IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

SARD: Towards Scale-Aware Rotated Object Detection in Aerial Imagery

YASHAN WANG[1,2,3], YUE ZHANG[1,3], YI ZHANG[1,3], LIANGJIN ZHAO[1,3], XIAN SUN[1,3], AND ZHI GUO[1,3]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; zhangyue@aircas.ac.cn (Y.Z.); sunxian@mail.ie.ac.cn (X.S.)
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

Corresponding author:Yue Zhang (e-mail: zhangyue@aircas.ac.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 41801349 and Grant 61725105.

ABSTRACT Multi-class object detection in remote sensing imagery is an important and challenging topic in computer vision. Compared with the object detection of natural scenes, remote sensing object detection has some challenges such as scale diversity, arbitrary directions and densely packed objects. To resolve these problems, this paper presents a scale-aware rotated object detection. Firstly, we propose a novel feature fusion module, which takes full advantage of high-level semantic information and low-level high resolution feature. The new feature maps are more suitable for detecting objects with a large difference in scale. Meanwhile, we design a specific weighted loss, which contains an intersection-over-union (IoU) loss and a smooth L1 loss to further address the scale diversity. Besides, in order to detect oriented and densely packed objects more accurately, we propose a normalization strategy for the representation of rotating bounding box. Our method is evaluated on two public aerial datasets DOTA and HRSC2016, and achieves competitive performances. On DOTA, we boost the mean Average Precision (mAP) to 72.95% on oriented object detection.

INDEX TERMS object detection, remote sensing, convolution neural network, rotation region.

I. INTRODUCTION

Multi-class object detection in aerial imagery is intended to locate objects of interest on the ground in aerial images and identifying their categories. And it has been playing an important role in computer vision over the years and can apply to many aspects such as cargo transportation and military uses. Although the object detection algorithm of the natural scene has been well developed, object detection in remote sensing images is full of challenges.

In recent decades, methods of object detection based on deep convolutional neural networks (CNNs) have been greatly developed. These methods are also applied to the field of remote sensing and have achieved good performances. But objects in remote sensing images have considerable scale differences, and objects of different scales are unevenly distributed. So detecting multi-scale objects in remote sensing imagery is a challenging task. In the natural scene object detection field, some studies have shown the multi-scale detector performs better for multi-scale objects than the single-scale detector. Multi-scale detectors use multiple feature layers for detecting objects of different scales. However, when we apply this method to the field of remote sensing, we find that this method is not suitable for predicting objects with unbalanced distribution of different scales. Take FPN (Feature Pyramid Network) [21] as an example, Region-of-Interests (RoIs) of different scales are assigned independently to different feature layers. Since objects in the remote sensing images are unbalanced in size and large objects only make up a small proportion, the feature layers of large aerial objects are not adequately trained with FPN.

To address the multi-scale problems, some researchers have also discovered the effect of the loss function on multi-scale objects [1], [35]. The traditional loss functions L1 and L2 use the distance of the coordinates to measure the effect of prediction. It can lead to a greater loss for large objects, and the network tends to ignore the small objects because of their smaller loss. Taking Fig. 1 as an example, when IoU is used as an indicator to evaluate the prediction effect, the prediction result of object 2 is much better than that of object 1. However, due to the larger scale of object 2, it will produce a larger loss than object 1 when measured by traditional L1 loss. Therefore, small objects are ignored in the network

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access

FIGURE 1. IoU versus L1norm for large and small objects. IoU is the indicator to evaluate the prediction effect, and the prediction result of object 2 is much better than that of object 1. However, due to the larger scale of object 2, it will produce a larger loss than object 1 when measured by traditional L1 loss.

optimization process. Unitbox [1] proposes that the IoU can be a good measurement of the object regression effect since it is independent of the object size. But the Unitbox network is a single-stage network structure that detected four corner points. The IoU loss proposed by it cannot be directly applied to the two-stage structure of this paper.

Besides the problem of various scales, object detection in remote sensing imagery also faces the problem of arbitrary directions. Aerial images are typically taken with bird views. Therefore objects in aerial images are always arbitrary oriented. Thus, using a horizontal box to locate an object would result in much mismatched space between the object and the prediction bounding box. At the same time, for some densely arranged and tilted objects, using the traditional horizontal RoI for detection is likely to filter out the correct object during Non-Maximum Suppression (NMS), resulting in missed objects [2]. In the field of text detection, rotation detection is proposed to detect oriented and dense text objects. They use rotated boxes with angle to detect objects instead of traditional horizontal RoI box [17], [27], [30]. Some researchers have applied the rotational detection to the area of remote sensing and have achieved some results [39], [40]. But most of them are designed for specific objects, such as vehicle detection [31], ship detection [39], [40], aircraft detection [22] and so on. When we apply the R^2CNN [17] structure to object detection in remote sensing imagery, angle predictions of some objects are inaccurate.

To address the aforementioned problems, we propose a scale-aware rotated object detector, called SARD. We firstly propose a novel feature fusion module, which takes full advantage of high-level semantic information and low-level high resolution feature. It is designed to detect objects in remote sensing images which have considerable scale differences, and objects of different scales are unevenly distributed. Then we propose a new weighted loss, which is developed on the traditional L1 loss and the IoU loss. This new weighted loss possesses the advantages of the above two losses and solves the problem that traditional loss is unfriendly to small objects. Finally, we propose a normalization strategy for the representation of rotating bounding box and improve the accuracy of angle regression.

In conclusion, the contributions of this paper are mainly three points:

- We design a novel scale-aware detector which aims at multi-category, and multi-scale oriented objects in large-scale remote sensing images. And we achieve competitive performance on both DOTA and HRSC2016 datasets.
- We propose a novel feature fusion module in the RCNN stage to make the feature fit for detecting multi-scale objects, and solve the problem of unbalance training. Meanwhile, we propose a weighted loss combining the traditional smooth L1 loss and the IoU loss, to balance the loss of large objects and small objects.
- We propose a normalization strategy for the representation of rotating bounding box to detect oriented and densely packed objects more accurately.

The rest of this article is organized as follows. The various parts of the scale-sensitive rotation detector are shown in Section 2. The results of our method and other methods in the DOTA data set are given in Section 3. By comparing with the other methods, we analyze the shortcomings and shortcomings of the network, while Section 5 gives our conclusions and future work.

