remote sensing

MDPI

Article

Contextual Region-Based Convolutional Neural Network with Multilayer Fusion for SAR Ship Detection

Miao Kang, Kefeng Ji *, Xiangguang Leng and Zhao Lin

School of Electronic Science and Engineering, National University of Defense Technology, Sanyi Avenue, Changsha 410073, China; kangmiao15@163.com (M.K.); luckight@163.com (X.L.); lzkmylz@163.com (Z.L.)
* Correspondence: jikefeng@nudt.edu.cn; Tel.: +86-731-8457-6384

Academic Editor: Xiaofeng Yang
Received: 21 July 2017; Accepted: 9 August 2017; Published: 20 August 2017

Abstract: Synthetic aperture radar (SAR) ship detection has been playing an increasingly essential role in marine monitoring in recent years. The lack of detailed information about ships in wide swath SAR imagery poses difficulty for traditional methods in exploring effective features for ship discrimination. Being capable of feature representation, deep neural networks have achieved dramatic progress in object detection recently. However, most of them suffer from the missing detection of small-sized targets, which means that few of them are able to be employed directly in SAR ship detection tasks. This paper discloses an elaborately designed deep hierarchical network, namely a contextual region-based convolutional neural network with multilayer fusion, for SAR ship detection, which is composed of a region proposal network (RPN) with high network resolution and an object detection network with contextual features. Instead of using low-resolution feature maps from a single layer for proposal generation in a RPN, the proposed method employs an intermediate layer combined with a downscaled shallow layer and an up-sampled deep layer to produce region proposals. In the object detection network, the region proposals are projected onto multiple layers with region of interest (ROI) pooling to extract the corresponding ROI features and contextual features around the ROI. After normalization and rescaling, they are subsequently concatenated into an integrated feature vector for final outputs. The proposed framework fuses the deep semantic and shallow high-resolution features, improving the detection performance for small-sized ships. The additional contextual features provide complementary information for classification and help to rule out false alarms. Experiments based on the Sentinel-1 dataset, which contains twenty-seven SAR images with 7986 labeled ships, verify that the proposed method achieves an excellent performance in SAR ship detection.

Keywords: context information; convolutional neural network (CNN); ship detection; synthetic aperture radar (SAR); Sentinel-1

1. Introduction

With the rapid development of spaceborne SAR, such as TerraSAR-X, RADARSAT-2 and Sentinel-1 [1–3], synthetic aperture radar (SAR) ship detection has been playing an increasingly essential role in marine monitoring and maritime traffic supervision [4–6]. Many investigations relating to ship detection in SAR imagery have been carried out. Traditional methods [7–9] detect targets after sea–land segmentation and utilize the hand-crafted features for discrimination, which has poor performance on nearshore areas and has difficulty ruling out false alarms, such as icebergs and small islands. Additionally, the existence of speckle noises and motion blurring in SAR images causes undesirable differences between ships, which creates difficulty for traditional SAR ship detection

methods in extracting effective features for discrimination. Therefore, it is necessary to develop detectors with strong feature extraction capabilities to obtain better performances in SAR ship detection.

Deep neural networks are capable of feature representation and have been widely applied for object detection [10,11]. They provide a highly promising approach for end-to-end object detection. Since the breakthroughs made by the region-based convolutional neural network (R-CNN) [12] using the PASCAL VOC dataset, the process followed by a region-based proposal extractor with a detection network has been intensively investigated in recent years [13,14]. Ren et al. [15], introduced a Region Proposal Network (RPN) to replace the typical region proposal methods, which achieves end-to-end object detection and shares full-image convolutional features with a RPN and Fast R-CNN. Deep transfer learning algorithms [16–18], which tune the model with rich labeled source domains and small-scale target domains, are widely used to reduce the demand of labeled data and accelerate the convergence of networks.

Despite being capable of extracting discriminative representation, the sharing CNN has a tradeoff between the spatial resolution of the network and the semantic distinction of features. Specifically, the shallow layers of CNN have a higher spatial resolution but more coarse features. The feature maps of intermediate layers are complementary with a passable resolution. Moreover, with the depth of layer increasing, the feature map becomes highly semantic but abstract. Taking VGG16 [19] for example, a 32×32 pixel object will shrink to 2×2 when it comes to the last convolutional layer. In general, the mean area of the majority of ships on SAR images from Sentinel-1 is smaller than 32×32, which means that the ship detection on Sentinel-1 belongs to small-sized object detection. Therefore, when the bounding box predictions map to the last feature maps by ROI pooling, small-sized objects have little information for location refinement and classification, which naturally degrades the performance of detection.

In order to cover the shortage of small-sized object detection, experiments have been conducted by utilizing the different layers of CNN. SSD [14], MS-CNN [20] and FPN [21] predict objects on multiple layers and fuse the output in the end, which also consumes more time for training and testing. Tao Kong et al. [22] proposed a HyperNet to incorporate the intermediate layer with the downscaling shallow layer and up-sampled deep layers, and compress them into a uniform space, which obtained a comprehensive and relatively high resolution framework. MultiPath Network [23] and U-Net [24] utilize skip connection between different layers to provide better feature representation at the cost of a complex network structure.

