smart cities

MDPI

Article

Small-Scale Ship Detection for SAR Remote Sensing Images Based on Coordinate-Aware Mixed Attention and Spatial Semantic Joint Context

Zhengjie Jiang [1], Yupei Wang [1,*], Xiaoqi Zhou [2], Liang Chen [1], Yuan Chang [1], Dongsheng Song [1] and Hao Shi [1]

[1] School of Information and Electronics, Beijing Institute of Technology (BIT), Beijing 100081, China; chenl@bit.edu.cn (L.C.)
[2] Chongqing Innovation Center, Beijing Institute of Technology, Chongqing 401135, China
* Correspondence: 7520190118@bit.edu.cn

Abstract: With the rapid development of deep learning technology in recent years, convolutional neural networks have gained remarkable progress in SAR ship detection tasks. However, noise interference of the background and inadequate appearance features of small-scale objects still pose challenges. To tackle these issues, we propose a small ship detection algorithm for SAR images by means of a coordinate-aware mixed attention mechanism and spatial semantic joint context method. First, the coordinate-aware mixed attention mechanism innovatively combines coordinate-aware channel attention and spatial attention to achieve coordinate alignment of mixed attention features. In this way, attention with finer spatial granularity is conducive to strengthening the focusing ability on small-scale objects, thereby suppressing the background clutters accurately. In addition, the spatial semantic joint context method exploits the local and global environmental information jointly. The detailed spatial cues contained in the multi-scale local context and the generalized semantic information encoded in the global context are used to enhance the feature expression and distinctiveness of small-scale ship objects. Extensive experiments are conducted on the LS-SSDD-v1.0 and the HRSID dataset. The results with an average precision of 77.23% and 90.85% on the two datasets show the effectiveness of the proposed methods.

Keywords: Synthetic Aperture Radar (SAR); ship detection; small object detection; attention mechanism; context

Citation: Jiang, Z.; Wang, Y.; Zhou, X.; Chen, L.; Chang, Y.; Song, D.; Shi, H. Small-Scale Ship Detection for SAR Remote Sensing Images Based on Coordinate-Aware Mixed Attention and Spatial Semantic Joint Context. *Smart Cities* **2023**, *6*, 1612–1629. https://doi.org/10.3390/smartcities6030076

Academic Editor: Pierluigi Siano

Received: 10 April 2023
Revised: 17 May 2023
Accepted: 29 May 2023
Published: 15 June 2023

1. Introduction

Transportation is an indispensable and important part of modern city life. With the increasing construction and improvement of urban transportation networks, transportation targets have shown diversity, including vehicles, bridges, ships, airports, airplanes, etc. However, with the modernization, informatization, and intelligence development of the transportation industry, various foreseeable and unforeseeable risks, as well as challenges, are also constantly increasing, such as the excessive speed of cars, insufficient road capacity, difficulty in vessel monitoring, etc. For conventional technical methods, it's challenging to timely obtain information from massive data and grasp the spatial distribution as well as dynamics changes of urban traffic targets from a macro perspective. Therefore, there is an urgent need for a transportation target processing method that adapts to the diverse characteristics of the stereoscopic space of smart cities.

Due to the all-weather imaging characteristics, Synthetic Aperture Radar (SAR) has been widely used in military and civil fields, such as airport reconnaissance, urban traffic control, port vessel surveillance, and maritime emergency rescue [1]. Object detection with accurate localization and classification is a key link in the intelligent interpretation of remote sensing images, providing a technical basis for urban transportation planning and promoting the orderly development of smart cities. As a basic task, SAR ship detection

plays an important role in maritime traffic control. Traditional SAR ship detection generally adopts the Constant False Alarm Rate (CFAR) algorithm based on the background clutter statistical distribution [2–4] or the methods that manually extract texture features [5–7]. Nevertheless, the CFAR conducts the same statistical model for all clutters and cannot fully utilize the feature clues of images. The texture-based methods are able to better exploit the image features but the manual extraction of features requires a complex design and has poor generalization ability.

Taking advantage of the rapid development of Convolutional Neural Networks (CNN) at present, many object detectors [8–17] that are outstanding in natural scenes have emerged. These object detection models based on CNN adaptively learn the advanced semantic expression of images via various network structures and powerful training strategies, becoming the mainstream algorithm for object detection. The current object detectors can be divided into two-stage models represented by Faster R-CNN [11] and single-stage models such as the YOLO series [8–10]. The two-stage detection framework uses a separate Region Proposal Network (RPN) to search for candidate boxes in the first stage and it implements the finer detection in the second stage. The single-stage models directly collect features from the input images and use predefined anchor boxes or key points to locate objects. As a result, two-stage detection networks require longer processing time and have complex model architectures, while single-stage detectors with lower detection accuracy are superior in terms of real-time performance and network lightweight design.

Owing to the unique imaging mechanism, within ship detection in SAR images exists significant barriers compared to natural scenes or optical remote sensing images. First, SAR images contain a lot of clutters, such as islands and waves, which seriously interfere with the perception of object features. In comparison to optical images, ships in SAR images lack sufficient recognizable texture information, making it difficult to distinguish the real targets from the background. Moreover, the low resolution of SAR images results in small object sizes. With limited visualization information and fuzzy boundaries of small-scale objects, the detection process is fragile and sensitive. The general detectors adopt a multi-level down-sampling paradigm to achieve high-level semantic clues. However, the small-scale objects lose most of their salient features in this framework, which further increases the difficulty of detection. Therefore, there is still a wide gap for these general object detectors in SAR image tasks.

In order to alleviate the impact of complex background and noise interference, inspired by the selective attention of human beings, some studies have proposed an attention mechanism to guide the sight of models to the foreground regions. Researchers [18,19] utilized channel attention to enhance the semantic category features of SAR ship objects, thus, automatically highlighting the important regions. However, the summarization of global features on each channel completely discards the spatial details of objects, leading to the inaccurate localization of the focus areas. Gao et al. [20] parallelly extracted spatial and channel attention features, and directly fused them through addition without considering the feature alignment between the two types of attention. For this reason, the integrated attention features cannot accurately describe the spatial and semantic characteristics. To promote the correlation between the two types of attention, studies [21–24] leveraged the CBAM (Convolutional Block Attention Module) mixed attention mechanism to capture spatial attention based on the channel attention features. Nevertheless, spatial details are lost during channel attention extraction, so the spatial attention mined on this basis cannot guarantee the positional feature alignment either.

Another line of works tried to exploit contextual information to intensify the feature expression of small-scale objects. Context is the relationship between objects and the environment. For instance, a rectangular object can be roughly identified as a ship rather than a car according to its sea background. Kang et al. [25] added contextual features to the corresponding regions of interest to enhance the distinguishability of ship features. Zheng et al. [26] expanded the effective receptive fields with the double dilated convolutions, so as to strengthen the context information of shallow-level features and enrich

the characteristic representation. Lim et al. [27] used the additional features in different levels as context to focus on targeted objects. Works [28,29] captured the global contextual correlation and used the environmental inductive clues to guide the detection of small-scale objects. However, these works mentioned above lack effective incorporation of a local and global context and, thus, cannot fully utilize the essential correlation between spatial and semantic clues provided by the environment, leading to sub-optimal solutions.

To this end, in this paper, we propose a coordinate-aware mixed attention mechanism and a spatial semantic joint context method to improve the small-scale ship detection performance in SAR images. Specifically, the coordinate-aware mixed attention mechanism solves the problem of feature misalignment between channel and spatial attention. By capturing channel attention in both horizontal and vertical directions, coordinate information is automatically encoded into the channel attention. Spatial attention is then extracted based on this. In this manner, channel attention features with coordinate perception ability can be easily aligned with spatial attention features. With finer spatial granularity of attention features, small-scale objects can be more precisely located and highlighted and background clutters can be more effectively suppressed. Furthermore, the spatial semantic joint context method simultaneously explores multi-scale local spatial clues and the global semantic context of objects. As a result, the spatial details of the adjacent area of objects and the semantic summaries of the entire scene constitute multiple environmental auxiliary information, which can enrich the feature representation and promote the detection performance of small-scale ships in SAR images.