II. RELATED WORKS

A. REGION BASED FRAMEWORK

Algorithms based on deep convolutional neural networks (CNNs) can be divided into two groups: region-based algorithms as well as regression-based algorithms. In 2015, the emergence of region-based detection algorithms [3] greatly improves the accuracy of detection. After that, Faster R-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access

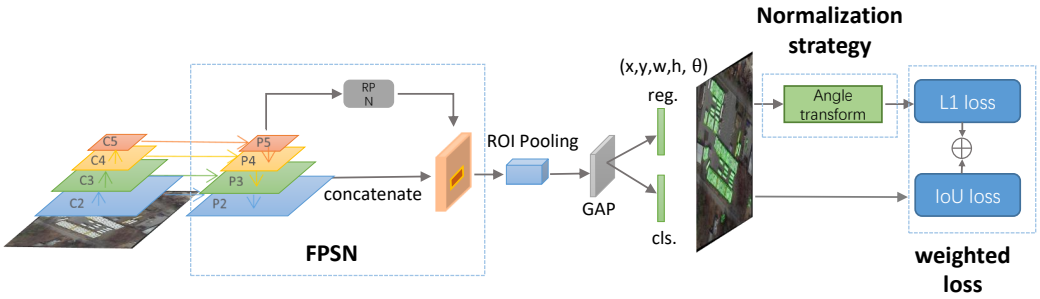


FIGURE 2. An overview of the SARD.

CNN detection framework [4] basically achieves end-to-end training, which generates proposal regions from the region proposal network (RPN) and afterward sends them to the R-CNN network for classification and regression. Meanwhile, methods based on regression directly return the position of the objects in the image, which has an advantage in speed, but the accuracy is not as high as the region-based detection algorithms. You Only Look Once [28]and Single Shot Multibox Detector [24] are classic regression-based object detection methods. When these generic object detection algorithms are introduced into the field of remote sensing, corresponding improvements have been achieved for the traits of remote sensing imagery [7], [12], [13], [19], [32], [33], [35], [36]. Xu et al. [34] introduce deformable convolution layers [10] to further improve detection accuracy. Han raises an efficient framework based on Faster R-CNN [15] which realizes the integration process of regional proposal generation stage and RCNN stage by sharing features of the region proposal generation stage and the RCNN stage. Region-based detection algorithms have become the mainstream for solving the difficulty of detecting multi-class objects in remote sensing imagery since these detectors achieve superior performance in accuracy.

B. MULTI-SCALE DETECTOR

The aforementioned methods are all single-scale detectors which typically use features from one layer to represent objects. At the same time, some methods have been raised to work out multi-scale object detection by using feature from several layers of CNN, such as HyperNet [18] and ION [6]. These methods usually combine some layers via skip connections to integrate deep, intermediate features and shallow features before making a prediction. Among them, FPN [21] has become the mainstream of multi-scale object detection for its character of easy to transplant. It generates feature maps of different scales through a top-down architecture and lateral connections.

C. ROTATIONAL REGION CNN

Rotational Region CNN(R^2CNN) [17] outputs five parameters with an angle in the RCNN stage to get the rotating bounding box. RRPN [27] takes another approach, which generates a rotating anchor during the RPN phase and picks them up. However, when extracting the features of the RoI, some information is lost since the ROI is a smaller area with an angle, therefore the detection effect is not as good as the R^2CNN. In the field of remote sensing, there are also researches of multi-class and rotation object detection based on CNN network. Cheng et al. [42] chooses to train a rotation invariant layer to rotate the result. Li et al. [43] applies the RRPN method to the remote sensing image to obtain an angled detection frame. Combined with the analysis of R^2CNN and RRPN above, we chose to apply the R^2CNN to the remote sensing images.

III. METHODOLOGY

In large-scale remote sensing images, the dense and multi-scale objects can result in significant difficulties to object detection. In this regard, we propose a scale-aware rotation detector. Our improvement mainly includes the feature fusion structure FPSN, the weighted loss which is divided into two parts and a normalization strategy for the representation of rotating bounding box. The architecture is shown in Fig. 2.

A. FPSN: FEATURE PYRAMID STACK NETWORK

The FPN structure uses different feature maps to extract features for objects of all sizes. In the face of objects with significant scale differences in remote sensing images, this method is unbalanced to train different feature layers, and cannot obtain the most suitable features. To address the question, we propose the FPSN structure, canceled the assignment of feature maps of FPN. We can not only obtain more suitable features for training, but also avoid the training imbalance caused by separating training of different feature maps.

In the FPN structure, the anchors were originally produced with a definite size on each feature map in the first regression stage. Then proposals extract features on the corresponding

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access

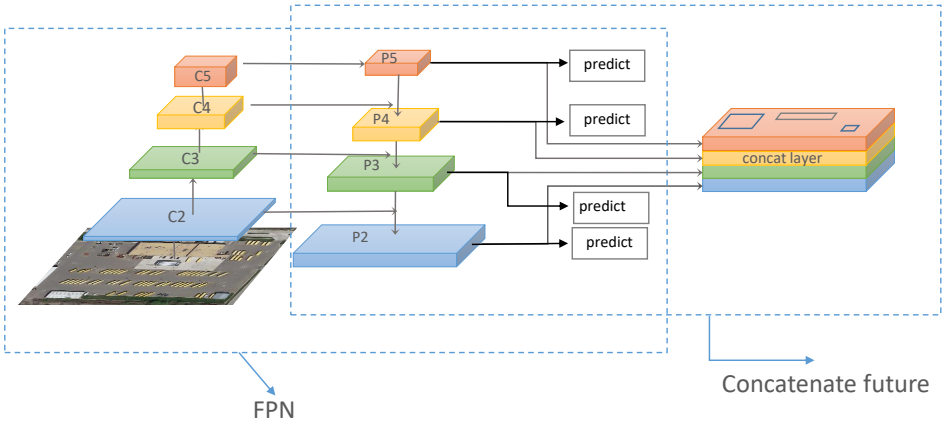


FIGURE 3. Schematic diagram of our FPSN structure. At the RPN stage, five pyramid features P2-P6 were fused from C2-C5. Then at the RCNN stage, the P2-P5 feature maps are resized to the P2 size, then concatenated together for use as a new feature map.

