remote sensing

MDPI

Article

Generating Anchor Boxes Based on Attention Mechanism for Object Detection in Remote Sensing Images

Zhuangzhuang Tian [1], Ronghui Zhan [1], Jiemin Hu [2], Wei Wang [1,*] and Zhiqiang He [1] and Zhaowen Zhuang [1]

[1] College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China; tzz14@nudt.edu.cn (Z.T.); zhanrh@nudt.edu.cn (R.Z.); hezhiqiang@nudt.edu.cn (Z.H.); zwzhuang@nudt.edu.cn (Z.Z.)

[2] School of Informatics Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; hujiemin@zstu.edu.cn

* Correspondence: wewan@kth.se

Received: 1 July 2020; Accepted: 24 July 2020; Published: 28 July 2020

Abstract: Nowadays, object detection methods based on deep learning are applied more and more to the interpretation of optical remote sensing images. However, the complex background and the wide range of object sizes in remote sensing images increase the difficulty of object detection. In this paper, we improve the detection performance by combining the attention information, and generate adaptive anchor boxes based on the attention map. Specifically, the attention mechanism is introduced into the proposed method to enhance the features of the object regions while reducing the influence of the background. The generated attention map is then used to obtain diverse and adaptable anchor boxes using the guided anchoring method. The generated anchor boxes can match better with the scene and the objects, compared with the traditional proposal boxes. Finally, the modulated feature adaptation module is applied to transform the feature maps to adapt to the diverse anchor boxes. Comprehensive evaluations on the DIOR dataset demonstrate the superiority of the proposed method over the state-of-the-art methods, such as RetinaNet, FCOS and CornerNet. The mean average precision of the proposed method is 4.5% higher than the feature pyramid network. In addition, the ablation experiments are also implemented to further analyze the respective influence of different blocks on the performance improvement.

Keywords: remote sensing; object detection; convolutional neural network; attention mechanism

1. Introduction

As an important tool of earth observation, remote sensing has been widely used in civil and military fields, such as geological monitoring, climatic forecast, ecological environment monitoring, and battle reconnaissance. With the continuous development of remote sensing technology, rapidly increasing images make manual interpretation a tremendous task, and as a result, automatic interpretation has drawn considerable research attention. Object detection plays an important role in image interpretation, the purpose of which is to detect the locations of the objects of interest and identify their corresponding categories. For optical remote sensing images, researchers have done a lot of work on object detection [1–4]. With the advancement of computer vision and machine learning, especially deep learning, object detection for natural images is more and more mature. Many researchers have drawn on the successful experience of natural images, and applied deep-learning-based methods to remote sensing images [5–8].

The object detection methods based on deep learning can be divided into one-stage and two-stage methods according to the detection process. One-stage methods, such as you only look once (YOLO) [9–11], single shot multi-box detection (SSD) [12] and RetinaNet [13], sample dense boxes at different positions of the images. Multiple scales and aspect ratios are used for sampling. The convolutional neural network (CNN) is used to extract features, classify and locate the objects directly. The whole process needs only one step, and therefore, the detection is efficient and fast. However, a considerable disadvantage of one-stage methods is that sampling dense boxes makes the positive and negative samples extremely imbalanced. Actually, most scenes of remote sensing images are large and the distribution of the contained objects is sparse, which exacerbates the imbalance problem. Therefore, in such a case, the performance of one-stage methods is usually unsatisfactory. The current object detection methods for remote sensing images are mostly two-stage methods, which are proposed on the basis of region CNN (RCNN) [14].

Regarding the RCNN-based methods, object detection includes two stages: the generation of proposal boxes, and the regression and classification of the boxes. In the first stage, according to the feature maps extracted from the backbone network, a series of class-agnostic proposal boxes are generated for each image by the proposal generation algorithms, such as selective search algorithm [15] and region proposal network (RPN) [16]. In the second stage, the feature maps are cropped according to the proposal boxes and then resized into the same size. These feature maps are used not only to predict the objects' categories, but also to further fine-tune the proposal boxes to make the predicted boxes more accurate.

The success of the RCNN-based methods is largely attributed to the generation of proposal boxes. The proposal boxes make the positive and negative samples used for training more balanced, and the double box regressions also make the predicted boxes more accurate. In addition, the cropped regional feature maps also enable the network of the second stage to avoid the influence of the background. For natural images, the methods based on RCNN have achieved satisfactory results. However, compared with natural images, remote sensing images have more diverse objects with different scales and aspect ratios, and more complex backgrounds. In this paper, in order to solve these problems, we make a variety of improvements and the main contributions are as follows:

1. In general, remote sensing images contain large scenes and sparse objects. In the CNN for feature extraction, the weights of different positional features are the same and the network pays uniform attention to them. However, there are many background areas in the scene, which may interfere with object detection and lead to false predictions. Therefore, we apply the attention mechanism into the proposed method to adjust the weights of the features. The generated attention map can be regarded as the spatial weights of different positions on feature maps. It can make the network pay more attention to the object area rather than the background area. The weighted feature maps are used to model the channel-wise dependencies, which can selectively enhance the informative feature channels, thereby improving the feature maps.
2. In most RCNN-based methods, the anchor boxes used to generate proposal boxes are preset with sliding windows and fixed shapes. However, the fixed shapes can not fit the diverse objects in remote sensing images. In this paper, we think that the generated attention map can also reflect the location of the object, and use the attention map to predict the positions of the anchor boxes. The guided anchoring method is incorporated into the proposed detection framework to predict the corresponding shapes of the anchor boxes.
3. In order to make the feature maps match better with the generated anchor boxes, as well as get more effective regional features, we adopt a modulated feature adaptation module (MFAM) to transform the feature maps. The module first calculates the offset and modulation scalar according to the predicted shape maps of anchor boxes. The modulated deformable convolutional layer [17] is used to produce the transformed feature maps, which are more matched with the anchor boxes.

2. Related Work

The task of object detection in remote sensing images has been extensively studied since the 1980s [18]. Object detection methods based on machine learning are usually divided into three stages: region proposal generation, feature extraction and classification. Region proposals are the regions that may contain the objects, and they are usually obtained by certain algorithms such as sliding window. Feature extraction is used to transform the image pixels to discriminative and representative features. The widely used features for object detection mainly include: histogram of oriented gradients feature [4], Bag-of-words feature [3] and texture features [19]. After obtaining the features of each region proposal, a classifier is needed to classify the features to determine whether the region is object or background. The commonly used classifiers include support vector machine [2], AdaBoost [1], *k*-nearest-neighbor [20], etc. Besides the optical and infrared sensors, synthetic aperture radar (SAR) is also widely used in remote sensing applications due to its capability of producing all-weather, all-time and high-resolution images. There are also many segmentation and object detection methods for SAR images [21–24]. These methods have achieved certain results in their respective fields, but they still need to manually select feature extraction methods and classifiers.