The main contributions of this work are listed as follows:

1. To overcome the interference of background noise in SAR images, we first propose a coordinate-aware mixed attention mechanism. The effective combination of coordinate-aligned channel and spatial attention features contributes to more accurate identification of small-scale ships from the background.
2. We design a spatial semantic joint context method to construct multiple environmental additional information by composing a local spatial and global semantic context, so as to intensify the feature expression and distinguishability of small-scale ships.
3. Experimental results demonstrate that our proposed methods not only significantly increase the detection performance of small-scale ship detection in SAR images but also gain superiority over other state-of-the-art models.

2. Related Work

2.1. Object Detection

In recent years, with the vigorous development of deep learning, intelligent image interpretation algorithms have also emerged. Due to its powerful feature extraction and description capabilities, as well as its robustness and adaptability to different scenarios, deep learning methods can not only avoid complex manual work but also possess high accuracy and generalization. With the help of natural bionics characteristics, deep-learning-based CNN has become the mainstream algorithm in the field of object detection. Currently, object detection networks can be primarily divided into two-stage and single-stage frameworks. The two-stage detectors typically exhibit good performance and stability by generating accurate bounding boxes, where region proposals are generated in the first stage while bounding box location regression and classification are completed in the second stage, such as R-CNN [14], Fast R-CNN [15], Faster R-CNN [11], and Cascade R-CNN [16]. In contrast, the single-stage networks adopt a dense sampling method to directly extract features from the input images and use predefined anchor boxes or key points with multiple scales and aspect ratios to locate targets. Representative frameworks include SSD [17], YOLO series [8–10], FCOS [13], etc. Although the single-stage framework has a slightly lower detection accuracy, it leads the two-stage detectors in terms of real-time performance and network simplification design. Despite the remarkable results achieved, general object detection networks cannot be directly applied to ship detection in SAR images. First, in comparison to optical images, SAR images can only reflect the electromagnetic scattering

intensity of objects, and hence, only the grayscale information can be represented. However, in SAR images there are plenty of background clutters such as ports, islands, waves, etc., which exhibit a similar scattering intensity and appearance characteristics with targets. These noises can seriously interfere with ship detection, resulting in a large number of false alarms. In addition, due to resolution limits, the object pixel area in SAR images is extremely small. Insufficient appearance features such as texture and contour make it difficult for detectors to extract recognizable information. The general paradigm of the current mainstream object detection framework is to down sample images several times to obtain semantic-rich deep feature maps. Small-scale ship objects with less information will further lose useful information in this process, leading to miss detection. Therefore, special network structures are necessary for the task of small-scale ship detection in SAR images.

2.2. Ship Detection under Complex Background in SAR Images

Complex background objects and clutter interference bring massive false alarms and, thus, affect detection accuracy. In response to this issue, many experts and scholars have conducted in-depth research based on the idea of CNN and achieved significant results. Yue et al. [30] designed a feature extraction network on the basis of VGG and dilated convolution, evidently improving inference speed and precision. In order to alleviate the problem of noise interference in a complex environment, Zhang et al. [31] proposed a SAR ship detection network by means of a Fully Convolutional Network (FCN), which transforms the object detection task into an image pixel classification problem. The experimental results show that this algorithm can not only effectively reduce false alarms, but also enhance the network robustness. Aiming at the issue of confusion between ships and onshore objects in SAR images, Fu et al. [32] introduced a nearshore SAR ship detection algorithm based on scene classification. This method can achieve better detection results with more land scenes. To further improve the performance of nearshore ship detection, Liu et al. [33] designed a novel sea–land segmentation method based on FCN which fully integrates global and local information of SAR images and has a high detection accuracy. Works [18–24] are the most relevant research to this paper, which adopted channel attention or a channel-spatial mixed attention mechanism to automatically intensify the object features and suppress background noise. Nevertheless, these attention modules used either completely abandon the spatial information or fail to consider the feature alignment between channel and spatial attention. Hence, the fused features cannot provide an accurate spatial and semantic description of objects, resulting in low efficiency of the attention mechanism. In this paper, we propose a coordinate-aware mixed attention mechanism that solves the problem and, thus, achieves a more refined attention effect granularity. In addition, via precision positioning and the highlighting of objects, background noise can be effectively filtered.

2.3. Small-Scale Ship Detection in SAR Images

In order to enhance the detection performance of small-scale ship objects, the key is to ensure that small targets will not lose information in high-level features. In the computer vision field, the common solution is feature fusion, which strengthens the feature expression by fully integrating the position information of low-level features and the semantic clues in high-level features. Research [34] adopted a feature pyramid fusion method to improve the detection accuracy of small-scale ships. Hu et al. [35] introduced the feature pyramid structure and balance factor on the basis of YOLOv3, effectively optimizing the weight of small objects in the loss function. Cui et al. [24] proposed a dense attention pyramid network that utilizes a CBAM module to fully connect the top and bottom features. This method extracts rich features including spatial clues and semantic information and solves the problem of multi-scale ship detection. The experimental results demonstrate that this algorithm has a high detection accuracy but poor adaptability to different scenarios. For the sake of further improving the detection performance of small-scale objects, other works [25–29] attempted to mine the contextual information and use auxiliary cues from the

environment to enrich the object's features. Compared to the simple feature fusion method that can only combine existing features, contextual approaches are able to capture more additional information. However, due to the lack of effective aggregation of local and global contexts, this solution cannot fully associate the spatial clues and semantic relationship provided by the scene, limiting the performance gains. In contrast, our proposed spatial semantic joint context method not only explores the multi-scale spatial information and global semantic cues of environments but also organically assembles these two types of contexts. The multiple spatial and semantic auxiliary information is fully utilized to boost the detection performance of small-scale ships.

3. Materials and Methods

In this section, we first illustrate the overall architecture of our model. Then, based on the analysis of various common attention mechanisms, the proposed coordinate-aware mixed attention mechanism is introduced. Last, we present the spatial semantic joint context method.

3.1. Overall Architecture of Network

As shown in Figure 1, the proposed network takes YOLOX [10] as the basic framework. The original YOLOX utilizes four Cross Stage Partial blocks (CSP) to implement the feature extraction. Our proposal embeds the Coordinate-aware Mixed Attention module (CMA) into the feature extraction block to form the CMA-CSP. The deepest two of the three extracted feature layers are then fed into the Spatial Semantic Joint Context module (SSJC) to capture the local spatial and global semantic contextual cues. Next, the top-down and bottom-up pathways are used to fully interact with the location-rich features in the lower layers and the semantic-rich features in the deeper layers. Last, we output the predicted results with the bounding box heads of YOLOX.

3.2. Coordinate-Aware Mixed Attention Mechanism

An attention mechanism is generally used to emphasize important features, so as to lead models to focus on targeted regions. As such, irrelevant information or background interference can be suppressed. Three common attention modules are shown in Figure 2. The SE (Squeeze-and-Excitation) channel attention mechanism [36] compresses features on each channel into an inductive neuron via Global Average Pooling (GAP) and builds the global associations among all semantic channels through non-linear connections. However, this method using GAP ignores the position distribution information, and cannot satisfy the high positioning requirements of remote sensing small-scale objects. In order to introduce spatial information, the CBAM mixed attention mechanism [37] adds a spatial attention module behind the channel attention extraction process. Nevertheless, the Global Max Pooling (GMP) and GAP adopted in the channel attention module still summarize channel features in a rough form, resulting in an insufficient spatial refinement of channel attention features. This makes it difficult for the subsequent spatial attention to align with the channel attention features. To bring the spatial information into channel attention, the CA (Coordinate Attention) mechanism [38] compresses channel features in both horizontal and vertical directions, enabling channel attention to possess a spatial guidance capability. However, the spatial guidance is still inadequate for the accurate positioning of small-scale objects. In order to further refine the granularity of the attention mechanism, we propose the CMA, as shown in Figure 1.

