remote sensing

MDPI

Article

Adaptive Dual-Domain Dynamic Interactive Network for Oriented Object Detection in Remote Sensing Images

Yongxian Zhao [1,2], Tao Yang [3,4], Shuai Wang [1], Hailin Su [1,2] and Haijiang Sun [1,*]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; zhaoyongxian20@mails.ucas.ac.cn (Y.Z.); wangshuai@ciomp.ac.cn (S.W.); suhailin22@mails.ucas.ac.cn (H.S.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Innovation Center for Control Actuators, Beijing 100076, China; yangtao@sia.cn
[4] Beijing Institute of Precision Mechatronics and Controls, Beijing 100076, China
* Correspondence: sunhj@ciomp.ac.cn

Abstract: Object detection in remote sensing images is an important research topic in the field of remote sensing intelligent interpretation. Although modern object detectors have made good progress, high-precision oriented object detection still faces severe challenges due to the large-scale variation, strong directional diversity and complex background interference of objects in remote sensing images. Currently, most remote sensing object detectors focus on modeling object characteristics in the spatial domain while ignoring the frequency domain information of the object. Recent studies have shown that frequency domain learning has brought substantial benefits in many visual fields. To this end, we proposed an adaptive dual-domain dynamic interaction network (AD^{3}I-Net) for oriented object detection tasks in remote sensing images. The network has three important components: a spatial adaptive selection (SAS) module, a frequency adaptive selection (FAS) module, and a dual-domain feature interaction (DDFI) module. The SAS module adaptively models spatial context information and dynamically adjusts the feature receptive field to construct more accurate spatial position features for objects of different scales. The FAS module uses the transformation from the spatial domain to the frequency domain to adaptively learn the frequency information of the object, to model direction features, and to make up for the lack of spatial domain information. Finally, through the DDFI module, the features extracted from the two domains are interactively fused, thereby bridging the complementary information to enhance the feature expression of the object and give it rich spatial position and direction information. The AD^{3}I-Net we proposed fully exploits the interaction between the different domains and improves the model's ability to capture subtle object features. Our method has been extensively experimentally verified on two mainstream datasets, HRSC2016 and DIOR-R. The experimental results demonstrate that this method performs competitively in oriented object detection tasks.

Keywords: remote sensing images; deep learning; frequency analysis; adaptive selection

check for updates

Received: 28 January 2025
Revised: 21 February 2025
Accepted: 6 March 2025
Published: 7 March 2025

Citation: Zhao, Y.; Yang, T.; Wang, S.; Su, H.; Sun, H. Adaptive Dual-Domain Dynamic Interactive Network for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 950. https://doi.org/10.3390/rs17060950

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid development of satellite technology, the scale and quality of remote sensing data are also increasing. In recent years, the task of intelligent interpretation of remote sensing images based on deep learning has received widespread attention [1–3]. As the core issue in the task of intelligent interpretation of remote sensing images, object detection plays an important role in the fields of civil life and military, including object

monitoring, disaster relief, and intelligence reconnaissance. The core of remote sensing image object detection technology lies in how to accurately locate and accurately identify the object of interest in the image. The objects of interest in remote sensing image object detection usually include various types of objects such as ships, aircraft, vehicles, oil tanks, ports, bridges, etc. Due to the unique perspective of remote sensing scenes, object types often exist in multiple scales, dense distributions, and arbitrary direction arrangements, combined with different imaging conditions, complex backgrounds, and environmental factors. In Figure 1, there are three remote sensing scene images showing the characteristics of objects in remote sensing bird's eye view scenes.



Figure 1. Three images of remote sensing scenes. They have the characteristics of complex background, large object scale variation, dense distribution, and arbitrary orientation.

To obtain a more accurate position representation of the object, a large number of studies are currently devoted to designing various oriented bounding box detection algorithms. The mainstream algorithms are divided into two-stage and single-stage detection algorithms. For the two-stage oriented object detection algorithm, PRN is generally used to generate a rotation region of interest, and then the features are extracted by pooling the region of interest. Finally, the RCNN detection head is used to predict the position and category information of the object. The single-stage oriented object detection algorithm directly regresses the position parameters of the object to improve the detection efficiency. In view of the unique characteristics of remote sensing scenes, the use of rotating bounding boxes to detect multi-directional objects can reduce the impact of the background on the object and obtain the orientation information of the object. Different representation forms of rotating bounding boxes will affect the network structure's design and the object's positioning effect. The appropriate representation form of the rotating box can make it easier for the network to learn directional features and have a lower model complexity. The rotated rectangular bounding box is the most popular representation method in the oriented object detection task. The model obtains five parameters representing the rotated rectangular box by regression prediction. The method can obtain angle information simply and intuitively. In order to avoid the boundary mutation problem caused by angle periodicity, Yang et al. [4] proposed a new angle encoding method for smooth labels to convert the angle regression problem into a classification problem. In addition, in order to solve the limitations of the classic norm loss function in oriented object detection algorithms, Yang et al. [5] proposed the GWD algorithm, which uses the wasserstein distance to measure the distance between object Gaussian distributions for loss optimization in training. These methods focus on improving the accuracy of oriented object detection by improving regression parameters and optimizing loss functions, but they often ignore the importance of feature extraction.

Feature extraction is a crucial step in the object detection task which directly affects the performance of the model. Remote sensing images have rich environmental information, appearance information, and contextual information, but they are still not fully developed.

Recently, some methods have made corresponding improvements on the basic feature extraction network, feature fusion network, and feature refinement module by designing novel deep learning networks. These methods have significantly enhanced the feature extraction ability and improved the detection performance by constructing multi-scale features, generating enhanced features, and using mainstream attention mechanisms. Zhang et al. [6] proposed a spatial and scale-aware attention module to guide the network to focus on areas with rich information and more appropriate feature scales. However, these attention modules are limited in that they only focus on the location of the object and ignore the unique directional characteristics of the object in the remote sensing images, resulting in a decrease in the accuracy of oriented bounding box detection. From the previous studies, it is not enough to capture the information of the object in remote sensing images only in the spatial domain. Using the coarse architecture of enhanced features such as attention mechanisms, it is difficult to extract the potential directional information in the image. In addition, operations such as convolution and downsampling in the spatial domain will lead to the loss of high-frequency detail information in the image. The high-frequency components in the frequency domain correspond to the edge and texture features of the object. By learning the frequency information, the directional changes of the target can be effectively captured, which helps the detection network to extract the directional information of the object. In addition, feature learning in the frequency domain can effectively reduce the impact of object rotation on feature expression, thereby improving the robustness of the detection model to objects in any direction to a certain extent. Therefore, we proposed a new method, called an adaptive dual-domain feature interaction network, for oriented object detection in remote sensing images. We extracted direction information by converting the spatial domain into the frequency domain, and we designed a frequency adaptive selection module to enable the network to dynamically learn the frequency characteristics of the object. At the same time, to ensure that the model could have different receptive fields when dealing with objects of different scales, we designed a spatial adaptive selection module. Subsequently, the proposed dual-domain feature interaction module interacted with the features extracted from the spatial domain and the frequency domain to fuse the subtle relationship between spatial position information and direction information, effectively enhancing the feature capture ability of the object. Finally, the rich feature information was input into the feature pyramid and the detection head to obtain more accurate oriented object detection performance. Therefore, the contributions of this research mainly include the following important aspects:

(1) We proposed a spatial adaptive selection module to extract features of different scales of the object so that the model could dynamically learn contextual information according to the object characteristics to match the receptive field size more suitable for the object itself, thereby constructing more accurate spatial position information;
(2) In order to make up for the shortcomings of single domain information, we proposed a frequency adaptive selection module to extract direction information by converting spatial domain features into the frequency domain, effectively enhancing the network's ability to model direction diversity;
(3) In the dual-domain feature interaction module, we interactively fused the features extracted from the spatial domain and the frequency domain to bridge the complementary information and to achieve the purpose of generating enhanced features, effectively improving the expressiveness of object features;
(4) The AD^{3}I-Net we proposed fully exploited the interaction relationship between the different domains, improved the ability to capture object features, and gave them rich spatial position and direction information. The performance of this method on the HRSC2016 dataset and the DIOR-R dataset was better than many advanced

methods and had a certain competitiveness. At the same time, this also confirmed the effectiveness of frequency domain learning in the task of oriented object detection in remote sensing images.