With the development of object detection technology based on deep learning, its applications in the field of remote sensing are increasing. Researchers have tried to improve the deep-learning-based methods from many aspects in recent years. In [5], Gong Cheng et al. proposed rotation-invariant CNN, which imposes a rotation-invariant regularizer and a Fisher discrimination regularizer on the CNN features. R^2-CNN [25] was proposed to tackle the object detection in large-scale remote sensing images, and it designs a lightweight residual structure called Tiny-Net to improve the speed of feature extraction. In order to detect the rotated objects, Jian Ding et al. [7] proposed an RoI Transformer to transform the horizontal boxes into rotated boxes. In this section, we mainly review common methods about the attention mechanism and the generation of proposal boxes.

2.1. Attention Mechanism

Humans usually pay more attention to some specific parts of the visual scene according to their needs, while ignoring the other parts. The above phenomenon is often referred to as the attention mechanism. For machine learning, by applying the attention mechanism, we can learn the importance of each element in the feature, and obtain its corresponding weight coefficient.

Wang et al. [26] proposed the non-local operation to model the pixel-level pairwise relations and computed the response at a position as a weighted sum of the features at all positions. The attention weight means the captured long-range dependency that is not constrained by the distance. However, the generation of an attention map has a high computation complexity. Therefore, Huang et al. [27] proposed a criss-cross network (CCNet) to model the pixel-level pairwise relations in a resource-saving way. For each position, CCNet obtains the contextual information of the surrounding points on the criss-cross path through a criss-cross attention module. The pixel-level pairwise relations between other points can be obtained after a recurrent operation. Thus, each position in the final output feature maps can capture the long-range dependencies from all points. In order to further reduce the amount of computation, Cao et al. [28] tried to use the query-independent attention map instead of the query-specific attention map used in non-local block, to aggregate the features of all positions together. They used a convolutional layer with 1×1 convolutional kernel and softmax function to calculate the global attention map and share it with all query positions. The query-independent attention map not only has a lower computation cost, but also maintains the accuracy.

The above methods are mainly used to calculate the attention weight in the spatial dimension. In addition, the correlation between different channels of feature maps is also important to be explored. To better build the dependencies between the channels, squeeze-excitation (SE) network [29] obtains the global distribution of channel-wise responses through the squeeze operation, and produces the weight value of each feature channel through excitation operation. The weight values are used to selectively emphasize informative features and suppress useless ones.

2.2. Generation Methods of Proposal Box

In the RCNN-based approaches, the first stage produces the proposal boxes that may contain objects. The most straightforward way to generate proposal boxes is the sliding window [30]. The sliding window method uses the boxes of different scales and aspect ratios to search all possible objects over the entire image, then the classifier is used to identify all the boxes and leave the boxes with high scores. Obviously, this approach produces too many redundant boxes and it is complex and unfeasible. Therefore, researchers have proposed a variety of region proposal algorithms to promote efficiency.

Selective search [15] is a region proposal algorithm used in R-CNN [14]. Selective search uses a graph-based segmentation method [31] to initialize the segmentation regions. Then, the similarities between all regions are measured, and the corresponding regions with the greatest similarity are merged as a proposal box. The similarity measures used in selective search include color, texture, size and shape compatibility. Selective search does not require training, but it has a large amount of calculation and the speed of generating region proposal is slow. Therefore, it is difficult to meet the needs of real-time detection.

In order to speed up the generation of proposal box and further use the features extracted by CNN, Faster R-CNN [16] directly uses RPN to obtain the proposal boxes. RPN slides a small network over the feature maps output by the backbone network, and each location of the sliding window corresponds to several anchor boxes of different scales and aspect ratios. Each sliding window is mapped to a low-dimensional feature vector. The feature vector is used to predict the score of being an object and regression values between the ground truth box and the anchor box. However, the application of the anchor box introduces more hyperparameters and design choices, such as the sizes and aspect ratios of anchor boxes.

In order to avoid setting anchor boxes, Hei Law et al. [32] used a single CNN detector named CornetNet to detect an object bounding box as a pair of key points, namely the top-left corner and the bottom-right corner. Transforming the detection of bounding box into the detection of key points can eliminate the dependence on the design of the anchor boxes; therefore, the methods like CornetNet are also known as anchor-free methods. Other anchor-free methods include CenterNet [33] and ExtremeNet [34]. Anchor-free methods have achieved good results in natural images, and they are also getting more and more attention from researchers.

In the proposed method, we used the attention mechanism to enhance the feature maps and make the network pay more attention to the object areas. Then, we adopted the generated attention map and guided anchoring method to obtain more adaptive anchor boxes. In addition, the modulated feature adaptation module was used to make the feature maps better match with the anchor boxes.

3. Method

3.1. Overall Detection Framework

Firstly, we clarify the overall detection framework in this subsection. The proposed method takes the whole image as input, and applies the backbone network to extract the feature maps. The attention mechanism was adopted to enhance the feature maps according to the generated attention maps. The attention maps were also used to represent the locations of the anchor boxes. The enhanced feature maps were then utilized to generate the shape maps of the anchor boxes. The attention maps and shape maps can obtain the anchor boxes based on the guided anchoring method. The modulated feature adaptation module can obtain the offset and modulated scalar from the shape maps, and transform the feature maps. The transformed feature maps and the generated anchors were finally used to predict the categories and locations of the objects. The schematic of the proposed detection method is shown in Figure 1.

In the following subsections, we will introduce the attention mechanism, anchor generation and modulated feature adaptation module in detail.

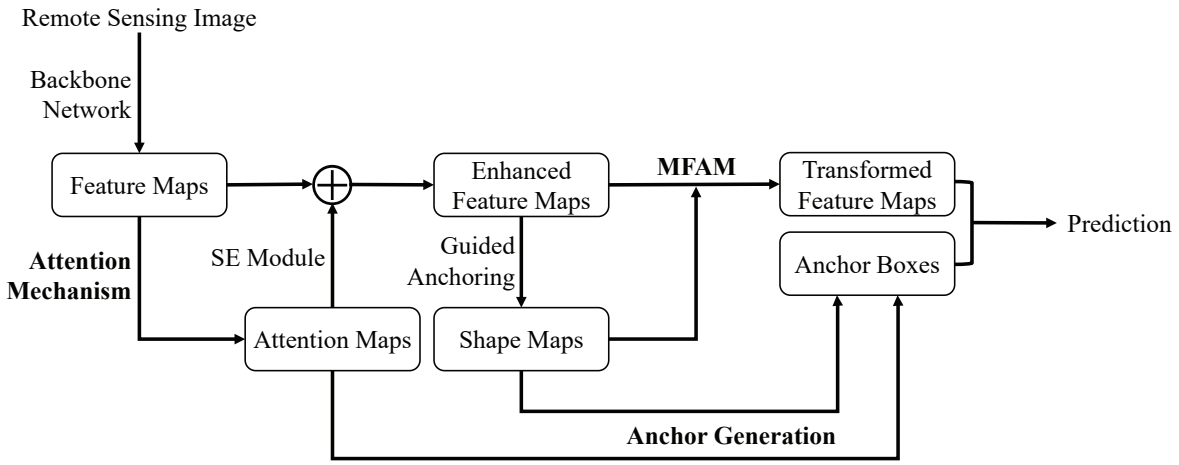

Figure 1. The schematic of the proposed detection framework.

