applied sciences

MDPI

Article

ReBiDet: An Enhanced Ship Detection Model Utilizing ReDet and Bi-Directional Feature Fusion

Zexin Yan [1], Zhongbo Li [1,*], Yongqiang Xie [1,*], Chengyang Li [1,2], Shaonan Li [1] and Fangwei Sun [1]

[1] Institute of System Engineering, Academy of Military Sciences, Beijing 100000, China; yanzexin@outlook.com (Z.Y.); chengyang_li@stu.pku.edu.cn (C.L.); sfw1208@outlook.com (F.S.)
[2] School of Computer Science, Peking University, Beijing 100871, China
* Correspondence: zbli2021@outlook.com (Z.L.); yqxie2021@outlook.com (Y.X.)

Abstract: To enhance ship detection accuracy in the presence of complex scenes and significant variations in object scales, this study introduces three enhancements to ReDet, resulting in a more powerful ship detection model called rotation-equivariant bidirectional feature fusion detector (ReBiDet). Firstly, the feature pyramid network (FPN) structure in ReDet is substituted with a rotation-equivariant bidirectional feature fusion feature pyramid network (ReBiFPN) to effectively capture and enrich multiscale feature information. Secondly, K-means clustering is utilized to group the aspect ratios of ground truth boxes in the dataset and adjust the anchor size settings accordingly. Lastly, the difficult positive reinforcement learning (DPRL) sampler is employed instead of the random sampler to address the scale imbalance issue between objects and backgrounds in the dataset, enabling the model to prioritize challenging positive examples. Through numerous experiments conducted on the HRSC2016 and DOTA remote sensing image datasets, the effectiveness of the proposed improvements in handling complex environments and small object detection tasks is validated. The ReBiDet model demonstrates state-of-the-art performance in remote sensing object detection tasks. Compared to the ReDet model and other advanced models, our ReBiDet achieves mAP improvements of 3.20, 0.42, and 1.16 on HRSC2016, DOTA-v1.0, and DOTA-v1.5, respectively, with only a slight increase of 0.82 million computational parameters.

Keywords: artificial intelligence; deep learning; remote sensing images; ship detection; bi-directional feature fusion; feature pyramid network; anchor size; K-means; sampler

check for updates

Citation: Yan, Z.; Li, Z.; Xie, Y.; Li, C.; Li, S.; Sun, F. ReBiDet: An Enhanced Ship Detection Model Utilizing ReDet and Bi-Directional Feature Fusion. *Appl. Sci.* **2023**, *13*, 7080. https://doi.org/10.3390/app13127080

Academic Editor: Luis Javier Garcia Villalba

Received: 14 April 2023
Revised: 7 June 2023
Accepted: 9 June 2023
Published: 13 June 2023

1. Introduction

With the advancement of space remote sensing technology, the observation capability of optical remote sensing satellites has significantly improved, leading to an increase in the spatial resolution of remote sensing images. Currently, the spatial resolution can reach 0.3 m per pixel [1], providing high-quality data for various scientific research applications. Computer vision technology plays a crucial role in extracting information from these images for tasks such as geological mapping, terrain measurement, and land cover change detection [2]. Object detection and recognition in remote sensing images are particularly important, especially in identifying ships, which hold strategic value in both economic and military domains. The accurate detection and recognition of ships in port areas and key waterways are vital for civilian and military purposes. However, the presence of complex backgrounds and variations in object scales in remote sensing images [3] pose significant challenges to achieving satisfactory ship detection accuracy.

1.1. Difficulties in Optical Remote Sensing Image Object Detection

Remote sensing images exhibit complexity, covering a wide range with large image scales. While they provide valuable visual features for detection models, they also introduce complex and irrelevant background information that hinders accurate detection. Firstly,

background interference arises from similar objects in the scene, leading to potential misjudgment. For example, small islands with similar colors and thin buildings on land can be mistakenly identified as ships. Secondly, self-interference occurs due to the diverse textures, similar colors, and wake features of ships, affecting the final recognition results. Thirdly, the random orientation of ships in remote sensing images poses another difficulty. Unlike natural scene images captured from a horizontal perspective, satellite images are taken from a bird's-eye view, resulting in ships appearing at any position with any orientation, making detection challenging.

Furthermore, vehicles, boats, helicopters, airplanes, and large structures such as stadiums appear much smaller in remote sensing images compared to natural scene images captured from a horizontal viewpoint. Consequently, existing visual models often struggle to achieve accurate detection for these objects. For instance, in the optical remote sensing image dataset DOTA [4], small objects, such as cars, occupy approximately 30×50 pixels, large trucks occupy around 35×90 pixels, and the smallest yacht of interest is merely about 35×60 pixels in size.

1.2. The Significant Advantages of Deep Learning Techniques

Traditional ship detection methods [5–8] based on artificial feature modeling rely on manually designed algorithms to extract shallow features. These algorithms often focus on specific scenes and exhibit limited generalization abilities. They still face challenges in terms of accuracy, computational efficiency, and robustness in complex environments, making them less applicable [9–11].

In recent years, significant progress has been made in the field of object detection in natural images, largely attributed to the rapid development of deep learning technology [12]. The introduction of deep learning models has greatly improved the accuracy of object localization and recognition in images. Convolutional neural networks (CNNs) have emerged as powerful tools for feature extraction, enabling automatic learning and extraction of features from images. The renowned Faster R-CNN model [13] has become a representative model in natural image object detection, opening possibilities for applying deep learning-based object detection models to remote sensing images.

1.3. Related Works

After years of research, several mature methods have been developed for ship detection. Early studies predominantly focused on ship detection methods based on artificial feature modeling. However, since 2012, deep learning-based detection methods have emerged as a transformative force in the field of ship detection. Recent years have witnessed the proposal of numerous methods for ship detection in remote sensing images. Leveraging the powerful feature expression capabilities of convolutional neural networks (CNNs), these deep learning-based object detection methods can extract higher-level semantic features of ships, delivering significantly improved performance in complex remote sensing scenarios compared to traditional approaches. Nevertheless, most of these methods originated from object detection models designed for natural images captured from horizontal viewpoints and are limited in their ability to detect objects with arbitrary orientations. Given that ships possess relatively large aspect ratios and can exhibit random orientations in images, employing horizontal detection boxes would result in significant background information being included in the detection boxes. This issue becomes particularly pronounced when detecting densely distributed ships in a harbor, where interference from neighboring objects leads to poor detection performance.

To address these challenges, researchers have proposed models suitable for detecting objects with arbitrary orientations. For instance, the rotational R-CNN (R2CNN) model [14], initially developed for scene text detection, has shown excellent performance in detecting slanted text and has been adapted for ship detection. Liu et al. [15] enhanced this approach by introducing a rotation region of interest (RoI) pooling layer and a rotation bounding box regression module based on R2CNN. This improvement enables accurate feature extraction

and localization, resulting in highly efficient and accurate ship detection. The rotated region proposal network (RRPN) [16] was designed as an enhancement of R2CNN to improve the quality of candidate regions generated by the region proposal network (RPN). RRPN generates 54 rotating anchor boxes with 3 scales, 3 aspect ratios, and 6 directions at each point in the feature map. Yang et al. [17] incorporated the dense feature pyramid network (FPN) based on RRPN to integrate low-level positional information and high-level semantic information, effectively detecting densely distributed ships using prow direction prediction. They also reduced redundant detection areas using rotational non-maximum suppression (NMS) to obtain more accurate object positions. Liu et al. [16] proposed the one-stage model CHPDet, which utilizes an orientation-invariant model to extract features. It selects the peak value of the feature map as the center point and then regresses the offset, size, and head point on the corresponding feature map of each center point, achieving impressive results in terms of speed and accuracy. Ding et al. [18] introduced a module called RoI Transformer based on the Faster R-CNN [13]. This module predicts a rough rotating RoI in the first stage based on the horizontal RoI generated by RPN. In the second stage, RoI Align [19] extracts the feature map of the RoI and refines the prediction result of the first stage to obtain a more accurate rotating RoI. Compared to RRPN, RoI Transformer significantly reduces the number of rotating anchors, leading to improvements in computational efficiency and detection accuracy. Oriented R-CNN [20], building upon the Faster R-CNN, introduced the midpoint offset representation and expanded the output parameters of the RPN regression branch from four to six to achieve skewed candidate boxes, resulting in substantial improvements in detection accuracy and computational efficiency.

Most object detection models are based on traditional CNNs, which possess translational equivariance but lack rotational equivariance. To enhance the adaptability of CNNs to rotational changes, numerous rotation-based object detection models have made significant efforts in data augmentation, rotation anchor boxes, or generating rotation RoIs to enhance the feature extraction of rotational objects in remote sensing images. Although these models have achieved some performance improvements, the issue of CNNs lacking rotational equivariance remains unaddressed.

To tackle this problem, ReDet [21] introduces the concept of E(2)-Equivariant Steerable CNNs [22] (E2-CNNs) into object detection by building upon the RoI Transformer [23]. ReDet adopts the *e2cnn* extension [22] for PyTorch [24] and employs it to reconfigure the architecture of ResNet-50 [25], renaming it ReResNet50. Additionally, the RiRoI Align module is redesigned to align channels and spatial dimensions, enabling the extraction of rotation-invariant features and further enhancing detection accuracy.

1.4. Goal of the Research

Although the methods mentioned above demonstrated good performance in ship detection, they still encounter several challenges. This study primarily focuses on two key issues:

(1) High-resolution optical remote sensing images contain rich pixel information of ships and other objects, but they also introduce a considerable amount of irrelevant, redundant, and potentially interfering noise. In specific scenarios, such as variations in lighting conditions and the limitations of satellite-borne sensors, the color contrast between ships and the background in some images may be low, leading to visual similarity to the human eye. Moreover, in harbor images, the shape of the dock can resemble that of large ships moored nearby. These factors contribute to the potential confusion of ships with complex backgrounds by detection models, resulting in missed detections and false alarms.

(2) Ships are typically sparsely and unevenly distributed in optical remote sensing images. Inappropriate sampling strategies may lead to the insufficient learning of ship features by the detection model. For example, the random sampling strategy commonly employed by many models, including ReDet, may exclude certain potential high-quality positive samples from the training process.

The ReDet model fails to address the aforementioned challenges. The upsampling operation in its FPN module may lead to the loss of crucial information, resulting in incomplete feature fusion and ultimately affecting the accuracy of the detection results. Additionally, the random sampler used in the RPN stage may lead to a small number of certain classes or samples due to class imbalance or sample imbalance, impacting the effectiveness of model training.

The objective of this study is to propose a solution based on ReDet to tackle the aforementioned problems, specifically focusing on ship object detection in remote sensing images. We introduce the ReBiDet model, which encompasses the following contributions: Firstly, we design a feature pyramid module called ReBiFPN based on rotational equivariance theory. ReBiFPN ensures balanced output levels and replaces the FPN module in the ReDet model. By integrating high-resolution features from higher layers with detailed information from lower layers, it enhances the detection performance of ships at multiple scales. Secondly, we propose the positive reinforcement learning sampler and an anchor optimization module based on K-means clustering. These components are utilized to balance the difficulty and proportion of positive samples during the training process, as well as optimizing the sizes and aspect ratios of anchors, further enhancing the model's detection performance. Thirdly, we validate and compare the generalization performance of ReBiDet by selecting images from different optical remote sensing datasets. Finally, following the practices of other researchers in the field of optical remote sensing image object detection, we conduct experiments on the HRSC2016 and DOTA datasets to demonstrate the effectiveness of the proposed modules and the performance advantages of ReBiDet.

