remote sensing

MDPI

Article

BoxPaste: An Effective Data Augmentation Method for SAR Ship Detection

Zhiling Suo, Yongbo Zhao *, Sheng Chen and Yili Hu

National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China
* Correspondence: ybzhao@xidian.edu.cn

Abstract: Data augmentation is a crucial technique for convolutional neural network (CNN)-based object detection. Thus, this work proposes BoxPaste, a simple but powerful data augmentation method appropriate for ship detection in Synthetic Aperture Radar (SAR) imagery. BoxPaste crops ship objects from one SAR image using bounding box annotations and pastes them on another SAR image to artificially increase the object density in each training image. Furthermore, we dive deep into the characteristics of the SAR ship detection task and draw a principle for designing a SAR ship detector—light models may perform better. Our proposed data augmentation method and modified ship detector attain a 95.5% Average Precision (AP) and 96.6% recall on the SAR Ship Detection Dataset (SSDD), 4.7% and 5.5% higher than the fully convolutional one-stage (FCOS) object detection baseline method. Furthermore, we also combine our data augmentation scheme with two current detectors, RetinaNet and adaptive training sample selection (ATSS), to validate its effectiveness. The experimental results demonstrate that our newly proposed SAR-ATSS architecture achieves 96.3% AP, employing ResNet-50 as the backbone. The experimental results show that the method can significantly improve detection performance.

Keywords: synthetic aperture radar; ship detection; data augmentation; target detection

Citation: Suo, Z.; Zhao, Y.; Chen, S.; Hu, Y. BoxPaste: An Effective Data Augmentation Method for SAR Ship Detection. *Remote Sens.* **2022**, *14*, 5761. https://doi.org/10.3390/rs14225761

Academic Editor: Gerardo Di Martino

Received: 28 September 2022
Accepted: 9 November 2022
Published: 15 November 2022

1. Introduction

Monitoring and identifying marine ships is a crucial task guaranteeing national security. Specifically, it plays a vital role in monitoring and managing fishing ships, combating smuggling, and protecting marine resources [1]. Synthetic Aperture Radar (SAR) is an appropriate sensor for ship detection [2–4] because it can create high-resolution images (Figure 1), regardless of the altitude and weather conditions, making ship detection a computer vision task.

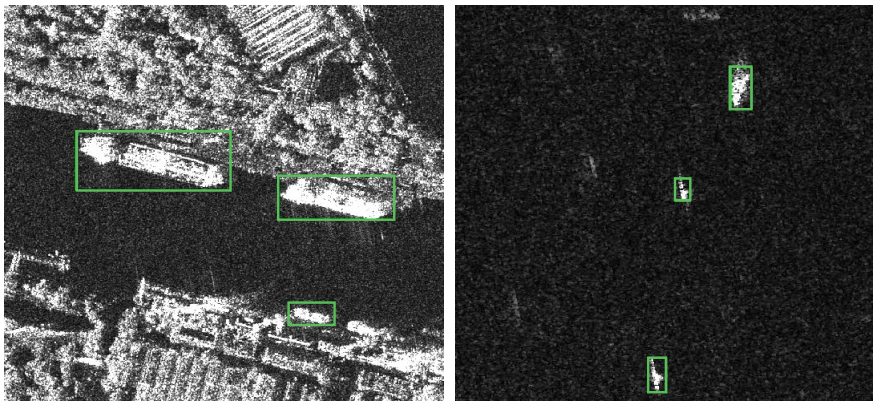


Figure 1. Synthetic Aperture Radar (SAR) ship detection images from the SAR Ship Detection Dataset (SSDD) [5]. Green boxes are ground-truth labels manually annotated.

In the past decades, the literature has suggested several SAR ship detection methods, mainly divided into two categories: traditional and convolutional neural network (CNN)-based methods. In the former category, the Constant False-Alarm Rate (CFAR) algorithm and its variants are the primary representative techniques of traditional SAR ship detection [6–10]. Technically, this approach establishes a threshold to identify targets statistically exceeding the background pixel level while retaining a low false alarm rate. However, traditional algorithms are not robust to light and weather condition variations. The second category involves CNNs, which have recently presented great success in object detection [11–17]. Employing CNNs for SAR ship detection is also becoming a trend, with [18] designing a modified faster region-based CNN (R-CNN) scheme involving a densely connected network to solve the scale variance issue in SAR ship detection. Furthermore, [19] introduces a R-CNN to detect ships within SAR imagery. The issue of small ship detection is solved by aggregating contextual features from different layers and achieving improved performance. Commonly, the detection speed of ships within a SAR image is often neglected and thus [20] suggests a lightweight network with fewer parameters by mainly using depthwise separable CNN (DS-CNN) to achieve high-speed SAR ship detection. Although the detection performance and speed continuously improved, their scopes are limited to modifying the network structure.

Opposing previous works, this paper notices that the statistical characteristics of SAR ship data are significantly different from the general object detection data. Considering the SAR Ship Detection Dataset (SSDD) [5] as an example, we first calculate the number of images *with regard to* the number of ships (Figure 2a). Figure 2a highlights that the dataset involves more than 700 images, each of which contains only one ground-truth target. The average number of ground-truth targets per image is 2.19, indicating that objects are sparsely distributed (see also Figure 2a), as most pixels in SAR imagery are background, revealing its relatively low information density. We also calculate the size of the ships in pixels. The corresponding results in Figure 2b indicate that the areas occupied by most ships are smaller than 2500 px (around 50×50), which is about 0.95% of the picture, highlighting that SAR imagery objects are very small compared to other general objects.

