remote sensing

MDPI

Technical Note

Ship Detection via Dilated Rate Search and Attention-Guided Feature Representation

Jianming Hu †, Xiyang Zhi *,†, Tianjun Shi, Lijian Yu and Wei Zhang

Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China; hujianming@stu.hit.edu.cn (J.H.); 20S021010@stu.hit.edu.cn (T.S.); 19B921013@stu.hit.edu.cn (L.Y.); wzhang@hit.edu.cn (W.Z.)

* Correspondence: zhixiyang@hit.edu.cn; Tel.: +86-0451-86414883

† These authors contributed equally to this work.

Abstract: Due to the complexity of scene interference and the variability of ship scale and position, automatic ship detection in remote sensing images makes for challenging research. The existing deep networks rarely design receptive fields that fit the target scale based on training data. Moreover, most of them ignore the effective retention of position information in the feature extraction process, which reduces the contribution of features to subsequent classification. To overcome these limitations, we propose a novel ship detection framework combining the dilated rate selection and attention-guided feature representation strategies, which can efficiently detect ships of different scales under the interference of complex environments such as clouds, sea clutter and mist. Specifically, we present a dilated convolution parameter search strategy to adaptively select the dilated rate for the multi-branch extraction architecture, adaptively obtaining context information of different receptive fields without sacrificing the image resolution. Moreover, to enhance the spatial position information of the feature maps, we calculate the correlation of spatial points from the vertical and horizontal directions and embed it into the channel compression coding process, thus generating the multi-dimensional feature descriptors which are sensitive to direction and position characteristics of ships. Experimental results on the Airbus dataset demonstrate that the proposed method achieves state-of-the-art performance compared with other detection models.

Keywords: deep network; dilated rate selection; attention-guided feature representation; ship detection; optical remote sensing

Citation: Hu, J.; Zhi, X.; Shi, T.; Yu, L.; Zhang, W. Ship Detection via Dilated Rate Search and Attention-Guided Feature Representation. *Remote Sens.* **2021**, *13*, 4840. https://doi.org/10.3390/rs13234840

Academic Editors: Mi Wang, Hanwen Yu, Jianlai Chen and Ying Zhu

Received: 11 October 2021
Accepted: 26 November 2021
Published: 29 November 2021

1. Introduction

Ship detection is of great significance to maritime transportation, port management, disaster rescue and other activities. At present, optical remote sensing satellite image has become the main input data of ship detection models because of its wide detection range, rich spectral characteristics and high image resolution [1]. Although researchers have done lots of work on ship detection algorithms in optical images, there are still many challenges in ship detection applications, mainly due to the complexity of scene interference factors, such as clouds, waves and so on [2,3]. In addition, the variability of the scale and shape of the ship target also puts forward new requirements for the detection model. Therefore, in the face of complex environmental interference and the multi-scale change of the ship itself, the detection models still have great research potential.

With the development of deep network, the learning-based method based on a convolutional neural network (CNN) has become the main idea in the field of ship detection [4]. At present, various network architectures have been proposed and have achieved excellent performance in detection tasks under different application conditions [5]. Actually, the deep network usually employs a convolution kernel for multiple down-sampling, thus having the advantage of multi-level extraction of target features [6]. However, the design of down-sampling in classical CNN structure also brings some defects in term of information

transmission. Specifically, although the application of down-sampling operations can expand the receptive field and reduce the calculation amount of data processing, it leads to the loss of scene information in spatial dimensions. This loss is difficult to ignore especially for small-scale object detection because after multi-layer sampling, the information of small targets cannot be reconstructed theoretically.

In order to expand the receptive field without sacrificing the resolution, the dilated convolution strategy [7] is applied to the object detection model in recent years. Compared with the normal convolution kernel, the improvement of the dilated convolution is defining a spacing between the values in a kernel. When calculating, since only non-zero elements in the kernel actually affect the output, the dilated convolution obtains a larger receptive field than the normal convolution on the premise of avoiding the down-sampling operation. Moreover, by setting different dilated rates, we can obtain multi-scale context information, which has been proved to be conducive to the localization and classification of small-scale targets [8].

However, adopting the dilated convolution to solve the problem of multi-scale object detection in remote sensing images still faces major challenges. Specifically, since the positions and scales of the detected objects in the image are quite different, it is challenging to select the appropriate kernel size and dilated rate for the convolution operation when designing the dilated convolution. The images with more global information distribution are more suitable when using a large dilated rate, and those with more local distribution tend to apply a small dilated rate. This situation leads to the fact that the previous excellent dilated convolutions are often designed by experienced researchers. In addition, many designed frameworks are often applied for specific scenes, which have poor applicability to other remote sensing application tasks. Obviously, it is difficult to meet the above conditions in many practical applications. As we know, simply stacking convolution layers of different sizes will waste computing resources, and the networks with many layers have difficulty avoiding the problem of overfitting. Therefore, how to adaptively select the appropriate parameters of the dilated convolution in remote sensing ship detection applications is an urgent problem to be solved.

In addition, after extracting features of different scales, the mainstream models widely adopt an attention mechanism to strengthen the response of feature maps in different dimensions. This strategy can make the detection model learn what content and where to pay more attention, which is crucial for target detection under complex environmental interference. Nevertheless, these attention models either focus on capturing the correlation of channel dimensions or mining the long-range dependence of spatial dimensions, rarely organically integrating the features of the two dimensions based on information coding mechanism. Therefore, it is necessary to study how to make full use of the characteristics of different dimensions, thus providing more effective information for subsequent object classification.

To solve these problems mentioned above, we propose a novel ship detection method combining the dilated rate selection and attention-guided feature representation strategies. Specifically, we employ a dilated convolution parameter search strategy to adaptively select the dilated rate, obtaining the multi-scale and multi-receptive-field characteristic information of the ship target. On this basis, to enhance the spatial position information of the feature maps, we calculate the correlation between two orthogonal directions in the spatial domain and embed it into the channel information coding, thus generating the multi-dimensional feature descriptors which are sensitive to direction and position characteristics of ships. Extensive experiments conducted on a high-quality dataset demonstrate the effectiveness of our method.

The main contributions of our work are summarized as follows:

- A novel framework for ship detection is proposed, which can efficiently detect ships of different scales under the interference of complex environments such as clouds, sea clutter and mist.

- A multi-branch dilated rate search architecture is presented, which adaptively captures target context information of different scales and different receptive fields.
- An attention-wise feature extraction strategy is adopted, which enhances the representation of feature map by encoding the spatial position information.