2. The Proposed Method

This section presents the architecture of ReBiDet. Figure 1 illustrates the five modules of ReBiDet: the feature extraction module ReResNet [21], the feature pyramid construction module ReBiFPN, the bounding box generation module RPN, the RoIAlign module for extracting feature maps of horizontal bounding boxes and performing preliminary classification and rotation bounding box generation, and the RiRoIAlign [21] module for extracting feature maps of rotation bounding boxes and performing classification refinement and rotation bounding box parameter refinement.

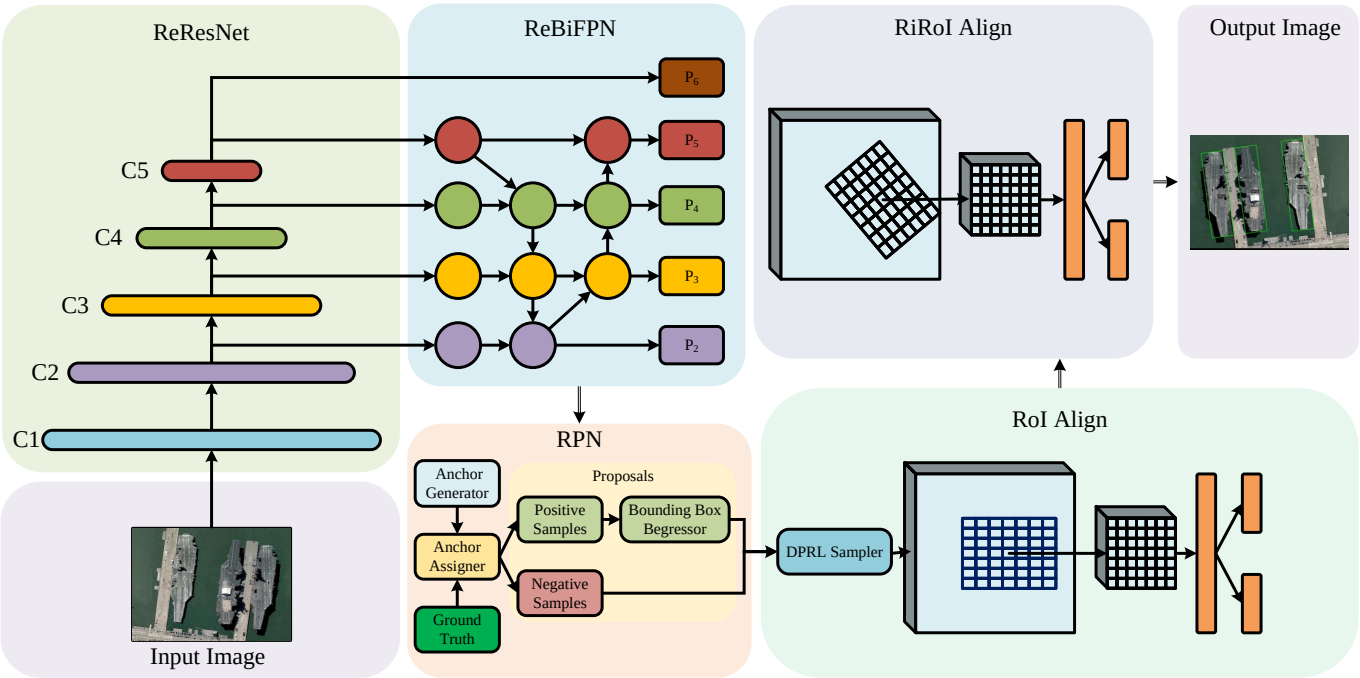


Figure 1. Architecture of rotation-equivariant bidirectional feature fusion detector (ReBiDet).

ReBiDet is an improved version of ReDet that specifically addresses the characteristics of optical remote sensing images. The objective is to tackle the challenge of detecting ships in complex scenes with scale differences and enhance the detection accuracy of ships in remote sensing images using specialized modules. Three changes were made to the original ReDet architecture. Firstly, we introduced ReBiFPN, constructed based on *e2cnn* [22], to enable the network to capture multi-scale feature information from both bottom–up and top–down paths, facilitating more comprehensive fusion of low-level positional information and high-level semantic information. Secondly, we utilized the K-means [26] algorithm to cluster the aspect ratios of ground truth boxes in the dataset and adjusted the size of the anchor boxes generated by the anchor generator accordingly. Finally, in the RoIAlign stage, we employed the difficult positive reinforcement learning (DPRL) sampler to address class imbalance in the dataset and make the model more sensitive to challenging positive samples. Experimental results demonstrate the effectiveness of these improvements in enhancing the detection performance of small objects in optical remote sensing images.

2.1. Rotation-Equivariant Networks

Convolutional neural networks (CNNs) typically employ fixed-size, two-dimensional convolution kernels to extract features from images. During convolution, these kernels apply the same weights to every position in the image, thereby allowing the detection of the same feature, even if the image is horizontally or vertically shifted. This property is known as translation invariance. However, when an image undergoes a rotation, the pixel arrangement of all objects in the image changes, causing the features that can be extracted by the CNN to also change and potentially become distorted. This lack of rotation equivariance in CNNs poses challenges in accurately recognizing rotated objects.

The absence of rotation equivariance implies that as an object in the input image changes its orientation, the features extracted by the CNN also change, adversely affecting the accuracy of the rotated object detection. This effect is particularly pronounced for elongated objects, such as buses and ships.

Traditional CNNs are designed based on the assumption of translation invariance in input data. However, for objects with rotational symmetry, the extracted features may change, resulting in the loss of important information and poor training outcomes. E2-CNNs provide a solution to this problem. Their core is rooted in group theory and convolution operations, enabling them to extract features from an image in multiple directions simultaneously, thereby ensuring rotation equivariance throughout the convolution process. This property ensures that objects yield the same feature output regardless of their orientation or angle. Since rotation involves a continuous operation that entails substantial floating-point computation and information storage, it is impractical and inefficient to include the entire 2D rotation group in the computation. Thus, optimization becomes necessary in this context.

The backbone network employed in our model, ReResNet [21], is constructed based on E2-CNNs. Specifically, the 2D rotation group is discretized initially, establishing a discretized 2D rotation group with eight discrete parameters, representing eight rotation angles. This discretization significantly reduces computational complexity and memory resource consumption. Subsequently, each convolutional kernel extracts features based on this discretized 2D rotation group, resulting in feature maps (K, N, H, W) as depicted in Figure 2, where K denotes the number of convolutional kernels, N represents the number of directions (eight), and H and W denote the height and width of the feature maps, respectively.

2.2. Rotation-Equivariant Bidirectional FPN

In the feature extraction network, lower-level features contain more detailed and positional information but have lower resolution. On the other hand, higher-level features have higher resolution and more semantic information but lack detail. In ReDet's ReFPN module, inspired by the traditional FPN [27], a top–down feature fusion approach is

adopted as shown in Figure 3. However, the traditional FPN has a relatively simple feature fusion method, performing simple feature upsampling from high to low levels. To provide a more intuitive understanding, we simplify the convolution layer and upsampling layer, representing them with Equation (1):

$$P_i = \sum_{n=i}^{l_{max}} C_n (i = 2, 3, 4, 5), \tag{1}$$

here, C_n denotes the feature map extracted by the backbone network, P_i represents the feature map output by the FPN module, and i_{max} is the total number of layers inputted into FPN from the backbone network. The output P_i fuses only the feature maps from layer i to i_{max}, ignoring the importance of feature maps below layer i and not fully utilizing the feature information from different levels, which may lead to information loss. Typically, top-level feature maps with large strides are used to detect large objects, while bottom-level feature maps with small strides are used to detect small objects [28]. However, in the case of P_3, if only the features of C_5, C_4, and C_3 from top to bottom are fused, the importance of C_2, dedicated to detecting small objects, is ignored. Additionally, both P_5 and P_6 are extracted from the C_5 layer inputted into the backbone network, but P_6 only uses 1×1 convolution downsampling with stride 2, which may result in the loss of significant feature information. In summary, the FPN design in the ReDet model suffers from unbalanced feature fusion, which may affect the detection performance of small objects in the overall model.

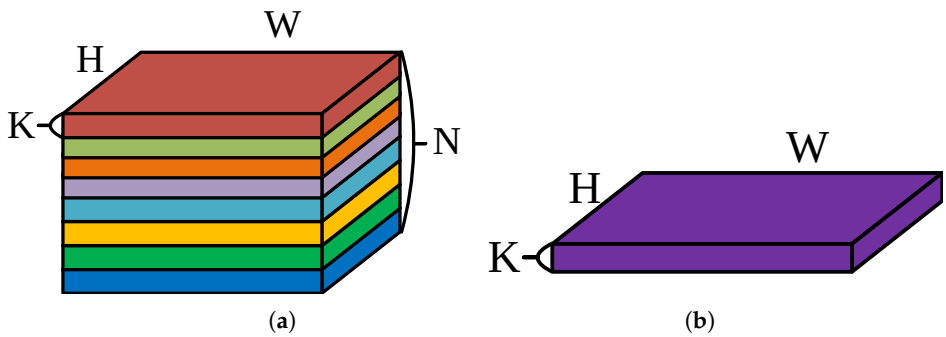


Figure 2. Difference between the feature maps extracted by ReResNet and ResNet networks. (**a**) Feature maps extracted by the ReResNet network, denoted as (K, N, H, W), where N is an additional dimension compared to traditional ResNet feature maps shown in (**b**). Here, N = 8 indicates that discrete rotation group convolution can extract feature maps at 8 different rotation angles.

To address these issues, we introduce the rotation-equivariant bidirectional feature fusion feature pyramid network module, named ReBiFPN, based on the idea of PaNet [29]. Figure 4 illustrates the ReBiFPN module. ReBiFPN balances the output levels by integrating the feature maps of each resolution. It combines the high-resolution of the high-level features with the high-detail information of the low-level features, enhancing the network's ability to extract and integrate multi-scale features through bidirectional path feature fusion. This approach captures richer multi-scale features and improves the detection performance of small objects. For simplicity, we represent the core idea of ReBiFPN with Equation (2):

$$P_i = \sum_{n=l_{min}}^{l_{max}} C_n (i = 2, 3, 4, 5), \tag{2}$$

where C_n represents the n_{th} feature map extracted by ReResNet, P_i denotes the feature map output by the ReBiFPN module, l_{max} is the total number of layers inputted into ReBiFPN from ReResNet, and l_{min} is the number of the lowest layer inputted into ReBiFPN from ReResNet. Each level of the output P_i balances the features fused from $\{C_2, C_3, C_4, C_5\}$. The P_6 layer is excluded from the feature fusion because experimental results indicate that downsampling directly from the C_5 layer to obtain P_6 improves the detection accuracy slightly compared to downsampling from the P_5 layer. This improvement may be due to

the weakening of object features after multiple convolution operations. A 3×3 convolution is added before all the final output feature maps $\{P_2, P_3, P_4, P_5\}$ to reduce the aliasing effect of upsampling.

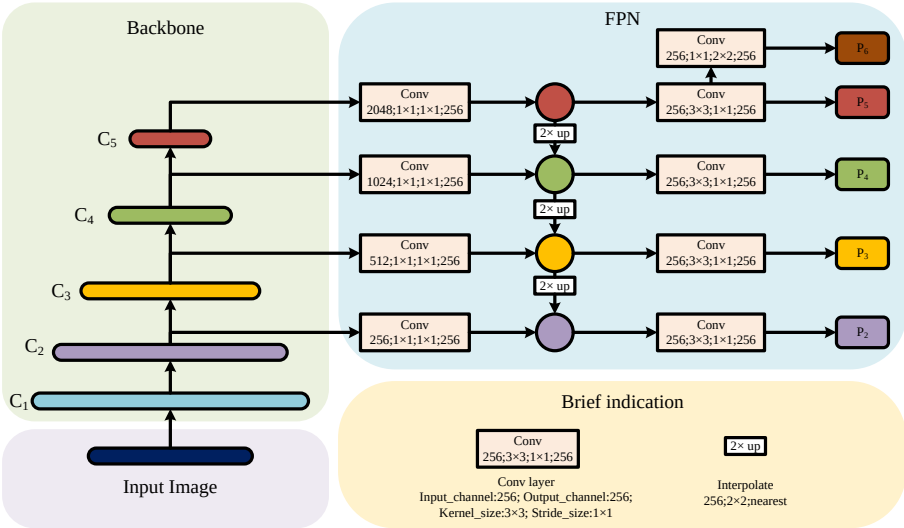


Figure 3. Traditional feature pyramid network (FPN) structure schematic.