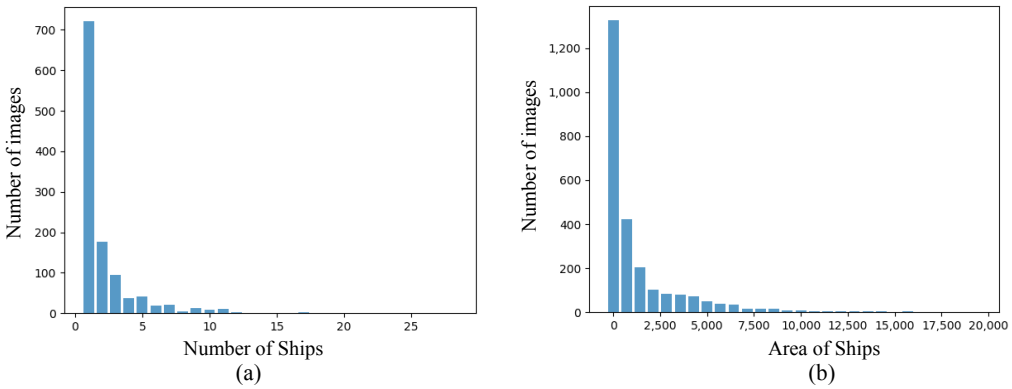


Figure 2. Statistics of the SSDD dataset. (**a**) is the number of images *with regard to* the number of ground-truth ships. (**b**) shows the number of ships *with regard to* the area of ground-truth ships.

Previous research [21] demonstrates that training an object detector is primarily a learning process to identify the objects of interest. Therefore, increasing the object density in each image should be beneficial for detection performance. Knowing these statistical characteristics in a SAR ship detection task, we propose a simple but effective data augmentation method named BoxPaste that increases the object density in each SAR image. Concretely, during training, we crop objects from one image using bounding box annotations and paste them into another SAR image. Opposing current data augmentation methods such as random flipping and color jittering, the proposed BoxPaste is specially designed to increase the number of ground-truth ships per image and ultimately enhance training efficiency. Our experiments demonstrate that BoxPaste greatly improves the detection performance on

anchor-free (Fully Convolutional One-Stage (FCOS) [15]) and anchor-based (RetinaNet [14]) detectors by 4.7% and 3.2% AP, respectively.

In addition, SAR ship detection is a single-class detection task. However, object detectors such as RetinaNet, FCOS, and Faster R-CNN [11] are designed for general object detection tasks, and thus directly applying those detectors to SAR imagery leads to overfitting. Therefore, we also introduce a principle for designing an appropriate SAR ship detector governed by the concept that the larger model is not always the better. Following this principle, we modify the well-known anchor-free object detector FCOS and develop its lighter variant entitled SAR-FCOS. The latter is twice as fast as its FCOS baseline and achieves a better detection performance, demonstrating the effectiveness of our modification.

In summary, our contributions from this work are three-fold:

- Proposing BoxPaste, an easy but powerful data augmentation strategy for SAR ship detection.
- Developing a principle to design a SAR image detector and proposing a modified detector SAR-FCOS.
- Combining the two previous contributions to achieve a great detection performance on the SSDD dataset employing ResNet-50 [22] as the backbone.

The rest of this article is organized as follows. The second section introduces the related work of SAR ship detection. Section 3 details the proposed methods, including SAR-FCOS and our proposed BoxPaste data enhancement strategy. In Section 4, the experimental results and corresponding analysis are provided, and some conclusions are made in Section 5.

2. Related Works

2.1. Traditional Methods

The most widely used traditional ship detection algorithms are the CFAR algorithm and its improved variants [6–10]. These methods set a threshold to detect statistically significant targets exceeding the background pixel while maintaining a constant false alarm rate. Concretely, [6] proposes a technique that computes the cross-correlation values between two images extracted by sliding a small-sized window on the multi-view SAR intensity (or amplitude) imagery, producing a coherent image. Furthermore, ref. [7] suggests a new CFAR-based ship detection algorithm that considers the normal distribution of two-dimensional joint logs, while [8] develops a ship detection method based on feature analysis for high-resolution SAR images. The author of [9] introduces a bilateral CFAR algorithm for ship detection in SAR images, reducing the influence of SAR ambiguities and sea clutter by combining the SAR images' intensity and spatial distribution. Due to the high similarity between the harbor's and ship's body gray and texture features, traditional methods cannot effectively detect inshore ships. Thus, ref. [10] presents a novel saliency and context information approach dealing with this issue. Since CFAR and its improved variants severely rely on the preset distribution or manually defined characteristics, their adaptive ability is weak.

2.2. Deep Learning-Based Methods

With the development of deep learning technology and the establishment of a SAR ship database [5,23,24], many ship detection algorithms based on convolutional neural networks have emerged. For example, ref. [25] introduces Faster R-CNN for ship detection in SAR imagery and solves the issue of small ship detection by aggregating contextual features from different layers, achieving improved performance. The work of [26] introduces a new network architecture, named You Only Look Once version2 (YOLOv2)-reduced, which has a lower detection time than YOLOv2 [27] on an NVIDIA TITAN X GPU. Aiming at the problem that the detection speed of SAR ships is often neglected at present, a brand-new lightweight network [20] is established with fewer network parameters by mainly using DS-CNN to achieve high-speed SAR ship detection, which can achieve high-speed and accurate ship detection simultaneously compared with other methods. In [28], the authors develop a

novel ship detection method based on a high-resolution ship detection network (HR-SDNet) appropriate for high-resolution SAR images. This method is more accurate and robust for inshore and offshore ship detection of high-resolution SAR imagery. A two-staged detector named Attention Receptive Pyramid Network (ARPN) [29] is suggested to improve detecting multi-scale ships in SAR images by enhancing the relationships among non-local features and refining information at different feature maps. This strategy is effective for scenes of various sizes and complex backgrounds. To alleviate the excessive computational burden and increased hyper-parameter cardinality problems, ref. [30] suggests an efficient and low-cost ship detection network for SAR imagery. This work utilizes an anchor-free SAR ship detection framework comprising a bounding box regression sub-net and a score map regression sub-net based on a simplified U-Net. This pipeline achieves a very competitive detection performance while being extremely lightweight. An improved algorithm based on CenterNet [31] has also been proposed [32] that is significantly better than CenterNet for small ship detection in low-resolution SAR imagery, adding low-level feature representation to the pyramids for small object detection and optimizing the head of detector to effectively distinguish foreground from background. Finally, ref. [33] introduces an anchorless convolution network aggregating an intensive attention function that obtains higher precision and is faster to execute than the mainstream detection algorithms.

