remote sensing

MDPI

Article

Object Detection in Remote Sensing Images Based on Adaptive Multi-Scale Feature Fusion Method

Chun Liu, Sixuan Zhang, Mengjie Hu and Qing Song *

Pattern Recognition and Intelligent Vision (PRIV), Beijing University of Posts and Telecommunications, Beijing 100876, China; chun.liu@bupt.edu.cn (C.L.); zhangsixuan@bupt.edu.cn (S.Z.); mengjie.hu@bupt.edu.cn (M.H.)
* Correspondence: priv@bupt.edu.cn

Abstract: Multi-scale object detection is critical for analyzing remote sensing images. Traditional feature pyramid networks, which are aimed at accommodating objects of varying sizes through multi-level feature extraction, face significant challenges due to the diverse scale variations present in remote sensing images. This situation often forces single-level features to span a broad spectrum of object sizes, complicating accurate localization and classification. To tackle these challenges, this paper proposes an innovative algorithm that incorporates an adaptive multi-scale feature enhancement and fusion module (ASEM), which enhances remote sensing image object detection through sophisticated multi-scale feature fusion. Our method begins by employing a feature pyramid to gather coarse multi-scale features. Subsequently, it integrates a fine-grained feature extraction module at each level, utilizing atrous convolutions with varied dilation rates to refine multi-scale features, which markedly improves the information capture from widely varied object scales. Furthermore, an adaptive enhancement module is applied to the features of each level by employing an attention mechanism for feature fusion. This strategy concentrates on the features of critical scale, which significantly enhance the effectiveness of capturing essential feature information. Compared with the baseline method, namely, Rotated FasterRCNN, our method achieved an mAP of 74.21% (0.81%) on the DOTA-v1.0 dataset and an mAP of 84.90% (+9.2%) on the HRSC2016 dataset. These results validated the effectiveness and practicality of our method and demonstrated its significant application value in multi-scale remote sensing object detection tasks.

Keywords: feature fusion; remote sensing; object detection; attention mechanism

check for updates

Citation: Liu, C.; Zhang, S.; Hu, M.; Song, Q. Object Detection in Remote Sensing Images Based on Adaptive Multi-Scale Feature Fusion Method. *Remote Sens.* **2024**, *16*, 907. https://doi.org/10.3390/rs16050907

Academic Editor: Hossein M. Rizeei

Received: 16 January 2024
Revised: 26 February 2024
Accepted: 28 February 2024
Published: 4 March 2024

1. Introduction

With the rapid development of remote sensing technology, object detection in remote sensing images has emerged as a burgeoning research area in computer vision. Various studies have focused on utilizing deep-learning-based object detection methods in the domain of remote sensing [1–6]. However, detecting targets in these images has shown itself to be challenging due to the objects' varying scales and resolutions. Consequently, multi-scale object detection has become a key research focus in remote sensing image analysis. It aims to improve the accuracy and efficiency of object detection methods by integrally processing images across multiple scales.

Deep-learning-based object detection methods can be categorized based on two primary criteria. Methods are divided into two-stage and single-stage detection, depending on whether the extraction of regions of interest (RoIs) is required. On the other hand, methods are classified into anchor-based detection and anchor-free detection, depending on the necessity for predefined anchor boxes. Similarly, deep-learning-based methods for detecting rotated objects in remote sensing images also follow these classification criteria.

Anchor-based methods utilize many anchor boxes with different-sized ratios to localize objects in images, subsequently generating candidate boxes. In two-stage anchor-based methods, such as SCRDet [7] and Oriented RCNN [8], region candidate boxes are initially produced via a region proposal network (RPN), followed by classification and regression on these boxes. SCRDet [7] is a novel multi-category rotation detector for small, cluttered, and rotated objects. Oriented RCNN [8] comprises an oriented region proposal network (oriented RPN) and introduces an oriented RCNN head for refining oriented regions of interest (oriented RoIs) and recognizing them. Compared with two-stage methods, single-stage anchor-based methods eliminate the step of generating proposals, instead directly performing regression and classification on the feature maps. Recently, significant research has been conducted on anchor-free detection methods. These approaches eschew the use of manually designed anchor boxes, thus typically offering faster performance and enhanced detection capabilities for objects with larger aspect ratios, such as an object-wise point-guided localization detector (OPLD) [9] and a fully convolutional one-stage object detector (FCOS) [10].

Advancements in object detection have significantly propelled progress in remote sensing. However, the direct application of algorithms intended for natural images to the remote sensing field encounters notable challenges. The challenge arises from remote sensing images often exhibiting significant scale variation between objects, which result from resolution differences and inherent size disparities across categories. Several effective methods were proposed to tackle the challenge, which have significantly advanced the object detection field. For example, a feature pyramid network (FPN) [11] achieves feature fusion by establishing cross-scale connections between different network layers, which can effectively solve the problem of object scale changes in object detection. Pyramid convolution [12] uses techniques such as multi-scale feature fusion in the network to further enhance the performance of object detection. In addition, InceptionNets [13,14] can improve the network's ability to extract multi-scale features using convolution kernels of different sizes. The global-to-local scale-aware detection network (GLSANet) [5] excavates and enhances the high-level semantic information within deep feature maps, optimizes the feature pyramid network, and boosts the performance of multi-scale detection in remote sensing images.

Motivations: We were inspired by the coarse and fine matching processes in image matching [15], which provided insights into handling features within network hierarchies. Our method fuses features across layers of different levels and explores detailed features within each level. Furthermore, inspired by the concept of state merging in automata theory, this paper proposes an innovative method to address the challenge of detecting objects at various scales. We aimed to treat objects of the same category at different scales in the feature map as equivalent states, allowing for some degree of state merging during feature extraction to improve the detection robustness. Combining the two aspects above, our proposed method achieved more accurate and robust object detection in remote sensing.

In this paper, in addition to dealing with the cross-scale feature fusion between different network layers of the backbone network, we also focus on the adaptive enhancement and extraction of multi-scale features at the same level. We hope the extracted multi-scale features are scale invariant within the same hierarchical level. What is scale invariance? It refers to the ability to extract identical features from the same object, regardless of size.

To achieve this goal, this paper introduces multiple shared parameter atrous convolutions to extract and fuse contextual features after the backbone network. Subsequently, an attention mechanism was applied for adaptive feature enhancement. This methodology aimed to optimize the extraction and utilization of contextual information, thereby improving the accuracy and effectiveness of object detection.

