*electronics*

MDPI

*Article*

# Arbitrary-Oriented Object Detection in Aerial Images with Dynamic Deformable Convolution and Self-Normalizing Channel Attention

**Yutong Zhang [1,2], Chunjie Ma [1,2], Li Zhuo [1,2,*] and Jiafeng Li [1,2]**

[1] Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China
[2] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
* Correspondence: zhuoli@bjut.edu.cn

**Abstract:** Objects in aerial images often have arbitrary orientations and variable shapes and sizes. As a result, accurate and robust object detection in aerial images is a challenging problem. In this paper, an arbitrary-oriented object detection method for aerial images, based on Dynamic Deformable Convolution (DDC) and Self-normalizing Channel Attention Mechanism (SCAM), is proposed; this method uses ReResNet-50 as the backbone network to extract rotation-equivariant features. First, DDC is proposed as a replacement for the conventional convolution operation in the Convolutional Neural Network (CNN) in order to cope with various shapes, sizes and arbitrary orientations of the objects. Second, SCAM embedded into the high layer of ReResNet-50, which allows the network to enhance the important feature channels and suppress the irrelevant ones. Finally, Rotation Regions of Interest (RRoI) are generated based on a Region Proposal Network (RPN) and a RoI Transformer (RT), and the RoI-wise classification and bounding box regression are realized by Rotation-invariant RoI Align (RiRoI Align). The proposed method is comprehensively evaluated on three publicly available benchmark datasets. The mean Average Precision (mAP) can reach 80.91%, 92.73% and 94.1% on DOTA-v1.0, DOTA-v1.5 and HRSC2016 datasets, respectively. The experimental results show that, when compared with the state-of-the-arts methods, the proposed method can achieve superior detection accuracy.

**Keywords:** aerial images; arbitrary-oriented object detection; dynamic deformable convolution; self-normalizing channel attention; ReResNet-50



**Citation:** Zhang, Y.; Ma, C.; Zhuo, L.; Li, J. Arbitrary-Oriented Object Detection in Aerial Images with Dynamic Deformable Convolution and Self-Normalizing Channel Attention. *Electronics* **2023**, *12*, 2132. https://doi.org/10.3390/electronics12092132

Academic Editor: Silvia Liberata Ullo

Received: 23 March 2023
Revised: 19 April 2023
Accepted: 27 April 2023
Published: 6 May 2023



## 1. Introduction

Object detection in aerial images is used to locate objects of interest on the ground and identify their categories; this has become an important research topic in the field of computer vision. Objects in natural images can maintain their orientations due to gravity, while objects in aerial images often have arbitrary orientations. The shape and scale of objects in these aerial images change dramatically, making object detection in aerial images a challenging problem [1–3]. In recent years, the Convolutional Neural Network (CNN) has made important breakthroughs. The CNN is widely used in various visual tasks, especially in the field of aerial images [4–8]. Correspondingly, several aerial image datasets have been released; these promote the continuous advance of related research work.

Existing aerial image object detection methods are generally based on the object detection framework used for natural images [9–12]. By elaborately designing specific mechanisms to cope with object rotation changes, including loss functions [13,14], enlarging the scale of training samples with various rotation changes [4,15], rotation invariant and rotation variant feature extraction, detection robustness and accuracy have been improved significantly. These methods usually adopt convolution operations with fixed weights; these make the network unable to cope with the drastic changes in the scale, orientation

and shape of objects effectively. In addition, the categories of objects in aerial images are complex and diverse, and the semantic feature representation capabilities of the existing detection methods are insufficient, which often affect the detection performance.

With the development of remote sensing technology, the resolution and the file sizes of aerial images are constantly increasing. Due to the limited budget, limited logistical resources and the power consumption in some aerospace systems, including satellites and aircraft, Zhang et al. [16] proposed a hardware architecture for the CNN-based aerial images object detection model. To see the issue from a different perspective, Li et al. [17] proposed a lightweight convolutional neural network. Recent advancements in remote sensing have widened the range of applications for 3D Point Cloud (PC) data. This data format poses several new issues concerning noise levels, sparsity and required storage space; as a result, many recent works address PC problems using deep learning solutions due to their capability to automatically extract features and achieve high performances [18,19].

In light of the above problems, an arbitrary-oriented aerial image object detection method based on Dynamic Deformable Convolution (DDC) and the Self-normalizing Channel Attention Mechanism (SCAM) is proposed. This method adopts ReResNet-50 as the backbone network for rotation-invariant feature extraction [8].

The main contributions of this study are summarized as follows:

- DDC is proposed. This can dynamically adjust the weights of convolution kernels according to the input image. The conventional convolution operation is replaced with DDC to cope with arbitrary-oriented objects.
- SCAM is proposed to enhance the important feature channels while suppressing the irrelevant ones. It is placed at the higher layer of the backbone network in order to enhance the semantic feature representation capability and, thus, improve detection accuracy.
- Experimental results on three challenging datasets (DOTA, HRSC2016 and UCAS-AOD) show that the proposed method can achieve state of the art detection performance.

A brief overview of the related work is given in Section 2. Section 3 introduces the proposed arbitrary-oriented object detection method. Section 4 reports the experimental results and analysis. Finally, conclusions are drawn in Section 5.

## 2. Related Work

Most object detection methods use a Horizontal Bounding Box (HBB) to denote the location of objects in aerial images. However, because of the dense distribution of objects in aerial images, the large vertical-horizontal ratio and arbitrary orientations, the use of HBB always contains some background regions; this causes interference in classification tasks, and the predicted object position is not accurate enough as a result. To cope with these challenges, aerial image object detection is usually formulated as an oriented object detection task by using an Oriented Bounding Box (OBB). The comparison of HBB and OBB is shown in Figure 1.

It can be seen from Figure 1 that, when compared with HBB, OBB can denote the position of objects with arbitrary orientations more precisely. Therefore, OBB is usually used for arbitrary-oriented object detection in aerial images.

Current mainstream arbitrary-oriented object detectors can be divided into three categories: single-stage detectors [20–23], two-stage detectors [24–27] and refine-stage detectors [28–31]. These are introduced separately below.

### 2.1. Single-Stage Object Detector

Single-stage object detectors have a high detection speed that is generally based on the YOLO series [11], SSD [32], and other single-stage frameworks. Yang et al. [13] proposed a regression loss based on Gaussian Wasserstein distance to solve the problems of boundary discontinuity and its inconsistency between detection performance evaluation and loss function in arbitrary-oriented object detection. The authors further simplified the network model [33] based on the Gaussian model and the Kalman filter, in which a loss function

was proposed for rotating object detection. The model can achieve trend-level alignment with SkewIoU loss instead of the strict value level identity.

Aerial images often use OBB for object detection. This leads to a large number of rotation-related parameters and anchor configurations in the anchor-based detection methods. Zhao et al. [27] proposed a different polar detector, which located an object by its center point, directed it by four polar angles and measured it using the polar ratio system. Yi et al. [26] applied the horizontal keypoint-based object detector to arbitrary-oriented object detection tasks. The experimental results showed that these two different methods can achieve the rapid detection of arbitrary-oriented objects, but that the detection accuracy needs to be improved.


