remote sensing

MDPI

Article

Improved YOLOv3 Based on Attention Mechanism for Fast and Accurate Ship Detection in Optical Remote Sensing Images

Liqiong Chen [1], Wenxuan Shi [2,*] and Dexiang Deng [1]

[1] School of Electronic Information, Wuhan University, Wuhan 430072, China; liqiongchen@whu.edu.cn (L.C.); ddx@whu.edu.cn (D.D.)
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China
* Correspondence: shiwx@whu.edu.cn

Abstract: Ship detection is an important but challenging task in the field of computer vision, partially due to the minuscule ship objects in optical remote sensing images and the interference of clouds occlusion and strong waves. Most of the current ship detection methods focus on boosting detection accuracy while they may ignore the detection speed. However, it is also indispensable to increase ship detection speed because it can provide timely ocean rescue and maritime surveillance. To solve the above problems, we propose an improved YOLOv3 (ImYOLOv3) based on attention mechanism, aiming to achieve the best trade-off between detection accuracy and speed. First, to realize high-efficiency ship detection, we adopt the off-the-shelf YOLOv3 as our basic detection framework due to its fast speed. Second, to boost the performance of original YOLOv3 for small ships, we design a novel and lightweight dilated attention module (DAM) to extract discriminative features for ship targets, which can be easily embedded into the basic YOLOv3. The integrated attention mechanism can help our model learn to suppress irrelevant regions while highlighting salient features useful for ship detection task. Furthermore, we introduce a multi-class ship dataset (MSD) and explicitly set supervised subclass according to the scales and moving states of ships. Extensive experiments verify the effectiveness and robustness of ImYOLOv3, and show that our method can accurately detect ships with different scales in different backgrounds, while at a real-time speed.

Keywords: ship detection; optical remote sensing images; multi-class ship detection; dilated attention module; real-time speed

Citation: Chen, L.; Shi, W.; Deng, D. Improved YOLOv3 Based on Attention Mechanism for Fast and Accurate Ship Detection in Optical Remote Sensing Images. *Remote Sens.* **2021**, *13*, 660. https://doi.org/10.3390/rs13040660

Academic Editor: Paolo Addesso

Received: 6 January 2021
Accepted: 9 February 2021
Published: 11 February 2021

1. Introduction

With the rapid advancement of space remote sensing technology, high-resolution and large-scale remote sensing images acquired from spaceborne and airborne sensors are constantly enriched and facilitate a wide range of applications, such as natural disaster assessment [1], urban planning [2], traffic management [3,4], and environment monitoring [5]. In these applications, automatic ship detection in remote sensing images has attracted increasing interests due to its important value in both civil and military fields, such as national defense construction, maritime security, harbor surveillance, and fishery management. For decades, many studies [6–10] in this field have prioritized synthetic aperture radar (SAR) images since they are little affected by weather and time. However, the noisy response and low resolution of SAR images [11] may limit their utilizations. In recent years, attributing to the improvement of high-resolution technology, some researchers [12–16] have paid more attention to using optical remote sensing images because they could provide more details and spatial contents for detecting ships than SAR images.

Extensive studies have been carried out about ship detection in optical remote sensing images in the past decades. Many traditional methods take a hierarchical paradigm via extracting handcrafted features, such as shape [12,17], texture [12,15], local binary patterns (LBP) [18], and histogram of oriented gradients (HOG) [13], followed by certain classification operations, e.g., support vector machine (SVM) [19,20], extreme learning

machine (ELM) [14], and Adaboost [21]. Most of the traditional algorithms make great success for ideal-quality images. However, they may encounter bottlenecks when ships are under complex weather conditions. Furthermore, the establishment of handcrafted features excessively relies on expert experience, so its generalization ability is weak. In this paper, we concentrate on ships at sea since the land can be removed by prior geographic information. In addition, we focus on detecting ships in panchromatic band of optical remote sensing images since it has higher image resolution and more detailed visual content than other bands.

Recently, driven by the excellent performance of convolutional neural networks (CNNs) [22–25], deep learning object detectors have become new options to tackle the ship detection problem [26–32]. The existing object detection methods based on deep learning can be roughly grouped into two streams on the basis of whether generating region proposals or not. They are two-stage detection methods, such as R-CNN [33], Fast R-CNN [34], and Faster R-CNN [35], and one-stage detection methods, such as SSD [36], YOLO series [37–39], and RetinaNet [40]. Broadly speaking, two-stage detectors offer high positioning accuracy, whereas one-stage detectors have an absolute advantage in terms of speed.

Many recent works have exploited state-of-the-art two-stage detectors for ship detection in remote sensing field. For instance, Yao et al. [32] utilized region proposal network (RPN) [35] to discriminate ship targets and regress the detection bounding boxes, in which the anchors were designed by intrinsic shape of ship targets. Yang et al. [30] presented dense feature pyramid network (DFPN) by introducing rotation anchor strategy and multi-scale region of interest (RoI) align into the Faster R-CNN baseline. Due to the massive amount of generated ROIs, the computation is time-consuming and the model efficiency may be reduced. Chen et al. [16] also designed a hierarchical detection process which applied discrete wavelet transform (DWT) to extract ship candidate regions and proposed deep residual dense network (DRDN) to improve the ship detection accuracy. The aforementioned two-stage detection methods mainly focus on improving the ship detection accuracy; however, they may ignore the detection speed. As a matter of fact, in some practical applications such as real-time ocean observation and timely ship rescue, increasing ship detection speed is as important as boosting the detection accuracy.

To achieve real-time ship detection in remote sensing images, the methods based on one-stage detectors have been gradually explored. For instance, Tang et al. [41] proposed HSV-YOLO which applied the difference of HSV color space among remote sensing images to extract the RoIs and sent them into the YOLOv3 [39] network. This approach is similar to saliency-based approaches [13,42]. However, the HSV-YOLO is not end-to-end due to the complex preprocessing based on the HSV difference. Van Etten [43] extended the YOLOv2 [38] network and implemented an end-to-end object detection framework optimized for overhead imagery: You Only Look Twice (YOLT). It realized rapid multi-scale object detection in satellite images. Nina et al. [44] compared YOLOv3 [39] and YOLT [43] for ship detection, without modification of the original architecture. Although these YOLO series could be seen as first choice for real-time applications in computer vision, they provided much lower detection accuracy when compared with two-stage detectors, notably in detecting small ships. In addition, some common challenges still arise for both two-stage and one-stage ship detection methods. First, due to the overhead perspective and imaging characteristics of optical remote sensing images, they usually suffer from the complex weather condition such as clouds, mists, and waves [45]. Second, the variant appearances and sizes of ships make it difficult for existing methods to accurately locate and detect targets, especially small ships.

To solve the aforementioned problems, especially for small ships detection and the interference of complex backgrounds, we propose a novel one-stage ship detection method named improved YOLOv3 (ImYOLOv3), intending to achieve the best trade-off between detection accuracy and speed. The ImYOLOv3 combines attention mechanism and dilated convolution for fast and accurate ship detection in optical remote sensing images. Due to

the speed, accuracy, and flexibility of original YOLOv3 [39], we adopt it as the inspiration of our ship detection framework. We argue that the YOLO series or other one-stage detectors remain appealing for real-time applications in remote sensing thanks to their high efficiency. To address the problems that YOLOv3 struggles with small objects and may produce more false alarms in complex backgrounds (such as clouds occlusion and strong waves), we integrate the attention mechanism into the detection network to highlight the difference between the ships and background. More specifically, we introduce a lightweight dilated attention module (DAM) which uses dilation convolution to enlarge the receptive fields and integrates both channel and spatial attention modules to extract salient features. To overcome the multi-scale characteristics of ship targets, we construct a multi-level feature pyramid and use the attention modules to obtain the salient features of different levels. At the end of the framework, we fuse the feature maps of different levels to further improve the detection accuracy and assign them to predict different scales of ship targets. In addition, we explicitly provide finer ship category information based on their sizes and moving states. Using these novel techniques, the proposed method is more suitable for ship detection task than original YOLOv3.