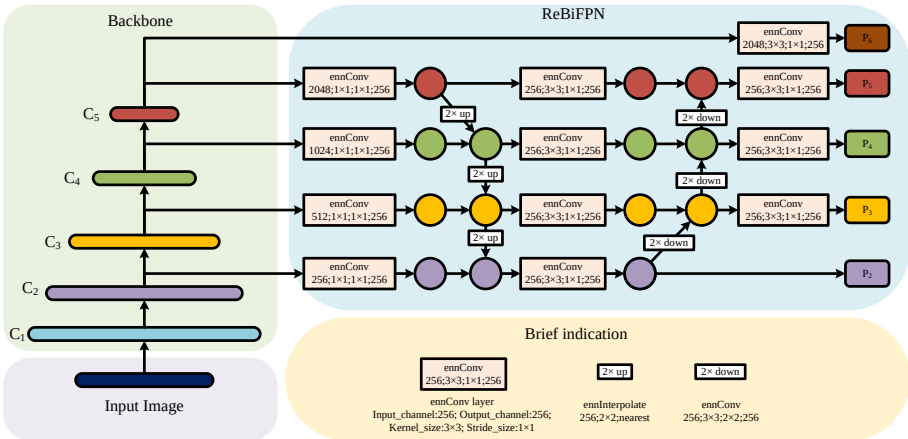


Figure 4. Structure diagram of rotation-equivariant bidirectional feature fusion feature pyramid network (ReBiFPN), where all convolution and interpolation operations are based on E(2)-equivariant steerable CNNs [22] (E2-CNNs).

Our proposed ReBiFPN achieves the following objectives:

(1) Multi-scale information fusion: By leveraging the structure of the feature pyramid, we extract features at different scales. The fusion of features from various scales enhances the network's receptive field, enabling it to detect ships of different scales. This capability allows the network to better adapt to variations in ship size and shape, thereby improving the robustness of ship detection.

(2) Bi-directional feature propagation: The use of a bi-directional feature propagation mechanism facilitates the exchange and interaction of information between different levels of the feature pyramid. This bidirectional propagation enables high-level semantic information to propagate to lower levels, enriching the semantic representation of lower-level features and enhancing their expressive capacity. Simultaneously, lower-level features can also propagate to higher levels, providing more accurate positional information, which aids in precise target localization.

(3) Multi-level feature fusion: We adopt a multi-level feature fusion strategy that progressively merges features from different layers. This approach allows the network to fuse semantic and positional information at different levels, resulting in more comprehensive

and accurate feature representations. Through multi-level feature fusion, the finer details and contextual information of ships can be better captured, thereby improving the accuracy of ship detection.

From a theoretical standpoint, our proposed ReBiFPN demonstrates superior adaptability to scale variations compared to FPN. It effectively extracts richer semantic information and accurately localizes targets, ultimately leading to improved precision in object detection. Subsequent experiments provide further validation of these findings.

2.3. Anchor Improvement

ReBiDet is a two-stage object detection model that follows an anchor-based approach, where anchors are predefined boxes placed at various positions in an image [13]. The selection of anchor sizes and aspect ratios should be based on the distribution of object annotation boxes in the dataset, which can be considered as prior knowledge [30]. Improper anchor settings can hinder the network from effectively learning features of certain objects or result in the misidentification of adjacent objects. Therefore, it is crucial to carefully consider the size and shape distribution of objects in the training dataset and experiment with different numbers and ratios of anchors to find the most suitable settings for improving the network's detection performance. To accomplish this, we employ the K-means algorithm to group the aspect ratios of ship bounding boxes in the HRSC2016 dataset into multiple clusters and determine anchor sizes based on the average aspect ratio of each cluster.

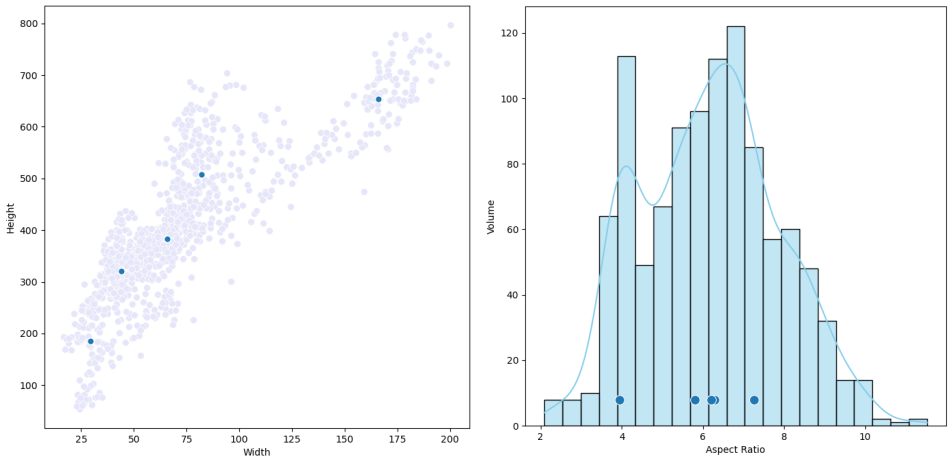
The K-means algorithm, presented in Algorithm 1, is an unsupervised learning technique. It starts by taking a set of aspect ratio data from the HRSC2016 dataset, randomly selecting K (set to 5 in our case) initial cluster centers. Each data point is then assigned to the nearest cluster center, and the cluster centers are iteratively recalculated and adjusted until they converge or reach a preset number of iterations. The objective is to maximize similarity within each cluster while minimizing similarity between clusters.

Algorithm 1 K-means used to cluster the aspect ratios of ground truth boxes.

Input: Set of bounding box aspect ratios $R = r_1, r_2, \ldots, r_n$, number of clusters K
Output: Set of K cluster centers $m_1, m_2, \ldots, m_K$ and cluster labels for each bounding box aspect

```
Randomly initialize K centroids m_1, m_2, ..., m_K from R
repeat
    for i = 1 to n do
        Assign r_i to the nearest centroid m_j, j in [1, K]
    for j = 1 to K do
        Recompute m_j as the mean of all r_i assigned to centroid m_j
until convergence or maximum number of iterations is reached
return Set of K cluster centers m_1, m_2, ..., m_K and cluster labels for each bounding box aspect
```

Figure 5 illustrates the aspect ratio data of bounding boxes in the HRSC2016 dataset, grouped into five categories. The deep blue dots represent the cluster centroids of each aspect ratio group, which are {3.9, 5.8, 6.2, 6.3, 7.3}. Although these centroids span from 3.9 to 7.3, they are concentrated around 6. This suggests that the aspect ratios of objects in the HRSC2016 dataset tend to cluster around the value of 6. After conducting multiple experiments, aspect ratios of 7.3 and above did not yield satisfactory results in practical detection. Analysis indicates that such aspect ratios are too large. Although they appear in the clustering results, many ground truth boxes in the dataset are not strictly horizontal or vertical, but rather arbitrary. Large aspect ratios result in low intersection over union (IoU) values between anchor boxes and ground truth boxes, thus failing to enhance the model's performance and introducing numerous negative samples that adversely impact performance. Consequently, we set the ratios of the RPN anchor generator as {1/6, 1/4, 1/2, 1, 2, 4, 6}.



(**a**) (**b**)

Figure 5. Aspect ratio distribution and K-means clustering results of labeled boxes in the HRSC2016 dataset. (**a**) Scatter plot of the distribution of the length and width values of all labeled boxes. (**b**) Histogram of the aspect ratio distribution of all labeled boxes.

2.4. Difficult Positive Reinforcement Learning Sampler

The RPN module generates anchor boxes of varying sizes and aspect ratios on the feature map. It assigns positive and negative samples to these anchors based on the IoU threshold with the ground truth boxes. Specifically, an anchor box with an IoU greater than 0.7 is labeled as a positive sample, while an IoU less than 0.3 results in a negative sample. Any anchors falling between these thresholds are ignored. However, training the model with all candidate boxes becomes impractical due to their large number. To expedite the training process, ReDet adopts a random sampler that randomly selects a small subset of candidate boxes for training [21]. Nevertheless, the random sampler has notable drawbacks. The ratio of positive to negative samples is often imbalanced, and the random sampling strategy may excessively emphasize easy samples, impeding the model's ability to learn from challenging samples such as small objects and overlapping objects. This not only affects the learning efficiency but also hampers the final detection accuracy.

Inspired by the concept of online hard example mining (OHEM) sampler [31], we propose selecting the candidate boxes with the highest loss for positive and negative samples during training to address the aforementioned issues with the random sampler. However, in practice, the distribution of negative samples is exceptionally complex, as depicted in Figure 6. During the sample selection stage in remote sensing ship detection, some anchor boxes may contain ship features but fall below the preset IoU threshold, while others may contain ships that were not annotated by the dataset's original author. Despite these cases being classified as negative samples with high classification loss, they actually contain object features and belong to the category of hard samples. Prioritizing such negative samples with high classification loss during training could significantly affect the detection accuracy as confirmed by our experiments. As it is not possible to entirely avoid incorrect answers, we focus on learning the correct answers presents a more viable solution. Furthermore, since the number of negative samples typically exceeds the number of positive samples, randomly sampling negative samples may have a higher likelihood of selecting samples that only contain background and lack ship instances. Based on these observations, we propose the difficult positive reinforcement learning (DPRL) sampler to address this problem. The DPRL sampler concentrates on learning challenging positive samples to enhance the model's performance, while continuing to employ the traditional random sampling approach for negative samples.

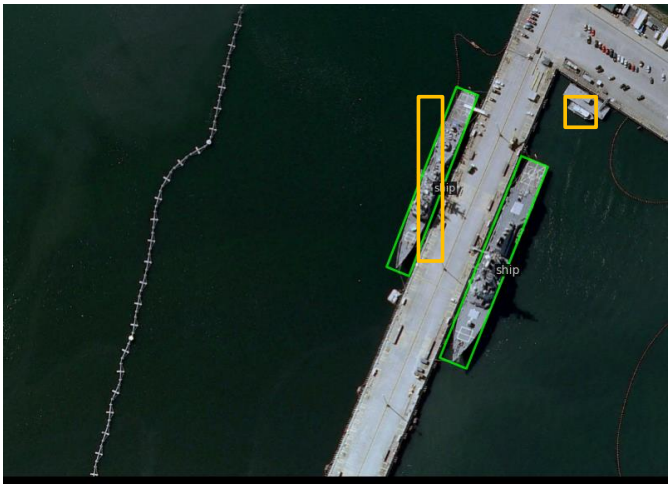


Figure 6. Illustration of the intersection over union (IoU) between anchor boxes and ground truth boxes. The green box represents the ground truth box, while the yellow box represents the anchor box generated by the model with a 1:6 ratio.