2. Related Works

2.1. Oriented Object Detection

Remote sensing images face great challenges in oriented object detection due to the large differences in object scales, the distribution of object directions at arbitrary angles, and complex backgrounds. In response to the problem of the diversity of object directions to be detected, some studies [7,8] first use the simplest method to expand the angle of the object by image rotation enhancement operations to improve the model's adaptability to objects in arbitrary directions. Although this method can play a certain role in the learning process of big data, it cannot fundamentally solve the problem of the model's poor sensitivity to direction. Therefore, Han et al. [9] proposed the ReDet algorithm to encode the object's rotation equivariant features and rotation invariant features so that it can adaptively extract the rotation invariant features from the equivariant features and improve the detection accuracy of multi-directional objects. In view of the dense arrangement of objects in remote sensing images, the conventional use of horizontal detection bounding boxes will lead to a large amount of interference from irrelevant backgrounds in the area, and even the high overlap between different object bounding boxes will cause object omissions and misdetections during post-processing. For the above reasons, a large number of studies have begun to design more accurate bounding box regression methods. Xie et al. [10] proposed a simple and effective two-stage framework in which high-quality anchor proposals are generated by oriented RPN in the first stage, and the oriented R-CNN detection head is used to refine the oriented region of interest in the second stage. In addition, another scheme completely abandons the traditional anchor box approach and uses arbitrary anchor points that can be learned. Dai et al. [11] dynamically collected object contour information using corner points and then gradually evolve to generate a quadrilateral bounding box of the object. Subsequently, Cheng et al. [12] abandoned the prior horizontal anchor box and proposed to generate oriented proposal bounding boxes in the RPN stage of the two-stage detector without anchor boxes. Li et al. [13] also proposed to generate high-quality oriented bounding boxes without a large number of preset anchor boxes and designed a novel bounding box representation method based on polar coordinates, which greatly improved the detection accuracy. Li et al. [14] proposed a novel key point detector which selects the most representative samples through a quality evaluation and sample allocation strategy to capture the object features from adjacent objects and complex background information, greatly improving the detection accuracy of densely arranged multi-directional objects.

In addition to the above CNN-based network architecture, some researchers have also introduced the Transformer architecture. Ma et al. [15] first used the DETR model for oriented object detection tasks in remote sensing images and proposed a more efficient encoder that used deep separable convolution instead of a self-attention mechanism to reduce memory usage. Dai et al. [16] simplified the Transformer detector and proposed a directional candidate box generation mechanism to provide better prior information for fusion features. At the same time, they introduced an adaptive candidate box refinement module and increased matching loss to ensure the correctness of the predicted bounding box. Zhou et al. [17] proposed to use point prediction to replace the bounding box regression. The distribution of points can better reflect the location information of the oriented bounding box, decouple the query features, reduce the number of object queries of the decoder, and balance the model accuracy and efficiency to a certain extent. In addition, Wang et al. [18] proposed a new architecture focused on remote sensing object detection.

By recognizing the background area, the category and bounding box of the predicted object in the area containing only the foreground are predicted. This method reduced redundant calculations and achieved higher quality detection results.

However, the above methods are only based on spatial domain features and cannot break away from the limitations of spatial domain information. To this end, we proposed an oriented object detection algorithm based on joint dual-domain features which used the spatial domain and the frequency domain to extract the rich characteristics of the object, greatly improving the perception ability of the object and thereby enhancing the robustness and generalization of the model.

2.2. Frequency Domain Learning in Image Processing

Frequency learning has always been a basic analysis tool in the field of traditional signal processing. Long-term research has shown that frequency domain analysis is also a very effective technology in image processing. In recent years, frequency domain learning has been introduced by a large number of scholars into the field of deep learning for application and has achieved excellent results. Ehrlich et al. [19] introduced frequency analysis methods in convolutional neural networks to solve the JPEG encoding problem. Subsequently, Qin et al. [20] proposed to use discrete cosine transform as a general representation of global average pooling, and they introduced more frequency components to make up for the shortcomings of insufficient feature information in existing channel attention. Rao et al. [21] proposed a global filtering network using a two-dimensional Fourier transform to learn long-term spatial dependencies from frequency domain features and demonstrated good performance of the model. Zhong et al. [22] introduced frequency domain features as additional clues in the camouflaged object recognition task and enriched the features by designing a frequency domain enhancement module to distinguish camouflaged objects from the background. Yang et al. [23] used wavelet transform features in the segmentation task to decompose features into high-frequency and low-frequency features and combined them with spatial features to cope with the challenges of significant grayscale change areas. Chen et al. [24] also realized the importance of spectral analysis and proposed to use frequency to improve dilated convolution. They designed a frequency adaptive dilated convolution to adjust the dilation rate to dynamically supplement high-frequency information. Similarly, Finder et al. [25] proposed a new convolutional layer, WTConv, by combining wavelet transform to enable the convolutional neural network to obtain a global receptive field, and they used wavelet transform to decompose it into different frequency bands to enhance the model's response to shape features. In addition, Kong et al. [26] proposed a frequency-domain-based self-attention decoder and a discriminative feedforward network which used the frequency domain characteristics of the image to effectively remove image blur. Patro et al. [27] used the spectral layer of Fourier transform and combined it with a multi-head attention mechanism to capture relevant features, verifying that frequency analysis is also effective in the Transformer architecture. Yao et al. [28] combined wavelet transform and self-attention to enhance output features and achieve low-cost downsampling. Cao et al. [29] proposed a dual-frequency attention fusion network to improve the ability of classification discriminative feature learning, and they enhanced the feature expression by combining CNN features and frequency features. Wu et al. [30] proposed a joint space–frequency rotation-invariant channel feature which has good adaptability to any angle and significantly improves the performance of the model.

The above exploration work shows the efficiency of frequency domain learning in the field of vision. Inspired by the above research work, we designed an adaptive frequency selection module for the task of oriented object detection in remote sensing images, aiming

to improve the network's ability to capture rich and detailed information of the object to improve detection performance.

2.3. Feature Fusion Network

After the basic feature extraction, the input image underwent certain feature fusion operations in the neck network to fully fuse high-level semantic information and underlying detail features. Considering the characteristics of the object in the remote sensing image, the researchers have made relevant improvements to the feature fusion network from the aspects of feature enhancement and efficient feature fusion. Li et al. [31] designed a feature enhancement module in the feature pyramid network which effectively reduced the aliasing effect in the feature fusion process through nearest neighbor interpolation and convolution operations. The method also designed a multi-layer mixed attention module which established the spatial position dependency by stacking the channel attention and position attention modules. Zhen et al. [32] proposed an adaptive multi-level feature fusion network which constructed a multi-scale feature fusion module at the top level of the feature pyramid and used multi-scale information to compensate for the loss of semantic information of small objects. Sun et al. [33] designed a feedback connection method in the feature fusion network for the efficient fusion of different feature layers, enhanced the representation ability of multi-scale features, and reduced the loss of high-level semantic information during feature fusion. The above studies fully show that object feature enhancement and efficient fusion of different scale features are important measures for feature fusion. However, enhancement and fusion only through spatial domain features cannot stimulate the expression of all the feature information of the object, resulting in the loss of important object information. Based on this, we introduced a spatial-frequency domain feature interaction enhancement module to fuse the feature information between the two different domains, thereby making up for the deficiency of single domain information.

3. Methods

In this section, we propose an efficient AD^{3}I-Net structure for oriented object detection in remote sensing images. Its main idea is to make full use of the rich feature information in the spatial domain and frequency domain to improve the object feature expression. The specific method we propose will be described in detail in the following subsections, mainly including the overall network structure, the spatial adaptive selection module, the frequency adaptive selection module, and the dual-domain feature interaction module.

3.1. Overall Structure

In this section, we will introduce the overall network framework of AD3I-Net and its important components in detail. As shown in Figure 2, we integrated the core method proposed in this paper into a common detection model to improve the performance of the network in the task of oriented object detection in remote sensing images. Similarly, this model has three important components: backbone, neck, and head. Among them, the backbone network is responsible for extracting basic features, the neck network is mainly used to further process and optimize features and finally outputs the prediction results of object type and position through the detection head. The neck network is usually located between the backbone network and the detection head to play a bridge role. The design of its structure affects the fusion and transmission mode of the features, which plays an important role in detection performance. Therefore, we introduced a modeling method of spatial-frequency domain interactive enhancement in the neck, dynamically modeling spatial and frequency domain information of different scales, so that the network could

adaptively select appropriate features. Then, the spatial domain and frequency domain features were interactively fused to fully explore the complementary information between the two domains and to enhance the diversity and robustness of object features. Different from the previous single spatial domain modeling method, we also took frequency domain information into consideration and used frequency domain information to better model directional change information. Specifically, the main process of the algorithm was as follows: First, four feature maps of different scales were obtained through the backbone network, namely [12], whose sizes were {1/4, 1/8, 1/16, 1/32} of the input image size; then, the obtained feature maps were input into three important spatial–frequency domain interaction enhancement components to enrich the feature information of the object, and a feature pyramid was generated through a top-down path, namely {P2, P3, P4, P5}; finally, the fused and enhanced features were passed through the detection head to achieve high-precision oriented object detection. In particular, the core method proposed in this paper can be integrated into most detection models to improve the performance of remote sensing image object detection algorithms.