The contributions of this article are summarized as follows:

1. We propose a novel multi-scale feature extraction method, which includes both inter-level and intra-level feature extraction and fusion.
2. Within the hierarchy, the extracted features are scale-invariant, and objects of different scales in the feature map are expected to achieve a uniform feature representation.

3. A scale selection mechanism is introduced, which discriminates between the multi-scale features extracted within the hierarchy. This method assigns greater weight to features from scales that are most critical and suitable for subsequent tasks, effectively focusing on the most relevant aspects for enhanced performance.
4. In the experiment, our proposed method achieved a 74.21% mAP result on the DOTA-v1.0 dataset and a 84.90% mAP result on the HRSC2016 dataset, which demonstrate the effectiveness and superiority of the improvement.

Section 2 introduces the related work involved in this study, including research progress of object detection in remote sensing images, multi-scale feature extraction, and the feature attention mechanism. Section 3 explains the adaptive enhancement method we propose, including network architecture and feature extraction fusion. Section 4 presents the dataset, implementation details, and results. Section 5 analyzes and discusses the experimental results. Section 6 summarizes the entire paper and looks forward to future research directions.

2. Related Work

2.1. Object Detection in Remote Sensing Image

Recent research revealed that rotating object detection possesses reliable analytical capabilities for objects in remote sensing images. Various studies have focused on utilizing deep-learning-based object detection methods in remote sensing. In addition to classification and localization, object detection in this domain also necessitates predicting the orientation of the actual detection boxes. To solve the problem of predicting the target direction, an RRPN (rotation region proposal network) [16] directly generates some rotated anchor boxes by sliding on the input image, subsequently producing rotated proposals for target localization and classification. However, this method generates a substantial number of anchor boxes, with as many as 54 directional anchor boxes at a single position. This leads to an abundance of negative samples and entails considerable computational expense. To reduce the large number of rotated anchor boxes and the mismatch between features and objects, RoI Transformer [17] learns the oriented RoI from the horizontal RoI generated by an RPN, greatly improving the detection accuracy. However, incorporating fully connected layers and RoI alignment operations in this method also substantially increases the computational cost. Yang et al. [7] proposed SCRDet for rotated targets, which is based on FasterRCNN [18], which is an improved version of RCNN (regions with CNN features) [19]. Xu et al. [20] proposed a gliding vertex method for rotated object detection by learning the four vertex sliding offsets on the FasterRCNN head regression branch. However, both methods adopt horizontal RoI for classification and rotated bounding box regression, which still results in insufficient matching between objects and features. Circular smooth labels (CSLs) [21] and densely coded labels (DCLs) [22] no longer regard the angle prediction of the rotated boxes as a regression task but as a classification task, essentially avoiding the problem of discontinuous boundary regression. Different from the bounding box regression method, Oriented RepPoints [23] provides an adaptive point learning method that utilizes adaptive point representations to capture geometric information of instances in any orientation accurately.

To address the issue of large target scale changes in remote sensing images, Deng et al. [24] proposed a CNN-based multi-scale object proposal network (MS-OPN) method. This method generates pre-selected boxes on feature maps of different depths and then sends these candidate areas to an accurate object detection network (AODN) for classification and regression. Objects with different scales are mapped by feature maps of varying depths. Combining feature maps of different depths retains more details and provides better results.

2.2. Extraction and Fusion of Multi-Scale Features

Multi-scale feature extraction and fusion are common methods used to improve the performance of detection algorithms. Due to the variability in scales and sizes of targets, relying solely on single-scale feature extraction may lead to the inadequate capture of both global and local details of the target. Employing multi-scale approaches addresses this limitation by comprehensively analyzing features at various scales.

To solve this problem, researchers have proposed various multi-scale feature extraction methods. The usual approach involves extracting feature maps of different depths using CNNs and performing fusion operations on them. This process combines shallow-layer texture information with deep-layer semantic information. Objects predicted from feature maps of different scales, when mapped back to the original image, correspond to targets of varying sizes. An FPN [11] introduces a top-down path to merge multi-scale features. A path aggregation network (PANet) [25] extends this method by adding an additional bottom-up path aggregation network on top of an FPN. YOLOv4 [26] slightly modifies the feature fusion method based on a PANet using concatenation instead of element-wise addition. NAS-FPN [27] adopts a neural architecture search to automatically design the feature network topology. EfficientDet [28] achieves multi-scale feature extraction through a bidirectional feature pyramid network (BiFPN), which produces better performance and efficiency. An informative feature pyramid network (Info-FPN) [6] was proposed to address channel information loss, feature misalignment, and aliasing effects.

To improve the performance of object detection tasks, many studies have considered multi-scale feature fusion. This is because images contain objects of different scales and sizes, and feature maps have distinct key features at multiple scales. By utilizing a multi-scale feature fusion approach, it is possible to effectively capture the characteristic information of objects at various scales, thereby enhancing the accuracy and robustness of object detection.

2.3. Attention Mechanism

Attention mechanisms have been widely used and studied in object detection tasks. By introducing an attention mechanism, a model can allocate different weights to each position or feature channel, effectively focusing on target-related regions and features in the image, thereby improving the accuracy and robustness of the object detection.

Spatial transformer networks [29] utilize a spatial attention mechanism to determine the importance of each position by calculating the relationship between features at different positions in the image. This can help the model to better focus on the area where the target is located and avoid background interference. Later, an SENet (squeeze-and-excitation network) [30] utilizes a channel attention mechanism, which weights different channels in the input feature map to better focus on important channels and suppress unimportant channels, thereby better capturing the characteristics and contextual information of the target. A CBAM [31] (convolutional block attention module) is an attention mechanism that combines spatial attention and channel attention to simultaneously focus on the spatial dimensions and channel dimensions of the image, thereby more comprehensively capturing important information in the image.

In addition, there is also a branch attention mechanism. For example, in an SKNet (selective kernel network) [32], different sizes of convolution kernels are selectively applied to dynamically adjust the receptive field size. This captures features at different scales in various branches, which are then fused through a weighted attention mechanism. This branch attention mechanism effectively processes multi-scale information, overcoming the limitations of traditional fixed-scale convolutions. It offers a more flexible and detailed approach to capturing multi-scale information in images.

