remote sensing

MDPI

Article

Multiscale Object Detection in Remote Sensing Images Combined with Multi-Receptive-Field Features and Relation-Connected Attention

Jiahang Liu *, Donghao Yang and Fei Hu

College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; yangdonghao@nuaa.edu.cn (D.Y.); fei_hu@nuaa.edu.cn (F.H.)
* Correspondence: jhliu@nuaa.edu.cn

Abstract: Object detection is an important task of remote sensing applications. In recent years, with the development of deep convolutional neural networks, object detection in remote sensing images has made great improvements. However, the large variation of object scales and complex scenarios will seriously affect the performance of the detectors. To solve these problems, a novel object detection algorithm based on multi-receptive-field features and relation-connected attention is proposed for remote sensing images to achieve more accurate detection results. Specifically, we propose a multi-receptive-field feature extraction module with dilated convolution to aggregate the context information of different receptive fields. This achieves a strong capability of feature representation, which can effectively adapt to the scale changes of objects, either due to various object scales or different resolutions. Then, a relation-connected attention module based on relation modeling is constructed to automatically select and refine the features, which combines global and local attention to make the features more discriminative and can effectively improve the robustness of the detector. We designed these two modules as plug-and-play blocks and integrated them into the framework of Faster R-CNN to verify our method. The experimental results on NWPU VHR-10 and HRSC2016 datasets demonstrate that these two modules can effectively improve the performance of basic deep CNNs, and the proposed method can achieve better results of multiscale object detection in complex backgrounds.

Keywords: convolutional neural networks (CNNs); multi-receptive-field feature extraction; multi-scale object detection; relation-connected attention; remote sensing images

Citation: Liu, J.; Yang, D.; Hu, F. Multiscale Object Detection in Remote Sensing Images Combined with Multi-Receptive-Field Features and Relation-Connected Attention. *Remote Sens.* **2022**, *14*, 427. https://doi.org/10.3390/rs14020427

Academic Editor: Józef Lisowski

Received: 30 November 2021
Accepted: 14 January 2022
Published: 17 January 2022

1. Introduction

With the development of remote sensing technology, object detection in remote sensing images has become a popular topic. Satellite remote sensing is not restricted by airspace and can continuously observe the Earth's surface dynamically, which has become the primary technique of the dynamic detection, tracking and recognition of time-sensitive targets. Remote sensing technology can quickly obtain the location, attributes, distribution and movement characteristics of objects, thus providing support for relevant decision-making. Object detection and identification are widely used in applications such as dynamic port surveillance, traffic monitoring, territorial defense and naval warfare. SAR can work all day long under various weather conditions and is suitable for object detection under various atmosphere and environment. In recent years, satellite remote sensing technology has significantly advanced, and high-resolution optical remote sensing images can provide more detailed and richer information than SAR images [1], which has attracted great attention in the field of object detection. In comparison to SAR images, optical remote sensing images can provide clear details in terms of geometric shape, structure, color and texture—intuitive information which is easier for human understanding and interpretation [2]. Furthermore, the utilization of a large number of optical satellites and UAVs allows for the acquisition

of high-quality and high-frequency optical remote sensing images. Optical remote sensing images play a vital part in object detection, which is a valuable supplement to object detection in SAR images.

Object detection in remote sensing images is very different from those of natural scenery. Remote sensing images are more complex and are shot from a great distance. The scales of the objects with the same class vary in a large range and the same object changes greatly in scale under different resolutions. The scale of the objects changes in a large range, which will reduce the performance of the detection algorithm, especially for very small or very large objects. Although deep learning methods have shown good performance on natural images, they do not perform well on optical remote sensing images, especially for small targets as their accessible characteristics are restricted. Great efforts have been made in the field of object detection in remote sensing images, but the detection performance degradation caused by large variation in object sizes and similarity among objects with similar scales is still a challenging problem for object detection. Figure 1 shows challenges of object detection in remote sensing images. On the one hand, the scale changes of objects in remote sensing images have a large range, as shown in Figure 1a, which often make the performance degradation under complex backgrounds. The receptive field of the common detector is difficult to effectively cover the various object sizes, and small targets are often submerged in the interference of large targets and background. On the other hand, remote sensing targets with the same scale often have similar appearance and characteristics, such as the tennis court vs. basketball court in Figure 1b. The common detectors will be confused and it is difficult to effectively distinguish them.

(a) (b)

Figure 1. Challenges of object detection in remote sensing images: (**a**) large variations in object scales; and (**b**) similarity between objects of similar scales.

In this paper, we propose two useful modules to help improve the performance of object detection tasks in remote sensing images. These are integrated into the Faster R-CNN framework to deal with the large variation in object sizes and feature similarity between objects of similar scales in remote sensing images. The experimental results demonstrate that MR-FEM and RC-AM are effective plug-and-play blocks, which are easily inserted into basic deep CNNs to improve their performance. The main contributions of this paper are as follows:

(1) A new Multi-Receptive-Field Feature Extraction Module (MR-FEM) was constructed and integrated into the feature pyramid network (FPN), which enables the network to extract multiscale object features that aggregate multi-receptive-field information, providing a powerful feature representation for multiscale objects.

(2) A Relation-Connected Attention Module (RC-AM) is proposed to automatically select and refine the features that may be similar among objects of the same scale, which can reduce the interference of redundant features. This module obtains global information by stacking the feature itself and the relation-feature between features, and then mines the

global attention from them. This can effectively enhance the foreground information while weakening the background information, making the features more distinguishable.

The rest of this paper is organized as follows: Section 2 reviews related works and describes the challenges in multiscale object detection. In Section 3, the framework of our object detection model is introduced in detail. We present the dataset used for the experiments, the details of the experimental setup, the evaluation metrics, the results of the experiments and the experimental analysis in Section 4. Finally, the conclusions are provided in Section 5.

2. Related Work

2.1. Deep-Network-Based Object Detection

In recent years, convolutional neural networks (CNNs) have confirmed their performance in object detection, and deep learning object detectors for object detection are now the most common technical approach. Existing object detection methods based on deep learning can be divided into two types based on whether region proposals are generated. R-CNN [3], Fast R-CNN [4] and Faster R-CNN [5] are common two-stage detection methods, while SSD [6], YOLO series [7–9] and RetinaNet [10] are typically one-stage detection methods. In general, two-stage detection methods offer high accuracy, while one-stage detection methods have a great advantage in terms of speed. Deep networks can automatically learn basic features from enormous amounts of image data with more accuracy and robustness than traditional methods. In the two-stage method, the input image is used in the first stage to create category-independent region proposals, and features of these regions are extracted, after which the objects are classified and regressed using category classifiers and regressors in the second stage. Finally, discriminative techniques such as non-maximal suppression (NMS) are used to remove redundant bounding boxes from the final detection results. The pioneering work of the Region-based CNN (R-CNN) [3] and its upgraded version SPP-Net [11], Fast R-CNN [4], make it possible to simplify learning and increase operation efficiency. By sharing convolutional weights, Faster R-CNN [5] combines the Region Proposal Network (RPN) and Fast R-CNN into a single network, allowing object detection to be speedy and precise end to end. FPN [12], R-FCN [13], Mask R-CNN [14] and other high-performance detection algorithms have also been presented to date. One-stage strategies treat object detection as a regression problem without going through the proposal generation process, resulting in almost real-time performance. YOLO [9] and SSD [6] are two popular one-stage methods that guarantee accuracy while maintaining real-time performance. As a solution to overcome the class imbalance problem of one-stage approaches, RetinaNet [10] introduces a new focal loss function that applies a modulation term to the cross-entropy loss in order for the network to focus on tough negative samples. RefineDet [15], inspired by the two-stage method, enhances accuracy by using cascade regression and the anchor refinement module (ARM) to adjust anchor sizes and locations before filtering out redundant negative anchors.

