*remote sensing*

MDPI

*Article*

# AFRE-Net: Adaptive Feature Representation Enhancement for Arbitrary Oriented Object Detection

**Tianwei Zhang [1,2,3,4], Xu Sun [2,*], Lina Zhuang [2], Xiaoyu Dong [5], Jianjun Sha [6,7], Bing Zhang [4] and Ke Zheng [8]**

1 Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; zhangtianwei20@mails.ucas.ac.cn
2 Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
3 International Research Center of Big Data for Sustainable Development Goals, Chinese Academy of Sciences, Beijing 100094, China
4 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
5 Department of Complexity Science and Engineering, The University of Tokyo, Tokyo 277-8561, Japan
6 College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China; shajianjun@hrbeu.edu.cn
7 Qingdao Innovation and Development Center, Harbin Engineering University, Qingdao 266000, China
8 College of Geography and Environment, Liaocheng University, Liaocheng 252059, China
* Correspondence: sunxu@aircas.ac.cn

**Abstract:** Arbitrary-oriented object detection (AOOD) is a crucial task in aerial image analysis but is also faced with significant challenges. In current AOOD detectors, commonly used multi-scale feature fusion modules fall short in spatial and semantic information complement between scales. Additionally, fixed feature extraction structures are usually used following a fusion model, resulting in the inability of detectors to self-adjust. At the same time, feature fusion and extraction modules are designed in isolation and the internal synergy between them is ignored. The above problems result in feature representation deficiency, thus affecting the overall detection precision. To solve these problems, we first create a fine-grained feature pyramid network (FG-FPN) that not only provides richer spatial and semantic features, but also completes neighbor scale features in a self-learning mode. Subsequently, we propose a novel feature enhancement module (FEM) to fit FG-FPN. FEM authorizes the detection unit to automatically adjust the sensing area and adaptively suppress background interference, thereby generating stronger feature representations. Our proposed solution was tested through extensive experiments on challenging datasets, including DOTA (77.44% mAP), HRSC2016 (97.82% mAP), UCAS-AOD (91.34% mAP), as well as ICDAR2015 (86.27% F-score) and its effectiveness and high applicability are verified on all the above datasets.

**Keywords:** deep learning; object detection; remote sensing; feature representation enhancement



**Citation:** Zhang, T.; Sun, X.; Zhuang, L.; Dong, X.; Sha, J.; Zhang, B.; Zheng, K. AFRE-Net: Adaptive Feature Representation Enhancement for Arbitrary Oriented Object Detection. *Remote Sens.* **2023**, *15*, 4965. https://doi.org/10.3390/rs15204965

Academic Editors: Yanni Dong and Xiaochen Yang

Received: 15 August 2023
Revised: 30 September 2023
Accepted: 8 October 2023
Published: 14 October 2023



## 1. Introduction

As a fundamental task in remote sensing image understanding, arbitrary-oriented-object-detection (AOOD) is attracting the attention of researchers more and more. At the same time, with the rapid development of convolutional neural network (CNN)-based methods [1–5], many outstanding AOOD detectors stand out [6–12]. However, different from object detection in natural images, AOOD is more challenging mainly due to the following two reasons:

1. Objects in remote sensing images tend to have random orientation and larger aspect ratios, which increase the feature representation complexity of detectors.
2. Remote sensing images, due to their wide imaging range, contain complex and diverse ground objects and scenes, resulting in increased interference targets and features.

However, the existing design of AOOD detectors cannot adapt to the feature representation of remote sensing objects very well. Although AOOD detectors use the oriented bounding box (OBB) as the object's marker, which can better fit the object's spatial contour, the feature representation ability of each detection unit (DN) (i.e., feature point in multi-scale detection layers) does not change.

Take the classic anchor-based object detector as an example, as shown in Figure 1, at each position of the multi-scale detection layers, a certain number of anchors will be preset for overlap calculation with GT (ground truth). When an anchor and GT meet certain position and overlap conditions (i.e., label assignment strategy), it will be determined as positive or negative. However, no matter whether HBB (horizontal bounding box) or OBB (oriented bounding box) is used as the labeling of GT, the effective receptive field (ERF) [13] of each DN does not change; that is, no matter what shape and aspect ratio of the object appears at the current position, existing detectors use a fixed feature vector to represent it. This means that for the red high potential DN in shallow feature layers shown in Figure 1a, its feature learning area is limited and does not coincide with the space occupied by the target. This issue has been discussed by some scholars [14,15] and summarized as a feature misalignment problem; however, these researches have not conducted in-depth summary and analysis of internal causes.

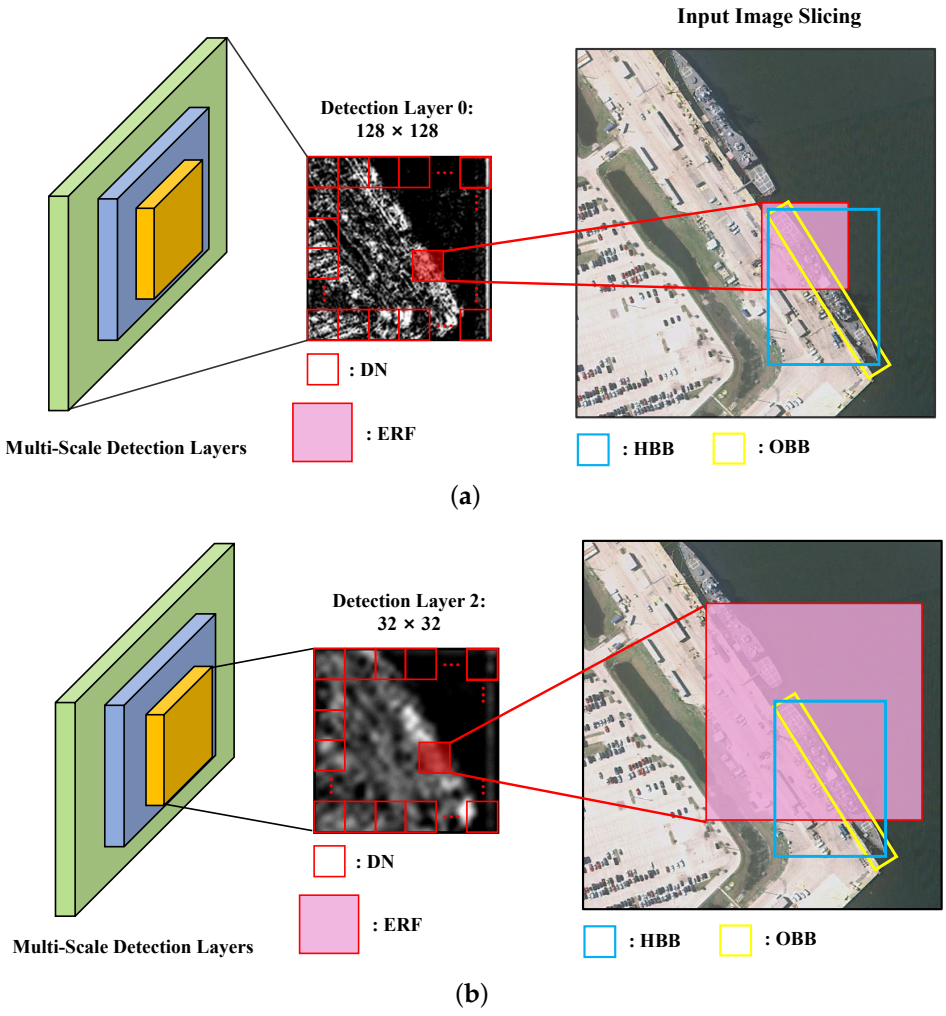



**Figure 1.** Illustration of the relationship between DN and GT and ERF. A part of the input image is captured inside the detector (RetinaNet-R [16]). The HBB and the OBB are the predicted box. The red box is virtual and represents only one pixel in detection layers. The ERF is calculated according to [13]. It should be noted that this is only a local scene captured from a large remote sensing image input. (**a**) Detection Layer 0; (**b**) Detection Layer 2.

For this issue, one intuitive solution is using multi-scale feature representations to compensate for the uncertainty caused by the change of image and target size. However, another problem arises, as shown in Figure 1b. With the deepening of the network and the down sampling operation, the ERF of DN has expanded. In the detection layer 2 with the size of 32 × 32, the marked DN expands its knowledge learning range but also receives more complex background information. The case in Figure 2 shows the negative impact of the disorderly expansion of ERF, which occurs in a real application scenario. Because the containers and cargo ships in the port have very similar characteristics, they are easily confused when they appear in the ERF of the same DN. Therefore, the container on the shore is also mistakenly identified as a ship with a high confidence. To deal with those problems, an ideal situation is that the field of vision focused by each DN is the whole body of the target, and does not contain additional background. However, due to the randomness of target size and input image size, it is difficult to achieve the above situation. More importantly, through the above case study, we observed that multi-scale feature fusion and feature extraction units are mutually constrained and auxiliary, because they jointly affect the ERF of each DN.

In summary, we need multi-scale fusion models to provide rich feature information to meet the size transformation of the target, and feature extraction operators to achieve the adaptive adjustment of ERF to suppress background information and highlight the key areas. However, the existing feature fusion models, such as FPN [17] and FPN-variants [18–21], cannot realize the information supplement between neighbor-scale features. The existing feature adaptive learning operators based on deformable convolution (DCN) [14,22,23] cannot achieve the synergy between the capture of key areas and background suppression, and their design is mostly separated from the fusion model, which does not form a good chemical reaction.