3. Methods

3.1. The Overall Network Structure

Figure 1 delineates the network architecture described in this paper. This architecture comprises a suite of modules, each with a specialized function. The backbone network serves as the foundational structure from which feature extraction commences. The innovative adaptive multi-scale feature enhancement and fusion module (ASEM) is integrated within this structure and is pivotal for refining and combining features at various scales. Additionally, the architecture includes classification and positional regression components, which enable precise object detection and localization within the input image.

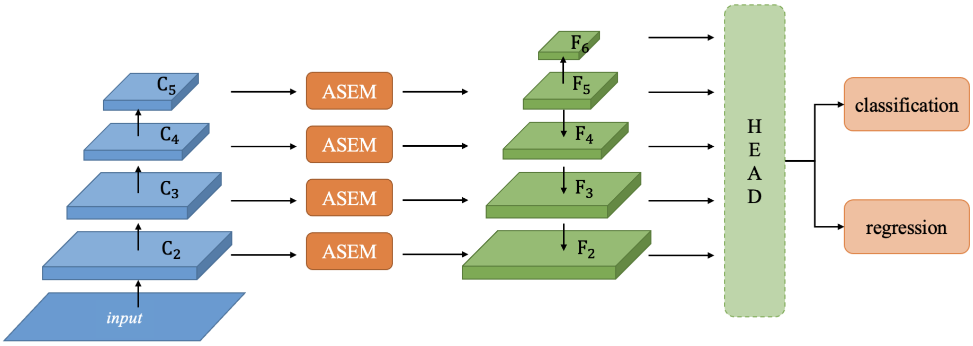

Figure 1. The architectural composition of the network. It encompasses a series of interconnected modules, including the backbone network, an adaptive multi-scale feature enhancement and fusion module (ASEM), and networks dedicated to classification and positional regression tasks.

The backbone network generates features at different stages, denoted as C2 to C5, which this paper refers to as inter-level features. Multi-scale feature extraction is conducted for each level, including C2, to extract scale-invariant features. These features serve as the foundation for subsequent calculations by the attention module. The features outputted by the ASEM are subjected to a dimensionality transformation through a 1 × 1 convolution, followed by an inter-level feature fusion that emulates the top-down pathway of the FPN.

3.2. Multi-Scale Feature Extraction and Fusion

Multi-scale feature extraction actually extracts features within receptive fields of different sizes. The most intuitive method is to apply convolution kernels of different sizes to simultaneously extract features of different scales. Another lightweight approach is to use dilated convolutions with different dilated rates to extract features of different scales.

Although this approach considers receptive fields of different sizes, it is a contextual feature enhancement method. By introducing more contextual information, we can determine the target class, as shown in Figure 2.

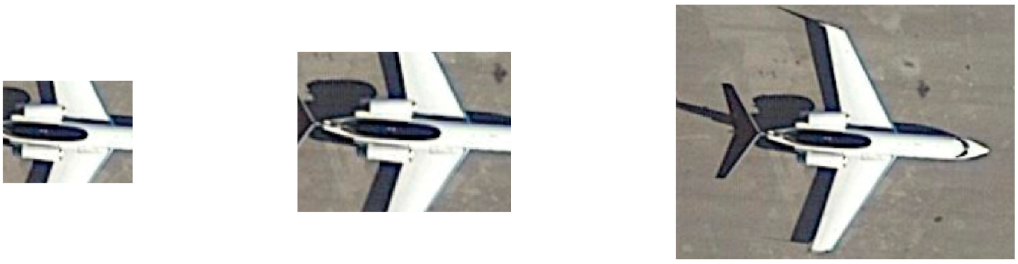

Figure 2. Images of the same object under different receptive fields.

While the contextual specifics of the object are critical, the primary focus of this module is to accentuate the object's scale, as depicted in Figure 3. Objects of the same category, despite varying in size, theoretically possess identical characteristics. This module aims to address this aspect by emphasizing scale variation between similar objects.

This section specifically introduces how to extract multi-scale features within the hierarchy. The module structure is shown in Figure 4.

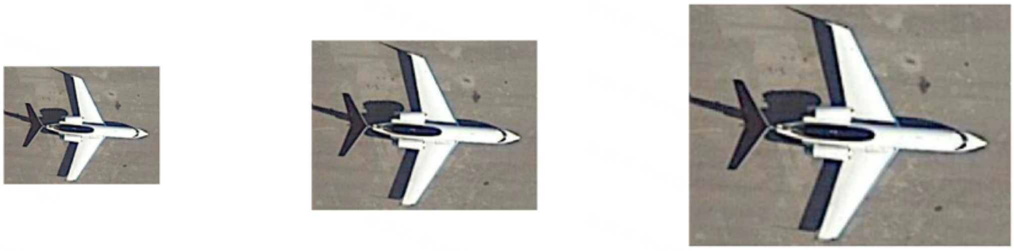

Figure 3. Images of objects of different sizes under different receptive fields.

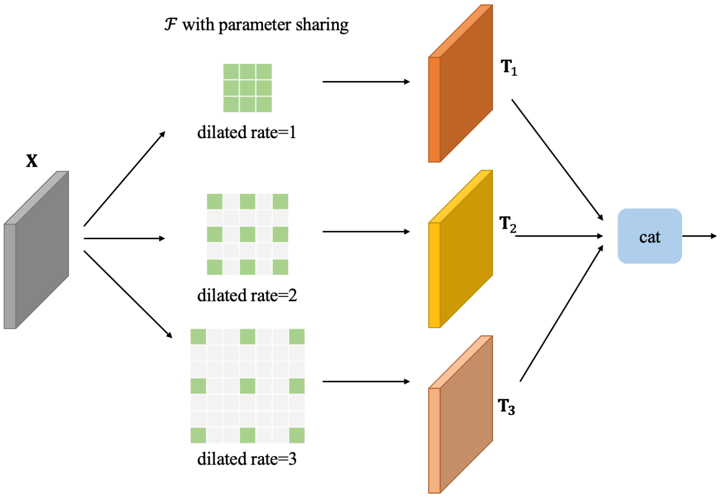

Figure 4. Schematic figure of multi-scale feature extraction structure.