(**a**) HBB (**b**) OBB

**Figure 1.** The comparison of (**a**) HBB and (**b**) OBB.

### *2.2. Two-Stage Object Detector*

Compared with single-stage detectors, two-stage object detectors often have high detection accuracy but with a lower detection speed. Currently, two-stage object detectors have become the mainstream in arbitrary-oriented object detectors.

In order to eliminate the loss discontinuity at the boundary of rotating object, Yang et al. [28,30] proposed an IoU-smooth L1 loss by the combination of IoU and smooth L1 loss. It is a rotating IoU loss without differentiability. Inspired by this, Yang et al. [34] further proposed a new rotation detection baseline to address the boundary problem by transforming angular prediction from a regression problem to a classification task with little accuracy degradation.

Ding et al. [4] proposed a multi-stage detector based on Cascade RCNN, which contains Rotation Regions of Interest Learner (RRoI Learner) and RRoI warping, to transform HRoI to RRoI. Han et al. [8] proposed Rotation-invariant RoI Align (RiRoI Align) to extract rotation-invariant features from rotation-equivariant features according to the orientation of RoI. These methods lead to confused sequential marking points when using rotating anchors. Therefore, Xu et al. and Wang et al. [6,7,35] employed quadrilateral masks to describe arbitrary-oriented objects precisely; they also used sequential label points to solve the above problems.

Xie et al. [31] proposed a two-stage arbitrary-oriented object detection framework that includes oriented RPN, an oriented RCNN header and a detection header that can refine RROI.

In general, the two-stage object detector can effectively deal with objects with various rotation angles, can improve the detection robustness and accuracy by designing the network structure and can accommodate loss function, feature fusion strategy, attention mechanism and so on.

### *2.3. Refine-Stage Object Detector*

To obtain higher detection accuracy, many refined one-stage or two-stage object detectors are proposed; these can not only improve detection speed, but also obtain higher detection accuracy.

To address the problem of feature misalignment, Yang et al. [21] designed a Feature Refining Module (FRM) that uses feature interpolation to obtain the position information of refining anchor points and reconstructed feature maps to realize feature alignment. Han et al. [36] proposed a single-shot alignment network for oriented object detection that aims at alleviating the inconsistency between the classification score and location accuracy via deep feature alignment. To overcome the boundary discontinuity issue, Yang et al. [37] proposed a regression-based object detector that uses Angle Distance and Aspect Ratio Sensitive Weighting (ADARSW) to make the detector sensitive to angular distance and object aspect ratio. Different from refined one-stage detectors, the second stage of a refined two-stage detector is used for proposal classification and regression, allowing it to obtain a higher detection accuracy.

These methods can improve detection robustness and accuracy by elaborately designing network structure, loss function and feature extraction strategy to effectively cope with the various rotation angles of objects. As large numbers of methods are constantly proposed, the experimental data begin to randomize, seriously affecting the accuracy of the experimental results. To address the problem of experimental data, Giordano et al. [38] go in depth regarding methods, resources, experimental settings and performance results to observe and study all the aspects that derive from the stages. However, there are still some problems to be solved. When designing the network structure, more complex modules, such as feature fusion and attention mechanism, are usually adopted, inevitably increase model complexity. To solve the above problems, in this paper, a two-stage arbitrary-oriented object detection method is proposed based on DDC and SCAM. This method can dynamically adjust convolution kernel parameters according to the input image and enhance the semantic feature representation capability, thus improving detection performance.

## 3. Proposed Method

The following section will describe the overall architecture of the proposed method and the implementation details of the ReResNet-50, DDC, SCAM and RoI-wise classification and bounding box regression.

### *3.1. Overall Framework*

The framework of the proposed two-stage arbitrary-oriented object detection method in aerial images is shown in Figure 2. For the input image, ReResNet-50 is used as the backbone network to extract rotation-equivariant features. The conventional convolution operation is replaced with the proposed DDC, which dynamically adjusts the offset weights of convolution kernels by obtaining the weights of convolution kernels and increases the offset direction. DDC is used to cope with the drastic orientation, scale and shape variations of objects, and to enhance the representation capability of features. Under the cyclic group $C_N$, the rotation-equivariant feature maps with the size $(K,\ N,\ H,\ W)$ have $N$ orientation channels, and each orientation channel corresponds to an element in $C_N$. SCAM is proposed and introduced into the high layer of the backbone network to improve the semantic feature representation capability. Then, RPN is used to generate HRoIs, followed by an RT that transforms HRoIs to RRoIs. Finally, RiRoI alignment with rotation invariance is used to realize object orientation classification and bounding box regression, which includes spatial alignment and orientation alignment to ensure that RRoIs with different orientations produce completely rotation-invariant features.

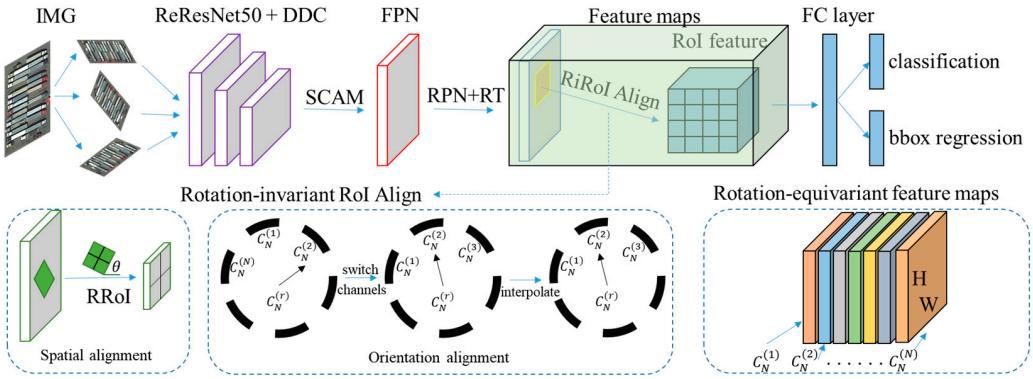


**Figure 2.** The overall architecture of the proposed method. ReResNet-50 is adopted as the backbone network to extract rotation-equivariant features and the conventional convolution operation is replaced with the proposed DDC. Under the cyclic group $C_N$, the rotation-equivariant feature maps with the size $(K,\ N,\ H,\ W)$ have $N$ orientation channels, and each orientation channel corresponds to an element in $C_N$. SCAM is proposed to enhance the representative capability of high layer semantic features. Finally, RPN and RT are adopted to generate RRoI, followed by a RiRoI Align to realize RoI-wise classification and bounding box regression.

The following section will describe the implementation details of the ReResNet-50, DDC, SCAM and RoI-wise classification and bounding box regression.

### *3.2. ReResNet-50 Network*

Existing object detectors usually adopt CNN as the backbone network to automatically extract multi-scale features. As shown in Figure 2, ReResNet-50 with rotation-equivariance is used as the backbone network; this is based on ResNet-50.