2.2. Multiscale Object Detection in Remote Sensing Images

In order to address the problems of multiscale object detection in remote sensing images, many object detection algorithms based on deep learning have been proposed. Multiscale image pyramids is an effective method to deal with large-scale variations. In order to implement multiscale training more efficiently, a scale normalization method is proposed in SNIP [16,17] for training objects that match the specific scale range for each scale during multiscale training. However, this method is time-consuming and requires more computational resources. Another way to solve the multiscale problem is to perform object detection on multiple layers for objects of different scales. SSD [6] directly uses multiscale feature maps from different layers to detect targets at different scales to alleviate scale variation. FMSSD [18] improves the performance of detection by integrating contextual information into the feature map using the dilated space pyramid module. In recent years, several methods have demonstrated that combining deep and shallow feature maps at different scales can help to solve the scale problem of object detection. A classical

means is that of FPN [12]. To enhance the loss of semantics in lower feature maps, FPN adds a top–down path and lateral connections to propagate semantic information between deep and shallow layers. PANet [19] improves the feature fusion structure in FPN with an additional bottom–up path and proposes adaptive feature pooling to fuse features from different layers. To improve the multiscale detection performance in [20], a feature fusion module is introduced in Faster R-CNN, which jointly uses the semantic information obtained at the high level and the detailed information captured at the low level to generate a refined feature map, effectively improving the performance of multiscale object detection. Yang et al. [21] proposed SCRDet, a multi-classification detection method based on the Faster R-CNN framework, that combines finer feature fusion, multi-dimensional attention networks and constant factors in the loss function to decrease the sensitivity of the rotation angle. TridentNet [22] constructs a parallel multi-branch architecture with a scale-aware training scheme to specifically train each branch by using appropriately scaled objects.

2.3. Attention Mechanism

The idea of the attention mechanism is derived from human visual attention. The attention mechanism is to accurately extract the most valuable information from a large amount of input data and ignore the irrelevant information. The key information is in the dominant position for subsequent information processing, analysis and decision-making, while other unfocused information is redundant, which will be ignored and suppressed. An attention mechanism was widely used in computer vision tasks such as object detection and semantic segmentation, which achieved remarkable results. There are two main types of attention mechanism commonly used in computer vision: channel attention and spatial attention.

SENet [23] is a classic work based on the channel attention mechanism. Its innovation is that it pays attention to the connections between channels. It uses the channel attention mechanism to learn the weights of each channel in the feature layer, and automatically obtains the value of each channel. Feature recalibration is performed to enhance the high-contributing features and suppress the low-contributing features according to their importance. After SENet [23], many effective methods based on the attention mechanism were proposed. In CBAM [24], the author combined the spatial attention and the channel attention to guide the training, so that the network can not only learn what to focus on, but also learn where to focus. ECANet [25] conducts the information interaction between adjacent channels instead of all channel interactions in SE, which significantly reduces the complexity of the model. Wang et al. [26] proposed the non-local (NL) module to generate the attention map by obtaining information from all positions in the feature map, which can capture the long-distance dependence.

3. Proposed Method

3.1. Overview

In this section, we present details of the proposed method. Since our proposed multi-receptive-field feature extraction module (MR-FEM) and relation-connected attention module (RC-AM) are functional modules that are used to improve the multiscale object detection capability of the detector, they cannot complete the detection task independently. For the convenience of description and experimental validation, we integrate them into the Faster R-CNN framework with the ResNet as the backbone network. Specifically, ResNet is utilized as the backbone for the extraction of basic features of images, and the basic feature maps C2, C3, C4 and C5 are the output for each stage of the backbone, corresponding to 4, 8, 16 and 32$\times$ downsampling, and the number of channels are 256, 512, 1024 and 2048, respectively. Figure 2 depicts the overall structure of the network. After obtaining the fundamental feature maps of each stage, the multi-receptive-field feature extraction module (MR-FEM) is applied to further extract the contextual feature information with multi-receptive fields, and the output feature maps from MR-FEM are unified into 256 channels using 1 $\times$ 1 convolution. The features at each level are then fused using FPN. Specifically, the higher-level feature map output is summed with the lower-level

feature map, which is 2× up-sampled first, to obtain the outputs F2, F3, F4 and F5 at each level. Then, F2, F3, F4 and F5 are applied as inputs to the relation-connected attention module (RC-AM) to generate the final feature maps P2, P3, P4 and P5 for detection. Global information is generated by RC-AM using global relation modeling, which can further enhance the features and lead the network to better focus on the foreground and suppress the background. Finally, the refined feature maps are fed into the detection head to obtain the class scores and regression bounding box.

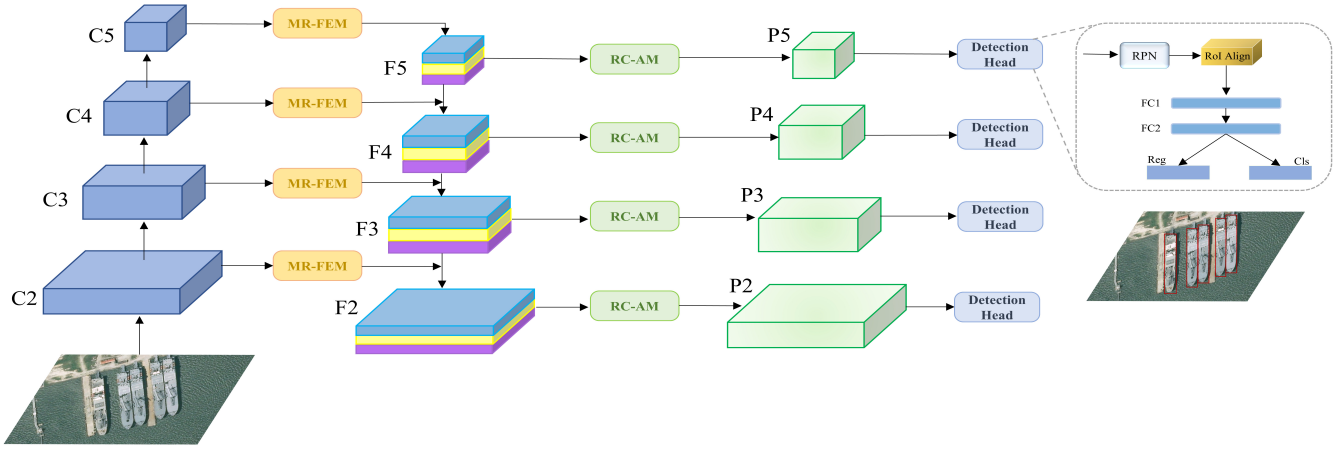


Figure 2. The overall framework of the detection model.

3.2. Multi-Receptive-Field Feature Extraction Module

With the development of deep learning techniques in recent years, many convolutional neural networks have overcome the problems of gradient descent and explosion caused by increased network depth and demonstrated robust feature extraction capabilities. However, features in deep layers are rich in semantic information, but spatial location information is severely compromised. In object detection, location information is crucial. Features in shallow layers, on the other hand, are weak in semantic information but sensitive to location information. Therefore, combining deep and shallow multiscale features with FPN may effectively fuse semantic and location information, achieving improved network performance. The scales of objects in optical remote sensing images change greatly, and receptive fields that are too large or too small cannot meet the needs of different receptive fields for the scale variation of objects, therefore feature extraction with multiscale receptive fields is necessary. The classic FPN directly employs the features extracted from the multiple levels of the backbone network for feature fusion, which can enhance the features to some extent by fusing information of different scales. However, for multiscale objects, this type of simple fusing strategy cannot adapt to the multiscale changes of the objects. As a result, this paper proposes a multi-receptive-field feature extraction module (MR-FEM), which aggregates multi-receptive-field object characteristics to produce a more informative feature representation via dilated convolution. As shown in Figure 2, the MR-FEM is embedded into FPN to adapt to multiscale objects.

Dilated convolution, also known as atrous convolution [27], was first used in semantic segmentation tasks to merge large-scale contextual information [28,29]. It expands the size of the convolution kernel with the original weights by performing convolution operations at sparsely sampled locations, thus increasing the size of the receptive field without additional parameter costs. Dilated convolution has also been widely used in the field of object detection. The proposed multi-receptive-field feature extraction module is shown in Figure 3.

To cover different sizes of receptive fields, the module adopts dilated convolution with the same kernel size but different dilation rates to extract multiscale context features. First, the features extracted by ResNet at each level are used as input, and atrous convolutions with the same size but different dilation rates of 3, 5 and 7 were employed to ensure that the

final feature maps contain scale and shape invariance characteristics. By inserting $(d-1)$ zeros between the values of the normal convolution kernel, dilated convolution extends the kernel size without increasing the parameters or computing cost. A 3×3 convolution with a dilation rate of d, for example, can have the same receptive field as a normal convolution with the kernel size of $3 + 2 \times (d-1)$. Finally, cross-channel concatenation is used to integrate features from convolution layers with different dilation rates and we then apply a 1×1 convolution to reduce the dimension. At this point, we finished the extraction of multi-receptive-field features using the proposed MR-FEM, and then the obtained features are fused via a side connection in FPN.