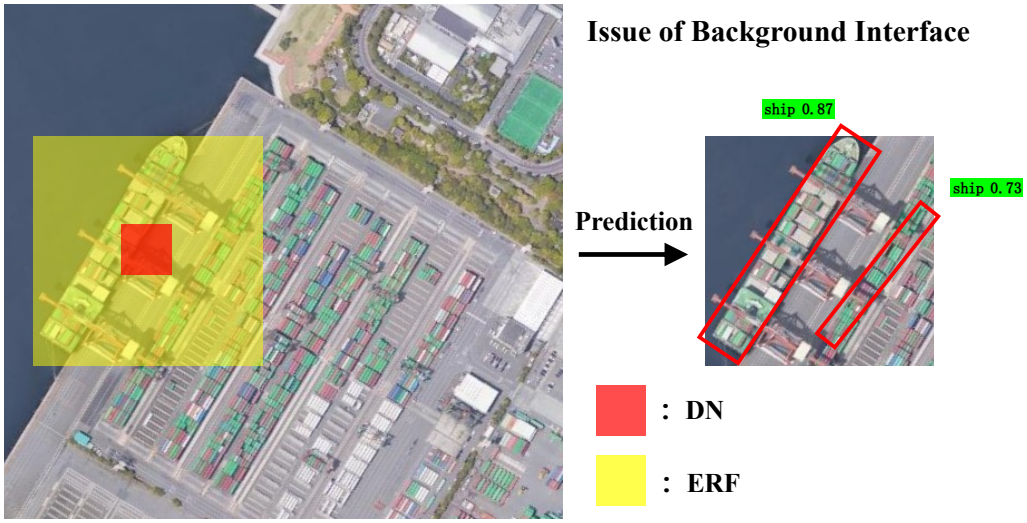



**Figure 2.** Example of wrong detection caused by background interference. In this case, RetinaNet-R is used. The prediction confidence threshold is 0.3. The container at the port is mistaken as a freighter because they have similar features.

Seeing the above challenges, we propose an innovative AOOD detector called AFRE-Net (adaptive feature representation enhancement network), which effectuates adaptive feature representation enhancement for DNs in multi-scale detection layers. AFRE-Net is committed to achieving feature relevance learning between adjacent scales and end-to-end ERF transformation, so as to strengthen the feature representation in the detection layer. The overall architecture of the proposed AFRE-Net is shown in Figure 3, which consists of four modules: (1) The backbone for basic feature extraction; (2) An FG-FPN (fine-grained feature pyramid network) for providing finer multi-dimensional feature maps and performing feature fusions; (3) A feature enhancement module (FEM) and (4) a rotation detection

module for category prediction (CP) and bounding box regression (BBR). As opposed to the regular feature pyramid network (FPN) [17], FG-FPN is designed to make better use of the low-dimensional feature maps rich in spatial information, and it uses a more fine-grained fusion method to provide a basis of features for subsequent FEM. In FEM, we apply the ERF transformation based on DCN, and invented a background suppression and foreground enhancement algorithm named relative-Conv, to achieve automatic and adaptive object representation enhancement. Extensive experimental tests on three benchmark remote sensing datasets (DOTA, HRSC2016, UCAS-AOD), as well as a text recognition dataset (ICDAR2015) demonstrate the state-of-the-art performance of our AFRE-Net.

The contributions of our work can be concluded as follows:

1. Our systematic analysis has mined three aspects that need to be considered together to improve the detector's feature representation ability: fusion module, receptive field adjustment, and background suppression.
2. We propose a novel FG-FPN to provide finer features and fuse them in a more efficient manner. Different from FPN and its modifications, we focus on neighbor-scale information supplement to fullfil all-scale features.
3. A novel background suppression and foreground enhancement convolution module called relative conv is proposed to encourage DNs to learn the key areas adaptively.
4. We propose a new ERF transformation algorithm to make the sampling position more accurately located on the main body of the target, obtaining stronger semantic features.

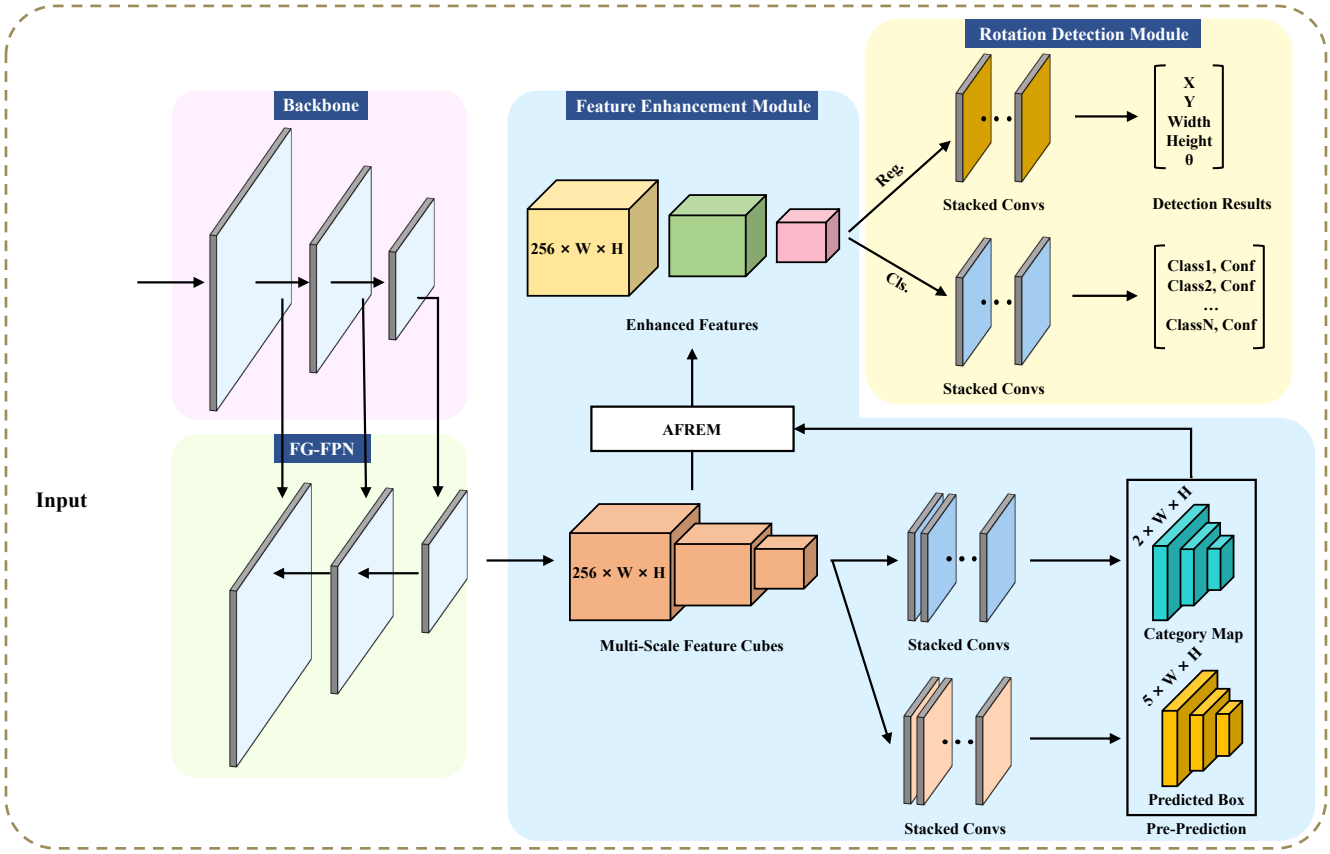



**Figure 3.** The overall architecture of the AFRE-Net. AFRENet is composed of a backbone, a fine-grained feature pyramid network, a feature enhancement module, and a rotation detection module. AFPEM denotes adaptive feature representation enhancement module. PB and CM denote the predicted box and classification map, respectively.

## 2. Related Work

### *2.1. Arbitrary Oriented Object Detection*

AOOD is an extension of object detection tasks in natural scenes [24,25], which follows the basic natural object detector pipeline. Concretely, detectors can be divided into anchor-based methods and anchor-free methods. For anchor-based detectors, YOLO [26,27] series lead the one-stage field and have achieved remarkable effects by designing regression models that balance accuracy and speed. R-CNN series [28,29] represent two-stage detectors and use region proposal network (RPN) to filter potential DNs. The latter often achieves higher detection accuracy due to its ability to control positive and negative samples well; however, the authors of ATSS [30] point out that the sample learning strategy is the key factor. After that, plenty of intelligent positive and negative sample learning strategies have been proposed [31,32].

Additionally, numerous scholars have attempted to tackle the practical issues that arise in AOOD. For example, refs. [10–12,33] have focused on solving the discontinuous problem of angle parameter regression in the training process, and constantly refining the loss function to improve performance. Some other scholars attempt to use a target representation vector that can eliminate such boundary discontinuities to represent instances, such as polar coordinates [12], ellipse bounding box [34], and middle lines of boxes [35]. Additionally, to obtain better-refined rotated anchor boxes, RR-CNN [36], R3Det [37], and CFC-Net [38] focus on spatial alignment and anchor refinement to guide the training process. However, none of them try to explain the limitations of detection units on target feature expression in a deeper perspective.

### *2.2. Feature Fusion Module*

Feature pyramid network (FPN) [17] is the most frequently used feature fusion structure, because it can well integrate low-level spatial information with high-level semantic information, and can well cope with the feature differences caused by the change of target size. After FPN is proposed, PANet [18] proposed a dual-path fusion mode of top-down plus bottom-up to enhance the semantic representation of low-level feature maps. BiFPN [20] refined the design of this pattern and further improved its performance. Recursive-FPN [21] adds the additional feedback connection of FPN to the bottom-up backbone layer and it convolves features with different void rates based on switchable atrous convolution. However, these methods are not designed from the perspective of scale information supplement. Considering that no feature pyramid can completely cover the full size of all targets and input images, we need to mine the feature correlation of critical scale as far as possible to make up for this deficiency.