For the given feature map $X \in \mathbb{R}^{H \times W \times C}$, three transformations are performed on it with the help of the dilated convolutions $\mathcal{F}_1 : X \rightarrow T_{\mathbf{1}} \in \mathbb{R}^{H' \times W' \times C'}$, $\mathcal{F}_2 : X \rightarrow T_{\mathbf{2}} \in \mathbb{R}^{H' \times W' \times C'}$, and $\mathcal{F}_3 : X \rightarrow T_{\mathbf{3}} \in \mathbb{R}^{H' \times W' \times C'}$. Here, H, W, and C represent the height, width, and channel, respectively. Similarly, H', W', and C' represent the same. This is described by the following equation:

$$T_k = \mathcal{F}_k(X) \tag{1}$$

The symbol $\mathcal{F}$ stands for an atrous convolution, and the variable k represents the dilated rate, which can be set to either 1, 2, or 3. It is important to note that the kernel size for the atrous convolution is 3×3, and when the dilated rate is set to 1, it is just a standard convolution. To make the features extracted from the same type of targets of different sizes as consistent as possible, the convolution kernel parameters of these three atrous convolutions are kept shared.

The features of the scale information of different branches are extracted for the input of the subsequent attention module. To obtain the importance of the information of different branches, all the information needs to be observed. Here, a Cat operation is used to fuse the information of these branches:

$$U = \mathrm{Cat}(T_1, T_2, T_3) = [T_1, T_2, T_3] \tag{2}$$

Here, Cat represents catenate, i.e., merging on the dimension of channel C'. At this point, multi-scale information is also integrated within each level.

3.3. *Adaptive Multi-Scale-Feature-Enhanced Module*

Through the merging operation Cat, the scale information of these three branches is fused into the feature U. If the feature U is directly used as the input of the subsequent module, then the contributions of these three branches to the subsequent network are equal. However, we hope to adaptively enhance or suppress them according to their importance.

To allow the network to adaptively extract target information of an appropriate scale, an attention mechanism is added to dynamically adjust the contribution of the channels of each branch so that more important branches and channels receive higher weights, while unimportant branches and channels then obtain a lower weight to achieve the adaptive weighting of features. The module structure is shown in Figure 5.

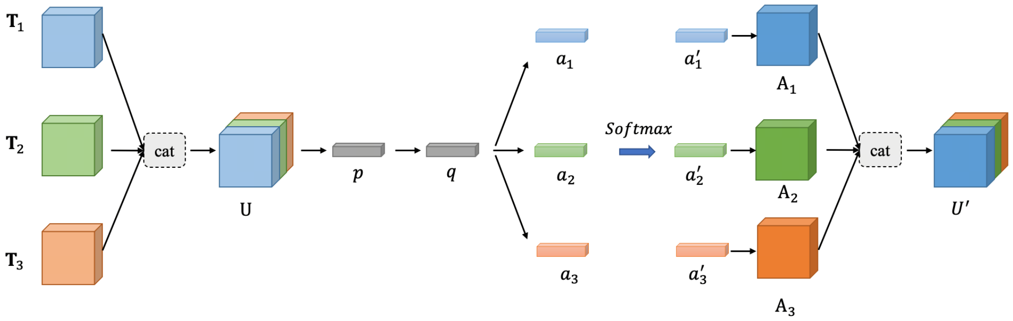

Figure 5. Adaptive multi-scale enhanced feature module structure diagram.

3.3.1. Global Pooling

Initially, we need to count the global channel information of the feature $p \in \mathbb{R}^{3\times C'}$. This step is straightforward. Compute the average U along the two dimensions of W' and H', effectively applying global average pooling. For the *k-th* channel, the elements are calculated using the following equation:

$$p_k = \mathcal{G}(U_k) = \frac{1}{H' \times W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} U_k(i,j) \tag{3}$$

where p_k is an element within p that represents the statistical information of the *k-th* channel; $\mathcal{G}(\cdot)$ represents the global average pooling transformation; and H' and W' represent the width and height of the feature map, respectively. The approach of channel dimension attention often requires such a process, as seen in networks like SENet [30] and SKNet [32], which first necessitate the statistical analysis of global information.

3.3.2. Attention Vector

To fully utilize the information of the feature U, we can use a fully connected layer to establish the correlation between its channels and convert it into a vector q. To make full use of the global channel information, the transformation $\mathcal{D}(\cdot)$ is used to establish the correlation between its channels and convert it into a vector $q \in \mathbb{R}^d$ to provide more accurate and adaptive choices for subsequent processing. To simplify this process, a fully connected layer is used to implement the transformation $\mathcal{D}(\cdot)$:

$$q = \mathcal{D}(p) = W_p \cdot p + B_p \tag{4}$$

In the fully connected layer transformation $\mathcal{D}(\cdot)$, p refers to the input feature; W_p denotes the weight matrix; B_p represents the bias, which is a vector; and $\cdot$ denotes element-wise multiplication.

The channel information between branches is relatively independent, and thus, three independent transformations, namely, $E_1(\cdot)$, $E_2(\cdot)$, and $E_3(\cdot)$, are used to transform q into the attention vectors a_1, a_2, and a_3 of each branch. Similarly, for simplicity, the transformation $E(\cdot)$ here also uses a fully connected layer.

$$\begin{aligned} a_1 &= E_1(q) = W_1 \cdot q + B_1 \\ a_2 &= E_2(q) = W_2 \cdot q + B_2 \\ a_3 &= E_3(q) = W_3 \cdot q + B_3 \end{aligned} \tag{5}$$

By establishing correlations with fully connected layers, channel information can be independently processed among the branches and interwoven within each branch. This method ensures that every branch obtains the most pertinent and efficient channel information, which enhances the overall accuracy and performance of the system.

3.3.3. Feature Fusion

First, the attention vectors of the three branches are merged to obtain the vector $a \in \mathbb{R}^{3\times d}$:

$$a = \mathrm{Cat}(a_1, a_2, a_3) = [a_1, a_2, a_3] \tag{6}$$

Then, a softmax operation is executed on a along the merged dimension to derive the weighted scores for each branch and channel:

$$a' = \mathrm{softmax}(a) = \left[\frac{e^{a_1}}{e^{a_1}+e^{a_2}+e^{a_3}}, \frac{e^{a_2}}{e^{a_1}+e^{a_2}+e^{a_3}}, \frac{e^{a_3}}{e^{a_1}+e^{a_2}+e^{a_3}}\right] \tag{7}$$

Note that to collectively evaluate the information from the three branches, the softmax operation is applied to the merged features rather than executing separate softmax operations on the attention vectors of each branch.