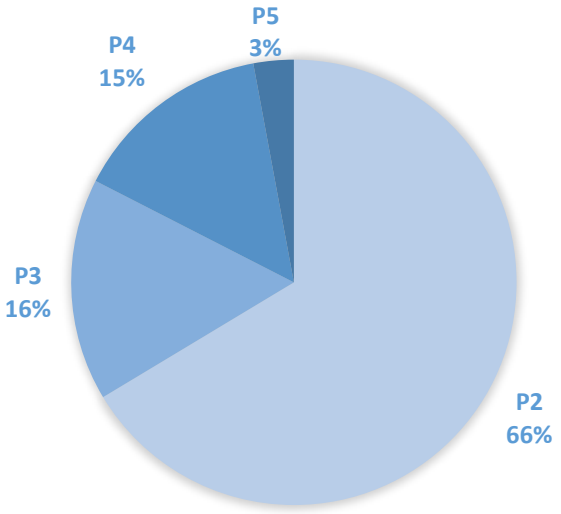


FIGURE 4. The proportion of proposals each original P feature used.

feature map according to its size. The chosen rule follows Eqn 1.

$$k = \lfloor k_0 + \log_2(\sqrt{wh}/224) \rfloor, k_0 = 4 \tag{1}$$

k presents which feature map to choose, w, h present the width and height of the proposal. The formula can be understood as follows: the k_0 means the P4 feature and proposals with $\sqrt{wh}$ less than 224 but bigger than 112 will extract features on it. It can be inferred from the formula that proposals with $\sqrt{wh}$ greater than 224 and less than 448 will all extract features on P5, and proposals of other sizes will follow the same analogy.

We use the formula to test the use of the feature maps on remote sensing dataset DOTA. The specific distribution is shown in Fig. 4. It can be observed that the distribution of the proposal on different feature maps is unbalanced. Thus, we believe that the assignment of feature maps has some defects on the detection of multi-scale objects for the following two reasons:

- Fig. 5 shows the difference between P2-P5 features and features after concatenated. The first set of graphs mainly analyzes the impact on large objects, when using the method of assigning levels, the objects in the graph will use the features of P4 or P5. It can be observed that the P5 feature loses many details, which is unfavorable for judging the category of objects. The second group analyzes the small objects, the P2 feature layer that the small objects originally used contains much low-level information, and little semantic information is taken into consideration.
- As depicted in Fig. 4, the scales of objects in remote sensing images vary greatly, and the distribution is unbalanced. The number of small objects ranks first (reaching 66%). Therefore, the features of small objects would take more training than big objects', resulting in the corresponding reduction in the effect of the big objects.

To solve these problems, we aim to cancel the feature assignment and design a new feature map which all proposals crop feature on it. In the second stage, we resize the previously used P2-P5 layer into a P2 feature size using bilinear interpolation, and then we concatenate them together, stack into a new feature layer(see Fig. 3). It is appropriate for proposals of all sizes, and the new feature map takes full advantage of different feature maps at all scales.

It can be inferred from Fig. 5 that features of C2 should be taken seriously in the new fusion features, because most proposals used to extract features from this feature map. Thus, we resize the previously used P2-P5 layer into a P2 feature size, therefore all the information about the P2 characteristics is preserved. Then we concatenate them together to obtain a new feature map. At this time, all proposals regardless of the size of them are cropped on this new feature map for the second stage. The information about the new

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access

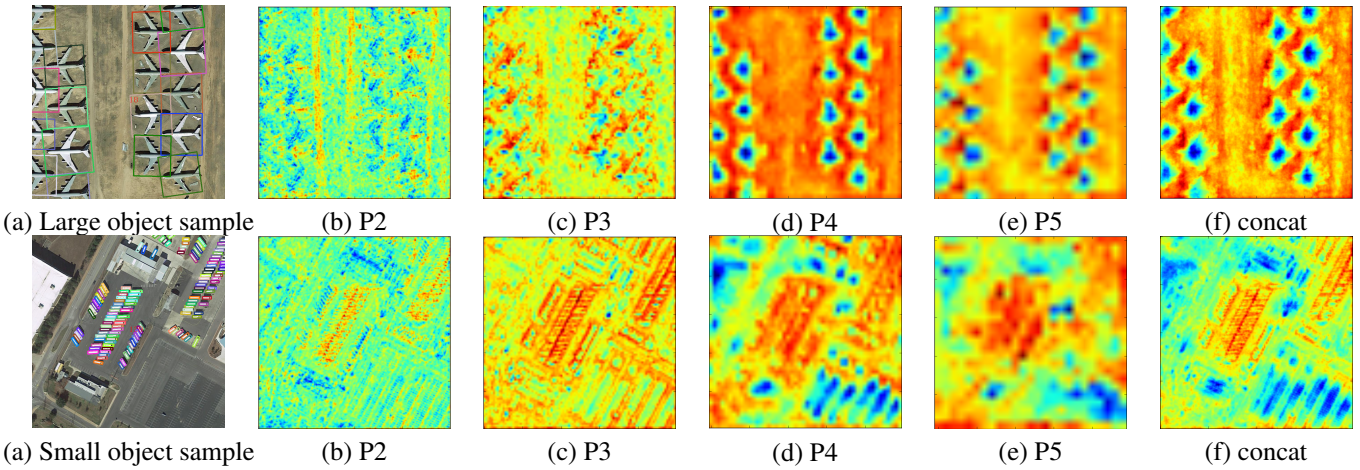


FIGURE 5. There are two sets comparison of original maps and concatenated feature map. (a) This is the original image, and (b)-(f) are feature maps P2-P5 and new feature map after being concatenated. As we can see from the figures, the underlying feature map shows the verse information of the picture, the high-level features are mainly semantic information, but the feature map after concatenated has both advantages.

Algorithm 1 Combined loss

Input: $\tilde{x}$ as the ground truth
Input: x as the prediction result
Parameter: u as the threshold of object size
Output: L as regression loss

for each det **do**
  if $\tilde{w} * \tilde{h} \leq u$ **then**
    $\alpha = 0.25$.
    $\beta = 0.75$.
  else
    $\alpha = 0.75$.
    $\beta = 0.25$.
  end if
  $L_{reg} = \alpha * L_{L1loss} + \beta * L_{IoUloss}$.
end for
return L

feature map can be observed in Fig. 5. For large objects, the new feature map (f) is significantly more detailed than the original P5(e), helping to classify objects. Meanwhile for small objects, the new feature map (f) contains more semantic information than the original P2(b). Moreover, proposals of different scales extract features from one feature map, which solves the problem of training unevenly. At the same time, this feature map after concatenated is suitable for all sizes of proposals, and will not generate unsuitable problems caused by discontinuity of feature maps.