### *2.3. Feature Enhancement Module*

FEM in AOOD has a broader scope of reference and is not limited to using a certain class of methods or specific means to be called feature representation enhancement. For example, some scholars [14,39] focus on solving feature misalignment between classification and box regression (localization), arguing the classification score of anchor boxes cannot reflect real localization accuracy. FFN [40] enhances the model's feature expression ability for sub-optimal detection units in a creative way, but its design is too complex. Han et al. [14] were the first to attempt to alleviate this inconsistency through deformable convolutions [22] in AOOD; however, they ignore the adaptive feature representation learning. Ref. [39] proposed rotated align convolution (RAC) to improve feature representation of ship targets. However, they did not consider that the selection of sampling location should fall on the main body of the target body, and did not analyze the impact of background information interference within the detection unit on feature expression.

In addition, the performance of CNN in rotational feature extraction is known to be subpar [41]. Therefore, the research on rotation invariant feature extraction, as highlighted by works such as [42–44], plays a crucial role in improving CNN-based detectors. To achieve enhanced rotation invariant features, spatial and channel interpolation techniques

are often employed [45,46]. For instance, Cheng et al. [6] were the pioneers in utilizing a fully connected layer as the rotation invariant layer, constraining it within the loss function. Furthermore, ReDet [8] takes a different approach by adaptively extracting rotation-invariant features from equivariant features based on the orientation of the Region of Interest (RoI). This adaptive extraction process contributes to the overall effectiveness of the detector. Oriented Reppoints [23] proposes an adaptive point set feature representation for AOOD tasks based on [47]. Although it can realize the adaptive transformation of ERF, its learning process is unordered and it does not realize the adaptive background suppression within a single feature extraction operator. Moreover, the above methods did not consider how to integrate the design of feature fusion with subsequent FEM as a whole. Our experiments show when the feature fusion module and ERF transformation are more fit, the feature representation ability of the detector is stronger.

## 3. Methodology

In this section, we will introduce the design of each independent module in AFRE-Net in detail. The overall pipeline of our detector is first introduced in Section 3.1. Then, we detail the FG-FPN in Section 3.2. Later, the adaptive feature representation enhancement module is unveiled in Section 3.4. Finally, experiment details about label assignment strategy are presented in Section 3.5.

### *3.1. Overall Pipeline*

AFRE-Net is built based on RetinaNet-R [16] (baseline detector), which has classic object detector architecture and is easy to transfer. The overall pipeline of our AFRE-Net is shown in Figure 3. AFRE-Net consists of four main parts:

(1): A backbone network (ResNet [48] in our experiments) for basic feature extraction.

(2): A feature pyramid network for multi-scale feature fusions. We replace FPN [17] with our FG-FPN. FG-FPN is more capable of taking advantage of lower dimensional feature maps, which contain richer spatial information, can provide more fine-grained feature vectors for subsequent FEM, and enhance feature capture for small objects. FG-FPN fuses the feature maps with a top-down-top (rather than top-down only in FPN) pathway so the network constructs both rich semantic features and spacial information.

(3): A feature enhancement module (FEM). FEM is designed to reconstruct the feature vectors of DNs in the detection layer, including the automatic transformation of the FRF and adaptive background suppression.

(4): A rotation detection module (RDM). RDM converts semantic features into predicted bounding boxes and confidence of prediction categories for regions. RDM is a multi-task module and the regression targets are obtained in it. In our experiments, we adopt the five-parameter method to describe the bounding box, which is denoted as:

$$\{(x, y, w, h, \theta)\}, \tag{1}$$

where $x, y, w, h$ are coordinates of the bounding box center, the width, and the height, respectively. Parameter $\theta \in \left[-\frac{\pi}{4}, \frac{3\pi}{4}\right]$ denotes the angle from the position direction of $x$ to the direction of $w$. We have

$$\begin{aligned} t_x &= (x - x_a)/w_a, \quad t_y = (y - y_a)/h_a, \\ t_w &= \log(w/w_a), \quad t_h = \log(h/h_a), \\ t_\theta &= \tan(\theta - \theta_a). \end{aligned} \tag{2}$$

### *3.2. FG-FPN*

#### 3.2.1. Overall Architecture

It can be seen in Figure 4, RetinaNet-R employs the $C_3$, $C_4$, $C_5$ in the backbone network as the foundation for the following feature fusions. $C_3$, $C_4$, and $C_5$ are performed by $1 \times 1$

convolution to obtain the same channel dimension (256 in the baseline). Then, $P_5$ is obtained through a convolution layer of $3 \times 3$. The top-level feature $P_5$ is transmitted in a top-down manner and fused with low-level features layer by layer to obtain $P_4$ and $P_3$. In this way, layer $P_i$ ($i = 3, 4, 5$) is the same size as layer $C_i$. Based on $P_5$, $P_6$ is obtained through a $3 \times 3$ convolution with stride set to 2. $P_7$ is obtained through a Relu activation function and a $3 \times 3$ convolution with stride set to 2.

However, the structure of FPN has the following defects:

(1): The utilization efficiency of low-level features is insufficient. It is necessary to add lower-level features rich in spatial information to participate in the fusion process, in order to improve the feature perception ability of small objects.
(2): There are barriers between high-level features and low-level feature maps, as using only the top-down linking makes it impossible for high-dimensional feature maps to communicate directly with low-level feature maps (such as C3 and P5).
(3): Lack of mining for the correlation of features between adjacent scales.

Therefore,

(1): We re-enable the $C_2$ layer in ResNet (yellow layer shown in Figure 5);
(2): After performing the top-down fusion, we perform the down-top feature fusion as well, which means using top-down-top structure.
(3): We design an attention mechanism for mining inter-scale correlations, achieving the goal of simulating full-scale pyramid layers.

It should be noted that both the size and the numbers of the detection layer in FG-FPN have not changed, and $C_2$ is only used to generate polished $P_2$ (blue layer in Figure 5). All the feature channels are set to 256. Additionally, in order to better ensure the integrity of low-level spatial characteristics, depthwise convolution is used when $P_2$ $P_3$ is down-sampled and pointwise convolution is used when $P_4$ $P_5$ is down-sampled. In this way, high-level features can extract richer spatial information, and, most importantly, can greatly enhance the effectiveness of the proposed FEM in later processing stages.

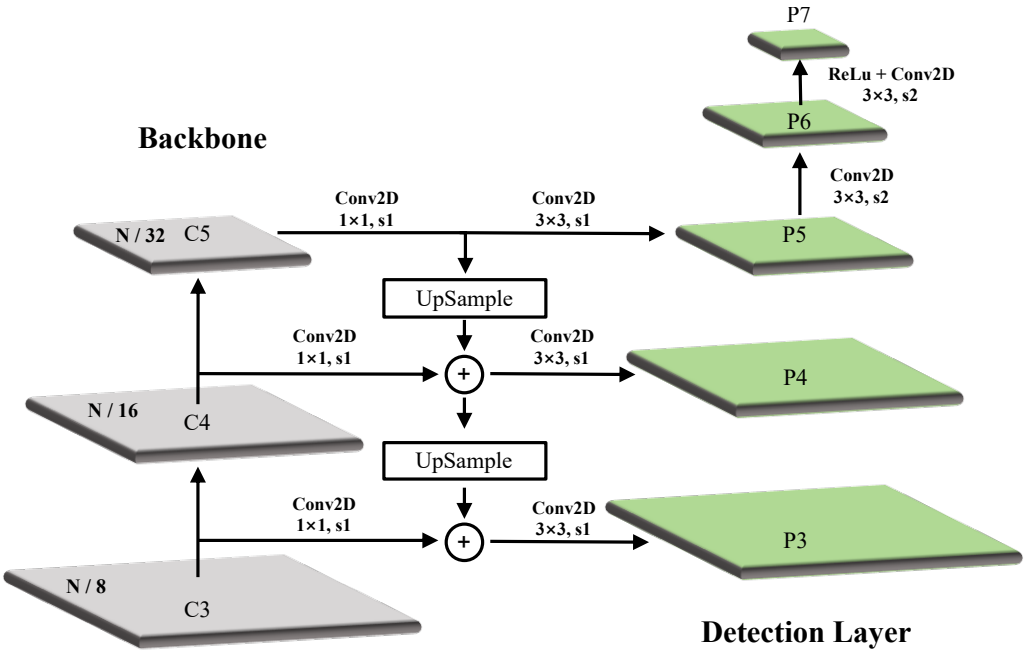



**Figure 4.** FPN structure of the RetinaNet. Conv2D $1 \times 1$, s1 refers to $1 \times 1$ convolution with stride set to 1. N denotes input image size. N/4 denotes the feature map resolution. $\oplus$ means addition.