The remainder of this paper is organized as follows: we introduce the development of a dilated convolution strategy and attention mechanism in a learning-based network in Section 2. Section 3 gives the design details of two main modules in the proposed model. In Section 4, we conduct comparative experiments based on a high-quality public dataset and analyze the experimental results of different compared methods. Section 5 provides the final conclusion of our work.

2. Previous Related Research

In this section, we first make a brief review of the development of dilated convolution strategy for receptive field problem; then, we introduce the improvement of attention-based representation, which significantly enhances the optimization of feature maps. By discussing the motives and shortcomings of these existing studies, we illustrate the differences between our method and other advanced methods.

2.1. Dilated Rate Strategy for Object Detection

In the past decade, due to the significant advantages of feature representation, a variety of deep neural networks have been proposed to solve different detection tasks. Compared with other deep neural networks [9,10], the convolution operation in CNN architecture can preserve the local correlation of image pixels (i.e., context information) and the spatial invariance of features. Moreover, due to the weight sharing strategy, this architecture reduces the number of model parameters, thus optimizing the complexity of the whole model. In addition, from the perspective of neuroscience, the multi-level extraction mechanism of the CNN model is in line with the habit of biological vision system, that is, we first pay attention to the shallow structural features such as edges and corners and then focus on more complex details such as texture, forming the overall concept of the object. Therefore, the CNN-based architecture has received the enthusiasm of researchers [11,12] and is widely used in image understanding and object classification applications.

In remote sensing images, there is usually a correlation between objects and the background environment. For target detection issue, mining this characteristic is conducive to improving the performance of object location and classification. Actually, before the rise of deep learning, studies have proved that appropriate context modeling can improve the performance of detection algorithms, especially for small-scale targets, which have no salient appearance characteristics. With the popularity of CNN-based networks, researchers have also proposed various models to expand the receptive field, integrating different scales of context information into the feature extraction network.

The receptive field can be regarded as the regional range of the input image learned by the unit of the CNN-based network. The larger the receptive field, the more global and semantic features can be learned by the neuron unit. The core idea of the dilated convolution is adding some zero values that do not participate in the calculation into the normal convolution kernel. In other words, the dilated convolution does not increase the amount of calculation involved in feature extraction. Moreover, because there is no down-sampling step, the convolution processing does not affect the resolution of the output image on the premise of increasing receptive field. In summary, through dilated convolution, we can achieve that the sizes of the output feature remain unchanged, and the output maps integrate multi-scale information. Because of the above advantages, dilated convolutions are used to deal with different detection tasks. In [13], a receptive field block based on the dilated convolution was proposed, which imitates the attention habit of the human visual system to enhance the representation ability of feature maps, achieving satisfactory object classification performance. In [14], an architecture with five dilated branches was designed manually to solve the receptive field issue, providing an input

including multi-scale information for subsequent feature fusion. In [15], a detector was employed, which configures the receptive field enhancement block after the multi-scale extraction module, bringing significant improvement to the performance of small target detection. In [16], an object classification backbone architecture was designed to improve the resolution change of the sampling process, thus solving the problem of multi-scale object recognition. In [17], a dense dilated architecture was applied to generate the initial target regions, and then a probability regularized walking strategy was used to reduce the influence of object region fracture in the initial extraction.

Generally speaking, most previous methods consider the problem of learning multi-scale features, and they usually manually design complex multi-scale extraction branches to mine the context information of input images. However, these models seldom consider adaptively setting the dilated parameter of dilated convolution based on target characteristics, which limits the contribution of the extracted features to subsequent object classification to a certain extent. Different from other methods, we propose an automatic dilated rate selection strategy in a dilated convolution kernel based on training data, which can obtain optimized effective receptive fields for ships of different scales. The experimental results show that it is conducive to improving the performance of ship detection in high-resolution remote sensing images.

2.2. Attention-Wise Design in Learning Network

When observing images, humans can select valuable regions from a large amount of irrelevant background interference through neural signal processing. The researchers of deep learning network have borrowed ideas from this efficient processing approach. In the deep networks presented in recent years, attention model, as a lightweight component to enhance feature representation, is widely used and plays a significant role in various computer vision tasks. Most of the current research methods tend to apply a weight mask to form the attention mechanism. Their principle is to generate a new weight distribution map by calculating the dependency relationship between channels or features, highlighting the significant features in the input feature map. Generally speaking, attention-based feature enhancement mainly includes two steps: first, extracting the distribution of input features, and then calculating the correlation between features according to the distribution information.

The attention mechanism for the image applications is mainly used to capture the perspective field on the image. In 1988, a representative work of applying attention mechanism to visual image was presented [18], which aggregates multi-scale features into a unified saliency map, and extracts potential areas according to the saliency score. Another milestone work that attracted extensive concern to the attention mechanism was the research on image classification in 2014 [19], which applies the recurrent model to learn the distribution areas of suspected target from image sequence. In 2017, a recurrent attention model was presented [20], which iteratively samples the whole image to generate multiple local feature regions and integrates the prediction of each region to gain the final classification result. In 2018, a significant design named squeeze-and-excitation networks (SE-Net) was proposed [21], which adaptively adjusts the feature response in channel dimension by using the pooling and nonlinear transformation strategies. It is worth noting that this work achieves significant performance improvement with low computing cost, leading a wave of research on channel domain processing. Considering that SE-Net makes less use of the correlation of spatial location information, and location details are actually conducive to object detection tasks, CBAM was proposed in 2018 [22], which introduces the correlation of spatial dimension to allocate the weight of feature maps. In addition to the channel and spatial dimensions, the follow-up research also expanded the attention mechanism to the time [23] and category dimensions [24]. On the whole, some of the above methods ignore the spatial location information, and some models that focus on global correlation are less specially designed combined with the characteristics of the detected target.

Different from these previous methods, we provide a new perspective to capture the global dependency in the feature map. Actually, starting from the narrow and long characteristics of typical ships, we extract the spatial correlation by calculating the dependency between each query point and points in the same column and row, which has been proved in subsequent ablation experiments to improve the accuracy of ship detection.

3. Proposed Method

3.1. Method Overview

The overall architecture of the proposed method is depicted in Figure 1. It mainly consists of two modules: dilated rate selection (DRS) module and attention-wise feature representation (AFR) module. By using the first module, we construct a search space to realize the mapping between the standard convolution and dilated convolution. On this basis, we gain the optimized dilated rate by constraining the loss function of training data. After feature extraction, the second module is utilized to enhance feature representation by mining the correlation between two orthogonal directions in spatial domain.