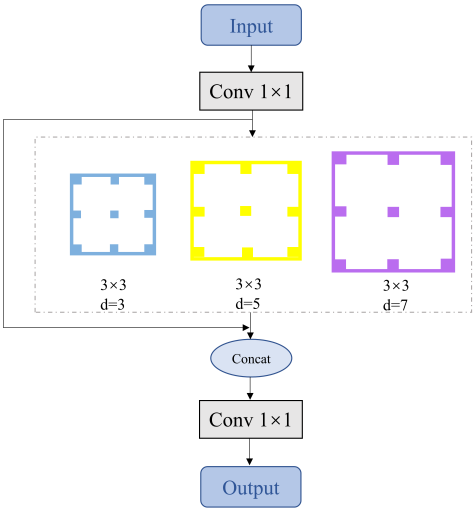


Figure 3. Structure of the multi-receptive-field feature extraction module. It contains three 3×3 convolutional layers with different dilation rates d.

3.3. Relation-Connected Attention Module

Attention is an important tool for visual perception that allows you to focus on the most relevant element of the input signal, which is critical in object detection. SENet [23] employs a channel attention method to learn the weights of each channel in the feature map and calculates the relevance of each channel automatically. The non-local Net [26] provides a global attention method that can capture long-range relational dependencies and collect information from all locations by learning pairwise relations in a deterministic manner to improve the object's location characteristics. Inspired by SENet and the non-local Net, we designed a relation-connected attention module (RC-AM) which can mine more valuable information from the features themselves and the relations between features, and acquire attention in a learning manner. RC-AM can effectively refine the features and distinguish similar features between objects of the same scale. As depicted in Figure 4, this module is mainly composed of two parts: Relation-Connected Spatial Attention and Relation-Connected Channel Attention. These two sub-modules are combined through residual connection.

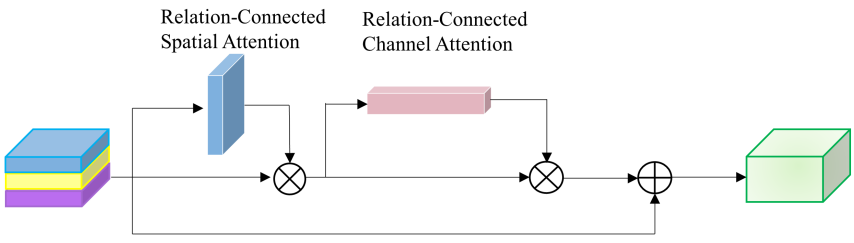


Figure 4. Architecture of RC-AM residual connection: $\otimes$ denotes element-wise multiplication; and $\oplus$ denotes element-wise summation

3.3.1. Main Idea of RC-AM

The goal of the attention mechanism is to train a series of attention masks defined by a set of parameters $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$ and use the masks to reweigh these N features depending on their importance for a given set $X = \left\{ x_i \in \mathbb{R}^d, i = 1, \cdots, N \right\}$ of N relevant features of dimension d. Figure 5a,b illustrated two common approaches of learning the attention weights. In this part, a feature vector is also regarded as a feature unit. (a) Local attention: the weights of a feature unit are locally calculated, and they are formed using the individual feature via a shared transformation function. However, this local strategy does not take the correlations between features into consideration, thus ignoring the global scope of information. (b) Global attention: all feature units are employed together to jointly learn attention, for example, utilizing a fully connected operation. However, this method is inefficient and difficult to optimize, particularly when the number of features N is large, resulting in a huge number of parameters being generated.

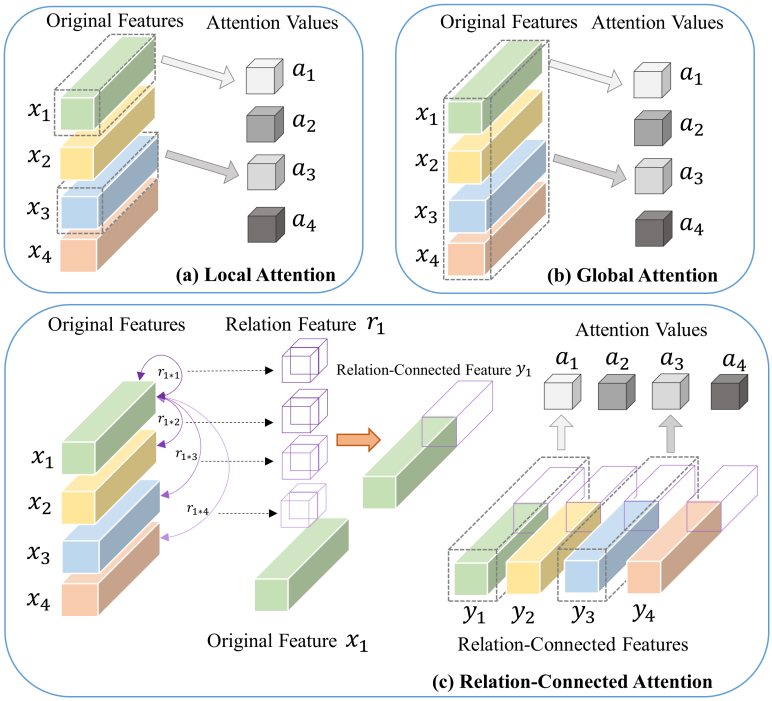


Figure 5. Different methods for the assignment of attention weights. $r_{i*j} = \left[r_{i,j}, r_{j,i} \right]$ is the relation vector.

To solve these problems, we propose a relation-connected attention module that explores global feature information to jointly learn attention weights by fusing features and their relationships. Figure 5c shows the main idea for our relation-connected attention. We first investigated the relationship between the feature unit itself and other feature units by calculating their affinity. Then, they were concatenated together to reflect the current feature unit's global information. Specifically, the module uses $r_{i,j} = R(i, j)$ to indicate the affinity between the ith and the jth feature unit. For the ith feature unit, its affinity vector is $\mathbf{r}_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,N}, r_{1,i}, r_{2,i} \cdots, r_{N,i}] = [R(i,:), R(:,i)]$. $R(i,:)$ which represents the ith row of the affinity matrix, which indicates the relations between the i^{th} feature unit and all other feature units. $R(:,i)$ denotes the ith column of the affinity matrix, which indicates the relations between all other feature units and the ith feature unit. $(r_{i,j}, r_{j,i})$ describes the bi-directional relations between the ith feature unit and the jth feature unit. All the affinity vectors form the affinity matrix $R \in \mathbb{R}^{N \times N}$. The feature itself and the affinity relations are then combined to generate $\mathbf{y}_i = [\mathbf{x}_i, \mathbf{r}_i]$, which is used to infer attention in a learned function. This method for learning attention weights is applied to both spatial and channel attention to obtain more discriminative features.

3.3.2. Relation-Connected Spatial Attention Module

Given a tensor $X \in \mathbb{R}^{C\times H\times W}$ of width W, height H and C channels, we construct a relation-connected spatial attention module named RC-SAM to learn a spatial mask of size $N = H \times W$. The C-dimensional feature tensor at each location is taken as a feature unit, and there are a total of N such feature units. As shown in Figure 6, the feature map is divided into N units in the spatial dimension and we set their label numbers as $1, \cdots, N$. We denote the N feature units as $\mathbf{x}_i \in \mathbb{R}^C$, where $i = 1, \cdots, N$.

The relation $r_{i,j}$ from feature unit i to feature unit j is described as a dot-product affinity as

$$r_{i,j} = F_s(\mathbf{x}_i, \mathbf{x}_j) = G_s(\mathbf{x}_i)^T H_s(\mathbf{x}_j) \tag{1}$$

where G_s and H_s are the transform functions which consist of a 1×1 convolution operation, batch normalization (BN) and ReLU activation, i.e., $G_s(\mathbf{x}_i) = \mathrm{ReLU}(BN(W_G\mathbf{x}_i))$, $H_s(\mathbf{x}_j) = \mathrm{ReLU}\big(BN\big(W_H\mathbf{x}_j\big)\big)$, where $W_G \in \mathbb{R}^{\frac{C}{s_1}\times C}$ and $W_H \in \mathbb{R}^{\frac{C}{s_1}\times C}$. s_1 is a dimension reduction ratio which is a pre-defined hyperparameter. Similarly, we can express the relationship from feature unit j to feature unit i as $r_{j,i} = F_s(\mathbf{x}_j, \mathbf{x}_i)$. Thus, $(r_{i,j}, r_{j,i})$ can be utilized to express the inter-relation between $\mathbf{x}_i$ and $\mathbf{x}_j$, and an affinity matrix $R_S \in \mathbb{R}^{N\times N}$ is used to represent the interaction among all feature units.