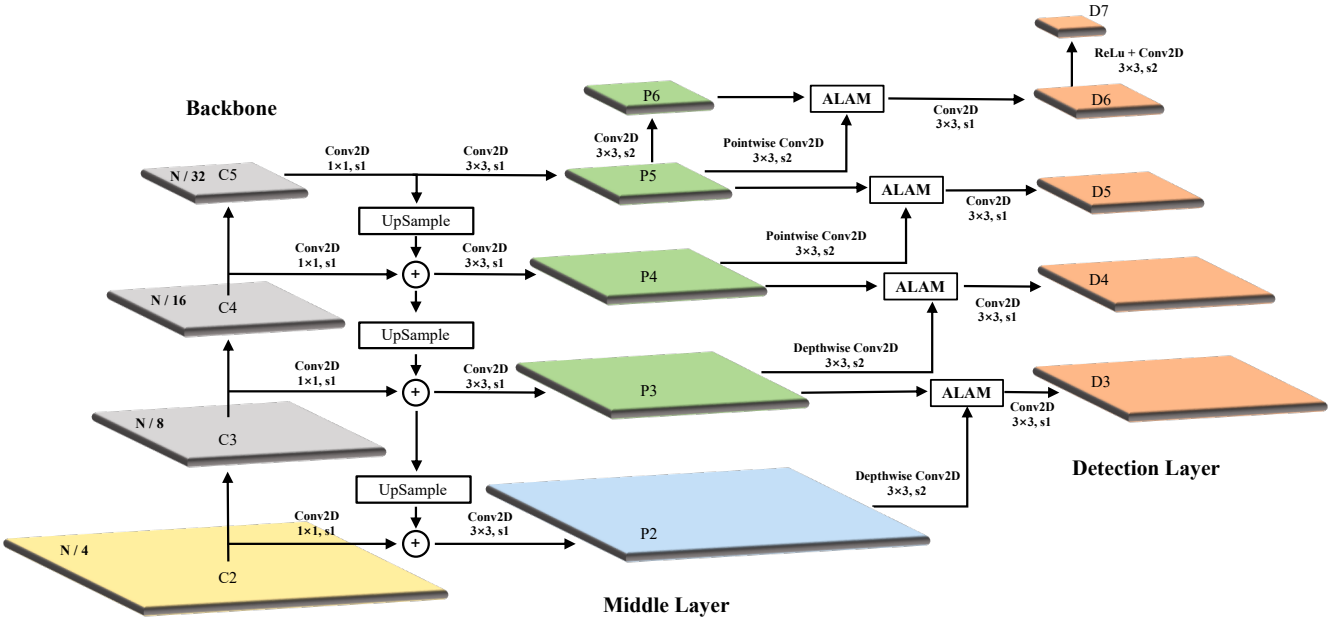



**Figure 5.** Structure of FG-FPN. Conv2D 1 × 1, s1 refers to 1 × 1 convolution with stride set to 1. N denotes input image size. N/4 denotes the feature map resolution. ⊕ means addition. ALAM refers to arhat layer attention module.

After strengthening the feature map of each scale, we tried to find a way to compensate for the differences between scales. Considering such an extreme situation, by building a feature pyramid with countless scales, we can cover the object scale transformation of various sizes, but this is obviously unrealistic. Therefore, we try to use the attention mechanism to generate a feature map with inter scale correlation. It is equivalent to further enhancing representation between scales.

#### 3.2.2. ALAM

The Arhat layer attention module (ALAM) is proposed to create a strong correlation between neighbor scales. Specifically, we regard the $\mathbf{P_i}$ as the query, and $\mathbf{P^D_{i-1}}$ (down-sampling output of $\mathbf{P_{i-1}}$) as the Key and Value. $\mathbf{P^D_{i-1}}, \mathbf{P_{i-1}} \in \mathbb{R}^{\mathbf{H}\times\mathbf{W}\times\mathbf{C}}$. Layer output $AHT_i$ is calculated as:

$$AHT_i = softmax(\frac{\mathbf{P_i} \otimes \mathbf{P^D_i} \cdot T}{\sqrt{H \cdot W \cdot C}}) \otimes \mathbf{P^D_{i-1}}, \tag{3}$$

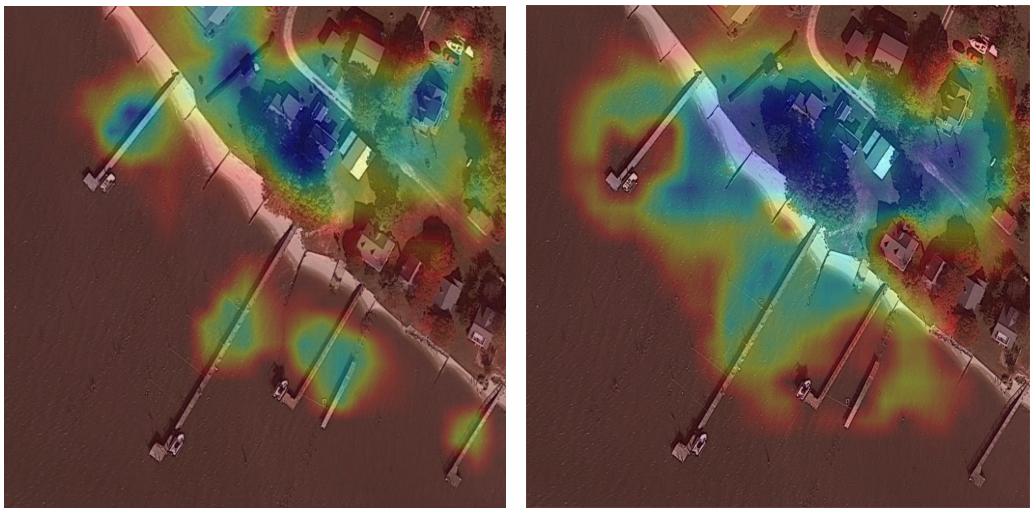
where, $\otimes$ denotes element-wise multiplication, $\mathbf{P^D_i} \cdot T$ denotes the transposition function, $i \in \{3, 4, 5\}$. $\mathbf{D_i}$ is generated from $ReLU(Conv2D(AHT_i))$. Through ALAM, $\mathbf{D_i}$ is equivalent to a correlation map between neighbor scales, which are significantly different from FPN and its variants. Because we use the attention mechanism to connect the upper and lower levels, we enrich them by building a pseudo pyramid to cover more scale information.

As shown in Figure 6, to further demonstrate the effectiveness of FG-FPN embedded with ALAM, we visualized the highest-resolution feature layers of BiFPN and FG-FPN after the same training iterations. It can be seen that FG-FPN has a more divergent heatmap distribution, but has a stronger response to the target. This is because ALAM allows for more implicated neurons in a single feature unit, which can cover a larger range of receptive fields.

### *3.3. FEM*

FEM is a multi-stage feature enhancement module because we perform a pre-prediction before accurate detection. As shown in Figure 3, FEM takes FG-FPN output $X_{LI}$ as the module input, where $LI$ denotes the layer index of different feature maps, assigning multi-scale feature cubes into two separate task brunches (i.e., classification and bounding box

regression). Classification and regression subnetworks are fully convolutional networks with a fixed number of stacked convolution layers (two in our experiments.) Note that here we only set one anchor in each DN and focal loss and smooth L1 loss are used for classification and bounding box regression, respectively. Softmax is used to generate category confidence and then the classification map (CM) is obtained. CM has the shape of $2 \times W \times H$. For each point $(i, j)$ in $CM$, $CM(i, j)$ saves category labels and corresponding confidence, which are recorded as $CM^L(i, j)$ and $CM^C(i, j)$, respectively. For predicted box (PB), $PB(i, j)$ preserves the box position vector and has a shape of $5 \times W \times H$. Finally, FEM re-inputs $X_{LI}$, PB, and CM into AFREM to complete feature representation enhancement.



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

**Figure 6.** Feature heatmap visualization of P3 in BiFPN [20] and D3 in FG-FPN. This experiment scenario is selected from baseline and AFRE-Net after 10K iterations of training on the DOTA dataset, respectively. It can be seen that FG-FPN has a stronger response to the targets' proposals, and can better distinguish the foreground and background. (**a**) Feature heatmap from BiFPN; (**b**) Feature heatmap from FG-FPN.

### *3.4. AFREM*

The pipeline of AFREM is shown in Figure 7. To adaptively achieve ERF transformation, we design the ERF transformation algorithm based on deformable convolutional networks (DCN) [22], making it more suitable for objects in AOOD scenes. Relative convolution network (Relative Conv) is proposed to mitigate the impact of background features and magnify foreground information in a self-learning manner.

#### 3.4.1. ERF Transformation

DCN pioneered the idea of changing the sampling position of the convolution kernel for each feature point in a self-learning manner. Outstanding works like oriented-reppoints and align-conv both achieve effective ERF transformation for remote sensing objects. However, the lack of control in the training process of the former leads to slow convergence of the network, and the lack of refined design of the latter leads to the inability of the sampling point to fall on the key area of the target of interest. Therefore, we try to use a more convenient way to make up for the two shortcomings at the same time.

Taking the convolution of $3 \times 3$ as an example, sampling offsets in both directions (18 values) are claimed at each point, making it possible to expand the ERF of DNs by providing learning capabilities in the sampling process. However, when applying DCN in AOOD, due to the larger object ratios and random orientations, intelligent ERF transformation faces challenges. As shown in Figure 8a, since PB is derived by OBB-guided regression, it basically fits the contour of the target. However, we hope that our sampling points (dark blue points in Figure 8a) can evenly fall on the target body, rather than just be limited to

a certain part of the target (ERF box with mapping to original input in Figure 8a) or the boundary of the target. Alignment convolution in [14] ignores the solution to this problem. In our experiments, we try our best to place each sampling point on the main body of the object, as this ensures that the learned knowledge is focused on the object itself rather than the background features. Hence, we have designed a sample point reassignment strategy, as shown in Figure 8b. We use a shrunk PB to constrain the sampling position, making the sampling point better located inside the rotated bounding box.