The main contributions of our paper are as follows:

(1) In view of the complex backgrounds encountered in optical remote sensing images, an end-to-end network structure named ImYOLOv3 is proposed for fast and accurate ship detection. We integrate the attention mechanism into the network to obtain discriminative feature maps at different levels and fuse corresponding multi-scale features, which ensures the effectiveness of detecting multi-scale ships in complex backgrounds.
(2) We design a novel and lightweight DAM which consists of three crucial cores: dilated block, channel attention sub-module, and spatial attention sub-module. The DAM can help our model to enlarge the receptive fields and highlight the difference between the ships and backgrounds, which overcomes the difficulty in detecting small ships.
(3) The proposed network is based on a one-stage object detection algorithm and can achieve high ship detection accuracy while maintaining a fast speed. Consequently, it can support real-time ship detection.
(4) We validate the proposed network on a challenging multi-class ship dataset (MSD) with huge scale variation. Additionally, unlike other ship detection datasets, our MSD includes four supervised categories, namely big ship, middle ship, small ship and moving ship, to investigate the detection effect of ships with different scales.

The reminder of this paper is organized as follows. The related work about CNN-based ship detection algorithms, attention mechanism, dilated convolution and classification strategy are described in Section 2. The framework and details of our proposed method are introduced in Section 3. The dataset setup and implementation details are shown in Section 4. Then, in Section 5, we present experiments conducted on MSD with the proposed ship detection framework. Finally, Section 6 concludes this paper.

2. Related Work

2.1. CNN-Based Ship Detection Methods

Researchers in the remote sensing community have made a lot of attempts to exploit CNN-based ship detection frameworks [26–32,43,45,46]. For instance, Zou et al. [45] presented SVDNet, which leveraged CNN to extract features and adopted feature pooling operation and linear SVM to determine the positions of ships. To further enhance the detection accuracy, Li et al. [27] introduced a hierarchical selective filtering (HSF) layer into Faster R-CNN [35] to generate features for multiscale ships and used an end-to-end training strategy by defining a joint loss. More recently, increasing attention has been paid to the time efficiency of ship detection algorithms. Some researchers try to transfer regression-based detection methods developed for natural images to remote sensing images. Liu et al. [31] designed an arbitrary-oriented ship detection framework based on YOLOv2 [38], which achieved robust and real-time detection in complex scenes. However, they may

struggle with ships of various scales due to a single-scale prediction. Chang et al. [46] also constructed a YOLOv2-based end-to-end training convolutional neural network to detect ships. In addition, they introduced a compact YOLOv2-reduced which had fewer layers for faster ship detection. Despite the improvement of detection speed, the YOLO-based methods suffer from a drop of the localization accuracy compared with two-stage detectors, especially for small ships. To achieve better trade-off between speed and accuracy, we propose DAM and incorporate it into YOLOv3 to improve the detection performance while maintaining its fast speed.

2.2. Attention Mechanism

Attention is an important tool in human perception [47,48] to bias the allocation of available processing resources towards the most salient part of input signals. Owing to its benefits, attention mechanism has been widely used in a lot of applications, from natural language processing [49] to image classification [50–54], image captioning [55] and image question answering [56]. In computer vision, Wang et al. [50] built "Residual Attention Network" by stacking attention modules with encoder-decoder style, which could quickly collect global information of the whole image and gradually refine the feature maps. Hu et al. [51] proposed a compact "Squeeze-and-Excitation" (SE) block to recalibrate the inter-channel attention. Closer to our work, Lin et al. [57] inserted SE module [51] into Faster R-CNN and designed rank operation to improve performance in SAR ship detection task. Beyond the channel attention, spatial attention is also explored in [52,53] to emphasize meaningful features along the spatial axis. However, Lin et al. [57] ignored the importance of spatial attention to emphasize salient objects. In this paper, we sequentially apply channel and spatial attention modules in our DAM, so that each module can learn what and where to attend in the channel and spatial axes respectively. Different from the attention modules in [52,53], we add dilated blocks in our DAM and leverage dilated convolutions [58] to increase receptive fields and adaptively adjust the receptive field sizes for objects of different scales, aiming to address the problem arising from scale variation and small ship instances.

2.3. Dilated Convolution

Dilated convolution, also named atrous convolution [59,60], was first introduced in the semantic segmentation task to incorporate large-scale context information [58,61]. It enlarges the convolutional kernel sizes with original weights by performing convolution at sparsely sampled locations, which increases the receptive field size without additional parameter cost. Dilated convolution has also been widely used in the field of object detection. To maintain the spatial resolution, DetNet [62] employed dilated convolution to design a specific detection backbone network and enlarge the receptive field. TridentNet [63] constructed a parallel multi-branch architecture with different receptive fields by using dilated convolution and generated scale-specific feature maps with a uniform representation power. In our ship detection framework, we use dilated convolution in our multi-branch architecture with different dilation rates to adapt the receptive fields for ships of different scales.

2.4. Classification Strategy

The main aim of ship detection is to locate and recognize true ships in remote sensing images. Many previous studies [15,30,45] in this field consider ship detection as a binary classification problem: ship and non-ship. However, as observed in [12], there usually exist large differences between the objects of interest within the same class. Simply using binary classification may have some detrimental impact on the detection performance. To address this problem, Qi et al. [13] divided ships into big ships and small ships according to their sizes. Li et al. [27] collected a ship detection dataset that included two categories, namely moving ship and static ship. Similarly, Shao et al. [64,65] also adopted multi-class classification for ship detection in visible light video images. They built a large-scale

ship dataset, called SeaShips, which covered six ship types (ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat, and passenger ship). Compared with video images, it is difficult to tell the exact categories of ships in remote sensing images due to the overhead perspective and long shooting distance. Consequently, we intuitively classify ship objects into four classes based on ship sizes and their moving states. The purpose is to alleviate the problem of scale variation and evaluate the detection accuracy for ships with different sizes.

3. Proposed Method

In this section, we elaborate the architecture of proposed ImYOLOv3 for fast and accurate ship detection in optical remote sensing images.

3.1. Network Architecture

As illustrated in Figure 1, our ImYOLOv3 mainly consists of three components: feature extraction module (FEM), dilated attention module (DAM), and class and bounding box prediction module (CBM). Firstly, the input images are resized to an appropriate size (taking 608 as an example). We employ Darknet-53 [39] as the backbone of FEM and propose DAM to extract salient features at different branches. Then, we construct a three-level feature pyramid sharing a similar concept with FPN [66]. Specifically, we upsample the salient feature maps of low resolution by a factor of 2 and merge them with previous feature maps of the same spatial size via concatenation. The feature fusion allows us to obtain more meaningful semantic information from the up-sampled feature maps and finer-grained information from the earlier feature maps. Our ImYOLOv3 predicts bounding boxes at three different scales. Behind the base FEM, we add convolutional sets to make further feature extraction based on these fused feature maps. The three branches have the same DAM structure but different receptive fields with the help of dilated convolutions [58,63]. Finally, we add a few more convolutional layers behind DAM of each branch and the last 1×1 convolutional layer in CBM is used to predict a 3D tensor encoding parameters of bounding box, object, and class predictions. In the multi-class ship detection experiments on MSD, we predict three boxes at each scale. Hence, the size of the predicted tensor is $N \times M \times [3 \times (4+1+4)]$ for four bounding box offsets, one object confidence, and four class probabilities, where N and M denote the height and width of the tensor, respectively. The final detection results are obtained after filtering the predicted boxes via non-maximum suppression (NMS).