The relation vector $\mathbf{r}_i = [R_s(i,:), R_s(:,i)] \in \mathbb{R}^{2N}$ for the i^{th} feature unit is created by concatenating its affinities with all the units in a specified order, i.e., $j = 1, 2, \cdots, N$. Specifically, $\mathbf{r}_3 = [R_s(3,:), R_s(:,3)]$, the third row and third column of the affinity matrix is the relation vector used to mine the attention of the third position, as illustrated in Figure 6.

To draw the attention of the ith feature unit, we also incorporated the local information of the feature itself, in addition to the global information derived by the relations, to explore both the global relational information and original local information. To effectively aggregate the information of the feature map and reduce the amount of parameters, we used global average pooling and global maximum pooling on the original feature maps to aggregate spatial information. For the extraction of spatial information, a common method is global average pooling, which responds to all pixels in the spatial position. In addition, we also used global maximum pooling to extract information. This will form the representations of object features that are different from average pooling, which is helpful for obtaining a more refined attention map. The local and global information of feature units are concatenated together to obtain the relation-connected spatial feature $\mathbf{y}_i$:

$$\mathbf{y}_i = [Avgpool_c(P_s(\mathbf{x}_i)), Maxpool_c(P_s(\mathbf{x}_i)), Q_s(\mathbf{r}_i)] \tag{2}$$

where $\mathbf{y}_i \in \mathbb{R}^{2+\frac{N}{s_1}}$, P_s and Q_s are transform functions which consist of a 1×1 convolution operation, batch normalization (BN) and ReLU activation, i.e., $P_s(\mathbf{x}_i) = \mathrm{ReLU}(BN(W_P\mathbf{x}_i))$, $Q_s(\mathbf{r}_i) = \mathrm{ReLU}\big(BN\big(W_Q\mathbf{r}_i\big)\big)$, where $W_P \in \mathbb{R}^{\frac{C}{s_1}\times C}$, $W_Q \in \mathbb{R}^{\frac{2N}{2s_1}\times 2N}$. $Avgpool_c(\cdot)$ and $Maxpool_c(\cdot)$ denote the global average pooling and global maximum pooling along the channel dimension, respectively. Finally, a learnable model is used to mine valuable knowledge from the extracted global scope information, and the spatial attention weight is inferred through a modeling function as

$$a_i = Sigmoid\Big(W^{''}\mathrm{ReLU}\Big(W^{'}\mathbf{y}_i\Big)\Big) \tag{3}$$

where $W^{'}$ and $W^{''}$ are operated by 1×1 convolution and BN. $W^{'}$ reduces the channel in the ratio of s_2, and $W^{''}$ reduces the number of channels to 1.

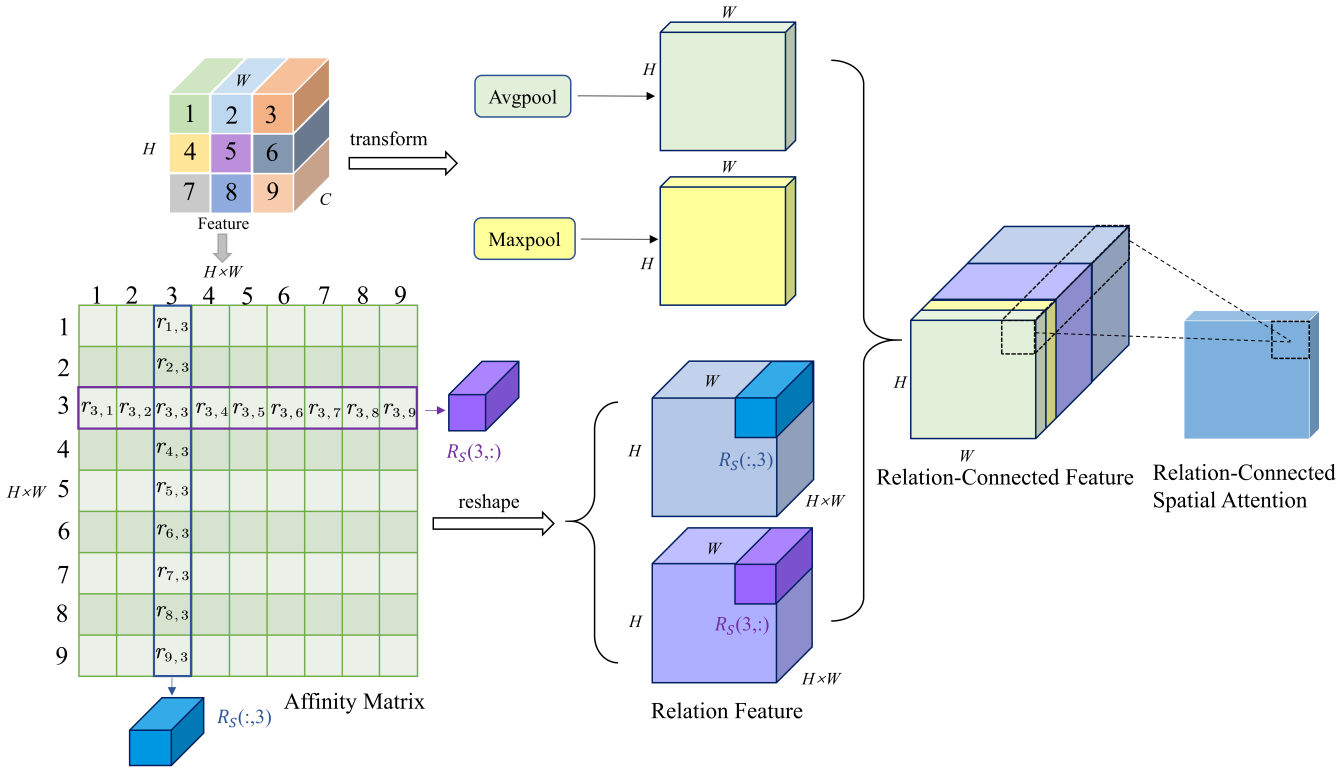


Figure 6. Diagram of our proposed Relation-Connected Spatial Attention Module.

3.3.3. Relation-Connected Channel Attention Module

In addition to spatial attention, we also designed a channel attention model, namely RC-CAM, to learn the C-dimensional channel attention weights. Given a feature tensor $X \in \mathbb{R}^{C \times H \times W}$ with width W, height H and C channels. We consider the feature map of dimension $H \times W$ on each channel to be a feature unit, and there are C feature units for C channels in total. As shown in Figure 7, the C channels are divided into C feature units and their labels are $1, \cdots, C$. The C feature units are denoted as $\mathbf{x}_i \in \mathbb{R}^{H \times W}$, where $i = 1, \cdots, C$.

The relation $r_{i,j}$ from feature unit i to feature unit j, such as spatial attention, is characterized as a dot-product affinity as

$$r_{i,j} = F_c(\mathbf{x}_i, \mathbf{x}_j) = G_c(\mathbf{x}_i)^T H_c(\mathbf{x}_j) \tag{4}$$

where G_c and H_c are transform functions consisting of a 1×1 convolution operation, BN and ReLU activation, which are shared among feature units. An affinity matrix $R_c \in \mathbb{R}^{C \times C}$ is used to express the global information among all feature units. An affinity matrix $R_c \in \mathbb{R}^{C \times C}$ is used to express the global information among all feature units.

The relation vector $\mathbf{r}_i = [R_c(i,:), R_c(:,i)] \in \mathbb{R}^{2C}$ for the ith feature unit is generated by concatenating its relevant relationships with all the units to express global scope information.

To derive the attention weights of the ith channel, we combine the feature itself $\mathbf{x}_i$ and its relation vector $\mathbf{r}_i$ to generate the channel relation-connected feature $\mathbf{y}_i$ and then use Equations (2) and (3) to calculate the channel attention weights a_i. All of the transform functions are shared among channels/units.