Algorithm 2 outlines the calculation process of the DPRL sampler, which can be summarized as follows. For each batch of training samples $D_{batchsize}$, the forward propagation is initially performed to generate a set of candidate boxes using RPN, followed by loss calculation. The DPRL sampler maintains positive and negative samples in L_{pos} and L_{neg}, respectively. The samples in L_{pos} are then sorted in descending order based on the loss value, and the top $pos_{expected}$ samples are selected. Simultaneously, $neg_{expected}$ negative samples are randomly chosen from L_{neg}. These selected samples are combined with L_{pos} to form a new sample set L, which is used for backward propagation and updating the neural network parameters. This method intelligently selects the most challenging positive samples for training, thereby mitigating the issue of neglecting difficult samples and effectively enhancing the training efficiency of the neural network.

Algorithm 2 Difficult positive reinforcement learning (DPRL) sampler.

Input: image set $D_{batchsize}$, samples expected $num_{expected}$, positive sample ratio pos_{ratio}
Output: The collection of selected samples L

$pos_{expected} = num_{expected} \times pos_{ratio}$
$neg_{expected} = num_{expected} \times (1 - pos_{ratio})$
Clear d_{pos} and d_{neg}
for j in range[len($D_{minibatch}$)] **do**
    RPN generate a set of positive samples $mini_{pos}$
    RPN generate a set of negative samples $mini_{neg}$
    $d_{pos} = d_{pos} + mini_{pos}$
    $d_{neg} = d_{neg} + mini_{neg}$
if $d_{pos} \leqslant pos_{expected}$ **then**
    $L_{pos} = d_{pos}$
else
    calculate the loss value for each sample in d_{pos}
    sort all elements in d_{pos} in descending order of their classification loss value
    extract the top $pos_{expected}$ samples, and form a new positive sample set L_{pos}
if $dpos \leqslant neg_{expected}$ **then**
    $L_{neg} = d_{neg}$
else
    randomly shuffle the order of each sample in d_{neg}
    extract $neg_{expected}$ samples, and form a new negative sample set L_{neg}
merge L_{pos} and L_{neg} into L
return L

3. Experimental Results

3.1. Datasets

HRSC2016 [32] was released by Northwestern Polytechnic University in 2016 and is a unique dataset focusing on ship detection in remote sensing images. It consists of a total of 1680 images captured from 6 renowned ports available on Google Earth. Among these images, 619 do not contain any ships and serve as background images. The image sizes range from 300 × 300 to 1500 × 900. The dataset is divided into training, validation, and test sets, with 436, 181, and 444 images, respectively. The annotation format for ship bounding boxes is oriented bounding boxes (OBB) rather than horizontal bounding boxes (HBB), and a total of 2976 ships are annotated. HRSC2016 has become one of the most widely used benchmark datasets in remote sensing detection. The dataset presents several challenges for ship detection: (1) The majority of ships are docked closely together, exhibiting a dense arrangement of ships side by side. (2) Some ships are integrated with long piers or located in shipyards or ship repair yards, where the similarity between the ships and the nearby textures is high, making their features less distinguishable. (3) Certain land buildings bear resemblance to luxury yachts, leading to potential misclassification. (4) The scale of ships varies greatly, and an image may contain an aircraft carrier as well as a very small civilian ship simultaneously. The dataset provides annotations for ship classification at three levels; however, due to their low frequency, detecting and classifying ship subcategories is challenging. Therefore, our RePaDet model focuses only on training and inferring the first-level ship category, following the approach of most researchers.

DOTA [4,33], publicly released by Wuhan University in November 2017, has become the most commonly used public dataset for object detection in remote sensing. DOTA-v1.0 [4] consists of 2806 images obtained from various platforms, including Google Earth, JL-1, and GF-2 satellites. The dataset contains a total of 188,282 annotated objects belonging to 15 different categories, as represented by the abbreviations in the tables presented later: airplane (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). Unlike HRSC2016, the annotation boxes in DOTA-v1.0 are quadrilaterals defined by four points $\{x_1, y_1; x_2, y_2; x_3, y_3; x_4, y_4\}$. DOTA-v1.5 [33], an extension of DOTA-v1.0, increases the number of bounding boxes to 402,089. It includes additional annotations for small objects (around 10 pixels or less) that were missed in DOTA-v1.0 and introduces a new category, container crane (CC), making it a more challenging dataset. In addition to the challenges faced by the HRSC2016 dataset, DOTA presents additional difficulties, including (1) a wide range of image pixel sizes, from 800 × 800 to 4000 × 4000, making it challenging to train on GPUs such as GTX 3090ti, and (2) small object sizes, with 98% of the objects in the dataset being smaller than 300 pixels, and 57% smaller than 50 pixels [4], resulting in significant scale differences between tiny and large objects, thereby complicating detection. It is worth noting that the DOTA dataset does not release annotations for the test set; instead, researchers are required to upload their detection results for evaluation, which limits comprehensive analysis of experimental results.

3.2. Setup

The experimental setup employed a workstation equipped with an AMD Threadripper PRO 5995WX processor, 2 NVIDIA GTX3090Ti graphics cards, and 128 GB of memory. The workstation ran on the Ubuntu20.04 operating system and was installed with various software, including CUDA 11.8, Python 3.8, PyTorch 1.11.0 [24], and torchvision 0.12.0. The MMRotate 0.3.2 [34] framework was used for the entire experimental process and result generation. Both the hardware and software versions were relatively new and exhibited strong compatibility, allowing for easy replication in various environments.

During training, data augmentation techniques, such as horizontal and vertical flips, were applied, with a batch size of 2. The entire network was optimized using the stochastic gradient descent algorithm, with a momentum of 0.9 and a weight decay of 0.0001. The IoU

threshold for positive samples in the RPN was set to 0.7. The horizontal NMS threshold was set to 0.7, and the oriented NMS threshold was set to 0.1.

For the HRSC2016 dataset, the longer side of the images was adjusted to 1280 pixels, while maintaining the aspect ratio. If the shorter side was larger than 1280 pixels, it was reduced to 1280 pixels as well. During training, the blank regions were filled with black, and the model was trained for 12 epochs. The initial learning rate was set to 0.01, which was divided by 10 at epoch 9 and epoch 11.

For the DOTA dataset, the original images were cropped to a size of 1024 × 1024, with a stride of 824 pixels, resulting in overlapping patches by 200 pixels. RePaDet was trained for 12 epochs with an initial learning rate of 0.01, which was also divided by 10 at epoch 9 and 11. Figure 7 illustrates the loss curve of our model during training on the DOTA dataset, with 3200 iterations per epoch. A significant decrease in loss was observed at the 8th epoch (25,600 iterations). To ensure a fair comparison with other models, multi-scale training was conducted by randomly cropping and rotating the dataset at three scales {0.5, 1.0, 1.5} for both training and testing.

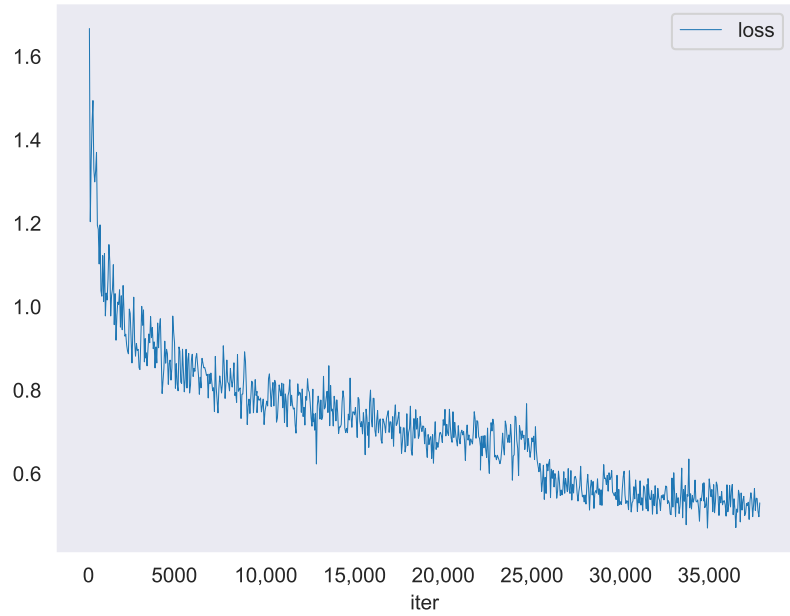


Figure 7. Train loss curve of ReBiDet on DOTA.

3.3. Results and Analysis

3.3.1. Analysis Based on HRSC2016 Dataset

Firstly, we conducted ablation experiments on the HRSC2016 dataset to evaluate the aforementioned modules. The average precision (AP) was used as the main evaluation metric, calculated using the 11-interpolated precision method proposed in VOC2007 [35]. AP50 represents the AP value at an IoU threshold of 0.5. As shown in Equation (3),

$$AP50 = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} \max_{\tilde{r}: \tilde{r} \geqslant r} p(\tilde{r}), \tag{3}$$

where r is the recall rate and $p(\tilde{r})$ is the precision at recall rate $\tilde{r}$. AP75 represents the AP value at an IoU threshold of 0.75. Its calculation formula is the same as that of AP50, except that the IoU threshold is changed to 0.75. AP50 and AP75 are commonly used to evaluate model performance [36]. The mean average precision (mAP) is the average AP calculated at different IoU thresholds ranging from 50% to 95% with a step size of 5%. Its calculation formula is given in Equation (4):

$$mAP = \frac{1}{|T|} \sum_{t \in T} AP_t \tag{4}$$

here, T represents the set of IoU thresholds, and AP_t is the average precision at IoU threshold t. This calculation method was initially used in the COCO detection challenge (bounding box) and is now commonly used in object detection evaluation [36].

Table 1 presents the performance of each module in ReDet. ReDet represents the experimental results reported by the authors, while ReDet* represents our best performance achieved in the MMRotate framework [34] for comparison. The three added modules showed varying degrees of improvement compared to the baseline. Although AP50 showed a small improvement and even a 0.1% drop after the DPRL module was added, mAP significantly improved. This suggests that the accuracy of the detection results under high IoU thresholds is greatly improved. Our ReBiDet model achieved a 1.27% improvement in mAP compared to the baseline, demonstrating the effectiveness of our proposed method.

Table 1. The ablation study results on the HRSC2016 dataset. Noted that mean average precision (mAP) refers to the average precision over all possible IoU thresholds. AP50 refers to the average precision at an IoU threshold of 0.5, and AP75 refers to the average precision at an IoU threshold of 0.75.

Method	ReBiFPN	DPRL Sampler	Anchor Improvement	mAP	AP50	AP75
ReDet	-	-	-	70.41 [1]	90.46 [1]	89.46 [1]
ReDet *	-	-	-	72.34 [2]	90.30 [2]	88.50 [2]
	✓	-	-	72.80	90.40	89.20
	-	✓	-	72.86	90.20	89.10
	-	-	✓	72.41	90.40	89.10
ReBiDet (ours)	✓	✓	✓	73.61	90.50	89.80

* Represents the detection results achieved by reproducing this model on the experimental platform we built. [1] The detection results published by the original paper authors. [2] The best result we obtained by reproducing ReDet. However, there are some differences from the results of the original authors, which cannot be ignored.

We further analyzed the detection results of the ReBiDet model on the HRSC2016 dataset and generated a confusion matrix as shown in Figure 8. The test set contained 1228 annotated ground truth boxes. Our model correctly detected 1178 ships, missed 50 ships, and misclassified 106 background images as ships. We were surprised by the misclassification of the background images, but upon visualizing the results, we discovered that incomplete annotations in the HRSC2016 dataset might be the main reason for this. The original author did not fully annotate all objects that should have been labeled as ships in this dataset. Incomplete annotations in the test set can negatively impact model performance during testing. Figure 9 illustrates that the model detected some ships that were not fully labeled during testing. However, due to the incomplete annotation information, the validation program could not accurately determine if these ships were correctly detected, leading to their misclassification as false positives and a decrease in the model's mAP.