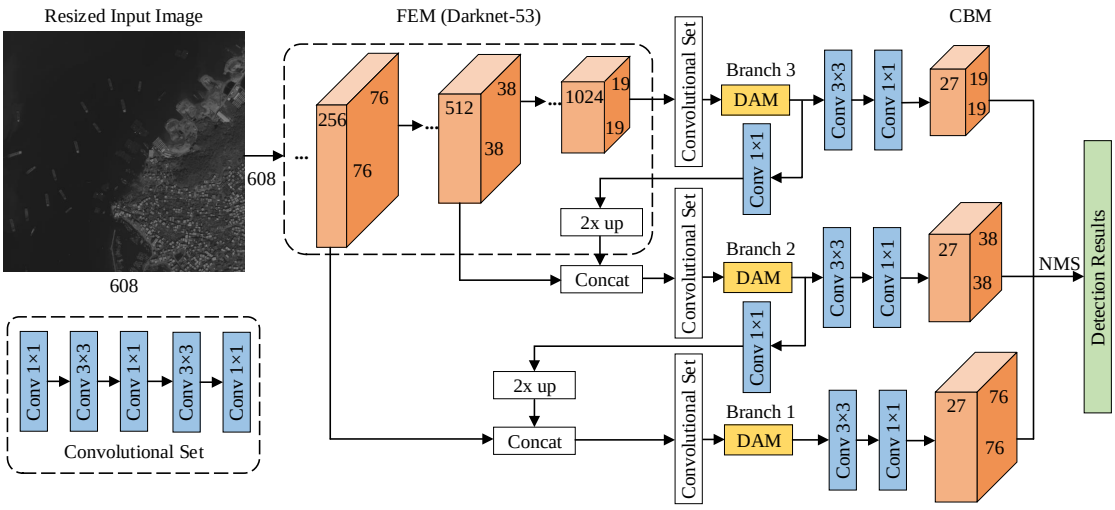

Figure 1. Architecture of ImYOLOv3 for ship detection.

3.2. Dilated Attention Module

As investigated in [63], the detection performance on objects of different scales is impacted by the receptive field of a network. Larger receptive field can capture a wider range of context information, which is beneficial for distinguishing small ship targets from complex backgrounds. In this section, we propose DAM which uses dilation convolution to enlarge the receptive fields and integrates both channel and spatial attention modules to extract salient feature maps, aiming at enhancing regions of ship targets and suppressing the interference of clouds and waves. As depicted in Figure 2, the DAM is mainly composed of three parts: dilated block, channel attention sub-module, and spatial attention sub-module.

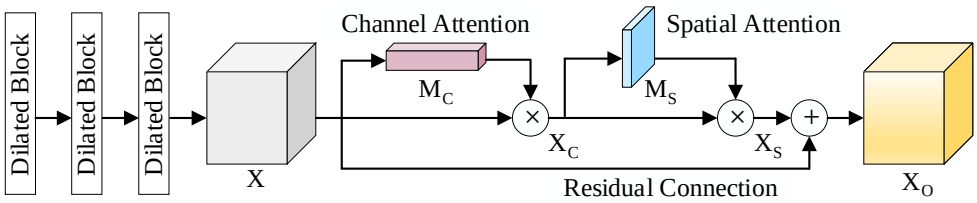

Figure 2. Dilated attention module (DAM) architecture.

The detailed structure of three sub-modules is illustrated in Figure 3. The dilated block shares similar style with the bottleneck [23] which consists of three convolutions with kernel size 1×1, 3×3, and 1×1. The difference is that we set different dilation rates for the 3×3 convolutional layers. Dilated convolution with dilation rate d inserts $(d-1)$ zeros between consecutive filter values, enlarging the kernel size without increasing the number of parameters and computation costs. Specifically, a dilated 3×3 convolution could have the same receptive field as the standard convolution with kernel size of $3 + 2 \times (d-1)$. Between two convolutional layers are batch normalization (BN) layer and leaky rectified linear unit (ReLU) layer sequentially.

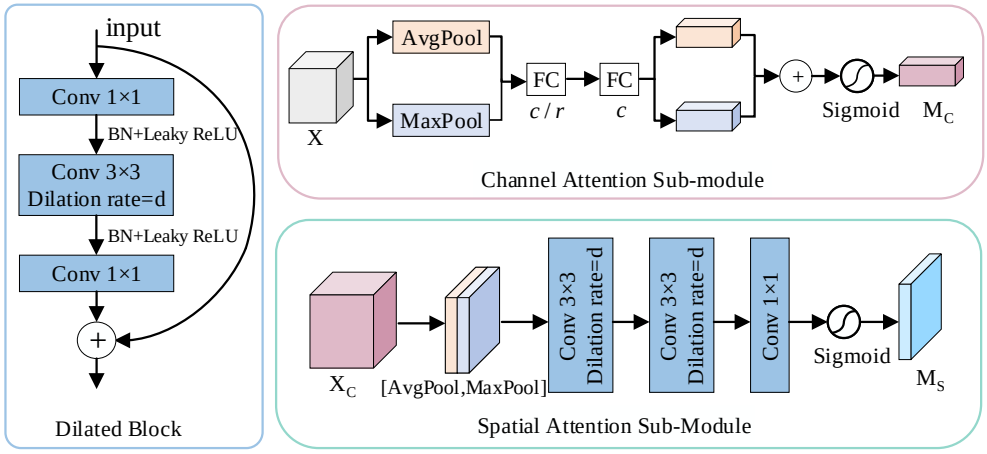

Figure 3. Diagram of each sub-module: dilated block, channel attention sub-module, and spatial attention sub-module.

Stacking dilated blocks allows us to modulate receptive fields of different branches and obtain an intermediate feature map $X \in \mathbb{R}^{C \times H \times W}$, where C is the channel, H is the height, and W is the width of feature maps. The subsequent two attention modules separately infer a 1D channel attention map $M_C \in \mathbb{R}^{C \times 1 \times 1}$ and a 2D spatial attention map $M_S \in \mathbb{R}^{1 \times H \times W}$. The whole process can be summarized as:

$$X_C = M_C \otimes X \tag{1}$$

$$X_S = M_S \otimes X_C \tag{2}$$

where $\otimes$ denotes element-wise multiplication, $X_C \in \mathbb{R}^{C\times H\times W}$ represents the channel-wise refined feature maps, and $X_S \in \mathbb{R}^{C\times H\times W}$ represents the spatial refined feature maps. Both M_C and M_S are expanded to $\mathbb{R}^{C\times H\times W}$ before multiplication. We compute the final refined feature map X_O via residual connection along with the attention mechanism to alleviate the vanishing-gradient problem as follows:

$$X_O = X_S + X \tag{3}$$

The following sections describe the detailed structure of each attention module.

Channel Attention Sub-module. Motivated by the success of attention mechanism in general deep neural networks, we adopt the channel attention module [52] after the dilated blocks to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. As shown in the top of Figure 3, the channel attention sub-module is fed with the intermediate feature map $X \in \mathbb{R}^{C\times H\times W}$ and produces a channel vector $M_C \in \mathbb{R}^{C\times 1\times 1}$. We first squeeze X along the spatial dimension $H \times W$ by using both average pooling and max pooling operations to obtain two channel descriptors ($\mathbb{R}^{C\times 1\times 1}$). The average pooling is to learn the extent of the target objects effectively and the max pooling is to gather important features about distinctive objects, which are helpful for ship detection task. Both descriptors are then forwarded to a shared network that consists of two fully connected (FC) layers. To reduce model complexity, the activation size of the first FC layer is set to $\mathbb{R}^{C/r\times 1\times 1}$, where r is the reduction ratio. After the shared network is applied to each channel descriptor, we merge the two output vectors via element-wise summation and employ a simple gating mechanism via sigmoid activation to obtain the final output M_C.

Spatial Attention Sub-module. Different from the channel attention that focuses on what is important in a given input image, we utilize the spatial attention as a supplement to learn where to attend for ship detection task. The structure of spatial attention sub-module is shown at the bottom of Figure 3. After obtaining the channel-wise refined feature map M_C, we also apply both average pooling and max pooling operations along the channel dimension and concatenate them to generate an efficient feature descriptor, which proves valid in highlighting informative regions [67]. Different from that in [52,53], we employ two 3×3 dilated convolutions to further enlarge the receptive fields and effectively leverage contextual information. Finally, the features are squeezed to $\mathbb{R}^{1\times H\times W}$ with 1×1 convolution, and the sigmoid function is used to obtain the final spatial attention map M_S.