3.2. Attention Mechanism

Generally, the scene of the remote sensing images is large, while the object areas occupy only a small part of the whole scene. The information in the background area may interfere with object detection. Therefore, it would be beneficial if suppressing the feature of background areas and making the network more focused on the feature of the object areas. In order to achieve this end, the attention mechanism is applied to the network.

The calculation of the values in the attention mechanism mainly involves queries and keys. However, Xizhou Zhu [35], Cao [28] et al. find that the attention weights are uncorrelated with the queries. They are only related to the keys. Therefore, the proposed method adopts the query-independent attention mechanism. Referring to [35], the attention weight can be modeled as:

$$A(q, k, z_q, x_k) = \text{softmax}(u^{\mathrm{T}} V^{C} x_k), \tag{1}$$

where z_q denotes the content of the q-th query element, x_k denotes the content of the k-th key element. $A(q, k, z_q, x_k)$ is the attention weight in the attention sub-network, and V^C denotes the learnable embedding matrix for the key content, and u is a learnable vector.

The final formula of attention weights is:

$$y_q = W[\sum_{k \in \Omega_q} \text{softmax}(u^{\mathrm{T}} V^{C} x_k) \otimes x_k]. \tag{2}$$

where Ω_q is the key region for the query, W is the learnable weight. y_q is the output of the attention mechanism. $\otimes$ means element-wise multiplication.

As described in Section 2, the squeeze-excitation module is proposed to capture the interdependencies between feature channels. It is worth noting that the final output y_q is also used to model the channel-wise dependencies in most attention mechanisms. Therefore, we introduce the SE module into the attention mechanism. In specific, the learnable weight W in Equation (2) is replaced by SE module. In [29], the squeeze operation is global pooling. However, the attention mechanism has multiplied feature map x_k by attention map A. Therefore, the squeeze operation can be regarded as the weighted pooling in the proposed method. Moreover, we add the layer normalization [36] into the SE module to make the network easier to optimize. Specifically, layer normalization directly calculates the statistics from the hidden units of each layer without introducing any new dependencies between training cases. The stochasticity from the statistics can serve as a regularizer during training, thus to enhance the generalization of the network.

The final calculation flowchart of attention weights is shown in Figure 2.

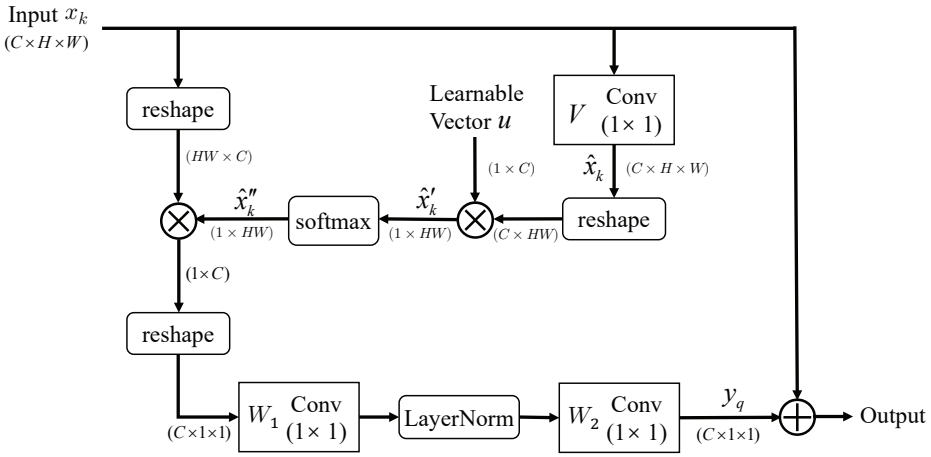

Figure 2. The flowchart of attention weight calculation in the proposed method. The numbers in parentheses denote the shape of the feature map.

According to Figure 2, the input feature maps x_k can generate the feature maps $\hat{x}_k$ by convolution. $\hat{x}_k$ can be weighted and summed up to get $\hat{x'}_k$ by multiplying with the learnable vector u. $\hat{x'}_k$ is normalized by the softmax function and obtains the final attention weights $\hat{x''}_k$. According to the learned attention map $\hat{x''}_k$, the input feature maps x_k are element-wise weighted and summed up to obtain the channel-level dependencies. This operation corresponds to the squeeze in the original SE module. The dependencies then generate the final output y_q through the excitation operation consisting of two convolutions. Subsequently, y_q is broadcast in each spatial dimension to match the input features. The enhanced feature maps are generated by element-wise adding the input feature maps and the broadcast channel-level dependencies together.

The attention map represents the area that the network should pay attention to in the global scene. For the task of object detection, we think that the most noteworthy areas are the regions where the objects are located. In practice, the query-independent attention maps of the images are visualized in Figure 3. It also confirms our view that the values of the object area in the attention map are usually higher than that of the background area. Therefore we try to use the attention map to predict the positions of objects. Specifically, attention maps are incorporated into the guided anchoring method to obtain more adaptive anchor boxes. The procedures will be described in the next section.

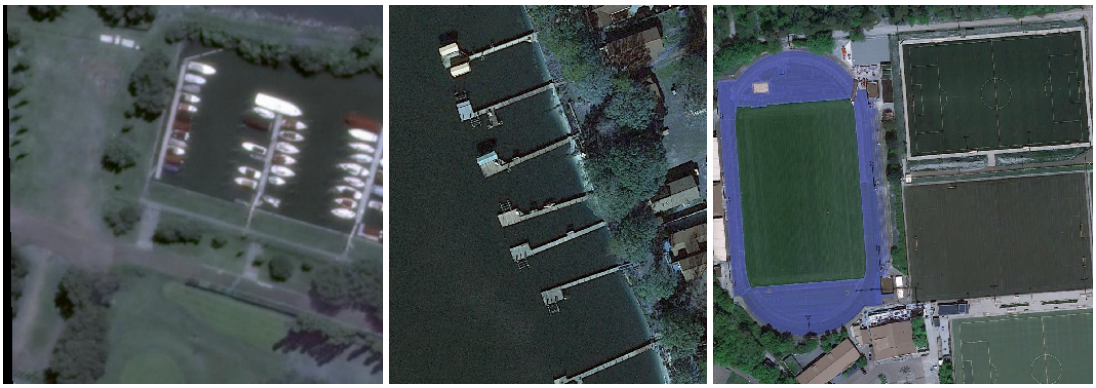

Figure 3. *Cont.*

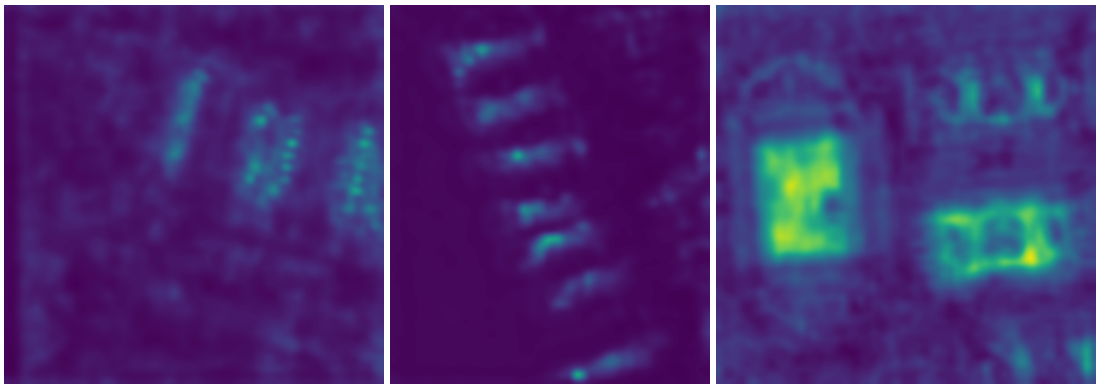

Figure 3. An example of query-independent attention map. Top row is the remote sensing images, and bottom row is the corresponding attention map.