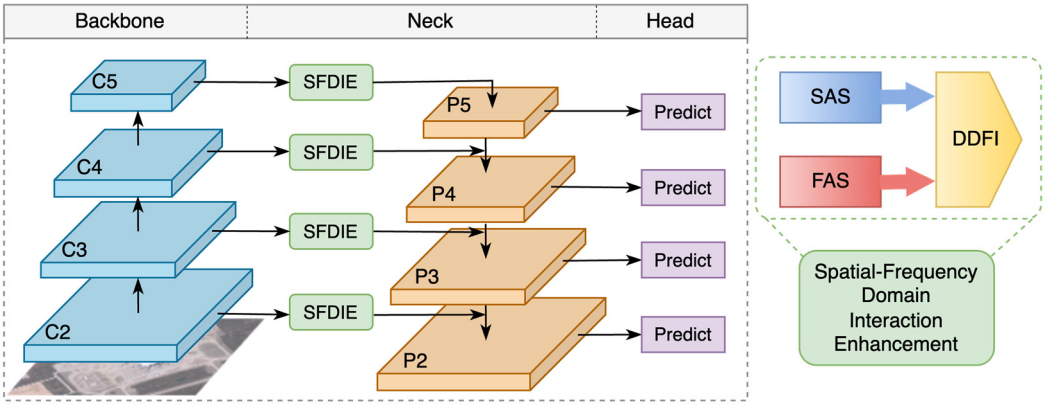


Figure 2. The overall structure of AD^{3}I-Net. The model is mainly composed of three parts: backbone, neck, and head. The core idea we proposed is the spatial-frequency domain interaction enhancement method in the Neck structure, which mainly includes three core components: the spatial adaptive selection (SAS) module, the frequency adaptive selection (FAS) module, and the dual-domain feature interaction (DDFI) module.

3.2. Spatial Adaptive Selection Module

In this paper, we proposed a spatial adaptive selection module to dynamically extract contextual information at the spatial domain level so that the network can adaptively select the receptive field size that is conducive to object recognition for objects of different scales. As shown in the network structure in Figure 3, we used multiple DW Conv with different kernel sizes to obtain multi-scale features in the spatial domain and filter the spatial contextual information that best matched the object through a selection mechanism. Specifically, we used a series of DW Conv with kernel sizes of {3, 5, 7, 9, 11} to explicitly aggregate object features under different receptive fields. In this way, the receptive field size can be gradually expanded to obtain contextual information from different perspectives. Assuming the input feature is $X \epsilon \mathbb{R}^{C \times H \times W}$, the process is as follows:

$$X_1^s = DWConv_3(X) \tag{1}$$

$$X_2^s = DWConv_5(X_1^s) \oplus DWConv_7(X_1^s) \oplus DWConv_9(X_1^s) \oplus DWConv_{11}(X_1^s) \tag{2}$$

where $DWConv_k$ represents the depth-wise separable convolution with kernel size k, and $\oplus$ represents element-wise addition.

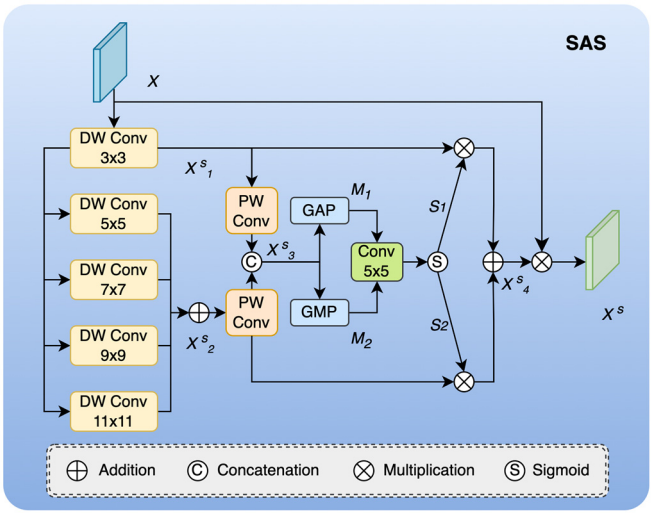


Figure 3. Network structure of the spatial adaptive selection module. X represents the input feature map, X^s represents the output feature map, and the others are intermediate processing features. *GAP* represents global average pooling, *GMP* represents global max pooling, DW Conv represents depth-wise separable convolution, PW Conv represents point-wise convolution, and Conv is a normal 2D convolution.

Then, we used PW Conv to reduce the number of feature channels to reduce complexity. After concatenating the feature map X_1^s of the small receptive field and the feature map X_2^s of the large receptive field, we effectively modeled the features along the channel dimension by applying average pooling and max pooling operations. The process is as follows:

$$X_3^s = Concat(PWConv(X_1^s); PWConv(X_2^s)) \tag{3}$$

$$M_1 = GAP(X_3^s) \tag{4}$$

$$M_2 = GMP(X_3^s) \tag{5}$$

For the modeling features obtained above, we spliced them together to allow interactive mixing of description information in different spaces. Then, we used 2D convolution to convert them into spatial attention maps, and we used the Sigmoid activation function to obtain dynamically selected spatial mask information. The specific process is as follows:

$$M = Concat(M_1; M_2) \tag{6}$$

$$[S_1, S_2] = Sigmoid(Conv_5(M)) \tag{7}$$

The spatial mask information obtained was used to adaptively select feature maps from different receptive field sizes, thereby dynamically optimizing the network receptive field according to the object characteristics. The adaptive selection operation used a weighted approach to calibrate the feature map. The specific process is as follows:

$$X^s = X \cdot ((S_1 \otimes X_1^s) \oplus (S_2 \otimes X_2^s)) \tag{8}$$

where $\otimes$ represents element-wise multiplication, and $\oplus$ represents element-wise addition.

The above are the specific steps to implement the spatial adaptive selection module. The SAS module we proposed uses the spatial selection mechanism to select the most refined spatial feature map that best meets the object characteristics from different receptive fields so that the network can focus more on the most effective spatial context information and enhance the network's semantic extraction ability.

3.3. Frequency Adaptive Selection Module

As an indispensable mathematical tool in frequency domain analysis, the Fourier transform usually plays an important role in the field of image processing. Fourier transform converts images from the time domain to the frequency domain, and it filters, enhances, and compresses the frequency domain information of the image by analyzing the different frequency components of the image. In practical image processing applications, two-dimensional discrete Fourier transform is often used to extract image features such as texture, edge, and noise. At present, the existing methods for oriented object detection in remote sensing images often only model the object characteristics in the spatial domain, which greatly limits the possibility of mining more object feature information. To this end, we proposed a frequency adaptive selection module, based on convolutional neural networks, which aimed to explore the natural characteristics of the object from the frequency domain and extract more robust image feature representations, thereby improving the network's perception of object direction diversity.

As shown in Figure 4, the frequency adaptive selection module is mainly composed of a 2D fast Fourier transform, an adaptive filter, and a 2D inverse Fourier transform. Specifically, the image features were converted to the frequency domain through fast Fourier transform, the adaptive filter generated by the dynamic learning of the neural network was used for feature extraction, and they were then converted back to the spatial domain through inverse fast Fourier transform. The proposed FAS module enhanced the feature expression ability of the object to a certain extent, improved the network's modeling ability of the object's directional diversity, and thus improved the accuracy of oriented object detection in remote sensing images.

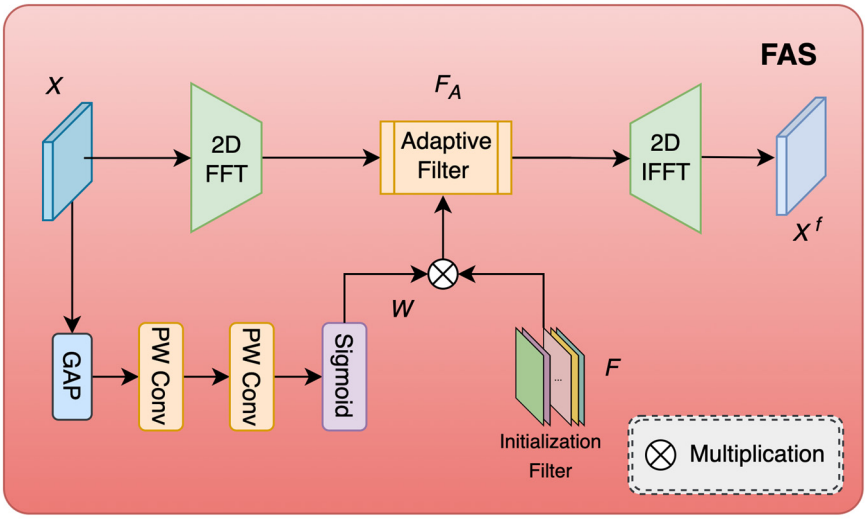


Figure 4. Network structure diagram of the frequency domain adaptive selection module. X represents the input feature map, X^f represents the output feature map, F_A represents the adaptive filter, F represents the initialized static filter, and W represents the dynamic weight parameter.