All layers of the backbone are re-implemented with rotation-equivariant networks based on e2CNN [39], including convolution, pooling, normalization and non-linearities. Considering the computational complexity, ReResNet-50 and ReFPN are only equivariant to the discrete features. Unlike the conventional feature maps, the rotation-equivariant feature maps $\Gamma$ with the size of $(K, N, H, W)$ have $N$ channels:

$$\Gamma = \left\{ \Gamma^{(i)} | i\epsilon\{1, 2, \ldots, N\} \right\} \tag{1}$$

where the feature maps of each orientation channel $\Gamma^{(i)}$ correspond to an element in $C_N$.

Compared with a conventional CNN backbone network, a rotation equivariant backbone network can obtain abundant directional information by extracting features from different directions and share convolution kernels with different rotation weight coefficients; this makes the model more robust and smaller in size.

As shown in Figure 2, DDC and SCAM are proposed and introduced into the low and high layer of the backbone network, respectively, to improve the feature representative capability of the network.

### *3.3. Dynamic Deformable Convolution*

In the field of object detection, deformation modeling is a fundamental problem. It aims to produce translation-invariant and rotated-invariant features. A Deformable Convolutional Network (DCN) [40] is a simple, efficient and end-to-end solution for modeling dense spatial transformations; it tends to obtain the offset by adding a standard convolutional layer branch whose convolution kernel has the same spatial resolution as the current convolutional layer. To further prove the validity of deformable convolution, DCNv2 [41] further improves the modeling capability and shows better performance in object detection tasks.

Inspired by DCN, DDC is proposed in this paper. DDC aims to dynamically integrate multiple convolution kernels, generate new weight parameters and better learn the offset when compared to DCN. Its network structure is shown in Figure 3.

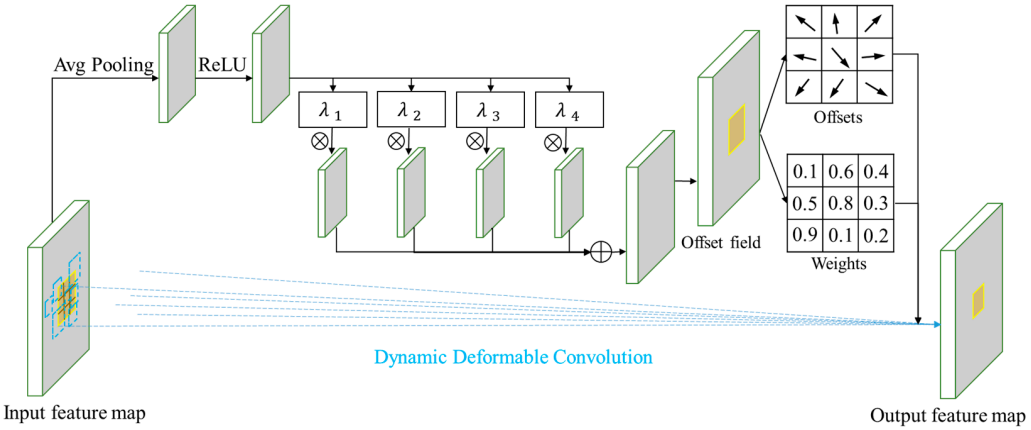


**Figure 3.** The structure of DDC. Input feature maps are fed into a set of average pooling, convolutional layers and Rectified Linear Units (ReLU) to obtain the weight $\lambda$, add weighted conventional convolution layer branch, which dynamically adjusts the offset weights of convolution kernels by obtaining the weights of convolution kernels and increasing the offset direction.

As illustrated in Figure 3, input feature maps are fed into a set of average pooling, convolutional layers and Rectified Linear Units (ReLU) to obtain the weight $\lambda$. Compared with DCN, a weighted conventional convolution layer branch is added that dynamically adjusts the offset weights of convolution kernels by obtaining the weights of convolution kernels and increasing the offset direction. The specific calculation method of offset weights is as follows:

$$offset(D, W) = x\left(\sum_{l=1}^{4} \lambda_i \times \omega_{Fl}\right) \tag{2}$$

where $x$ represents the input feature maps, $\omega$ represents the parameters of the convolutional layer, $\lambda$ represents the learned weights and $offset(D, W)$ represents the learned offset directions and weights.

The DDC modulation process of each point $\theta$ on the output feature map $y$ is expressed by the following equation:

$$y(\theta) = \sum_{k=1}^{K} x(\theta + \Delta\theta_k + offset(D)) \times \omega(m) \times offset(W) \tag{3}$$

where $x$ and $y$ represent the input and output feature maps and the values of $offset(D)$ and $offset(W)$ are produced by a branch of the $offset(D, W)$, which represents the learned offset directions and weights, respectively. $k$ is the $k$-th point in the convolution kernel; the range of its value is $[1, K]$. $\Delta\theta_k$ and $\omega(m)$ are the learnable offsets and modulation parameters of the $k$-th location; the range of $\omega(m)$ value is $[0, 1]$ and $\Delta\theta_k$ is a real number with unconstrained range. Through the above operation, the output feature maps can keep the direction deviation and the adjustable weights are added to the convolution kernel parameters; this can effectively cope with various complex objects and enhance the feature representation capability.

### *3.4. Self-Normalizing Channel Attention Mechanism*

SCAM is proposed to further improve the representative capability of semantic features; its network structure is shown in Figure 4. SCAM aims to enhance the feature channels important in the object detection tasks while suppressing the irrelevant ones. SCAM uses Scaled exponential Linear Units (SeLU) [42] and Global Average Pooling

(GAP); these can avoid possible gradient explosion. To avoid high complexity, SCAM only considers direct interaction between each channel and its S-nearest neighbors. Meanwhile, one-dimensional convolutional kernel size is adaptively selected.

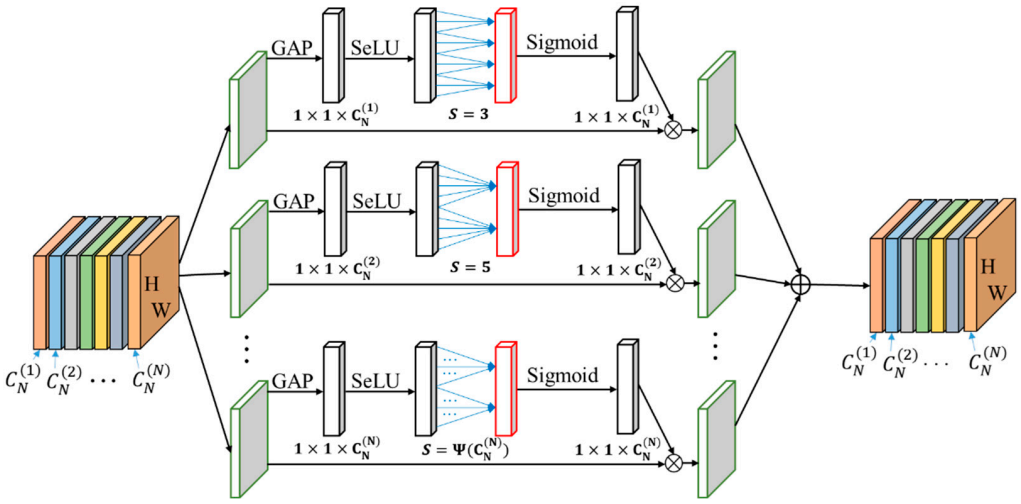


**Figure 4.** The structure of SCAM. Given the aggregated features obtained by SeLU and GAP, SCAM generates channel weights of size $S$, where $S$ is adaptively determined via a mapping of channel dimension C.