3.3. Anchor Generation

As is well known, the anchor box prediction is a preliminary work for the proposal box. One task of the proposal box is to find the location of the potential object; therefore, the positions of the anchor boxes only need to be concentrated on the regions where the objects are located. In [28], the guided anchoring method predicts the probability of the object at each position through an independent subnetwork and obtains a probability map. Since the predicted position map reflects the potential object locations throughout the scene, it is similar to the attention map. We reuse the attention map in the guided anchoring method to predict the anchor boxes. A threshold for the attention map is used to determine whether the region contains the objects. The regions above the threshold are where the generated anchor boxes should be located.

Besides the location, the generation of the anchor box also needs to predict its corresponding shape. However, different from the bounding box, the anchor box does not have a proposal box to provide prior information, which is also the reason why the conventional methods need to slide the window and preset the shape of the anchor box. To solve the problem, we adopt a guided anchoring method to directly predict the shapes of the anchor boxes according to the feature maps. In specific, the width and height of the anchor boxes are predicted by an independent subnetwork. The adopted subnetwork is a simple single-layer convolutional layer. Its output channel number is 2, representing the predictions of width and height, respectively. The value of each location in the output feature map is the predicted width and height of the anchor box in the corresponding position.

As with the proposed box, the shape prediction of the anchor box also depends on the regression. The most straightforward idea is to directly predict the width and height of the anchor box through the network. However, the scales and aspect ratios of the objects have a large variation range, but the value ranges of different locations are close due to the shared parameters in the convolutional layer. Therefore, it is hard to directly regress the original widths and heights of the anchor boxes. It should be noted that the proposed method is based on feature pyramid network (FPN), which produces multiple feature levels. The lower-level feature map has higher resolution and can be used to detect small objects, while the coarser-resolution feature map is used to detect large objects. To avoid the direct regression, different regression ranges can be set for different feature levels. Specifically, the guided anchoring method adopts a nonlinear mapping function to transform the width and height into a narrower range. The nonlinear function is:

$$\begin{aligned} dw &= \ln(\frac{w}{\sigma s}), \\ dh &= \ln(\frac{h}{\sigma s}), \end{aligned} \tag{3}$$

where s is the stride of the feature map relative to the original image, σ is the preset scale factor, w and h are the original width and height of the ground truth box. The diagram of the function is shown in Figure 4, where the scale factor is 8 and the strides are 4, 8, 16, 32 and 64.

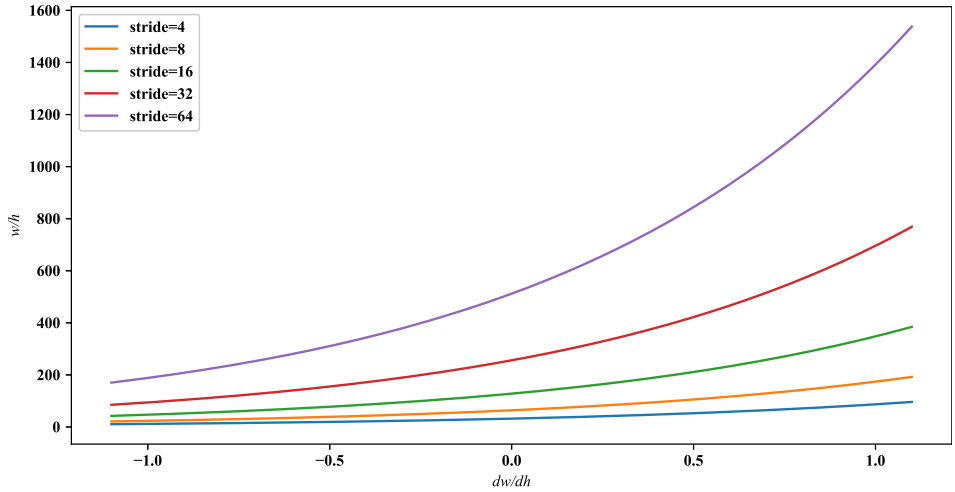

Figure 4. The illustration of the nonlinear function mapping.

As shown in Figure 4, by selecting the appropriate scale factor and stride, dw and dh in the range of $[-1, 1]$ can be converted to any value in the range of $[12, 1391]$. dw and dh as the regression target can significantly reduce the difficulty of regression. The guided anchoring method uses a subnetwork to predict dw and dh, which are then converted to w and h according to the inverse transformation of Equation (3).

3.4. Modulated Feature Adaptation Module

Due to the adoption of guided anchoring, the shapes of the anchors are different for each position. The different shapes result in that the receptive field of the convolutional layer in subsequent steps does not match well with the anchors in different locations. In order to solve this problem, Wang et al. [37] tried to use a feature adaptation module to transform the feature maps extracted from the backbone network based on the shapes of the predicted anchors. The feature adaptation module enhances the perception capability of the convolutional kernel in different spatial positions by adjusting offsets in perceiving input features. To further strengthen this capability, we propose to use a modulated feature adaptation module to regulate the amplitudes of different spatial locations in the input feature map.

The core of the proposed MFAM is modulated deformable convolutional layer [17] of 3×3 kernel size, and the calculation formula of modulated deformable convolution is:

$$\mathbf{y}(p_0) = \sum_{p_n \in \mathcal{R}} \mathbf{k}(p_n) \cdot \mathbf{x}(p_0 + p_n + \Delta p_n) \cdot \Delta m_n, \tag{4}$$

where $\mathbf{x}$ is the input feature map, $\mathbf{y}$ is the output feature map, $\mathbf{k}$ is the convolutional kernel, and p_0 demotes the position to be calculated in the output feature map, $\mathbf{p}_n$ means the n-th position of the convolutional kernel, $\mathcal{R}$ denotes the set of the positions in the convolutional kernel. Take the convolutional kernel of 3×3 as an example, $\mathcal{R} = \{(-1,-1), (-1,0), \cdots, (0,1), (1,1)\}$. $(p_0 + p_n)$ is the position of the input feature map corresponding to the n-th position of the convolutional kernel in the conventional convolutional layer.

As shown in Equation (4), the deformable convolution adds Δp_n and Δm_n, which denotes the offset and the modulation scalar of the n-th position in the convolutional kernel. Compared with the conventional convolution, the offset of the modulated deformable convolution can adjust the perceivable positions in input feature maps, and the modulation is applied to change the amplitudes

of different spatial locations. Therefore, modulated deformable convolution has the capability to manipulate spatial support regions.