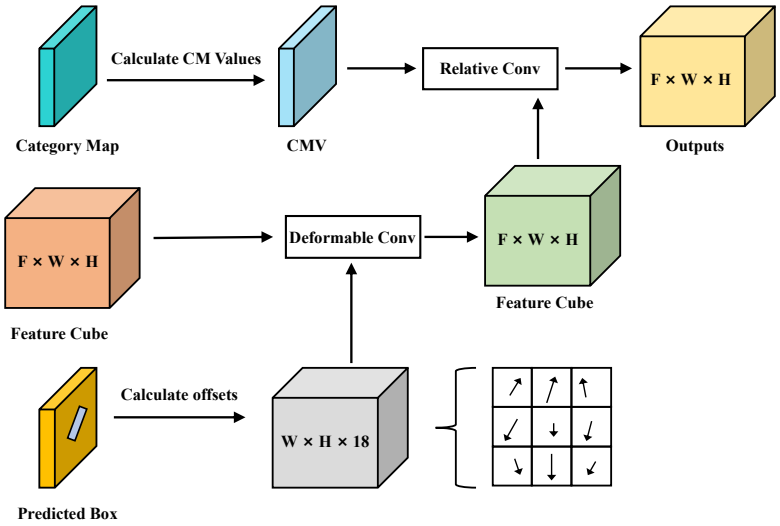



**Figure 7.** Pipeline of AFREM.

Given an $F \times W \times H$ ($F$ denotes feature channels) feature cubes $X_{LI}$, for each position $\mathbf{P}(x_0, y_0)$, we obtain ERF transformation results $FE_{LI}$ by

$$FE_{LI}(\mathbf{P}) = \text{Deformable}(X_{LI}, \text{offsets}(\mathbf{P})), \tag{4}$$

where offsets($\mathbf{P}$) is the position bias with the size of $W \times H \times 18$. The original offsets of $3 \times 3$ convolutions can be defined as $\mathcal{O}_\mathcal{G}$ = {(−1, -1), (−1, 0), (−1, 1), (0, −1), . . . , (1,1)},

$$\text{offsets}(\mathbf{P}) = \mathcal{O}_\mathcal{G} + \sigma_S, \tag{5}$$

where $\sigma_S$ is the shifting vector from the original sampling box to shrunk PB. As shown in Figure 8b, PB is scaled down with a shrinkage coefficient $\alpha$ to obtain shrunk PB (SPB). Let $(x, y, w, h, \theta)$ represent the PB in position $P$, shrunk box can be defined as $(x, y, \alpha w, \alpha h, \theta)$. We set $\alpha = 0.85$ in our experiments to suitably fit the target body. Since PB is derived from horizontal anchor box $V_a(w_a, h_a)$ by

$$SPB = \alpha V_a \cdot t_{w,h} \cdot R^T(\theta), \tag{6}$$

the sampling DN $S_P$ in position $P$ can be calculated as

$$S_p = \frac{1}{S}(\mathbf{P} + \frac{1}{K} \cdot SPB), \tag{7}$$

where $K$ is the kernel size and $S$ is the down-sampling strides of current feature maps. Then, we can obtain offsets($\mathbf{P}$) by

$$\text{offsets}(\mathbf{P}) = S_p - \mathcal{O}_\mathcal{G} - \sigma_S. \tag{8}$$

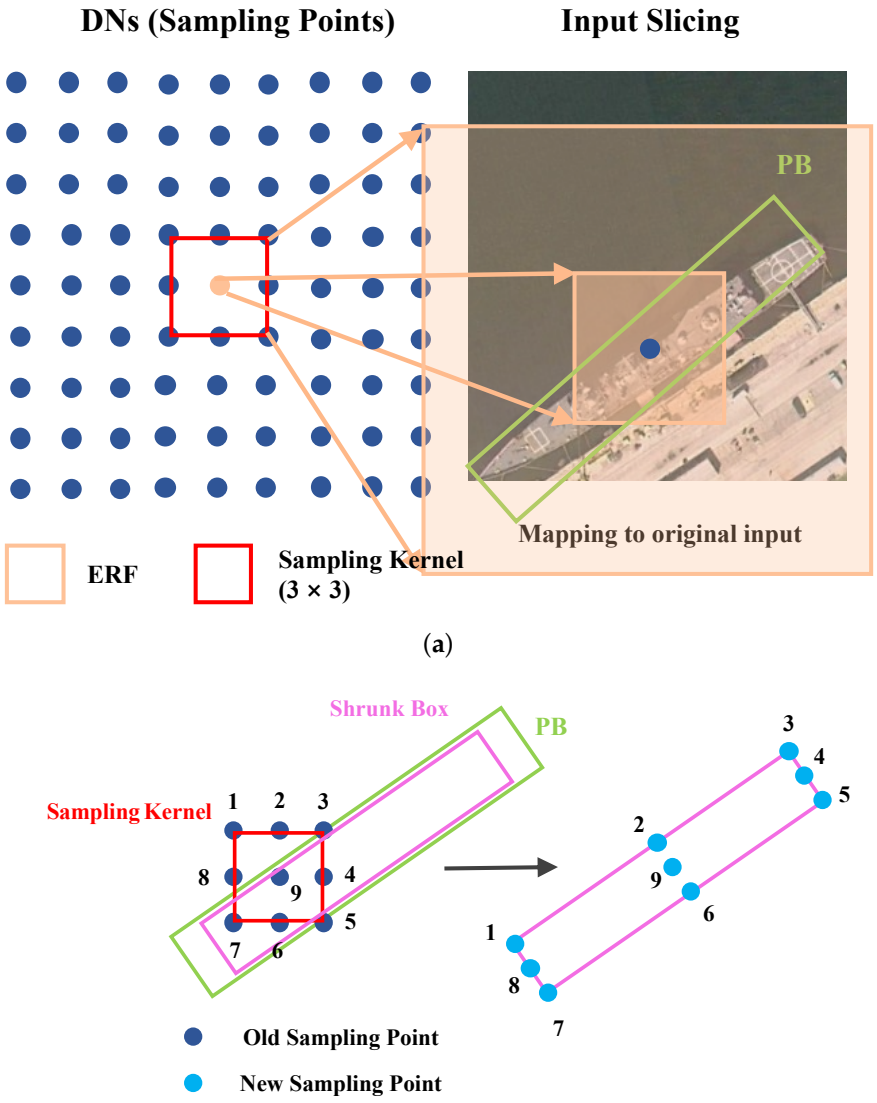



**Figure 8.** Illustration of ERF transformation. (**a**) Explanation diagram of the relationship between PB, sampling point, and ERF; (**b**) Identify sampling points through shrunk PB.

#### 3.4.2. Relative ConV

For a DN, to automatically suppress background and highlight foreground targets, it is necessary to enable it to learn which regions are more important during back-propagation. In a standard 2D convolution, the output feature $\mathbf{Y}$ can be obtained by

$$\mathbf{Y}_{\mathbf{LI}}(\mathbf{P}) = \sum_{\mathbf{g} \in \mathcal{O}_{\mathcal{G}}} \mathbf{W}(\mathbf{g}) \cdot \mathbf{X}_{\mathbf{LI}}(\mathbf{P} + \mathbf{g}), \tag{9}$$

where $\mathbf{P}$ is the position of each DN, $\mathbf{P} \in \{0, 1, \ldots, W-1\} \times \{0, 1, \ldots, H-1\}$, and $\mathbf{W}(\mathbf{g})$ is the kernel weights. $\mathbf{W}(\mathbf{g})$ is the weights in an ordinary $3 \times 3$ convolution layer, which is updated and iterated with the optimization of the model. In relative convolution, we define the output $\mathbf{Y}_R$ by

$$\mathbf{Y}_R(\mathbf{P}) = \sum_{\mathbf{g} \in \mathcal{O}_{\mathcal{G}}} (\delta RJ(\mathbf{g}) + 1)\mathbf{W}(\mathbf{g}) \cdot \mathbf{X}(\mathbf{P}), \tag{10}$$

and

$$RJ(\mathbf{g}) = \begin{cases} 0, & CM^L(\mathbf{g}) \neq CM^L((0,0)) \\ 1, & CM^L(\mathbf{g}) = CM^L((0,0)) \\ & \text{and} \quad CM^L(\mathbf{g}) \neq -1 \end{cases} \tag{11}$$

$CM^L$ is the category label obtained in FEM, $\delta$ is the accommodation coefficient and can be calculated by

$$\delta = CM^C + \eta, \tag{12}$$

where $CM^C$ is the category prediction confidence. Symbol $\eta = -0.2$ is used to control the learning intensity. Figure 9 illustrates the operating mode of the relative convolution. By using it, detectors can obtain enhanced representations by strengthening learning about foreground targets.

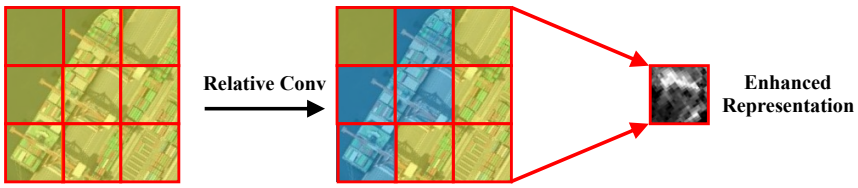



**Figure 9.** Illustration of Relative ConV.

### *3.5. Label Assignment Strategy*

Label assignment strategy (LAS) has a significant impact on the overall accuracy of the model since it encourages the model to select and refine positive samples reasonably and effectively during training. To enhance the robustness of our AFRE-Net, we have decided to optimize the LAS in our detector. In the FEM, we utilize intersection over union (IoU) as the matching metric. Specifically, we set the foreground and background thresholds for determining whether an anchor is positive or negative to 0.5 and 0.4, respectively. These thresholds help us differentiate between foreground and background regions effectively. In the RDM, we employ dynamic anchor learning (DAL) [31] for intelligent anchor selection, which aims to activate more positive anchors during the refinement process. The matching degree, denoted as $md$, is defined as follows:

$$md = \alpha \cdot sa + (1 - \alpha) \cdot fa - u^{\gamma} \tag{13}$$

$$u = |sa - fa| \tag{14}$$

where $sa$ denotes IoU of the anchor input, and $fa$ represents the IoU between the GT box and the regression box. The term $u$ is the absolute difference between the IoU of the anchor and the IoU between the GT box and the regression box. In our experiments, we set $\alpha$ to 0.3, and $\gamma$ to 5. If the IoU of an anchor is greater than $md$, it will be classified as positive; otherwise, it will be classified as negative. This approach allows us to effectively determine the positive anchors based on their IoU values.

## 4. Experimental Results and Analysis

### *4.1. Datasets*

Our AFRE-Net was assessed on three publicly available and challenging datasets, namely DOTA [49], HRSC2016 [50], and UCAS-AOD [51].