4. Dataset and Implementation Details

4.1. Dataset

Recently, large-scale datasets have played important roles in data-driven research. To this end, we collect seven panchromatic images from GF-1 satellite and six panchromatic images from GF-2 satellite for training and testing. These images contain many kinds of landscapes, such as the ocean, the harbor, and the island, which are taken under different light and weather conditions. First, we cut the large panchromatic images into 1000×1000 sized patches with $0.2\times$ overlap and filter out images that do not contain ships. Then, these patches with ships are annotated by experts in aerial image interpretation via an open annotation tool LabelImg. Some example images are listed in Figure 4, which shows some typical scenes in optical remote sensing images, such as quiet sea, sea under cloud coverage, inshore land, moving ships with wakes, and sea with waves.

To facilitate multi-scale ship detection, we set four ship categories according to the length and moving state of targets, namely big ship (length more than 100 pixels), middle ship (length about 50–100 pixels), small ship (length about 10–50 pixels), and moving ship (ship with long wakes). Those extremely small ships (length less than 10 pixels) are excluded from the targets in our dataset. When labeling moving ships, the main parts of ships are inside the bounding boxes. In total, there are 1015 images in our multi-class ship dataset (MSD). We randomly choose 80% of the images for training and the remaining 20% for testing. An overview of MSD can be found in Table 1. To better show the scale variation

of ship targets, we draw a scatter plot according to the width and height of the ship targets in MSD dataset, as shown in Figure 5.

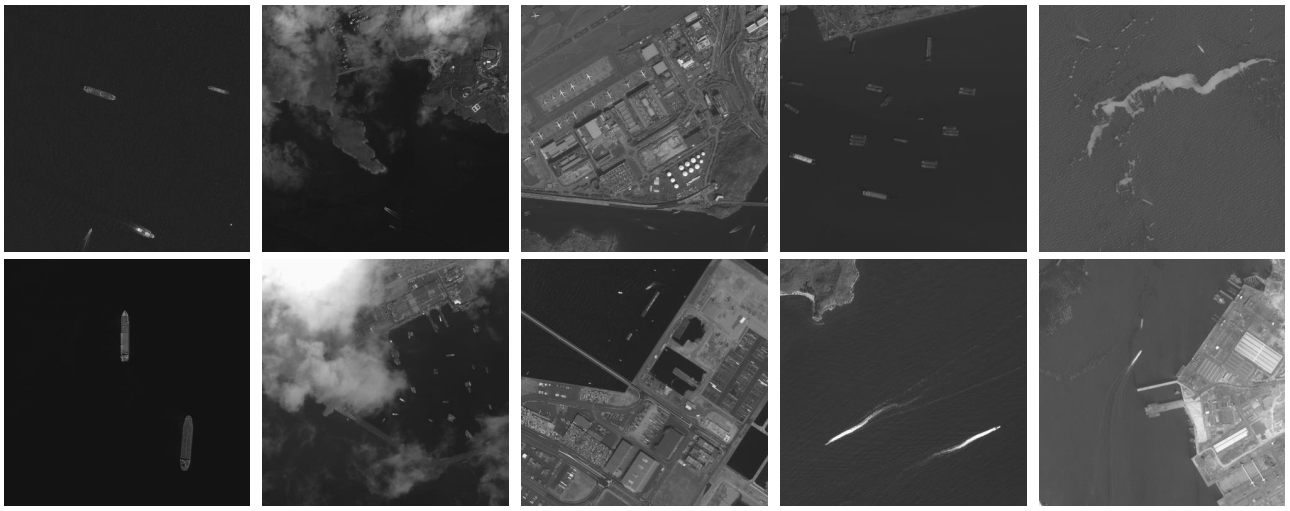

Figure 4. Some example images in our MSD dataset.

Table 1. Overview of our MSD.

Data Type	Images	Big Ship	Middle Ship	Small Ship	Moving Ship
Train	812	425	730	1240	367
Test	203	114	165	295	97

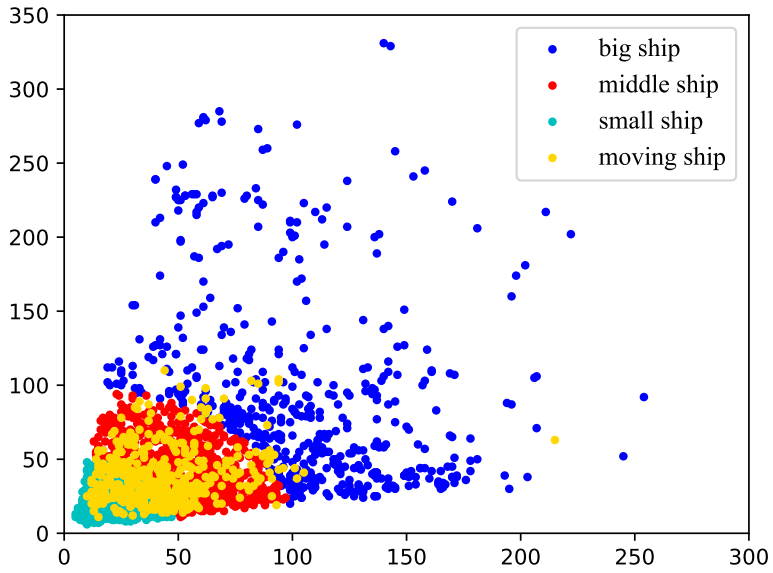

Figure 5. The sizes of ship targets in our dataset. As illustrated, ships have different scales and there exists large scale variations in MSD.

4.2. Evaluation Protocol

To quantitatively evaluate the detection performance of the network, we follow the same protocol as used by PASCAL VOC [68], which is briefly described below. Given the intersection-over-union (IoU) threshold, the detector will decide whether the detected box belongs to the background or not. The recall and precision are calculated as follows:

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{4}$$

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{5}$$

where N_{TP}, N_{FP}, and N_{FN}, respectively, stand for the number of accurately detected ship targets (true positive), the number of incorrectly detected targets (false positive), and the number of missing ship targets (false negative). For each category in our dataset, we can draw a precision–recall curve according to recall and precision values. The average precision (AP) means the area surrounded by the curve.

$$AP = \int_0^1 P(R)dR \tag{6}$$

Each class i corresponds to an AP value AP_i and mAP denotes their mean, which assesses the detection effect of the model.

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{7}$$

where n denotes the number of ship categories. As for the speed, we adopt FPS (frame per second) as indictor which symbolizes the frames that model can detect in one second.

4.3. Implementation Details

In our experiment, the dilation rates of dilated blocks are set to 1, 2, and 3 in Branch-1, Branch-2, and Branch-3, respectively. The reduction ratio r in the channel attention sum-module is set to 16 following Woo et al. [52]. All experiments were executed on a PC with NVIDIA GeForce GTX 1080ti. The initial learning rate was set to 0.001. We used stochastic gradient descent (SGD), with a weight decay of 0.0001 and momentum of 0.9. Warm-up was introduced during the initial training stage to avoid gradient explosion. All models were trained for 21,000 iterations, and the learning rate was divided by 10 at 7000 and 14,000 iterations. The IoU threshold of NMS was set to 0.5. For evaluation, we used mAP and FPS as the metrics.

5. Experimental Results and Discussion

5.1. Ablation Experiments

Number of dilated blocks. We conduct an ablation study to explore how many dilated blocks are needed for ImYOLOv3. The results are shown in Figure 6 as follows. When the number of dilated blocks grows beyond 3, the performance of ImYOLOv3 becomes stable. It indicates the robustness of our method with regards to the number of dilated blocks when the disparity of receptive fields between branches is large enough. Consequently, we set three dilated blocks in our DAM.

Performance of each branch. Our ImYOLOv3 has a multi-branch architecture, where each branch is responsible for detecting ships with different scales. Each branch has a different receptive field with the help of dilated convolution. Here, we conduct an additional experiment to study the mechanism about how to adjust the receptive field adaptively. Table 2 shows the results of three single branches and our three-branch method. As expected, through our training, Branch-1 with smallest receptive field achieves good results on small ships, Branch-2 works well on ships with middle scale, and Branch-3 with largest receptive field is good at detecting big ships. As a result, our three-branch ImYOLOv3 inherits the merits from three single branches and achieves the best result.