B. WEIGHTED LOSS FUNCTION

Another problem caused by the large difference in the scale of objects is that small objects tend to be ignored. The traditional loss function tends to favor large objects and to ignore small objects because it uses the distance of the ground truth and the prediction object to measure the loss. We introduce a size-insensitive loss function to raise a new loss function that can treat objects with large scale differences more equitably. Mainstream object detection algorithms generally adopt smooth L1 loss to measure the effect of the regression task (see formula 2).

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{2}$$

Using the traditional smooth L1 loss produces a problem: when IoU is the same, the loss of large objects will be much larger than small objects. As a result, the network is easily dominated by large objects, and the gradient direction is more inclined to the regression of large objects. Compared to the smooth L1 loss function, IoU is also a good indicator to measure the effect of regression: IoU is between 0 and 1, the larger the IoU, the higher the accuracy of the prediction result. IoU is defined as shown in the formula below.

$$I = area(GT \cap det) \tag{3}$$

$$U = area(GT \cup det) \tag{4}$$

$$IoU = \frac{I}{U} \tag{5}$$

GT represents the ground truth box and det is the result of the network. Contrary to the smooth L1 loss, the IoU loss of a small object is larger than that of a large object when the coordinate offsets are the same.

$$L_{IoUloss} = -\ln IoU \tag{6}$$

The gradient descent formula of IoU loss:

$$\frac{\partial L}{\partial x} = -\frac{I * (\bigtriangledown x - \bigtriangledown I) - U * \bigtriangledown I}{U^2 * IoU} = \frac{1}{U} \bigtriangledown x - \frac{U + I}{U * I} \bigtriangledown I \tag{7}$$

However, the smooth L1 loss is more concise than IoU loss when guiding network learning. Smooth L1 loss is to guide the prediction of the four coordinate points to the true value point, IoU loss is designed to make the ground truth and the prediction box overlap more, what is more complicated for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access

the network. Smooth L1 loss learns better under the same conditions. Based on the characteristics of the above two loss functions on the size of the object, we propose a weighted loss: a loss consisting of smooth L1 loss and IoU loss. It is dominated by the object size.

When the object size is less than one–tenth of the image, the L1 loss occupies 25%, and the IoU loss occupies 75%; otherwise, the weight of L1 loss is 0.75, and the weight of IoU loss is 0.25.

$$L_{reg} = \alpha * L_{smooth_{L1}loss} + \beta * L_{IoUloss} \tag{8}$$

In addition, the loss of the classification task uses the softmax cross entropy. The L_{cls} is the logarithmic loss of two categories, and N_{cls} is the total number of anchors. L_{cls} and N_{cls} are defined same as suggested in Fast R-CNN. The overall loss formula is as follows:

$$L = \frac{1}{N_{cls}} \sum L_{cls} + \frac{1}{N_{reg}} \sum L_{reg} \tag{9}$$

C. ROTATION BRANCH

When using the R^2CNN structure to obtain a rotating bounding box, we find that some objects have inaccurate angle predictions. We think this phenomenon is caused by the boundary problem for the rotation angle, as shown in Fig. 6. It shows that an ideal form of regression (the yellow box rotates counterclockwise to the blue box), but the loss of this situation is very large due to the periodicity of the angle. We have normalized the representation of the rotating bounding box to avoid the periodicity of the angle, thus reducing the inaccuracy of the angle prediction.

We use the R^2CNN structure to get rotating objects, which is a two-stage detection framework. In the RPN phase, our structure is similar to a conventional detector, except that the anchor aspect ratio and size for large-scale complex objects are carefully selected. In the Fast-RCNN phase, the traditional horizontal bounding box is generated by four coordinate points $(x_{min}, y_{min}, x_{max}, y_{max})$.In our case, we use five parameters including the angle (x, y, w, h, θ) to get rotating bounding box (see Fig. 6). The Angle (θ) is the counterclockwise direction of the horizontal (x) axis and is used to handle the rectangle on the edge of the first corner. This side is defined as width w; and the other one is height h. The scope of angle is $[-\pi/2, 0)$. The proposal generated by the RPN phase is horizontal, and the five regression parameters obtained through the prediction phase are converted into a rotating bounding box (see Fig. 6). The proposal is transformed according to the following formula to get the final bounding box, where x stands for the prediction box and x_p stands for proposal.

$$t_x = (x - x_p)/w_p, t_y = (y - y_p)/h_p \tag{10}$$

$$t_w = log(w/w_p), t_h = log(h/h_p) \tag{11}$$

$$t_\theta = \theta - \theta_p \tag{12}$$

When we define and optimize the detection madness of rotation using the above definition, we find that a small number of abnormal results appear in the prediction results. The predicted values of these angles are about 45 degrees from the true value, and the length and width are different. Our analysis suggests that this phenomenon may be due to the angular definition of the rotating frame. In the above settings, we set the angle to the range of $[-\pi/2, 0)$ for the sake of unification, but we cannot guarantee that the θ parameters given by the network are within the range in the prediction. A situation will appear as shown in Fig. 6: The prediction result is a yellow box and the blue box is a ground truth box. The figure implies that the predicted result is good. However, because θ is bigger than $-\pi/2$ in the given result, there will be a big difference between t, height, width and the ground truth, resulting in a big loss.

To preserve this better result as feedback, we limit the predicted angle to the $[-\pi/2, 0)$ range as a normalized representation. For instance, we rotate the case where θ is greater than $-\pi/2$ in the prediction result by 90 degrees (see Fig. 6). At the same time, there are other types with predicted angles less than -90. For example, if the angle prediction is within $[-\pi, -\pi/2)$, we also reverse the angle, but the length and width are unchanged. After such conversions, the abnormal loss value caused by such an angle will become smaller. Thereby it reduces the inaccuracy of predictions that angle rotates by 45 degrees and the length and width are similar, the effect is shown in Fig. 7.

In addition, the IoU calculation method of the rotating bounding box has to be changed, and we use the trigonometric calculation method proposed by [31]. Considering the diversity of object shapes in the dataset, we set different R-NMS thresholds for different categories [37].