Next, the weighted features of each branch are calculated through the element-wise dot multiplication operation:

$$\begin{aligned} A_1 &= T_1 \cdot a'_1 \\ A_2 &= T_2 \cdot a'_2 \\ A_3 &= T_3 \cdot a'_3 \end{aligned} \tag{8}$$

Finally, the weighted features A_1, A_2, and A_3 of each branch are fused together to obtain U':

$$U' = \mathrm{Cat}(A_1, A_2, A_3) = [A_1, A_2, A_3] \tag{9}$$

The merging method is still used here to preserve the detailed information in the vector space against degradation and to align with the prior fusion operations of the adaptive module.

After undergoing several steps, including global pooling, computing attention vectors, and performing feature-weighted fusion, the result is an adaptively enhanced multi-scale feature U', which is then suitable for input into subsequent networks. These processes effectively harness global channel information, inter-branch scale information, and inter-channel correlations, thereby offering more precise adaptive features.

4. Experiments and Results

4.1. Datasets

DOTA-v1.0 [33] dataset: This is a large-scale dataset for object detection in aerial images. The dataset comprises 2806 aerial images, which are divided into 1411 for training, 937 for validation, and 458 for testing. The resolution of the images is high, ranging from 800 × 800 to 4000 × 4000 pixels. It contains objects of various scales, orientations, and shapes, and is annotated with directed bounding boxes, as shown in Figure 6. The dataset encompasses 15 common object categories: plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC).

HSRC2016 [34] dataset: This dataset is a high-resolution satellite image collection designed explicitly for ship detection and identification tasks. It comprises images from the Gaofen-3 satellite, which are characterized by their high spatial resolution. The dataset includes comprehensive annotation information, such as detailed target location, size, and orientation information, as shown in Figure 7. It consists of 1061 remote sensing images extracted from Google Earth, showcasing 2976 ship instances. Image resolutions span from 300 × 300 to 1599 × 900 pixels. This collection is systematically divided, offering

436 images for the training set, 181 images for validation, and 444 images designated for the test set.

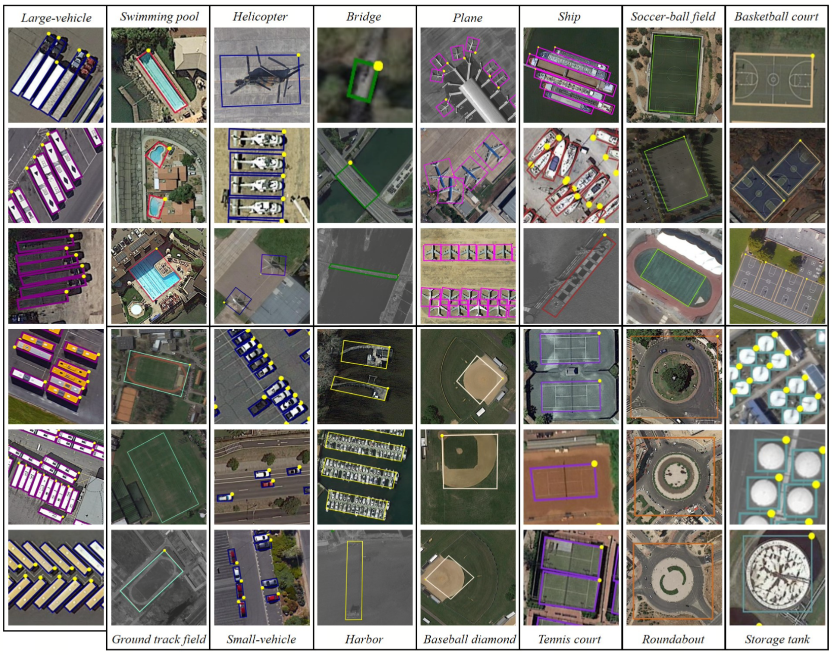

Figure 6. Samples of DOTA-v1.0 [33].

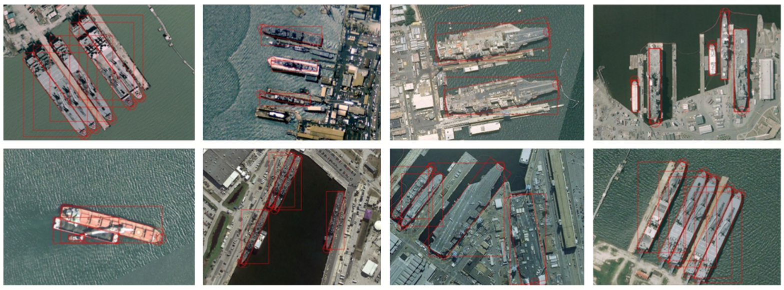

Figure 7. Samples of HSRC2016 [34].

4.2. Implementation Details

The server utilized in the experiments operated on an Ubuntu 20.04 LTS system that was equipped with an NVIDIA GeForce RTX 3090 GPU. The software environment was based on Python 3.8.18 and PyTorch 1.10.0.

The experimental code presented in this article was developed by the MMRotate [35] framework, which is based on PyTorch. The backbone network for feature extraction, namely, ResNet, was initially pre-trained on the ImageNet dataset and subsequently fine-tuned on the remote sensing dataset employed in our experiments while training. The final four layers of the backbone network served as the input for the ASEM, with the output channel set to 256.

Due to the high resolution of DOTA, input images were cropped to 1024 × 1024-pixel patches by a sliding window with an overlap of 200 pixels before being used for training and prediction. To enhance the robustness of the model, we performed data augmentation by randomly flipping images horizontally, vertically, or diagonally with a 25% probability. For the HRSC2016 dataset, the images were resized to 512 × 800 pixels. To improve the model's robustness, we employed data augmentation by randomly flipping the images horizontally, vertically, or diagonally with a 25% probability and rotating the images with a 50% probability.