Considering these findings, we thoroughly examined the HRSC2016 test set and manually re-annotated a portion of the missing targets, adding 80 new target annotations to the existing 1228 annotations. We then re-evaluated the detection performance of ReBiDet on the newly annotated test set. To ensure fairness, all compared models were baseline models from the MMRotate framework, replicated by the authors based on the original papers and the HRSC2016 dataset. Among them, S^2A-Net-TO [18] was the best model obtained after adjusting the hyperparameters by us. The final experimental results are presented in Table 2. On the new test set, ReBiDet showed slight improvements in mAP, AP50, and AP75. In terms of target detection statistics, at an IoU threshold of 0.5, ReBiDet correctly detected 1256 targets, missed 52 targets, and misdetected 29 targets. Its precision, recall, and accuracy were significantly better than those of ReDet and other detection models.

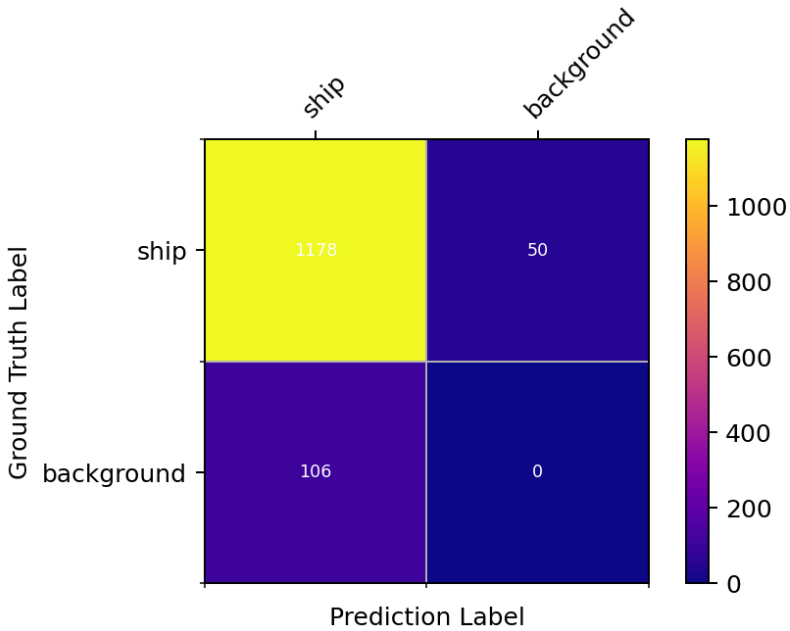


Figure 8. Confusion matrix of the detection results of the ReBiDet model on the HRSC2016 dataset. The numbers in the figure refer to the number of targets.

(**a**) (**b**)

Figure 9. Comparison between the ground truth label and detection result of an image in the test set of HRSC2016. (**a**) Image annotated with the ground truth label. (**b**) Our detection result on the same image.

Table 2. Comparison of detection results on the re-annotated HRSC2016 test set. All models compared are baseline models from the MMRotate framework.

Method	mAP	AP50	AP75	TP	FN	FP	TN	Precision	Recall	Accuracy
RetinaNet-O [37]	52.51	86.80	59.20	997	311	67	0	93.70	76.22	72.51
RetinaNet-O [37] + KLD [38]	55.43	88.70	62.00	1054	254	36	0	96.70	80.58	78.42
S^2A-Net [18]	55.31	89.80	58.90	1185	123	118	0	90.94	90.60	83.10
S^2A-Net-TO [18]	61.00	90.20	71.10	1185	123	63	0	94.95	90.60	86.43
ReDet [21]	72.46	90.70	89.70	1243	65	44	0	96.58	95.03	91.94
ReBiDet (ours)	73.66	90.80	90.00	1256	52	29	0	97.74	96.02	93.94

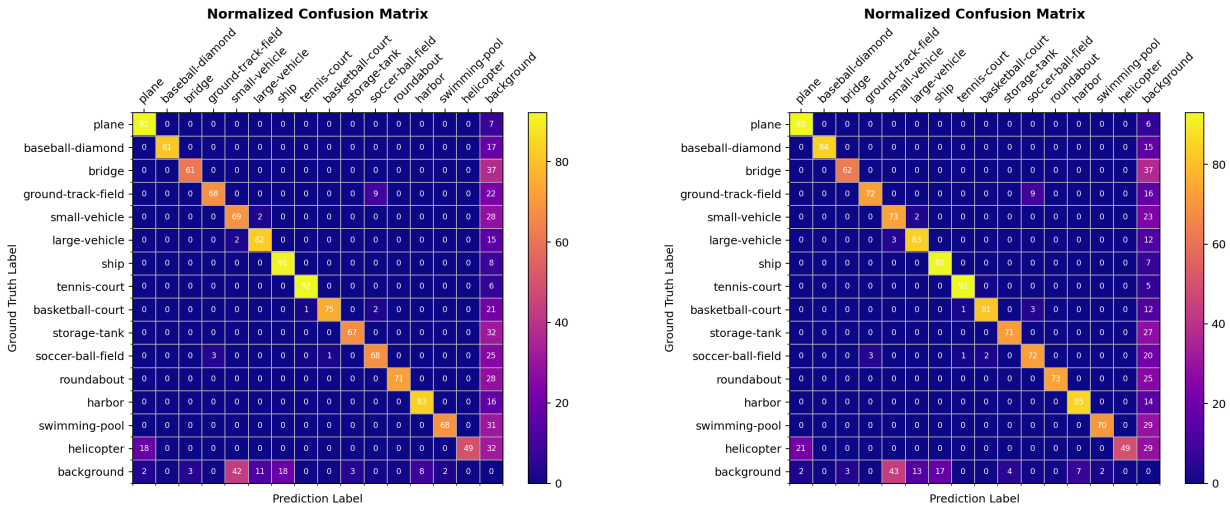
Note. The model configurations and parameters of RetinaNet OBB (RetinaNet-O) [37], RetinaNet-O [37]+ KLD [38], S^2A-Net [18], and ReDet [21] were all trained by the MMRotate team based on HRSC2016. We did not make any modifications to them. S^2A-Net-TO is a model that we optimized based on the MMRotate framework.

In terms of Precision, ReBiDet outperformed S^2A-Net by 6.8% at an IoU threshold of 0.5, but the difference in AP50 was only 1%. This can be attributed to the lower precision of ReBiDet at certain recall rates, affecting its overall AP50. Compared to RetinaNet + KLD, ReBiDet had only a 1.04% higher precision, but its precision at certain recall rates was much higher, resulting in a 2.1% higher AP50 than RetinaNet+KLD. This demonstrates that AP

and precision are not directly related [36]. AP is a more comprehensive indicator, as it considers multiple factors and evaluates the model's performance under different conditions [36]. Compared to the baseline model ReDet, ReBiDet consistently outperformed in AP50, AP75, and mAP, indicating its superior detection accuracy under high IoU threshold conditions. This further validates the superiority of our proposed ReBiDet model.

3.3.2. Analysis Based on DOTA-v1.0

Earlier, we noted that the authors of DOTA-v1.0 did not provide ground truth annotation files for the test set. Instead, participants were required to submit their test results to a designated server, receiving mAP and AP values for each class as feedback. However, these metrics were insufficient for a comprehensive performance analysis. In our experiments, we trained ReBiDet and ReDet models on the training set and evaluated them on the validation set, which was not used for training. The MMRotate framework was employed for training and evaluation. The ReDet model's learning rate was set to 0.01, consistent with ReBiDet. Other settings for ReDet remained at their default values. To facilitate a more intuitive comparison of detection performance, we generated confusion matrices for both models. Due to the large number of objects per class in the test set, we normalized the confusion matrix rows to present the correct and false recognition rates as percentages. The diagonal line in the matrix represents the proportion of true positive (TP) detections, while the rightmost column represents the number of false negative (FN) detections. Other areas denote false positive (FP) detections. Figure 10 illustrates that ReBiDet achieved a 92% TP rate for the ship class, surpassing ReDet by 1%, while also exhibiting a 7% FN rate, 1% lower than ReDet. For other classes, ReBiDet demonstrated either higher or equivalent TP numbers compared to ReDet, along with generally significantly lower FN numbers. This further confirms the effectiveness of our proposed modules.



(**a**) (**b**)

Figure 10. Normalized confusion matrix on the DOTA dataset. (**a**) The detection results of ReDet reproduced by us. (**b**) The detection results of ReBiDet. The unit of the numbers in the figure is percentage (%), representing the percentage of ground truth label targets.

Figure 11 compares the ground truth labels of some validation set images with the detection results of the ReBiDet model trained solely on the training set. The images demonstrate the superior ability of ReBiDet to detect objects not annotated in the original dataset. Figure 11a,b showcase the strong performance of ReBiDet in detecting densely packed vessels, while Figure 11c,d exemplify its recognition capability for vessels in motion. Figure 11e,f,k,l demonstrate the effectiveness of ReBiDet in detecting small objects with in-

distinct features. Furthermore, Figure 11g–j demonstrate the robust detection performance of ReBiDet, even in complex scenes with cloud and mist interference.

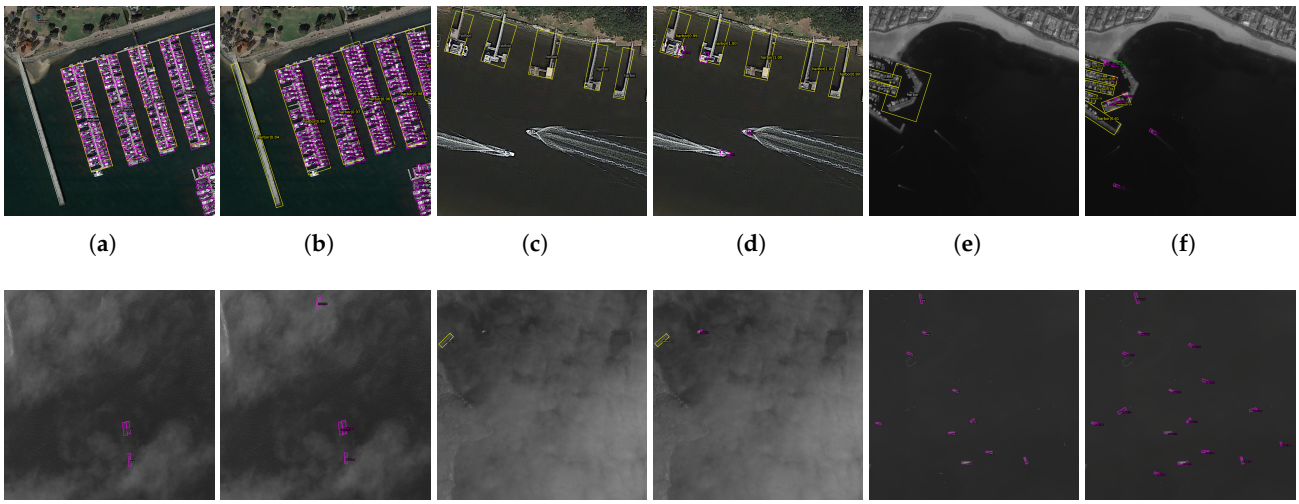


Figure 11. Comparison between the ground truth label and detection result of images in the validation set of DOTA-v1.0. (**a**,**c**,**e**,**g**,**i**,**k**) are image annotated with the ground truth label. (**b**,**d**,**f**,**h**,**j**,**l**) are our detection result on the same image.