First, the input feature $X \epsilon \mathbb{R}^{C \times H \times W}$ passed through two branches, one for feature processing converted to the frequency domain, and the other for the generation of frequency domain adaptive filters. The adaptive filter was dynamically learned by the neural network and used for global filtering of image features. Specifically, it was constructed of an initialized static filter and dynamic weight parameters obtained by network learning. The dynamic weight parameters were determined by a series of operations such as pooling and convolution. We define the dynamic weight parameter W as follows:

$$W = Sigmoid(PWConv(PWConv(GAP(X)))) \quad (9)$$

Therefore, using the dynamic weight parameter W obtained above and the initialized static filter F, it was easy to obtain an adaptive filter F_A with dynamic adaptability, which is specifically defined as follows:

$$F_A = F \otimes W \tag{10}$$

The frequency domain information was globally filtered using the adaptive filter obtained above, thereby extracting the detailed features of the image. Specifically, the input spatial domain features were transformed into the frequency domain through the 2D fast Fourier transform algorithm, filtered with the adaptive filter, and then the 2D inverse Fourier transform algorithm was used to convert the frequency domain features back to the spatial domain features, thereby improving the perception of the global and local details of the object. We define this process as follows:

$$X^f = IFFT(F_A \odot FFT(X)) \tag{11}$$

where X^f represents the output feature, FFT represents the 2D fast Fourier transform, IFFT represents the 2D inverse Fourier transform, and $\odot$ represents the element-wise product.

The fast Fourier transform mentioned above is a practical and efficient algorithm for the discrete Fourier transform, which uses the symmetry and periodicity of the data to reduce the necessary amount of calculation and reduce the time complexity. The mathematical expression of the discrete Fourier transform is as follows:

$$X\left[h', w'\right] = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x[h, w] e^{-j2\pi\left(\frac{h'h}{H} + \frac{w'w}{W}\right)} \tag{12}$$

where $X \epsilon \mathbb{R}^{C \times H \times W}$ represents the input image feature, H and W represent the height and width of the feature, respectively, and h and w represent the coordinates on the feature.

Similarly, we could easily obtain spatial domain features by discrete inverse Fourier transform from frequency domain features. The mathematical expression is as follows:

$$x[h, w] = \frac{1}{HW} \sum_{h'=0}^{H-1} \sum_{w'=0}^{W-1} X\left[h', w'\right] e^{-j2\pi\left(\frac{h'h}{H} + \frac{w'w}{W}\right)} \tag{13}$$

The above are the specific steps to implement the frequency adaptive selection module. The FAS module we proposed used an adaptive filter to dynamically filter the frequency domain features, better perceive and model the direction of the object, and effectively improve the network's ability to extract object detail features.

3.4. Dual-Domain Feature Interaction Module

Through the spatial adaptive selection module and frequency adaptive selection module introduced in the above two sections, we captured the feature information of different aspects and attributes of the object in the spatial domain and the frequency domain, respectively. In order to make full use of the semantic complementarity of the features between the two domains, we designed a dual-domain feature interaction module to better promote their fusion and enhance the semantic integrity of the object. As shown in Figure 5, first, the spatial domain features and frequency domain features were mapped to a unified scale as the input features of the next stage. Then, the features between the two domains were semantically fused in a manner similar to the attention mechanism. Specifically, by calculating the key value of the query itself and the attention between the corresponding objects, the features were weighted so as to achieve interactive alignment of the features between the two domains and bridge the loss of semantic information. Finally, the channel attention mechanism was used to eliminate redundant information so that the

feature elements were better focused and the interference caused by useless information was avoided.

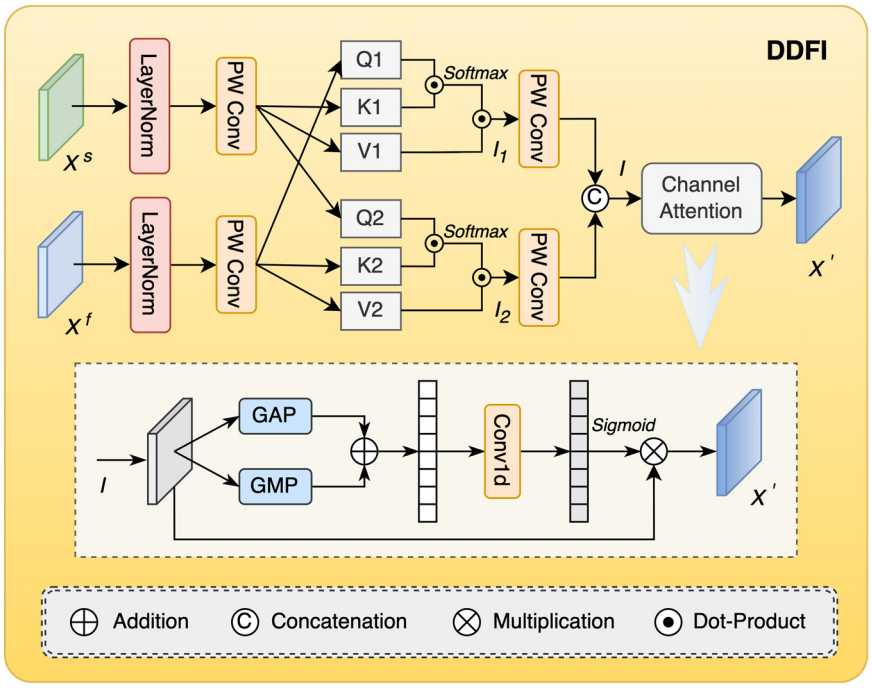


Figure 5. Network structure diagram of the dual-domain feature interaction module. X^s represents the input spatial domain feature map, X^f represents the input frequency domain feature map, I represents the fused feature, and X' is the final output feature.

For the obtained spatial domain features $X^s \epsilon \mathbb{R}^{C \times H \times W}$ and frequency domain features $X^f \epsilon \mathbb{R}^{C \times H \times W}$, we input them into the dual-domain feature interaction module to obtain the query object and key-value pair, respectively. The specific definitions are as follows:

$$[Q, K, V] = PWConv(LN(x)) \tag{14}$$

where x represents any feature map of the input, LN represents the layer normalization operation, and Q, K, V represents the output matrix, where Q is the query, K is the key, and V is the value.

In the DDFI module, the input spatial domain feature X^s and frequency domain feature X^f are operated by Formula (14) to obtain two sets of matrices, which are defined as follows:

$$[Q2, K1, V1] = PWConv(LN(X^s)) \tag{15}$$

$$[Q1, K2, V2] = PWConv\left(LN\left(X^f\right)\right) \tag{16}$$

Then, based on the obtained $[Q, K, V]$, weighted fusion between features was performed through dot product attention, and its expression is as follows:

$$Attn(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{17}$$

where $d = C \times H \times W$. The corresponding results are as follows:

$$I_1 = Attn(Q1, K1, V1) \tag{18}$$

$$I_2 = Attn(Q2, K2, V2) \tag{19}$$

Next, the above two elements were first subjected to PW Conv to adjust the number of channels and then concatenated to obtain the fused feature I, which is defined as follows:

$$I = Concat(PWConv(I_1); PWConv(I_2)) \tag{20}$$

where I_1, $I_2 \epsilon \mathbb{R}^{C/2 \times H \times W}$, and $I \epsilon \mathbb{R}^{C \times H \times W}$.

So far, we obtained the features after the fusion of the spatial domain features and the frequency domain features, which complemented the semantic loss in a single domain, maximized the mining of effective information in the image, and enriched the feature expression of the object. Finally, we introduced the channel attention module to ensure that while complementing the semantic information, the redundant features caused by feature consistency and the invalid features caused by non-focused objects were reduced. The channel attention mechanism is a method that focuses on the importance of each channel in the image feature map. Usually, in convolutional neural networks, each channel of the feature map represents a specific feature information, and different channels play different roles in various tasks. The channel attention module aims to dynamically assign different weights according to the contribution of each channel of the feature map in the specified task, thereby suppressing irrelevant or redundant feature channels, allowing the network to focus on the most critical features for the task, and ultimately improving the overall performance of the model. It mainly includes three main steps: feature conversion, attention weight calculation, and feature reweighting.