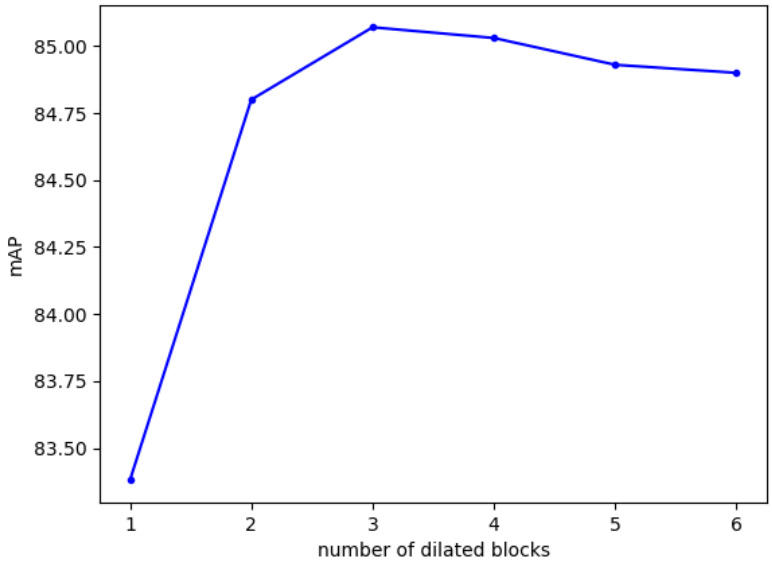

Figure 6. Ship detection results on the MSD dataset using different number of dilated blocks in DAM.

Table 2. Ship detection results of each branch in ImYOLOv3 evaluated on the MSD dataset. The dilation rates of three branches in ImYOLOv3 are set as 1, 2, and 3. The entries with the best APs for each object category are boldfaced and the suboptimal entries are underlined.

Method	Branch	Big Ship	Middle Ship	Small Ship	Moving Ship	mAP (%)
YOLOv3(baseline)	-	93.74	80.14	72.97	70.67	79.38
ImYOLOv3	Branch-1	78.90	84.37	79.83	72.59	78.92
	Branch-2	94.15	86.45	75.50	75.30	82.85
	Branch-3	94.86	77.95	63.46	71.29	76.89
	3 Branches	**95.01**	**86.69**	**80.97**	**77.60**	**85.07**

Performance of attention module. To verify the effectiveness of DAM, we conduct comparative experiments with other attention modules, and the results are listed in Table 3. We adopt YOLOv3 as the baseline and keep other hyper-parameters consistent in the experiments. It can be seen that adding the DAM into YOLOv3 brings 5.69% mAP increment (79.38% vs. 85.07%), which especially increases the AP for middle ship (80.14% vs. 86.69%), small ship (72.97% vs. 80.97%), and moving ship (70.67% vs. 77.60%). It implies that our DAM can extract discriminative and robust features for ship targets in remote sensing images, which effectively boosts the detection performance of small ships. When comparing the performance difference between YOLOv3 with SE block [51] and YOLOv3 with our DAM, the improvement of mAP (81.43% vs. 85.07%) indicates that adding spatial attention module is also very helpful for ship detection. In addition, we compare our DAM with the similar modules BAM [53] and CBAM [52] with both channel and spatial attention. The YOLOv3 with DAM achieves higher accuracy than YOLOv3 with BAM (85.07% vs. 82.55%) and YOLOv3 with CBAM (85.07% vs. 82.91%). The improvement is attributed to the benefits that our DAM enlarges receptive fields of the network through dilated convolutions and obtains richer feature representation. The above experiments demonstrate that our DAM is more suitable for the ship detection task than other off-the-shelf attention modules.

Table 3. The performance of attention modules on ship detection performance. The entries with the best APs for each object category are boldfaced.

Methods	Big Ship	Middle Ship	Small Ship	Moving Ship	mAP(%)
YOLOv3 (baseline) [39]	93.74	80.14	72.97	70.67	79.38
YOLOv3+SE [51]	93.45	82.40	76.83	73.04	81.43
YOLOv3+BAM [53]	94.05	83.59	77.61	74.95	82.55
YOLOv3+CBAM [52]	94.26	84.30	77.84	75.24	82.91
YOLOv3+DAM (ours)	**95.01**	**86.69**	**80.97**	**77.60**	**85.07**

In addition, as shown in Table 3, YOLOv3+DAM achieves larger AP increment for small ship (8.00%), moving ship (6.93%), and middle ship (6.55%) than for big ship (1.27%), compared with the baseline YOLOv3. Since the intention of our DAM is to boost the performance of original YOLOv3 for small objects, we further show some detection results compared with original YOLOv3 [39] to verify the validity of our model. Figure 7 shows the detection results of original YOLOv3 (the first row) and our ImYOLOv3 (the second row). The boxes in different colors represent different ship categories and the yellow circles denote missing ships. As is shown in the first column of Figure 7, YOLOv3 fails to detect some small ships and moving ships while our ImYOLOv3 can detect them successfully. Detecting ships under clouds coverage or with waves remains a challenging problem in remote sensing images. In the second and third columns of Figure 7, YOLOv3 misses the small ship under cloud coverage and the middle ship surrounded by waves. However, our ImYOLOv3 can effectively distinguish them from the complex background, which shows that our method is very robust in complex scenes. The proposed attention modules can greatly help to highlight the difference between the ships and background. The performance benefits not only from the proposed dilated attention module but also from the better feature representation for ship targets.

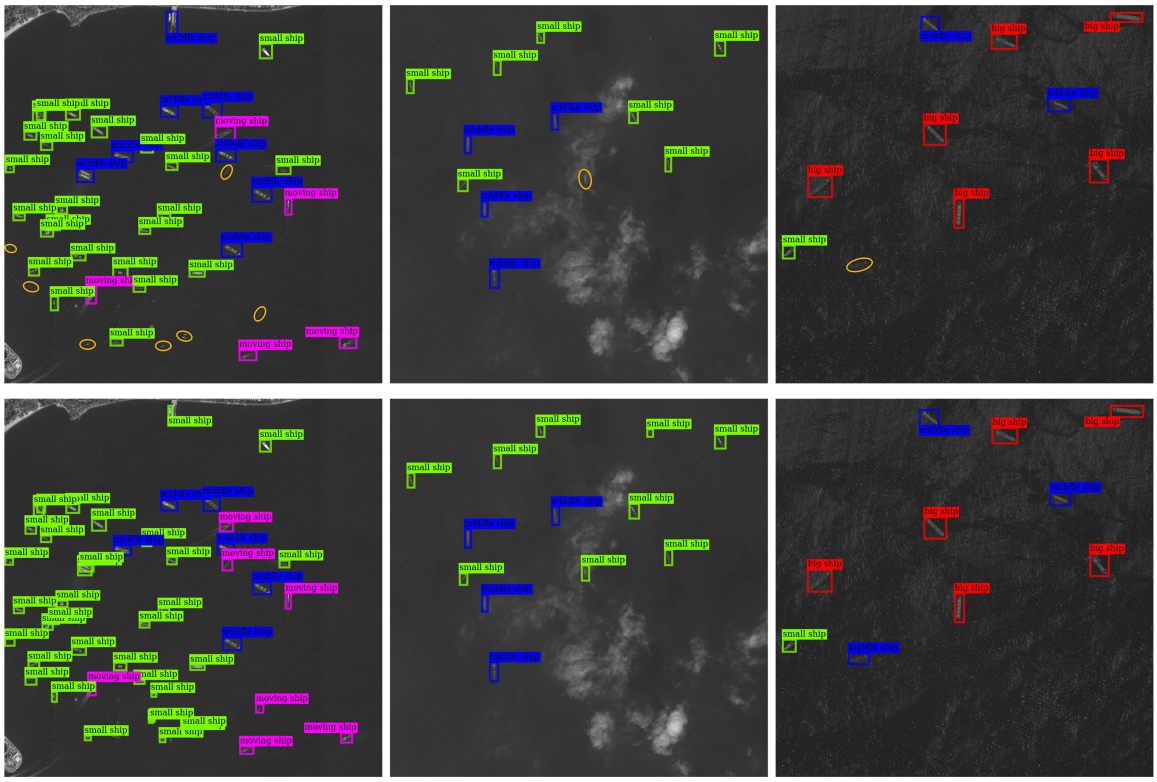

Figure 7. Detection results of YOLOv3 (the first row) and our ImYOLOv3 (the second row). Yellow circles denote missing ships.

