Learning Point-guided Localization for Detection in Remote Sensing Images

Qing Song, Fan Yang, Lu Yang, Chun Liu, Mengjie Hu, Lurui Xia,

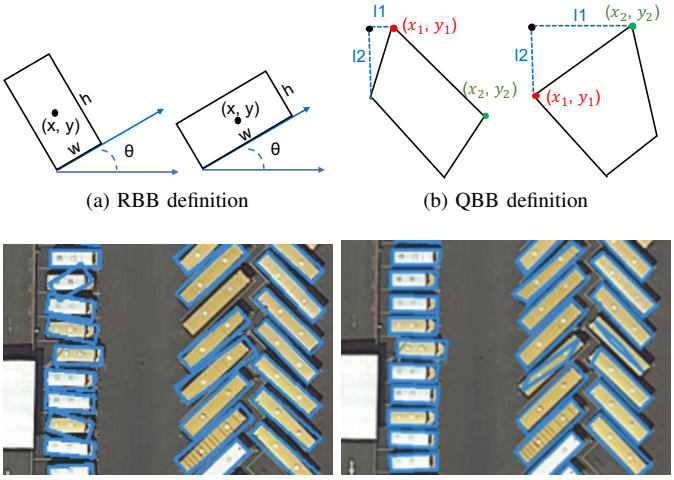
***Abstract*—Object detection in remote sensing images is challenging due to the dense distribution and arbitrary angle of the objects. It is a consensus that the oriented bounding box (OBB) is more suitable to represent the aerial objects. However, there are some extreme cases in regression-based OBB detection that make the regression target discontinuous, resulting in poor performance. In this paper, an analysis of the formats of OBB and the problems in its regression is presented, following with an exploration of transform localization from regression to keypoint estimation, which could be applied to avoid the problem of discontinuous regression target. Our novel method is called Object-wise Point-guided Localization Detector (OPLD). Continuously, a new prediction of center-point is introduced to refine the results, as the truncation problem caused by the cut graph. Lastly, in order to figure the problem of inconsistency between the localization quality and the classification score, both the endpoint scores and the classification score are adopted weighting as a result score. Experimental results are based on two widely used datasets, i.e., DOTA and HRSC2016. OPLD achieve 76.43% mAP and 78.35% mAP in oriented bounding boxes (OBB) and horizontal bounding boxes (HBB) tasks of DOTA-v1.0, which achieves state-of-the-art performance, respectively. Project page at https://github.com/yf19970118/OPLD-Pytorch.**

***Index Terms*—convolutional neural network, deep learning, oriented object detection, remote sensing**

I. Introduction

OBJECT detection is an essential task in computer vision, which can be decoupled into object classification and location. In recent years, many detectors based on deep convolution neural networks have made great progress in the field of natural images. According to the different localization methods, it can be roughly divided into regression-based methods and keypoint-based methods. The regression-based method [1]–[4] obtains the starting point of regression through manual setting or model detection, which can be the bounding box or the center-point of the bounding box, one or more refinement through the predicted offset will be implemented. The keypoint-based method [5]–[7] detects all points on the entire image used to represent the bounding box and then groups them to obtain the final result.

Unlike natural images that are usually taken from a horizontal angle, remote sensing images are bird's-eye views, and the objects in the images are arbitrary oriented and may be clustered. The horizontal bounding box (HBB) used by the general detection cannot accurately calibrate the position of the object, and when the object appears densely, the adjacent HBB will be filtered by NMS [8]. Many works [9]–[12] have explored the use of oriented bounding box (OBB) and have made significant progress. Among these methods, regression-based detection is the mainstream.



(a) RBB definition (b) QBB definition

(c) extreme cases in RBB regression (d) extreme cases in QBB regression

Fig. 1. Commonly used OBB formats in aerial objects detection and their problems in regression-based methods.

However, the regression-based OBB detector always suffers from the regression targets discontinuity problem. The OBB formats commonly used in the remote sensing field are rotated bounding box (RBB) and quadrilateral bounding box (QBB), which set up rules to uniquely determine an OBB for each set of parameters, as shown in Fig. 1.(a) and (b) respectively. Yet, there are some extreme cases in these rules, in which a slight change in angle may lead to a completely different expression. If the boxes before and after the change are simulated as candidates and match the same ground-truth, the completely different expressions cause a sudden change in regression targets, which is not conducive to the learning of the detector. Fig. 1.(c) and (d) show unsatisfied detection in these extreme cases.