D. NETWORK OVERVIEW

Fig. 2 presents an overview of SARD for multi-class object detection. Given an input image, we use ResNet [41] as backbone network to extract different levels of feature maps. ResNet is a popular backbone network in recent years. It has a great network depth and can obtain multiple feature maps, but reduces the amount of parameters by avoiding the use of fully connected layers. Input a raw picture through the multiple convolution of ResNet, we get the feature maps of different levels. Then feature maps are fused by the structure of FPN to obtain the P2-P5 feature maps. The channel numbers of the feature maps are (128, 256, 512, 1024), and the feature size is reduced by 0.5, which is the same as FPN. Then in the second stage, we resize all P feature maps to the P2 size and then concatenate them together to get a new feature map. Then proposals extract features on this new feature map for classification and regression in the second phase.

IV. EXPERIMENTS AND ANALYSIS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access

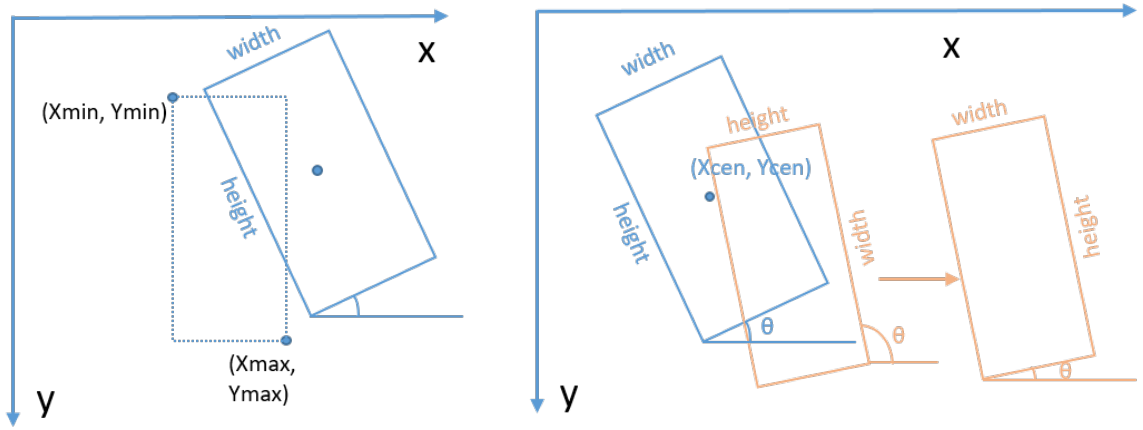
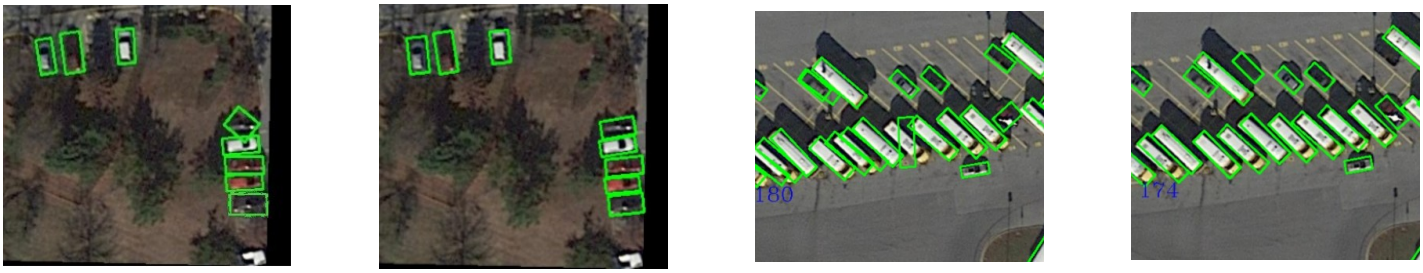


FIGURE 6. (a) The representation of horizontal bounding box and rotating bounding box. Horizontal bounding boxes are represented using the top left and bottom right dots. Rotating bounding boxes are represented using coordinates of the center point (x, y), width w, height h, and θ. Rotation Angle (θ) is the horizontal axis (the X-axis) counterclockwise to deal with the rectangle on the edge of the first Angle. (b) Normalization strategy for the representation of rotating bounding box. We limit the predicted angle to the $[-\pi/2, 0)$ range.



(a) (b) (c) (d)

FIGURE 7. There are two groups of comparisons between the optimized rotational regression strategies used before and after. The results of multiple predictions that were inaccurate were corrected.

TABLE 1. R-NMS threshold for each category

class	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC
threshold	0.3	0.3	0.0001	0.3	0.2	0.1	0.05	0.3	0.3	0.2	0.3	0.1	0.0001	0.1	0.2

TABLE 2. Ablation Studies of each components in our proposed method with ResNet50d on DOTA

method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC	mAP
baseline	88.96	78.36	45.96	64.19	69.45	60.21	72.69	90.61	81.56	84.14	50.16	63.47	57.46	63.74	56.03	68.47
+Normalization strategy	89.06	80.00	46.36	63.26	67.06	76.80	72.76	90.37	81.90	84.05	54.93	63.57	61.43	66.93	54.47	68.92
+weighted loss	88.98	79.15	46.99	64.61	69.44	58.43	73.72	90.68	78.82	84.21	55.39	61.28	57.12	66.59	60.81	69.08
+FPSN	88.78	80.22	46.04	65.22	69.00	59.06	73.73	90.83	78.95	84.47	55.09	62.70	56.97	67.49	62.21	69.38
+both	89.00	77.93	50.79	67.25	72.31	61.98	74.61	90.89	86.58	85.25	61.45	61.93	63.92	68.79	60.81	71.57

TABLE 3. Comparative experiment of our method with ResNet101d for oriented object detection on DOTA

method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC	mAP
SSD [24]	39.83	9.09	0.64	13.18	0.26	0.39	1.11	16.24	27.57	9.23	27.16	9.09	3.03	1.05	1.01	10.59
YOLOv2 [28]	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.35	34.65	16.02	37.62	47.23	25.5	7.45	21.39
R-FCN [9]	37.80	38.21	3.64	37.26	6.74	2.60	5.59	22.85	46.93	66.04	33.37	47.15	10.60	25.19	17.96	26.79
FR-H [4]	47.16	61.00	9.80	51.74	14.87	12.80	6.88	56.26	59.97	57.32	47.83	48.70	8.23	37.25	23.05	32.29
FR-O [2]	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
R-DFPN [39]	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^2CNN [17]	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 [27]	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 [5]	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
RoI Trans [11]	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
R^2CNN++ [37]	89.66	81.18	45.50	71.02	67.42	59.18	66.83	90.90	79.00	84.43	61.17	63.48	65.34	68.01	62.05	71.09
SARD	89.93	84.11	54.19	72.04	68.41	61.18	66.00	90.82	87.79	86.59	65.65	64.04	66.68	68.84	68.03	72.95