Another way to improve the performance is to add contextual features for small-sized objects. Research shows that contextual information around the objects in input images can provide a valuable cue for object detection [25,26]. Especially for ship detection, the ocean surroundings can help detectors to better rule out false alarms on land. Thus, adding context information to deep object detection networks is a way to improve their distinction of small-sized ship detection. In ParseNet [27], global context features are appended to help clarify the local confusion. With the contextual information about the whole image, it has limited effects on object detection. Inside–Outside Net (ION) [28] integrated the contextual information outside the region of interest by using spatial recurrent neural networks with multiple layer feature maps. In order to obtain a better performance, the IRNN which is Recurrent Neural Networks with ReLU recurrent transitions, needs to be trained on extra semantic segmentation labels, which increases the difficulty of training. Chenchen Zhu et al. [29] presented a face object detection network named CMS-RCNN, which combined multi-scale information with body contextual information, for real-world face detection. However, this approach only builds fused feature maps, which have the same resolution as the deepest layer and the small-sized objects have little information for bounding box prediction.

This paper proposes a contextual convolutional neural network with multilayer fusion for SAR ship detection. Similar to R-CNN, the proposed network is composed of a RPN with high resolution and an object detection network with contextual features. Instead of using low-resolution feature maps from a single layer for proposal generation, the proposed method employs an intermediate layer

combined with a downscaled shallow layer and an up-sampled deep layer to predict the bounding box. In this way, the fused feature maps integrate semantic, complementary, and high-resolution CNN features. The spatial resolution of a RPN is raised to the same level as the intermediate layer, which enlarges the response area of small-sized ships in feature maps. In the object detection network, region proposals are projected onto multiple layers with ROI pooling to extract the corresponding features. Contextual features around the ROI contain the environmental information of candidates, which can complement the computation of a confidence score and bounding box regression.

The rest of this paper is organized as follows. Section 2 introduces the details of the proposed method. Section 3 presents three experiments conducted on Sentinel-1 dataset to validate the effectiveness of the proposed framework. Section 4 discusses the results of the proposed method. Finally, Section 5 concludes this paper.

2. Proposed Method

In order to improve the performance of ship detection, the proposed network consists of a RPN with higher resolution and an object detection network with contextual features. As is shown in Figure 1, in a RPN, 13 convolutional layers of VGG16 [19] are employed for shared feature extraction. All convolutional layers adopt very small 3×3 filters in order to reduce the number of parameters and to decrease the demand of labeled data. In this paper, conv1_2, conv2_2, conv3_3, conv4_3 and conv5_3 of VGG16 are called conv1, conv2, conv3, conv4 and conv5 respectively. In order to improve the resolution of the network, a shallow layer and a deep layer ("conv1" and "conv3") are downscaled with max pooling and up-sampled with deconvolution respectively. Then, they are concatenated with the intermediate layer ("conv2") and compressed into a uniform space with l_2 normalization [26], which obtains the same resolution of the intermediate layer and more detailed information for region generation. The fused cube is reshaped to the same dimension as the intermediate layer and fed into bounding box regression for the sake of the region proposal. The predicted bounding boxes are mapped to different layers of VGG16 by ROI-pooling operations to obtain ROI features. Simultaneously, contextual features around each ROI are extracted. After normalization, concatenation and dimension reduction, the ROI features and contextual features are imported into two fully connected layers ("fc"). Finally, two flattened vectors are concatenated for classification and location refinement.

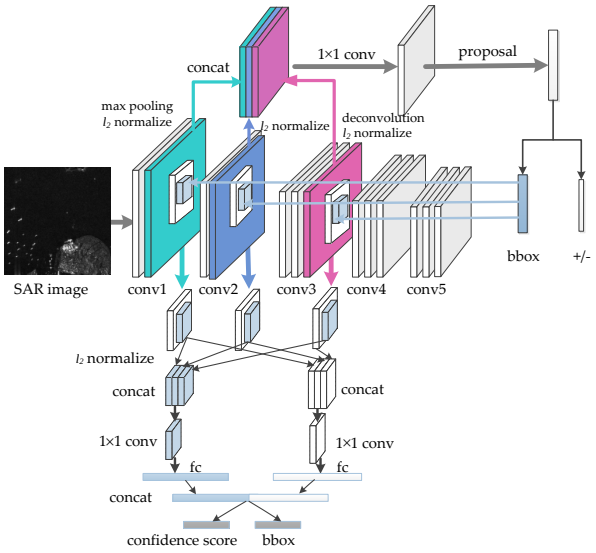

Figure 1. The architecture of the proposed network. SAR images are fed into VGG16 which has five sets of layers for feature extraction. The upper part is the RPN of the network. The white blocks and light blue blocks represent contextual features and ROI features respectively, which are processed concurrently in the object detection network.

The rest of this section introduces the details of the proposed method and explains the motivation of our design.

2.1. Concatenation of Multiple Layers

In order to reduce the number of parameters in the neural network, CNN always shrinks its feature maps by using the max pooling operation after convolution. That is, one pixel on the feature map corresponds to several pixels in the input image and the numerical correspondence is defined as the resolution of a network. Some feature maps are displayed in Figure 2, which shows that in VGG16 shallow layers keep more details of the input image. With the increase of resolution, the feature map becomes smaller and more abstract, and while small-sized objects hardly have responses on the deeper layers.

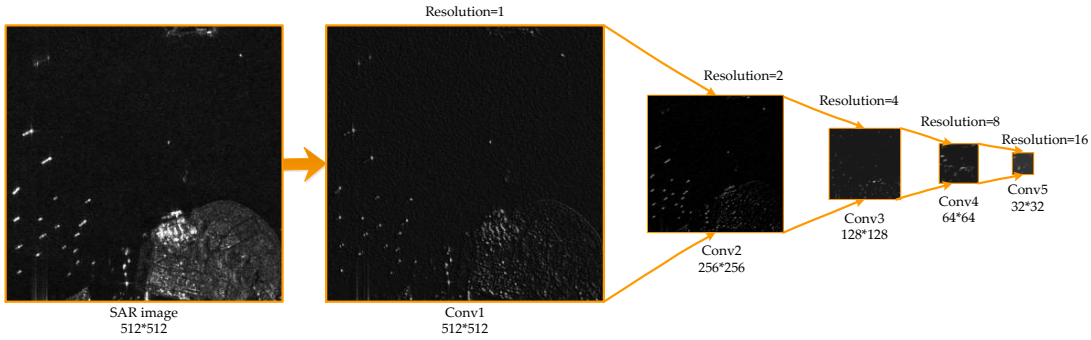

Figure 2. The feature maps of different layers in VGG16. With the increasing of resolution, the feature map becomes smaller and more semantic.

