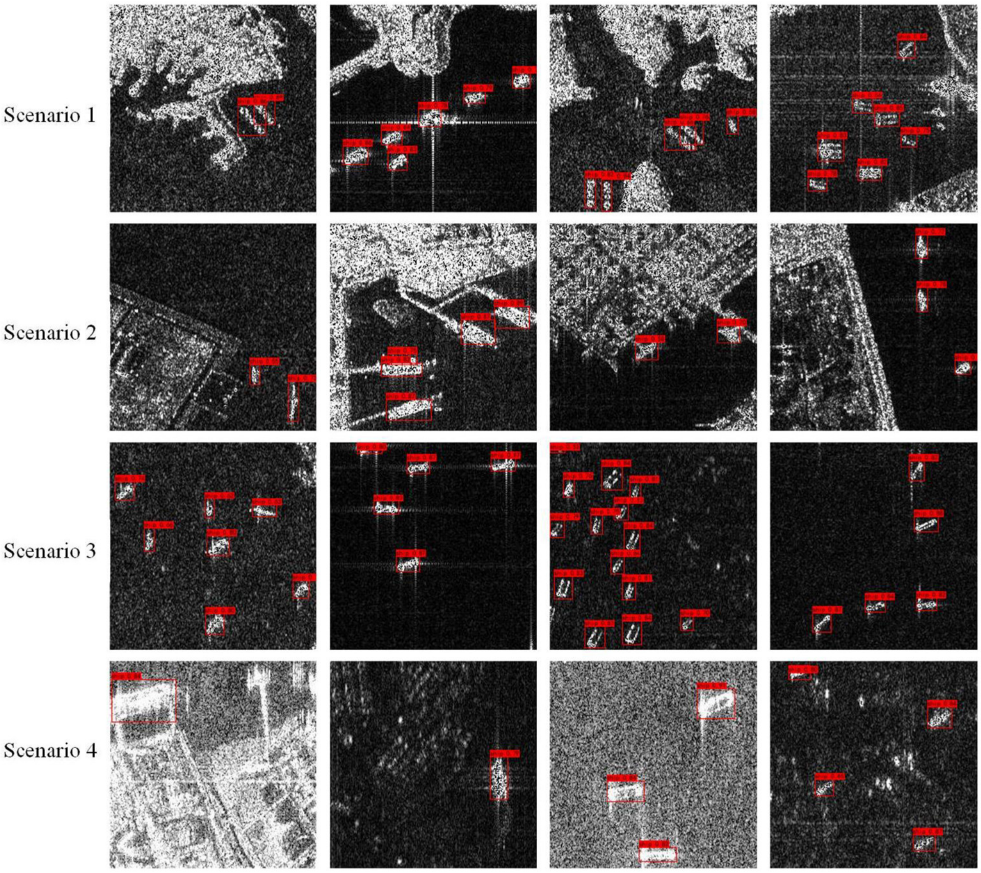

FIGURE 11
Detection effects in different scenarios.

which need to be alleviated by using lightweight models or other techniques.

5. Conclusions

To ensure the accuracy of SAR ship target recognition under complicated situations, in this study, we have suggested a more extended ST-YOLOA ship target identification model. To begin with, the feature extraction section adds the Patch Embedding module after the input layer to chunk and flatten the input image and then produces feature maps of varying sizes using Swin Transformer Blocks and the Patch Merging layer. A coordinated attention mechanism is designed at the end to simultaneously capture position information and channel relationships, which significantly improves the performance of downstream tasks. Second, to effectively use semantic and localization information, the PANet is employed to thoroughly fuse high-level and low-level feature information. Finally, a decoupled detection head in the target detection section is used to significantly speed up model convergence and improve the position loss function, both of which improve model performance. This model is more suited for ship target detection in challenging surroundings and complex circumstances because it can extract more potent semantic characteristics and can better learn global features than other detection models.