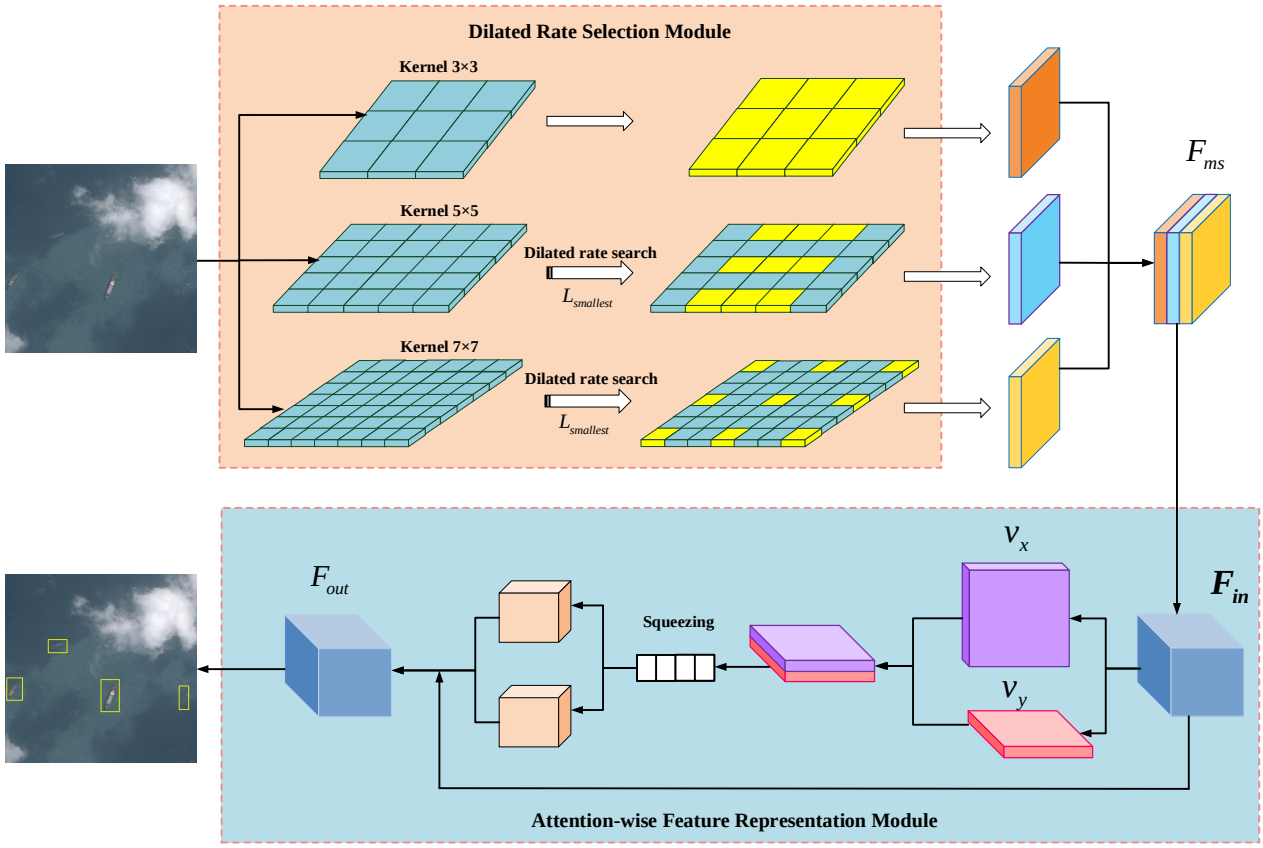

Figure 1. Flow chart of the proposed work.

3.2. Dilated Rate Selection for Multi-Scale Extraction

In order to expand the receptive field of the image, dilated convolution is often employed to replace the down-sampling operation, avoiding the defect of resolution reduction. By setting different dilated rate parameters, the network can have excellent feature extraction effect for objects with different scales.

Taking a one-dimensional signal as an example, the input signal x is acted by the filter W, and the output response y is thus generated as

$$y(i) = \sum_{m=1}^{M} x(i + r \cdot m)w(m) \tag{1}$$

where $x(i)$ is a point of the input signal. r indicates the sampling step of the input signal.

Extending the dimension to two dimensions, the input–output relationship of dilated convolution can be characterized as

$$y(i,j) = \sum_{m=1}^{M} \sum_{n=1}^{N} x(i + r_i \cdot m, j + r_j \cdot n) w(m,n) \quad (2)$$

where $x(i,j)$ is the coordinate of the input image. r_i and r_j represent the expansion rate of convolution in two directions, respectively. $w(m,n)$ means a convolution kernel with length m and width n. Obviously, for detecting ships of different sizes in practical applications, it is inaccurate and cumbersome to manually design the dilated rates of each convolution kernel. Therefore, inspired by [25], we seek a strategy of parameter search to automatically assign the optimized rates (as illustrated in Figure 2).

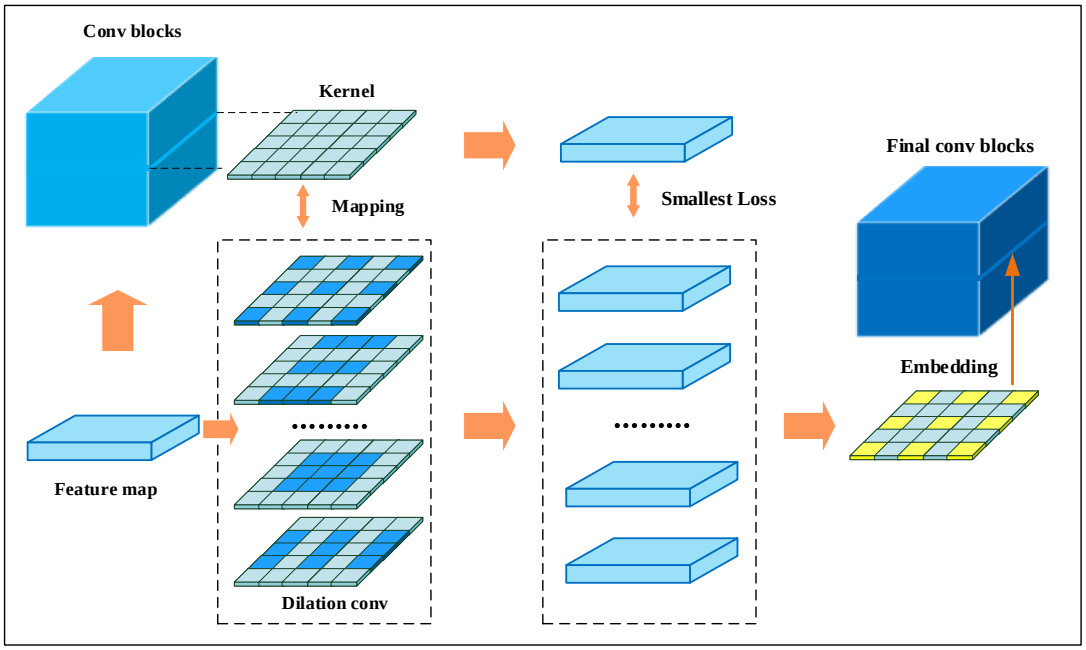

Figure 2. Design of the dilated rate search module for multi-scale extraction.