In this paper, the regression targets discontinuity problems of two mainstream OBB formats in a regression-based detector are discussed. After that, a keypoint-based detection method [13] is introduced. By directly detecting the skew box's endpoints, the original offset regression is transformed into a unique point-guided keypoint estimation. The direct prediction of bounding box endpoints on the whole image poses a problem of difficult combination due to the obscure

Manuscript received

Q. Song, F. Yang, L. Yang, C. Liu and M. Hu are with the Pattern Recognition and Intelligent Vision Laboratory(PRIV), Beijing University of Posts and Telecommunications(BUPT), Beijing 100876, China.

L. Xia is with the Space Engineering University, Beijing 101416, China.

object features and a large number of similar objects in remote sensing images. With the help of RPN, we first obtain the horizontal circumscribed rectangle of each object, then classify and detect the keypoints inside the proposal. At this time, each endpoint corresponds to an object uniquely, avoiding point combination in the mainstream keypoint-based detector. Our method is entirely dependent on the visible appearance of the object. For the truncation problem caused by the prevailing image clipping of remote sensing images, an extra center-point prediction is added for post-processing. For the inconsistency between the final score and location quality caused by using classification confidence as the score of detection results, the final score is corrected by the endpoint's location score. Through introducing a new pipeline with potential solutions to the problems mentioned above, the method in this paper has achieved satisfactory results in DOTA [14] and HRSC2016 [15].

The major contributions of this study are as follows:

(1) We introduce a detection pipeline that directly predicts the endpoint of the quadrilateral. By converting the original offset regression into unique point-guided keypoint estimation for localization, our method avoids discontinuous regression targets and achieves state-of-the-art performance.

(2) We add the center point prediction in the network and use it for center-point post-processing to solve the endpoint loss problem caused by remote sensing image cropping.

(3) We use the average localization score of the OBB endpoints to correct the classification score to obtain the final score, which improves the correlation between classification confidence and localization quality.

II. Related Work

Current object detection consists of anchor-based and anchor-free detector. Further, the former one can be divided into one stage and two stages methods, and the latter one falls into keypoint-based and center-based methods. The keypoint-based detecter is based on keypoint estimation, while the others are based on regression.

A. General object detection

Faster R-CNN [16] dominates the two-stage detector, which can be divided into a region proposal network(RPN) and a region-wise prediction network (R-CNN). Since then, there have been many works that have improved their performance in different ways. OHEM [17] selects some ROIs with the largest loss as training samples to optimize network parameters. HyperNet [18] and FPN [19] integrates feature maps at different levels so that high-level semantic information features and low-level detailed features complement each other. Cascade R-CNN [3] reforms the traditional cascade connection, the output of each stage got optimized corresponding different IoU thresholds. SNIP [20] introduces an image pyramid to obtain images of different sizes, and only the gradient of the RoI corresponding to the size of the training data of the pre-training model is returned. Mask R-CNN [21] and Parsing R-CNN [22], [23] added an extra task branch to enriching supervising information. TridentNet [24] built three parallel branches with different receptive fields, each branch is responsible for samples within a certain scale. After the appearance of YOLOv2 [1] and SSD [25], the one-stage detectors show their great advantage on the computational efficiency. DSSD [26] upsamples feature maps and detects small objects on lower layers to improve performance for small objects. RefineDet [27] put two-step regression into a one-stage framework, generating more accurate refined anchors to improve detector performance. In recent years, the anchor-free detector has become popular. The center-based method obtains the initial regression state through the model rather than uses hyperparameters to generate. GA-RPN [28] defines the pixels in the center region of the object as positives to predict the location, width, and height of proposals. FCOS [4] regards all the locations inside the object bounding box as positives with four distances to four borders and a novel centerness score to detect objects.