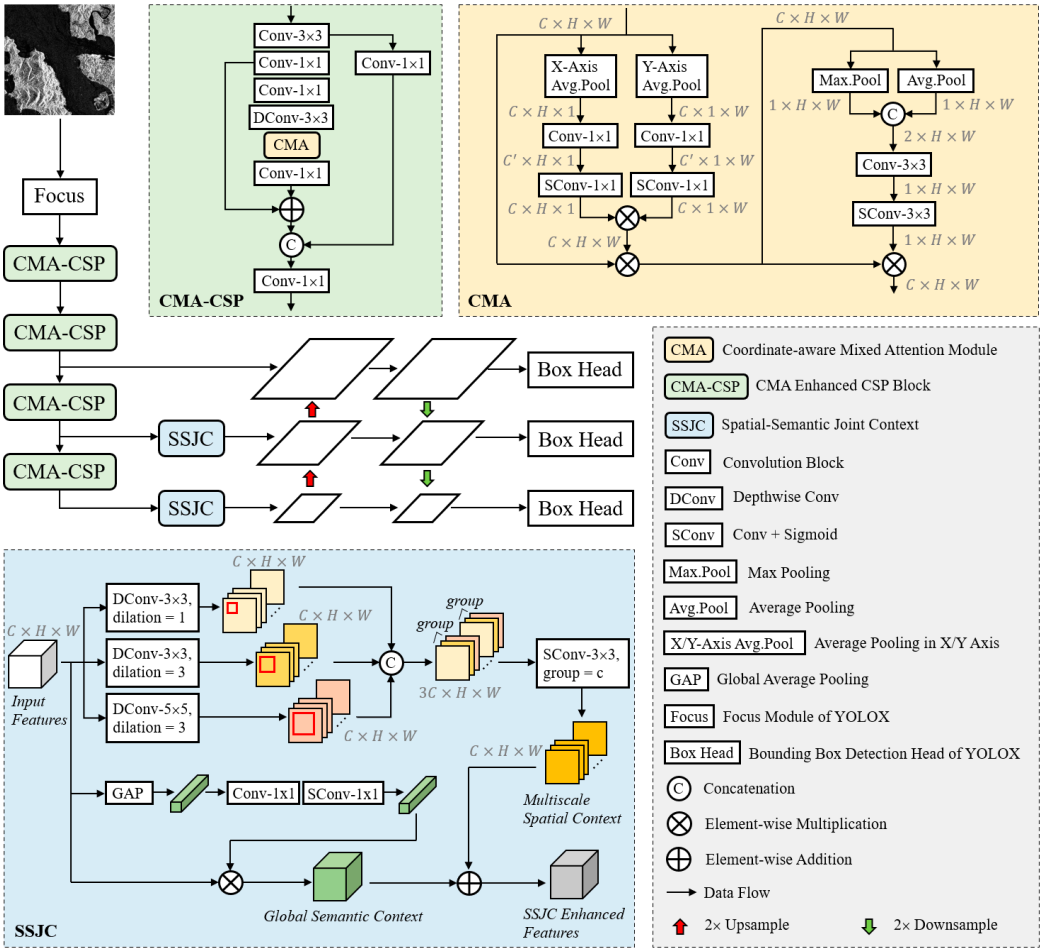


Figure 1. Network architecture of our methods. The whole frame is based on YOLOX. The CMA is embedded into feature extraction block CMA-CSP to enhance the focusing ability on small objects. The SSJC is designed to exploit environmental information to enrich the feature expression.

Similar to the CA module, CMA embeds spatial information into channel attention to preserve the position correlation of channel attention features. Specifically, the CMA firstly implements a one-dimensional average pooling on input features in the x- and y-axis, respectively. The produced two two-dimensional feature maps, which represent the coordinate-aware summarized features in horizontal and vertical directions, are then independently fed into two Conv-1 $\times$ 1 layers and a sigmoid function. In this way, features across channels interact and global dependency of features in the two directions is established. Next, we multiplicatively integrate the two attention features as the coordinate-aware channel attention and use it to strengthen the input features. For spatial attention, we use the one-dimensional max and average pooling to extract the important spatial cues on the coordinate-aware channel-enhanced features. Through channel-wise concatenation, the two resulting spatial feature maps are incorporated. The spatial attention is then calculated by means of two consecutive Conv-1 $\times$ 1 layers with a sigmoid function, and mixed with the channel attention features to construct the coordinate-aware mixed attention. Since we excavate the spatial attention based on the channel attention features that retain coordinate information, the spatial distribution of channel and spatial attention features can be correctly aligned when mixing, thus, reducing the information loss and making the effect of the granularity of the attention mechanism more fine-grained.

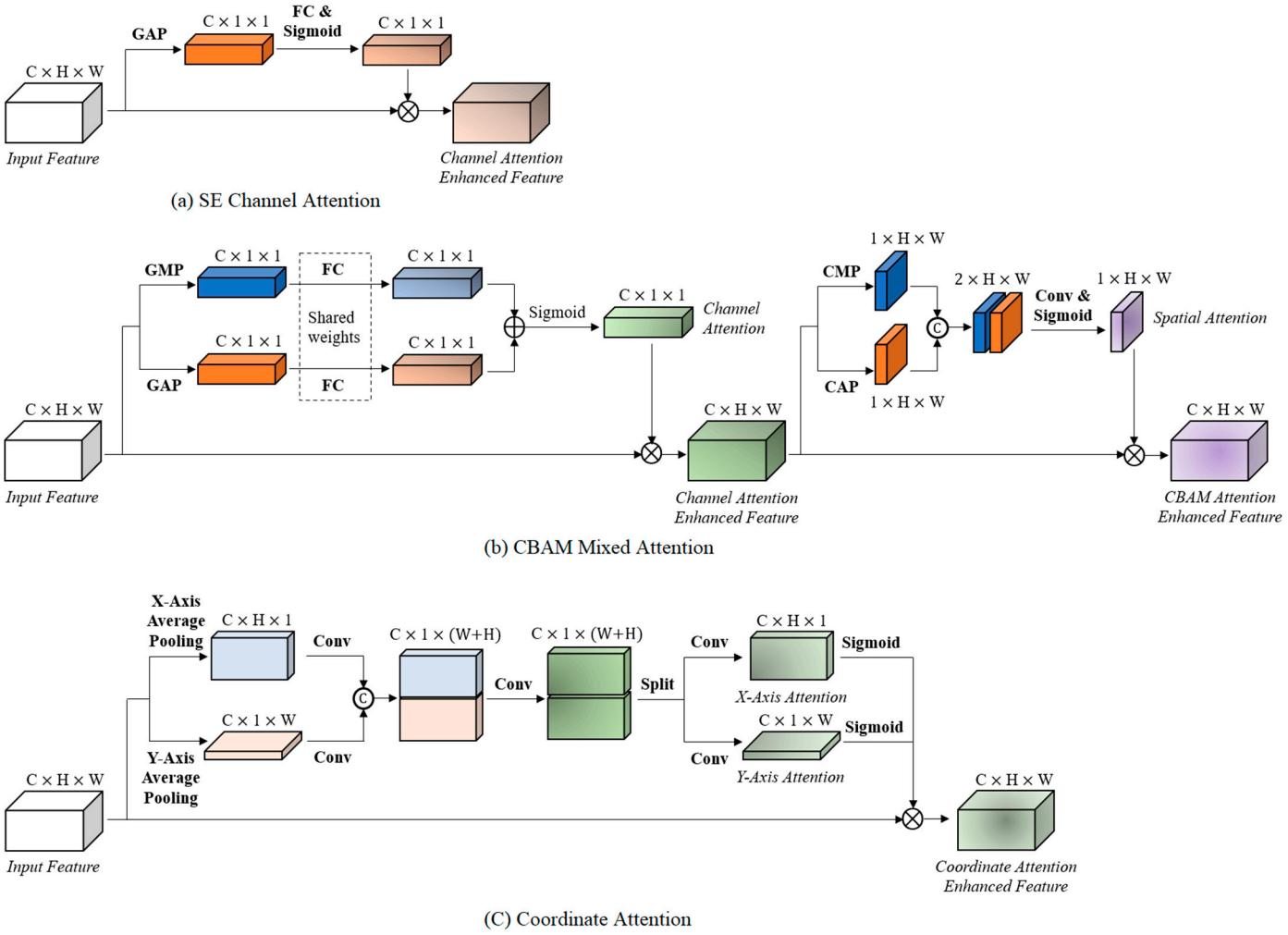


Figure 2. Module structures of different attention mechanisms.