Due to the respective merits that different layers possess, multiple layers fusion is a popular way to enhance the performance of detection in the current top-performance detector. As CMS-RCNN [29] did, the first way is to integrate down-sampled earlier layers with the last layer of the sharing CNN. Despite the fact that the feature map information is increased, small-sized objects still only cause responses on a tiny area in a fused feature map. Another way is to increase the resolution of a network by up-sampling the deeper layer and connecting them with the shallower layer as proposed in this paper. With the integration of conv1, conv2 and conv3, the resolution of network changes from 16 to 2, which means a 32×32 sized object in the input image will have a 16×16 sized response on the fused feature map rather than a 2×2 sized response. The increase of resolution will naturally provide more detailed information for the following bounding box prediction.

2.2. Layer Up-Sampling with Deconvolution

Deconvolution, also known as transposed convolution, is extensively used in feature visualization, image generation and up-sampling [21,30,31]. Since a naive up-sampling inadvertently loses details, for feature map rescaling, a better option is to have a trainable up-sampling convolutional layer, namely a deconvolution layer [21,22], whose parameters will change during training. The implementation of deconvolution consists of two operations as shown in Figure 3. The first step is to insert zeros between the consecutive inputs according to the resolution requirements. After that, an operation similar to convolution is conducted, that is, defining a kernel of an appropriate size and sliding it with a stride to get a higher resolution output compared with the inputs. Since such an operation simply reverses the forward and backward passes of convolution, up-sampling with deconvolution is able to be performed in-network for end-to-end learning by backpropagation.

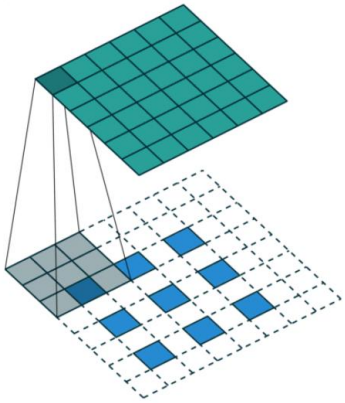

Figure 3. An illustration of deconvolution. It consists of inserting zeros and convolution operation [32].

2.3. L_2 Normalization

In general, with the depth of the network increasing, the scale and norm of feature maps always have a tendency to decrease. Concatenating multiple feature maps directly will lead to the dominance of shallow layers [27] and degrade the generalization ability of the model. Although the weights of layers are able to be tuned during the training, it takes a long time for the network to fill the dramatic gap in scale of value and it requires elaborate tricks to achieve a good performance. With the limited labeled data, overtraining will make the model learn the detail and noises in the training data and put the network at risk of overfitting. Therefore, a desirable approach is to employ l_2 normalization to constrain the scale of value of the different feature maps to the same level before integration.

l_2 normalization is applied to every pixel of the feature maps. For a layer that has d-channel feature maps sized with (w, h), l_2-norm for a d-channel vector is represented with Equation (1)

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{d} |x_i|^2\right)^{1/2} \tag{1}$$

Per d-dimension pixels vector $\mathbf{x}$ of a layer is normalized as in Equation (2)

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \tag{2}$$

where $\hat{\mathbf{x}}$ is the d-dimension normalized pixels vector.

In order to accelerate the training, the scale value of layers is always rescaled with a factor μ for each channel i.

$$y_i = \mu x_i \tag{3}$$

The scale factor μ is able to be updated with the backpropagation and chain rule [29]. In this paper, a fixed scale factor, which makes the fused feature maps have the same mean level as the replaced layer in Faster RCNN, is adopted [28].

2.4. ROI Pooling to Multiple Layers

In Faster R-CNN, the prediction of the bounding box will be projected onto the last convolutional layer. Since region proposals are extracted from the fusion layer in this paper, projecting the region proposals to appropriate layers and fusing the region features as the fused layer will generate more accurate and comprehensive features for classification.

As shown in Figure 1, bounding box predictions are mapped to conv1, conv2 and conv3 respectively instead of a single layer. The corresponding regions on feature maps are normalized, rescaled and fused together.

2.5. Integrating Contextual Information

When searching for ships in a SAR image with the visual system of a human, context information is able to help us to increase the confidence of decision. For instance, an object located on land is highly unlikely to be considered a ship, while an object with bright intensity in the ocean area is prone to be affirmed as a positive object. In order to mimic the visual effect of a human being in a computer vision field, context information is always added into the deep neural network to recognize the small-sized objects [27,29,33].

As shown in Figure 4, the proposed method takes the surrounding pixels of the proposal as context information. In order to keep the same quantitative relation when the bounding boxes are projected to multiple layers to obtain contextual features, we keep $w_{context} = \lambda \times w_{proposal}$ and $h_{context} = \lambda \times h_{proposal}$, where w and h represent the width and height of the bounding box. After l_2 normalization and concatenation, the contextual features are flattened to a vector in the fully connected layers, which are combined with ROI features in a new fused vector for the final output.