Specifically, we first fix a basic architecture of multi-scale feature extraction. Considering the residual network is a widely used high-quality structure, we search the parameters based on the ResNet backbone. In order to make the dilated convolution have a comparison benchmark and gain the initial weight assignment, we establish a kernel search space to support the mapping of weight parameters between dilated convolution and the pre-trained standard convolution.

Let us assume that the search space of the network structure is S [26]. Theoretically, S is a set containing all potential convolution kernel scale designs and corresponding weight allocation, but considering the limitation of actual computing resources, S can often be regarded as a sampled subset of the theoretical set S. Specifically, we apply the commonly used 3×3, 5×5 and 7×7 kernels as the elements of the search space. Similarly, convolution kernels of other scales can continue to be set according to computing resources.

Assuming that the ResNet structure has l layers, then on the l-th lay of the residual architecture, we can characterize S by the size of convolution kernel k and all candidate dilated rates r as

$$S = f(k,r) \quad (3)$$

Based on training data, taking the best detection accuracy as the evaluation index, we then train a set of relatively optimal convolution kernels and weight parameters by setting

different search elements, i.e., different convolution kernels, for each layer. This process can be characterized as

$$W_{sc} = \arg\min L(D_{\text{train}}, W) \tag{4}$$

where $L(\cdot)$ represents the mean squared error loss function. W indicates the weight parameter. W_{sc} represents the weight learned from the training set D_{train} by the standard convolution of the ResNet architecture.

Through the above training processes, we obtain the kernel parameters of the standard convolution architecture in each layer, which are a set of parameter combinations that deeply learn the characteristics of training image data. The results actually provide a useful reference for the subsequent dilated convolution search. For the convenience of the subsequent search, it is necessary to provide an initial weight for each convolution element. We next clarify the corresponding pixel-level relationship between the standard convolution and dilated convolution, and the maximum dilated rate of the i-th lay is calculated as

$$r_{l,\max} = \frac{k_{l,sc} + 1}{k_{l,dc}} \tag{5}$$

where $k_{l,sc}$ represents the standard convolution kernel scale of the l-th lay. $k_{l,dc}$ means the dilated convolution kernel scale. By setting the values of $k_{l,sc}$ and $k_{l,dc}$, we make $r_{l,\max}$ be an integer which is not less than 1, thus constructing the pixel-level mapping of two convolutions.

Obviously, from the pixel mapping we established, we can see that the pixel region included in the dilated convolution is a subset of the pixel region of the standard convolution. Inspired by [27], we directly copy the weights of the pre-trained standard convolution to the non-zero element position of the dilated convolution, which is equivalent to the dilated convolution retaining some feature information extracted by the original convolution.

Finally, for each layer, we input the same feature map and calculate the loss function between the output of standard convolution and all possible dilated convolutions. The dilated convolution which has the minimum loss is selected as the output convolution. The search process can be modeled as an optimization issue [28], which is expressed as

$$r_{\text{out}} = \arg\min_{S,W} L(S, D_{\text{train}}, W_{sc \to dc}) \tag{6}$$

where r_{out} means the output dilated rate parameter of the selected dilated convolution.

Specifically, we define that $W_{l,i}$ means the weight of the i-th convolution at the l-th lay in the standard convolution architecture, $W_{l,i}^{r_x,r_y}$ means the weight of the corresponding dilated convolution architecture, (r_x, r_y) represents the dilated rate of convolution operation in direction x and direction y, and X is the input of the l-th lay. Then, the loss function can be calculated as

$$L_1 = \left| W_{l,i} \otimes X - W_{l,i}^{r_x,r_y} \otimes X \right| = \left| \left(W_{l,i} - W_{l,i}^{r_x,r_y} \right) \otimes X \right| \tag{7}$$

Since X is independent of $W_{l,i}$ and $W_{l,i}^{r_x,r_y}$ [25], the objective function can be expressed as

$$\min_{r_x,r_y} \left| \left(W_{l,i} - W_{l,i}^{r_x,r_y} \right) \right|, \text{ s.t. } r_x, r_y \in \{0, 1, \ldots, r_{l,\max}\} \tag{8}$$

We traverse the combinations of r_x and r_y and select the one corresponding to the smallest loss, thus gaining the final dilated convolution of the search space.

3.3. Attention-Wise Feature Representation

Through the extraction module based on the multiple dilated convolution design, we finely characterize the multi-scale characteristics of the input scene by apply the pyramid pooling operation [29]. In addition, the receptive field expansion caused by dilated convolution retains more context information than ordinary convolution. In order to make

better use of the spatial position details between these extracted features, we employ the attention mechanism to further strengthen the representation of target features and reduce the response of feature map to background interferences [30].

As we know, the previous attention models mainly mine the correlation between spatial points from the global perspective in the spatial domain. Nevertheless, considering the characteristic that most typical ships have a long and narrow shape, we creatively introduce a direction sensitive attention model to enhance the feature representation of ship targets. The structure of the employed attention-wise module is shown in Figure 3. The main idea of this module is as follows: The distribution characteristics of two spatial-domain orthogonal directions in the feature map are extracted to obtain the direction and position-sensitive feature vectors. On this basis, we obtain the weight integrating the correlation of spatial dimension and channel dimensions through vector aggregation and channel squeezing operation.