Keypoint-based methods follow the standard keypoint estimation pipeline, that is, detecting all keypoints of different objects and then grouping them. CornerNet [5] uses the upper left and lower right corners of HBB to represent an object. After detecting the upper left and lower right corners of all objects in the image, it determines whether the two corners are from the same object by embedding. CenterNet [6] adds extra center-point prediction based on CornerNet. By judging whether there is a center keypoint of the same category in the center area of a pair of corner points, the box with an incorrect group is filtered. ExtremeNet [7] uses four poles (left-most, top-most, right-most, and bottom-most) and the center-point to represent an object. Given four extreme points, if their geometric center is predicted with the high response in the center map, then commits the extreme points as a valid detection. Reppoints [29] represents objects as a set of sample points and learns to arrange themselves. There are often a large number of dense [30] and similar-looking objects in remote sensing images, which brings great difficulty to group keypoints. Our method uses the four endpoints of OBB as key points, and the proposals generated by RPN ensure that every four endpoints directly correspond to a particular object, avoiding the problem of grouping.

B. Oriented Object detection in remote sensing images

In addition to horizontal object detection [31], [32], some works have explored oriented object detection. The rotated bounding box(RBB) or quadrilateral(QBB) is often used to represent the object in the remote sensing image, and the mainstream detector takes localization as a regression task. FR-O [14] and ICN [33] use RPN [16] generation horizontal proposals, directly regress the offset of OBB relative to HBB, and the enormous gap makes performance unsatisfactory. R-DFPN [34] refers to RRPN [9], generates a large number of proposals with angle information, and then returns the offset of OBB relative to these oriented RoIs. Although regression is simpler, the exponentially increasing proposals and the corresponding skew IoU bring a huge calculation burden. ROI Transformer [10] proposes RoI learner that converts the horizontal RoIs generated by RPN into oriented RoIs and

then performs feature extraction and refinement on it. Gliding Vertex [12] uses QBB to represent objects, regressions the offset from the four endpoints of QBB to the corresponding endpoints of its horizontal circumscribed rectangle. Some works have also borrowed ideas from semantic segmentation [35]. APE [36] generates candidate bounding boxes from the shrunk segmentation map of the OBB, which is the same as EAST [37]. Segmentation maps with 8 channels are predicted in the RPN stage to represent rotated proposals for regression. Unlike the regression-based method mentioned above. Our method directly predicts the four endpoints of the quadrilateral, changing the localization from a regression problem to a keypoint estimation problem.

C. Detection Score Correction

Previous works have proved that the classification score and localization quality of the two-stage network are not strongly related. Some works are devoted to correct the final detection score. Tychsen-Smith et al. [38] regards the Intersection of Union (IoU) between the predicted box and ground truth as a classification task, and uses the predicted IoU to correct the detection score. IoU-Net [39] directly regresses IoU that is used for both score and bounding box correction. SoftNMS [40] uses the IoU between predicted boxes to corrects the box with a low score by replacing the original score with a slightly lower score instead of directly setting zero. CPM R-CNN [41] proposed a Fused Scoring Network to predict the IoU score and combined it with the classification score. The methods above introduce a large amount of calculation in the process of calculating IoU. OPLD uses class agnostic keypoints estimation to obtain endpoints of the OBB. High response in the heatmap means that the corresponding position is highly likely to be an endpoint. Therefore, the localization quality of the box is measured by the mean response of four endpoints inside, which can be combined with the classification score to provide a more reasonable detection score.

III. Method

A. Motivation

The mainstream regression-based aerial object detector matches a ground truth for each proposal during training and encodes the offset between them as the supervision information of regression. Whether RBB or QBB, there will be extreme cases in this process, which cause the regression targets discontinuity problem.

RBB is an oriented rectangle that can be determined by (x, y, w, h, θ), its determination rules are shown in 1.(a). The center-point of RBB is the same as that of OBB. Take the lowest point of OBB as the origin, rotate the horizontal axis counterclockwise, the first side touched is w, the other side is h, and the angle rotated is θ. To obtain more accurate prediction results through larger weighting, the regression objectives of RBB are as follows:

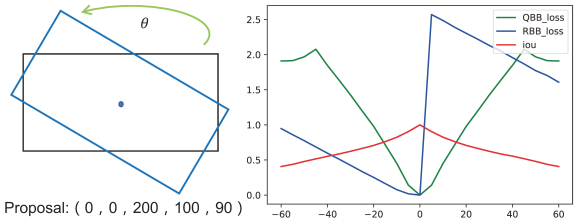


Fig. 2. The discontinuity of regression target in regression-based detection. Given a certain proposal(black box), we rotate it counterclockwise by an angle as its target(blue box). The graph on the right shows the IoU of the two boxes and the smooth L_1 Loss calculated using the offsets between them.