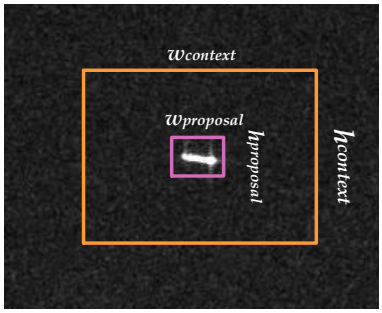

Figure 4. An illustration of context information. The bounding box in purple represents the region proposal of the network and the outer orange bounding box is the boundary of context information.

3. Experiments and Results

In this section, experiments are carried out to evaluate the performance of the proposed method. Two experiments are designed to explore the effect of different layer fusions and the influence of contextual features. Besides, the comparison with other methods indicates the outperformance of the proposed method.

3.1. Experiment Dataset and Settings

The dataset used in this paper is Sentinel-1, provided by the European Space Agency (ESA) on the Internet [34] for free, which was collected in Interferometric Wide swath (IW) mode. Compared with Extra-Wide swath (EW) mode, IW mode, as the main operational mode of Sentinel-1, is able to acquire more and higher resolution images. Full resolution Level-1 Ground Range Detected (GRD) products with 10 m pixel spacing were obtained. We labeled the location and the box of the ships on SAR images with ship detection software [35] and visual interpretation. Some of them were verified with Automatic Identification System (AIS) information [36]. Twenty-seven SAR images with 7986 labeled ships were utilized in this paper and seven of them, containing 1502 ships, were used for testing. Five-sixths and one-sixth of the remainders were used for training and validation sets respectively.

The histogram of the ship area is shown in Figure 5, according to the labeled ships that provided AIS information. More than 85% of ships have an area smaller than 8000 m^2, that is, around 80 pixels on a SAR image, which is less than the object size of the ImageNet dataset (more than 80% of objects have sizes between 40 and 140 pixels) [33]. Additionally, the ships which offer AIS information have an average length of 168.3 m. Furthermore, the average area is around 51 pixels which is far less than the area that is able to cause a response on the last convolutional layer of VGG16.

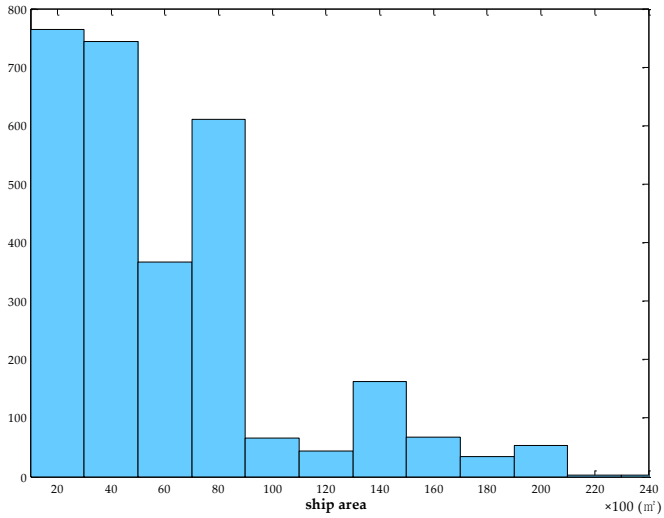

Figure 5. The distribution of the ship area in the Sentinel-1 dataset (only the ships have Automatic Identification System (AIS) information). More than 85% of ships have an area smaller than 80 pixels on SAR images.

The labeled SAR images were cut into 512 × 512 sized patches without overlap and the coordinates of the labeled bounding boxes were transformed into the location of the corresponding patch. Those patches with the labeled ship were selected to feed into the proposed network for training. The testing images are processed in the same way and are combined together for the detection result display.

All experiments are implemented in the Tensorflow deep learning framework [37] and are executed on a PC with an Intel single Core i7 CPU, NVIDIA GTX-1070 GPU (8 GB video memory), and 64 GB RAM. The PC operating system was Ubuntu 14.04.

As is common practice, the pre-trained model on the ImageNet dataset of VGG16 was used to initialize the model. According to the calculation of the mean norm of conv5 and pool5 of Faster RCNN, which is trained on the Sentinel-1 dataset, the scale factor μ for a RPN and object detection network is initialized to 20 and 40 respectively. The learning rate was set to 1×10^{-4} initially and the maximal iteration was 10,000.

At the same time, we define the target detection probability as

$$p_d = \frac{N_{td}}{N_{ground\_truth}} \tag{4}$$

where N_{td} is the number of detected targets and N_{ground_truth} denotes the total number of ground truth and in this paper we have N_{ground_truth} = 1502. Similarly, the estimation of the false alarm probability is defined as (5), where N_{fd} denotes the number of false detected targets of all testing images and N_{total_target} denotes the total number of detected ships of all testing images.

$$p_f = \frac{N_{fd}}{N_{total\_target}} \tag{5}$$

In order to evaluate the overall performance of the detector, F_1 score which is defined as (6) is adopted in this paper. It reaches its best value at 1 and worst at 0.

$$F_1 = 2 \times \frac{p_d \times (1 - p_f)}{p_d + (1 - p_f)} \tag{6}$$

3.2. Influence of Different Layer Combination Strategies

As mentioned before, feature maps from different layers differ in terms of spatial resolution and semantic distinction, giving them comparative advantages and disadvantages. Therefore, layer selection has a great impact on the performance of the detection system. In this section, four models with different layer combination strategies are trained for exploring the influence of different layer selections. Specifically, the first model combines the conv3, conv4 and conv5 of VGG16 together for region proposal. The second model integrates conv1, conv3 and conv5, and the final model selects conv1, conv2 and conv3. The baseline method is a model with a single layer conv5. All models have the same object detection network as the proposed method. The influence of different λ for contextual features will be discussed in Section 3.3 and in this section we take λ equal to 3 for all models to explore the effects of different layer fusion strategies.