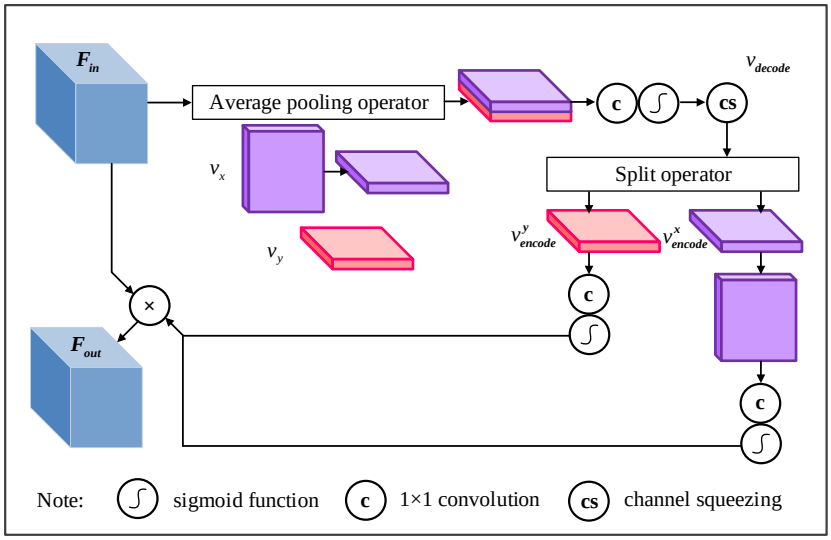

Figure 3. Design of the attention-wise representation module.

Specifically, for an input feature map $F_{in} \in \mathbb{R}^{c\times h\times w}$, c represents the channel number, and h and w represent the pixel lengths in two directions of the spatial dimension. We collect feature information along the two directions in the spatial dimension, and the feature vectors $v_{\mathbf{x}} \in \mathbb{R}^{c\times h\times 1}$ and $v_{\mathbf{y}} \in \mathbb{R}^{c\times 1\times w}$ can be expressed as

$$v_x(k,i) = \frac{1}{w}\sum_{i=1}^{w} F_{in}(k,i,j) \tag{9}$$

$$v_y(k,j) = \frac{1}{h}\sum_{j=1}^{h} F_{in}(k,i,j) \tag{10}$$

where $v_x(k,i)$ indicates the feature vector v along the horizontal direction in c-th channel. Similarly, $v_y(k,j)$ indicates the feature vector along the vertical direction.

Through the above processing steps, we integrate the features along the two orthogonal directions. Because the features aggregate the spatial information along the orthogonal direction, they can effectively capture the distribution relationship of the position and direction of the potential region. To further embed the spatial information into the channel dimension, the concatenating operation is applied to obtain the aggregated feature vector $v_{x,y} \in \mathbb{R}^{c\times 1\times(h+w)}$. Moreover, inspired by [21], we then encode the feature vector to compress the channel of the input vector by α times, where α is a positive integer representing the channel compression multiple. This squeezing step not only reduces the amount

of calculation parameters, but also integrates cross-channel information, facilitating the representation of channel correlation.

$$v_{\text{encode}} = sigmoid\left(\text{conv}_{1\times1}\left(v_{x,y}\right)\right) \tag{11}$$

where $\text{conv}_{1\times1}$ denotes the 1×1 convolutional transformation. *Sigmoid* means a non-linear activation function. The obtained $v_{\text{encode}} \in {}^{c/\alpha\times1\times(h+w)}$.

Actually, through above encoding process, the vector v_{encode} extracts cross-channel correlation and embeds spatial location information at the same time. As we know, the spatial position characteristic is beneficial to locate the interested region in the object detection task.

After obtaining the feature vector v_{encode} containing accurate location information through the spatial position encoding, in order to apply the decoded attention weight to the input feature map F_{in}, we split the vector v_{encode} along the vertical and horizontal directions, thus obtaining

$$v^{x}_{\text{encode}}, v^{y}_{\text{encode}} = split(v_{\text{encode}}) \tag{12}$$

where *split(·)* represents the dimension splitting operation. The vector $v^{\boldsymbol{x}}_{\text{encode}} \in \mathbb{R}^{c/\alpha\times1\times h}$, $v^{\boldsymbol{y}}_{\text{encode}} \in \mathbb{R}^{c/\alpha\times1\times w}$.

For the split vectors, 1×1 convolutional transformation is used to restore the influence of pervious channel reduction in the encoding stage, yielding the vectors with the same channel number as the input feature map. The decoded attention weights along different spatial directions can be given as

$$v^{\mathrm{x}}_{\text{decode}} = sigmoid(\text{conv}_{1\times1}(v^{\mathbf{x}}_{\text{encode}})) \tag{13}$$

$$v^{\mathrm{y}}_{\text{decode}} = sigmoid\left(\text{conv}_{1\times1}\left(v^{\mathbf{y}}_{\text{encode}}\right)\right) \tag{14}$$

where $v^{\mathbf{x}}_{decode}$ and $v^{\mathbf{y}}_{decode}$ are attention weights embedding the vertical and horizontal spatial directions, respectively. By applying the decoded attention weights, the final output feature map $F_{out} \in \mathbb{R}^{c\times h\times w}$ can be given as

$$F_{\text{out}}(k,i,j) = v^{x}_{\text{decode}}(k,i) \times v^{y}_{\text{decode}}(k,j) \times F_{\text{in}}(k,i,j) \tag{15}$$

Then, similar to the Faster R-CNN method [31], we input the obtained features into the fully connected layers and achieve the final classification.

4. Experimental Analysis

4.1. Experimental Setup

In order to illustrate the application effect of our method, the comparative experiments are constructed on the widely used Airbus dataset. This dataset is a high-quality image dataset recognized by the ship detection research community. The image scene covers various sea and air elements, such as waves, clouds and reefs, and includes a variety of practical situations, both single-target and multi-target. The image size is 768×768 pixels. The whole dataset includes more than two hundred thousand images, of which more than forty thousand images contain ships (a total of 81,689 ship instances). We randomly selected eight thousand images from the dataset, including six thousand as the training set and the rest as the test set.

To measure the accuracy of the proposed model, the commonly used average-precision (AP) is selected as the evaluation index, which is a comprehensive characterization of the precision and recall of the model. Actually, the AP describes the area under the precision-recall curve as

$$AP = \int_0^1 precision(recall)d(recall) \tag{16}$$

Specifically, AP@α means the AP value at the intersection over union (IOU) threshold of α. The IOU is a parameter to calculate the overlap of the ground truth box and the prediction box of the detection algorithm, which is defined as

$$IOU = \frac{B_{gt} \cap B_{pd}}{B_{gt} \cup B_{pd}} \tag{17}$$

where B_{gt} and B_{pd} represent the ground truth box and the predicted box, respectively.

In addition, the false alarm rate (FAR) of the algorithm directly determines the effect of practical applications. Based on this consideration, we also take the FAR as an evaluation index of different detection methods, which is defined as

$$FAR = \frac{\text{number of detected false alarms}}{\text{number of detected candidates}} \tag{18}$$

4.2. Implementation Details

Our detector is end-to-end trained on a workstation with an Nvidia RTX 2080 GPU. The proposed model is implemented with the PyTorch framework. Before the network training stage, we applied the anchor box parameter optimization strategy to automatically generate a set of anchors that are more suitable for the employed dataset, thus improving the efficiency and accuracy of ship detection. The stochastic gradient descent strategy is used to optimize the objective function in the training process. The learning rate begins from 0.001, and the weight decay and momentum are set to 0.0005 and 0.9, respectively. We apply 250 epochs to learn the characteristics of our training data, and the train batch size is 4. The variation curves of loss function during training are shown in Figure 4.