TABLE I
THE REGRESSION TARGET CHANGES DRAMATICALLY WHEN THE SUPERVISION POINTS ARE MISPLACED. A POSITIVE ANGLE REPRESENTS A COUNTERCLOCKWISE ROTATION

angle of rotation	regression targe
-5	(0.4, -0.5, 0.9, -0.1, -0.4, 0.45, -1, 0.1)
0	(0, 0, 0, 0, 0, 0, 0, 0)
5	(6, -3.3, 14.4, 6.55, -6.2, 3.35, -14.3, -6.65)

$$
\begin{aligned}
t_x &= 10 * \frac{x_g - x_p}{w_p}, t_y = 10 * \frac{y_g - y_p}{h_p} \\
t_w &= 5 * \log \frac{w_g}{w_p}, t_w = 5 * \log \frac{h_g}{h_p} \\
t_\theta &= 5 * \frac{(\theta_g - \theta_p) * \pi}{180}
\end{aligned} \tag{1}
$$

where x, y, w, h, θ denote the RBB's center-point coordinates, width, height, and angle, respectively. Variables x_g and x_p correspond to ground-truth and proposal respectively.

QBB is an irregular quadrilateral determined by four coordinate vectors, and the first point is determined under specific rules. Fig. 1.(b) shows the starting point determination rule used in DOTA [14], that is, the closest point to the top left corner is the first point. In this paper, the regression objectives of QBB are as follows:

$$
t_{x_i} = 10 * \frac{x_{g_i} - x_{p_i}}{w_{p_i}}, t_{y_i} = 10 * \frac{y_{g_i} - y_{p_i}}{h_{p_i}}, i \in [0, 1, 2, 3] \tag{2}
$$

To express the discontinuity of the regression target more intuitively, we use smooth L_1 loss that is commonly used in bounding box regression to calculate the loss. In the case of different rotation angles, the IoU between two boxes and the resulting loss is shown in Fig. 2.

The extreme situation of RBB regression occurs when the bounding box near the horizontal. When an OBB reaches the level and continues to rotate counterclockwise, the box's length and width will be reversed, and the angle will change from $90°$ to $0°$. The extreme situation of QBB regression occurs at around $45°$. At this time, the distance between the two endpoints and the upper left corner of the horizontal circumscribed rectangle is almost identical. A slight change of the angle will confuse the order of endpoints, resulting in the supervision points' dislocation. We rotate the box in Fig. 2

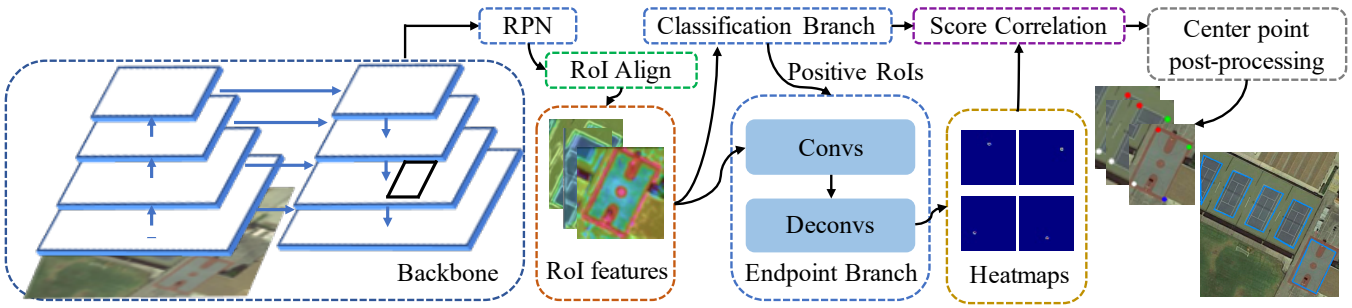


Fig. 3. The pipeline of our method. OPLD mainly consists of four parts: backbone with FPN for feature extraction, RPN for candidates proposals prediction, classification branch for background filtering and the endpoint branch for generating heatmaps to locate the endpoints of OBB. 'P' and 'C' represent the number of endpoints to predict and the number of channels used to predict each point.

by $45°$ as a new proposal, and use the same method to generate its targets. The change of supervision information can be seen in Table I.