We leverage the CMA to construct the coordinate-aware mixed attention feature extraction block, i.e., the CMA-CSP. Inspired by the CSP structure, the CMA-CSP separates the features into two parts, which are fed into different processing pathways and merged to obtain richer gradient combination information. Given an input feature map, CMA-CSP first adjusts the channels and scale of the feature map via a Conv-3 × 3 with stride 2. Next, the obtained features are sent to two processing paths, one consisting of a Conv-1 × 1 and N residual blocks with CMA, and the other only including a Conv-1 × 1. In the residual block, a 1 × 1 convolution is first executed to reduce the channel number, so as to reduce the computational complexity. After the 1 × 1 depth-wise convolution, features are then sent into the CMA to generate the coordinate-aware mixed attention features. Recovering the channels via a Conv-1 × 1, the feature map is additively fused with the input features of the residual block. Last, we stack the features with the output on another CSP path and then smooth the concatenated feature map using a 1×1 convolution. The feature extraction module with coordinate-aware mixed attention learns to pay more attention to important regions during feature extraction and increases the positioning accuracy of small-scale objects with finer attention granularity. As a result, when suppressing background interference in SAR images, the detection performance of small-scale ships can also be improved.

3.3. Spatial Semantic Joint Context Method

Limited by the number of pixels, small-scale objects carry inadequate information and, thus, lack crucial appearance characteristics apart from the background or similar objects.

Therefore, it is difficult for general detectors to identify small ships. Contextual information brings the environment around the objects to assist in inferring the location and category of the targets. The SSJC proposed in this paper on the one hand extracts multi-scale local spatial features to obtain the spatial clues of the adjacent areas of objects. On the other hand, it captures global semantic features and establishes a semantic correlation between objects and the entire scene. By integrating the two types of contextual information, multiple environmental auxiliary clues composed of local positioning and global semantics can be mined to intensify the feature expressiveness and recognition of small objects.

The structure of the SSJC is illustrated in Figure 1. Concretely, input features are first fed into three parallel depth-wise dilated convolution workflows to attain the spatial features. The depth-wise convolution not only contributes to reducing the computation cost but also to using the channel independency to lead the focus of the model on the spatial context. Owing to the different kernel sizes (3×3, 3×3, 5×5) and dilation rates (1, 3, 3) implemented, multi-scale receptive fields are produced. Then, the three contextual feature maps are concatenated according to the corresponding channels, where everythree adjacent slices is integrated into one through a 1×1 group convolution with a sigmoid function to reinforce the correlation among the multi-scale spatial context in the same channel. Meanwhile, we capture the semantic contextual information of the input features through a global average pooling and two Conv-1×1 layers with sigmoid activation. Through multiplicative fusion, the semantic context is generated and then added to the multi-scale spatial context to form the SSJC-enhanced features. The multi-scale spatial context represents the adjacent environmental information in different spatial ranges around objects, while the semantic context refers to the global semantic association between objects and the entire scene. Through their effective combination, features exhibit a more detailed expression and, thus, small-sized objects are more recognizable.

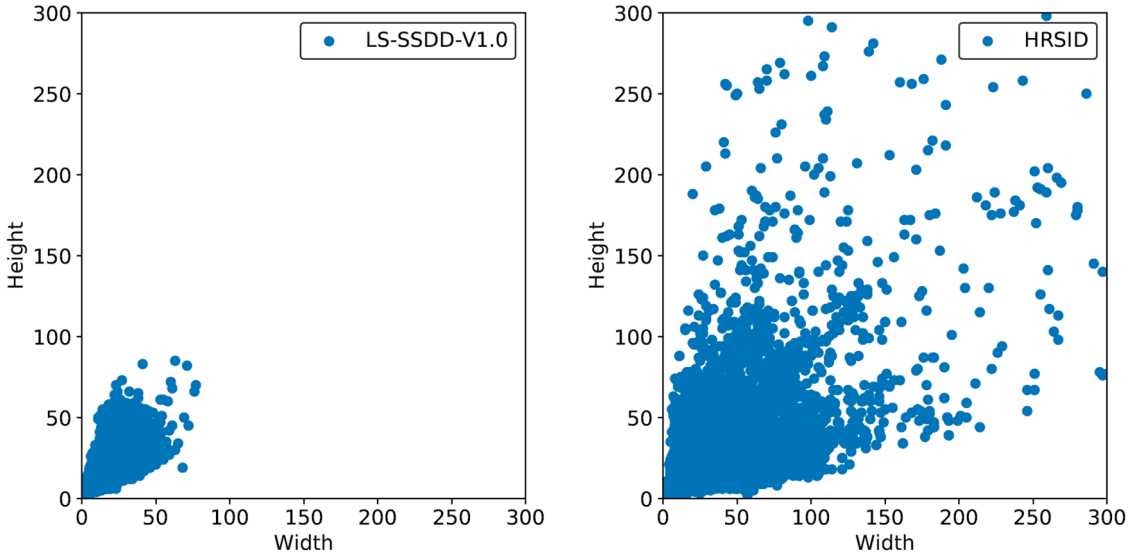
4. Experiments and Results

In this section, we conduct extensive experiments on the LS-SSDD-v1.0 [39] and the HRSID [40] dataset. First, we provide a brief introduction to the two datasets, experimental settings, and evaluation metrics. Then, a series of ablation studies are implemented to verify the effectiveness of our proposed CMA and SSJC. Last, we demonstrate the advantages of our methods compared to other state-of-the-art models through quantitative analysis and visualization results.

4.1. Dataset Description

LS-SSDD-v1.0 and HRSID are both high-resolution SAR ship detection datasets. LS-SSDD-v1.0 contains 15 large-scale SAR images of size $24{,}000 \times 16{,}000$, which are split into 9000 sub-images of size 800×800 for network training and inference. There are 6015 ship objects in total, and each object box covers an average area of 381 pixels. HRSID includes 5604 images of size 800×800, with 16,951 ship objects and an average pixel amount of 1808. Figure 3 shows the distribution of object scales in both datasets. It can be seen that the objects in the LS-SSDD-v1.0 dataset exhibit a concentrated distribution in the area of both a width < 100 pixels and height < 100 pixels, while objects in the HRSID dataset are provided with multi-scale characteristics. In order to gain a more intuitive understanding of these two datasets, we use the definition standard from the COCO dataset regulation [41] to determine the small objects, medium objects, and large objects. According to the definition, targets with rectangular box areas <0.37% proportion among the total image pixels (i.e., 640,000 owing to 800×800 image size) are regarded as small ones; a 0.37% to 3.29% proportion among the total pixels are medium ones; and areas >3.29% proportion among the total pixels are large ones. On this basis, 99.8%, 0.2%, and 0% of all ship objects in the LS-SSDD-v1.0 dataset are made up of small, medium, and large objects, while 91.42%, 7.9%, and 0.65% of all ship targets in the HRSID dataset are small, medium, and large sizes, respectively. Due to the overwhelming proportion of small object quantities, the LS-SSDD-v1.0 is used to verify the detection performance of the proposed methods for

small-scale ships in SAR images. The HRSID is leveraged to test the generalization ability of multi-scale ship objects. Table 1 shows the details of the two different datasets, including collecting satellite, location, resolution, etc.



(**a**) (**b**)

Figure 3. Distribution of object scales. (**a**) LS-SSDD-v1.0 dataset; (**b**) HRSID dataset.

Table 1. Descriptions of LS-SSDD-v1.0 and HRSID.