During the training of the DOTA dataset, both the training and validation sets were utilized together, and the batch size was set to 2. The initial learning rate was set to 0.005, the momentum was set to 0.9, and the weight decay was set to 0.0001. The number of training epochs was set to 12, and the learning rate linearly decayed from the 8th epoch to the 11th epoch. When the training was complete, the experimental results needed to be submitted to the official online evaluation platform provided by DOTA for testing and evaluation. The test platform returned the mean average precision (mAP) and accuracy of each category. In the training of the HRSC2016 dataset, both the training and validation sets were also used together with a batch size of 2. The initial learning rate was set to 0.01, the momentum was set to 0.9, and the weight decay was set to 0.0001. The dataset was trained for 36 epochs, and the learning rate was set to decay between the 24th and 33rd epochs. The NMS threshold was set to 0.7. For further details, please refer to https://github.com/Christy99cc/SeRotate, (accessed on 26 February 2024).

4.3. Results

In this study, Rotated FasterRCNN was utilized as the baseline model, with ResNet50 as the backbone network. The rationale for selecting this as the baseline extended beyond convention and was founded on the following considerations:

Rotated FasterRCNN is an improved version of FasterRCNN [18] for detecting rotated objects. Accurate detection of rotated targets can be achieved by introducing the representation of rotated bounding boxes and the corresponding loss function. In the field of remote sensing, numerous studies have built upon Rotated FasterRCNN, showcasing its technological maturity, broad applicability, and reliability. Compared with alternative methods, Rotated FasterRCNN is more extensively supported by empirical evidence demonstrating its effectiveness. Selecting Rotated FasterRCNN as the baseline offers a solid foundation for research, leveraging its widespread recognition and validation to ensure a stable starting point for further investigation.

For the DOTA-v1.0 dataset, the experimental results are shown in Table 1, where R-FR represents Rotated FasterRCNN.

Table 1. Experimental results of different detection algorithms on DOTA-v1.0 dataset.

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
FR-O [33]	R101	79.42%	77.13%	17.70%	64.05%	35.30%	38.02%	37.16%	89.41%	69.64%	59.28%	50.30%	52.91%	47.89%	47.40%	46.30%	54.13%
ICN [36]	R101-FPN	81.40%	74.30%	47.70%	70.30%	64.90%	67.80%	70.00%	90.80%	79.10%	78.20%	53.60%	62.90%	67.00%	64.20%	50.20%	68.20%
RoI-Trans. [17]	R101-FPN	88.64%	78.52%	43.44%	**75.92%**	68.81%	73.68%	83.59%	90.74%	77.27%	81.46%	58.39%	53.54%	62.83%	58.93%	47.67%	69.56%
CADNet [37]	R101-FPN	87.80%	82.40%	49.40%	73.50%	71.10%	63.50%	76.60%	**90.90%**	79.20%	73.30%	48.40%	60.90%	62.00%	67.00%	62.20%	69.90%
CenterMap [38]	R50-FPN	88.88%	81.24%	**53.15%**	60.65%	78.62%	66.55%	78.10%	88.83%	77.80%	83.61%	49.36%	66.19%	**72.10%**	**72.36%**	58.70%	71.74%
SCRDet [7]	R101-FPN	**89.98%**	80.65%	52.09%	68.36%	68.36%	60.32%	72.41%	90.85%	**87.94%**	**86.86%**	**65.02%**	**66.68%**	66.25%	68.24%	**65.21%**	72.61%
R-FR [18]	R-50-FPN	89.25%	**82.45%**	49.95%	69.36%	78.17%	73.60%	**85.92%**	**90.90%**	84.04%	85.48%	57.58%	60.98%	66.25%	69.23%	57.74%	73.40%
Ours	R-50	89.26%	82.26%	51.33%	68.49%	**78.88%**	**74.14%**	85.59%	90.88%	84.94%	85.73%	60.78%	64.76%	65.72%	71.32%	59.08%	**74.21%**

For the HRSC2016 dataset, the experimental results are shown in Table 2.

Compared with the baseline Rotated FasterRCNN, the method proposed in this paper exhibited an increase in mAP by 0.81% on the DOTA-v1.0 dataset and 9.2% on the HRSC2016 dataset. These results show that the incorporation of multi-scale feature fusion and attention mechanisms improved the multi-scale remote sensing object detection.

Moreover, Figures 8 and 9 display the visualization results on the DOTA test set and HRSC2016 test set, respectively.

Table 2. Experimental results of different detection algorithms on HRSC2016 dataset.

Method	Backbone	mAP
RRPN [16]	R101	79.08%
R2CNN [39]	R101	73.07%
R-FR [18]	R-50-FPN	75.70%
Ours	R-50	**84.90%**

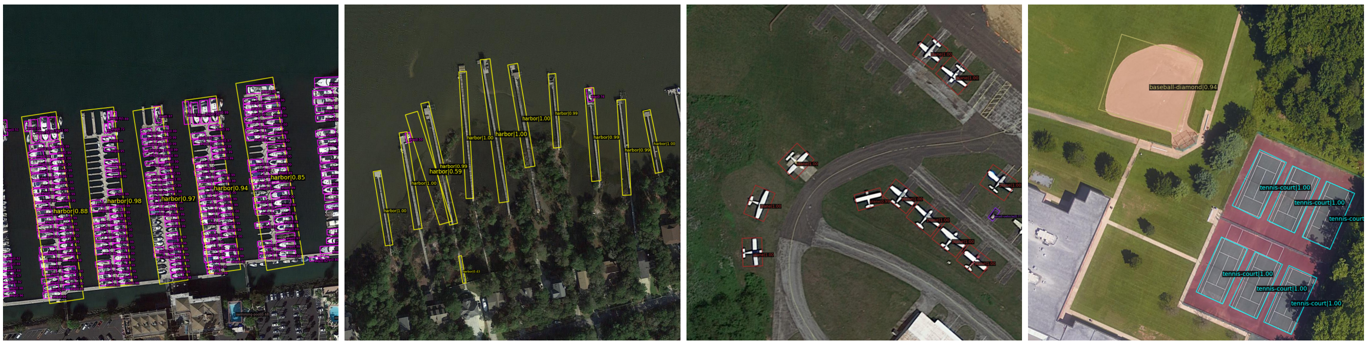

Figure 8. Visualization on DOTA-v1.0 test set.