2.3. Data Augmentation for Object Detection and Instance Segmentation

Reference [34] leverages segmentation annotations to increase the number of object instances by appropriately modeling the visual context surrounding objects. The work of [35] automatically extracts object instance masks and renders them on random background images. Mixup [36] randomly extracts two images from the training set and then performs a linear weighted summation of the pixel values of the extracted image data. At the same time, the One-hot vector labels corresponding to the samples are also weighted and summed. In this way, a new image with a fuzzy classification boundary can be obtained, enhancing the generalization ability of the model. CutMix [37] replaces the removed regions with a patch from another image and changes the ground truth labels by the number of pixels of the combined images. By requiring the model to recognize the target from a local perspective, the localization ability can be enhanced. CutMix is usually used for classification tasks and is not suitable for detection tasks because it usually crops image patches randomly, which requires that the image does not contain too much context. While [38] highlights that CopyPaste, i.e., simply pasting objects randomly, provides solid gains on the detectors' performance. While being similar to BoxPaste in this work, we argue that our work is the first to migrate the key ideology of CopyPaste neglecting the unnecessary usage of instance mask annotations in the SAR ship detection task.

3. Methods

This section first revisits FCOS [15], a well-known one-staged anchor-free object detector, which will be used as our baseline in this paper. Considering that objects in a SAR ship detection scenario are statistically small and sparse, we lighten the structure of FCOS from the Feature Pyramid Networks (FPN) [39] to the detection head and suggest SAR-FCOS. Finally, this section introduces our proposed BoxPaste, a powerful data augmentation strategy for SAR ship detection.

3.1. Revisiting FCOS

Although anchor-based object detectors have achieved massive success on many object detection datasets, they suffer from the requirement to design anchor boxes. Furthermore, the enormous amount of detection proposals dramatically slows down the post-processing method, prohibiting the anchor-based mechanism from real-time applications. Recently, anchor-free object detectors have shown great potential in general object detection. They usually attain higher performance than their counterpart anchor-based detectors while

enjoying a more straightforward architectural design. Hence, this work considers the well-known anchor-free object detector FCOS as the baseline method.

Like other standard object detectors, FCOS comprises three parts: a backbone for feature extraction, FPN for feature integration, and a detection head for prediction. Figure 3 illustrates the overall FCOS structure. As the backbone structure, Visual Geometry Group (VGG) [40], ResNet [22], Inception [41], or any other well-known architectures designed for classification can be exploited.

FPN comprises a sequence of top-down layers and several shortcut layers to combine the knowledge encoded at different layers, which is also a broadly used structure in object detection. The predicted results encoded by the detection head of FCOS are different from other anchor-based object detectors. Unlike RetinaNet, Single Shot MultiBox Detector (SSD) [13], YOLOv3 [12], and Faster R-CNN, which use anchor boxes, FCOS directly views location points as training samples and learns to predict the four offsets from each location to the bounding boxes, i.e., left, top, right, and bottom (l^*, t^*, r^*, b^*). Concretely, the bounding box regression targets for location (x, y) is defined as:

$$\begin{aligned} l^* &= x - x_0, r^* = x_1 - x, \\ t^* &= y - y_0, b^* = y_1 - y, \end{aligned} \tag{1}$$

where (x_0, y_0), (x_1, y_1) are the coordinates of the left-top and right-bottom corners of the ground-truth bounding box. In addition to the regression prediction, one also needs to know each location's category. Regarding classification, if a location falls into the ground-truth bounding box, that location is considered a positive sample and is responsible for predicting that ground truth. Moreover, the FPN in the standard FCOS typically contains five levels, i.e., from P3 to P7, while to construct valid receptive field scales for the neurons at different FPN levels, different FPN levels are regressing different objects sizes. Given the regression targets l^*, t^*, r^*, and b^* for a location, the object scale per feature pyramid level follows the following constraint:

$$\begin{aligned} max(l^*, t^*, r^*, b^*) &> m_i, \\ max(l^*, t^*, r^*, b^*) &< m_{i+1}, \end{aligned} \tag{2}$$

where $m_i, i \in 2, 3, 4, 5, 6, 7$ in the original FCOS are set as 0, 64, 128, 256, 512, and ∞, respectively.

To suppress the low-quality predicted bounding boxes generated by the locations far away from the objects' center, FCOS adopts the centerness branch. The centerness score of each location x_i, y_i and the corresponding bounding box at feature level i is defined as:

$$centerness^* = \sqrt{\frac{min(l^*, r^*)}{max(l^*, r^*)} \times \frac{min(t^*, b^*)}{max(t^*, b^*)}}. \tag{3}$$

The centerness score is then multiplied by the classification score to provide the final predicted confidence used by the Non-Maximum-Suppression (NMS). It should be noted that the pre-processing (e.g., normalizing images) and post-processing (e.g., decoding outputs, NMS) methods of FCOS are the same as in other standard object detectors.

3.2. SAR-FCOS

Unlike general object detection tasks, objects in SAR ship detection scenarios usually have two features: the objects' sizes are relatively small, and there is only one object category. Such features require designing a new network structure, and thus this section simplifies the FCOS structure and proposes SAR-FCOS. Compared to the original FCOS, the complexity of SAR-FCOS is severely reduced, aiming at preventing the network from overfitting. Specifically, we only modify the FPN and detection head structure, as these two parts are the main difference between detection and performing other tasks such as image

classification and segmentation. Moreover, we wish to emphasize that our motivation is not to design a lightweight detector for ship detection in SAR images but to highlight the critical merit of designing an appropriate detector for the SAR ship detection task.

3.2.1. Light FPN

In [39], the authors developed the FPN to handle the large-scale variance in general object detection tasks. The critical insight behind FPN is that the neurons' valid receptive fields at the deep layers are significant, and at the shallow layers, these are typically small. Hence classifying different object sizes on different feature pyramid levels can benefit from the scale alignment between the objects' sizes and neurons' valid receptive fields.