5.2. Comparison with the State-of-the-Art Methods

In this section, we compare our ImYOLOv3 with other state-of-the-art object detection methods, including single shot multi-box detector (SSD) [36], feature pyramid network (FPN) [66], and RetinaNet [40]. The training settings follow the original references. Table 4 shows the comparative results, while Figure 8 shows the precision–recall curves of each ship category.

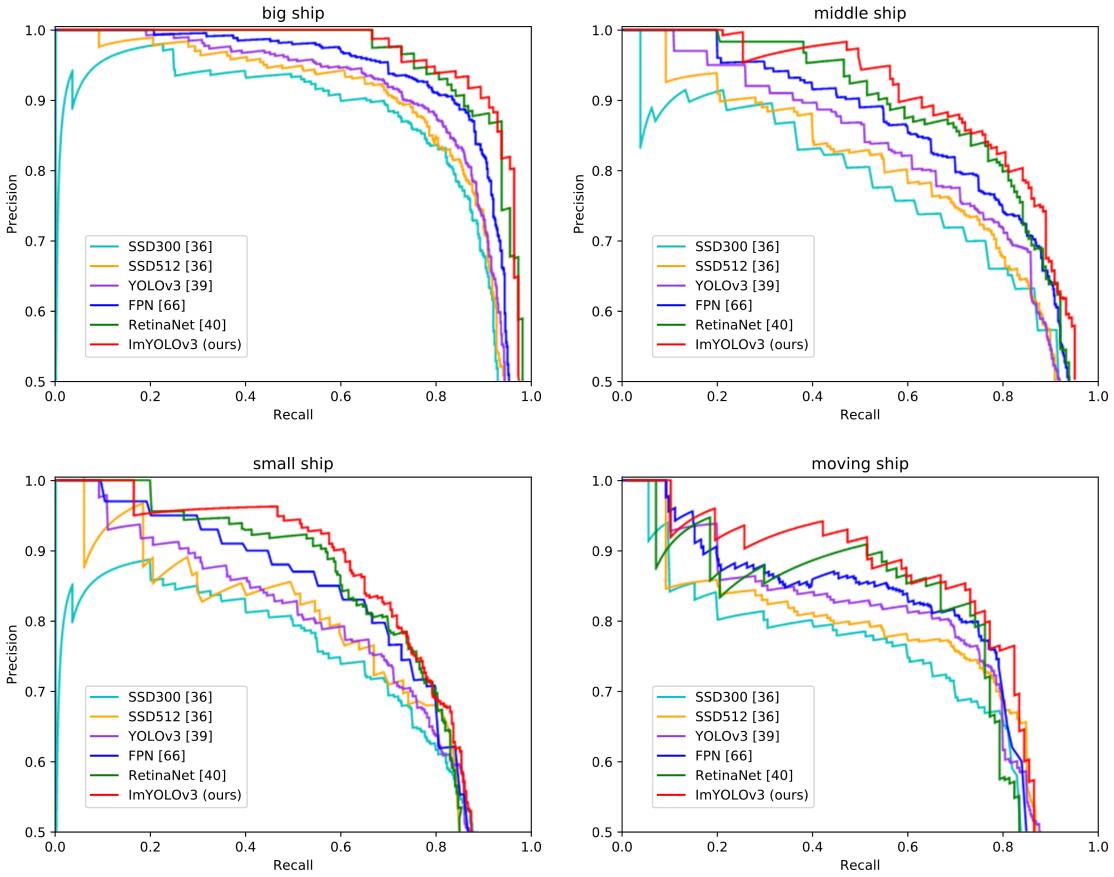

Figure 8. Precision–recall curves of different detection methods on four ship types.

As shown in Table 4, our ImYOLOv3 clearly outperforms SSD300 by 9.64% mAP (85.07% vs. 75.43%) and SSD512 by 7.10% mAP (85.07% vs. 77.97%). As for the detection speed, ImYOLOv3 is a little slower than SSD300 (28 FPS vs. 46 FPS), but it is faster than SSD512 (28 FPS vs. 19 FPS). Compared with two-stage model FPN, our ImYOLOv3 achieves 3.28% mAP increments, which improves big ship AP by 0.44%, middle ship AP by 3.4%, small ship AP by 2.93%, and moving ship AP by 6.34%, while being much faster (28 FPS vs. 6 FPS). Moreover, our model surpasses one-stage RetinaNet by 1.52% mAP, while being 2.8× faster. It should be noted that, because FPN and RetinaNet achieve better performance than original YOLOv3 (79.38% mAP), the gains of ImYOLOv3 can be only attributed to the effectiveness of the proposed DAM. To sum up, our ImYOLOv3 achieves a better trade-off between the detection accuracy and detection speed than the state-of-the-art methods. The contribution of ImYOLOv3 is not only a more powerful attention module but also a better understanding of what (channel-wise attention) and where (spatial attention) the ship object is in an input image. We believe that the efficiency and simplicity of our proposed method will benefit future research and ship detection applications.

Table 4. Detection average precision (%) and speed of different methods on MSD. The entries with the best APs for each object category are boldfaced.

Methods	Big Ship	Middle Ship	Small Ship	Moving Ship	mAP(%)	FPS
SSD300 [36]	89.83	77.50	68.95	65.44	75.43	**46**
SSD512 [36]	91.94	78.20	72.79	68.95	77.97	19
YOLOv3 [39]	93.74	80.14	72.97	70.67	79.38	29
FPN [66]	94.57	83.29	78.04	71.26	81.79	6
RetinaNet [40]	94.96	86.07	79.70	73.46	83.55	10
ImYOLOv3	**95.01**	**86.69**	**80.97**	**77.60**	**85.07**	28

5.3. Detection Performance in Different Image Backgrounds

To show the robustness of our model, we test it in some typical remote-sensing scenes and the detection results are shown in Figure 9. The different environmental conditions include quiet sea, sea with waves, sea under clouds coverage, and inshore land (from top to the bottom in Figure 9). The boxes in different colors represent different ship categories. As shown in the first row, we can see that our ImYOLOv3 performs well in quiet sea, which can successfully detect ships with different scales. The second row represents results for ship target detection in sea with waves. The third row shows results for ship detection under clouds coverage, which are difficult problems for ship detection in remote sensing images due to the interference of strong waves and clouds. However, our proposed method still effectively distinguishes ship targets even with the interference of waves and clouds and accurately locates ship targets with long wakes. The fourth row presents some detection results for inshore land, where some buildings or harbors may have similar shapes and sizes to ship targets. It is clear that our ImYOLOv3 accurately recognizes ship targets in such complicated scenes even though some ships are very small. In short, our method is very robust in various complex scenes, which benefits from the proposed dilated attention module and better feature representation for multi-scale ship detection.

5.4. Generalization Ability Testing

To further illustrate the generalization ability of the proposed method, we evaluate its performance on SeaShips dataset [64], which consists of 31,455 images and cover six common ship types (ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat, and passenger ship). The similarity between SeaShips and MSD is that both datasets provide fine-grained ship category information to overcome the intra-class differences because different ship types vary greatly in shape and appearance. The difference between SeaShips and our MSD is that all of the images in SeaShips dataset are from real-world video segments, which are acquired by the monitoring cameras in a deployed coastline video surveillance system, while our MSD is based on optical remote sensing images.