Let $\gamma \epsilon \mathbb{R}^{W \times H \times C_N}$ denote the output of one convolution layer, where $H, W$ and $C_N$ are height, width and number of feature channels, respectively. The specific calculation of the channel weights in SCAM is expressed as follows:

$$\theta = Sigmoid(F_{SCAM}(W_1, W_2) \times GAP(\gamma)) \tag{4}$$

where $GAP(\gamma)$ is a GAP operation in the channel direction and $F_{SCAM}(W_1, W_2)$ represents SCAM operation. $W_1$ and $W_2$ represent model parameters. To make parameters self-normalized, the specific equation of $F_{SCAM}(W_1, W_2)$ and $GAP(\gamma)$ are as follows:

$$F_{SCAM}(W_1, W_2) \times GAP(\gamma) = SeLU\left(W_1 \times \frac{1}{WH} \times \sum_{i=1,j=1}^{W,H} \gamma_{ij}\right) \times W_2 \tag{5}$$

where $SeLU\left(W_1 \times \frac{1}{WH} \times \sum_{i=1,j=1}^{W,H} \gamma_{ij}\right)$ represents $SeLU$ operation in the channel direction.

Next, by analyzing the Efficient Channel Attention (ECA) module [43], including a squeeze module for aggregating global spatial information and an efficient excitation module for modeling cross-channel interaction, this paper proposes a module that can adaptively select one-dimensional convolutional kernel sizes. The simplest mapping is a linear function $\psi_1(S) = k \times S - b$. According to the channel dimension, which is usually set to power of two, extend the linear function $\psi_1(S)$ to a non-linear one, denoted as $\psi_2(S) = 2^{k \times S - b}$,to overcome the deficiency of linear function representation capability.

Then, given channel dimension ${C_N}^{(n)}$, the specific calculation of the kernel size S is expressed as follows:

$$S = \left| \frac{\log_2 {C_N}^{(n)}}{a} + \frac{k}{a} \right|_{odd} \tag{6}$$

where $|t|_{odd}$ represents the nearest odd number of $t$. In this paper, $k$ and $a$ are set to two and one, respectively. High-dimensional channels can have longer range interaction when using this formula, while low-dimensional channels undergo shorter range interaction by using non-linear mapping. As a result of the interaction between feature channels, the representation capability of feature channels can be improved.

### *3.5. RoI-Wise Category Classification and Bounding Box Regression*

As shown in Figure 2, RPN is adopted to generate Horizontal RoI (HRoI) parallel to the coordinate axis; then, an RoI Transformer (RT) is adopted to convert HRoI into RRoI with rotation characteristics [4]. Traditional RoI pooling can only deal with candidate regions parallel to coordinate axes. In this paper, RRoI Pooling is further used to pool the rotating bounding box. Rotation-invariant features cannot be extracted from rotational equivariant features by using RRoI warping directly. As can be seen in Figure 2, RiRoI Align [8] is adopted. According to the RRoI bounding box of the spatial dimension, it can align the features of the orientation dimension by cyclically switching the orientation channel and interpolating the features.

The RoI transformer contains two parts: RRoI Learner and RRoI warping. RRoI learner attempts to learn RRoI from the HRoI. During the model's training, the input HRoI is matched with the rotating OBB and the related parameters of RRoI are decoded from it. RRoI Warping extracts rotation-invariant features through RRoI parameters. RiRoI Align includes two parts: spatial alignment and directional alignment. For an $RRoI\ (x,\ y,\ w,\ h,\ \varphi)$, spatial alignment warps it from the feature maps $f$ to produce rotation-invariant region features $f_R$ in the spatial dimension, which is consistent with RiRoI Align. To ensure that RRoI with different orientations can produce rotation-invariant features, it performs orientation alignment in the orientation dimension. The specific calculation of the output region features $\hat{f_R}$ is expressed as follows [8]:

$$\hat{f_R} = \mathcal{I}(CH(f_R, [\varphi N/2\pi]), \varphi) \tag{7}$$

where $\varphi$ is an index set to $[\theta N/2\pi]$ and $CH$ and $\mathcal{I}$ represent the switching channels and feature interpolation operations, respectively. For each location in the feature maps, both the orientation with the strongest response and the features from all orientations are preserved using this formula.

## 4. Experimental Results and Analysis

In order to verify the effectiveness of the proposed method, massive comparative experiments are conducted on three benchmark datasets (DOTA [1], HRSC2016 [2] and UCAS-AOD [3]). In this section, the datasets and evaluation criteria are introduced, then the experimental results are reported, and finally the results are analyzed.

### *4.1. Datasets and Evaluation Criteria*

The following section will describe the details of the datasets and evaluation criteria.

#### 4.1.1. Datasets

DOTA is the largest dataset for arbitrary-oriented object detection in aerial images; it is comprised of 2806 large aerial images from different sensors and platforms. Objects in DOTA exhibit a wide variety of scales, orientations and shapes. As can be seen in Table 1, the fully annotated DOTA benchmark dataset contains 188,282 instances, each of which is labeled by an arbitrary quadrilateral. These images are then annotated by experts using 15 object categories. The short names for categories are defined as: 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). Half of the original images are randomly selected as the training set, 1/6 as the validation set and 1/3 as the testing set. A series of $1024 \times 1024$ patches are cropped from the original images with a stride of 824.

**Table 1.** Comparison among DOTAv1.0, HRSC2016 and UCAS-AOD datasets in aerial images.

| Dataset | Category | Image Quantity | | | Instance Quantity | Image Size |
|---|---|---|---|---|---|---|
| | | Train | Val | Test | | |
| DOTAv1.0 [1] | 15 | 1404 | 467 | 935 | 188,282 | 800 × 800–4000 × 4000 |
| HRSC2016 [2] | 1 | 436 | 181 | 444 | 2976 | 300 × 300–1500 × 900 |
| UCAS-AOD [3] | 2 | 1004 | 106 | 400 | 14,596 | 659 × 1280 |

HRSC2016 is a challenging ship detection dataset with OBB annotations. The dataset contains 1061 aerial images, with sizes ranging from 300 × 300 to 1500 × 900. It includes 436, 181 and 444 images in the training, validation and test set, respectively. All images are resized to 1024 × 1024.

UCAS-AOD is an aerial aircraft and car detection dataset. UCAS-AOD contains 1510 aerial images, with sizes around 659 × 1280 pixels, with two categories of 14,596 instances. In line with [1,44], 1110 images are randomly selected for training and 400 for testing.