Considering that our model focuses on improving SAR ship detection accuracy in complex environments, the vital index of the number of parameters of the model is ignored to a certain extent. In the future, we will further conduct model optimization and carry out research on model lightweight by adjusting hyperparameters and model compression methods, such as quantization, distillation, and pruning, and further analysis on lightweight Swin Transformer to achieve lower model parameter computation, faster training speed, and maintain previous accuracy.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization: XY, KZ, and RL. Methodology and writing–original draft preparation: KZ and RL. Software: RL. Investigation: JF and SW. Resources and visualization: KZ. Writing–review and editing: XY, KZ, and SW. Supervision: KZ and SW. Project administration and funding acquisition: XY. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62276274 (Corresponding author: RL).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

An, Q., Pan, Z., Liu, L., and You, H. (2019). DRBox-v2: an improved detector with rotatable boxes for target detection in SAR images. *IEEE Trans. Geosci. Remote Sensing* 57, 8333–8349. doi: 10.1109/TGRS.2019.2920534

Bochkovskiy, A., Wang, C. Y., and Liao, H-, Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv* 2004, 10934. doi: 10.48550/arXiv.2004.10934

Cumming, I. G., and Wong, F. H. (2005). Digital processing of synthetic aperture radar data. *Artech House* 1, 108–110.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q., et al. (2019). "Centernet: Keypoint triplets for object detection", in *Proceedings IEEE/CVF,* 6569–6578. doi: 10.1109/ICCV.2019.00667

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2009). The pascal visual object classes (voc) challenge. *Int. J. Computer Vision* 88, 303–308. doi: 10.1007/s11263-009-0275-4

Gao, G., Gao, S., He, J., and Li, G. (2018). Adaptive ship detection in hybrid-polarimetric SAR images based on the power–entropy decomposition. *IEEE Trans. Geosci. Remote Sensing* 56, 5394–5407. doi: 10.1109/TGRS.2018.2815592

Gao, W., Liu, Y., Zeng, Y., Li, Q., and Liu, Q. (2022). Enhanced attention one shot SAR ship detection algorithm based on cluster analysis and transformer in *Second International Conference on Digital Signal and Computer Communications (DSCC 2022): SPIE,* 290–295. doi: 10.1117/12.2641456

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv*:2107, 08430.

Girshick, R. (2015). "Fast r-cnn", in *Proceedings of the IEEE International Conference on Computer Vision,* 1440–1448. doi: 10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 580–587. doi: 10.1109/CVPR.2014.81

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., et al. (2022). Attention mechanisms in computer vision: a survey. *Comput. Visual Media* 8, 331–368. doi: 10.1007/s41095-022-0271-y

Guo, W., Shen, L., Qu, H., Wang, Y., and Lin, C. (2022). Ship detection in SAR images based on adaptive weight pyramid and branch strong correlation. *J. Image Graphics* 27, 3127–3138. doi: 10.11834/jig.210373

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 13713–13722.

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 7132–7141.

Jiang, X., Cai, W., Yang, Z., Xu, P., and Jiang, B. (2022). Infrared dim and small target detection based on YOLO-IDSTD algorithm. *Infrared Laser Eng.* 51, 502–511. doi: 10.3788/IRLA20210106

Jocher, G. (2020). *YOLOv5.* Available online at: https://github.com/ultralytics/yolov5 (accessed December 5, 2022).

Li, K., Zhang, M., Xu, M., Tang, R., Wang, L., Wang, H., et al. (2022). Ship detection in SAR images based on feature enhancement Swin transformer and adjacent feature fusion. *Remote Sensing* 14, 3186. doi: 10.3390/rs14133186

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S., et al. (2017a). "Feature pyramid networks for object detection", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2117–2125. doi: 10.1109/CVPR.2017.106

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Int.* 8, 2999–3007. doi: 10.1109/ICCV.2017.324

Liu, C., Xie, N., Yang, X., Chen, R., Chang, X., Zhong, R. Y., et al. (2022). A domestic trash detection model based on improved YOLOX. *Sensors* 22, 6974. doi: 10.3390/s22186974