As shown in Figure 6, in the open water areas, models have a comparative performance. Conv5 misses most of the ships around the tiny harbor where the denser ships berth. The situation improves greatly when conv3, conv4 and conv5 are combined. With the improvement of network resolution, more small-sized targets are picked up with the combination of conv1, conv3 and conv5. When the resolution of the network increases to the same level of conv2 in the model of conv 1+2+3, the best performance is achieved and only few tiny weak targets are missing. The comparison of the performance indicates that the detection performance of dense tiny ships improves dramatically with the fusion of layers and the increase in network resolution.

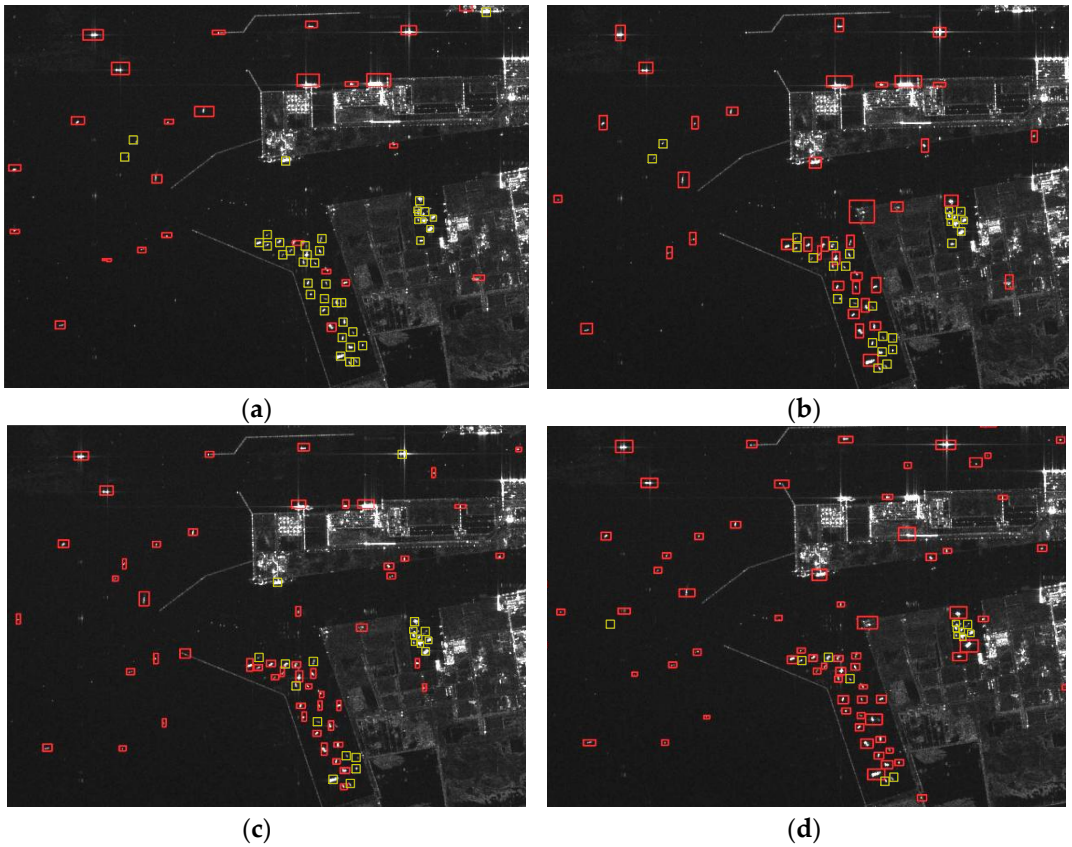

Figure 6. The comparisons of detection results with different layer combination strategies. The red and yellow rectangles represent the detection results and missing ships of detectors respectively. (**a**) conv5; (**b**) conv 3+4+5; (**c**) conv 1+3+5; (**d**) conv 1+2+3.

Table 1 displays the detection probability, false alarm probability and F_1 scores of different layer combination strategies. Compared with the performance on a single layer conv5, the networks with

combined layers achieve higher detection probability and lower false alarm probability. With the combination of feature maps and a slight increase of resolution, the model with layer conv 3+4+5 detects more targets and obtains the lowest false alarm probability. The fusion of conv1, conv3 and conv5 promotes the detection probability to 80.43%. Conv 1+2+3 has the highest resolution compared with the other structures, which leads to a 12.71% increase in P_{d} compared with a single layer. Compared with other fusion structures, conv 1+2+3 also has a slight but acceptable increase in false alarm probability, since the feature maps from shallow layers have lower semantic distinction. The highest F_1 score also indicates that the combination of conv1, conv2 and conv3 has the best performance in SAR ship detection.

Table 1. Detection performance with different layer combination strategies.

Layers	$N_{total\_targets}$	N_{td}	N_{fd}	P_d (%)	P_f (%)	F_1
conv5	1463	1106	357	73.64	24.4	0.746
conv 3+4+5	1355	1160	195	77.23	**14.39**	0.812
conv 1+3+5	1418	1208	210	80.43	14.81	0.827
conv 1+2+3	1540	1297	243	**86.35**	15.78	**0.853**

In summary, the increase of network resolution can dramatically improve the performance of detectors, especially in small-sized targets detection. Different layer combination strategies have a great impact on detection performance. As for SAR ship detection on Sentinel-1, since the sizes of most targets are smaller than 32 $\times$ 32 and the features of ships are relatively simple in intensity imagery, the combination of shallow layers from VGG16 is semantic enough to detect a ship in the background. In other words, resolution improvement plays a more important role than semantic feature for ships detection in SAR imagery.