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access

TABLE 4. Comparison of the accuracy and speed of different methods on the HRSC2016

Method	Backbone	Image Size	mAP	Speed
R^2CNN [17]	ResNet101	800*800	73.07	2fps
RC1 [25] & RC2 [25]	VGG16	-	75.7	<1fps
RRPN [27]	ResNet101	800*800	79.08	3.5fps
R2PN [40]	VGG16	-	79.6	<1fps
RRD [20]	VGG16	384*384	84.3	<1fps
SARD	ResNet101	800*800	85.4	1.5fps

A. DATASETS

To assess SARD, we conducted experiments on two well-known datasets (DOTA, HRSC2016) for oriented object detection in aerial images.

DOTA [2].The dataset consists of 2,806 aerial images, each of which was approximately 4000*4000, covering a wide range of scales, locations, and shapes. It is much larger than the other datasets for object detection in aerial images with oriented bounding box annotations. These images are divided into 15 categories, including swimming pool (SP), baseball diamonds (BD), small vehicles (SV), large vehicles (LV), storage tanks (ST), tennis courts (TC), basketball courts (BC), football field (SBF), roundabout (RA), ground track (GTF), and helicopter (HC). The annotated complete data set includes 188,282 instances, each labeled with an arbitrary quad. The DOTA dataset provides an evaluation server. We cut the original image into a small image of 800*800 in steps of 600 and cut the original 2806 images into 18,000 images.

HRSC2016 [26].It collected 1061 aerial images from Google Earth, and is a very classical dataset for detecting oriented ships. The sizes of images range from 300*300 to 1500*900. We used the training set plus the validation set (436+181=617) for training and tested on a test set containing 444 images.

B. EXPERIMENTAL FRAMEWORK

The baseline of the experiment is the R^2CNN structure with FPN. We use the pre-trained ResNet-50d for ablation study and ResNet-101d models provided by gluonCV for comparison experiment. The difference between ResNet50 and ResNet101 lies in the number of convolution layers. In this paper, ResNet50 is selected to perform the ablation experiment in order to save the experiment time. In the comparison experiment, ResNet101 is selected to obtain better experimental data for comparison to other methods. All experiments are implemented on TensorFlow, and the network is trained on a Tesla P100 GPU. And we employ Momentum Optimizer as optimizer. The images in the DOTA dataset are large. So in the data preprocessing stage, we cut the image into small images of 800*800 in size, overlapping 200 pixels. In the stage of making anchors, an anchor of one size is drawn on each feature map P, and each type of anchor is designed with nine anchor ratios:[1, 1 / 2, 2., 1 / 3., 3., 5., 1 / 5., 7., 1 / 7.]. In the RPN [4] phase, the batch size is set to 512. During training, we take the highest score of 12,000 bounding boxes for NMS, and then 2000 of them were left for subsequent training. The number of RoIs is defined same as suggesting in Faster-RCNN[15]. When testing, these two parameters are set to 10000, 1000. The initial learning rate is 0.0003. We conducted a total of 300k rounds' training on DOTA data set, and the learning rate is 10 times attenuation at 100k rounds and 200k rounds.

C. ABLATION STUDIES

The following ablation experiments were carried out using the above baseline. All parameters that do not involve improvement are all consistent. The experimental results of the DOTA dataset were evaluated using the DOTA official evaluation server.

FPSN. According to Table 2, concatenating the feature maps increases the mAP by 0.9% compared to the baseline. In the performance of the single category, the AP of baseball diamond, soccer ball field and swimming pool classes have significantly improved by two percentage points, five percentage points and four percentage points respectively. It proves that the new fusion feature improves the mAP significantly compared to the baseline method.

Weighted loss. In the comparison experiment, the new loss function has improved the detection effect: mAP has increased by 0.6 points. The performance of small object detection such as swimming pool and helicopter are improved by three percentage points and four percentage points. The use of a new detector effectively prevents the optimization weight imbalance due to object size disproportion.

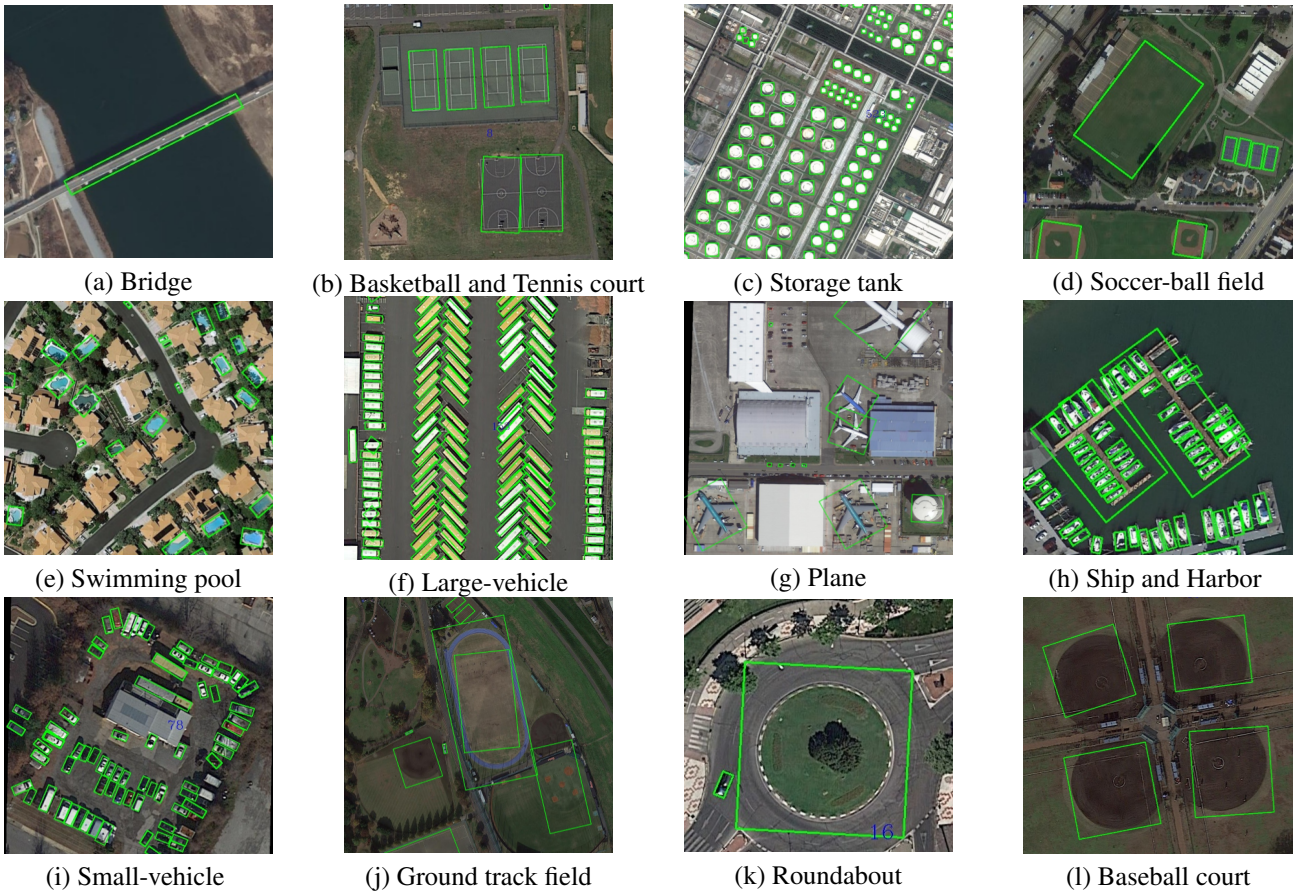
Normalization strategy. The normalized representation for the rotation frame has not much improvement in the data. Fig. 7 shows that the regression angle is more accurate. But this regression anomaly does not affect the classification, and the IoU value is big enough to get a good mAP.