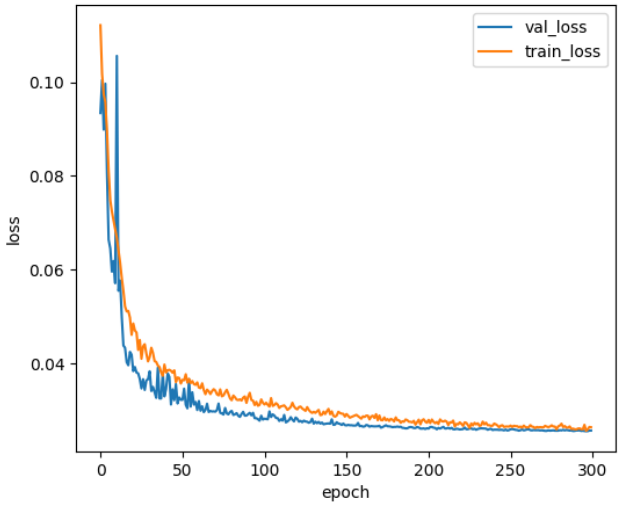

Figure 4. Loss curve of the image training process.

In addition, it should be noted that detection algorithms may have different requirements for the input image size. For example, the default input size of the Faster R-CNN model is 800 × 600 pixels, while the YOLOv4's requirement is 608 × 608 pixels [32]. In order to make our experimental images (they are 768 × 768 pixels) meet the input requirements of various compared models, we pre-processed the images of the whole dataset before inputting different models. Specifically, we performed a padding operation for the side length shorter than the input requirements of the models and random clipping operation for the side length longer than the input sizes.

4.3. Ablation Analysis

To assess the effectiveness and necessity of our employed modules, we perform extensive ablation experiments. Since our proposed multi-scale extraction module can

be regarded as a modified version of the ResNet architecture, the typical ResNet50 and ResNet101 architectures are applied as the baseline backbone.

In order to illustrate the details of the dilated rate selection, we calculate the dilated convolution proportions of different shapes in ResNet50 and ResNet101 architectures after selection operation. According to the height h and width w of the convolution kernels, the shapes of dilated convolution are divided into square ($h = w$), vertical ($h > w$) and horizontal ($h < w$). For the fairness and accuracy of the statistical results, we conduct ten experiments to obtain the average values. As can be seen from Table 1, the selected convolution shapes are mainly square. This is because the distribution directions of ships are random in the remote sensing image, the square convolution is more suitable for feature extraction.

Table 1. Statistical results of dilated convolution shapes.

Shapes	ResNet50 + Proposed DRS (%)	ResNet101 + Proposed DRS (%)
Square	42.00	46.53
Vertical	28.00	24.75
Horizontal	30.00	28.72

Table 2 reports the ablation experimental results on the Airbus dataset. When the Faster R-CNN detection framework is limited, the AP values obtained by configuring the original ResNet50 and ResNet101 as the multi-scale extraction module are 68.52 and 69.38, respectively. From the AP measure, we can find that when we embed the multi-scale dilated convolution module into the Faster R-CNN detection framework, the AP value reaches 70.72. After further strengthening the characterization with the attention-based feature enhancement module, the final AP value reaches 72.68, showing a significant performance improvement compared with the original detection framework.

In order to intuitively illustrate the role of each module, we select a typical example and add each module in turn. The detection results of the example are shown in Figure 5. The result of only using the detection framework configured with the ResNet101 is shown in Figure 5b. It can be found that for the three ships in the image, this method misses the smallest ship and mistakenly detects two objects that are actually clouds. This may be because the detection model has some limitations in the feature representation of the small target through multiple down-sampling operations. After embedded the proposed DRS module, the algorithm successfully detects the smallest target and removes a false alarm (as shown in Figure 5c). When we further configure the attention module to the Faster R-CNN framework, we can observe that we detect all the targets and effectively constrain false alarms (as shown in Figure 5d). Therefore, the above ablation experiments illustrate the validity and effectiveness of the proposed module.

Table 2. Evaluation results of ablation experiments.

Methods	Multi-Scale Extraction Module	Attention Module	AP@0.5 (%)	AP@0.75 (%)	AP (%)
Faster R-CNN	ResNet50	-	83.50	72.79	68.52
Faster R-CNN	ResNet101	-	84.44	73.87	69.38
Faster R-CNN	ResNet50 + proposed DRS	-	84.71	74.32	69.94
Faster R-CNN	ResNet101 + proposed DRS	-	86.08	76.98	70.72
Faster R-CNN	ResNet50	Proposed AFR	85.37	75.72	69.71
Faster R-CNN	ResNet101	Proposed AFR	86.14	77.85	71.19
Faster R-CNN	ResNet50 + proposed DRS	Proposed AFR	87.23	76.43	71.56
Faster R-CNN	ResNet101 + proposed DRS	Proposed AFR	87.65	79.81	72.68

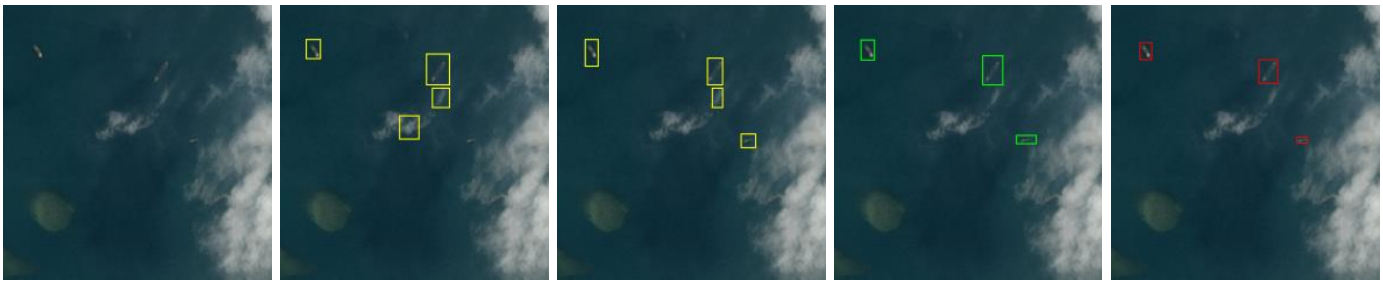

Figure 5. Detection examples of the ablation experiment: (**a**) original image, (**b**) Faster R-CNN with ResNet101, (**c**) Faster R-CNN with the proposed DRS module, (**d**) Faster R-CNN with the two proposed modules and (**e**) ground truth.