Parameters	LS-SSDD-v1.0	HRSID
Satellite	Sentinel-1	Sentinel-1, TerraSAR-X
Band	C	C, X
Location	Tokyo, Adriatic Sea, etc.	Houston, Sao Paulo, etc.
Resolution (m)	5 × 20	0.5, 1, 3
Polarization	VV, VH	HH, HV, VV
Image size (pixel)	800 × 800	800 × 800
Object Quantity	-	-
Total	6015	16,951
Small	6003 (99.8%)	15,510 (91.42%)
Medium	12 (0.2%)	1344 (7.9%)
Large	0 (0%)	111 (0.65%)
Object Area (pixel)	-	-
Smallest	6	3
Largest	5822	522,400
Average	381	1808

In addition, with the prior knowledge that the background in inshore-waters is much more complex than that in offshore-waters, the validation set of LS-SSDD-v1.0 is marked with entire scenes, inshore scenes and offshore scenes, to verify the adaptability of the algorithm in different background complexities.

4.2. Experiment Settings

In this article, all experiments were performed with the Linux operating system equipped with an Intel Xeon Gold 5218 CPU (Intel Corporation) and NVIDIA TITAN RTX GPU (Nvidia Corporation). Pytorch 1.7 was used for training, and CUDA 10.2 was leveraged to speed up the calculations. Experiments used a multi-scale training strategy with a scaling factor range of 0.8 to 1.2. To avoid overfitting, random rotation and flipping

were used to augment the training data. The batch size was set to 16, and the final network was obtained after 300 training epochs.

4.3. Evaluation Metrics

For a quantitative evaluation of the performance of object detectors, the evaluation metrics such as Intersection over Union (IoU), Precision (P), Recall (R), and Average Precision (AP) are the normative means. During positioning, the overlap rate of the predicted result and ground truth is the measurement of the correlation between the two; a higher degree of overlap indicates a better correlation and a more precise prediction. As shown in formula (1), the bounding box IoU is defined by the overlap rate of the predicted bounding box as the ground truth bounding box:

$$IoU = \frac{Bbox_{pd} \cap Bbox_{gt}}{Bbox_{pd} \cup Bbox_{gt}} \tag{1}$$

During classification, objects may be misjudged by the background. There are four classification results: True Position (TP), True Negative (TN), False Negative (FN), and False Positive (FP). TP denotes the amount of correctly classified positive samples; TN refers to the amount of correctly classified negative samples; FN represents the amount of missed positive samples; FP indicates the number of false alarms in the background. The precision and recall are defined by these criteria, as shown in Equations (2) and (3).

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

Based on the quantities of precision and recall, AP is defined. In the Cartesian coordinate system, if the horizontal coordinate is a recall value and the vertical coordinate is a precision value, the area under the recall–precision curve is an AP value, as shown in formula (4):

$$AP = \int_0^1 P(R)dR, \tag{4}$$

4.4. Ablation Study

To verify the effectiveness of the proposed CMA and SSJC, we conducted ablation studies on the LS-SSDD-v1.0 dataset using YOLOX as the baseline network. The results are shown in Table 2.

Table 2. Ablation study results on the LS-SSDD-v1.0 dataset.

Methods	Entire Scenes			Inshore Scenes			Offshore Scenes		
	P	**R**	**AP**	**P**	**R**	**AP**	**P**	**R**	**AP**
YOLOX	66.00	77.38	73.96	42.00	52.89	46.15	81.92	91.84	89.25
YOLOX + SE	63.65	77.54	74.29	38.16	54.36	47.59	83.22	91.24	88.91
YOLOX + CBAM	71.70	77.04	73.64	51.38	52.77	45.12	82.79	91.37	89.04
YOLOX + CA	68.01	79.65	75.71	47.07	59.12	50.82	81.86	91.77	88.98
YOLOX + CMA	75.32	79.31	76.41	57.53	57.08	51.50	84.89	92.44	90.26
YOLOX + SSJC	69.91	78.64	75.25	49.45	55.83	48.94	82.06	92.11	89.65
YOLOX + CMA + SSJC	70.44	80.36	77.23	50.29	58.66	52.26	82.77	93.18	90.94

Firstly, in order to prove the advantages of CMA, we compare it with three other commonly used attention mechanisms; namely SE, CBAM, and CA. SE is a pure channel attention mechanism that completely discards spatial information. CBAM excavates spatial

attention based on pure channel attention features. CA encodes spatial cues into channel attention. Our proposed CMA extracts coordinate-aligned spatial attention based on the coordinate-aware channel attention features.

As we can see in the results, the utilization of almost all attention methods has improved the detection performance. For example, in entire scenes, SE, CA, and CMA have brought AP gains of 0.33%, 1.75%, and 2.45%, respectively, proving the positive effect of the attention mechanism in SAR small-scale object detection. Comparing the baseline model, although SE in entire scenes has a slight recall increase of 0.16%, it shows a significant precision drop of 2.35%. This indicates that the introduction of SE does highlight more areas with real targets but the complete abandonment of spatial cues makes the positioning of interest regions inaccurate. Therefore, it also brings more false alarms, resulting in a decrease in precision. Comparing CA and SE, it can be found that in entire inshore and offshore scenes, the AP of CA is 1.43%, 3.23%, and 0.07% higher than SE, and the recall is 2.11%, 4.76%, and 0.53% higher, and precision is 4.36%, 8.91%, and 1.67% higher, separately. This implies that encoding spatial information into pure channel attention effectively gains the positioning accuracy of regions of interest. Hence, the model can not only find more true samples but also accurately identify foreground objects out of the background noise. It is worth noting that in all three scenes CBAM results in a slight performance decrease with 0.32%, 1.03%, and 0.21% in AP. It shows that the incorrect feature alignment of channel and spatial attention indeed causes the problem of information loss and performance decline. This issue may not be reflected in natural scenes but it is particularly prominent in SAR small-scale ship detection, which is sensitive to spatial position. Our proposed CMA has fixed this defect. In all three scenes, the AP of CMA is 2.27%, 6.38%, and 1.22% higher than that of CBAM, indicating that the mixed attention mechanism with spatial coordinate alignment is pivotal for small-scale ship detection. In comparison with CA, our CMA has an AP superiority of 0.7%, 0.68%, and 1.28% in three scenes, respectively. In entire scenes, the recall of CMA and CA is almost the same, while the precision of CMA is 7.31% higher. It proves that the extraction of spatial attention based on the channel attention features embedded with coordinate information can indeed further refine the spatial granularity of attention. In this way, the model can achieve a more accurate localization of interest regions and distinguish them from a complex background.

In addition, relative to the baseline model, the use of SSJC also promotes the AP by 0.82%, 2.47%, and 0.18% in entire inshore and offshore scenes, separately. To be specific, in the three scenes, precision increases by 3.91%, 7.45%, and 0.14%, while recall raises by 1.26%, 2.94%, and 0.27%. The local spatial context and global semantic correlation in SSJC aggregate the environmental clues of the adjacent areas around objects and the entire image. As we can see from the results, the multiple environmental auxiliary information does enrich the feature expression of small-scale objects and facilitate the saliency and distinguishability of small objects in the complex background. Therefore, the detector based on SSJC can not only detect more real targets with a higher recall but also correctly recognize small-scale ships from a large number of false alarms with higher precision.

By merging the two methods, we boost more significant performance gains. The AP in entire inshore and offshore scenes with 77.23%, 52,26%, and 90.94% achieves 3.27%, 6.11%, and 1.69% improvements, showing the effective combination of CMA and SSJC. It can also be seen that the performance increase in inshore scenes is obviously higher than that in offshore scenes, which can be explained by the complexity of the scenes. The simple background in offshore scenes limits the further growth of performance, while the complex inshore scenes with the land, waves, or reef bring more possibilities for algorithm optimization. Our proposed methods with remarkable performance benefits in inshore scenes prove the robustness against complex backgrounds and noise.