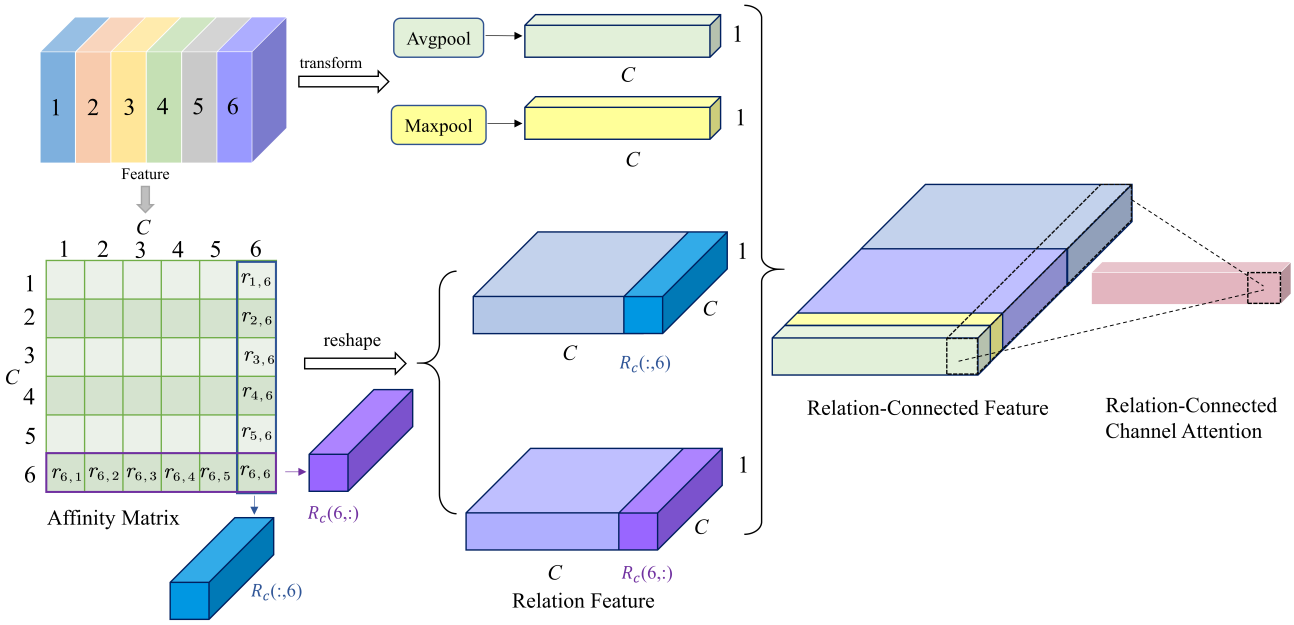


Figure 7. Diagram of our proposed Relation-Connected Channel Attention Module.

3.4. Loss Function

In the proposed model, a multi-task loss is utilized to guide the training of the network. This loss function consists of three parts: regression loss, classification loss and attention loss. It is defined as follows:

$$L = \lambda_1 L_{Reg} + \lambda_2 L_{Cls} + \lambda_3 L_{Att} \tag{5}$$

λ_1, λ_2 and λ_3 are the balance parameters for multi-task loss.

Specifically, L_{Reg} is defined as

$$L_{Reg} = \frac{1}{N} \sum_i p_i' \sum_{j \in \{x,y,w,h,\theta\}} L_{reg}\left(v_{ij}', v_{ij}\right) \tag{6}$$

where N is the number of proposal boxes, p_i' denotes the probability of different classes calculated by the softmax function. v_{*j}' denotes the predicted offset vectors and v_{*j} denotes the offset vector of the ground truth. When $p_i' = 1$, it represents the foreground, and it represents the background when $p_i' = 0$. The regression loss L_{reg} is a $smooth_{L1}$ function which is defined as follows:

$$L_{reg} = smooth_{L1}\left(v' - v\right) \tag{7}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, if|x| < 1 \\ |x| - 0.5, otherwise \end{cases}$$

Furthermore, the classification loss is the softmax cross-entropy loss function, which is consistent with Faster R-CNN. Due to the complexity of remote sensing images, the obtained feature map often contains a lot of noise information, which will make the object features become blurred. Thus, we used a supervised learning way to obtain the attention values, which is beneficial for specific tasks. We introduced the attention loss (L_{Att}) to guide the training of the attention module. Specifically, we first generated a binary map as the mask (label) according to the ground truth. Then, we used the pixel-wise cross-entropy loss of the spatial attention map and the binary map as the attention loss, which is defined as follows:

$$L_{Att} = \frac{1}{h \times w} \sum_{i}^{h} \sum_{j}^{w} L_{att}\left(u_{ij}^{'}, u_{ij}\right) \tag{8}$$

where $u_{ij}^{'}$ is the pixel of the learned spatial attention map and u_{ij} is the pixel of the mask (label).

4. Experiments and Results

4.1. Dataset

(1) NWPU VHR-10 Dataset: To validate the performance of our method, we conducted experiments on the NWPU VHR-10 dataset. It contains ten classes objects which are airplane; ship; storage tank (ST); baseball diamond (BD); tennis court (TC); basketball court (BC); ground track field (GTF); harbor; bridge; and vehicle. There are a total of 800 high-resolution remote sensing images collected from Google Earth and Vaihingen dataset [30].

(2) HRSC2016 [31] is a public dataset of optical remote sensing images for ship detection which is also used in this study to evaluate and analyze the performance of the proposed model. The dataset, which contains 1070 images with resolutions ranging from 0.4 to 2 m, was gathered from famous harbors in Google Earth. There are a total of 2976 ship targets in HRSC2016, with image sizes ranging from 300 × 300 to 1500 × 900. There are various sorts of ships, such as warships, aircraft carriers and cargo ships. Some ships have a big rotation angle, a large aspect ratio and a lot of diversity in appearance.

4.2. Implementation Details

We adopted Adam [32] as the model optimizer in the training stage with a total of 300 epochs. The training batch size is set to 6, and the learning rate is 1×10^{-4} for the first 100 epochs and 1×10^{-6} for the next 200 epochs. Our experiments are implemented in Pytorch 1.5.0 on a NVIDIA Titan XP GPU. The FPN-based Faster R-CNN is used as the baseline and the backbone for initialization in the end-to-end training is ResNet-50 pre-trained on ImageNet. Before training, the images are normalized to a size of 608 × 608 for NWPU VHR-10 and 800 × 800 for HRSC2016 while maintaining their original aspect ratio to prevent the distortion of the target and facilitate image batch training. On the HRSC2016 dataset, the scales of the anchor are set to (1/1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/9) in order to cover as many ship scales as possible. In the RPN training stage, when the IoU > 0.7, the anchor is regarded as a positive sample, and as a negative sample when IoU < 0.3. The other anchors will not be used in the training. Furthermore, in Equation (5), the balance hyperparameters are taken as 4, 2 and 1, respectively. s_1 and s_2 are set to 8.

4.3. Evaluation Criteria

For the algorithm performance, the mean average precision (mAP) is a widely used evaluation metric. mAP is calculated by precision (P) and recall (R). The precision (P) and recall (R) metrics are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

where TP denotes a true positive, indicating the number of correctly detected ships; FP denotes a false positive, indicating the number of incorrectly detected ships; and FN denotes a false negative, indicating the number of undetected ships. We also utilize P and R to obtain the mean average precision (mAP) for each class. It is defined as follows:

$$mAP = \frac{1}{N_{cls}} \int_{0}^{1} P(R) dR \tag{11}$$

where $P(R)$ denotes the P–R function curve. N_{cls} denotes the number of classes.

The $F1$-Score is the harmonic mean of P and R, which is formulated as

$$F1 = \frac{2PR}{P+R} \tag{12}$$

4.4. Comparisons with State-of-the-Art Methods

4.4.1. NWPU VHR-10 Dataset Results

Table 1 reports the comparisons of our method and the state-of-the-art detectors on the HBB prediction task of the NWPU VHR-10 dataset. It contains ten classes of objects which are airplane; ship; storage tank (ST); baseball diamond (BD); tennis court (TC); basketball court (BC); ground track field (GTF); harbor; bridge; and vehicle. We compared our method with state-of-the-art methods which are RICNN, SSD512 [6], R-FCN [13], Faster R-CNN [5] and FMSSD [18].

Table 1. Comparisons with state-of-the-art methods on the NWPU VHR-10 dataset.

Method	Airplane	Ship	ST	BD	TC	BC	GTF	Harbor	Bridge	Vehicle	*mAP*
RICNN [30]	88.35	77.34	85.27	88.12	40.83	58.45	86.73	68.60	61.51	71.70	72.63
SSD512 [6]	90.40	60.90	79.80	89.90	82.60	80.60	98.30	73.40	76.70	52.10	78.40
R-FCN [13]	81.70	80.60	66.20	90.30	80.20	69.70	89.80	78.60	47.80	78.30	76.30
Faster R-CNN [5]	94.60	82.30	65.32	95.50	81.90	89.70	92.40	72.40	57.50	77.80	80.90
FMSSD [18]	99.70	89.90	90.30	98.20	86.00	96.80	99.60	75.60	80.10	88.20	90.40
Ours	99.50	88.40	90.20	98.70	89.20	95.40	99.20	89.60	82.20	92.90	92.50

Our method achieves the highest *mAP* and significantly outperforms the other methods in object detection for small targets. FMSSD [18] achieves good performance on NWPU VHR-10 by using context information in different feature maps. Compared with FMSSD, our method has a similar performance in large-size objects while better results in small objects such as vehicles. In addition, Figure 8 shows some detection results from the NWPU VHR-10 dataset. Our method can effectively detect objects of different classes and scales in remote sensing images.