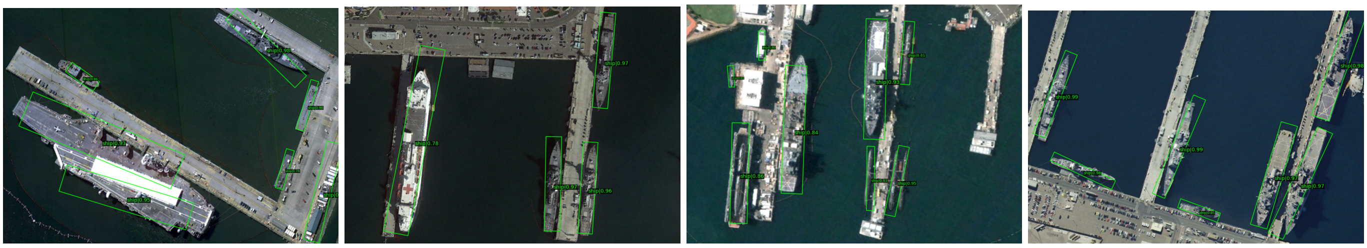

Figure 9. Visualization on HRSC2016 test set.

4.4. Ablation Study

Based on the Rotated FasterRCNN as the baseline, ablation experiments were conducted on the DOTA-v1.0 dataset, and the experimental results are listed in Tables 3–6.

Table 3. Ablation experimental results on DOTA-v1.0 dataset.

Method	PL	BD	BR	GTF	SV	LV	SH	TC
R-FR+FPN	89.25%	82.45%	49.95%	**69.36%**	78.17%	73.60%	**85.92%**	**90.90%**
ADD	89.25%	**83.24%**	50.48%	66.61%	78.79%	**74.95%**	85.30%	**90.90%**
Ours	**89.26%**	82.26%	**51.33%**	68.49%	**78.88%**	74.14%	85.59%	90.88%
Method	**BC**	**ST**	**SBF**	**RA**	**HA**	**SP**	**HC**	**mAP**
R-FR+FPN	84.04%	85.48%	57.58%	60.98%	**66.25%**	69.23%	57.74%	73.40%
ADD	**85.02%**	85.50%	55.94%	**66.19%**	65.66%	71.26%	**60.54%**	73.98%
Ours	84.94%	**85.73%**	**60.78%**	64.76%	65.72%	**71.32%**	59.08%	**74.21%**

Table 4. Ablation experimental results on HRSC2016 dataset.

Method	Backbone	mAP
R-FR [18]	R-50-FPN	75.70%
ADD	R-50	82.90%
Ours	R-50	**84.90%**

Ablation study of ASEM: To investigate the effectiveness of the adaptive multi-scale feature enhancement module (ASEM), we conducted an ablation study on methods with and without this module; the results are presented in Tables 3 and 4. Through multi-scale feature extraction, fusion, and adaptive enhancement, ASEM achieved a gain of 0.81%, increasing from 73.40% to 74.21% on the DOTA-v1.0 dataset, while ASEM helped the model to gain 9.2% mAP, growing from 75.70% to 84.90% on the HRSC2016 dataset.

Ablation study of fusion methods in ASEM: To explore the effectiveness of the fusion method in this module, ablation experiments were conducted on the element addition fusion and Cat fusion methods; the results are shown in Tables 3 and 4. Compared with the baseline method, on the DOTA-v1.0 dataset, element addition fusion achieved an improvement of 0.58% (from 73.40% to 73.98%), and Cat fusion achieved a small improvement of 0.23% compared with additive fusion (from 73.98% to 74.21%). In the HRSC2016 dataset, element addition fusion achieved a 7.20% improvement over the baseline method (from 75.70% to 82.90%), and Cat fusion achieved another 2.0% improvement over element addition fusion (from 82.90% to 84.90%).

Furthermore, a detailed ablation analysis was conducted on the number of dilated convolutions used for scale-invariant feature extraction within the same hierarchical level. The number of atrous convolutions was correlated with the dilation rates, as described in Equation (1). The results of the ablation experiments are shown in Tables 5 and 6.

Table 5. Ablation study results on dilated convolutions of the DOTA-v1.0 dataset.

Method	Dilation Rate	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
R-FR+FPN	-	89.25%	82.45%	49.95%	**69.36%**	78.17%	73.60%	**85.92%**	**90.90%**	84.04%	85.48%	57.58%	60.98%	66.25%	69.23%	57.74%	73.40%
Ours	(1, 2)	89.13%	**83.97%**	50.23%	67.77%	78.84%	**75.45%**	85.30%	90.89%	**85.84%**	85.48%	55.09%	**66.72%**	**66.67%**	70.90%	**60.06%**	74.16%
	(1, 2, 3)	89.26%	82.26%	**51.33%**	68.49%	**78.88%**	74.14%	85.59%	90.88%	84.94%	**85.73%**	**60.78%**	64.76%	65.72%	71.32%	59.08%	**74.21%**
	(1, 2, 3, 4)	**89.39%**	81.00%	49.95%	66.19%	78.73%	74.85%	85.33%	**90.90%**	84.63%	85.64%	55.02%	66.59%	65.29%	**71.96%**	59.47%	73.66%

Table 6. Ablation study results on dilated convolutions of the HRSC2016 dataset.

Method	Dilation Rate	Backbone	mAP
R-FR [18]	-	R-50-FPN	75.70%
Ours	(1, 2)	R-50	82.40%
	(1, 2, 3)	R-50	**84.90%**
	(1, 2, 3, 4)	R-50	83.30%

Ablation study results about dilated convolutions: To determine the appropriate number of scale levels of features divided within the hierarchical level, we conducted several experiments to investigate the number of dilated convolutions utilized for scale-invariant feature extraction within this hierarchical level. In the DOTA-v1.0 dataset, compared with the baseline, our method yielded improvements. Specifically, a dilation rate configuration of (1, 2) enhanced the mAP by 0.76%, a setup of (1, 2, 3) achieved a 0.81% mAP increase, and a configuration of (1, 2, 3, 4) led to a 0.26% improvement. In the HRSC2016 dataset, relative to the baseline, our approach facilitated enhancements, with a dilation rate configuration of (1, 2) resulting in a 6.7% mAP increase, a setup of (1, 2, 3) securing a 9.2% mAP enhancement, and a configuration of (1, 2, 3, 4) achieving a 7.6% mAP improvement. The experimental evidence suggests that a dilation rate configuration of (1, 2, 3) represents the most optimal setting currently available.