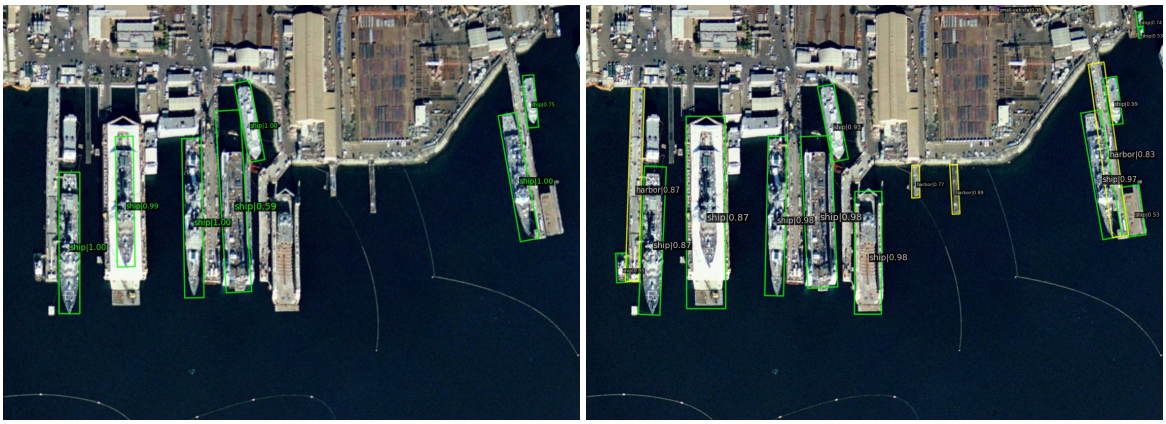
Additionally, when comparing ReBiDet with other models on a single RTX4090 computing platform, ReBiDet exhibited a slightly higher computational parameter count while achieving superior performance. Table 3 presents the comparison of ReBiDet with other models in terms of computational parameters, inference speed, and other dimensions. Within the MMRotate framework, ReBiDet had 32.56 million computational parameters, merely 0.92 M more than the 31.64 M of ReDet. This less than 1 M increase resulted in an additional 2.5 ms per image during model inference, corresponding to a decrease of 1.4 img/s in FPS. Thus, ReBiDet incurred a small computational cost, while significantly improving detection accuracy. The table reveals that computational parameters and inference speed are not completely correlated, with models having similar parameters but different inference speeds, and vice versa. Specifically, although ReBiDet had a relatively small parameter count, its inference speed was not fast enough. This may be attributed to the backbone network ReResNet50 and ReBiFPN, which are constructed based on E2CNN. These networks utilize rotation-equivariant convolutional layers, involving additional operations, such as rotation, interpolation, and grouping, which can reduce the calculation speed. Additionally, unlike traditional ResNet employing PyTorch's built-in convolutional layers, ReResNet50 and ReBiFPN require calling the *e2cnn* library for rotation-equivariant network calculations, potentially affecting the computational efficiency.

Table 3. Comparison of ReBiDet with other models on computational parameters and inference speed.

Method	Backbone	SH	mAP	Params	FPS	TPI (ms)
RoI Transformer * [23]	ResNet-101	83.59 [1]	69.56 [1]	74.12	30.00	33.30
R^3Det [39]	ResNet-152	78.21 [1]	73.74 [1]	76.54	20.40	49.10
S^2A-Net [18]	ResNet-50	87.25 [1]	74.12 [1]	87.25	32.10	31.20
Oriented R-CNN [20]	ResNet-50	88.20 [1]	75.87 [1]	41.14	39.30	25.40
Oriented R-CNN [20]	ResNet-101	87.52 [1]	76.28 [1]	60.13	32.00	31.30
ReDet [21]	ReResNet-50	88.04 [1]	76.25 [1]	31.64	24.70	40.50
ReBiDet	ReResNet-50	88.41	76.50	32.56	23.30	43.00

* indicates multi-scale training and testing. [1] The detection results published by the original paper authors. Note. SH refers to the AP of the ship category. Params refers to the computational parameter quantity of the model, in million units. FPS refers to the number of images detected per second by the model during inference. TPI is an abbreviation for times per image.

It is important to note that ReBiDet also exhibits excellent generalization performance. Figure 12 presents the detection results when applying the ReBiDet model trained on the DOTA dataset to image from the HRSC2016 dataset. The ReBiDet trained on the DOTA dataset outperformed the one trained on HRSC2016, showing improved alignment between the labeled detection boxes and the actual targets. ReBiDet detected more ships, including three additional small-sized vessels. This improvement can be attributed to the HRSC2016 training dataset's incomplete annotation, which hinders model training completeness and leads to subpar detection performance. Additionally, the ReBiDet model underwent specific optimizations for ship detection tasks, particularly enhancing feature extraction capabilities through the ReBiFPN module. These improvements contribute to better generalization performance following thorough training.



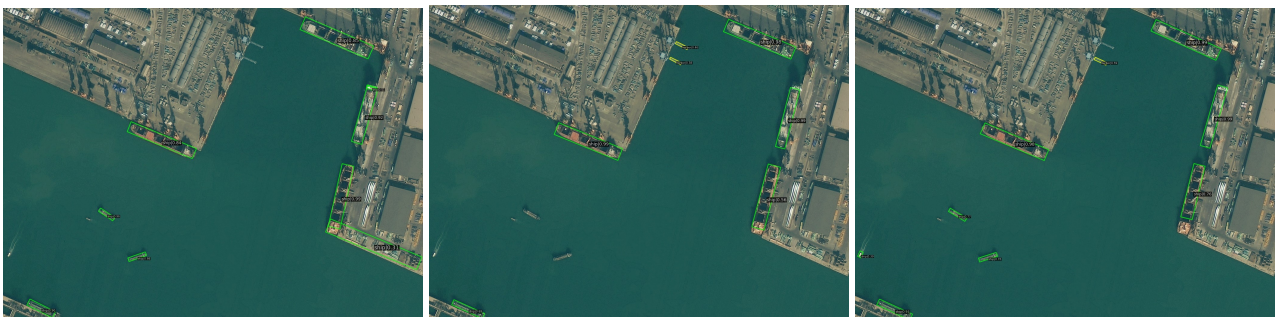
(**a**) (**b**)

Figure 12. Illustration of the generalization ability of the ReBiDet model trained on the DOTA dataset, using an original image from the HRSC2016 test set. Subfigure (**a**) displays the detection results obtained by the ReBiDet model trained on the HRSC2016 dataset. Subfigure (**b**) illustrates the detection results achieved by the ReBiDet model trained on the DOTA dataset.

To further assess the models' generalization capability, we selected an image from the FAIR1M [40] dataset. To ensure fairness, all compared models were baseline models from the MMRotate framework, reconstructed by the authors based on the original papers and the DOTA dataset. Figure 13 reveals that only ReBiDet and RoI Transformer [23] detected the small vessel on the left side of the image. However, RoI Transformer mistakenly identified three land objects as ships. Although R^3Det [39] detected more ships, most of the annotated detection boxes had confidence scores below 0.5. S^2A-Net [18] exhibited a similar situation, with confidence scores around 0.5 for the detected large ships, indicating uncertainty in the results and potentially affecting the AP metric. Oriented R-CNN [20] falsely detected a land object, and ReDet [21] failed to detect an adequate number of ships. For this particular image, ReBiDet demonstrated superior generalization capability.

(**a**) (**b**) (**c**)

Figure 13. *Cont.*



(**d**) (**e**) (**f**)

Figure 13. Comparison of the generalization capability of ReBiDet and other five models trained on the DOTA dataset, using an original image from the FAIR1M [40] dataset. Subfigures (**a**–**c**) depict the detection results of RoI Transformer [23], R^3Det [39], and S^2A-Net [18], respectively, while subfigures (**d**–**f**) represent the detection results of Oriented R-CNN [20], ReDet [21], and ReBiDet, respectively. With the exception of ReBiDet, all other models were replicated by the authors of the MMRotate framework based on the original papers and the DOTA dataset.

3.3.3. Results

This section primarily presents the performance test results of ReBiDet. We follow the practices commonly employed by researchers in the field of optical remote sensing image object detection. We compare the ReBiDet model with state-of-the-art models reported in recent years on two benchmark datasets, namely HRSC2016 and DOTA, to evaluate and compare its performance. The selected models are from reputable journals and conferences, and we directly quote their performance metrics without modification. The backbone networks used by these models are denoted as R50, R101, R152, H-104, and DLA34, representing ResNet-50, ResNet-101, ResNet-152, 104-layer hourglass network, and 34-layer deep layer aggregation network, respectively.

Detailed comparison results for the HRSC2016, DOTA-v1.0, and DOTA-v1.5 datasets are reported in Tables 4–6. On the HRSC2016 dataset, ReBiDet outperforms all existing models, demonstrating outstanding performance compared to the state-of-the-art approaches. In the single-scale detection task on DOTA-v1.0, ReBiDet surpasses all compared methods with a mAP of 76.50%, which is 0.22 higher than the baseline model. In most fine-grained categories, ReBiDet ranks within the top three, except for the ship category, where it achieves the highest AP. In the multi-scale detection task, ReBiDet exhibits various degrees of improvement compared to the single-scale detection results across all categories. ReBiDet also demonstrates improvement in the PL, BD, GTF, SV, SH, TC, ST, SBF, and HA categories compared to the baseline model, with first-place rankings in GTF, TC, ST, SBF, and HA categories. On the DOTA-v1.5 dataset, ReBiDet showcases superior performance compared to the baseline method. This advantage can be attributed to the presence of numerous small objects, approximately 10 pixels or smaller, in DOTA-v1.5, which ReBiDet excels at learning and detecting. Furthermore, even when compared to RTMDet (a research collaboration among Shanghai AI Laboratory, Nanyang Technological University, Tianjin University, Shanghai Jiaotong University, and Northeastern University) [41], ReBiDet exhibits certain advantages in terms of performance and computational parameters. These results indicates the significant capability of ReBiDet in detecting large and elongated objects with ship features. Notably, despite utilizing a 50-layer backbone network, our model outperforms all compared methods with backbone networks of 101 layers or more. Figure 14 presents some results for the DOTA dataset.

Table 4. Comparison results on the HRSC2016 dataset.

Method	Backbone	mAP	AP50	AP75
RRPN [16]	R50-FPN	-	79.08	-
R^2PN [42]	R50-FPN	-	79.60	-
RRD [43]	VGG-16	-	84.30	-
RoI Transformer [23]	R101-FPN	-	86.20	-
Gliding Vertex [44]	R101-FPN	-	88.20	-
R^3Det [39]	R101-FPN	-	89.26	-
CSL [45]	R152-FPN	-	89.62	-
S^2A-Net [18]	ResNet50	-	89.75	-
ReDet [21]	ReR50-ReFPN	70.41	90.46	89.46
ReBiDet (Ours)	ReR50-ReBiFPN	73.61	90.50	89.80

Table 5. Comparisons with state-of-the-art methods on DOTA-v1.0 OBB task.