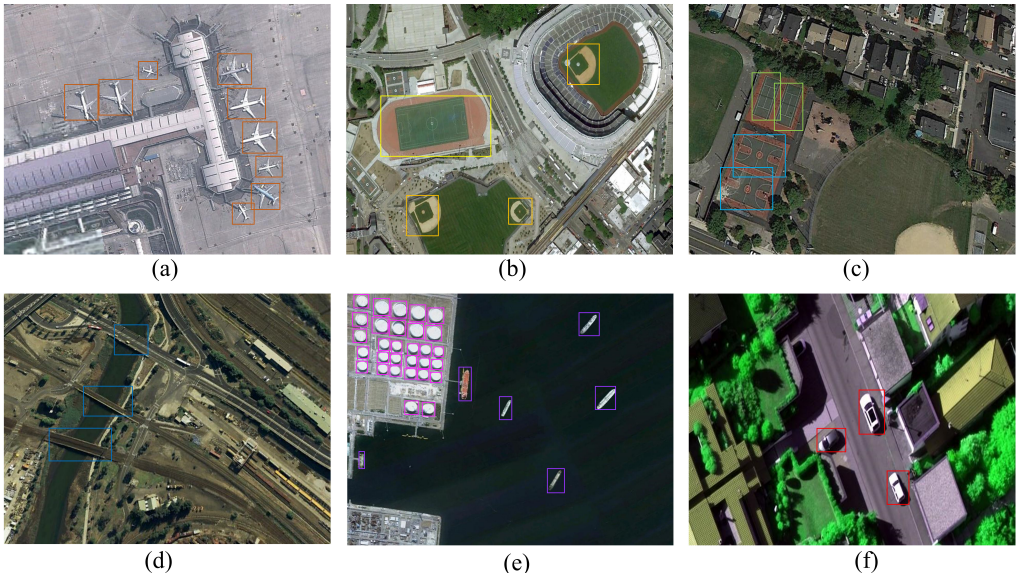


Figure 8. Some detection results of our method on the NWPU VHR-10 dataset: (**a**) Airplane; (**b**) GTF and BD; (**c**) TC and BC; (**d**) Bridge; (**e**) ST and ship; and (**f**) Vehicle.

4.4.2. HRSC2016 Dataset Results

To further illustrate the effectiveness of our proposed algorithm framework, we conducted comparison experiments with state-of-the-art methods on the HRSC2016 dataset, and the experimental results are shown in Table 2. We compared it with the following algorithms: R2CNN [33], RRPN [34], SCRDet [21], RoI Transformer [35] and Gliding Vertex [36], and the experiments show that our algorithm achieves the most competitive results.

Table 2. Comparison with other state-of-the-art methods.

Method	Backbone	Image Size	mAP
R2CNN [33]	ResNet101	800 × 800	73.07
RRPN [34]	ResNet101	800 × 800	79.08
SCRDet [21]	ResNet101	512 × 800	83.41
RoI Trans [35]	ResNet101	512 × 800	86.20
Gliding Vertex [36]	ResNet101	512 × 800	88.20
Ours	ResNet50	800 × 800	88.92
Ours	ResNet101	800 × 800	89.24

From the performance comparison results shown in Table 2, our model using ResNet50 as the backbone outperforms most of the models with ResNet101 as the backbone, which fully demonstrates that the model components proposed in this paper are very effective in improving the performance of the detector. The method we proposed has a large improvement over R2CNN and RRPN, 16.17% and 10.16%, respectively. R2CNN and RRPN, which are originally designed for text detection in arbitrary directions, have poor performance due to the complexity of remote sensing images, although they have similar characteristics to ship detection. RoI Trans [36] uses an RoI transformer in the RPN phase to transform the horizontal RoI into a rotational RoI by fully connected operation, which effectively improves the detection accuracy. Compared with RoI Trans, we achieved 3.04% higher in mAP. The performance of Gliding Vertex is comparable to ours in this paper, with a difference of only 1.04%. Gliding Vertex [36] describes oriented targets by sliding the four vertices of the horizontal bounding box on each corresponding edge, which facilitates the learning of offsets and avoids the confusing problem of having sequential labeling points for oriented targets. The method works very well for both remote sensing target detection and text detection. Our method benefits from the multi-level receptive field feature aggregation capability of MR-FEM and the feature refinement of RC-AM. Figure 9 shows comparative visual results with other methods. The red box represents the false detections and the yellow box represents the missed detections.

Figure 10 shows some of the visualization results of the method in this paper on the HRSC2016 dataset. From the demonstrated results, it can be seen that our method can effectively detect multiscale ship targets and is more friendly in the detection of small ships. It can detect inshore ships of different types and various sizes, as well as ships close to each other. For harbor, roof and other disturbances, our method can effectively distinguish them from ship targets with good robustness.

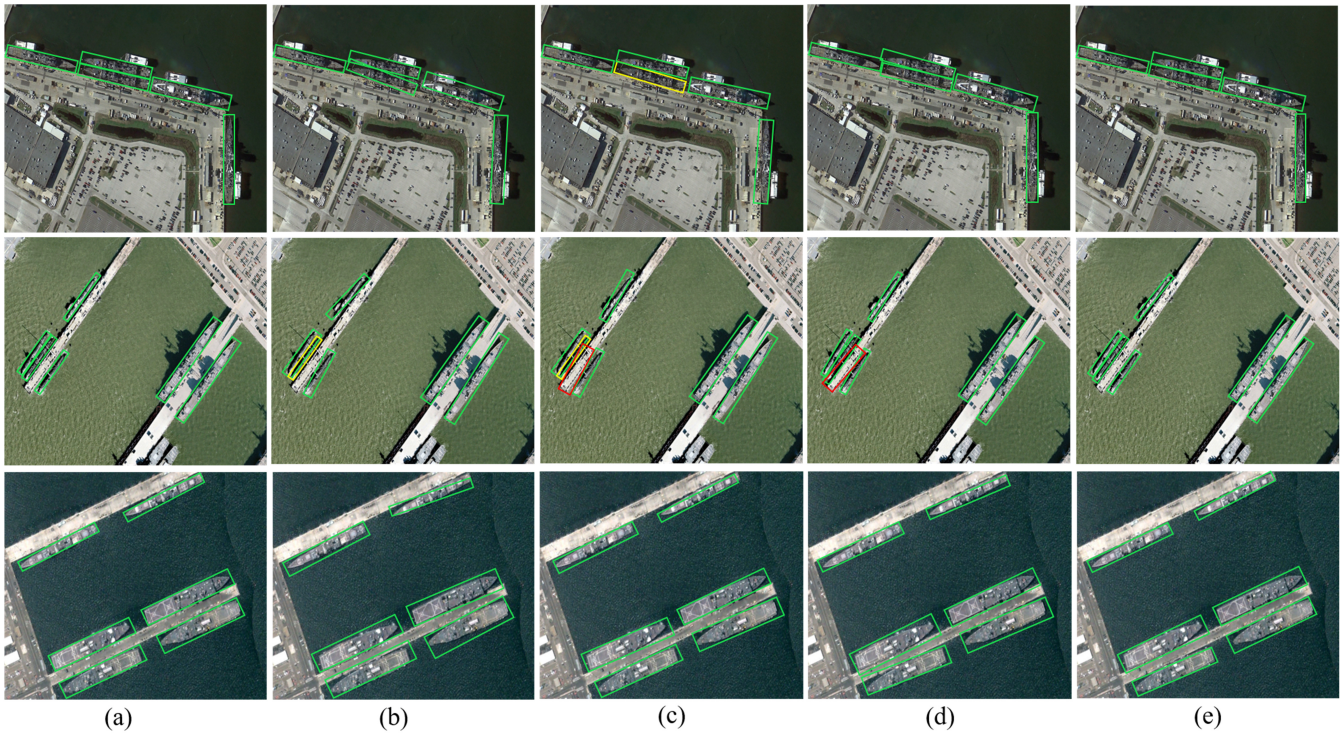


Figure 9. Comparative visual results with other methods on the HRSC2016 dataset: (**a**) ground truth; (**b**) R2CNN; (**c**) RRPN; (**d**) SCRDet; and (**e**) ours.