Specifically, we first transformed the above fused feature map and used global average pooling and global max pooling to obtain the fusion information of global context information and local salient information in each channel. Next, the channel weight was calculated by a one-dimensional convolution with an adaptively selected kernel size, and the normalized mapping was performed through the Sigmoid activation function. Finally, the reshaped weight value was multiplied by the original feature map to obtain the reweighted feature representation. Through this operation, the important channel features were given attention while the unimportant channel features were suppressed. The channel attention mechanism can effectively capture the interaction between channels and enhance the model's perception of key features. The channel attention calculation formula is as follows:

$$X' = I \otimes Sigmoid(Conv1d_k(GAP(I) \oplus GMP(I))) \tag{21}$$

where $Conv1d_k$ represents a one-dimensional convolution with a kernel size of k, and X' is the output feature.

3.5. Loss Function

The loss function mainly includes regression loss and classification loss. Among them, regression loss is used to guide the network to learn the location and size of the target, and classification loss is used to learn the target category. The total loss function is as follows:

$$Loss == \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg} + \frac{1}{N_{cls}} \sum_i L_{cls} \tag{22}$$

where N_{reg} and N_{cls} represent the number of positive samples in the two stages; p_i^* is an index which is 1 when the proposal is a positive sample and is otherwise 0; and λ is a balancing factor, which is 1 here. The regression loss L_{reg} uses SmoothL1 Loss, and the classification loss L_{cls} uses CrossEntropy Loss.

4. Experimental Results and Discussions

In order to verify the effectiveness of our proposed method, we conducted a large number of experiments to verify it. In this section, we mainly introduce the datasets, experimental details, evaluation indicators, and experimental results used in this experiment. The detailed experimental content is as follows.

4.1. Datasets

To ensure the fairness of the method comparison, we used two public remote sensing oriented object detection datasets, HRSC2016 [34] and DIOR-R [12], for experimental verification. The details of the two datasets are as follows.

4.1.1. HRSC2016 Dataset

The HRSC2016 dataset is a public dataset for ship detection in optical remote sensing images released by Liu et al. The dataset contains 1061 images and 2976 object instances. The remote sensing images in the dataset are all from the Google Earth platform, with image sizes ranging from 300 pixels × 300 pixels to 1500 pixels × 900 pixels and a spatial resolution between 0.4 m and 2 m, include images of ships at sea and near the coast

4.1.2. DIOR-R Dataset

The DIOR-R dataset is a large-scale optical remote sensing image rotation object detection dataset released by Cheng et al. All data come from the Google Earth platform. The dataset has a total of 23,463 images, the images are 800 pixels × 800 pixels, and the spatial resolution is between 0.5 m and 30 m. It has 192,518 instances, the instance size varies widely, and there are 20 categories. The images in this dataset show different imaging effects due to different imaging conditions, different weather conditions, and seasonal differences, which greatly enhance the diversity of the data.

4.2. Implementation Details

To ensure the consistency of the experiment, all the training and validation experiments were conducted on a single NVIDIA RTX3090 graphics card with a batch size of 2. The platform was built on the Ubuntu operating system and used the PyTorch 1.13.0 deep learning framework. In the experiment, we used the SGD optimizer with the initial learning rate set to 0.005, the momentum set to 0.9, and the weight decay parameter set to 0.0001. In the actual training and testing process of the model, the image size on the HRSC2016 dataset was uniformly adjusted to 800 × 512. It was trained for 36 epochs in total, and the learning rate was divided by 10 at the 24th and 33rd epochs. Among them, 617 images were used for training and 444 images were used for testing. The image size on the DIOR-R dataset was uniformly adjusted to 800 × 800, and it was trained for 12 epochs in total. The learning rate was divided by 10 at the 8th and 11th epochs. According to the official allocation, 11,725 images were used as the training set and 11,738 images were used as the test set.

Following the mainstream evaluation method of the rotation object detection model, we used the mean Average Precision (mAP) evaluation index to evaluate the performance. The Pascal VOC evaluation index usually includes two evaluation forms: VOC07 and VOC12. Among them, the VOC07 evaluation standard is calculated by 11-point interpolation, while the VOC12 evaluation standard is calculated by the area under the PR curve. In this experiment, the DIOR-R dataset used the VOC07 evaluation standard, and the HRSC2016 dataset used the VOC07 and VOC12 evaluation standards.

4.3. Ablation Study

In this section, we conducted ablation experiments on the DIOR-R dataset to verify the effectiveness of the proposed method in our model. The specific results of the ablation experiment are shown in Table 1. Compared with the HRSC2016 dataset, the DIOR-R dataset is more diverse and can better evaluate the actual effects of the different modules in this method. The baseline model uses Rotated Faster R-CNN, and the backbone network is ResNet50. In the ablation experiment, we compared the baseline and three other different combinations of modules to show their impact on model performance. The baseline model in the table has a mAP index of 63.42 when no components are added. When we added the SAS module alone, its mAP index increased by 0.78, indicating that the spatial adaptive selection method can effectively increase the model's adaptability to objects of different scales. When we added the FAS module alone, its mAP index increased by 1.15, proving that the frequency adaptive selection method can effectively improve the model's ability to model object features, thereby significantly improving the performance of the model. Finally, when we added the SAS module, the FAS module, and the DDFI module to the baseline method, its mAP index was improved by 2.23. From the results, it can be seen that the method proposed by us can effectively interactively fuse the spatial and frequency domain information to enhance the model's learning ability of the object features in the remote sensing image, thereby improving the performance of oriented object detection in remote sensing images. While improving the detection accuracy of the model, the addition of multiple modules also brought additional computational overhead, and the model's Params and FLOPs increased. However, these overheads are worthwhile for the improvement in the accuracy of the detection of the directional object. If the application is under resource-constrained conditions, the detection framework needs to be optimized to adapt.

Table 1. Ablation experimental results of the proposed method on the DIOR-R dataset.

Method	SAS	FAS	DDFI	mAP_{50}	Params(M)	FLOPs(G)
Baseline	-	-	-	63.42	31.95	126.71
AD^3I-Net	✓	-	-	64.20	32.81	129.35
	-	✓	-	64.57	32.34	128.08
	✓	✓	✓	65.65	33.42	134.26

In addition, we also made a visual comparison of the features extracted by the model. From Figure 6, it can be seen that after applying the three modules proposed in this paper, the model had a more refined extraction of the object position and an enhanced ability to perceive the direction. Through the analysis of qualitative and quantitative experiments, it can be proved that the method proposed in this paper has good performance in the oriented object detection task.

4.4. Experimental Results and Discussion

4.4.1. Results on the HRSC2016 Dataset

We compared the proposed method with some advanced algorithms on the HRSC2016 dataset. As shown in Table 2, these quantitative results demonstrate the effectiveness of our dual-domain feature interaction method in remote sensing-oriented object detection. Our proposed method achieved 90.70 and 97.85 in the evaluation indicators of VOC07 and VOC12, respectively. RTMDet-R uses the CSPNeXt backbone network pre-trained on the COCO dataset and achieved 90.60 on the VOC07 indicator and 97.10 on the VOC07 indicator. In comparison, our method surpassed RTMDet-R with mAP values higher by 0.10 and 0.75. Compared with the CGCDet method that also uses the ResNet50 backbone

network, although we were 0.01 lower than this method in the VOC12 evaluation indicator, we were 0.13 higher in the VOC07 indicator. From the results in the table, it can be seen that our proposed method is superior to most mainstream advanced methods and has certain advantages. We also present some qualitative results in Figure 7. As can be seen from the figure, even under the interference of different scales and cluttered backgrounds, our method could accurately detect the object and accurately locate the direction of the object.