4.5. Comparative Experiments

To verify the advantages of the proposal, we conduct the comparative experiments on the LS-SSDD-v1.0 and HRSID dataset with other state-of-the-art methods, i.e., Faster

R-CNN [11], CenterNet [12], FCOS [13], YOLOv3 [8], YOLOv5 [9], and YOLOX [10]. Faster R-CNN is a typical two-stage detection network. CenterNet and FCOS are representative of an anchor-free strategy. YOLOv3, YOLOv5, and YOLOX are excellent models in the YOLO series of single-stage frameworks. The results are in Tables 3 and 4.

Table 3. Comparative experiment results on the LS-SSDD-v1.0 dataset. **Bold** indicates the best performance.

Methods	Entire Scenes			Inshore Scenes			Offshore Scenes		
	P	R	AP	P	R	AP	P	R	AP
Faster R-CNN	67.17	74.18	69.47	49.31	48.70	39.90	76.05	89.23	85.53
CenterNet	61.52	76.83	72.56	40.77	51.76	43.22	74.09	91.64	88.47
FCOS	53.06	75.95	69.45	31.67	49.72	34.86	67.74	91.44	87.30
YOLOv3	58.98	77.92	72.58	42.04	56.51	45.42	69.26	90.57	87.11
YOLOv5	58.76	77.88	72.90	39.40	55.15	45.43	71.24	91.30	87.76
YOLOX	66.00	77.38	73.96	42.00	52.89	46.15	81.92	91.84	89.25
Ours	**70.44**	**80.36**	**77.23**	**50.29**	**58.66**	**52.26**	**82.77**	**93.18**	**90.94**

Table 4. Comparative experiment results on HRSID dataset. **Bold** indicates the best performance.

Methods	P	R	AP
Faster R-CNN	66.87	83.43	80.46
CenterNet	66.33	86.76	83.82
FCOS	76.21	85.07	82.06
YOLOv3	83.01	90.66	89.11
YOLOv5	83.13	91.30	89.52
YOLOX	82.61	90.39	88.91
Ours	**83.24**	**91.72**	**90.85**

Results of LS-SSDD-v1.0

As shown in Table 3, our method outperforms all other state-of-the-art models. Concretely, in entire scenes, our proposal is 7.78% in AP higher than FCOS with the worst results in this experiment, and 3.27% higher than YOLOX with the second-best performance. Compared to YOLOv3, YOLOv5, and YOLOX, the proposed algorithm not only has a nearly 3% advantage in recall but also shows a superiority of 11.68% over YOLOv5 in precision. This implies that the proposal can, on the one hand, locate more targets and, on the other hand, effectively reduce the negative impact of a complex background and noise interference in SAR images, so as to achieve more accurate identification. In inshore scenes, the AP of our network is 12.36%, 9.04%, and 17.4% higher than Faster R-CNN, CenterNet, and FCOS, separately. Compared to the YOLO series, the AP is also 6–7% higher. It demonstrates that in inshore scenes with more complex backgrounds, our method has better robustness than other models. In offshore scenes, the proposal has an advantage of 1.69%AP and 5.41%AP over the second and the worst performing model, i.e., YOLOX and Faster R-CNN. This shows that our network has good adaptability in different scenes.

In order to demonstrate the advantages of the proposed algorithm more concretely, we visualize the detection results of Faster R-CNN, YOLOX and our method, as shown in Figure 4. It can be found that the ships in the LS-SSDD-v1.0 dataset are extremely small, and the background clutters indeed makes the detection difficult. However, as can be seen, our model not only improves the recall but also alleviates the false alarms, especially in inshore scenes. As shown in Figure 4a,d,g, some small ships that almost blend with nearshore objects have been accurately detected, proving the strong detection performance of our method for small-scale ships in complex backgrounds. In addition, the densely distributed ship objects are also well detected by the proposal, such as Figure 4b,e,h. When missed detection occurs in other models, our network locates almost all the objects. Besides, Figure 4c,f,i also show the ability of our proposal to suppress irrelevant false alarms.

We conduct the same comparative experiments on the HRSID dataset, and the results are shown in Table 4. It can be seen that the proposed method in this article ranks first among all other models with an AP, R, and P of 90.85%, 91.72%, and 83.24%, which are 1.33%, 0.42%, and 0.11% higher than YOLOv5 with the second best results. As we know from Figure 3, the object size distribution in the HRSID dataset is relatively wide. Therefore, the experimental results indicate that our method can not only effectively improve the detection performance of small-scale ships in SAR images, but that they also satisfied adaptability to multi-scale object detection. Meanwhile, the proposal has also been proven to have sufficient robustness and generalization ability against other datasets.