3.3. Influence of Contextual Features

In order to identify the influence of contextual features, comparison experiments with different sizes of contextual features in the proposed network are conducted in this section. The network without contextual information means the object detection network only has one branch in the object detection network of Figure 1. In other models, λ changes from 2 to 7 to obtain different sized contextual features. The combination strategy of conv 1+2+3 is adopted and all models have the same experiment settings.

Table 2 shows that when the bounding box of context information is relatively small, additional contextual information improves the overall performance to different degrees with higher F_1 scores. When a bounding box of contextual information five times larger than normal is appended, the best performance is obtained and the F_1 score changes to 0.873. Compared with the model without any contextual information, the model with fivefold contextual features increases by 4.53% in detection probability and decreases by 3.34% in false alarm rate. That is, extra contextual features provide more information for the model to pick up more targets. Meanwhile, the additional surrounding information of proposals also successfully assists to discriminate targets from false alarms. However, when the size of the bounding box enlarges to 6 or 7, the detection probability begins to decrease. One of the possible reasons is that most of the bounding boxes are oversized when λ is too large, which leads to the dominance of contextual information in the concatenated features and aggravates the performance of the network. Thus, the size of contextual information should be moderated according to the detection task. Specifically, the proposed method possesses the best detection performance when fivefold contextual information is added and conv1, conv2 and conv3 are fused.

Table 2. Detection performance comparisons between different sized contextual features.

Context Size	$N_{total\_targets}$	N_{td}	N_{fd}	P_d (%)	P_f (%)	F_1
without context	1518	1259	259	83.82	17.06	0.834
2× context	1534	1285	249	85.56	16.23	0.847
3× context	1540	1297	243	86.35	15.78	0.853
4× context	1564	1280	284	85.22	18.16	0.836
5× context	1538	1327	211	**88.35**	**13.72**	**0.873**
6× context	1519	1228	291	81.76	19.16	0.813
7× context	1481	1231	250	81.95	16.88	0.825

3.4. Comparisons with Other Methods

In order to validate the effectiveness of the proposed method, Faster RCNN [15,38] and CMS-RCNN [29] are applied to Sentinel-1 dataset. CMS-RCNN, which has the same resolution as conv5, fuses conv3, con4 and conv5 by down-sampling. The other experiment settings of CMS-RCNN and Faster RCNN are the same as the proposed method.

Table 3 displays the performance of the three methods. Due to the increase of complexity in the network structure, the proposed method consumes more time in training. However, for a 512 × 512 sized image, the testing time of the proposed method remains at the same level as Faster RCNN and CMS-RCNN. With the layer fusion and the additional context information, the proposed network increases by 25.8% in detection probability and reduces the false alarm probability from 27.68 to 13.72% compared with Faster RCNN. Based on a higher network resolution than CMS-RCNN, the proposed method also promotes the detection performance significantly.

Table 3. Detection performance comparison between different methods.

Method	$N_{total\_targets}$	N_{td}	N_{fd}	P_d (%)	P_f (%)	F_1	Testing Time
Faster RCNN	1304	943	361	62.78	27.68	0.672	1.019 s
CMS-RCNN	1491	1126	365	74.97	24.48	0.752	1.064 s
Proposed method	1538	1327	211	**88.35**	**13.72**	**0.873**	2.180 s

By changing the confidence score threshold of detection results on one testing image, different values of p_d and p_f are obtained, which produces the performance curves of different methods in Figure 7. As shown in the figure, the proposed method has the highest detection probability in a given false alarm probability. Similarly, with a specific p_d, the proposed method has the lowest false alarm probability. Therefore, the proposed method performs better than Faster RCNN and CMS-RCNN.

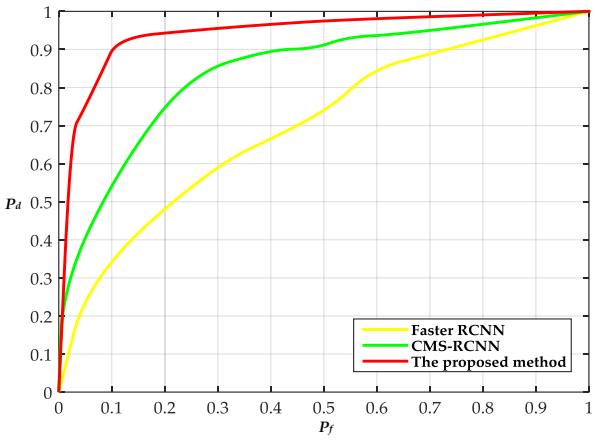

Figure 7. The performance curves of different methods. The yellow, green and red curves represent the performance of Faster RCNN, CMS-RCNN and the proposed method respectively. The x label and y label represent p_f and p_d respectively.

4. Discussion

Experiments on combination strategies and the influence of context information verify the effectiveness of the proposed method in ship detection, especially in small-sized targets detection. The comparisons with Faster RCNN and CMS-RCNN demonstrate the necessity of resolution improvement and additional context information.

Since the proposed method omits sea–land segmentation which traditional methods required, it provides the possibility for the network to detect ships nearshore, where traditional methods cannot perform well because of the limited accuracy of sea–land segmentation. The equal treatment of land and sea area also brings some undesirable false alarms on the land as shown in Figure 8. The red, yellow and purple boxes represent the detected target, the false alarms and missing targets respectively.