In [37], the offset is obtained by convolutional layer $\mathcal{C}_1$, the input of which is the predicted width and height. In this way, the feature adaptation module transforms the feature map using the generated offset to better match the shape of the anchor. Similar to the offset, the modulation scalar in the proposed method is also obtained through the other convolutional layer $\mathcal{C}_2$. In general, the input of $\mathcal{C}_2$ is the original feature map extracted from the backbone network. In the proposed method, the purpose of using the MFAM is to make the transformed feature maps more suitable for the various shapes of the anchor boxes. Therefore, it is a better choice to use the shape maps to predict the modulation scalar. Therefore, we also consider to directly use the shape maps as the input of $\mathcal{C}_2$. In addition, because the required weight has a value range of $[0, 1]$, the generated modulation scalar needs to be transformed by a sigmoid function.

It is worth noting that the subnetwork to predict the shape map is not trained by the loss from the MFAM. That is, loss only adjusts the modulated deformable convolutional layer and two convolutional layers $\mathcal{C}_1$ and $\mathcal{C}_2$ in the backward propagation. The diagram of the MFAM is shown in Figure 5.

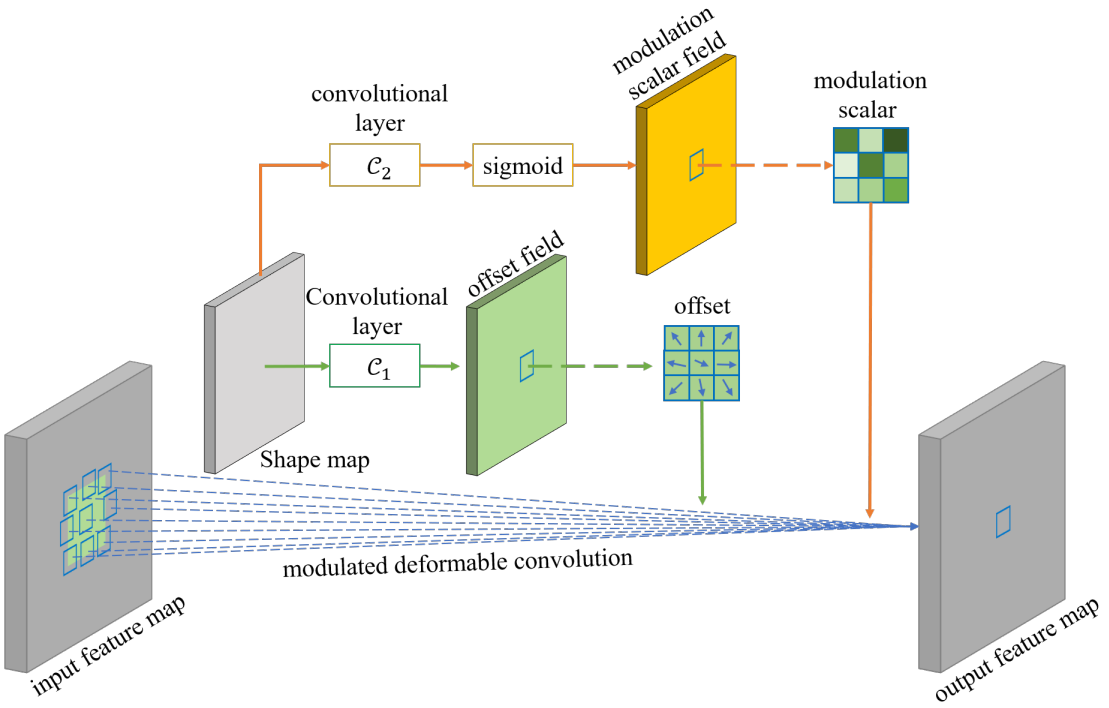

Figure 5. The diagram of modulated feature adaptation in the proposed method.

4. Experiments and Results

In order to verify the effectiveness of the proposed method, we conducted several experiments on the DIOR dataset [38]. The experimental results show that the proposed method has a significant improvement in the object detection task for remote sensing images compared with the conventional methods. All the experiments are conducted on a computer with a central processing unit (CPU) of Intel 6700K, a graphics processing unit (GPU) of NVIDIA GTX 1080Ti, and random access memory (RAM) of 32 GB. All experiments are implemented based on the open source code base mmdetection [39].

4.1. Data Preparation

The DIOR dataset consists of a large number of remote sensing images. The dataset includes 20 classes of objects, which are airplane, airport, baseball field, basketball court, bridge, chimney, dam,

expressway service area, expressway toll station, golf course, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle and windmill. There are 23,463 optical remote sensing images in total, and their sizes are all 800 × 800 pixels. The spatial resolutions of the images are from 0.5 to 30 m.

The DIOR dataset is randomly divided into two subsets, namely training set and testing set, to achieve more consistent distribution between two subsets. In which, the training and testing sets contain 11,725 and 11,738 images, respectively. The labeled bounding boxes in the training set are used to train the whole network. In the inference, the test set is used to evaluate the effectiveness of the trained network.

4.2. Evaluation Metrics

In order to evaluate the performance of the detectors, precision-recall (PR) curve and average precision (AP) are adopted as the evaluation metric.

In general, there are three types of detection results, including true positive (TP), false positive (FP) and false negative (FN). The IoU threshold is used to determine the specific type of the detected bounding box. TP usually refers to the result that the IoU between the predicted box and the ground truth box is greater than the preset threshold. FP is just the opposite. FN means the missed objects that are predicted as background. The IoU threshold is set as 0.5 in our experiments.

According to these three types of detection results, we can further obtain the precision and recall metrics. The former metric is used to measure the proportion of correct predictions in all predicted results, while the latter one can evaluate the proportion of detected objects in all positives. The precision $\mathcal{P}$ and recall $\mathcal{R}$ are defined as:

$$\begin{aligned} \mathcal{P} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}, \\ \mathcal{R} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}. \end{aligned} \tag{5}$$

The PR curve reflects the corresponding relationship between precision and recall, namely the precisions under different recalls. High precision means a low FP, and high recall means a low FN. A better method should have a higher precision under the same recall than the other methods.

The AP computes the average value of $\mathcal{P}$ over the interval from $\mathcal{R} = 0$ to $\mathcal{R} = 1$. Therefore, it can be regarded as the area under the PR curve, and an ideal detector should have a high area and vice versa. Compared with the PR curve, AP more directly reflects the overall performance of the detector, and it is commonly used for the evaluation of object detection results. In addition, because the objects have multiple classes in the experiments, the mean AP (mAP), which averages the APs of different classes, is generally used to evaluate the detection performance on multiple classes.

4.3. Implementation Details

ResNet [40] has been proved to be an effective backbone network in many tasks of object detection. ResNet uses the residual learning framework to address the degradation problem of the deep network, thereby easing the training of the network. In order to reduce the computation required by the backbone network, we adopt ResNet-50 as the backbone network in our experiments.