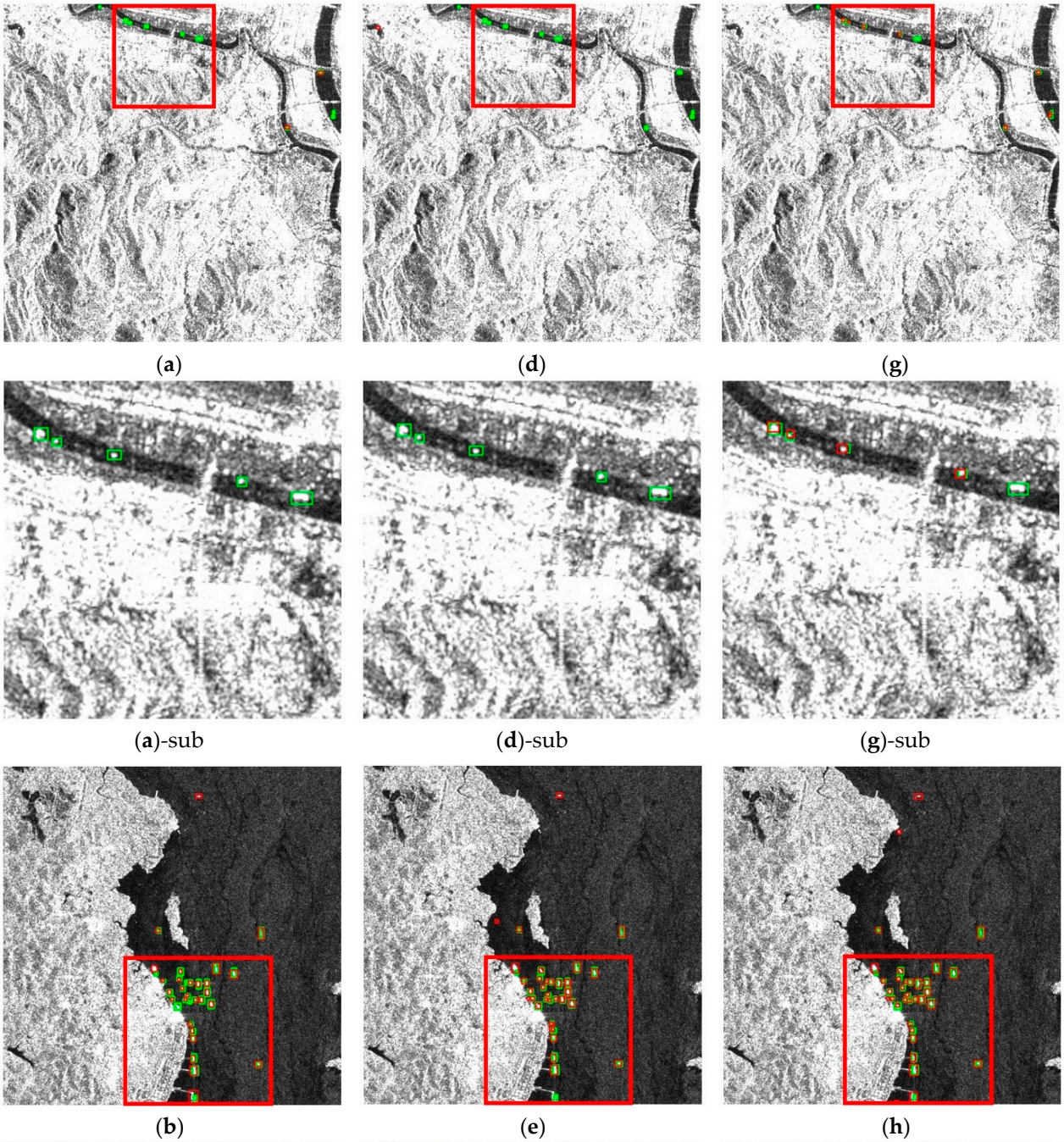


Figure 4. *Cont.*

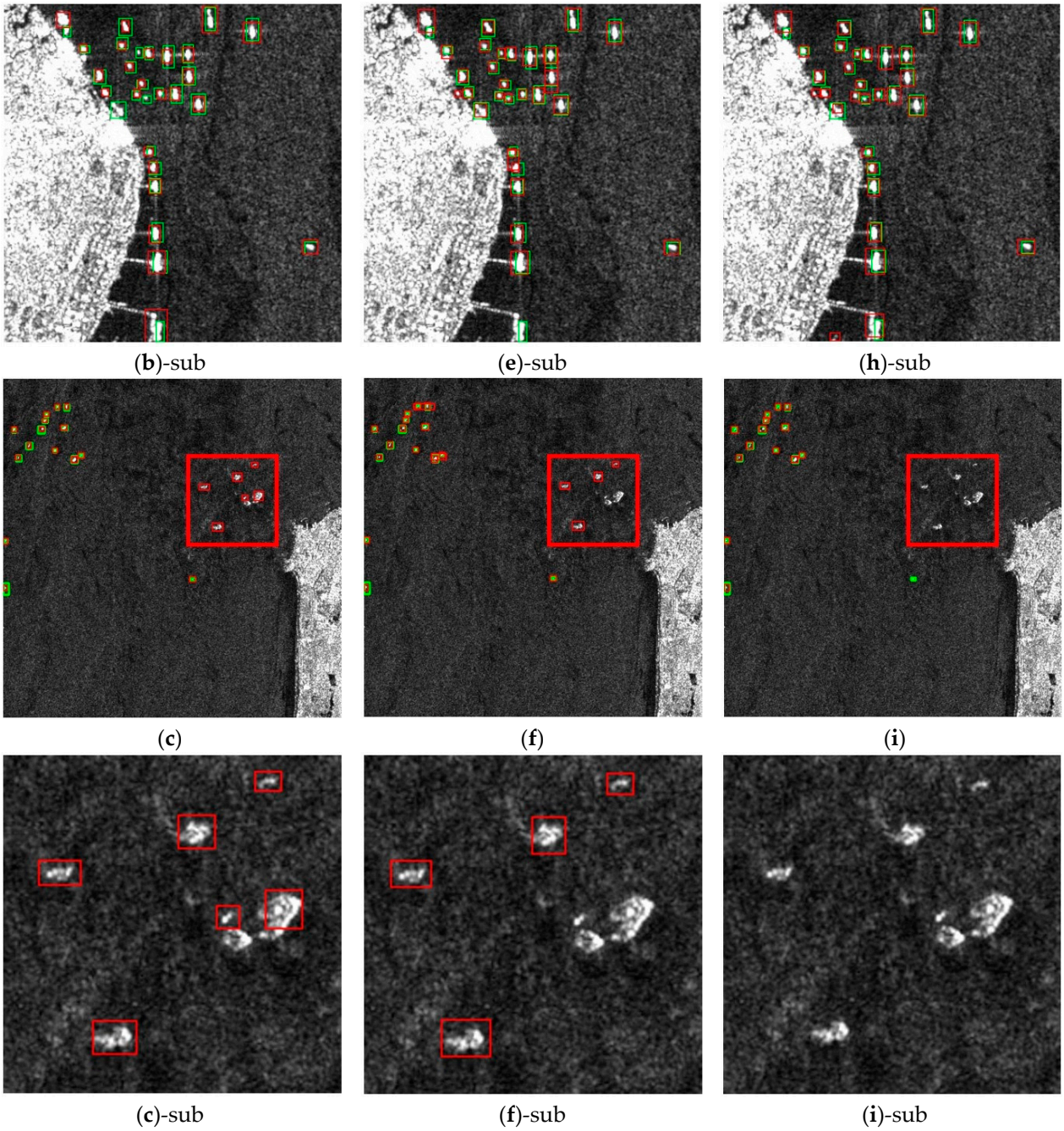


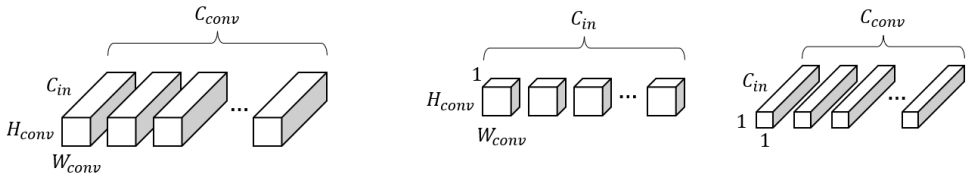
Figure 4. Visualization Results. (**a**–**c**): Faster R-CNN, (**d**–**f**): YOLOX, (**g**–**i**): our method. Image denoted as (·)-sub indicates the local enlarged drawing of its corresponding original image (·). Green box refers to ground truth, and red box represents detection result.4.5.2. Results on HRSID.

5. Discussion

In Section 4, we emphatically verified the effectiveness of the proposed method. The results show that our proposal not only alleviates the interference caused by complex backgrounds and noise in SAR images, significantly reducing the false alarms, but also enhances the performance of small-scale ship detection. Furthermore, we also test the adaptability and generalization capability of our algorithm in different environments and with multi-scale ship objects. In this section, we will focus on introducing the computational complexity and limits of this method.

The other models, such as Faster R-CNN, YOLOv3, YOLOv5, and YOLOX, adopt a standard convolution to construct the frameworks. In order to avoid a significant increase

in the calculation amount while introducing extra modules, we utilize the depthwise separable convolution. Figure 5 illustrates the computational complexity differences between standard convolution and depthwise separable convolution.



(a) Standard Convolution (b) Depthwise Separable Convolution

Figure 5. Computational Complexity Analysis. (**a**) Standard Convolution; (**b**) Depthwise Separable Convolution.

Given input feature maps with a size of $H_{in} \times W_{in} \times C_{in}$, convolutional kernels with a size of $H_{conv} \times W_{conv} \times C_{in}$ and number of C_{conv}, as well as the output feature maps in a size $C_{conv} \times H_{in} \times W_{in}$, the calculation amount of standard convolution $Amount_{standard}$ is:

$$Amount_{standard} = H_{in} \cdot W_{in} \cdot C_{in} \cdot H_{conv} \cdot W_{conv} \cdot C_{conv} \tag{5}$$

Using depthwise separable convolution, the amount of the required calculation $Amount_{depthwise}$ is:

$$Amount_{depthwise} = H_{conv} \cdot W_{conv} \cdot C_{in} \cdot H_{in} \cdot W_{in} + C_{conv} \cdot C_{in} \cdot H_{in} \cdot W_{in} \tag{6}$$

Thus, we achieve the ratio of the computational complexity of two convolutions:

$$ratio = \frac{Amount_{depthwise}}{Amount_{standard}} = \frac{1}{C_{conv}} + \frac{1}{H_{conv} \cdot W_{conv}} < 1 \tag{7}$$

From Equation (7), the depthwise separable convolution effectively reduces the amount of computation compared to standard convolution.

In order to verify the computational efficiency of our method intuitively, we compare the Floating Point Operations (FLOPs) and Parameter quantity (Params) with other state-of-the-art networks, as shown in Table 5. It can be seen that CenterNet has the smallest parameter count of 32.66 M and the lowest computational complexity of 170.74 GFLOPs. Among all YOLO models mentioned in this paper, YOLOv3 has the highest computation amount with 302.95 GFLOPs, while YOLOv5 possesses the most parameters at 76.8 M. Compared to the baseline model YOLOX, the parameters and computational complexity of our method have slightly increased by 4.8% and 1.3%, which is negligible considering the remarkable performance boost. In other words, the approach adopting depthwise separable convolution effectively introduced additional modules without the extra computational burden. However, while this method does not demonstrate outstanding computational efficiency advantages among other mainstream models, which is our current limitation, it also brings more possibilities for practical engineering work. Owing to the use of simple and general operators, the network has good potential for lightweight processing. Through routine pruning and quantization operations, our algorithm can achieve real-time inference on FPGA devices or AI chips. Nonetheless, in future work, we will strive to incorporate the concept of a lightweight design into the module as well as network construction. A fast and accurate SAR ship detection model is more adaptable for the high-speed operation of smart cities.