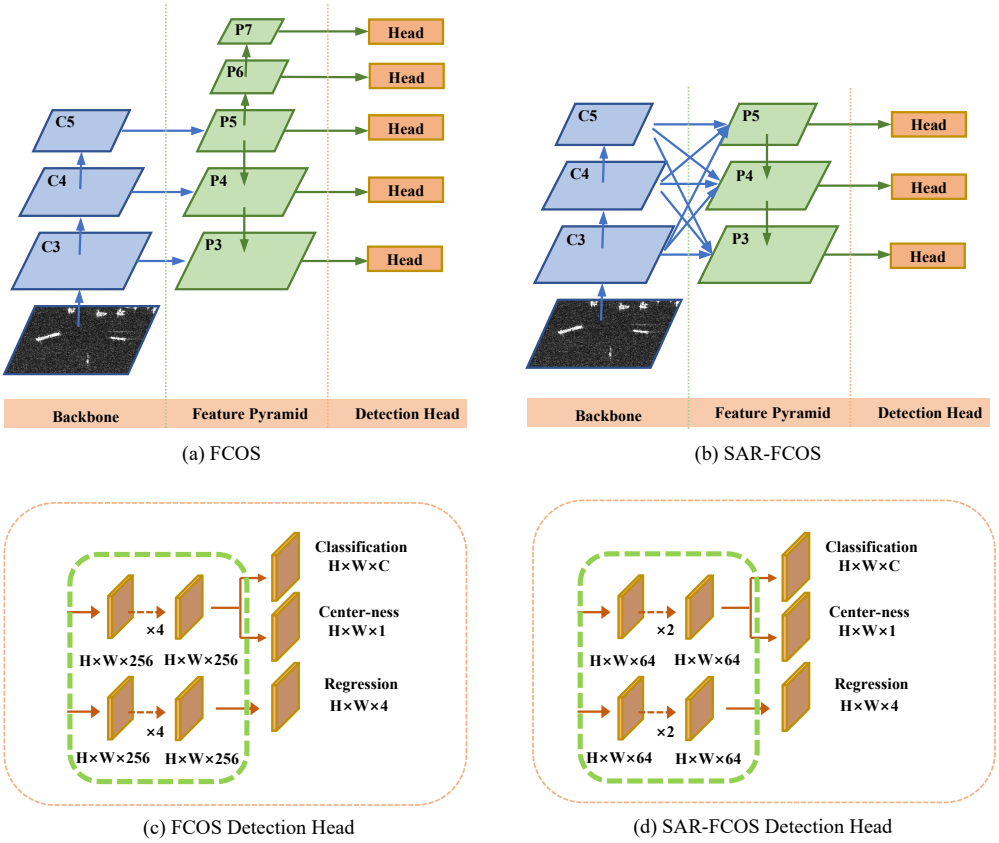


Figure 3. A detailed illustration of the structural difference between the Fully Convolutional One-Stage (FCOS) and SAR-FCOS.

However, as stated in Section 1, the ships' scales in a SAR detection task are small. Hence, by following the assigning rule of FCOS, only a few large objects will be assigned to P6 and P7, i.e., the feature maps P6 and P7 have little contribution to the accuracy during testing but impose a high computational cost. Therefore, P6 and P7 are redundant, and we remove both these layers and their corresponding detection head. Our model involves fewer FPN layers, and, therefore, our light FPN executes faster, exploiting some extra modules to enhance the feature integration between the left layers. Knowing that the ground-truth objects are mainly assigned to the P3, P4, and P5 layers, we use the Adaptive Spatial Feature Fusion (ASFF) [42] to fuse different knowledge encoded between different FPN layers.

3.2.2. Light Detection Head

The original structure and hyper-parameters are designed for the Common Objects in Context (COCO) [43] benchmark containing 80 categories of various objects. However, there is only one Ship category in the SSDD dataset, making the number of head channels

(CHead) and head blocks (NHead) redundant in the original network. Therefore, it is necessary to reduce the complexity of the detection head to prevent over-fitting. Our trials involve extensive experiments with different head parameters, and specifically, we reduce the CHead and NHead from 256/4 to 64/2, affording a better trade-off between latency and detection performance. Despite our modifications being simple, these are not trivial. In the experimental section, we demonstrate that the simplified detection head reduces the detector's complexity and surprisingly improves the detection performance, indicating the severe over-fitting of the original FCOS for the SAR ship detection tasks. Such a phenomenon reveals the importance of designing specific model structures and hyper-parameters for a specific task. Our modified FCOS is named SAR-FCOS and is illustrated in Figure 3.

3.3. BoxPaste

3.3.1. Revisiting CopyPaste

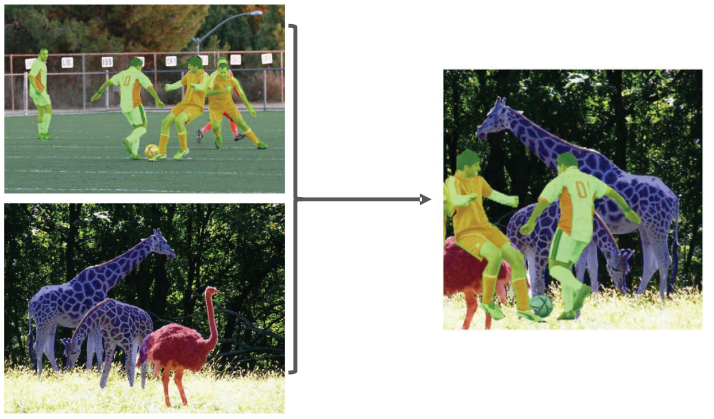
Data augmentation aims to increase the training dataset's variability, a critical component during object detector training, leading to significant improvements in object detection tasks. The most recent and effective augmentation method is CopyPaste. By randomly cropping objects from image A utilizing ground-truth masks and pasting them on image B, CopyPaste creates more training samples and increases the number of ground-truth samples per mini-batch, which are crucial for training object detectors [21]. A simple illustration of CopyPaste is presented in Figure 4. CopyPaste affords a remarkable performance gain on instance segmentation tasks, while additionally, it also significantly improves the performance of object detection. Spurred by the advantages of CopyPaste, naturally, the following question is raised: can we bring CopyPaste into the SAR ship detection task?