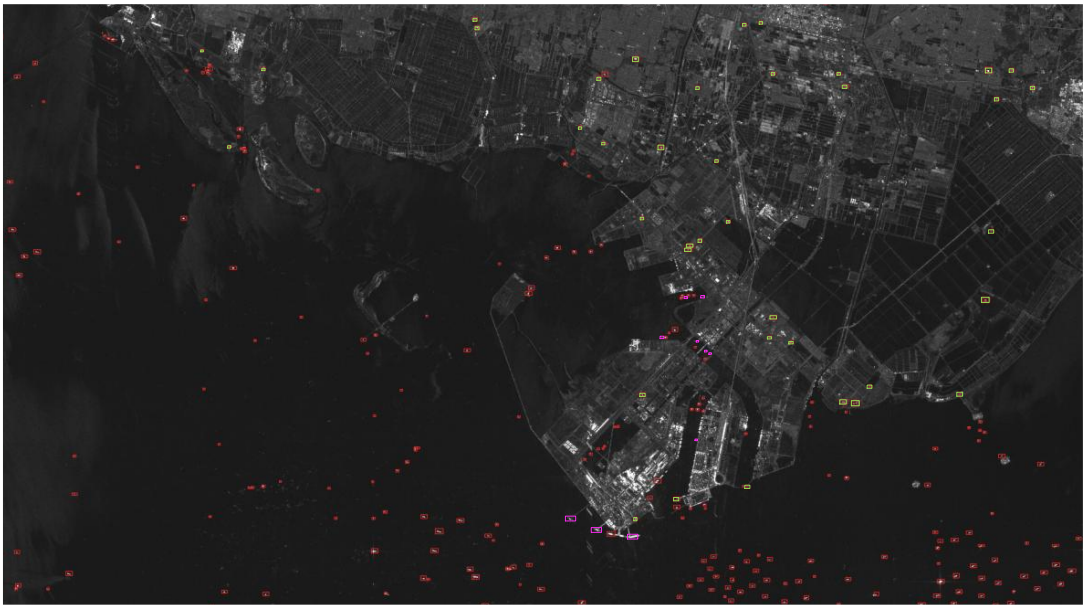

Figure 8. The detection results of the proposed method near the harbor area. The red, yellow and purple boxes represent the detected target, the false alarms and missing targets respectively.

Table 4 records the main categories of the false alarms in one test image. It is found that almost 65% of false alarms are building facilities on land, which are able to be ruled out with sea–land segmentation in image preprocessing. Some harbor facilities also are incorrectly detected as ships. While in the open ocean area, some noises, such as azimuth ambiguity and speckle, which have bright intensity will be picked up by the model. Islands, one of most annoying false alarms in the traditional method, are the least common false alarm category.

Table 4. The categories of false alarms.

Categories	Building	Harbor	Island	Noise	Total
Number	68	17	3	**18**	**106**

In order to analyze the characteristics of false alarms, some typical patches are displayed in the blue box of Figure 9. Visually, most of them are extremely similar to true positive targets. That is, they are brighter than their context and are shaped similar to ships, which means that the network values the visual features. Those kinds of false alarms are also hard to rule out by some hand-crafted methods. Therefore, some additional discrimination networks need to be trained, aimed at those false alarms and ships.

As shown in the purple box of Figure 9, some missing targets have weak or small intensity, which makes them cause few responses on the shallow layers and go undetected by the network. The missing label of weak and tiny targets on the training dataset is another possible reason for the missing detection, since the performance of the network is driven by the data which is fed into the network. Some of the missing targets are very near to the shore or to some other brighter targets, which makes the network assign them a low confidence score. Additionally, the motion blurring and cross sidelobe of ships also exert adverse effects on classification.

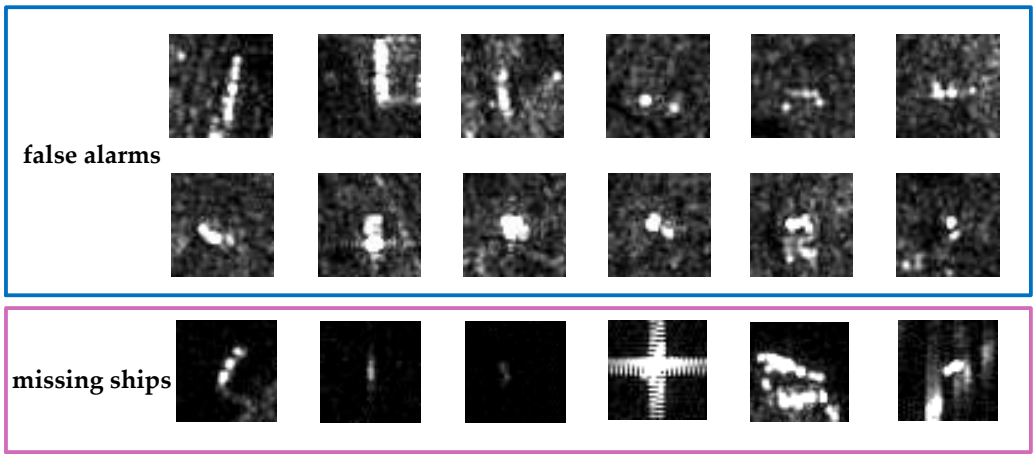

Figure 9. Some typical false alarms and missing detection targets among the detection results of the proposed method. The chips in the blue box and the purple box are false alarms and missing ships respectively.

5. Conclusions

With the labeled dataset on Sentinel-1, this paper opens up the possibility of utilizing deep neural networks for SAR ship detection. In order to improve the detection of ships on Sentinel-1 SAR imagery, where ships always appear small, layer fusion is employed in a contextual convolutional neural network to obtain semantic and high-resolution feature maps. Additionally, contextual information is added in the object detection network in order to help detectors to rule out false alarms. Experiments conducted in this paper demonstrate the effect of the layer fusion strategy and validate the influence of contextual information. More importantly, experiment results validate that the proposed method improves the detection performance dramatically.