#### 4.1.2. Evaluation Criteria

In the object detection evaluation criteria, mean Average Precision (mAP) is generally used to evaluate detection accuracy. When calculating mAP, some indicators, such as recall, precision and average precision (AP), are required. *Precision* and *Recall* can be formulated as:

$$Precision = TP/(TP + FP) \tag{8}$$

$$Recall = TP/(TP + FN) \tag{9}$$

where $TP$, $FP$ and $FN$ denote the number of true positives, false positives and false negatives, respectively.

$IoU$ refers to the intersection ratio between the prediction box and Ground Truth (GT), and is usually used to measure the overlapped degree between the prediction box and GT. $IoU$ can be formulated as:

$$IoU = \frac{Bounding\ Box\ \cap\ Ground\ Truth}{Bounding\ Box\ \cup\ Ground\ Truth} \tag{10}$$

where $Bounding\ Box\ \cap$ Ground Truth denotes the intersection of the predicted detection results and GT and $Bounding\ Box\ \cup$ Ground Truth denotes their union.

$IoU$ is set as the standard threshold for evaluating position accuracy. $AP_i$ of a certain detection category is calculated from the area of the precision–recall curve. The $AP_i$ can be formulated as:

$$AP_i = \int_0^1 P_i(r)dr \tag{11}$$

The detection performance of the detector is the average of the $AP_i$ of all categories. Therefore, mAP can be formulated as:

$$mAP = \sum_{i=1}^{N} AP_i/N \tag{12}$$

where $N$ is the total number of categories and $AP_i$ is the $AP$ of a certain detection category.

The Matthew's correlation coefficient (MCC) is calculated from the correlation coefficient between the true and the predicted value. It can denote the sensitivity of imbalanced data. $MCC$ can be formulated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{13}$$

### *4.2. Implementation Details*

The proposed network model was built, trained and tested on the Pytorch platform. The hardware configuration was: Ubuntu 16.04 with an Intel Xeon(R) E5-2602 v4 CPU, 16 G memory and an Nvidia RTX 2080Ti GPU. ReResNet-50 was implemented based on the mmclassification, available at https://github.com/open-mmlab/mmclassification (accessed on 30 July 2020). ReResNet-50, pre-trained on ImageNet-1K with an initial learning rate of 0.1, was used as the backbone network in this paper. All models were trained for 100 epochs and the learning rate was divided by 10 at {30, 60, 90} epochs; the batch size is set to 256 [8].

To enlarge the scale of the training samples, three-scale data samples are provided. Random rotation operation was adopted for training and testing to improve the training performance of the network.

### *4.3. Comparison with the State-of-the-Arts Methods*

In order to verify the effectiveness of the proposed method in this paper, the method was compared with state-of-the-arts methods on three publicly datasets (DOTAv1.0, HRSC2016 and UCAS-AOD).

#### 4.3.1. Evaluation on DOTA Benchmark Dataset

We compared the proposed arbitrary-oriented object detection method with 23 state of the art methods on the DOTA-v1.0 dataset, as categorized by single-, two-, and refine-stage methods. The single-stage methods included $O^2$-Dnet [24], DRN [25], BBAVectors [26], PolarDet [27], GWD [13] and KFIOU [33]. The two-stage methods included RoI-Trans [4], SCRDet [28], GlidingVertex [7], Mask-OBB [29], CenterMap [6], CSL [34], RSDet-II [35], SCRDet++ [30], ReDet [8] and Oriented RCNN [31]. The refine-stage methods included CFCNet [20], R3Det [21], CFA [22], DCL [37], RIDet [23], S2Anet [36] and KLD [14]. The comparison results using different methods are shown in Table 2, in which R-101 denotes ResNet-101 (likewise for R-50, R-152). RX-101, ReR-50 and H-104 denote ResNeXt101, ReReNet-50 and Hourglass-104, respectively. The top two detection accuracies are marked in red and blue. The experimental data of the other methods are cited in the references.

**Table 2.** Comparison results using state of the art methods on the DOTA-v1.0 dataset. The top two detection accuracies are marked in red and blue.

| | Method | Backbone | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP | HC | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single-stage | $O^2$-Dnet [24] | H-104 | 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 [25] | H-104 | 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 |
| | BBAVectors [26] | R-101 | 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 |
| | PolarDet [27] | R-101 | 89.65 | 87.07 | 48.14 | 70.97 | 78.53 | 80.34 | 87.45 | 90.76 | 85.63 | 86.87 | 61.64 | 70.32 | 71.92 | 73.09 | 67.15 | 76.04 |
| | GWD [13] | R-152 | 86.96 | 83.88 | 54.36 | 77.53 | 74.41 | 68.48 | 80.34 | 86.62 | 83.41 | 85.55 | 73.47 | 67.77 | 72.57 | 75.76 | 73.40 | 76.30 |
| | KFIOU [33] | R-152 | 89.46 | 85.72 | 54.94 | 80.37 | 77.16 | 69.23 | 80.90 | 90.79 | 87.79 | 86.13 | 73.32 | 68.11 | 75.23 | 71.61 | 69.49 | **77.35** |
| Refine-stage | CFCNet [20] | R-101 | 89.08 | 80.41 | 52.41 | 70.02 | 76.28 | 78.11 | 87.21 | 90.89 | 84.47 | 85.64 | 60.51 | 61.52 | 67.82 | 68.02 | 50.09 | 73.50 |
| | R3Det [21] | R-152 | 89.80 | 83.77 | 48.11 | 66.77 | 78.76 | 83.27 | 87.84 | 90.82 | 85.38 | 85.51 | 65.67 | 62.68 | 67.53 | 78.56 | 72.62 | 76.47 |
| | CFA [22] | R-152 | 89.08 | 83.20 | 54.37 | 66.87 | 81.23 | 80.96 | 87.17 | 90.21 | 84.32 | 86.09 | 52.34 | 69.94 | 75.52 | 80.76 | 67.96 | 76.67 |
| | DCL [37] | R-152 | 89.26 | 83.60 | 53.54 | 72.76 | 79.04 | 82.56 | 87.31 | 90.67 | 86.59 | 86.98 | 67.49 | 66.88 | 73.29 | 70.56 | 69.99 | 77.37 |
| | RIDet [23] | R-50 | 89.31 | 80.77 | 54.07 | 76.38 | 79.81 | 81.99 | 89.13 | 90.72 | 83.58 | 87.22 | 64.42 | 67.56 | 78.08 | 79.17 | 62.07 | 77.62 |
| | S2Anet [36] | R-50 | 88.89 | 83.60 | 54.74 | 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 |
| | KLD [14] | R-152 | 89.92 | 85.13 | 59.19 | 81.33 | 78.82 | 84.38 | 87.50 | 89.80 | 87.33 | 87.00 | 72.57 | 71.35 | 77.12 | 79.34 | 78.68 | **80.63** |
| Two-stage | RoI-Trans [4] | 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 |
| | SCRDet [28] | 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 |
| | GlidingVertex [7] | R-101 | 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 |
| | Mask-OBB [29] | RX-101 | 89.56 | 85.95 | 54.21 | 72.90 | 76.52 | 74.16 | 85.63 | 89.85 | 83.81 | 86.48 | 54.89 | 69.64 | 73.94 | 69.06 | 63.32 | 75.33 |
| | CenterMap [6] | R-101 | 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 [34] | R-152 | 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 |
| | RSDet-II [35] | R-152 | 89.93 | 84.45 | 53.77 | 74.35 | 71.52 | 78.31 | 78.12 | 91.14 | 87.35 | 86.93 | 65.64 | 65.17 | 75.35 | 78.74 | 63.31 | 76.34 |
| | SCRDet++ [30] | R-152 | 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 |
| | ReDet [8] | ReR-50 | 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 |
| | OrientedRCNN [31] | R-50 | 89.84 | 85.43 | 61.09 | 79.82 | 79.71 | 85.35 | 88.82 | 90.88 | 86.68 | 87.73 | 72.21 | 70.80 | 82.42 | 78.18 | 74.11 | 80.87 |
| | Ours | ReR-50 | 88.92 | 86.21 | 61.93 | 80.73 | 75.71 | 86.25 | 88.93 | 91.13 | 88.99 | 87.09 | 71.79 | 67.69 | 82.78 | 80.12 | 75.12 | **80.91** |