Table 5. Comparisons of computational efficiency. **Bold** indicates the best performance.

Methods	Params (M)	FLOPs (G)
Faster R-CNN	41.35	268.22
CenterNet	**32.66**	**170.74**
FCOS	45.66	374.74
YOLOv3	61.53	302.95
YOLOv5	76.8	174.06
YOLOX	54.15	243.13
Ours	56.74	246.36

6. Conclusions

Aiming at the problems of complex backgrounds and small object scales for ship detection in SAR images, this paper proposes a small-scale ship detection algorithm using a coordinate-aware mixed attention mechanism and spatial semantic joint context method. The coordinate-aware mixed attention mechanism integrates the coordinate-aware channel attention and the spatial attention to enhance the focusing ability on small objects and suppress the background noise. The spatial semantic joint context method effectively combines the local and global context, so as to enrich the spatial and semantic feature expression and, thus, make small objects more recognizable. The experiment results show that our proposed method has significant performance advantages in the detection of small ships in SAR images.

Author Contributions: Conceptualization, Z.J. and Y.W.; methodology, Z.J.; software, X.Z.; validation, X.Z.; formal analysis, Y.C.; investigation, D.S.; writing—original draft preparation, X.Z.; writing—review and editing, Z.J., Y.W. and H.S.; visualization, Y.C.; supervision, L.C.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by MYHT Program of China under Grant No. D040404.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dai, H.; Du, L.; Wang, Y.; Wang, Z. A Modified CFAR Algorithm Based on Object Proposals for Ship Target Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1925–1929. [CrossRef]
2. Xu, L.; Jun, H.; Xian, S.; Yi, Y. New CFAR Ship Detection Algorithm based on Adaptive Background Clutter Model in Wide Swath SAR Images. *Remote Sens. Technol. Appl.* **2014**, *29*, 75–81.
3. Huang, Y.; Liu, F. Detecting Cars in VHR SAR Images via Semantic CFAR Algorithm. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 801–805. [CrossRef]
4. Ai, J.; Yang, X.; Yan, H. Local CFAR Detector Based on Gray Intensity Correlation in Sar Imagery. In Proceedings of the IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 697–700.
5. Kaplan, L. Improved SAR target detection via extended fractal features. *IEEE Trans. Aerosp. Electron. Syst.* **2001**, *37*, 436–451. [CrossRef]
6. Charalampidis, D.; Kasparis, T. Wavelet-based rotational invariant roughness features for texture classification and segmentation. *IEEE Trans. Image Process.* **2002**, *11*, 825–837. [CrossRef]
7. Stein, G.W.; Charalampidis, D. Target detection using an improved fractal scheme. In Proceedings of the SPIE-The International Society for Optical Engineering, Orlando, FL, USA, 5 May 2006.
8. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. Available online: http://arxiv.org/abs/1804.02767 (accessed on 8 April 2018).
9. ultralytics/YOLOV5: V5.0-YOLOv5-P6 1280 Models, AWS, Supervise.ly. Available online: https://www.github.com/ultralytics/yolov5 (accessed on 8 April 2018).
10. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430. Available online: https://arxiv.org/abs/2107.08430 (accessed on 6 August 2021).
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]

12. Zhou, X.; Wang, D.; Kraehenbuehl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850. Available online: http://arxiv.org/abs/1904.07850 (accessed on 25 April 2019).
13. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019.
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
15. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
18. Shi, H.; Fang, Z.; Wang, Y.; Chen, L. An Adaptive Sample Assignment Strategy Based on Feature Enhancement for Ship Detection in SAR Images. *Remote Sens.* **2022**, *14*, 2238. [CrossRef]
19. Sun, Z.; Meng, C.; Cheng, J.; Zhang, Z.; Chang, S. A Multi-Scale Feature Pyramid Network for Detection and Instance Segmentation of Marine Ships in SAR Images. *Remote Sens.* **2022**, *14*, 6312. [CrossRef]
20. Gao, F.; He, Y.; Wang, J.; Hussain, A.; Zhou, H. Anchor-free Convolutional Network with Dense Attention Feature Aggregation for Ship Detection in SAR Images. *Remote Sens.* **2022**, *12*, 2619. [CrossRef]
21. Zhu, C.; Zhao, D.; Liu, Z.; Mao, Y. Hierarchical Attention for Ship Detection in SAR Images. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October2020; pp. 2145–2148.
22. Guo, Y.; Chen, S.; Zhan, R.; Wang, W.; Zhang, J. SAR Ship Detection Based on YOLOv5 Using CBAM and BiFPN. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2147–2150.
23. Zhang, L.; Chu, Z.; Zou, B. Multi Scale Ship Detection Based on Attention and Weighted Fusion Model for High Resolution SAR Images. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 631–634.
24. Cui, Z.; Li, Q.; Cao, Z.; Liu, N. Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8983–8997. [CrossRef]
25. Kang, M.; Ji, K.; Leng, X.; Lin, Z. Contextual Region-Based Convolutional Neural Network with Multilayer Fusion for SAR Ship Detection. *Remote Sens.* **2017**, *9*, 860. [CrossRef]
26. Zheng, C.; Zhang, Y.; Hu, H.; Wu, Y.; Huang, G. Object detection enhanced context model. *J. Zhejiang Univ. (Eng. Sci.)* **2017**, *54*, 529–539.
27. Lim, J.; Astrid, M.; Yoon, H.; Lee, S. Small Object Detection using Context and Attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 181–186.
28. Yao, C.; Bai, L.; Xue, D.; Lin, X.; Ye, Z.; Wang, Y.; Yin, K. GFB-Net: A Global Context-Guided Feature Balance Network for Arbitrary-Oriented SAR Ship Detection. In Proceedings of the 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, 26–28 July 2022; pp. 166–171.
29. Zou, B.; Qin, J.; Zhang, L. Vehicle Detection Based on Semantic-Context Enhancement for High-Resolution SAR Images in Complex Background. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
30. Yue, B.; Han, S. A SAR Ship Detection Method Based on Improved Faster R-CNN. *Comput. Mod.* **2019**, *9*, 90–101.
31. Zhang, Y.; Zhu, W.; Wu, X. Target Detection Based on Fully Convolutional Neural Network for SAR Images. *Telecommun. Eng.* **2018**, *58*, 1244–1251.
32. Fu, X.; Wang, Z. SAR Ship Target Rapid Detection Method Combined with Scene Classification in the Inshore Region. *J. Signal Process.* **2020**, *36*, 2123–2130.
33. Liu, L.; Chen, G.; Pan, Z.; Lei, B.; An, Q. Inshore ship detection in SAR images based on deep neural networks. In Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 25–28.
34. Chen, P.; Li, Y.; Zhou, H.; Liu, B.; Liu, P. Detection of Small Ship Objects Using Anchor Boxes Cluster and Feature Pyramid Network Model for SAR Imagery. *J. Mar. Sci. Eng.* **2020**, *8*, 112. [CrossRef]
35. Hu, C.; Chen, C.; He, C.; Pei, H.; Zhang, J. SAR Detection for Small Target Ship Based on Deep Convolutional Neural Network. *Chin. J. Inert. Technol.* **2019**, *27*, 397–405.
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
38. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.

39. Zhang, T.; Zhang, X.; Ke, X.; Zhan, X.; Shi, J.; Wei, S.; Pan, D.; Li, J.; Su, H.; Zhou, Y.; et al. LS-SSDD-v1.0: A Deep Learning Dataset Dedicated to Small Ship Detection from Large-Scale Sentinel-1 SAR Images. *Remote Sens.* **2020**, *12*, 2997. [CrossRef]
40. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
41. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Cham, Germany, 6–12 September 2014; pp. 740–755.