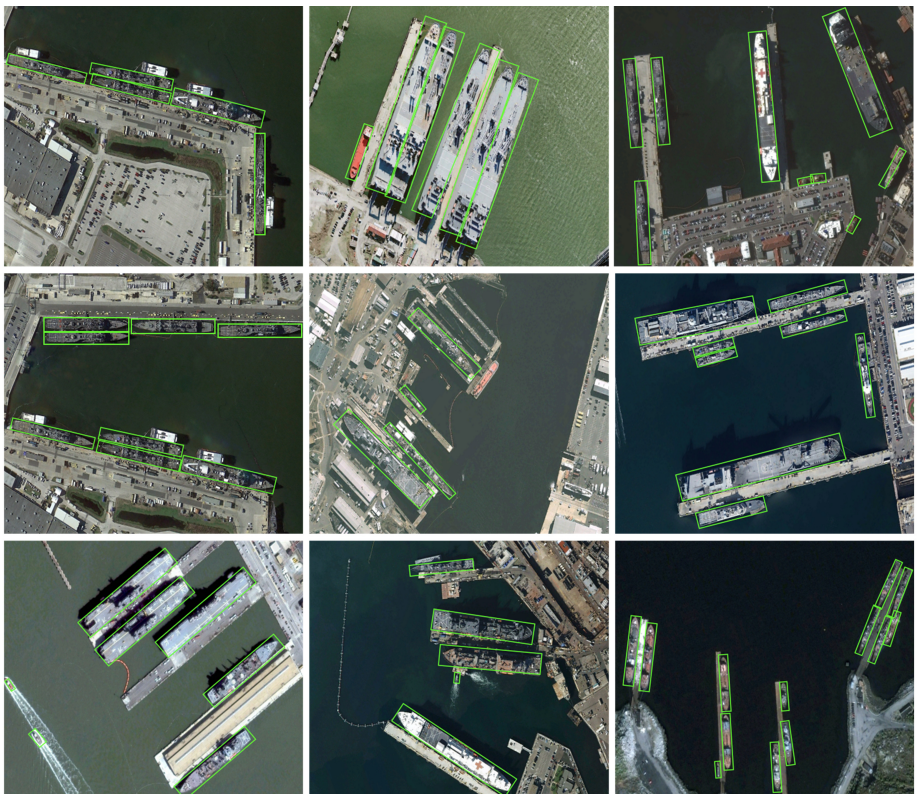


Figure 10. Visualization results of the method in this paper on the HRSC2016 dataset.

4.5. Discussion

4.5.1. Ablation Study on NWPU VHR-10 Dataset

In order to verify the importance of each different module of the proposed model in this paper, a series of ablation experiments were performed using the NWPU VHR-10 dataset. The categories are airplane; ship; storage tank (ST); baseball diamond (BD); tennis court (TC); basketball court (BC); ground track field (GTF); harbor; bridge; and vehicle. The main purpose was to investigate the impact of the proposed multi-receptive-field feature extraction module (MR-FEM) and relation-connected attention module (RC-AM) on the performance of the algorithm. In this experiment, the FPN-based Faster R-CNN algorithm was used as the baseline, and the pre-trained ResNet-50 was used as the backbone. All other settings were made the same to ensure fairness. The experimental results are shown in Table 3.

Table 3. Performance of different modules in our model on the NWPU VHR-10 dataset.

Method	Airplane	Ship	ST	BD	TC	BC	GTF	Harbor	Bridge	Vehicle	*mAP*
Baseline	94.60	82.30	65.32	95.50	81.90	89.70	92.40	72.40	57.50	77.80	80.90
+MR-FEM	95.84	86.36	80.64	95.62	82.40	89.82	92.62	82.10	64.20	86.42	85.60
+RC-AM	98.40	85.20	86.20	96.20	87.32	93.44	92.46	83.42	77.34	83.50	88.35
Ours	99.50	88.40	90.20	98.70	89.20	95.40	99.20	89.60	82.20	92.90	92.50

The experimental results show that the addition of MR-FEM is beneficial to improve the detection accuracy by 4.7% in terms of *mAP*. When RC-AM is added, the detection accuracy is improved by 7.45% compared to the baseline model. The performance is achieved is better while using both of them. The experiments fully demonstrate the effectiveness of the proposed method. MR-FEM aggregates the contextual information of different receptive fields on the feature map, which improves the detection accuracy of small targets such as ships and vehicles. BC and TC are different object classes which consist of a similar appearance and features. They achieve larger improvements with RC-AM. RC-AM enhances the feature selection and refinement capability by obtaining the global and local attention through relation modeling, which makes the features between similar targets more distinguishable and thus improves the detection performance. RC-AM not only enhances the distinguishability of features among different classes of objects, but also reduces the interference of background features, which is very helpful for object detection tasks.

4.5.2. Ablation Study on HRSC2016 Dataset

We also performed a series of ablation experiments using the HRSC2016 dataset on OBB detection task. In this experiment, the FPN-based rotated Faster R-CNN algorithm was used as the baseline, and the pre-trained ResNet-50 was used as the backbone. All other settings are made the same. The experimental results are shown in Table 4.

As can be seen from Table 4, both the proposed MR-FEM and RC-AM can effectively improve the performance of the baseline, which are 1.24% and 3.48% higher than the baseline model, respectively, and the AP then achieves 88.92% when using both of them. Our proposed MR-FEM and RC-AM were proven helpful for ship detection tasks. MR-FEM can extract multi-receptive-field features from the convolutional layers at each stage of the backbone. Through the convergence of multi-level receptive fields, the contextual information is increased in the feature map, which is very effective for the detection of multiscale ship targets. The RC-AM is an important module of the method in this paper, through which the relation between the features and the feature itself are combined to explore the global attention. RC-AM can automatically select and refine the features, so as to highlight the foreground and eliminate the effect of noise.

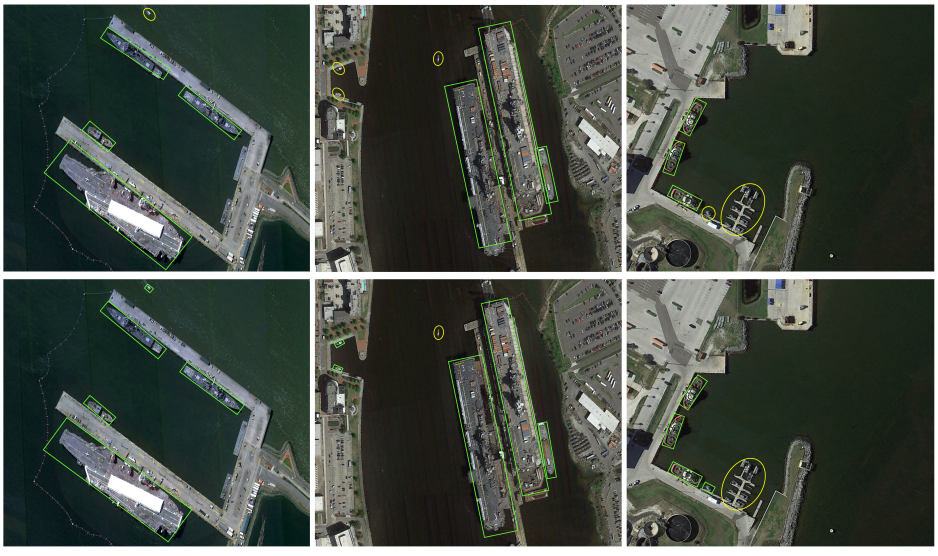
Table 4. Performance of different modules in our model on the HRSC2016 dataset.

Method	Precision	Recall	*F*1-Score	*mAP*
Baseline (ResNet-50)	85.42	86.88	86.14	84.20
+MR-FEM	87.30	88.25	87.77	85.44
+RC-AM	90.87	90.24	90.55	87.68
+MR-FEM+RC-AM	92.68	91.52	92.10	88.92

Figure 11 shows the visualization of ship detection results for this experiment. Figure 11a shows the comparison results between the baseline and the baseline adding MR-FEM, and it can be found that when MR-FEM is applied to the baseline, the model has better robustness for multiscale ship targets and is friendlier for small target detection. Figure 11b demonstrates that, after using RC-AM, the network is able to resist the interference of complex backgrounds and reduce the missed detections and false positives.