Despite the effectiveness of the proposed method, some weak and tiny targets remain undetected and false alarms on land are hard to rule out. Investigations into the detection of these targets and false alarm discrimination need to be carried out in the future.

Acknowledgments: This work is supported partly by the National Natural Science Foundation of China under Grant Nos. 61372163 and 61331015. The authors would like to thank European Space Agency (ESA) for providing free Sentinel-1 data online. The authors would also like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

Author Contributions: Miao Kang conceived and designed the experiments; Miao Kang performed the experiments and analyzed the data; Kefeng Ji, Xiangguang Leng and Zhao Lin contributed materials; Miao Kang and Kefeng Ji wrote the paper.

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

References

1. Brusch, S.; Lehner, S.; Fritz, T.; Soccorsi, M. Ship suveillance with TerraSAR-X. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 1092–1103. [CrossRef]

2. Crisp, D.J. A ship detection system for RADARSAT-2 dual-pol multi-look imagery implemented in the ADSS. In Proceedings of the 2013 IEEE International Conference on Radar, Adelaide, Australia, 9–12 September 2013; pp. 318–323.
3. Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. GMES Sentinel-1 mission. *Remote Sens. Environ.* **2012**, *120*, 9–24. [CrossRef]
4. Crisp, D.J. The state-of-the-art in ship detection in Synthetic Aperture Radar imagery. *Org. Lett.* **2004**, *35*, 2165–2168.
5. Marino, A.; Sugimoto, M.; Ouchi, K.; Hajnsek, I. Validating a Notch Filter for Detection of Targets at Sea with ALOS-PALSAR Data: Tokyo Bay. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 74907–74918. [CrossRef]
6. Pelich, R.; Longépé, N.; Mercier, G.; Hajduch, G.; Garello, R. AIS-Based Evaluation of Target Detectors and SAR Sensors Characteristics for Maritime Surveillance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3892–3901. [CrossRef]
7. Wang, C.; Bi, F.; Zhang, W.; Chen, L. An Intensity-Space Domain CFAR Method for Ship Detection in HR SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 529–533. [CrossRef]
8. Zhi, Z.; Ji, K.; Xing, X.; Zou, X.; Zhou, H. Ship Surveillance by Integration of Space-borne SAR and AIS—Review of Current Research. *J. Navig.* **2014**, *67*, 177–189. [CrossRef]
9. Fingas, M.F.; Brown, C.E. Review of Ship Detection from Airborne Platforms. *Can. J. Remote Sens.* **2001**, *27*, 379–385. [CrossRef]
10. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
11. Druzhkov, P.N.; Kustikova, V.D. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognit. Image Anal.* **2016**, *26*, 9–15. [CrossRef]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp. 779–788.
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Fu, C.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
15. 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] [PubMed]
16. Long, M.; Wang, J.; Jordan, M.I. Deep transfer learning with joint adaptation networks. *arXiv*, **2016**, arXiv:1605.06636.
17. Gupta, U.; Chaudhury, S. Deep transfer learning with ontology for image classification. In Proceedings of the 5th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Patna, India, 16–19 December 2015; pp. 1–4.
18. Ravishankar, H.; Sudhakar, P.; Venkataramani, R.; Thiruvenkadam, S.; Annangi, P.; Babu, N.; Vaidya, V. Understanding the Mechanisms of Deep Transfer Learning for Medical Images. *arXiv*, **2017**, arXiv:1704.06040.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv*, **2014**, arXiv:1409.1556.
20. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 354–370.
21. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
22. Kong, T.; Yao, A.; Chen, Y.; Sun, F. Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Seattle, WA, USA, 27–30 June 2016; pp. 845–853.
23. Zagoruyko, S.; Lerer, A.; Lin, T.Y.; Pinheiro, P.O.; Gross, S.; Chintala, S.; Dollár, P. A multipath network for object detection. *arXiv*, **2016**, arXiv:1604.02135.

24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
25. Divvala, S.K.; Hoiem, D.; Hays, J.H.; Efros, A.; Hebert, M. An empirical study of context in object detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 1271–1278.
26. Galleguillos, C.; Belongie, S. Context based object categorization: A critical survey. *Comput. Vis. Image Underst.* **2010**, *114*, 712–722. [CrossRef]
27. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv*, 2015; arXiv:1506.04579.
28. Bell, S.; Lawrence Zitnick, C.; Bala, K.; Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Seattle, WA, USA, 27–30 June 2016; pp. 2874–2883.
29. Zhu, C.; Zheng, Y.; Luu, K.; Savvides, M. CMS-RCNN: Contextual Multi-Scale Region-based CNN for Unconstrained Face Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Seattle, WA, USA, 27–30 June 2016.
30. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the 14th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
31. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*, **2015**, arXiv:1511.06434.
32. Dumoulin, V.; Visin, F. A guide to convolution arithmetic for deep learning. *arXiv*, **2016**, arXiv:1603.07285.
33. Hu, P.; Ramanan, D. Finding tiny faces. *arXiv*, **2016**, arXiv:1612.04402.
34. Sentinels Scientific Data Hub. Available online: https://scihub.copernicus.eu/ (accessed on 1 March 2017).
35. Leng, X.; Ji, K.; Zhou, S.; Zou, H. An adaptive ship detection scheme for spaceborne SAR imagery. *Sensors* **2016**, *16*, 1345. [CrossRef] [PubMed]
36. OpenSAR. Available online: http://opensar.sjtu.edu.cn/ (accessed on 29 March 2017).
37. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.; Jeffrey, D.; Devin, M.; et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv*, **2016**, arXiv:1603.04467.
38. Faster-RCNN_TF. Available online: https://github.com/smallcorgi/Faster-RCNN_TF (accessed on 24 May 2017).