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). "Path aggregation network for instance segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). "Ssd: Single shot multibox detector", in *Computer Vision–ECCV* 2016, 14th. *European Conference* (Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14: Springer), 21–37.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows", in *Proceedings of the IEEE/CVF International Conference Computer Vision,* 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Lu, R., Yang, X., Jing, X., Chen, L., Fan, J., Li, W., et al. (2020a). Infrared small target detection based on local hypergraph dissimilarity measure. *IEEE Geoscience Remote Sens Lett* 19, 1–5. doi: 10.1109/LGRS.2020.3038784

Lu, R., Yang, X., Li, W., Fan, J., Li, D., Jing, X., et al. (2020b). Robust infrared small target detection via multidirectional derivative-based weighted contrast measure. *IEEE Geosci. Remote Sensing Letters* 19, 1–5. doi: 10.1109/LGRS.2020.3026546

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K. P., et al. (2013). A tutorial on synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* 1, 6–43. doi: 10.1109/MGRS.2013.2248301

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 779–788. doi: 10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Proc. Syst.* 28, 1–47. doi: 10.1109/TPAMI.2016.2577031

Robey, F. C., Fuhrmann, D. R., Kelly, E. J., and Nitzberg, R. (1992). A CFAR adaptive matched filter detector. *IEEE Trans. Aerospace Electr. Syst.* 28, 208–216. doi: 10.1109/7.135446

Schwegmann, C. P., Kleynhans, W., and Salmon, B. P. (2015). Manifold adaptation for constant false alarm rate ship detection in South African oceans. *IEEE J. Selected Topics Appl. Remote Sensing* 8, 3329–3337. doi: 10.1109/JSTARS.2015.2417756

Tan, M., Pang, R., and Le, Q. V. (2020). "EfficientDet: Scalable and Efficient Object Detection", in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR).

Tian, Z., Shen, C., Chen, H., and He, T. (2019). "Fcos: Fully convolutional one-stage object detection", in *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 9627–9636.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Proc. Syst.* 30, 2.

Wang, C., Bi, F., Zhang, W., and Chen, L. (2017). An intensity-space domain CFAR method for ship detection in HR SAR images. *IEEE Geosci. Remote Sensing Letters* 14, 529–533. doi: 10.1109/LGRS.2017.2654450

Wang, C.Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., Yeh, I. H., et al. (2020). "CSPNet: A new backbone that can enhance learning capability of CNN", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* 390–391.

Wang, C. Y., Bochkovskiy, A., and Liao, H-, Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* 2207, 02696. doi: 10.48550/arXiv.2207.02696

Wang, J., Zhang, Z., Luo, L., Zhu, W., Chen, J., Wang, W., et al. (2021). SwinGD: A robust grape bunch detection model based on swin transformer in complex vineyard environment. *Horticulturae* 7, 492. doi: 10.3390/horticulturae7110492

Wang, Y., Wang, C., and Zhang, H. (2018). Combining a single shot multibox detector with transfer learning for ship detection using sentinel-1 SAR images. *Remote Sensing Lett.* 9, 780–788. doi: 10.1080/2150704X.2018.1475770

Wu, Z., Liu, C., Wen, J., Xu, Y., Yang, J., Li, X., et al. (2022). *Selecting High-Quality Proposals for Weakly Supervised Object Detection With Bottom-Up Aggregated Attention and Phase-Aware Loss. IEEE Transactions on Image Processing.* doi: 10.1109/TIP.2022.3231744

Xia, R., Chen, J., Huang, Z., Wan, H., Wu, B., Sun, L., et al. (2022). CRTransSar: a visual transformer based on contextual joint representation learning for SAR ship detection. *Remote Sensing* 14, 1488. doi: 10.3390/rs140 61488

Yuan, Y., and Zhang, Y. (2021). OLCN: An optimized low coupling network for small objects detection. *IEEE Geosci. Remote Sensing Letters* 19, 1–5. doi: 10.1109/LGRS.2021.3122190

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *Remote Sensing* 1710, 09412. doi: 10.48550/arXiv.1710.09412

Zhang, S., Chi, C., Yao, Y., Lei, Z., and Li, S. Z. (2020). "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 9759–9768.

Zhang, Y., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T., et al. (2022). Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* 506, 146–157. doi: 10.1016/j.neucom.2022.07.042