3.3.2. BoxPaste

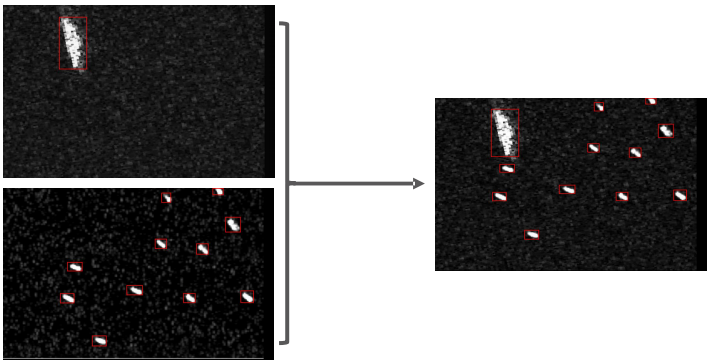
A straightforward method is applying CopyPaste without instance segmentation masks and employing bounding boxes to crop objects. However, the objects' scales in general object detection tasks vary greatly, and thus, the proposed method will lead to heavy occlusion between the original and pasted objects.

However, there are three properties of SAR ship detection tasks (considering the SSDD dataset as an example). (1) Most of the ships are small. (2) The number of target ships in each image is quite limited. Therefore, most pixels in an image are background, making each mini-batch less informative. (3) The diversity of backgrounds for SAR images is extremely less than natural RGB images; hence, SAR images can be easily converted to gray-scale images. Unlike general object detection tasks, these three features suggest that using bounding boxes to perform CopyPaste can create more realistic training samples. We name this method BoxPaste. A clear illustration of the difference between applying BoxPaste in SAR ship detection and general object detection is shown in Figure 4.

In the following experiment, we demonstrate that although the proposed data augmentation scheme BoxPaste is simple, it substantially improves the SAR ship detection performance, revealing its great value. To provide a clear image of the training samples created by BoxPaste, we present more examples in Figure 5.



(a) CopyPaste



(b) BoxPaste

Figure 4. An illustration of CopyPaste [38] and our proposed BoxPaste. (**a**) is borrowed from the original paper. (**b**) shows a combination of two training images in the SSDD dataset. Note that for CopyPaste, ground-truth instance masks are required. However, applying BoxPaste in the SAR ship detection task only requires the bounding box annotations.

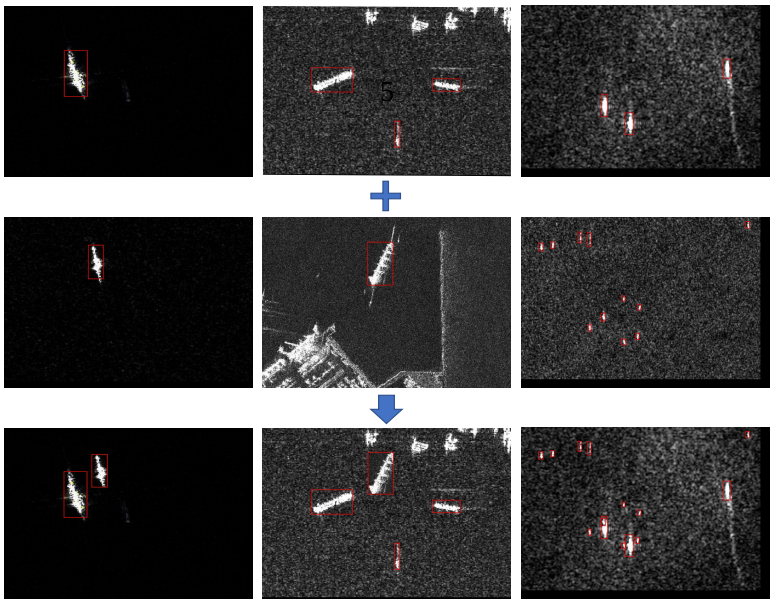


Figure 5. Examples of the created training images by BoxPaste. By copying the bounding boxes of the middle image and pasting them into the top image, we get the bottom image. One can see that BoxPaste can effectively increase the number of ground-truth ship objects per image.

4. Experiment

This section evaluates the effectiveness of our proposed SAR-FCOS network and the BoxPaste data augmentation method on the SSDD dataset.

4.1. Dataset and Training Details

SSDD [5] is a public dataset for SAR ship detection. It contains 1160 images presenting 2540 ships. As we mentioned at the beginning of this paper, the ship objects in this dataset are sparsely distributed, while according to our calculation, more than 700 images contain only one object, i.e., over 60% of all images. The average number of ships per image is 2.19. We also measure the overlap between ships using the Intersection over Union (IoU) metric. The SSDD has an average IoU involving overlapping ships of 0.0048, illustrating that the dataset's objects are sparse. We divide the training and test set according to the way the author of the SSDD dataset divided; that is, the training set and the test set were split according to the images' name, i.e., images with their name ending with 1 or 9 belong to the test set, while the remaining images belong to the training set. To evaluate the detector's performance, we adopt average precision (AP) and recall [44]. Precision is the proportion of accurately predicted ships in all forecasts, and recall is the proportion of accurately predicted ships in all ground-truth ships, both of which are the most widely used indicators. The AP metric is used to evaluate the comprehensive performance of the detector. It can be obtained by sorting the output results in descending order of the detection confidence, drawing the precision-recall curve, and calculating the area of the curve. In this paper, AP is calculated when the IoU threshold is 0.5.

Because our work focuses on the network designing strategy and data augmentation method, which is able to be applied to every detection algorithm, we choose FCOS as a simple baseline. The trials combine the proposed FCOS with ResNet-50 as our baseline backbone network, utilizing the pre-trained ones from ImageNet [45] as initial parameters. All models are trained on one NVIDIA GeForce GTX 1080Ti involving a stochastic gradient descent (SGD) for 12 epochs. The initial learning rate is 0.01, which reduces by 10 at the 8th and 11th epochs. The weight decay factor and momentum are set to 0.0001 and 0.9, respectively, while the input images are resized to [512, 512]. In BoxPaste, both the cropped image patches from the original images and the target images are randomly flipped horizontally at a ratio of 50%.