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
single-scale:																	
FR-O [4]	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 [46]	R101-FPN	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.20	53.64	62.90	67.02	64.17	50.23	68.16
CADNet [47]	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
DRN [48]	H-104	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70
CenterMap [49]	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 [50]	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^3Det [39]	R152-FPN	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74
S^2*A*-Net [18]	R50-FPN	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
Oriented R-CNN [20]	R50-FPN	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87
ReDet [21]	ReR50-ReFPN	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25
Oriented R-CNN [20]	R101-FPN	88.86	83.48	55.27	76.92	74.27	82.10	87.52	90.90	85.56	85.33	65.51	66.82	74.36	70.15	57.28	76.28
ReBiDet (Ours)	ReR50-ReBiFPN	89.36	83.77	53.04	71.55	79.01	83.56	88.41	90.89	87.31	86.00	65.45	61.79	76.73	70.30	60.36	76.50
multi-scale:																	
RoI Transformer * [23]	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
O^2-Dnet * [51]	H104	89.30	83.30	50.10	72.10	71.10	75.60	78.70	90.90	79.90	82.90	60.20	60.00	64.60	68.90	65.70	72.80
DRN * [48]	H104	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
Gliding Vertex * [44]	R101-FPN	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
BBAVectors * [52]	R101	88.63	84.06	52.13	69.56	78.26	80.40	88.06	90.87	87.23	86.39	56.11	65.62	67.10	72.08	63.96	75.36
CenterMap * [49]	R101-FPN	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
CSL * [45]	R152-FPN	90.25	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
SCRDet++ * [53]	R152-FPN	88.68	85.22	54.70	73.71	71.92	84.14	79.39	90.82	87.04	86.02	67.90	60.86	74.52	70.76	72.66	76.56
S^2A-Net * [18]	R50-FPN	88.89	83.60	57.74	81.95	79.94	83.19	89.11	90.78	84.87	87.81	70.30	68.25	78.30	77.01	69.58	79.42
ReDet * [21]	ReR50-ReFPN	88.81	82.48	60.83	80.82	78.34	86.06	88.31	90.87	88.77	87.03	68.65	66.90	79.26	79.71	74.67	80.10
ReBiDet * (Ours)	ReR50-ReBiFPN	89.39	85.19	60.44	82.17	79.48	85.74	88.54	90.90	88.62	87.85	72.24	66.32	79.32	78.06	73.52	80.52

* indicates multi-scale training and testing. Note. For the convenience of reading and comparison, we bolded and marked the red, yellow and green numbers in the column as the 1st, 2nd, and 3rd largest values in that column, respectively.

Table 6. Comparisons with state-of-the-art methods on DOTA-v1.5 OBB task.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP
single-scale:																	
RetinaNet-O [37]	71.43	77.64	42.12	64.65	44.53	56.79	73.31	90.84	76.02	59.96	46.95	69.24	59.65	64.52	48.06	0.83	59.16
FR-O [4]	71.89	74.47	44.45	59.87	51.28	68.98	79.37	90.78	77.38	67.50	47.75	69.72	61.22	65.28	60.47	1.54	62.00
Mask R-CNN [19]	76.84	73.51	49.90	57.80	51.31	71.34	79.75	90.46	74.21	66.07	46.21	70.61	63.07	64.46	57.81	9.42	62.67
HTC [54]	77.80	73.67	51.40	63.99	51.54	73.31	80.31	90.48	75.12	67.34	48.51	70.63	64.84	64.48	55.87	5.15	63.40
CMR [54]	77.77	74.62	51.09	63.44	51.64	72.90	79.99	90.35	74.90	67.58	49.54	72.85	64.19	64.88	55.87	3.02	63.41
DAFNe [55]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.76
FR-O [4] + RT [23]	71.92	76.07	51.87	69.24	52.05	75.18	80.72	90.53	78.58	68.26	49.18	71.74	67.51	65.53	62.16	9.99	65.03
ReDet [21]	79.20	82.81	51.92	71.41	52.38	75.73	80.92	90.83	75.81	68.64	49.29	72.03	73.36	70.55	63.33	11.53	66.86
ReBiDet (Ours)	80.54	82.90	53.62	74.55	52.55	79.65	87.53	90.84	84.57	72.93	65.02	73.05	75.87	65.56	65.18	7.32	69.48
multi-scale:																	
DAFNe * [55]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.99
OWSR * [56]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	74.90
RTMDet-R-tiny [41]	88.14	83.09	51.80	77.54	65.99	82.22	89.81	90.88	80.54	81.34	64.64	71.51	77.13	76.32	72.11	46.67	74.98
RTMDet-R-s [41]	88.14	85.82	52.90	82.09	65.58	81.83	89.78	90.82	83.31	82.47	68.51	70.93	78.00	75.77	73.09	47.32	76.02
RTMDet-R-m [41]	89.07	86.71	52.57	82.47	66.13	82.55	89.77	90.88	84.39	83.34	69.51	73.03	77.82	75.98	80.21	42.00	76.65
ReDet * [21]	88.51	86.45	61.23	81.20	67.60	83.65	90.00	90.86	84.30	75.33	71.49	72.06	78.32	74.73	76.10	46.98	76.80
ReBiDet * (Ours)	86.23	85.89	61.99	82.41	67.86	83.94	89.78	90.88	86.37	83.70	72.12	77.58	78.38	73.24	75.01	52.05	77.96

* indicates multi-scale training and testing. Note. The results of RetinaNet OBB (RetinaNet-O) [37], Faster R-CNN OBB (FR-O) [4], Mask R-CNN [19] and Hybrid Task Cascade (HTC) [54] are re-implemented [33] version for DOTA, and used by some scholars [21,41]. For the convenience of reading and comparison, we bolded and marked the red, yellow and green numbers in the column as the 1st, 2nd, and 3rd largest values in that column, respectively.

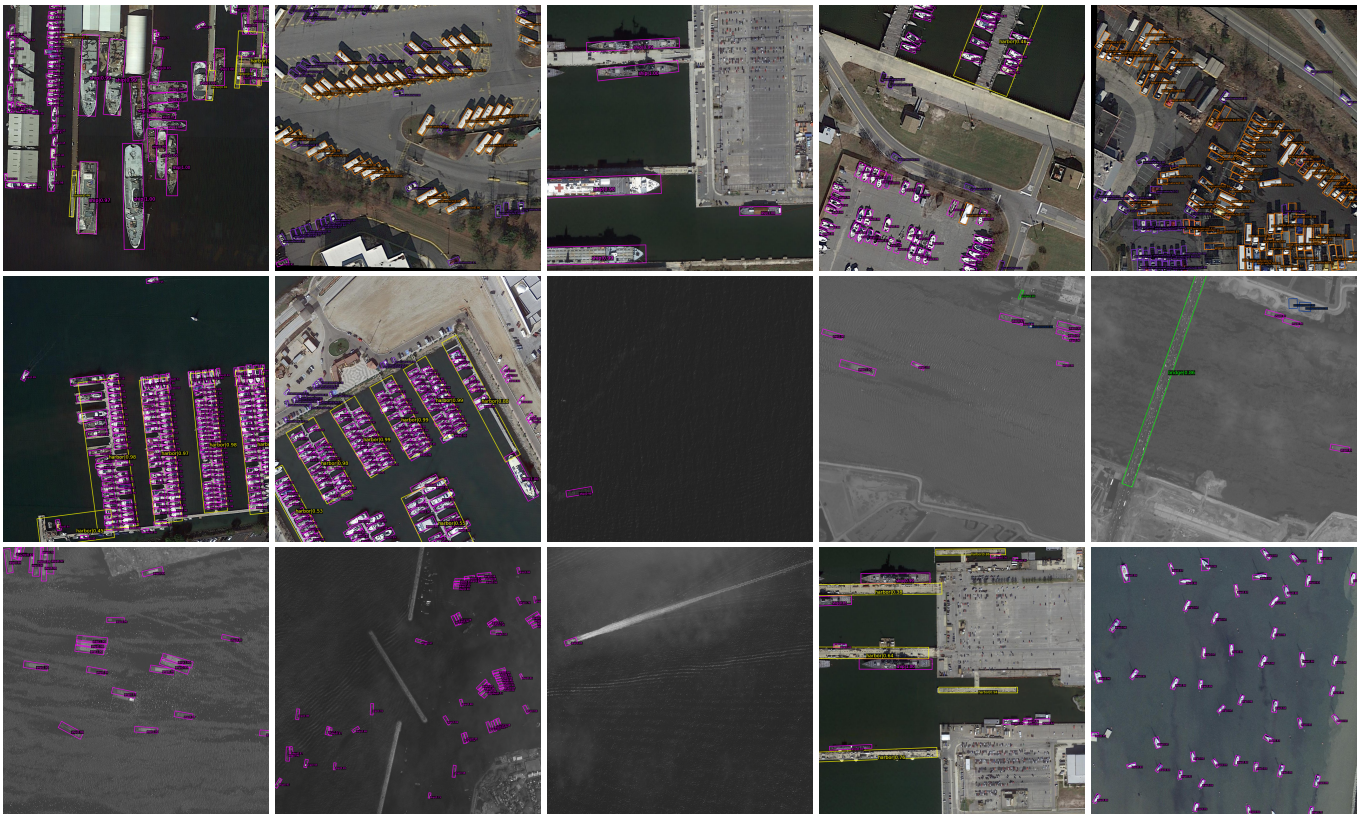


Figure 14. Examples of detection results on the DOTA dataset using ReBiDet.

4. Discussion

The experimental results validate the effectiveness of our proposed ReBiDet model in ship detection tasks, indicating the efficacy of our designed methods and modules. ReBiFPN efficiently performs cross-fusion between deep features, which possess large receptive fields and strong semantic representation ability, and shallow features that lack spatial geometric feature details and have low resolution. This approach exhibits significant performance advantages compared to traditional feature pyramid networks. Each layer acquires more comprehensive and rich feature information, thereby facilitating downstream task execution and considerably improving the model's detection ability. The anchor improvement method aggregates features from annotated boxes in the dataset, adjusting the parameters of the anchor generator accordingly, thereby enhancing the model's detection capability for specific objects. The DPRL sampler module enables the model to focus on learning difficult positive samples, while avoiding the interference of false difficult negative samples. In summary, these three proposed methods or modules hold value in improving the ship detection capability in optical remote sensing images.

Nevertheless, it is important to acknowledge that any method or research inherently possesses certain limitations. Our research primarily falls within the qualitative category, and there is still ample room for exploration in the quantitative aspects, which represents our future direction. The ReBiFPN module, inspired by PANet [29], indeed yielded positive results. The EfficientDet research [57] demonstrates that stacking multiple bi-directional feature fusion paths further enhances the detection accuracy on the COCO dataset [58]. Moreover, the detection accuracy varies with the number of stacked layers. Hence, in

our field of optical remote sensing detection, it is crucial to investigate whether a similar effect exists, which will be the focus of our forthcoming quantitative research. Regarding the anchor improvement method, we have only roughly aggregated the features of ships in the existing dataset and adjusted the parameters of the anchor generator, achieving a qualitative effect. To put it figuratively, this can be likened to "casting a wide net to catch more fish" (a common idiom). Additionally, the quantitative research lies in how to set the anchor generator reasonably and accurately to strike a balance between accuracy and computational complexity. After introducing the OHEM sampler module, we observed a decrease in detection accuracy instead of an increase. Through repeated experiments and in-depth analysis, we proposed the DPRL sampler module to enhance the learning of difficult positive samples, thereby beneficially improving the model's detection accuracy. As for low-loss positive samples, it raises the question of whether it is the correct choice to abandon direct learning. In terms of negative samples, ships that are only partially anchored and interference from land structures, roads, and vegetation are considered difficult negative samples. Although random sampling is a feasible approach, it remains uncertain whether more reasonable methods exist to address these challenges. We firmly believe that such methods do indeed exist.

For ship detection in optical remote sensing images, our research is still in its early stages, with quantitative research serving as the focal point of our future work. Concerning the issue of inaccurate detection accuracy evaluation due to incomplete object annotations in HRSC2016, we intend to manually re-annotate all ship targets in the HRSC2016 dataset in our next study. This will enable us to conduct a comprehensive comparison of various evaluation indicators based on the model's performance in existing detection models, ultimately leading to breakthroughs in improving detection performance. Furthermore, as illustrated in Figure 15, ReBiDet encounters an issue of inaccurate detection box positioning when detecting elongated objects such as ships. This issue primarily stems from inadequate training of the detection box during the alignment regression stage. We plan to explore potential solutions to this problem, including, but not limited to, modifying the loss function and the bounding box alignment regression module.