5. Discussion

The empirical research results show that our proposed method provided a significant improvement compared with the baseline method, namely, Rotated FasterRCNN, in the multi-scale remote sensing target detection task. Specifically, on the DOTA-v1.0 dataset, our method improved the mAP by 0.81% compared with the baseline method. This shows that the method proposed in this paper could more accurately locate and classify target objects in remote sensing images, providing a more reliable solution to the task. Similarly, experiments on the HRSC2016 dataset also showed promising results. Compared with the baseline method, the mAP of this method on this data set increased by 9.2%. This improvement further verified the effectiveness and practicability of this method, especially when dealing with ship target detection at sea.

These advantages are obtained from the adaptive multi-scale feature fusion method used in this method. In addition to the fusion of multi-scale features between levels, this study also paid special attention to the extraction and fusion of multi-scale features within the same level. Introducing a multi-scale feature enhancement module allows feature representations in different scale ranges to be more accurately explored and utilized. At the same time, this study adopted the strategy of feature fusion and attention allocation to further optimize the expression and attention mechanism of features. Through feature fusion, feature information from different scales is combined to make the feature expression more comprehensive. The attention distribution can adaptively guide the model to focus on and strengthen important features, playing a vital role in the target detection process. Although attention mechanisms have been utilized in many studies, this study focused on selecting features at the appropriate scale, which plays a crucial role in multi-scale object detection in remote sensing.

The ablation experiments revealed that the ASEM significantly impacted the results. The fusion method and the dilation rate settings also played some roles. Both element-wise addition and Cat (concatenation) fusion methods benefited the outcomes, but concatenation yielded better results. This was likely because concatenation combines features along the channel dimension, preserving spatial information to some extent.

As for the setting of dilation rates, the number of dilated convolutions with sharing parameters used determines the number of scale levels of features divided within the hierarchical level. This corresponds to how many scale levels of features are scale invariant. The experimental results indicate that incorporating dilated convolutions with shared parameters was beneficial. The experiments were designed with varying numbers of dilated convolutions sharing parameters. When the number of shared-parameter dilated convolutions was two or three, the results showed an improvement; however, with four, there was a slight downward trend. Therefore, we selected three as the optimal number of dilated convolutions for our experiments. Choosing three dilated convolutions also aligns with common sense in categorizing object sizes into large, medium, and small. When the number of dilated convolutions increases to four, it is considered that the excessive enlargement of the receptive field introduces unnecessary noise, leading to a decrease in the performance.

This design and strategy allow the method to adaptively concentrate on features of objects at various scales, resulting in substantial accuracy improvements. By fully leveraging the extraction and fusion potential of multi-scale features, the method effectively captures key characteristics in object detection tasks, leading to an enhanced detection performance.

6. Conclusions

We propose an adaptive multi-scale feature fusion method for remote sensing object detection to improve accuracy and robustness. In addition to the fusion of multi-scale features across levels, this study also focused on the extraction and fusion of multi-scale features within each level. Specifically, the extraction of scale-invariant features within levels effectively captured the characteristic information of targets at various scales. Furthermore, the proposed ASEM enabled adaptive focus on significant feature scales, thereby enhancing the precision of object detection.

Research indicates that leveraging multi-scale feature fusion and attention mechanisms can enhance the performance of multi-scale remote sensing object detection. Further research can expand our method, including adaptive feature fusion techniques, to address complex scenarios and object categories and enhance the performance efficiency.

Author Contributions: Conceptualization, C.L. and S.Z.; methodology, C.L. and S.Z.; software, S.Z.; validation, S.Z.; formal analysis, S.Z.; investigation, S.Z.; resources, M.H. and Q.S.; data curation, C.L. and S.Z.; writing—original draft preparation, S.Z.; writing—review and editing, C.L. and S.Z.; supervision, C.L., M.H., and Q.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Sagar, A.S.; Chen, Y.; Xie, Y.; Kim, H.S. MSA R-CNN: A comprehensive approach to remote sensing object detection and scene understanding. *Expert Syst. Appl.* **2024**, *241*, 122788. [CrossRef]
2. Zhang, X.; Zhang, T.; Wang, G.; Zhu, P.; Tang, X.; Jia, X.; Jiao, L. Remote Sensing Object Detection Meets Deep Learning: A metareview of challenges and advances. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 8–44. [CrossRef]
3. Yu, Y.; Da, F. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13354–13363.
4. Jiang, X.; Wu, Y. Remote Sensing Object Detection Based on Convolution and Swin Transformer. *IEEE Access* **2023**, *11*, 38643–38656. [CrossRef]
5. Gao, T.; Niu, Q.; Zhang, J.; Chen, T.; Mei, S.; Jubair, A. Global to local: A scale-aware network for remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615614. [CrossRef]
6. Chen, S.; Zhao, J.; Zhou, Y.; Wang, H.; Yao, R.; Zhang, L.; Xue, Y. Info-FPN: An Informative Feature Pyramid Network for object detection in remote sensing images. *Expert Syst. Appl.* **2023**, *214*, 119132. [CrossRef]
7. 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, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
8. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 3520–3529.
9. Song, Q.; Yang, F.; Yang, L.; Liu, C.; Hu, M.; Xia, L. Learning point-guided localization for detection in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1084–1094. [CrossRef]
10. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
11. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
12. Wang, X.; Zhang, S.; Yu, Z.; Feng, L.; Zhang, W. Scale-equalizing pyramid convolution for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13359–13368.
13. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 4278–4284. [CrossRef]
14. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
15. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-free local feature matching with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8922–8931.
16. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
17. 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, 16–20 June 2019; pp. 2849–2858.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]

19. 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.
20. 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]
21. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 677–694.
22. Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; Yan, J. Dense label encoding for boundary discontinuity free rotation detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15819–15829.
23. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 1819–1828.
24. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
25. 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, 18–22 June 2018; pp. 8759–8768.
26. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
27. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7036–7045.
28. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
29. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems, NIPS 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 2017–2025.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
31. 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.
32. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
33. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
34. 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 6th International Conference on Pattern Recognition Applications and Methods—Volume 1: ICPRAM, INSTICC, SciTePress, Porto, Portugal, 24–26 February 2017; pp. 324–331.
35. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C. Mmrotate: A rotated object detection benchmark using pytorch. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 7331–7334.
36. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 150–165.
37. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
38. Wang, J.; Yang, W.; Li, H.C.; Zhang, H.; Xia, G.S. Learning center probability map for detecting objects in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4307–4323. [CrossRef]
39. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.