The overall framework is based on the classic FPN [41]. FPN uses the bottom-up pathway to calculate the feature hierarchy, which is composed of feature maps of multiple scales. Then the top-down pathway is used to obtain high-resolution feature maps with stronger semantic information. The feature maps of the same size, which are from the bottom-up pathway and the top-down pathway, respectively, are merged by lateral connection. Finally, the merged feature maps pass through a convolutional layer with a convolutional kernel of 3 × 3 to obtain the final feature maps. In our experiments, the final feature maps have five scales, and the strides of the scales are 4, 8, 16, 32, and 64,

respectively. The feature maps are used for subsequent steps, such as the generation of proposal boxes, the classification and localization of RoIs.

In the guided anchoring method, it is impossible to calculate the IoUs between the anchor boxes and the ground truth boxes since there are no preset shapes. Therefore, a total of nine predetermined shapes of anchor boxes are taken to calculate the IoUs to obtain the suitable ground truth boxes in the experiments. The scales of the predetermined shapes are $\{2^0, 2^{1/3}, 2^{2/3}\}$ and the aspect ratios are $\{0.5, 1.0, 2.0\}$. The shapes are the combinations of all the scales and aspect ratios. RoI pooling uses the RoIAlign [42] with an output size of 7×7. The whole network adopts the stochastic gradient descent algorithm with momentum, with the learning rate of 0.002 and the momentum of 0.9. Due to the fact that the generated proposal boxes are usually redundant, one object may correspond to multiple predicted neighborhood boxes in the testing. Therefore, we adopt the classic non-maximum suppression (NMS) algorithm to filter the similar predicted boxes.

4.4. Comparison with the State-of-the-Art

To evaluate the detection performance of the proposed method quantitatively, we compared the experimental results to several representative deep-learning-based methods that are widely used for object detection in natural images and remote sensing images. The selected methods include one-stage and two-stage methods. Specifically, the compared methods include FPN, RetinaNet [13], fully convolutional one-stage detector (FCOS) [43], FoveaBox [44], modulated deformable convolutional network (MDCN) [17], CornerNet [32], SSD and YOLOv3. In which, the results of CornerNet, SSD and YOLOv3 are coming from [38]. Since their relevant data of PR curves are not provided in [38], our experimental results only list their AP values. Their detailed descriptions are as follows.

1. FPN. As the foundation of the proposed method, FPN is the baseline for comparison. Therefore, except for the improved parts, all the settings of FPN are the same as the proposed method. As for the generation of the proposal boxes, RPN is adopted in FPN.
2. RetinaNet. RetinaNet is a one-stage method based on FPN, and the backbone network is ResNet-50. The backbone network attaches two subnetworks, which are used to classify and regress anchor boxes, respectively. The anchor scales are $\{2^0, 2^{1/3}, 2^{2/3}\}$ and the aspect ratios are $\{0.5, 1.0, 2.0\}$. The main contribution of RetinaNet is that focal loss is proposed to ease the imbalance between positive and negative samples.
3. FCOS. FCOS is an anchor-free method, which directly regresses the target bounding box for each location. FCOS detects different sizes of objects on different levels of feature maps following FPN. In addition, it also adds a single-layer branch to predict the center-ness of the location to suppress the low-quality detected bounding boxes. The backbone network is ResNet-50.
4. FoveaBox. FoveaBox is also an anchor-free method. It directly learns the objects by predicting category-sensitive semantic maps for the object existing possibility and producing a category-agnostic bounding box for each position that potentially contains an object. FoveaBox divides the scales of objects into several bins according to the number of feature pyramid levels, and every feature pyramid learns to be responsive to objects of particular scales.
5. MDCN. Modulated deformable convolutional layer is an important component of the proposed method, and plays a key role in the transformation of the feature map. For this reason, we use MDCN [17], which proposes a modulated deformable convolutional layer as one of the comparison methods. It is worth noting that although modulated deformable convolutional layers are used in both the proposed method and MDCN, their locations and purposes are different. MDCN mainly applies the deformable convolution in the backbone network to make the extracted feature focus on pertinent image regions. However, the proposed method adopts the deformable convolution in the guided anchoring block to make the feature match better with the generated anchor boxes.
6. SSD. SSD is a one-stage method, which extracts the feature maps of different scales from the images for detection. The proposal boxes are obtained by densely sampling on the feature map.

There are multiple anchor boxes of different scales and aspect ratios in each sampled position. The features of the anchor boxes are classified and regressed by CNN. In this experiment, SSD uses VGG-16 [45] as the backbone network.

7. YOLOv3. YOLOv3 is an improved version of YOLO and it has a higher mAP than the original one. It predicts the width and height of the box as offsets from cluster centroids, and then predicts the center coordinates of the box relative to the location of filter application using a sigmoid function. After obtaining the predicted boxes, YOLOv3 also predicts the corresponding classes for the bounding boxes by a softmax classifier. YOLOv3 uses Darknet-53 as the backbone network.
8. CornerNet. CornerNet detects an object bounding box as a pair of key points, namely the top-left corner and the bottom-right corner. This method does not need to design anchor boxes commonly used in deep-learning-based object detection methods, but it is required to group the corners based on the predicted embedding vectors.

Table 1 lists the APs of all methods on different categories, and Figure 6 shows the PR curves of different categories. As we can see from Table 1 and Figure 6, the proposed method is obviously better than the other methods and has the highest mAP of 73.6%. In terms of specific categories, the AP values of the proposed method are the highest in 9 of the 20 categories, the second highest in seven categories and the third highest in three categories. In other words, except for the category of the chimney, the proposed method has achieved the top three results compared with the other methods.

Table 1. The mean average precisions (mAPs) of different methods, where BC means the Basketball Court, ESA denotes expressway service area, ETS denotes expressway toll station and GTF is ground track field. The red, orange and yellow numbers indicate that the AP values of the corresponding methods are the highest, the second highest and the third highest in this category.

Categories	FPN	RetinaNet	FCOS	FoveaBox	MDCN	CornerNet	SSD	YOLOv3	Ours
Airplane	60.7	66.1	61.1	66.9	61.1	58.8	59.5	72.2	70.5
Airport	77.0	78.4	82.6	79.6	81.2	84.2	72.7	29.2	81.9
Baseball Field	74.9	74.4	76.6	76.7	75.3	72.0	72.4	74	76.5
BC	87.9	87.8	87.6	87.6	88.6	80.8	75.7	78.6	89.3
Bridge	46.4	37.2	42.8	42.7	48.5	46.4	29.7	31.2	49.0
Chimney	80.2	80.2	80.6	79.8	79.6	75.3	65.8	69.7	79.5
Dam	60.0	63.1	64.1	60.6	70.1	64.3	56.6	26.9	66.0
ESA	76.7	76.7	79.1	81.1	81.1	81.6	63.5	48.6	85.2
ETS	70.6	58.8	67.2	66.4	74.4	76.3	53.1	54.4	71.9
Golf Course	77.3	79.9	82.0	74.5	77.4	79.5	65.3	31.1	81.2
GTF	83.2	76.4	79.6	80.3	84.0	79.5	68.6	61.1	83.3
Harbor	46.0	36.6	46.4	50.0	51.0	26.1	49.4	44.9	52.8
Overpass	60.0	56.6	57.8	57.6	62.7	60.6	48.1	49.7	62.2
Ship	75.1	68.7	72.1	73.1	75.1	37.6	59.2	87.4	77.1
Stadium	67.4	69.0	64.8	71.5	67.0	70.7	61.0	70.6	76.0
Storage Tank	61.1	43.1	63.4	60.0	63.5	45.2	46.6	68.7	72.4
Tennis Court	87.3	85.7	85.2	86.9	86.8	84.0	76.3	87.3	87.7
Train Station	58.8	58.4	62.8	54.5	65.2	57.1	55.1	29.4	64.1
Vehicle	45.0	40.7	43.8	42.7	46.1	43.0	27.4	48.3	55.0
Windmill	87.9	85.1	87.5	88.3	88.7	75.9	65.7	78.7	90.3
mAP	69.2	66.1	69.4	69.0	71.4	64.9	58.6	57.1	73.6