4.2. Ablation Study

4.2.1. Experiments on SAR-FCOS

Initially, we investigate the performance of SAR-FCOS from the perspective of evaluating the light detection head, light FPN, and the ASFF feature integration module. Table 1 highlights that the FCOS baseline achieves 90.8% mAP at 53.1 frames per second (FPS). Then, we reduce the number of the convolution layers in the detection head, i.e., NHead, from four to two, and find that mAP reduces only by 0.4% while affording a 13 FPS improvement, which is an acceptable trade-off between performance and latency. After that, we reduce the number of channels in both the detection head and FPN, i.e., CHead, from 256 to 128. As expected, reducing the parameters alleviates over-fitting, increasing the detection performance to 91.1% mAP.

Further reducing the CHead to 64 and 32 indicates that for CHead = 64 we obtain the best performance of 92.4% mAP at 101.2 FPS. These modifications highlight that altering only a few hyper-parameters in the detection head almost doubles the SAR ship detection speed while increasing performance by 1.6% mAP compared to the baseline network. Such a phenomenon strongly validates our design intuition for an appropriate SAR ship detector. Namely, the SAR ship detection dataset is easy to get over-fitted, therefore, it is better to use light models than heavy models. Note that we did not try different backbones because our goal is to express the key ideology of designing appropriate detectors for SAR ship detection, not to thoroughly explore every combination of different components in detectors.

Table 1. Ablation study on the light detection head and the light Feature Pyramid Network (FPN) in SAR-FCOS.

	NHead	CHead	FPN	AP (%)	Recall (%)	FPS
FCOS (baseline)	4	256	P3–P7	90.8	91.1	53.1
Light Head	2	256	P3–P7	90.4	90.5	66.3
	2	128	P3–P7	91.1	91.5	96.3
	2	64	P3–P7	92.4	92.4	101.2
	2	32	P3–P7	89.2	89.4	102.8
Light FPN	2	64	P3–P5	92.3	91.9	119.9
	2	64	P3–P4	90.1	90.3	127.6
Light Head + ASFF [42]	2	64	P3–P5	93.0	94.4	110.1

As we stated in Section 3, ship objects in the SSDD dataset are relatively small, and most of them are assigned to P3 to P5 levels during the FCOS label assignment. Hence, P6 and P7 levels contribute less to the final performance. The last three rows in Table 1 present our ablation study involving different FPN levels. Specifically, by removing the P6 and P7 levels, mAP drops by only 0.1%, but FPS increases by nearly 20. However, further removing the P5 level, the detection performance drops by 2.2%, indicating that the P5 level is essential. Affording a detector that is twice as fast as the baseline, we have the option to add extra modules such as ASFF or attention mechanisms [46,47]. This work enhances feature integration by adopting ASFF because the parameter reduction occurs in the FPN and detection head. Indeed, the last row in Table 1 highlights that by employing ASFF, our final SAR-FCOS model attains 93.0% mAP at 110.1 FPS on the SSDD dataset, which is faster and more robust than the original FCOS baseline.

4.2.2. Ablation Study on BoxPaste

We also investigate the effect of BoxPaste from two aspects: scale jittering and total training epochs. Performing scale jittering on two combined images is borrowed from the original CopyPaste paper. For example, if the scale jittering range is [0.1, 2], i.e., the sampled image's size is $[int(512 \times 0.1), int(512 \times 2)]$, we perform, if necessary, padding and cropping to align the image size to 512. It is well known that the greater the data augmentation, the longer the convergence time during training. Hence, to explore the upper bound of our BoxPaste on SAR-FCOS, we explore a various number of training epochs. Table 1 shows that our SAR-FCOS model achieves 93.0% AP and 94.4% recall. However, when applying BoxPaste, the AP metric improves by 1.1%, reaching 94.1% (Table 2). If we apply scale jittering from 0.5 to 1.5, the mAP further improves to 94.5%, and for scale jittering within the range [0.1, 2], our model achieves 94.6% AP and 95.5% recall. It should be noted that BoxPaste is only used during training. Hence, the testing FPS is not affected.

Then we investigate the effects of the total training epochs. The results demonstrate that training the detectors for 36 epochs yields the best results, while when exceeding 36 epochs, the model overfits, reducing performance. Figure 6 visualizes the detection results to provide an intuitive understanding of the improved performance, highlighting that the SAR-FCOS model combined with BoxPaste yields fewer missed detections and higher recall.