DOTA is an extensive dataset consisting of aerial images that capture complex scenes relevant to AOOD. It comprises a total of 2806 aerial images, with 1411 images for training, 458 images for validation, and 937 images for testing. These images contain a total of 188,281 instances belonging to 15 categories. The image size ranges from 800 × 800 to 4000 × 4000, and all instances are labeled with OBB, which exhibit variations in scales, aspect ratios, and orientations. To facilitate training, we divided the images into regular 1024 × 1024 patches with a stride of 200. The categories and corresponding IDs are as follows: Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Small

vehicle (SV), Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HA), Swimming pool (SP), and Helicopter (HC).

HRSC2016 is a high-resolution ship detection dataset that contains images collected from six international harbors. It consists of 1061 images, with image sizes ranging from 300 $\times$ 300 to 1500 $\times$ 900. The dataset includes 436 images for training, 541 images for validation, and 444 images for testing. All ship objects are labeled with OBB, and the substantial variation in ship sizes poses a significant challenge for detection.

UCAS-AOD is an aerial image dataset specifically designed for oriented aircraft and car detection. It comprises 1510 images, including 1000 airplane images and 510 car images. We randomly divided the dataset into training, validation, and test sets in a ratio of 5:2:3.

Additionally, to assess the scenario generalization capabilities of our AFRE-Net, we utilized the ICDAR-2015 [52] dataset as a benchmark for testing. This dataset consists of 1000 training images and 500 test images. The text boxes in this dataset are labeled with OBB and exhibit a very large aspect ratio, making them particularly challenging for detection.

### *4.2. Implementation Detail*

For all datasets, we only set one horizontal anchor with aspect ratios of {1}, and resize all images to 1024 $\times$ 1024. Data augmentation techniques, such as random flip, rotation, and HSV color space transformation, are employed. The training optimizer used is Adam, with the initial learning rate set to $5 \times 10^{-4}$. At each decay step, the learning rate is divided by six. We utilize ResNet50 as the backbone network, which has been pre-trained on ImageNet. For DOTA, the models are trained on a single RTX 3090, and the batch size is set to two. Regarding HRSC2016, the detector undergoes a total of 12 K iterations during training, with the learning rate decaying at 8 K and 11 K, respectively. We evaluate the performance using average precision (AP) as the metric, following the same definition as the PASCAL VOC 2012 object detection challenge [53]. Unless explicitly stated, mAP refers to $AP_{50}$.

### *4.3. Ablation Studies*

In this section, we conduct a series of experiments on DOTA and HRSC2016 to test our proposed AFRE-Net. We first verify the progressiveness of FG-FPN at the entire detector. Then, FEM is disassembled from the model to analyze its vital impact on overall performance. Finally, the respective capabilities of FRF expansion and RC are verified separately. Our ablation experiments demonstrate that when FG-FPN is combined with our meticulously designed FEM, our detector can achieve greater efficacy, thereby demonstrating the advantages of AFRE-Net.

To ensure fair comparisons, our baseline model adopts the same configuration as described in Section 4.2. Furthermore, we set the depth of the detection head (i.e., the rotation detection module in Figure 3) to a uniform value of 2, as it has a significant impact on the final detection result. In contrast to our AFRE-Net, which only utilizes one preset anchor with an aspect ratio of {1}, our baseline model employs three horizontal anchors with aspect ratios of {0.5, 1, 2} for matching objects. The results presented in Tables 1 and 2 demonstrate that our baseline model achieves an mAP of 68.2% on DOTA and 86.32% on HRSC2016.

#### 4.3.1. Effectiveness of Hyper-Parameter

The parameter $\eta$ in our model is used to deal with the weight imbalance caused by strengthening the learning of key areas in relative conv. As shown in Table 3, when $\eta$ is around $-0.2$, negative compensation can achieve better performance by relative conv.

**Table 1.** Comparison on DOTA test dataset. R-101 represents ResNet-101 (likewise for R-50), and H-104 denotes Hourglass-104.

| Methods | Backbone | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP | HC | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FR-O [49] | R-101 | 79.09 | 69.12 | 17.17 | 63.49 | 34.20 | 37.16 | 36.20 | 89.19 | 69.60 | 58.96 | 49.40 | 52.52 | 46.69 | 44.80 | 46.30 | 52.93 |
| R-DFPN [54] | R-101 | 80.92 | 65.82 | 33.77 | 58.94 | 55.77 | 50.94 | 54.78 | 90.33 | 66.34 | 68.66 | 48.73 | 51.76 | 55.10 | 51.32 | 35.88 | 57.94 |
| $R^2$CNN [36] | R-101 | 80.94 | 65.67 | 35.34 | 67.44 | 59.92 | 50.91 | 55.81 | 90.67 | 66.92 | 72.39 | 55.06 | 52.23 | 55.14 | 53.35 | 48.22 | 60.67 |
| RRPN [55] | R-101 | 88.52 | 71.20 | 31.66 | 59.30 | 51.85 | 56.19 | 57.25 | 90.81 | 72.84 | 67.38 | 56.69 | 52.84 | 53.08 | 51.94 | 53.58 | 61.01 |
| ICN [56] | R-101 | 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 |
| RetinaNet-O [16] | R-50 | 88.67 | 77.62 | 41.81 | 58.17 | 74.58 | 71.64 | 79.11 | 90.29 | 82.13 | 74.32 | 54.75 | 60.60 | 62.57 | 69.67 | 60.64 | 68.43 |
| RoI Trans. [57] | R-101 | 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 |
| CAD-Net [58] | R-101 | 87.80 | 82.40 | 49.40 | 73.50 | 71.10 | 63.50 | 76.70 | 90.90 | 79.20 | 73.30 | 48.40 | 60.90 | 62.00 | 67.00 | 62.20 | 69.90 |
| DRN [59] | 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 |
| $O^2$-DNet [35] | H-104 | 89.31 | 82.14 | 47.33 | 61.21 | 71.32 | 74.03 | 78.62 | 90.76 | 82.23 | 81.36 | 60.93 | 60.17 | 58.21 | 66.98 | 61.03 | 71.04 |
| DAL [31] | R-101 | 88.61 | 79.69 | 46.27 | 70.37 | 65.89 | 76.10 | 78.53 | 90.84 | 79.98 | 78.41 | 58.71 | 62.02 | 69.23 | 71.32 | 60.65 | 71.78 |
| SCRDet [60] | R-101 | 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 |
| $R^3$Det [37] | R-152 | 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 [14] | R-50 | 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 |
| $R^4$Det [61] | R-152 | 88.96 | 85.42 | 52.91 | 73.84 | 74.86 | 81.52 | 80.29 | 90.79 | 86.95 | 85.25 | 64.05 | 60.93 | 69.00 | 70.55 | 67.76 | 75.84 |
| Oriented Reppoints [23] | R-101 | 89.53 | 84.07 | 59.86 | 71.76 | 79.95 | 80.03 | 87.33 | 90.84 | 87.54 | 85.23 | 59.15 | 66.37 | 75.23 | 73.75 | 57.23 | 76.52 |
| AFRE-Net (ours) | R-101 | 89.34 | 85.74 | 53.23 | 75.96 | 79.22 | 81.03 | 87.88 | 90.86 | 83.82 | 87.08 | 65.95 | 67.33 | 76.52 | 73.06 | 64.52 | 77.44 |

[1] Best results for each category are in red. Second-best results achieved by our detector are labeled in blue.

**Table 2.** Performance comparisons with state-of-the-art AOOD methods on the test set of HRSC2016. NA denotes the number of preset anchors of RDM. mAP (07/12): VOC2007/VOC2012 metrics.

| Methods | Backbone | Size | NA | mAP (07) | mAP (12) | Params (M) | FLOPS |
|---|---|---|---|---|---|---|---|
| $R^2$CNN [36] | ResNet101 | 800 × 800 | 21 | 73.07 | - | - | - |
| * RRD [62] | VGG16 | 384 × 384 | 13 | 84.30 | - | 27.6 | 176 G |
| RoI Trans. [57] | ResNet101 | 512 × 800 | 20 | 86.20 | - | 55.1 | 200 G |
| $R^3$Det [37] | ResNet101 | 800 × 800 | 21 | 89.26 | 96.01 | 41.9 | 336 G |
| * R-RetinaNet [16] | ResNet101 | 800 × 800 | 121 | 89.18 | - | 35.8 | 236 G |
| * DCL [9] | ResNet101 | 800 × 800 | - | 89.46 | - | 49.6 | 472 G |
| GWD [33] | ResNet101 | 800 × 800 | - | 89.85 | 97.37 | 47.4 | 456 G |
| DAL [31] | ResNet101 | 800 × 800 | 3 | 89.77 | - | 36.4 | 216 G |
| DRN [59] | ResNet101 | 768 × 768 | - | 92.7 | - | - | - |
| S2A-Net [14] | ResNet50 | 512 × 800 | 1 | 90.17 | 95.01 | 38.6 | 198 G |
| AOPG [63] | ResNet101 | 800 × 800 | - | 90.34 | 96.22 | - | - |
| ReDet [8] | - | 800 × 800 | - | 90.46 | 97.63 | 31.6 | - |
| O-RCNN [64] | ResNet101 | 800 × 800 | - | 90.50 | 97.60 | 41.1 | 199 G |
| * Baseline | ResNet50 | 800 × 800 | 3 | 86.32 | 91.04 | 31.5 | 199 G |
| AFRE-Net (ours) | ResNet50 | 800 × 800 | 1 | **92.36** | 97.32 | 42.2 | 323 G |
| AFRE-Net (ours) | ResNet101 | 800 × 800 | 1 | 92.18 | **97.82** | 51.4 | 469 G |

[1] The instances of best detection performance are in **bold**. [2] * means that the precision and model complexity data source is from our local machine.