D. COMPARISONS WITH THE STATE-OF-THE-ART

We compared our proposed detectors to the latest methods on the public datasets DOTA and HRSC2016. The results are shown in Table 3 and Table 4.

Result on DOTA. Table 3 compares SARD and other methods for rotation object detection on the DOTA dataset. Here, we used the pre-training model ResNet101d provided by gluonCV. Compared with R^2CNN, SARD uses the FPSN and the weighted loss, has obvious improvement in small and intensive objects such as vehicles and ships, which verifies that our proposed weighted loss works. At the same time, we also compare SARD with the latest algorithms R^2CNN++ and RoI Transformer. R^2CNN++ adds the attention structure to predict the position of objects better. RoI Transformer designed a ROI learner to obtain an accurate rotation box. These two methods are the algorithms that have recently achieved the best results on the DOTA dataset rotation task.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2956569, IEEE Access



(a) Bridge (b) Basketball and Tennis court (c) Storage tank (d) Soccer-ball field

(e) Swimming pool (f) Large-vehicle (g) Plane (h) Ship and Harbor

(i) Small-vehicle (j) Ground track field (k) Roundabout (l) Baseball court

FIGURE 8. Prediction results for each category of our algorithm on the DOTA dataset.

Compare to them, our method also has a relatively better performance (6 percentage points higher in accuracy) in the detection of large objects (e.g. bridges and baseball courts). It demonstrates that our designed feature fusion structure has sufficient training for large objects. On the whole, we achieved state-of-the-art performance on the DOTA dataset and overall detection performance improved by 4.5% compared to the baseline network.

Result on HRSC2016. Table 4 gives an experiment comparison of our algorithm and other methods on the HRSC2016 data set. Compared with other methods [20], [25], we have achieved competitive performances. RC1 [25] uses a framework similar to RRPN. While the SARD is based on R^2CNN, a new feature fusion structure and loss function are used. So the SARD get a great improvement in accuracy. RRD [20] uses two branches to get different characteristics to deal with the oriented object detection. Compared with it, the accuracy of SARD improves with an acceptable speed.

V. DISCUSSION

The experimental results show the SARD can achieve desirable results. The FPSN structure proposed is more suitable for objects with large scale differences in remote sensing images, and achieves better overall detection performance. Compared with the work of Jian et al. [11], we have greatly improved on large objects such as bridges, baseball diamonds and soccer ball fields. It indicates that the FPSN works, avoiding less training of large objects. At the same time, it also has improvement in other categories, but not as much as above categories.

In addition, we have designed a new size-weighted loss function. As can be seen from Table 2, the use of this new loss function improves the overall performance of the detector. However, we found that the angle definition of the rotation detector is not rigorous, and the representation of the rotating bounding box regression is normalized to reduce the error loss. To some extent, this problem can also be solved by defining a more reasonable loss function. In the future, it will be possible to design a loss function that satisfies both of these demands, making the overall design of the detector more concise.

VI. CONCLUSION

In the paper, we evaluate various feature fusion methods and find that existing methods are not perfectly applicable to multi-scale objects in remote sensing images. Therefore, a scale-aware object detection algorithm is proposed, which can balance multi-scale objects. We propose a novel fusion feature which considers the multi-scales objects in remote

sensing imagery. Additionally, we propose a weighted loss based on object size to improve overall detection performance. Finally, we use data enhancement and multi-scale training to further improve performance. The prediction results on the DOTA dataset and the HRSC2016 dataset demonstrate that our method has more advanced detection performance than the baseline network and other published algorithms [11], [37]. In the future, we plan to explore the method of feature fusion, strive to get the right features at a lower cost, and try to further improve the accuracy of angle prediction with cascade-RCNN.

REFERENCES

[1] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In Proceedings of the 2016 ACM on Multimedia Conference, pages 516–520. ACM, 2016.

[2] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai. Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In Proc. CVPR, 2018.

[3] Ross Girshick. Fast r-cnn. In The IEEE International Conference on Computer Vision (ICCV), December 2015.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 91–99. Curran Associates, Inc., 2015.

[5] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. arXiv preprint arXiv:1807.02700, 2018.

[6] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2874–2883, 2016.

[7] Huiyuan Chen, Zeyu Liu, Weiwei Guo, Zenghui Zhang, and Wenxian Yu. Fast detection of ship targets for large-scale remote sensing image based on a cascade convolutional neural network". Journal of Radars, 8(R19041):413, 2019.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2017.

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems, pages 379–387, 2016.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 764–773, 2017.

[11] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for detecting oriented objects in aerial images. arXiv preprint arXiv:1812.00155, 2018.