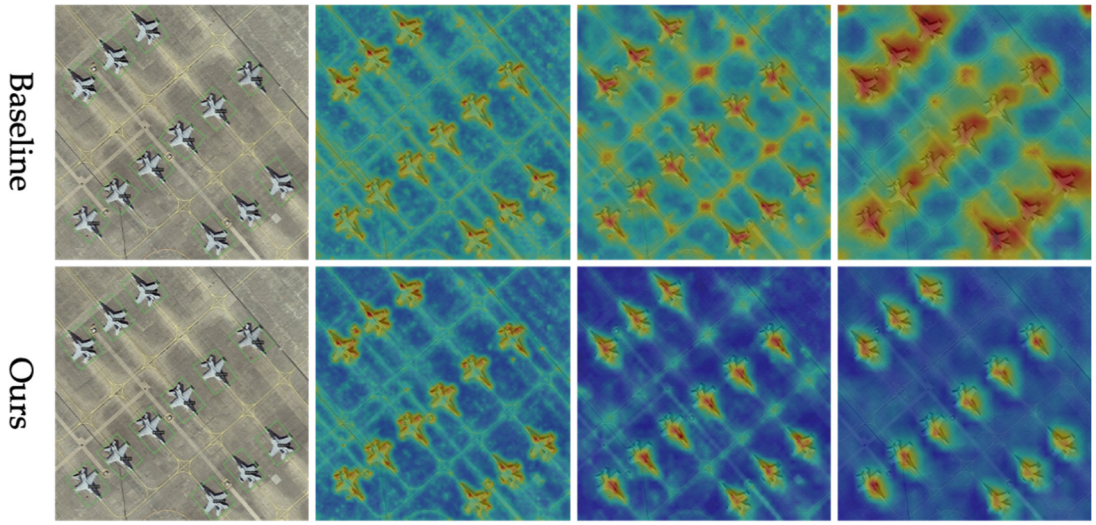


Figure 6. Feature visualization comparison of the baseline and our proposed method. The first column of images on the left is the detection result and the others are feature visualization effects at different stages.

Table 2. Results of the state-of-the-art methods compared on the HRSC2016 dataset. IN represents the backbone network pre-trained on the ImageNet dataset, and CO represents the backbone network pre-trained on the COCO dataset. mAP (07/12) represents the VOC2007 or VOC2012 evaluation index.

Method	Pretrained	Backbone	mAP(07)	mAP(12)
RetinaNet-O [35]	IN	Res50	73.42	77.83
RRPN [36]	IN	Res101	79.08	85.64
DRN [37]	IN	Houglass34	-	92.70
CenterMap [38]	IN	Res50	-	92.80
RoI Trans. [39]	IN	Res101	86.20	-
CFA [40]	IN	Res50	87.10	91.60
SASM [41]	IN	Res50	87.90	91.80
AO2-DETR [16]	IN	Res50	88.12	97.47
Gliding Vertex [42]	IN	Res101	88.20	-
R-DINO [43]	IN	Res50	88.80	95.24
PIoU [44]	IN	DLA34	89.20	-
R^3Det [45]	IN	Res101	89.26	96.01
DAL [46]	IN	Res101	89.77	-
GWD [5]	IN	Res50	89.85	97.37
S^2Anet [47]	IN	Res101	90.17	95.01
DODet [48]	IN	Res50	90.18	95.84
AOPG [12]	IN	Res50	90.34	96.22
Oriented R-CNN [10]	IN	Res101	90.40	96.50
ReDet [9]	IN	ReRes50	90.46	97.63
QPDet [49]	IN	Res101	90.52	96.64
CGCDet [50]	IN	Res50	90.57	97.86
RTMDet-R [51]	CO	CSPNeXt	90.60	97.10
AD^3I-Net (ours)	IN	Res50	90.70	97.85

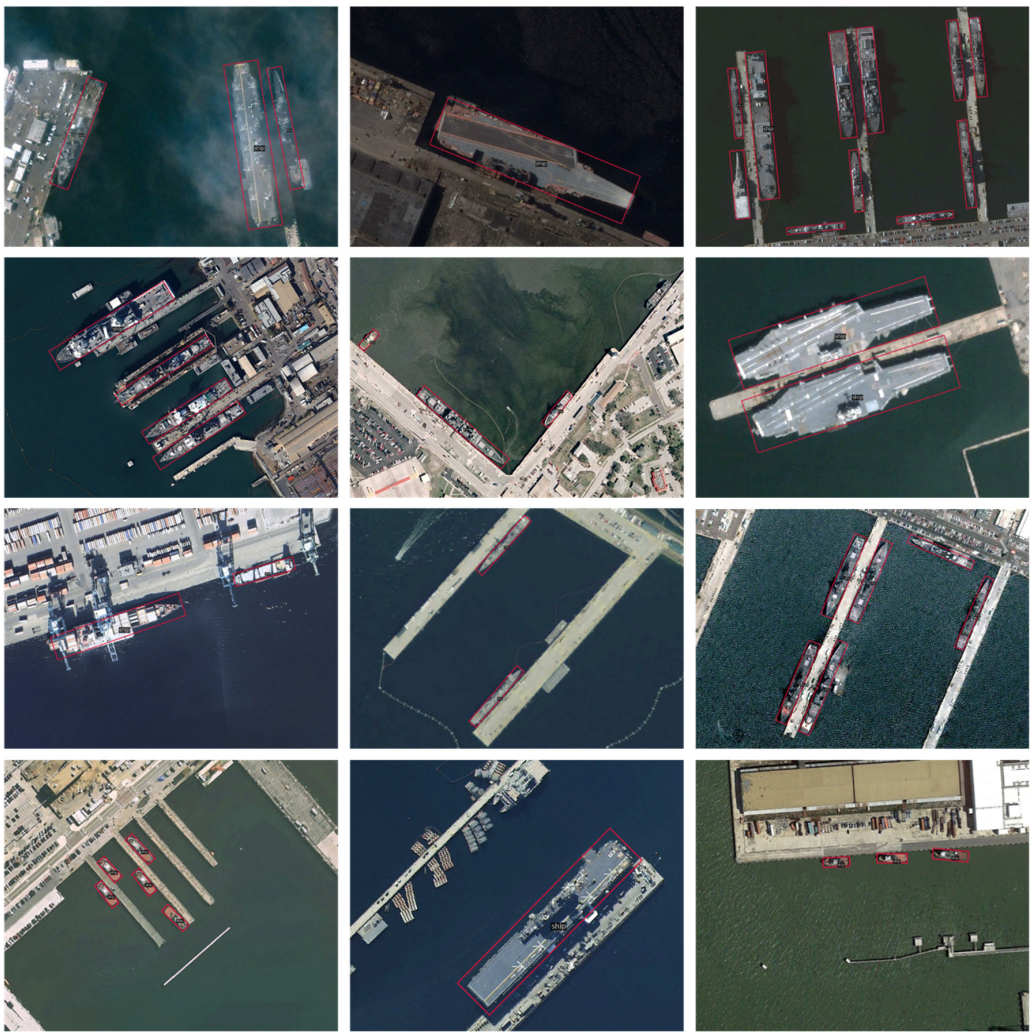


Figure 7. Prediction results of our proposed method on the HRSC2016 dataset. The red box in the figure shows the detection results.

4.4.2. Results on the DIOR-R Dataset

In Table 3, we show the performance of current advanced methods on the DIOR-R dataset. In contrast, the mAP index of our proposed method reached 65.65, which has a significant advantage. This index exceeds the previous best performer, CGCDet, at 64.88. Our method exceeded the performance indicators of all methods in the table in the three categories of BC, BF, and TS. Among them, the BC category reached 89.46, which is 0.91 higher than the QPDet method, which had the highest performance in this category (88.55). The TS category reached 63.79 in our method, which is 5.7 higher than the QPDet method, which previously had the highest performance in this category. In the BF category, we reached 80.40, which is 1.28 higher than the CGCDet method, which previously had the highest performance. As can be seen from the table, although our method did not achieve the best performance in other categories, our performance was very competitive overall. The detection effect of our method on the DIOR-R dataset is shown in Figure 8. These results prove the effectiveness of our proposed method. For the 20 common categories in remote sensing scenes, our method could accurately predict the location of the object category in different scenes and at different scales, thereby generating accurate oriented bounding box representations.

Table 3. Results of the state-of-the-art methods compared on the DIOR-R dataset.