As shown in Table 2, the proposed method achieves a mAP of 80.91%, outperforming the other methods. It exceeds the second Oriented RCNN method by 0.04%, exceeds ReDet

by 0.81% and obtains the optimal or sub-optimal results in 12/15 categories. Compared with the sub-optimal results, the AP values of eight object categories are increased, including baseball diamond (85.95 to 86.21), bridge (61.09 to 61.93), large vehicle (86.06 to 86.25), ship (88.82 to 88.93), basketball court (88.77 to 88.99), harbor (82.42 to 82.78), swimming pool (79.71 to 80.12) and helicopter (74.67 to 75.12).

In order to show the effectiveness of the proposed method more intuitively, the visualized detection results on the DOTA-v1.0 dataset are depicted in Figure 5. It can be clearly observed that the proposed method achieves accurate bounding box regression in detecting densely packed and arbitrary oriented objects, and that it can capture the edge information of the rotated objects better and obtain a higher detection accuracy.

**Figure 5.** Visualized detection results using different methods on DOTA dataset. The red and purple boxes in the figure represent the predicted results and the ground truth, respectively.

#### 4.3.2. Evaluation on HRSC2016 Benchmark Dataset

We compared the proposed method with 13 state of the art methods on the HRSC2016 dataset, including RC1&RC2 [45], RRPN [15], R2PN [46], RRD [47], RoI-Trans [4], Gliding Vertex [7], R3Det [21], CSL [34], DAL [48], GWD [13], S2anet [36], ReDet [8] and Oriented RCNN [31]. The comparison results using different methods are shown in Table 3. The results are all evaluated with the VOC2007 metric for fair comparison. The experimental data of the other methods are cited in the references.

**Table 3.** Comparison results using state of the art methods on the HRSC2016 dataset. The top two detection accuracies are marked in red and blue.

| Method | Backbone | mAP (07) |
|---|---|---|
| RC1 & RC2 [45] | VGG16 | 75.70% |
| RRPN [15] | ResNet101 | 79.08% |
| R2PN [46] | VGG16 | 79.60% |
| RRD [47] | VGG16 | 84.30% |
| RoI-Trans [4] | ResNet101 | 86.20% |
| Gliding Vertex [7] | ResNet101 | 88.20% |
| R3Det [21] | ResNet50 | 89.26% |
| CSL [34] | ResNet152 | 89.60% |

**Table 3.** *Cont.*

| Method | Backbone | mAP (07) |
|---|---|---|
| DAL [48] | ResNet101 | 89.80% |
| GWD [13] | ResNet101 | 89.85% |
| S2anet [36] | ResNet50 | 90.17% |
| ReDet [8] | ReResNet-50 | 90.46% |
| Oriented RCNN [31] | ResNet101 | 90.50% |
| Ours | ReResNet-50 | 92.73% |

As shown in Table 3, the proposed method achieves an mAP of 92.93% under VOC2007 metrics, outperforming the other methods. It exceeds the second Oriented RCNN method by 2.23%. The proposed method can improve detection accuracy significantly, especially for the type of ship.

#### 4.3.3. Evaluation on UCAS-AOD Benchmark Dataset

We compared the proposed method with eight state of the art methods on the UCAS-AOD dataset, including Yolov3 [11], RetinaNet-O [11], DAL [48], S2anet [36], RoI-Trans [4], R3Det [21], ReDet [8] and Faster-RCNN [10]. The comparison results using different methods are shown in Table 4. To ensure fair comparison, we reimplemented them with the same parameters. The experimental data of the other methods are cited in the references.

**Table 4.** Comparison results using state of the art methods on the UCAS-AOD dataset. The top two detection accuracies are marked in red and blue.

| | Method | Backbone | Plane | Car | mAP |
|---|---|---|---|---|---|
| Single-stage | Yolov3 [11] | DarkNet53 | 89.5% | 74.6% | 82.1% |
| | RetinaNet-O [11] | ResNet101 | 90.5% | 84.6% | 87.6% |
| | DAL [48] | ResNet101 | 90.5% | 89.3% | 89.9% |
| | S2anet [36] | ResNet50 | 96.5% | 83.5% | 90.0% |
| | R3det [21] | ResNet50 | 95.4% | 85.9% | 90.7% |
| Two-stage | Faster-RCNN-O [10] | ResNet50 | 89.9% | 86.9% | 88.4% |
| | RoI-Trans [4] | ResNet101 | 89.9% | 88.0% | 89.0% |
| | R-Faster-RCNN [10] | ResNet50 | 95.2% | 87.6% | 91.4% |
| | ReDet [8] | ReResNet-50 | 95.6% | 88.9% | 92.3% |
| | Ours | ReResNet-50 | 96.4% | 91.8% | 94.1% |

As shown in Table 4, the proposed method achieves an mAP of 94.1%, outperforming the other methods. It exceeds the second ReDet method by 1.8%. The proposed method obtains the best results for the type of car; this exceeds the second DAL method by 2.5%. The proposed method also obtains the sub-optimal results for the type of plane, which indicates that the proposed method is also robust for small objects. This also demonstrates the good generalization capability of the proposed method.

### *4.4. Ablation Studies*

In order to further verify the influence of each part of the proposed method on the final detection performance, we conducted ablation experiments on the HRSC2016 dataset. Table 5 shows the comparison results of the detection performance obtained by using different modules and network structures. The baseline method adopts ResNet-50 as the backbone network, then uses RPN and RT to generate RRoI, followed by RroI Align to produce features for RoI-wise classification and bounding box regression. The same parameter setting was used during the training process.