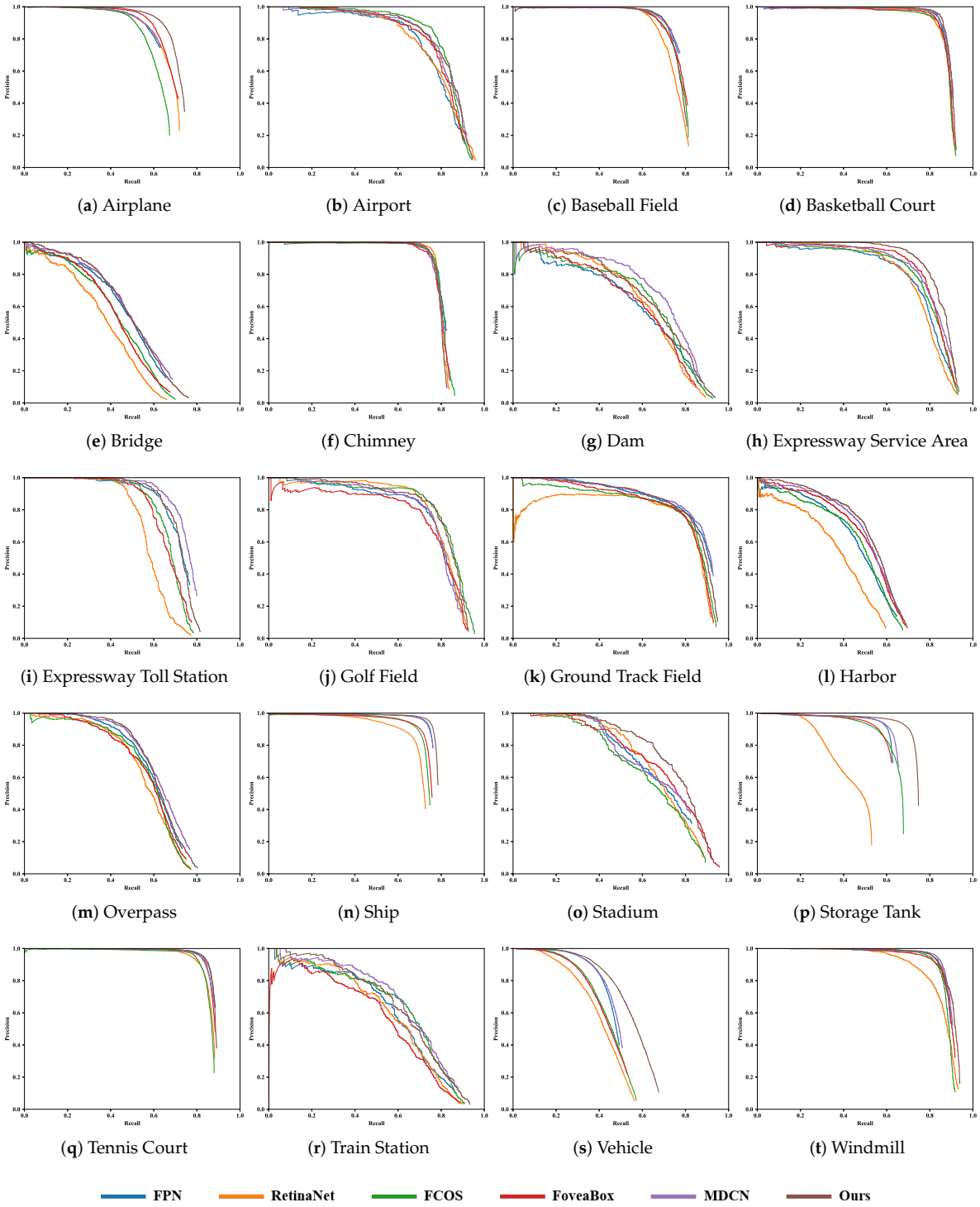

Figure 6. The precision-recall (PR) curves of different methods.

Several samples of the detection results obtained by the proposed methods are shown in Figure 7. The green rectangles denote the predicted results.

(**a**)Ship and harbor (**b**)Storage tank (**c**)Bridge

(**d**)Vehicle (**e**)Basketball, baseball and tennis court (**f**)Airplane

Figure 7. The samples of detection results of the proposed methods.

5. Discussion

In the proposed method, our contributions mainly include the following aspects. First, we adopt the attention mechanism in the proposed method to make the network pay more attention to the object areas rather than the background areas. Then, the generated attention map is used to predict the positions of the objects. Specifically, it is incorporated into the guided anchoring method to produce the anchor boxes, which are the preliminary boxes for object detection. Finally, we replace the deformable convolutional layer used in the guided anchoring method with the modulated deformable convolutional layer. Compared with the former, the latter adds a modulation scalar to adjust the amplitudes of different spatial locations in the input feature map.

In order to better analyze the impact of different aspects in the proposed method on the final detection performance. We progressively study the proposed method. In specific, the method will be studied from three parts: attention mechanism, guided anchoring and deformable convolution. The detailed results are shown in Table 2.

Table 2. The influence of different aspects on the detection results, where DC means deformable convolution, MDC denotes modulated deformable convolution.

Method	Attention Mechanism	The Prediction Way for Anchor Position	Deformable Convolution	mAP (%)
FPN	-	Sliding Window	-	69.1 (baseline)
FPN with Attention	√	Sliding Window	-	71.9 (+2.8)
Guided Anchoring	-	Independent Subnetwork	DC	72.1 (+3.0)
Ours without MDC	√	Attention Map	DC	73.3 (+4.2)
Ours	√	Attention Map	MDC	73.6 (+4.5)

In Table 2, the first method is the basic FPN, which is the baseline. FPN with Attention adds an attention mechanism compared to basic FPN. It can be seen that the addition of the attention mechanism increases the mAP of the network by 2.8%. This indicates that the attention mechanism, as the way to enhance the feature maps, can improve the performance of object detection. The third method is the guided anchoring method in [37], and the 3.0% improvement shows that the adaptive anchor boxes can generate better predicted boxes to improve the detection performance, thus improving the detection quality. In the fourth one, we directly use the attention map generated by the attention mechanism to replace the position map in the guided anchoring method to produce the locations of anchor boxes, but modulated deformable convolution is not used in this method. Compared with the guided anchoring method, the fourth method shows a 1.2% improvement in terms of the mAP value. This result means that the attention map can better represent the object location, so as to improve the final detection results. The final method is the proposed method, and it replaces the deformable convolution in the fourth method with modulated deformable convolution, which brings an mAP increase of about 0.3%. This shows that the modulation scalar of modulation deformable convolution plays a certain role in improving the detection performance.