We compare our ImYOLOv3 with other baseline detectors including Fast R-CNN [34], Faster R-CNN [35], SSD [36], and YOLOv3 [39]. The results are shown in Table 5. For Fast R-CNN, we adopt VGG16 net [25] as the backbone network. For Faster R-CNN [35], we utilize ZFNet [69], VGG16 net [25], and ResNet (ResNet50 and ResNet101) [23] as the base network architectures. For SSD300 and SSD512 [36], we use VGG16 net as the backbone network. To ensure fair comparison, we keep the same training settings in each experiment. As shown in Table 5, Fast R-CNN performs worse than other detectors by a wide margin in terms of mAP and has a lower speed (3 FPS). Faster R-CNN achieves better detection accuracy when we use deeper networks, and the Faster R-CNN series all outperform SSD series. Surprisingly, YOLOv3 with Darknet-53 obtains better performance than Faster R-CNN with ResNet101 by 0.62% mAP (93.02% vs. 92.40%), while being 4.8× faster (29 FPS vs. 6 FPS). We conjecture that this is due to the benefits of multi-scale training strategy and because the ship targets in SeaShips dataset are relatively large objects and easy to be detected. It should be noted that the average precision (AP) for six ship types increases when we add DAM into YOLOv3, which improves ore carrier AP by 0.79%, bulk cargo carrier AP by 0.61%, general cargo ship AP by 0.15%, container ship AP by 0.60%, fishing boat AP by 0.97%, and passenger ship AP by 0.92%. Moreover, our

ImYOLOv3 achieves a real-time speed (28 FPS). Considering the same experimental setup, the improvements can only be attributed to the better feature extraction of our DAM. The above experiments demonstrate that our ImYOLOv3 has great generalization ability and detection performance in both remote sensing images and natural images. Furthermore, our ImYOLOv3 achieves a better trade-off between the detection accuracy and speed than other detection methods.

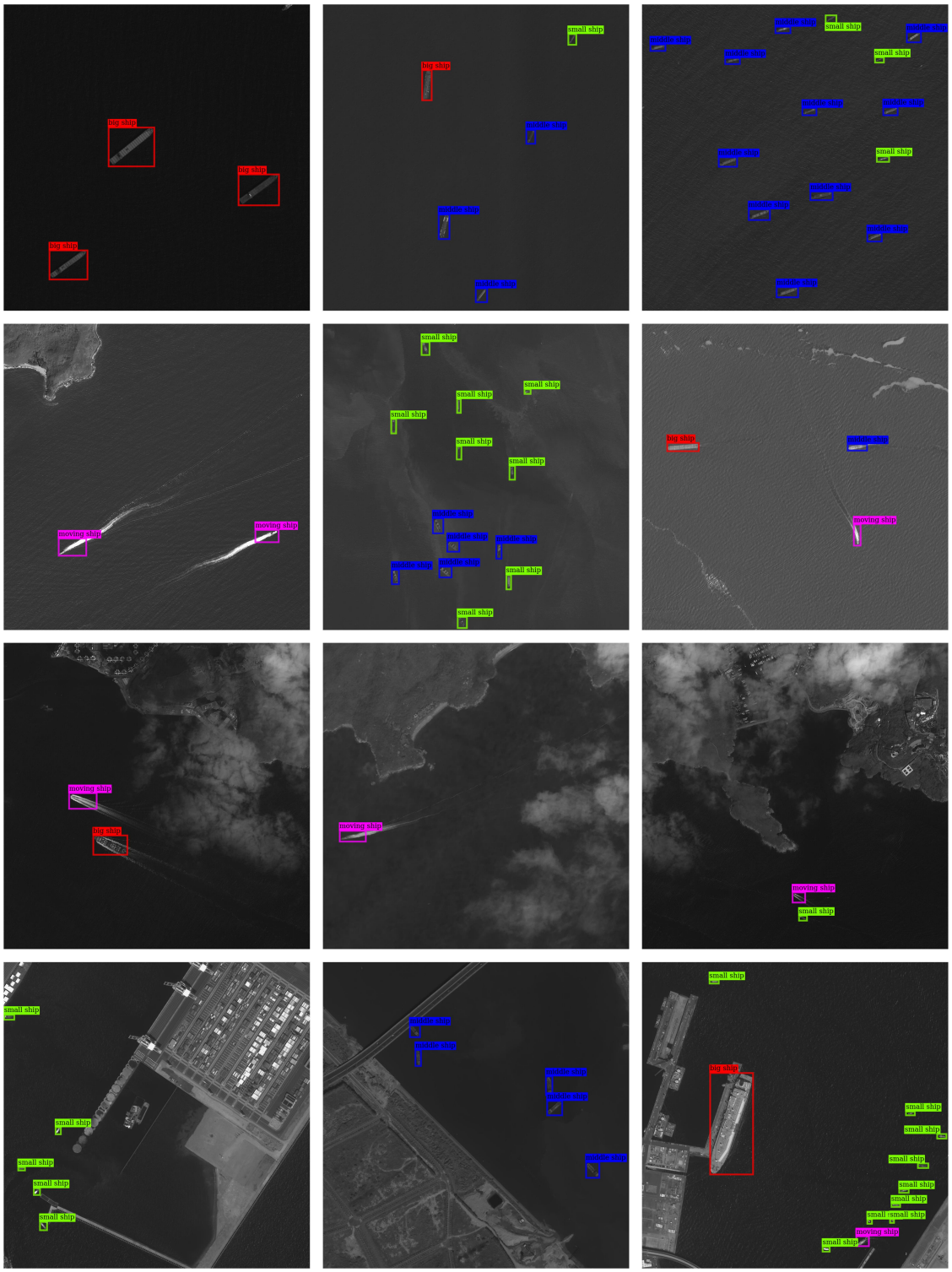

Figure 9. Ship detection results of our ImYOLOv3 in some typical scenes.

Table 5. Detection average precision (%) and speed of different methods on SeaShips. c1–c6 represent ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat, and passenger ship, respectively. The entries with the best APs for each object category are boldfaced.

Methods	Backbone	c1	c2	c3	c4	c5	c6	mAP(%)	FPS
Fast R-CNN [34]	VGG16	77.09	71.33	77.05	86.81	61.70	52.20	71.03	3
Faster R-CNN [35]	ZFNet	90.50	90.01	90.77	90.91	85.68	87.06	89.16	17
Faster R-CNN [35]	VGG16	89.44	90.34	90.73	90.87	88.76	90.57	90.12	5
Faster R-CNN [35]	ResNet50	92.38	90.88	92.46	92.91	89.27	90.93	91.65	7
Faster R-CNN [35]	ResNet101	93.68	90.22	93.87	93.41	89.96	91.78	92.40	6
SSD300 [36]	VGG16	75.03	76.66	87.66	90.71	71.79	74.35	79.37	**46**
SSD512 [36]	VGG16	83.99	83.00	87.08	90.81	85.85	89.65	86.73	19
YOLOv3 [39]	Darknet-53	94.55	93.47	95.99	97.47	89.28	87.34	93.02	29
ImYOLOv3 (ours)	Darknet-53	**95.34**	**94.08**	**96.14**	**98.07**	**90.25**	**88.26**	**93.69**	28

6. Conclusions

In this paper, we propose a novel one-stage ship detection algorithm based on attention mechanism called ImYOLOv3, aiming to support further research in optical remote sensing ship detection field. First, to realize fast ship detection, we adopt the off-the-shelf YOLOv3 as the basic detection framework due to its fast speed and utilize Darknet-53 for feature extraction. Then, to alleviate the problem of multi-scale ship detection, we construct a three-level feature pyramid sharing a similar concept with FPN, which combines meaningful semantic information from the up-sampled feature maps and fine-grained information from the earlier feature maps. Finally, to achieve accurate ship detection, we design DAM which can be integrated into the backbone framework to highlight the difference between ship targets and background. Our DAM can help to extract more discriminative features for ship targets and overcome the interference of clouds, mists, and waves. More specifically, the channel-wise attention is combined with spatial attention to learn what and where to focus or suppress in an input image and adaptively refine the intermediate features. With the assistance of dilated convolutions, we enlarge the receptive fields of our model to further improve the ship detection accuracy. In addition, we build a multi-class ship dataset to facilitate multi-scale ship detection. Extensive experimental results show that our ImYOLOv3 has promising potentials on detecting ships with different scales and different moving states in complex backgrounds, while achieving a fast speed. We believe that ImYOLOv3 could benefit current ship detection and other vision tasks. We would like to explore it in the future work.