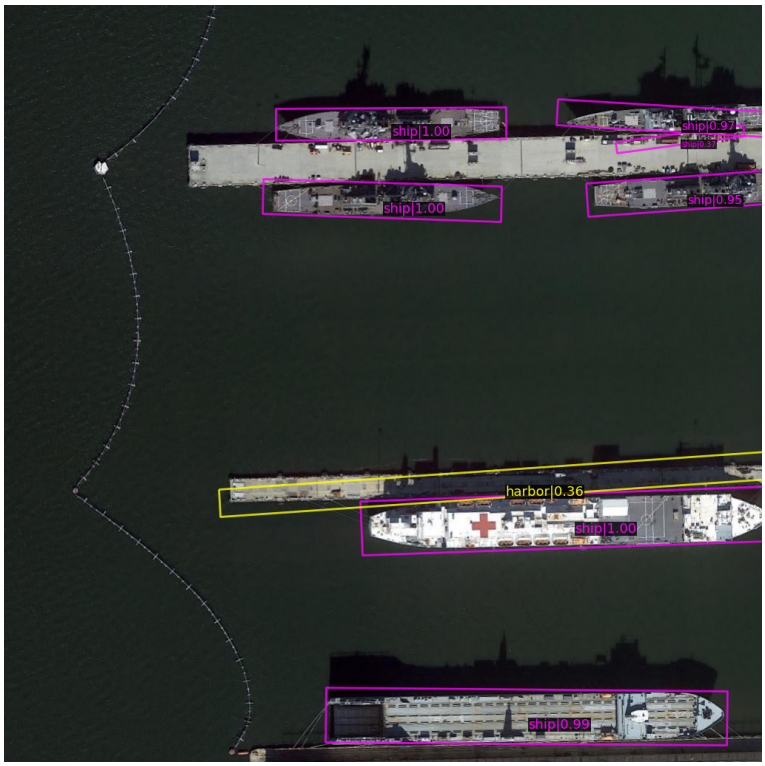


Figure 15. Illustration of inaccurate detection box positioning by ReBiDet.

5. Conclusions

This study is dedicated to addressing the ship detection problem in optical remote sensing images. The proposed ReBiDet, an enhanced version of ReDet, incorporates three improvements: ReBiFPN, DPRL sampler, and anchor improvement. These enhancements respectively bolster the network's ability to capture multi-scale features, learn difficult positive samples, and adapt to ship target aspect ratios. As a result, ReBiDet provides an effective and practically valuable solution for ship detection. The experimental results on HRSC2016 and DOTA datasets underscore the model's exceptional performance. Future research can concentrate on enhancing the model's performance by exploring more advanced network structures, loss functions, and optimization methods. Additionally, researchers can delve into additional optimization schemes for ship detection in remote sensing images to enhance accuracy and robustness. Moreover, it would be intriguing to explore the model's generalization capability on other datasets, such as FAIR1M [40].

Author Contributions: Methodology, Z.Y.; Software, Z.Y.; Supervision, Z.L. and Y.X.; Writing—Original Draft, Z.Y. and S.L.; Writing—Review and Editing, Z.Y., C.L. and Z.L.;Visualization, F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The HRSC2016 dataset used in this study was publicly released by Northwestern Polytechnical University in 2016 and includes complete sets of training, validation, and testing images and annotation files. It can be downloaded at the following address: https://aistudio.baidu.com/aistudio/datasetdetail/54106 (accessed on 11 June 2023). The DOTA dataset was publicly released by Wuhan University in November 2017 and includes complete sets of training, validation, and testing images, as well as annotation files for the training and validation sets. It can be downloaded at this address: https://captain-whu.github.io/DOTA/dataset.html (accessed on 11 June 2023). Since the downloaded package does not include annotation files for the testing set, users need to submit their model's inference results to the Wuhan University official Evaluation Server to obtain the accuracy of the inference results, at this address: https://captain-whu.github.io/DOTA/evaluation.html (accessed on 11 June 2023). Finally, we published some of the raw experiment results for this study at https://github.com/YanZeGit/ReBiDet (accessed on 11 June 2023). After this article is published, our code and other related materials will also be released at this address.

Acknowledgments: This work was supported by the Institute of Systems Engineering, Academy of Military Sciences, and We would like to express our gratitude to Yongqiang Xie and Zhongbo Li, for their support and guidance throughout the project. Their contributions have been invaluable to the success of this study.

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

Abbreviations

The following abbreviations are used in this manuscript:

ReBiDet	Rotation-equivariant Bidirectional Feature Fusion Detector
CNN	Convolutional Neural Networks
R2CNN	Rotational R-CNN
E2-CNNs	E(2)-Equivariant Steerable CNNs
RoI	Region of Interest
RRPN	Rotated Region Proposal Network
RPN	Region Proposal Network
NMS	Non-Maximum Suppression
IoU	Intersection over Union
FPN	feature pyramid network
ReFPN	Rotation-equivariant Feature Pyramid Network
ReBiFPN	Rotation-equivariant Bidirectional Feature Fusion Feature Pyramid Network
OHEM	Online Hard Example Mining

DPRL	Difficult Positive Reinforcement Learning
mAP	mean Average Precision
OBB	Oriented Bounding Boxes
TP	True Positive
FN	False Negative
FP	False Positive

References

1. Feng, X.; Zhang, W.; Su, X.; Xu, Z. Optical Remote Sensing Image Denoising and Super-Resolution Reconstructing Using Optimized Generative Network in Wavelet Transform Domain. *Remote Sens.* **2021**, *13*, 1858. [CrossRef]
2. Zhou, H.; Tian, C.; Zhang, Z.; Li, C.; Ding, Y.; Xie, Y.; Li, Z. Position-Aware Relation Learning for RGB-Thermal Salient Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 2593–2607. [CrossRef] [PubMed]
3. Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X.X. HSF-Net: Multiscale Deep Feature Embedding for Ship Detection in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7147–7161. [CrossRef]
4. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [CrossRef]
5. Shuai, T.; Sun, K.; Wu, X.; Zhang, X.; Shi, B. A ship target automatic detection method for high-resolution remote sensing. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1258–1261. [CrossRef]
6. Diao, W.; Sun, X.; Zheng, X.; Dou, F.; Wang, H.; Fu, K. Efficient Saliency-Based Object Detection in Remote Sensing Images Using Deep Belief Networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 137–141. [CrossRef]
7. Proia, N.; Pagé, V. Characterization of a Bayesian Ship Detection Method in Optical Satellite Images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 226–230. [CrossRef]
8. Ding, Z.; Yu, Y.; Wang, B.; Zhang, L. An approach for visual attention based on biquaternion and its application for ship detection in multispectral imagery. *Neurocomputing* **2012**, *76*, 9–17. [CrossRef]
9. Li, B.; Xie, X.; Wei, X.; Tang, W. Ship detection and classification from optical remote sensing images: A survey. *Chin. J. Aeronaut.* **2021**, *34*, 145–163. [CrossRef]
10. Yin, Y.; Cheng, X.; Shi, F.; Liu, X.J.; Huo, H.; Chen, S. High-order Spatial Interactions Enhanced Lightweight Model for Optical Remote Sensing Image-based Small Ship Detection. *arXiv* **2023**, arXiv:abs/2304.03812.
11. Ren, Z.; Tang, Y.; He, Z.; Tian, L.; Yang, Y.; Zhang, W. Ship Detection in High-Resolution Optical Remote Sensing Images Aided by Saliency Information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623616. [CrossRef]
12. Li, C.; Zhou, H.; Liu, Y.; Yang, C.; Xie, Y.; Li, Z.; Zhu, L. Detection-Friendly Dehazing: Object Detection in Real-World Hazy Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8284–8295. [CrossRef]
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
14. 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:abs/1706.09579.
15. Liu, Z.; Hu, J.; Weng, L.; Yang, Y. Rotated region based CNN for ship detection. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 900–904. [CrossRef]
16. Liu, W.; Ma, L.; Chen, H. Arbitrary-Oriented Ship Detection Framework in Optical Remote-Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 937–941. [CrossRef]
17. Yang, X.; Sun, H.; Sun, X.; Yan, M.; Guo, Z.; Fu, K. Position Detection and Direction Prediction for Arbitrary-Oriented Ships via Multitask Rotation Region Convolutional Neural Network. *IEEE Access* **2018**, *6*, 50839–50849. [CrossRef]
18. 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]
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
20. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3500–3509. [CrossRef]
21. Han, J.; Ding, J.; Xue, N.; Xia, G. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2785–2794. [CrossRef]
22. Weiler, M.; Cesa, G. General E(2)-Equivariant Steerable CNNs. In Proceedings of the Thirty-Third Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019. Available online: https://proceedings.neurips.cc//paper/2019/hash/45d6637b718d0f24a237069fe41b0db4-Abstract.html (accessed on 11 June 2023).
23. Ding, J.; Xue, N.; Long, Y.; Xia, G.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2844–2853. [CrossRef]

24. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Thirty-Third Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019. Available online: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (accessed on 11 June 2023).
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
26. Arthur, D.; Vassilvitskii, S. K-means++ the advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; pp. 1027–1035.
27. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
28. Gong, Y.; Yu, X.; Ding, Y.; Peng, X.; Zhao, J.; Han, Z. Effective Fusion Factor in FPN for Tiny Object Detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1159–1167. [CrossRef]
29. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [CrossRef]
30. Zhong, Y.; Wang, J.; Peng, J.; Zhang, L. Anchor Box Optimization for Object Detection. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1275–1283. [CrossRef]
31. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769. [CrossRef]
32. 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–ICPRAM, Porto, Portugal, 24–26 February 2017; pp. 324–331. [CrossRef]
33. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [CrossRef] [PubMed]
34. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. 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. [CrossRef]
35. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. Available online: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/ (accessed on 11 June 2023).
36. Padilla, R.; Netto, S.L.; Silva, E.A.B.d. A Survey on Performance Metrics for Object-Detection Algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 1–3 July 2020; pp. 237–242. [CrossRef]
37. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
38. 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 Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual, 6–14 December 2021. Available online: https://proceedings.neurips.cc/paper/2021/hash/98f13708210194c475687be6106a3b84-Abstract.html (accessed on 11 June 2023).
39. Yang, X.; Liu, Q.; Yan, J.; Li, A. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 8–9 February 2021; pp. 3163–3171. [CrossRef]
40. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
41. 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:abs/2212.07784.
42. Zhang, Z.; Guo, W.; Zhu, S.; Yu, W. Toward Arbitrary-Oriented Ship Detection With Rotated Region Proposal and Discrimination Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1745–1749. [CrossRef]
43. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.S.; Bai, X. Rotation-Sensitive Regression for Oriented Scene Text Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918. [CrossRef]
44. 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]
45. Yang, X.; Yan, J. Arbitrary-Oriented Object Detection with Circular Smooth Label. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 677–694.

46. 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 Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 150–165. [CrossRef]
47. 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]
48. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11204–11213. [CrossRef]
49. 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]
50. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8231–8240. [CrossRef]
51. Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; Sun, X. Oriented objects as pairs of middle lines. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 268–279. [CrossRef]
52. Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; Metaxas, D.N. Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 2149–2158. [CrossRef]
53. Yang, X.; Yan, J.; Yang, X.; Tang, J.; Liao, W.; He, T. SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2384–2399. [CrossRef] [PubMed]
54. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid Task Cascade for Instance Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4969–4978. [CrossRef]
55. Lang, S.; Ventola, F.G.; Kersting, K. DAFNe: A One-Stage Anchor-Free Deep Model for Oriented Object Detection. *arXiv* **2021**, arXiv:abs/2109.06148.
56. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Jie, Z.; Zhang, T.; Yang, J. Learning Object-Wise Semantic Representation for Detection in Remote Sensing Imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 20–27. Available online: http://openaccess.thecvf.com/content_CVPRW_2019/html/DOAI/Li_Learning_Object-Wise_Semantic_Representation_for_Detection_in_Remote_Sensing_Imagery_CVPRW_2019_paper.html (accessed on 11 June 2023).
57. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
58. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [CrossRef]