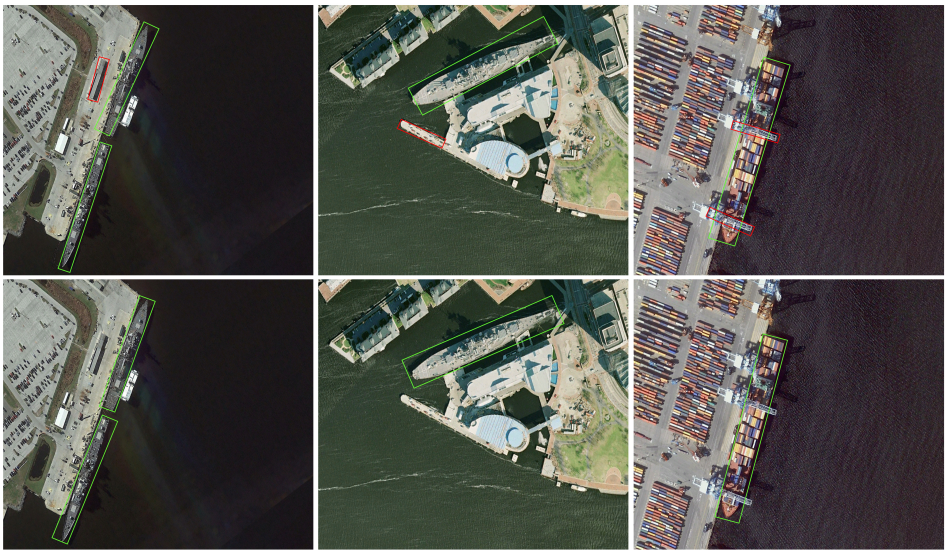
Studies have shown that the attention mechanism plays an important role in target detection. In this paper, in order to obtain more fine-grained ship features with more discriminative power, a relation-connected attention module is proposed, which consists of two parts: spatial attention and channel attention. To further illustrate the superiority of this module, it is compared with other commonly used attention methods, and the experimental results are shown in Table 5.

SENet [23] is a very classical and widely used attention module and has even become standard in some baseline models. In SE, they use spatial global average pooling features and utilize two fully connected (FC) nonlinear layers to compute channel attention. CBAM [24] designs a similar channel attention with reference to SE and uses a 7×7 filter to compute spatial attention. The ECA [25] module is modified from SE, proposing that only a few adjacent channel information interactions are required instead of interacting with all channels, which reduces the computational cost.



(**a**)

Figure 11. *Cont.*



(**a**)

Figure 11. Visualization of the ablation experiment results for MR-FEM and RC-AM: (**a**) The first row is the baseline method and the second row is the result of baseline+MR-FEM. The yellow circle represents the missed detection; (**b**) the first row is the baseline method and the second row is the result of baseline+RC-AM. The red box represents the misjudgment.

Table 5. Performance comparison for different attention modules on the HRSC2016 dataset.

Method	Precision	Recall	*F*1-Score	*mAP*
Baseline + MR-FEM	87.30	88.25	87.77	85.44
+SE [23]	87.92	89.02	88.47	87.29
+CBAM [24]	88.54	89.24	88.89	87.64
+Non-local [26]	89.81	90.36	90.08	87.95
+ECA [25]	88.63	89.30	88.96	87.70
+RC-AM (ours)	92.68	91.52	92.10	88.92

From Table 5, we can see that the RC-AM proposed in this paper has a 1.63%, 1.28% and 1.22% performance improvement over SE, CBAM and ECA, respectively. Non-local (NL) [26] uses pairwise relations as weights to obtain long-range information to reweigh the features. However, NL only uses them for weighted summation and ignores mining global range information from the relations. We improved the performance from 87.95% to 88.92% using RC-AM compared to NL. Thanks to our effective way of obtaining the attention values, RC-AM achieves the best performance among the compared attention modules.

The RC-AM proposed in this paper contains two parts: spatial attention and channel attention. To further investigate the respective roles of spatial attention, channel attention, their combinations and the effects of their connection methods, we conducted some further experiments. Table 6 shows the comparisons of our RC-CAM(RC-AM_C), RC-SAM(RC-AM_C) and their combinations (RC-AM_S//C, RC-AM_CS and RC-AM_SC).

As can be seen in Table 6, either RC-AM_C or RC-AM_S significantly enhances the performance over baseline by 1.62% and 0.76%, respectively, and the performance improvement is much greater when using their combined version (RC-AM_SC) from 85.44% in the baseline to 88.92%. In this experiment, we investigate three methods of combination: sequential spatial–channel (RC-AM_SC), sequential channel–spatial (RC-AM_CS) and parallel fusion (RC-AM_S//C). Compared with RC-AM_S and RC-AM_C, RC-AM_SC achieves the best performance, which is 2.72% and 1.87% higher in mAP. The experiments also show that the sequential connection of spatial and channel attention is better than the parallel connection, where RC-AM_SC is slightly better than RC-AM_CS. The comprehen-

sive experiments demonstrate that the relation-connected attention assignment strategy proposed in this paper is effective, and mining the global attention by modeling the relationship between features and stacking them together can significantly help improve the model performance.

Table 6. Performance comparisons of our models with the baseline, and the effectiveness of channel attention and spatial attention on the HRSC2016.

Method	Precision	Recall	*F*1-Score	*mAP*
Baseline + MR-FEM	87.30	88.25	87.77	85.44
+RC-AM_C	88.67	88.82	88.74	87.05
+RC-AM_S	88.25	88.70	88.47	86.20
+RC-AM_S//C	89.23	89.78	89.50	87.62
+RC-AM_CS	90.80	90.24	90.52	88.76
+RC-AM_SC (ours)	92.68	91.52	92.10	88.92

5. Conclusions

In this paper, we propose a novel method combined with multi-receptive-field features and relation-connected attention for multiscale object detection in optical remote sensing images. Considering the various scales of objects, a multi-receptive-field feature extraction module containing atrous convolution with different dilation rates is designed to extract multiscale context features, which effectively adapts to the scale changes of an object. To distinguish different objects with similar scales, a relation-connected attention module is proposed to dynamically select and refine features and make them more discriminative, which can mine a spatial and channel attention through relation modeling. RC-AM can effectively guide the network to focus on object regions and strengthen object features while suppressing redundant background information. It also makes the network more robust for object detection under complex background conditions in optical remote sensing images. MR-FEM and RC-AM are effective plug-and-play blocks to improve the performance of basic deep CNNs. The experimental results on the NWPU VHR-10 and HRSC2016 datasets show that the algorithm proposed in this paper can accurately detect objects of different scales and distinguish different object classes of similar scales, which achieves competitive results and better robustness.

Author Contributions: J.L. supervised the study, designed the architecture and revised the manuscript; D.Y. wrote the manuscript and designed the comparative experiments; F.H. made suggestions to the manuscript and assisted D.Y. in conducting the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the China High-Resolution Earth Observation System Program under Grant 41-Y30F07-9001-20/22, by the Innovative talent program of Jiangsu under Grant JSSCR2021501, and by the High-level talent plan of NUAA, China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

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

References

1. Su, H.; Wei, S.; Liu, S.; Liang, J.; Wang, C.; Shi, J.; Zhang, X. HQ-ISNet: High-quality instance segmentation for remote sensing imagery. *Remote Sens.* **2020**, *12*, 989. [CrossRef]
2. Dong, C.; Liu, J.; Xu, F. Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor. *Remote Sens.* **2018**, *10*, 400. [CrossRef]

3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
4. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
7. 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 30 2016; pp. 779–788.
8. 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.
9. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
10. 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.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
12. 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, 21–27 July 2017; pp. 2117–2125.
13. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016; pp. 379–387.
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy 22–29 October 2017; pp. 2961–2969.
15. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 28–22 June 2018; pp. 4203–4212.
16. Singh, B.; Davis, L.S. An analysis of scale invariance in object detection snip. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 28–22 June 2018; pp. 3578–3587.
17. Singh, B.; Najibi, M.; Davis, L.S. Sniper: Efficient multi-scale training. *arXiv* **2018**, arXiv:1805.09300.
18. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3377–3390. [CrossRef]
19. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 28–22 June 2018; pp. 8759–8768.
20. Guan, W.; Zou, Y.; Zhou, X. Multi-scale object detection with feature fusion and region objectness network. In Proceedings of the2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2596–2600.
21. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 8232–8241.
22. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6054–6063.
23. 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, 28–22 June 2018; pp. 7132–7141.
24. 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.
25. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.
26. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 28–22 June 2018; pp. 7794–7803.
27. 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.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
28. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
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. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
31. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, SCITEPRESS, Porto, Portugal, 24–26 February 2017; Volume 2, pp. 324–331.

32. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
33. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2cnn: Rotational region cnn for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
34. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimedia* **2018**, *20*, 3111–3122. [CrossRef]
35. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
36. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef] [PubMed]