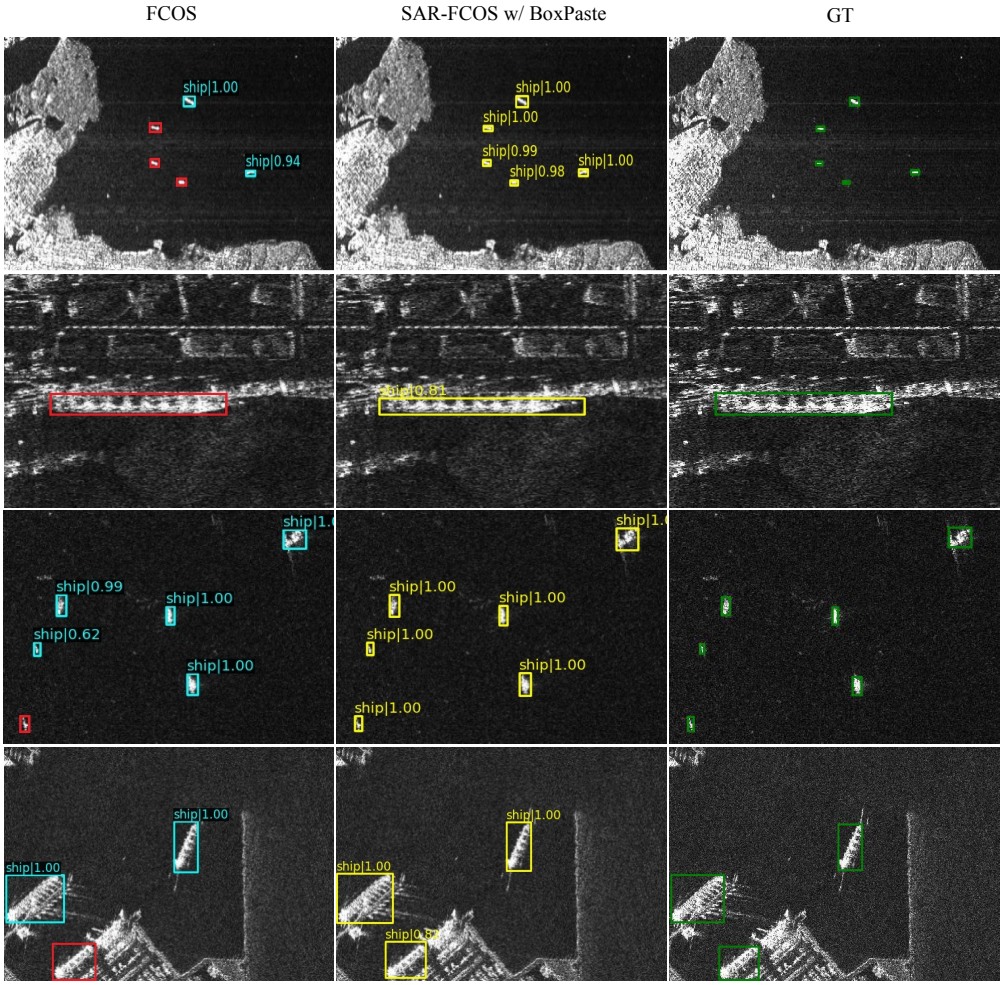


Figure 6. Visualization of the detection results of FCOS, SAR-FCOS with BoxPaste, and ground-truth. The cyan, red, yellow, and green boxes indicate the test results of FCOS, the ship object missed the test results of ours, and the ground truth.

Table 2. Ablation study on BoxPaste *with regard to* scale jittering and training epochs.

	Scale Jit. Range	Total Epochs	BoxPaste	AP (%)	Recall (%)	FPS
SAR-FCOS	-	12	-	93.0	94.4	110.1
	-	12	✓	94.1	95.1	
	[0.5, 1.5]	12	✓	94.5	95.2	
	[0.1, 2]	12	✓	94.6	95.5	
	[0.1, 2]	24	✓	95.2	96.5	
	[0.1, 2]	36	✓	95.5	96.6	
	[0.1, 2]	48	✓	95.0	96.1	

4.2.3. Comparing BoxPaste to CopyPaste

Since the SSDD dataset provides segmentation annotation, we also perform CopyPaste data augmentation on SSDD. The experimental results are shown in Table 3. It shows that CopyPaste marginally improves AP by 0.3%, which may be due to CopyPaste's precise cropping of the object. Meanwhile, because most of the SAR ship database background is simple and clean, the performance of using BoxPaste is very close to CopyPaste. Since the segmentation mask is more difficult to obtain than the object box, making BoxPaste more feasible in the real-world scenario.

Table 3. Performance comparison between BoxPaste+BBox-SSDD and CopyPaste+PSeg-SSDD.

	AP (%)	Recall (%)
BoxPaste+BBox-SSDD	95.5	96.6
CopyPaste+PSeg-SSDD	95.8	97.0

4.3. Wide Applicability

The trials presented in the previous sub-sections solely relied on FCOS. Hence, we validate our methods' broad applicability in this sub-section by combining it with RetinaNet and Adaptive Training Sample Selection (ATSS) [48]. While FCOS is an anchor-free one-stage detector, RetinaNet is anchor-based, and ATSS adopts an advanced label-assigning strategy. Because RetinaNet and ATSS leverage different label-assigning strategies that are probably more suitable in SAR ship detection tasks, therefore, their baseline performances are higher than FCOS. On these detectors, we apply both the light head/FPN and BoxPaste, while in any case, ResNet-50 is the backbone model. The corresponding experimental results are presented in Table 4, highlighting that RetinaNet and ATSS attain 92.6% and 94.4% AP, far better than FCOS (90.8% AP). Possibly due to RetinaNet being an anchor-based detector, while ATSS uses an advanced label-assigning strategy. Both detectors increase the number of positive training samples per mini-batch. Nevertheless, using the light head/FPN still improves their AP by 0.7% and 0.5%, respectively. After adopting BoxPaste, the detection performance is further improved to 95.8% and 96.3%, respectively, validating the applicability of our proposed methods on different detectors. The new detectors, i.e., SAR-RetinaNet and SAR-ATSS, are compared against other state-of-the-art SAR ship detectors on SSDD, utilizing ResNet-50 as the backbone network. The counterparts include one-stage detectors, such as ATSS, as well as powerful two-stage detectors, such as Faster R-CNN and Cascade R-CNN. The results in Figure 7 show that SAR-ATSS achieves 96.3% AP, surpassing all other previous work.

Table 4. The effects of light head/FPN and BoxPaste on RetinaNet and Adaptive Training Sample Selection (ATSS).

	RetinaNet			ATSS		
	Baseline	**+Light**	**+BoxPaste**	**Baseline**	**+Light**	**+BoxPaste**
AP (%)	92.6	93.3	95.8	94.4	94.9	96.3
Recall (%)	94.1	94.2	96.5	95.5	95.5	97.0

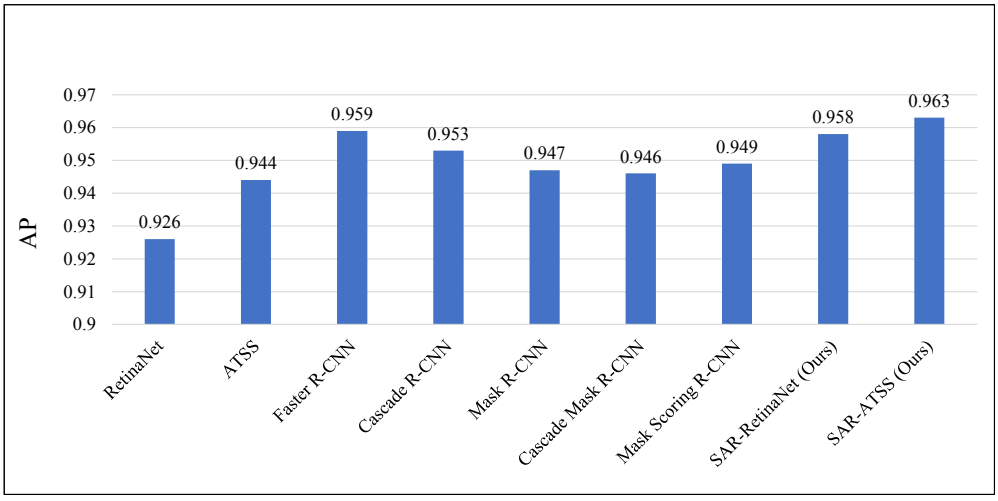


Figure 7. Combining the proposed methods on RetinaNet and ATSS and comparing them against current state-of-the-art methods.