[12] Kun Fu, Tengfei Zhang, Yue Zhang, Menglong Yan, Zhonghan Chang, Zhengyuan Zhang, and Xian Sun. Meta-ssd: Towards fast adaptation for few-shot object detection with meta-learning. IEEE Access, 7:77597–77606, 2019.

[13] Xun Gao, Xian Sun, Menglong Yan, Hao Sun, Kun Fu, Yue Zhang, and Zhipeng Ge. Road extraction from remote sensing images by multiple feature pyramid network. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, pages 6907–6910. IEEE, 2018.

[14] Wei Guo, Wen Yang, Haijian Zhang, and Guang Hua. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. Remote Sensing, 10(1):131, 2018.

[15] Xiaobing Han, Yanfei Zhong, and Liangpei Zhang. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sensing, 9(7):666, 2017.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2980–2988. IEEE, 2017.

[17] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. arXiv e-prints, page arXiv:1706.09579, Jun 2017.

[18] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 845–853, 2016.

[19] Hongyan Li, Chungeng Li, Jubai An, and Junli Ren. Attention mechanism improves cnn remote sensing image object detection. Journal of Image and Graphics, 24(8):1400–1408, 2019.

[20] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5909–5918, 2018.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, volume 1, page 4, 2017.

[22] Lei Liu, Zongxu Pan, and Bin Lei. Learning a rotation invariant detector with rotatable bounding box. arXiv preprint arXiv:1711.09405, 2017.

[23] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. arXiv preprint arXiv:1809.02165, 2018.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.

[25] Zikun Liu, Jingao Hu, Lubin Weng, and Yiping Yang. Rotated region based cnn for ship detection. In 2017 IEEE International Conference on Image Processing (ICIP), pages 900–904. IEEE, 2017.

[26] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geoscience and Remote Sensing Letters, 13(8):1074–1078, 2016.

[27] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. IEEE Transactions on Multimedia, 2018.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[29] Yun Ren, Changren Zhu, and Shunping Xiao. Deformable faster r-cnn with aggregating multi-layer features for partially occluded object detection in optical remote sensing images. Remote Sensing, 10(9):1470, 2018.

[30] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2550–2558, 2017.

[31] Tianyu Tang, Shilin Zhou, Zhipeng Deng, Lin Lei, and Huanxin Zou. Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks. Remote Sensing, 9(11):1170, 2017.

[32] Siyu Wang, Xin Gao, Hao Sun, Xinwei Zhang, and Xian Sun. An aircraft detection method based on convolutional neural networks in high-resolution sar images. Journal of Radars, 6(2095-283X(2017)02-0195-09):195, 2017.

[33] Gang Xu, Jiguang Yue, Yanchao Dong, Qijia Lou, Wencheng Xiong, and Yihuang Nie. Cement plant detection on satellite images using deep convolution network. Journal of Image and Graphics, 24(4):550–561, 2019.

[34] Zhaozhuo Xu, Xin Xu, Lei Wang, Rui Yang, and Fangling Pu. Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery. Remote Sensing, 9(12):1312, 2017.

[35] Jiangqiao Yan, Hongqi Wang, Menglong Yan, Wenhui Diao, Xian Sun, and Hao Li. Iou-adaptive deformable r-cnn: Make full use of iou for multi-class object detection in remote sensing imagery. Remote Sensing, 11(3):286, 2019.

[36] Zhiyuan Yan, Menglong Yan, Hao Sun, Kun Fu, Jun Hong, Jun Sun, Yi Zhang, and Xian Sun. Cloud and cloud shadow detection using multilevel feature fused segmentation network. IEEE Geoscience and Remote Sensing Letters, 15(10):1600–1604, 2018.

[37] Xue Yang, Kun Fu, Hao Sun, Jirui Yang, Zhi Guo, Menglong Yan, Tengfei Zhan, and Sun Xian. R2cnn++: Multi-dimensional attention based rotation invariant detector with robust anchor strategy. arXiv preprint arXiv:1811.07126, 2018.

[38] Xue Yang, Qingqing Liu, Junchi Yan, and Ang Li. R3det: Refined single-stage detector with feature refinement for rotating object. arXiv preprint arXiv:1908.05612, 2019.
[39] Xue Yang, Hao Sun, Kun Fu, Jirui Yang, Xian Sun, Menglong Yan, and Zhi Guo. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. Remote Sensing, 10(1):132, 2018.
[40] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. IEEE Geoscience and Remote Sensing Letters, (99):1–5, 2018.
[41] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778, 2016.
[42] Cheng Gong, Zhou Peicheng, Han Junwei. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 7405–7415, 2016.
[43] Li Ke, Cheng Gong, Bu Shuhui, You Xiong. Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 56(4): 2337-2348, 2017.

YASHAN WANG received the B.Sc. degree from the Beijing Institute of Technology, Beijing, China, in 2017. She is currently pursuing the M.Sc. degree with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision, and remote sensing image processing, especially on object detection .

YUE ZHANG (M'18) received the B.E. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 2012, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2017, where he is currently an Assistant Professor with the Institute of Electronics.

His research interest includes the analysis of optical and synthetic aperture radar remote sensing images.

YI ZHANG received the B.Sc. Degree from Xidian University, China, in 2000. He received the Ph.D. degree with the University of Science and Technology of China, China, in 2015. He is an assistant professor of Institute of Electrics, Chinese Academy of Sciences.

His research interests include remote sensing images and communication technology.

LIANGJIN ZHAO received the B.E degree in Automation from University of Electronic Science and Technology of China, Chengdu, China, in 2015, and the Master degree from Beijing Institute of Technology, Beijing, China, in 2018, where he is currently Research Assistant with Institute of Electronics, Chinese Academy of Sciences.

His research interest includes the target detection and recognition in unmanned aerial vehicle remote sensing images and simultaneous localization and mapping.

XIAN SUN received the B.Sc. degree from Beihang University, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2006 and 2009, respectively.

He is currently a Professor with the Institute of Electronics, Chinese Academy of Sciences. His research interests include computer vision and remote-sensing image understanding.

ZHI GUO received the B.Sc. degree from Tsinghua University, Beijing, China, in 1998, and the M.Sc. and Ph.D. degrees from the Chinese Academy of Sciences University, Beijing, in 2003.,He is currently a Professor with the Institute of Electronics, Chinese Academy of Sciences, Beijing.

His research interests include remote sensing information processing and application.