**Table 3.** Influence of hyperparameters $\eta$ on HRSC2016 dataset.

| $\eta$ | −0.45 | −0.35 | −0.25 | −0.20 | −0.15 | −0.10 | NaN |
|---|---|---|---|---|---|---|---|
| $AP_{50}$ | 87.26 | 89.33 | 91.08 | **92.16** | 88.54 | 89.34 | NaN |
| $\eta$ | 0.45 | 0.35 | 0.25 | 0.20 | 0.15 | 0.10 | 0.00 |
| $AP_{50}$ | 79.32 | 81.26 | 82.77 | 79.98 | 82.32 | 86.02 | 88.06 |

[1] The instances of best detection performance are in **bold**. [2] Both FG-FPN and DAL are used.

#### 4.3.2. Effectiveness of FG-FPN

Our baseline detector applies FPN as the neck to fuse multi-scale feature maps. As shown in Table 4, when replacing FPN with FG-FPN alone, the detector achieves accuracy gains of +1.01 and +0.92 on two datasets, respectively, proving that FG-FPN has a stronger feature fusion ability than FPN. In particular, it can provide low-level spatial information, which is very friendly for small targets. As can be seen in Table 1, the detection accuracy of SV has been greatly improved. However, at the same time, the replacement of FG-FPN also increased the number of parameters in the model, and we calculated that the size of the weight checkpoint file increased by 6.5 M. The detailed FG-FPN complexity is shown in Table 5. It can be seen that the introduction of FG-FPN has brought about a certain increase in model complexity, mainly caused by ALAM, as a large number of

intermediate parameters are generated during the calculation process of this attention mechanism, also leading to an increase in inference time. It should be noted that this group of testing experiments did not use any feature enhancement modules, including AFREM. As shown in Table 5, when FG-FPN is used alone, it can only slightly improve the overall detection accuracy of the model. However, when FG-FPN is combined with our proposed AFREM, it can fully release the model performance, as AFREM can utilize the features rich in low-level spatial information provided by FG-FPN, obtaining more robust target feature representations.

**Table 4.** Ablation study of embeddings in AFRE-Net on DOTA and HRSC2016 dataset.

| | **Baseline** | **Component Settings** | | | | **AFRE-Net** |
|---|---|---|---|---|---|---|
| FPN [17] | ✓ | | | ✓ | ✓ | |
| FG-FPN (ours) | | ✓ | ✓ | | | ✓ |
| FEM (ours) | | | ✓ | ✓ | | ✓ |
| DAL [31] | | | | | ✓ | ✓ |
| $mAP_{DOTA}$ | +0 | +1.01 | +6.92 | +4.81 | +1.54 | +9.01 |
| $mAP_{HRSC2016}$ | +0 | +0.92 | +5.12 | +3.79 | +1.26 | +6.04 |

[1] $\eta$ is set to $-2$.

**Table 5.** Module complexity on different datasets.

| **Datasets** | **Modules** | **Params (M)** | **FLOPS** | **Runtimes (s)** | **mAP (%)** |
|---|---|---|---|---|---|
| DOTA | B+FPN | 27.6 | 168 G | 0.42 | 71.27 |
| HRSC2016 | B+FPN | 27.6 | 168 G | 0.36 | 85.24 |
| UCAS-AOD | B+FPN | 27.6 | 168 G | 0.35 | 84.72 |
| DOTA | B+FG-FPN | 34.1 (+6.5) | 321 G | 0.79 | 72.28 (+1.01) |
| HRSC2016 | B+FG-FPN | 34.1 (+6.5) | 321 G | 0.68 | 86.16 (0.92) |
| UCAS-AOD | B+FG-FPN | 34.1 (+6.5) | 321 G | 0.68 | 86.88 (+2.16) |

[1] Runtimes refer to inference time, and B denotes backbone network. [2] AFREM was not used in this set of experiments.

#### 4.3.3. Effectiveness of FEM

FEM consists of two parts: ERF expansion and Relative ConV. We first verified how the overall detection accuracy of the detector changes when the entire FEM module is removed. As shown in the third and fourth control experiments in Table 4 (third and fourth columns), the use of different combinations of embeddings in detectors results in varying levels of detection accuracy. The combination of FG-FPN and FEM resulted in an astonishing mAP gain of +6.92 for the detector on DOTA, while FPN plus FEM achieves a +4.91 mAP improvement, which is also satisfactory. However, the former cannot be compared to the latter. Similar results also occurred on HRSC2016, where the combination of FG-FPN plus PEM achieves better performance, and improves the mAP by 5.12%.

In addition, to verify the contributions made by ERF transformation and Relative Conv in FEM, we conducted two comparative experiments, as shown in Table 6. It should be noted that both FG-FPN and DAL are used in these two comparative experiments. Inside FEM, since the size of feature cubes does not change, we only need to remove the other embedding when testing only one embedding. The experimental results show that when the two are combined, they can play a greater role. This is because by combining the two embeddings, DN not only can obtain the self-learning changes of the ERF, but also can adaptively learn the key areas in it and suppress the background information, thus obtaining better feature expression ability.

#### 4.3.4. Effectiveness of Label Assignment Strategy

In order to eliminate the impact of LAS in our experiment, we also conducted comparative experiments to verify the universality of our proposed methods. As shown in Table 6, when using DAL in the baseline, the accuracy of the detector on DOTA increased by 1.54%

and on HRSC06 by 1.26%, indicating that the optimization of LAS is significantly helpful in improving the overall detection accuracy. However, the experimental results show that the use of LAS does not affect the improvement of the model detection accuracy brought by FG-FPN and FEM. On the contrary, when the three are combined, the maximum gain can be achieved.

**Table 6.** Effects of FEM structure on DOTA an HRSC2016.

| | | **FEM Settings** | | **AFRE-Net** |
|---|---|---|---|---|
| ERF transformation | ✓ | | | ✓ |
| AlignConV [14] | | ✓ | | |
| Relative ConV | | ✓ | ✓ | ✓ |
| $mAP_{DOTA}$ | +3.77 | +5.65 | +2.89 | +9.01 |
| $mAP_{HRSC2016}$ | +2.85 | +4.26 | +4.33 | +6.04 |

[1] FG-FPN and DAL are both used in this experiments set.

### *4.4. Comparison with State-of-the-Art Detectors*

#### 4.4.1. Results on DOTA

We select some difficult scenarios as a demonstration of AFRE Net's detection capabilities. As shown in Figures 10 and 11, because our detector has improved the ability of feature expression, its confidence in the predicted output of the target has been greatly improved, and error detection has been effectively avoided (red circle in Figure 10a). Moreover, the detection ability for small targets has also been greatly improved.

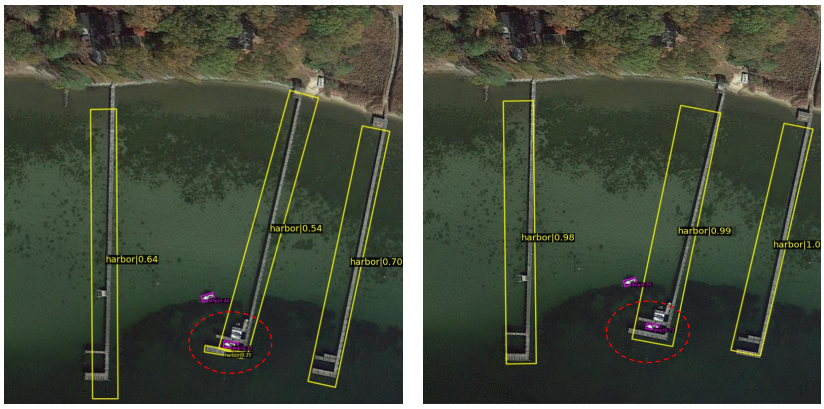
Figure 12 also shows some tough detection scenarios in AOOD (dense, small, large aspect ratio, chaotic, and orientation random). It can be seen that AFRE-Net is able to better cope with the above challenges. Compared with other state-of-the-art AOOD detectors shown in Table 1, our model outperforms the best $R^4$Det [61] by an mAP of 1.6%, and achieves mAP improvement of 9.01% over the baseline detector. Compared with the anchor-free reppoints, our AFRE-Net achieves better performance on most categories.

Surprisingly, the accuracy improvement ability of AFRE-Net on specific objects is impressive. For SP, HA, RA, SBF, and ST, AFRE-Net achieves improvements of 3.39%, 13.95%, 6.73%, 11.2%, and 12.76%, respectively over the baseline. This suggests that our proposed method has a more significant and prominent effect on improving the feature expression of targets with large aspect ratio scales. The first reason is that the target with a large aspect ratio is more likely to contain more area of background information within the rectangular box of its outer contour, increasing the likelihood of interference; Secondly, the original regular feature sampling mode makes it impossible to accurately collect all the spatial scale features of the target when representing the target with a large-scale aspect ratio. AFREM achieves finer feature extractions by accurately changing the sampling points. Lastly, it can be seen that AFRE-Net has a good accuracy improvement effect on small targets, because the application of FG-FPN improves the ability of the detector to capture features in a small space range.

#### 4.4.2. Results on HRSC2016

We evaluate the performance of our AFRE-Net on HRSC2016 with existing state-of-the-art AOOD detectors, which are divided into two categories, i.e., two-stage methods, such as $R^2$CNN [36], RRPN [55], $R^2$PN [65], RoI Trans. [57], and Gliding Vertex [66], and single-stage methods, such as DCL [9], DAL [31], DRN [59], and S2A-Net [14]. As shown in Table 2, our AFRE-Net outperforms all the detectors, especially towards two-stage methods, by a large gap up to 4.16%. Our AFRE-Net obtains an mAP of 92.36% under the condition that only ResNet50 is used, meaning that our model can achieve better feature extraction and detection results with fewer parameters of backbones. Compared with the baseline model, we improve 6.04% mAP with only one preset anchor in the FEM and RDM. In addition, Table 2 also shows the performance of our method in terms of model complexity and efficiency. The FG-FPN and AFREM has greatly improved the complexity of the model

and increased the inference time of the detector, but extensive experiments have proved that our method is powerful in improving the detection performance. We have achieved 15.8FPS inference speed on a single RTX3090, proving that the model has maintained certain efficacy while improving its performance.