Author Contributions: L.C. and W.S. conceived and designed the idea; L.C. performed the experiments; W.S. analyzed the data and helped with validation; L.C. wrote the paper; and D.D. supervised the study and reviewed this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by National Natural Science Foundation of China (No. 61501334).

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions.

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

References

1. Fan, Y.; Wen, Q.; Wang, W.; Wang, P.; Li, L.; Zhang, P. Quantifying Disaster Physical Damage Using Remote Sensing Data—A Technical Work Flow and Case Study of the 2014 Ludian Earthquake in China. *Int. J. Disaster Risk Sci.* **2017**, *8*, 1–18. [CrossRef]
2. Martinuzzi, S.; Gould, W.A.; González, O.M.R. Land development, land use, and urban sprawl in Puerto Rico integrating remote sensing and population census data. *Landsc. Urban Plan.* **2007**, *79*, 288–297. [CrossRef]
3. Chen, X.; Xiang, S.; Liu, C.; Pan, C. Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *11*, 1797–1801. [CrossRef]
4. Kalantar, B.; Mansor, S.B.; Halin, A.A.; Shafri, H.Z.M.; Zand, M. Multiple Moving Object Detection From UAV Videos Using Trajectories of Matched Regional Adjacency Graphs. *IEEE Trans. Geosci. Remote. Sens.* **2017**, *55*, 5198–5213. [CrossRef]

5. Durieux, L.; Lagabrielle, E.; Nelson, A. A method for monitoring building construction in urban sprawl areas using object-based analysis of Spot 5 images and existing GIS data. *Isprs J. Photogramm. Remote Sens.* **2008**, *63*, 399–408. [CrossRef]
6. 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]
7. Armando, M.; Maria, S.F.; Irena, H.; Kazuo, O. Ship Detection with Spectral Analysis of Synthetic Aperture Radar: A Comparison of New and Well-Known Algorithms. *Remote Sens.* **2015**, *7*, 5416–5439.
8. 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]
9. 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]
10. Huang, X.; Yang, W.; Zhang, H.; Xia, G.S. Automatic Ship Detection in SAR Images Using Multi-Scale Heterogeneities and an A Contrario Decision. *Remote Sens.* **2015**, *7*, 7695–7711. [CrossRef]
11. Chen, C.; He, C.; Hu, C.; Pei, H.; Jiao, L. A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios. *IEEE Access* **2019**, *7*, 104848–104863. [CrossRef]
12. Zhu, C.; Zhou, H.; Wang, R.; Guo, J. A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3446–3456. [CrossRef]
13. Qi, S.; Ma, J.; Lin, J.; Li, Y.; Tian, J. Unsupervised ship detection based on saliency and S-HOG descriptor from optical satellite images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1451–1455.
14. Tang, J.; Deng, C.; Huang, G.B.; Zhao, B. Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 1174–1185. [CrossRef]
15. Nie, T.; He, B.; Bi, G.; Zhang, Y.; Wang, W. A method of ship detection under complex background. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 159. [CrossRef]
16. Chen, L.; Shi, W.; Fan, C.; Zou, L.; Deng, D. A Novel Coarse-to-Fine Method of Ship Detection in Optical Remote Sensing Images Based on a Deep Residual Dense Network. *Remote Sens.* **2020**, *12*, 3115. [CrossRef]
17. Wang, W.; Fu, Y.; Dong, F.; Li, F. Remote sensing ship detection technology based on DoG preprocessing and shape features. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; pp. 1702–1706.
18. Yang, F.; Xu, Q.; Li, B. Ship detection from optical satellite images based on saliency segmentation and structure-LBP feature. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 602–606. [CrossRef]
19. Xia, Y.; Wan, S.; Yue, L. A novel algorithm for ship detection based on dynamic fusion model of multi-feature and support vector machine. In Proceedings of the 2011 Sixth International Conference on Image and Graphics, Hefei, China, 12–15 August 2011; pp. 521–526.
20. Xu, J.; Sun, X.; Zhang, D.; Fu, K. Automatic detection of inshore ships in high-resolution remote sensing images using robust invariant generalized Hough transform. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 2070–2074.
21. Shi, Z.; Yu, X.; Jiang, Z.; Li, B. Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 4511–4523.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
23. 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.
24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
25. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
26. Lin, H.; Shi, Z.; Zou, Z. Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1665–1669. [CrossRef]
27. Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X.X. HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7147–7161. [CrossRef]
28. Zhang, R.; Yao, J.; Zhang, K.; Feng, C.; Zhang, J. S-CNN-Based Ship Detection from High-Resolution Remote Sensing Images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, *423–430* . [CrossRef]
29. 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, 19–21 May 2017; pp. 1–4.
30. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
31. Liu, W.; Ma, L.; Chen, H. Arbitrary-oriented ship detection framework in optical remote-sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 937–941. [CrossRef]
32. Yao, Y.; Jiang, Z.; Zhang, H.; Zhao, D.; Cai, B. Ship detection in optical remote sensing images based on deep convolutional neural networks. *J. Appl. Remote Sens.* **2017**, *11*, 042611. [CrossRef]

33. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the CVPR, Columbus, OH, USA, 24–27 June 2014.
34. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
35. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
36. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Amsterdam, The Netherlands, 2016; pp. 21–37.
37. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
38. 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.
39. Joseph, R.; Ali, F. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
40. 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.
41. Tang, G.; Liu, S.; Iwao, F.; Claramunt, C.; Wang, Y.; Men, S. H-YOLO: A Single-Shot Ship Detection Approach Based on Region of Interest Preselected Network. *Remote Sens.* **2020**, *12*, 4192. [CrossRef]
42. Nie, T.; Han, X.; He, B.; Li, X.; Bi, G. Ship Detection in Panchromatic Optical Remote Sensing Images Based on Visual Saliency and Multi-Dimensional Feature Description. *Remote Sens.* **2020**, *12*, 152. [CrossRef]
43. Van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv* **2018**, arXiv:1805.09512.
44. Nina, W.; Condori, W.; Machaca, V.; Villegas, J.; Castro, E. Small Ship Detection on Optical Satellite Imagery with YOLO and YOLT. In *Future of Information and Communication Conference*; Springer: San Francisco, CA, USA, 2020; pp. 664–677.
45. Zou, Z.; Shi, Z. Ship detection in spaceborne optical image with SVD networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5832–5845. [CrossRef]
46. Chang, Y.L.; Anagaw, A.; Chang, L.; Wang, Y.; Hsiao, C.Y.; Lee, W.H. Ship detection based on YOLOv2 for SAR imagery. *Remote Sens.* **2019**, *11*, 786. [CrossRef]
47. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
48. Corbetta, M.; Shulman, G.L. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **2002**, *3*, 201–215. [CrossRef]
49. Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; Zhang, C. Disan: Directional self-attention network for rnn/cnn-free language understanding. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
50. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
51. 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, 18–22 June 2018; pp. 7132–7141.
52. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
53. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
54. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
55. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
56. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
57. Lin, Z.; Ji, K.; Leng, X.; Kuang, G. Squeeze and excitation rank faster R-CNN for ship detection in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 751–755. [CrossRef]
58. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the ICLR, San Juan, Puerto Rico, 2–4 May 2016.
59. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
60. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.

61. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
62. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Detnet: Design backbone for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 334–350.
63. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6054–6063.
64. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [CrossRef]
65. Shao, Z.; Wang, L.; Wang, Z.; Du, W.; Wu, W. Saliency-aware convolution neural network for ship detection in surveillance video. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 781–794. [CrossRef]
66. 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.
67. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
68. 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]
69. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Zurich, Switzerland, 2014; pp. 818–833.