The mAP of the proposed method finally increases to 73.6%, and it is increased by 4.5% compared with the baseline FPN. It can be seen from Table 1 that the accuracy of the proposed method is higher than that of the baseline method in almost every category, which fully indicates the superiority of the proposed method.

As can be seen from Figure 7, the proposed method can detect the vast majority of objects. It also has a good performance on the objects with large size differences in Figure 7a,c. In Figure 7e, the background is complex, but the proposed method can still detect the vehicles on the side of the house very well, which shows that the proposed method is robust to a certain extent.

However, we also find that there are still missed objects and repeated bounding boxes in the detection results. We guess the reason is that the proposed method predicts the positions of anchor boxes through the attention map instead of the sliding window. It cannot ensure that all object regions are completely covered, and causes the undetected objects. In future research, we will try to take measures to alleviate the problem. To be specific, the following aspects of work can be continued. Firstly, the attention mechanism includes many forms. The proposed method mainly focuses on the query-independent attention mechanism. Whether the other attention mechanism can predict more accurate positions is worthy of study. Secondly, a fixed threshold is applied to determine whether the position is the object in the proposed method, and missed or duplicated boxes may occur. A better way to determine the position should be researched.

6. Conclusions

In this paper, we propose an object detection method that incorporates the attention mechanism, which is used not only to enhance the features, but also to generate adaptive anchor boxes. First of all, we apply the attention mechanism to make the network pay more attention to the object regions and reduce the influence of the background on detection. Secondly, we regard the attention map as the distribution probability map of the anchor box positions. The adaptive anchor boxes are obtained by combining the predicted positions and the shapes generated by the guided anchoring method. The obtained anchor boxes are diverse and conform to the positions and shapes of the objects. Finally, we use the MFAM to transform the feature map. By adjusting the calculation position and amplitude of the convolutional kernel, the transformed feature maps are more suitable for anchor boxes. The quantitative comparison results on the public DIOR dataset demonstrate the superiority of the proposed method over several state-of-the-art methods. On the one hand, this superiority comes from the fact that the attention mechanism can enhance the extracted features and make it more focused on the object regions. On the other hand, the attention map can reflect the positions of objects, thereby generating more suitable anchor boxes by combining with the guided anchoring method.

In our future work, the global attention map will be further investigated to detect and generate more deformable bounding boxes such as rotated boxes and polygons.

Author Contributions: Conceptualization, Z.T.; methodology, Z.T. and R.Z.; software, Z.T.; validation, Z.T., Z.H. and J.H.; formal analysis, W.W.; investigation, R.Z.; writing—original draft preparation, Z.T.; writing—review and editing, W.W.; supervision, Z.Z.; project administration, R.Z.; funding acquisition, R.Z. and W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Nos. 61471370 and 61901500).

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

References

1. Shi, Z.; Yu, X.; Jiang, Z.; Li, B. Ship Detection in High-Resolution Optical Imagery Based on Anomaly Detector and Local Shape Feature. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4511–4523.
2. Zhang, W.; Sun, X.; Wang, H.; Fu, K. A generic discriminative part-based model for geospatial object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2015**, *99*, 30–44. [CrossRef]
3. Bai, X.; Zhang, H.; Zhou, J. VHR Object Detection Based on Structural Feature Extraction and Query Expansion. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6508–6520.
4. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Scalable multi-class geospatial object detection in high-spatial-resolution remote sensing images. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 2479–2482.
5. 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]
6. 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]
7. 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 Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
8. Li, S.; Xu, Y.; Zhu, M.; Ma, S.; Tang, H. Remote Sensing Airport Detection Based on End-to-End Deep Transferable Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1640–1644. [CrossRef]
9. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
10. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
11. Joseph, R.; Ali, F. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:cs.CV/1804.02767.
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
13. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
15. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 91–99.

17. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. *arXiv* **2018**, arXiv:cs.CV/1811.11168.
18. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
19. Grabner, H.; Nguyen, T.T.; Gruber, B.; Bischof, H. On-line boosting-based car detection from aerial images. *ISPRS J. Photogramm. Remote Sens.* **2008**, *63*, 382–396. [CrossRef]
20. Yang, J.; Yu, P.; Kuo, B. A Nonparametric Feature Extraction and Its Application to Nearest Neighbor Classification for Hyperspectral Image Data. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 1279–1293. [CrossRef]
21. Lang, F.; Yang, J.; Yan, S.; Qin, F. Superpixel Segmentation of Polarimetric Synthetic Aperture Radar (SAR) Images Based on Generalized Mean Shift. *Remote Sens.* **2018**, *10*, 1592. [CrossRef]
22. Ciecholewski, M. River Channel Segmentation in Polarimetric SAR Images. *Expert Syst. Appl.* **2017**, *82*, 196–215. [CrossRef]
23. Braga, A.M.; Marques, R.C.P.; Rodrigues, F.A.A.; Medeiros, F.N.S. A Median Regularized Level Set for Hierarchical Segmentation of SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1171–1175. [CrossRef]
24. Jin, R.; Yin, J.; Zhou, W.; Yang, J. Level Set Segmentation Algorithm for High-Resolution Polarimetric SAR Images Based on a Heterogeneous Clutter Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 4565–4579. [CrossRef]
25. Pang, J.; Li, C.; Shi, J.; Xu, Z.; Feng, H. $\mathcal{R}^2$ -CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5512–5524. [CrossRef]
26. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803. [CrossRef]
27. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-Cross Attention for Semantic Segmentation. *arXiv* **2018**, arXiv:cs.CV/1811.11721v1.
28. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. *arXiv* **2019**, arXiv:cs.CV/1904.11492.
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
30. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv* **2013**, arXiv:cs.CV/1312.6229v4.
31. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
32. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 765–781.
33. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. *arXiv* **2019**, arXiv:cs.CV/1904.08189v3.
34. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-Up Object Detection by Grouping Extreme and Center Points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
35. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An Empirical Study of Spatial Attention Mechanisms in Deep Networks. *arXiv* **2019**, arXiv:cs.CV/1904.05873.
36. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:stat.ML/1607.06450.
37. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region Proposal by Guided Anchoring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
38. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
39. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:cs.CV/1906.07155.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]

41. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
42. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
43. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. Available online: https://arxiv.org/abs/1904.01355v5 (accessed on 2 April 2019).
44. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Shi, J. FoveaBox: Beyond Anchor-based Object Detector. Available online: https://arxiv.org/abs/1904.03797v1 (accessed on 8 April 2019).
45. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition 2014. Available online: https://arxiv.org/abs/1409.1556 (accessed on 4 September 2014).