**Table 5.** The effect of different components of the proposed method on detection performance.

| Baseline | ReResNet-50 | RiRoI | SCAM | DDC | MCC | mAP (07) |
|---|---|---|---|---|---|---|
| √ | | | | | 65.05% | 88.03% |
| √ | √ | | | | 68.57% | 90.47% |
| √ | √ | √ | | | 72.10% | 91.53% |
| √ | √ | √ | √ | | 72.23% | 92.31% |
| √ | √ | √ | | √ | 69.43% | 92.27% |
| √ | √ | √ | √ | √ | 73.40% | 92.73% |

Compared with ResNet-50, ReResNet-50 obtained enriched orientation information by generating features from multiple directions using ReResNet-50 as the backbone network. It can be seen from Table 5 that mAP values can be improved by 2.44% and MCC values can be improved by 3.52%. RiRoI Align shows significant improvements due to its orientation alignment mechanism; when compared with RroI Align, mAP values can be improved by 1.06% and MCC values can be improved by 3.53%. RroI warping can only align features in the spatial dimension; though the orientation dimension remains misaligned, RiRoI Align can extract completely rotation-invariant features.

It can be seen in Table 5 that; SCAM contributes more to the detection performance improvement than DDC. With SCAM, mAP values can be improved by 0.78% and MCC values can be improved by 0.13%. This is because SCAM can enhance the important feature channels while suppressing the irrelevant ones, effectively enhancing the high-level semantic features of aerial images by using SeLU and GAP to avoid the possible gradient explosion. SCAM also can adaptively select one-dimensional convolutional kernel sizes. Compared with the conventional convolution operation, when the DDC layer obtains the offset by adding a standard convolutional layer branch, mAP can be improved by 0.64%; this proves that DDC can dynamically adjust the weights of convolution kernels according to the input image, effectively dealing with the arbitrary-oriented objects. When all modules are added simultaneously, the mAP value can be improved to 92.73% and the MCC value can be improved to 73.40%. The experimental results comprehensively prove the effectiveness of these proposed modules.

Additionally, in order to compare the effect of the different components more intuitively, the visualized detection results and feature maps on HRSC2016 dataset are depicted in Figure 6. It can be clearly observed that, when compared with the baseline, SCAM allows the network to enhance the feature channels important to the detection tasks, and that DDC can capture the edge information of the rotated objects better. The proposed method achieves accurate bounding box regression in detecting arbitrary oriented objects, allowing it to achieve better detection performance.

### *4.5. Discussion*

DOTA is a very challenging dataset. It includes the complexity of the aerial image and the large number of cluttered, rotated and small objects. We compared the proposed method with other state of the art methods on DOTA, as shown in Table 2. Since different methods use different image resolutions, network structures, training strategies and various tricks, we cannot make absolutely fair comparisons. In terms of overall performance, our method has achieved the best performance so far, at around 80.91%.

The HRSC2016 and UCAS-AOD datasets contain lots of large aspect ratio ship, plane and car instances with arbitrary orientation; this poses a huge challenge to the positioning accuracy of the detector. Experimental results in Tables 3 and 4 show that our model achieves state of the art performances, at around 92.73% and 94.1%, respectively.

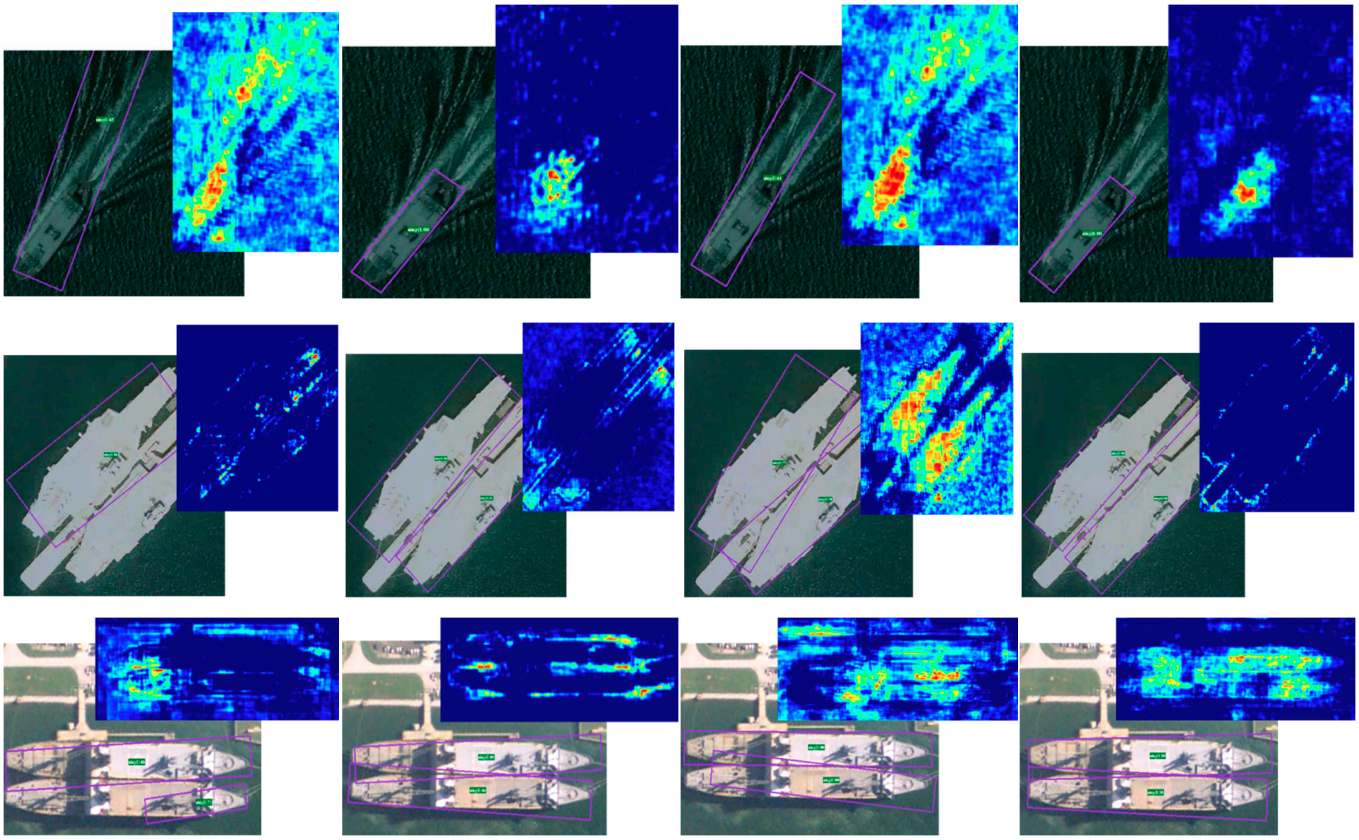


**Figure 6.** Qualitative comparisons using different components of the proposed method on HRSC2016. The purple boxes in the figure represent the predicted results. (**a**) the baseline method using ReResNet-50 as the backbone network, and using RiRoI Align to extract completely rotation-invariant features; (**b**) using the baseline method with the addition of the SCAM module; (**c**) using the baseline method with the addition of the DDC module; (**d**) using the baseline method with the addition of the SCAM module and DDC module.