4.4. Comparative Experiments

To further verify the overall performance of our algorithm, several advanced algorithms are adopted as baseline approaches for comparative experiments, including ship detection via visual saliency (SDVS) [33], YOLOv4 [32] and HSF-Net [34] methods. The SDVS method constructs a visual saliency model based on Fourier transform to locate the potential region of objects and utilizes the geometric prior features to achieve the final classification. The YOLOv4 model is a famous single-stage object detection method, which applies logistic regression to predict the bounding box of objects in each grid cell, thus determining the location and scale of ships. Moreover, this approach omits the step of generating a large number of proposals for the input image, thus having high computational efficiency. The HSF-Net approach adopts a hierarchical selective filtering architecture to learn features of different levels, thus providing the input for the accurate prediction of potential target regions. Then, a classifier is added to the detection head, producing the final detection results for targets with different scales.

In order to intuitively illustrate the performance effect of each method, we show the prediction results of four representative examples in Figure 6. The scenes of these examples not only include ship targets of different scales but also contain the interference of thin cloud, thick cloud, sea clutter and mist. The scene in the first column of Figure 6 includes a ship and some thin clouds. The SDVS and YOLOv4 methods mistakenly classify the clouds as targets and miss the real small-scale ship. In contrast, both of the HSF-Net method and our approach can effectively detect the ship which has low contrast with the neighborhood background. The scene in the second column of Figure 6 includes the interference of thick clouds and reefs. In addition, it is worth noting that one of the ships is in the shadow of clouds. One can see that the SDVS and the HSF-Net approaches correctly detect a ship, whereas the YOLOv4 model and our approach can successfully detect two ships, including the ship in the shadow. Unfortunately, there is also a false alarm in the result of our method. For the image scene with the sea clutter interference (as seen in the third column of Figure 6), although the three compared methods correctly detect the ship, they also generate additional false alarms. In contrast, the proposed method can detect the target efficiently. For the mist scene given in the fourth column of Figure 6, the SDVS and YOLOv4 approaches miss a small-scale ship located in the upper left corner of the image, whereas the HSF-Net method and our approach obtain satisfactory results.

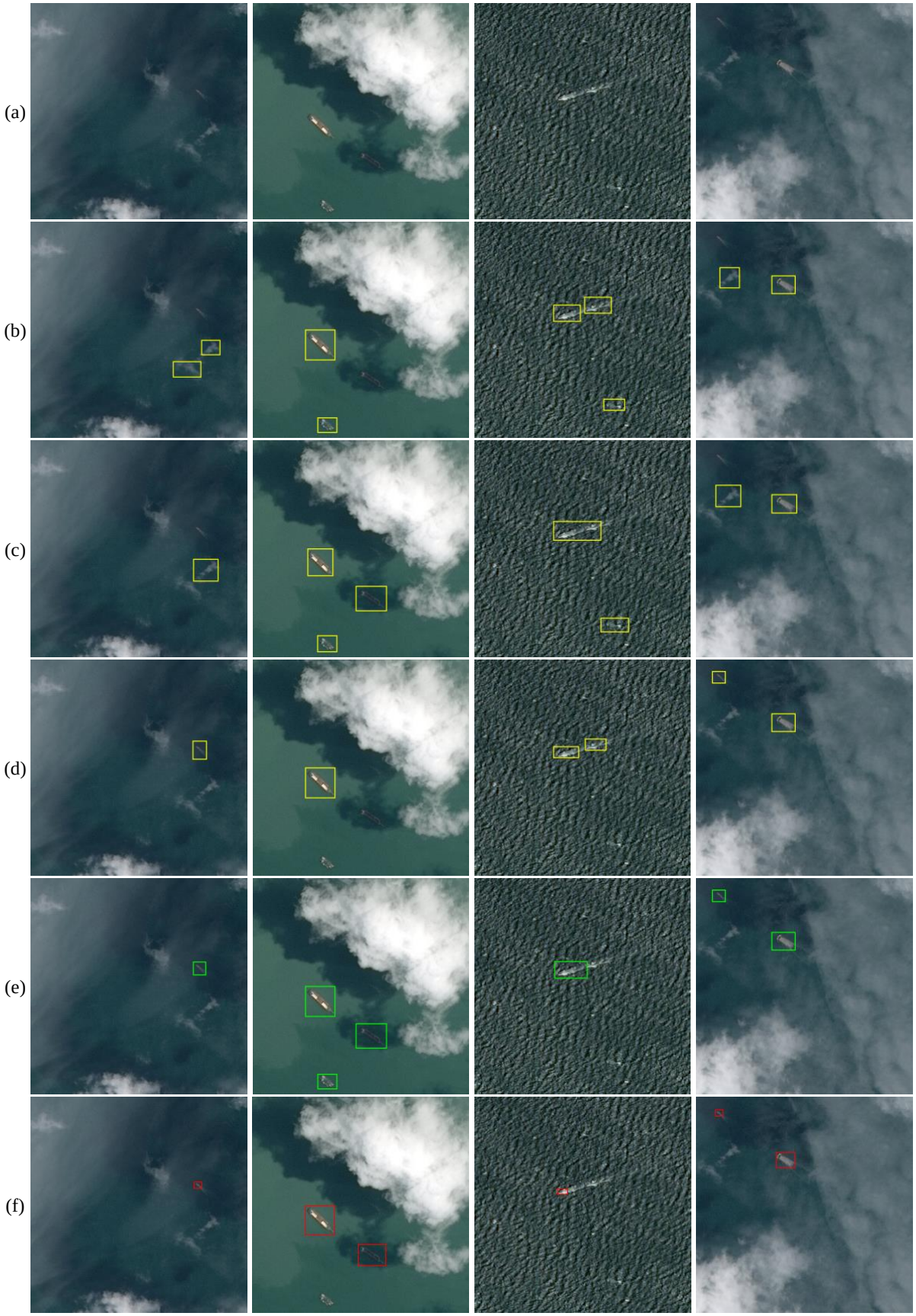

Figure 6. Detection result examples provided by different detection methods on the Airbus dataset: (**a**) original image, (**b**) SDVS, (**c**) YOLOv4, (**d**) HSF-Net, (**e**) proposed method and (**f**) ground truth.