5. Conclusions

The characteristics of the dataset are important for appropriate detection architecture design and training recipes. In SAR ship detection, object size, density, and background diversity are essentially different from general object detection, such as VOC and COCO, motivating us to explore domain-specific techniques in it. In this work, we first present BoxPaste, a simple but effective data augmentation method for SAR ship detection that crops the ship objects from one training image using bounding box annotations and pastes them on another image. Despite its simplicity, BoxPaste significantly improves its baseline by 4.7% mAP. Given the SAR ship image dataset characteristics, we also introduce a principle for designing a SAR ship detector, i.e., a larger model does not guarantee better performance. With this principle, we conduct thorough experiments on FCOS and propose SAR-FCOS, which runs twice as fast and achieves better detection performance. Thorough experiments are conducted that validate the effectiveness of our proposed methods.

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

Funding: This research was funded by the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project), grant number B18039.

Data Availability Statement: Not applicable.

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

References

1. Lee, H.J.; Huang, L.F.; Chen, Z. Multi-frame ship detection and tracking in an infrared image sequence. *Pattern Recognit.* **1990**, *23*, 785–798. [CrossRef]
2. Mingbo, Z.; Jianwu, Z.; Jianguo, H. Imaging simulation of sea surface with full polarization SAR. In Proceedings of the 2015 IEEE 5th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), Singapore, 1–4 September 2015; pp. 815–817.
3. Brusch, S.; Lehner, S.; Fritz, T.; Soccorsi, M.; Soloviev, A.; van Schie, B. Ship Surveillance with TerraSAR-X. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 1092–1103. [CrossRef]
4. Eldhuset, K. An automatic ship and ship wake detection system for spaceborne SAR images in coastal regions. *IEEE Trans. Geosci. Remote Sens.* **1996**, *34*, 1010–1019. [CrossRef]
5. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017; pp. 1–6.
6. Ouchi, K.; Tamaki, S.; Yaguchi, H.; Iehara, M. Ship detection based on coherence images derived from cross correlation of multilook SAR images. *IEEE Geosci. Remote Sens. Lett.* **2004**, *1*, 184–187. [CrossRef]
7. Ai, J.; Qi, X.; Yu, W.; Deng, Y.; Liu, F.; Shi, L. A new CFAR ship detection algorithm based on 2-D joint log-normal distribution in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 806–810. [CrossRef]
8. Wang, C.; Jiang, S.; Zhang, H.; Wu, F.; Zhang, B. Ship detection for high-resolution SAR images based on feature analysis. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 119–123. [CrossRef]
9. Leng, X.; Ji, K.; Yang, K.; Zou, H. A bilateral CFAR algorithm for ship detection in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1536–1540. [CrossRef]
10. Zhai, L.; Li, Y.; Su, Y. Inshore ship detection via saliency and context information in high-resolution SAR images. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1870–1874. [CrossRef]
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
12. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
13. 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, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
14. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
15. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9627–9636.
16. Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; Sun, J. OTA: Optimal Transport Assignment for Object Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 303–312.

17. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
18. Jiao, J.; Zhang, Y.; Sun, H.; Yang, X.; Gao, X.; Hong, W.; Fu, K.; Sun, X. A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection. *IEEE Access* **2018**, *6*, 20881–20892. [CrossRef]
19. Kang, M.; Leng, X.; Lin, Z.; Ji, K. A modified faster R-CNN based on CFAR algorithm for SAR ship detection. In Proceedings of the 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 18–21 May 2017; pp. 1–4.
20. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. Depthwise separable convolution neural network for high-speed SAR ship detection. *Remote Sens.* **2019**, *11*, 2483. [CrossRef]
21. Ge, Z.; Jie, Z.; Huang, X.; Li, C.; Yoshie, O. Delving deep into the imbalance of positive proposals in two-stage object detection. *Neurocomputing* **2021**, *425*, 107–116. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. A SAR dataset of ship detection for deep learning under complex backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]
24. Xian, S.; Zhirui, W.; Yuanrui, S.; Wenhui, D.; Yue, Z.; Kun, F. AIR-SARShip-1.0: High-resolution SAR ship detection dataset. *J. Radars* **2019**, *8*, 852–862.
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. Chang, Y.L.; Anagaw, A.; Chang, L.; Wang, Y.C.; Hsiao, C.Y.; Lee, W.H. Ship detection based on YOLOv2 for SAR imagery. *Remote Sens.* **2019**, *11*, 786. [CrossRef]
27. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
28. Wei, S.; Su, H.; Ming, J.; Wang, C.; Yan, M.; Kumar, D.; Shi, J.; Zhang, X. Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet. *Remote Sens.* **2020**, *12*, 167. [CrossRef]
29. Zhao, Y.; Zhao, L.; Xiong, B.; Kuang, G. Attention receptive pyramid network for ship detection in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2738–2756. [CrossRef]
30. Mao, Y.; Yang, Y.; Ma, Z.; Li, M.; Su, H.; Zhang, J. Efficient low-cost ship detection for SAR imagery based on simplified U-net. *IEEE Access* **2020**, *8*, 69742–69753. [CrossRef]
31. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
32. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]
33. 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.* **2020**, *12*, 2619. [CrossRef]
34. Dvornik, N.; Mairal, J.; Schmid, C. Modeling visual context is key to augmenting object detection datasets. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 364–380.
35. Dwibedi, D.; Misra, I.; Hebert, M. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1301–1310.
36. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, *arXiv:1710.09412*.
37. Yun, S.; Han, D.; Chun, S.; Oh, S.J.; Yoo, Y.; Choe, J. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6022–6031.
38. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2918–2928.
39. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
40. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
41. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-Frst AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
42. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
43. 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, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
44. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
45. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
46. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

47. 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.
48. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. 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, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.