Method	APL	APO	BC	BF	BR	CH	DAM	ESA	ETS	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP_{50}
RetinaNet-O [35]	59.54	25.03	81.01	70.08	28.26	72.02	21.26	55.35	56.77	65.70	70.28	30.52	44.37	77.02	59.01	59.39	81.18	38.43	39.10	61.58	54.83
SASM [41]	61.41	46.03	82.04	73.22	29.41	71.03	30.63	69.22	53.91	70.04	77.02	39.33	47.51	78.62	66.14	62.92	79.93	54.41	40.62	63.01	59.81
GWD [5]	69.68	28.83	81.49	74.32	29.62	72.67	27.13	76.45	63.14	77.19	78.94	39.11	42.18	79.10	70.41	58.69	81.52	47.78	44.47	62.63	60.31
R^3Det [45]	62.55	43.44	81.48	71.72	36.49	72.63	27.02	79.5	64.41	77.36	77.17	40.53	53.33	79.66	69.22	61.10	81.54	52.18	43.57	64.13	61.91
Gliding Vertex [42]	62.67	38.56	81.20	71.94	37.73	72.48	22.81	78.62	69.04	77.89	82.13	46.22	54.76	81.03	74.88	62.54	81.41	54.25	43.22	65.13	62.91
R-DINO [43]	44.50	52.70	80.60	71.00	44.40	73.00	29.20	72.50	83.10	72.40	76.50	43.50	55.30	80.70	61.80	69.60	81.20	58.00	51.70	61.20	63.10
Rotated FCOS [52]	62.31	42.18	81.32	75.34	39.26	74.89	26.00	77.42	68.67	73.94	78.73	41.28	54.19	80.61	66.92	69.17	87.20	52.31	47.08	65.21	63.21
Rotated ATSS [53]	62.19	44.63	81.42	71.55	41.08	72.37	30.56	78.54	67.50	75.69	79.11	42.77	56.31	80.92	67.78	69.24	81.62	55.45	47.79	64.10	63.52
ReDet [9]	63.22	44.18	81.26	72.11	43.83	72.72	28.45	79.10	69.78	78.69	77.18	48.24	56.81	81.17	69.17	62.73	81.42	54.90	44.04	66.37	63.81
RoI Trans. [39]	63.18	44.33	81.26	71.91	42.19	72.64	29.42	79.30	69.67	77.33	82.88	48.09	57.03	81.18	77.32	62.45	81.38	54.34	43.91	66.30	64.31
QPDet [49]	63.22	41.39	88.55	71.97	41.23	72.63	69.00	28.82	78.90	70.07	83.01	47.83	55.54	81.23	72.15	62.66	89.05	58.09	43.38	65.36	64.20
S^2ANet [47]	67.98	44.44	81.39	71.63	42.66	72.72	27.08	79.03	70.40	75.56	81.02	43.41	56.45	81.12	68.00	70.03	87.07	53.88	51.12	65.31	64.50
Oriented RCNN [10]	63.31	43.10	81.17	71.89	44.78	72.64	33.78	80.12	69.67	77.92	83.11	46.29	58.31	81.17	74.54	62.32	81.29	56.30	43.78	65.26	64.53
KLD [54]	66.52	46.80	81.43	71.76	40.81	78.25	29.01	79.23	66.63	78.68	80.19	44.88	57.23	80.91	74.17	68.02	81.48	54.63	47.80	64.41	64.63
RepPoints [14]	63.48	51.20	86.55	69.68	42.92	75.09	31.82	74.11	68.46	77.52	76.54	41.76	56.67	87.62	64.42	71.79	81.61	55.83	52.79	66.18	64.80
CGCDet [50]	68.46	38.34	86.21	79.12	38.97	73.52	26.84	74.72	66.00	67.49	84.45	48.02	56.05	81.25	79.32	72.17	88.56	50.69	51.72	65.84	64.88
AD^3I-Net (ours)	63.20	42.50	89.46	80.40	42.59	72.58	30.31	79.97	68.12	77.83	82.66	46.58	56.95	80.65	73.45	70.84	81.58	63.79	43.84	65.61	65.65

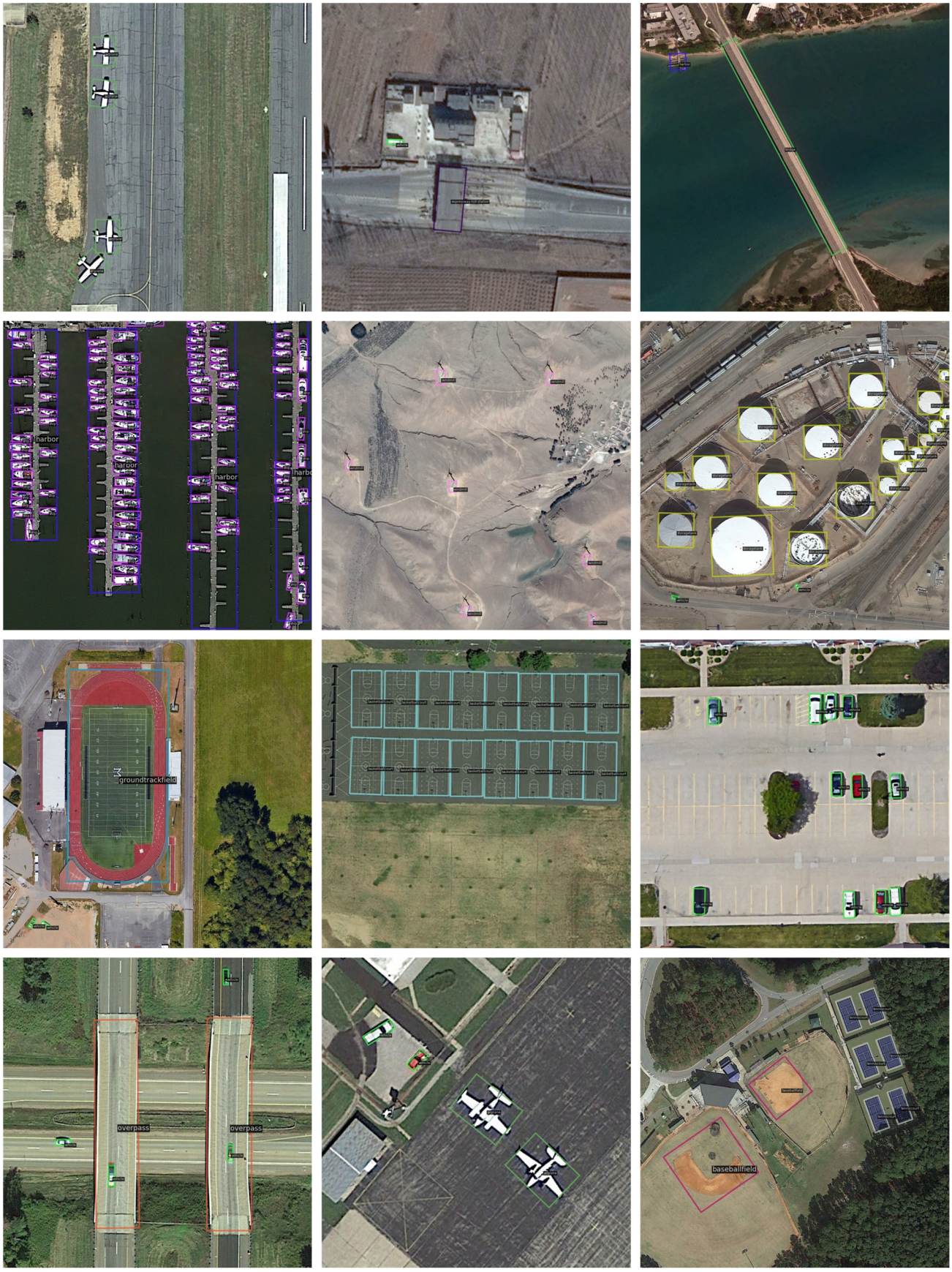


Figure 8. Prediction results of our proposed method on the DIOR-R dataset.

5. Conclusions

In this study, we explored the feasibility of frequency learning for oriented object detection in remote sensing images, and based on this, we proposed an adaptive dual-domain dynamic interaction network to cope with the challenges brought by complex backgrounds, different object scales, and changing directions in remote sensing images. We proposed three core components, namely the spatial adaptive selection (SAS) module, the frequency adaptive selection (FAS) module and the dual-domain feature interaction (DDFI) module. Among them, the SAS module uses the dynamic receptive field to adaptively learn the contextual information of the object and extract more accurate spatial position features for objects of different scales. The FAS module makes up for the shortcomings of learning in a single domain, using the rich information in the frequency domain to extract the direction features of the object and improve the network's ability to model the direction. Finally, we proposed a DDFI module to efficiently fuse the features of the above two modules and effectively enhance the features through the consistency and complementarity of information. This method fully exploits the spatial and directional information between the two domains, improves the network's ability to model object features, and thus improves the performance of oriented object detection in remote sensing images. The experiments on two mainstream remote sensing image object detection task datasets have demonstrated the effectiveness of our proposed method and its competitiveness. In future research, we will continue to expand and optimize the model to enable its application in tasks in more complex remote sensing scenarios.