Table 3 provides the quantitative evaluation results of different detection methods on the Airbus dataset. We can observe that the proposed method outperforms other compared approaches in terms of the AP and FAR indexes. In contrast, the SDVS method gains the lowest AP and the highest FAR. Actually, this SDVS method, which directly applies a visual saliency model to extract candidate regions, can easily produce false alarms in complex scenes. In addition, it only applies the geometric and texture features for the final target classification, which has disadvantages compared with deep networks in terms of the feature utilization. Besides, as we know that the YOLOv4 model has the advantage of fast detection, nevertheless, it lacks the direct extraction process of candidate regions, easily leading to incorrect output positions, especially for small ships. Among these considered approaches, the HSF-Net obtains the second highest AP value. Actually, both the technique and our method focus on multi-dimensional representation of the extracted features and enhancing information that contributes to the final classification. The experimental results prove the correctness of this strategy to a certain extent. Consequently, based on the above qualitative and quantitative evaluations, we can conclude that the proposed method can effectively detect ships with different scales under complex environmental interference conditions.

Table 3. Comparative experimental results on the Airbus dataset.

Methods	Backbone	Input Image Size	AP(%)	FAR(%)
SDVS	-	768 × 768	62.67	12.39
Faster R-CNN	ResNet50	1000 × 600	68.52	8.12
Faster R-CNN	ResNet101	1000 × 600	69.38	7.84
YOLOv4	CSPDarknet53	608 × 608	69.64	6.69
HSF-Net	VGG-16	500 × 500	71.37	5.84
Proposed	ResNet101 + proposed DRS	1000 × 600	**72.68**	**5.65**

5. Conclusions

In this paper, we present a ship detection architecture based on the dilated rate selection and attention-guided feature representation strategies, which can detect ships of different scales under complex scene conditions. Specifically, a dilated convolution selection strategy is applied to a multi-branch extraction network, adaptively extracting context information of different receptive fields without reducing the image resolution. Moreover, to enhance the feature characterization, we calculate the correlation of spatial points from the vertical and horizontal directions and embed it into the channel compression coding process, realizing the accurate capture of multi-dimensional characteristics. Experimental results on the Airbus dataset demonstrate the effectiveness of the proposed method in detecting ship targets under complex environmental interference.

Author Contributions: Investigation, literature analysis, methodology, writing—original draft, validation, J.H.; funding acquisition, project administration, X.Z.; supervision, W.Z.; revising and editing, T.S. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (61975043).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The research dataset was released in July 2018, which is available at https://www.kaggle.com/c/airbus-ship-detection/data, accessed on 23 September 2021.

Acknowledgments: The authors would like to thank the Kaggle competition organizers for providing the Airbus dataset.

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

References

1. Kanjir, U.; Greidanus, H.; Oštir, K. Vessel detection and classification from spaceborne optical images: A literature survey. *Remote Sens. Environ.* **2018**, *207*, 1–26. [CrossRef] [PubMed]
2. Li, L.; Zhou, F.; Zheng, Y.; Bai, X. Saliency detection based on foreground appearance and background-prior. *Neurocomputing* **2018**, *301*, 46–61. [CrossRef]
3. Hu, J.; Zhi, X.; Zhang, W.; Ren, L.; Bruzzone, L. Salient Ship Detection via Background Prior and Foreground Constraint in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3370. [CrossRef]
4. 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]
5. Hu, J.; Zhi, X.; Shi, T.; Zhang, W.; Cui, Y.; Zhao, S. PAG-YOLO: A Portable Attention-Guided YOLO Network for Small Ship Detection. *Remote Sens.* **2021**, *13*, 3059. [CrossRef]
6. 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]
7. Jiang, W.; Liu, M.; Peng, Y.; Wu, L.; Wang, Y. HDCB-Net: A Neural Network with the Hybrid Dilated Convolution for Pixel-Level Crack Detection on Concrete Bridges. *IEEE Trans. Ind. Inform.* **2020**, *17*, 5485–5494. [CrossRef]
8. Oliva, A.; Torralba, A. The role of context in object recognition. *Trends Cogn. Sci.* **2007**, *11*, 520–527. [CrossRef]
9. Jeon, M.; Jeong, Y.S. Compact and accurate scene text detector. *Appl. Sci.* **2020**, *10*, 2096. [CrossRef]
10. Vu, T.; Van Nguyen, C.; Pham, T.X.; Luu, T.M.; Yoo, C.D. Fast and efficient image quality enhancement via desubpixel convolutional neural networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 243–259.
11. Ji, Y.; Zhang, H.; Zhang, Z.; Liu, M. CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Inf. Sci.* **2021**, *546*, 835–857. [CrossRef]
12. Zhang, S.; Wu, R.; Xu, K.; Wang, J.; Sun, W. R-CNN-based ship detection from high resolution remote sensing imagery. *Remote Sens.* **2019**, *11*, 631. [CrossRef]
13. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
14. Xu, J.; Wang, W.; Wang, H.; Guo, J. Multi-model ensemble with rich spatial information for object detection. *Pattern Recognit.* **2020**, *99*, 107098. [CrossRef]
15. Qu, J.; Su, C.; Zhang, Z.; Razi, A. Dilated convolution and feature fusion SSD network for small object detection in remote sensing images. *IEEE Access* **2020**, *8*, 82832–82843. [CrossRef]
16. 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.
17. Mou, L.; Chen, L.; Cheng, J.; Gu, Z.; Zhao, Y.; Liu, J. Dense dilated network with probability regularized walk for vessel detection. *IEEE Trans. Med. Imaging* **2019**, *39*, 1392–1403. [CrossRef]
18. 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]
19. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2204–2212.
20. Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. IAUnet: Global context-aware feature learning for person reidentification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4460–4474. [CrossRef]
24. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 173–190.
25. Liu, J.; Li, C.; Liang, F.; Lin, C.; Sun, M.; Yan, J.; Ouyang, W.; Xu, D. Inception convolution with efficient dilation search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 11486–11495.
26. Guo, Z.; Zhang, X.; Mu, H.; Heng, W.; Liu, Z.; Wei, Y.; Sun, J. Single path one-shot neural architecture search with uniform sampling. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 544–560.
27. Fang, J.; Sun, Y.; Peng, K.; Zhang, Q.; Li, Y.; Liu, W.; Wang, X. Fast neural network adaptation via parameter remapping and architecture search. *arXiv* **2020**, arXiv:2001.02525.
28. Liu, Y.; Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Tan, K.C. A survey on evolutionary neural architecture search. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–21. [CrossRef]

29. 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.
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
32. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
33. Nie, T.; Han, X.; He, B.; Li, X.; Liu, H.; 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]
34. 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]