## 5. Conclusions

The objects in aerial images often have arbitrary orientations and variable shapes and sizes. Despite the performance of CNN in aerial images, object detection has made important breakthroughs; however, accurate and robust object detection in aerial images remains a challenging problem. In this paper, an arbitrary-oriented object detection method in aerial images based on DDC and SCAM is proposed. The experimental results demonstrate that:

(1) Compared with the ResNet-50, ReResNet-50 can obtain enriched orientation information by generating features from multiple directions, while using RiRoI Align extracts rotation-invariant features from features.

(2) Compared with the conventional convolution operation, DDC can dynamically adjust the weights of convolution kernels according to the input image in order to enhance the feature representation capability.

(3) SCAM can effectively improve the semantic representation capability of high-level features; this can improve detection performance of arbitrary-oriented objects in aerial images.

Extensive experiments demonstrate that our method can achieve state of the art performances on the DOTA-v1.0, HRSC2016, and UCAS-AOD datasets.

Starting from these premises, the analysis highlights that there is still plenty of room for improvements. In future work, we will extend the method to identify roof types and geomorphological types, improve the test images' georeferences and further improve the

generalization capability of the proposed method by making our method more sensitive to arbitrary-oriented objects' angular distance and aspect ratio. We will also seek to enhance the practical application value of the method, reduce the model size, test images obtained directly by flying the UAV to supplement the research and improve the ability of the algorithm to extract the real objects.

**Author Contributions:** Conceptualization, L.Z. and Y.Z.; methodology, L.Z., Y.Z., C.M. and J.L.; validation, J.L.; formal analysis, J.L.; investigation, Y.Z. and C.M.; resources, J.L.; data curation, L.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, L.Z.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the National Natural Science Foundation of China under Grant 61871006, in part by the R&D Program of Beijing Municipal Education Commission (No. KZ202210005007), in part by the Beijing Natural Science Foundation (No. L211017) and in part by the General Program of Beijing Municipal Education Commission (No. KM202110005027).

**Data Availability Statement:** In this study, the DOTA dataset was downloaded from https://captain-whu.github.io/DOTA/dataset.html (accessed on 26 January 2018), the HRSC2016 dataset from https://sites.google.com/site/hrsc2016/ (accessed on 26 December 2016), and the UCAS-AOD dataset from https://hyper.ai/datasets/5419 (accessed on 26 January 2015).

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

## References

1. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
2. 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 International Conference on Pattern Recognition Applications & Methods (ICPRAM), Porto, Portugal, 24–26 February 2017; pp. 324–331.
3. 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 IEEE International Conference on Image Processing (ICIP), Quebéc City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
4. Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
5. 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]
6. Wang, J.; Yang, W.; Li, H.-C.; Zhang, H.; Xia, G.-S. Learning Center Probability Map for Detecting Objects in Aerial Images. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *59*, 4307–4323. [CrossRef]
7. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.; Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef]
8. Han, J.; Ding, J.; Xue, N.; Xia, G.-S. Redet: A rotationequivariant detector for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 2786–2795.
9. Yang, X.; Yan, J.; Tao, H. On the arbitrary-oriented object detection: Classification based approaches revisited. *arXiv* **2020**, arXiv:2003.05597.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
11. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
12. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
13. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. *arXiv* **2021**, arXiv:2101.11952.
14. 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. *arXiv* **2021**, arXiv:2106.01883.
15. Ma, J.; Shao, W.; Hao, Y.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
16. Zhang, N.; Wei, X.; Chen, H.; Liu, W. FPGA Implementation for CNN-Based Optical Remote Sensing Object Detection. *Electronics* **2021**, *10*, 282. [CrossRef]

17. Li, W.; Zhang, L.; Wu, C.; Cui, Z.; Niu, C. A new lightweight deep neural network for surface scratch detection. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1999–2015. [CrossRef] [PubMed]
18. Gharineiat, Z.; Tarsha Kurdi, F.; Campbell, G. Review of automatic processing of topography and surface feature identification LiDAR data using machine learning techniques. *Remote Sens.* **2022**, *14*, 4685. [CrossRef]
19. Camuffo, E.; Mari, D.; Milani, S. Recent Advancements in Learning Algorithms for Point Clouds: An Updated Overview. *Sensors* **2022**, *22*, 1357. [CrossRef] [PubMed]
20. Ming, Q.; Miao, L.; Zhou, Z.; Dong, Y. CFC-Net: A Critical Feature Capturing Network for Arbitrary-Oriented Object Detection in Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
21. 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.
22. Guo, Z.; Liu, C.; Zhang, X.; Jiao, J.; Ji, X.; Ye, Q. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 8792–8801.
23. Ming, Q.; Miao, L.; Zhou, Z.; Yang, X.; Dong, Y. Optimization for arbitrary-oriented object detection via representation invariance loss. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
24. 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]
25. 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 Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
26. Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; Metaxas, D. Oriented object detection in aerial images with box boundary-aware vectors. *arXiv* **2020**, arXiv:2008.07043.
27. Zhao, P.; Qu, Z.; Bu, Y.; Tan, W.; Guan, Q. Polardet: A fast, more precise detector for rotated target in aerial images. *Int. J. Remote Sens.* **2021**, *42*, 5821–5851. [CrossRef]
28. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 29 October–1 November 2019; pp. 8231–8240.
29. Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; Yang, W. Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sens.* **2019**, *11*, 2930. [CrossRef]
30. 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. *arXiv* **2020**, arXiv:2004.13316. [PubMed]
31. 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 (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3500–3509.
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
33. Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J. The KFIoU Loss for Rotated Object Detection. *arXiv* **2022**, arXiv:2201.12558.
34. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 677–694.
35. Qian, W.; Yang, X.; Peng, S.; Guo, Y.; Yan, C. Learning modulated loss for rotated object detection. *arXiv* **2021**, arXiv:1911.08299.
36. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
37. Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; Yan, J. Dense label encoding for boundary discontinuity free rotation detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 15819–15829.
38. Giordano, M.; Maddalena, L.; Manzo, M.; Guarracino, M.R. Adversarial attacks on graph-level embedding methods: A case study. *Ann. Math. Artif. Intell.* **2022**, 1–27. [CrossRef]
39. Weiler, M.; Cesa, G. General e(2)-equivariant steerable cnns. In Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 9 December 2019; pp. 14334–14345.
40. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
41. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets V2: More Deformable, Better Results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308.
42. Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-normalizing neural networks. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 972–981.
43. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
44. Majid Azimi, S.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. In Proceedings of the 14th Asian Conference on Computer Vision (ACCV), Perth, Australia, 2–6 December 2018; pp. 150–165.

45. Liu, L.; Pan, Z.; Lei, B. Learning a rotation invariant detector with rotatable bounding box. *arXiv* **2017**, arXiv:1711.09405.
46. 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]
47. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.; Bai, X. Rotation-Sensitive Regression for Oriented Scene Text Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918.
48. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic anchor learning for arbitrary-oriented object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Hilton New York Midtown, NY, USA, 7–12 February 2020; pp. 2355–2363.