Author Contributions: Conceptualization, Y.Z. and H.S. (Haijiang Sun); methodology, Y.Z.; software, S.W. and H.S. (Hailin Su); validation, Y.Z., H.S. (Hailin Su) and S.W.; formal analysis, Y.Z.; investigation, Y.Z., H.S. and T.Y.; resources, H.S. (Haijiang Sun) and T.Y.; data curation, S.W. and H.S. (Hailin Su); writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.; visualization, Y.Z. and H.S. (Hailin Su); supervision, H.S. (Haijiang Sun) and T.Y.; project administration, H.S. (Haijiang Sun) and T.Y.; funding acquisition, H.S. (Haijiang Sun). 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 number 62475255.

Data Availability Statement: Two publicly available datasets, HRSC2016 and DIOR-R, were used in this study. These data can be found at https://sites.google.com/site/hrsc2016/ (accessed on 24 June 2024) and https://gcheng-nwpu.github.io/ (accessed on 15 July 2024).

Acknowledgments: This research was supported by the Jilin Provincial Key Laboratory of Machine Vision Intelligent Equipment.

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

Abbreviations

The following abbreviations are used in this manuscript:

AD^{3}I-Net	Adaptive Dual-Domain Dynamic Interaction Network
SAS	Spatial Adaptive Selection
FAS	Frequency Adaptive Selection
DDFI	Dual-Domain Feature Interaction
RCNN	Region Convolutional Neural Network
RPN	Region Proposal Networks
mAP	Mean Average Precision
GAP	Global Average Pooling
GMP	Global Max Pooling

References

1. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 431–435. [CrossRef]
2. Chen, W.; Miao, S.; Wang, G.; Cheng, G. Recalibrating Features and Regression for Oriented Object Detection. *Remote Sens.* **2023**, *15*, 2134. [CrossRef]
3. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 16748–16759.
4. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part VIII 16, Glasgow, UK, 23–28 August 2020; pp. 677–694.
5. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021.
6. 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]
7. Li, M.; Guo, W.; Zhang, Z.; Yu, W.; Zhang, T. Rotated region based fully convolutional network for ship detection. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 673–676.
8. Yan, Z.; Song, X.; Zhong, H.; Zhu, X. Object detection in optical remote sensing images based on transfer learning convolutional neural networks. In Proceedings of the 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, 23–25 November 2018; pp. 935–942.
9. Han, J.; Ding, J.; Xue, N.; Xia, G.-S. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2786–2795.
10. 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, BC, Canada, 11–17 October 2021; pp. 3520–3529.
11. Dai, P.; Yao, S.; Li, Z.; Zhang, S.; Cao, X. ACE: Anchor-free corner evolution for real-time arbitrarily-oriented object detection. *IEEE Trans. Image Process.* **2022**, *31*, 4076–4089. [CrossRef] [PubMed]
12. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625411. [CrossRef]
13. Li, J.; Tian, Y.; Xu, Y.; Zhang, Z. Oriented object detection in remote sensing images with anchor-free oriented region proposal network. *Remote Sens.* **2022**, *14*, 1246. [CrossRef]
14. 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, 18–24 June 2022; pp. 1829–1838.
15. Ma, T.; Mao, M.; Zheng, H.; Gao, P.; Wang, X.; Han, S.; Ding, E.; Zhang, B.; Doermann, D. Oriented object detection with transformer. *arXiv* **2021**, arXiv:2106.03146.
16. Dai, L.; Liu, H.; Tang, H.; Wu, Z.; Song, P. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2342–2356. [CrossRef]
17. Zhou, Q.; Yu, C.; Wang, Z.; Wang, F. D^2Q-DETR: Decoupling and Dynamic Queries for Oriented Object Detection with Transformers. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
18. Wang, X.; Chen, H.; Chu, X.; Wang, P. AODet: Aerial Object Detection Using Transformers for Foreground Regions. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4106711. [CrossRef]
19. Ehrlich, M.; Davis, L.S. Deep residual learning in the jpeg transform domain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3484–3493.
20. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
21. Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; Zhou, J. Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 980–993.
22. Zhong, Y.; Li, B.; Tang, L.; Kuang, S.; Wu, S.; Ding, S. Detecting camouflaged object in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4504–4513.
23. Yang, Y.; Yuan, G.; Li, J. SFFNet: A Wavelet-Based Spatial and Frequency Domain Fusion Network for Remote Sensing Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3000617. [CrossRef]
24. Chen, L.; Gu, L.; Zheng, D.; Fu, Y. Frequency-Adaptive Dilated Convolution for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 3414–3425.
25. Finder, S.E.; Amoyal, R.; Treister, E.; Freifeld, O. Wavelet Convolutions for Large Receptive Fields. In Proceedings of the 18th European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 363–380.

26. Kong, L.; Dong, J.; Ge, J.; Li, M.; Pan, J. Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 5886–5895.
27. Patro, B.N.; Namboodiri, V.P.; Agneeswaran, V.S. SpectFormer: Frequency and Attention is what you need in a Vision Transformer. *arXiv* **2023**, arXiv:2304.06446.
28. Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; Mei, T. Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning. In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 328–345.
29. Cao, Y.; Wu, Y.; Li, M.; Liang, W.; Hu, X. DFAF-Net: A Dual-Frequency PolSAR Image Classification Network Based on Frequency-Aware Attention and Adaptive Feature Fusion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224318. [CrossRef]
30. Wu, X.; Hong, D.; Tian, J.; Chanussot, J.; Li, W.; Tao, R. ORSIm Detector: A Novel Object Detection Framework in Optical Remote Sensing Imagery Using Spatial-Frequency Channel Features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5146–5158. [CrossRef]
31. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
32. Zhen, P.; Wang, S.; Zhang, S.; Yan, X.; Wang, W.; Ji, Z.; Chen, H.-B. Towards Accurate Oriented Object Detection in Aerial Images with Adaptive Multi-level Feature Fusion. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 6. [CrossRef]
33. Sun, P.; Zheng, Y.; Zhou, Z.; Xu, W.; Ren, Q. R^4 Det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. *Image Vis. Comput.* **2020**, *103*, 104036. [CrossRef]
34. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
35. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
36. 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]
37. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, 14–19 June 2020; pp. 11204–11213.
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.* **2021**, *59*, 4307–4323. [CrossRef]
39. Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; Lu, Q.; Soc, I.C. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2844–2853.
40. Guo, Z.; Liu, C.; Zhang, X.; Jiao, J.; Ji, X.; Ye, Q.; Ieee Comp, S.O.C. Beyond Bounding-Box: Convex-hull Feature Adaptation for Oriented and Densely Packed Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 8788–8797.
41. Hou, L.; Lu, K.; Xue, J.; Li, Y.; Assoc Advancement Artificial, I. Shape-Adaptive Selection and Measurement for Oriented Object Detection. In Proceedings of the 36th AAAI Conference on Artificial Intelligence/34th Conference on Innovative Applications of Artificial Intelligence/12th Symposium on Educational Advances in Artificial Intelligence, Virtual, 22 February–1 March 2022; pp. 923–932.
42. 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.* **2021**, *43*, 1452–1459. [CrossRef] [PubMed]
43. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
44. Chen, Z.; Chen, K.; Lin, W.; See, J.; Yu, H.; Ke, Y.; Yang, C. Piou loss: Towards accurate oriented object detection in complex environments. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part V 16, Glasgow, UK, 23–28 August 2020; pp. 195–211.
45. Yang, X.; Yan, J.; Feng, Z.; He, T.; Assoc Advancement Artificial, I. R^3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the 35th AAAI Conference on Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence, Virtual, 2–9 February 2021; pp. 3163–3171.
46. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L.; Assoc Advancement Artificial, I. Dynamic Anchor Learning for Arbitrary-Oriented Object Detection. In Proceedings of the 35th AAAI Conference on Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence, Virtual, 2–9 February 2021; pp. 2355–2363.
47. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602511. [CrossRef]

48. Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; Han, J. Dual-Aligned Oriented Detector. *Ieee Trans. Geosci. Remote Sens.* **2022**, *60*, 5618111. [CrossRef]
49. Yao, Y.; Cheng, G.; Wang, G.; Li, S.; Zhou, P.; Xie, X.; Han, J. On Improving Bounding Box Representations for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5600111. [CrossRef]
50. Wang, Y.; Zhang, Z.; Xu, W.; Chen, L.; Wang, G.; Yan, L.; Zhong, S.; Zou, X. Learning Oriented Object Detection via Naive Geometric Computing. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 10513–10525. [CrossRef]
51. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. Rtmdet: An empirical study of designing real-time object detectors. *arXiv* **2022**, arXiv:2212.07784.
52. 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 (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635.
53. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 9756–9765.
54. Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; Yan, J. Learning High-Precision Bounding Box for Rotated Object Detection via Kullback-Leibler Divergence. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021.

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.