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

**Figure 10.** Detection comparison between baseline detector and our AFRE-Net. AFRE-Net tends to obtain higher confidence score and more accurate predictions. (**a**) Wrong Detection by baseline; (**b**) Correct Detection by AFRE-Net.

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

**Figure 11.** Detection comparison between baseline detector and our AFRE-Net. AFRE-Net is more capable of generating small predictions. (**a**) Missed Detection by baseline; (**b**) Finer detection by AFRE-Net.

**Figure 12.** Visualization of some detection results on DOTA.

#### 4.4.3. Results on UCAS-AOD

The distribution of vehicle targets in UCAS-AOD is relatively dense, and the spatial size is small, making detection difficult. As shown in Table 7, our baseline detector only achieved an accuracy performance of 83.22% on car detection. However, after AFRE-Net is applied, the mAP is improved to 90.62%, and the overall mAP is promoted to 91.34%, which surpasses all other comparison methods.

**Table 7.** Results comparison with advanced detectors on UCAS-AOD dataset.

| Methods | Car | Airplane | mAP |
|---|---|---|---|
| Baseline | 0.8322 | 0.8643 | 0.8472 |
| * YOLOv3-O [26] | 0.7463 | 0.8952 | 0.8208 |
| Faster R-CNN-O [49] | 0.8687 | 0.8986 | 0.8836 |
| * DAL [31] | 0.8925 | 0.9049 | 0.8987 |
| * O-RCNN [64] | 0.8874 | 0.9123 | 0.9003 |
| * Oriented Reppoints [23] | 0.8951 | 0.9070 | 0.9011 |
| * ReDet [8] | 0.9034 | 0.9107 | 0.9079 |
| AFRE-Net (ours) | **0.9062** | **0.9143** | **0.9134** |

[1] The instances of best detection performance are in **bold**. [2] * means that the precision and model complexity data source is from our local machine.

#### 4.4.4. Results on ICDAR2015

To assess the robustness and generalization capability of our proposed AFRE-Net algorithm across various application scenarios, as well as to tackle annotation boxes with larger aspect ratio scales, we conducted training and testing on the ICDAR2015 dataset.

ICDAR2015 comprises challenging targets with significant variations in length and width, annotated in the oriented bounding box (OBB) format. As depicted in Table 8, our baseline model achieved an F-measure of 80.72 and a recall of 80.23%. Compared with other text detectors, such as EAST [67], $R^2$CNN [36], and $R^3$Det [37], AFRE-Net obtains the best recall performance at 88.82% and the best F-measure score at 86.27%, proving that our proposed solution has good migration application capabilities.

**Table 8.** Results comparison with text detector on ICDAR 2015.

| Methods | Recall | Precision | F-Measure |
|---|---|---|---|
| RRPN [55] | 82.17 | 73.23 | 77.44 |
| EAST [67] | 78.33 | 83.27 | 80.72 |
| $R^2$CNN [36] | 79.68 | 85.62 | 82.54 |
| $R^3$Det [37] | 81.64 | 84.97 | 83.27 |
| Baseline | 80.23 | 82.06 | 80.72 |
| AFRE-Net (ours) | 88.82 | 85.82 | 86.27 |

[1] Best results for each category are in red. Second-best results achieved by our detector are labeled in blue.

## 5. Conclusions

In this paper, we make several significant contributions to arbitrary-oriented object detection (AOOD). First, we identify the shortcomings and possible problems of the existing AOOD detectors in the structure design of feature extraction. Specifically, we point out that the existing models cannot automate DN learning and adjust ERF, and cannot adapt learning focus areas and suppress background information. To address these limitations, we conceive our detector AFRE-Net by designing a finer-grained feature fusion neck, and proposing ERF transformation and relative conv on this basis. These modifications enable the detector to acquire new capabilities for expressing object features. We validate the effectiveness of our algorithm on several remote sensing datasets and application scenarios. Extensive experimental results show that our method is effective and has a positive impact on the future design of feature representation enhancement strategies.

**Author Contributions:** Conceptualization, T.Z.; methodology, T.Z.; software, T.Z.; validation, T.Z.; formal analysis, T.Z.; investigation, T.Z.; resources, T.Z., X.S. and K.Z.; data curation, T.Z.; writing—original draft preparation, T.Z.; writing—review and editing, L.Z. and X.D.; visualization, T.Z.; supervision, J.S.; project administration, B.Z.; funding acquisition, B.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Key R&D Program of China (Grant No. 2021YFB3900502).

**Data Availability Statement:** For all source data and code, please contact us: zhangtianwei20@mails.ucas.ac.cn.

**Acknowledgments:** We sincerely appreciate the constructive comments and suggestions of the anonymous reviewers, which have greatly helped to improve this paper.

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

## References

1. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
2. Chen, Y.; Zhu, L.; Ghamisi, P.; Jia, X.; Li, G.; Tang, L. Hyperspectral images classification with Gabor filtering and convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2355–2359. [CrossRef]
3. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
4. Ma, T.Y.; Li, H.C.; Wang, R.; Du, Q.; Jia, X.; Plaza, A. Lightweight Tensorized Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5544816. [CrossRef]
5. Bai, L.; Liu, Q.; Li, C.; Ye, Z.; Hui, M.; Jia, X. Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620214. [CrossRef]
6. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
7. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
8. 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.
9. Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; Yan, J. Dense Label Encoding for Boundary Discontinuity Free Rotation Detection. *arXiv* **2021**, arXiv:2011.09670.
10. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 677–694.
11. Qian, W.; Yang, X.; Peng, S.; Guo, Y.; Yan, J. Learning modulated loss for rotated object detection. *arXiv* **2019**, arXiv:1911.08299.
12. Zhou, L.; Wei, H.; Li, H.; Zhao, W.; Zhang, Y.; Zhang, Y. Arbitrary-oriented object detection in remote sensing images based on polar coordinates. *IEEE Access* **2020**, *8*, 223373–223384. [CrossRef]
13. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
14. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. *arXiv* **2020**, arXiv:2008.09397.
15. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 2965–2974.
16. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
17. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
18. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9197–9206.
19. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 7036–7045.
20. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
21. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10213–10224.

22. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
23. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1829–1838.
24. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
25. Zhang, X.; Wan, F.; Liu, C.; Ji, R.; Ye, Q. Freeanchor: Learning to match anchors for visual object detection. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 147–155.
26. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
27. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
28. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
30. 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, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
31. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic anchor learning for arbitrary-oriented object detection. *arXiv* **2020**, arXiv:2012.04150.
32. Zhang, T.W.; Dong, X.Y.; Sun, X.; Gao, L.R.; Qu, Y.; Zhang, B.; Zheng, K. Performance releaser with smart anchor learning for arbitrary-oriented object detection. *Caai Trans. Intell. Technol.* **2022**. [CrossRef]
33. 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, PMLR, Virtual, 18–24 July 2021; pp. 11830–11841.
34. Liu, J.; Zheng, H. EFN: Field-Based Object Detection for Aerial Images. *Remote Sens.* **2020**, *12*, 3630. [CrossRef]
35. Wei, H.; Zhou, L.; Zhang, Y.; Li, H.; Guo, R.; Wang, H. Oriented objects as pairs of middle lines. *arXiv* **2019**, arXiv:1912.10694.
36. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2cnn: Rotational region cnn for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
37. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv* **2019**, arXiv:1908.05612.
38. Ming, Q.; Miao, L.; Zhou, Z.; Dong, Y. Cfc-net: A critical feature capturing network for arbitrary-oriented object detection in remote sensing images. *arXiv* **2021**, arXiv:2101.06849.
39. Yu, Y.; Yang, X.; Li, J.; Gao, X. A cascade rotated anchor-aided detector for ship detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 5600514. [CrossRef]
40. Zhang, T.; Sun, X.; Zhuang, L.; Dong, X.; Gao, L.; Zhang, B.; Zheng, K. FFN: Fountain Fusion Net for Arbitrary-Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5609913. [CrossRef]
41. Cohen, T.; Welling, M. Group equivariant convolutional networks. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 19–24 June 2016; pp. 2990–2999.
42. Worrall, D.E.; Garbin, S.J.; Turmukhambetov, D.; Brostow, G.J. Harmonic networks: Deep translation and rotation equivariance. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5028–5037.
43. Weiler, M.; Hamprecht, F.A.; Storath, M. Learning steerable filters for rotation equivariant cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 849–858.
44. Weiler, M.; Cesa, G. General e (2)-equivariant steerable cnns. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
45. Marcos, D.; Volpi, M.; Komodakis, N.; Tuia, D. Rotation equivariant vector field networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5048–5057.
46. Zhou, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Oriented response networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 519–528.
47. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point set representation for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9657–9666.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NA, USA, 27–30 June 2016; pp. 770–778.
49. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.

50. 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 2017), Porto, Portugal, 24–26 February 2017.
51. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; IEEE: New York, NY, USA, 2015; pp. 3735–3739.
52. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 competition on robust reading. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, 23–26 August 2015; IEEE: New York, NY, USA, 2015; pp. 1156–1160.
53. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
54. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
55. 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]
56. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 150–165.
57. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
58. 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]
59. 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, Seattle, WA, USA, 13–19 June 2020; pp. 11207–11216.
60. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
61. Sun, P.; Zheng, Y.; Zhou, Z.; Xu, W.; Ren, Q. R4 Det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. *Image Vis. Comput.* **2020**, *103*, 104036. [CrossRef]
62. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.s.; Bai, X. Rotation-sensitive regression for oriented scene text detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5909–5918.
63. 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]
64. 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.
65. 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]
66. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
67. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. East: An efficient and accurate scene text detector. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5551–5560.