We believe that a detector should extract similar features for similar objects. Considering that the starting point of QBB is completely determined by the spatial information, we introduce a direct prediction method of QBB endpoint based on keypoint estimation.

B. Overall Pipeline

The pipeline of OPLD is shown in Fig. 3, which is a two-stage detector. We employ a BackBone with FPN [19] to extract features of input images. Since the features of objects in bird's-eye view are not apparent and there may be a large number of similar targets in remote sensing images, RPN [16] is adopted to limit the region for keypoint estimation, which avoids the grouping in mainstream keypoint-based detection. Considering the computational burden brought by skew IoU calculation, what will be predicted in RPN is horizontal enclosing rectangle of the object rather than RBB in RRPN [9]. The proposals are sent to the classification branch that is entirely consistent with Faster R-CNN [16] to obtain category and classification confidence. After non-maximum suppression (NMS), the positive proposals that are predicted as objects are selected for keypoint estimation in the subsequent branch.

The endpoint branch is a fully convolutional architecture that can capture the spatial information explicitly. The RoI features from RoI align [21] are converted into feature maps with the shape of $14 \times 14 \times P \times C$ by N convolution layers, where P is the number of predicted points, and C is the number of channels used to predict each point. After up-sampling by two deconvolution layers, P heatmaps with a resolution of 56×56 are obtained, which are category agnostic and correspond to different endpoints. After softmax, the response $h_{p_{ij}} \in (0, 1)$ represents the probability that the position(i, j) on the p heatmap is p endpoint.

During training, the positive samples' decision condition of the endpoint branch is that the IoU between proposal and ground-truth is greater than 0.5, which is consistent with the classification branch. As shown in Fig. 4, a proposal (the blue bounding box) cannot cover the matched OBB' s endpoints, and the lack of supervision point leads to inefficient utilization of training samples. While in the inference stage,

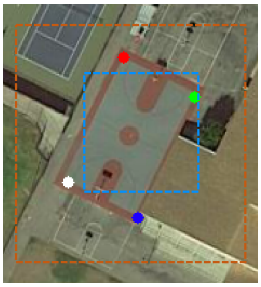


Fig. 4. To ensure the endpoints inside heatmap, the proposal (blue) will be expanded to double the size (orange) before mapping.

by simply choosing the maximum response on the heatmap, we may obtain a completely incorrect location due to the ground truth endpoint is outside the proposal. To ensure that all endpoints can be mapped on heatmap without introducing additional interference information, the side length of the proposal is doubled before mapping, while the region for RoI feature extraction is still the same. Then, each heatmap has a corresponding supervision map, the point (I_x, I_y) in origin image will be mapped to the point (H_x, H_y) in supervision map by:

$$\begin{aligned} H_x &= \frac{I_x - P_x}{w_h} * 2w_p \\ H_y &= \frac{I_y - P_y}{h_h} * 2h_p \end{aligned} \tag{3}$$

where (P_x, P_y) is the position of the upper left corner of the proposal in the input image, w_h and h_h are the width and height of the supervision map, w_p and h_p are the width and height of the proposal.

Following the above mapping, all target endpoints of positive proposals will be covered by the corresponding region of the supervision map. After getting endpoint (H_x, H_y), a positive region was determined with r as the radius, and the rest region was negative. We use BCE loss for optimization. Therefore, the training objectiveness of the whole network is:

$$L_{all} = \lambda_1 L_{rpn} + \lambda_2 L_{cls} + \lambda_3 L_{endpoint} \tag{4}$$

$$L_{endpoint} = BCELoss(h_g, h) \tag{5}$$

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/JSTARS.2020.3036685, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

where h_g represents the groud-truth heatmaps, h represents the predicted heatmaps. L_{rpn} and L_{cls} are consistent with Faster R-CNN [16].

In the Inference stage, we select the pixel with the highest confidence on each predicted heatmap and calculate the corresponding location on the original image as the endpoint. We can get the final detection by direct connecting the four endpoints.

C. Center-point Post Processing

Despite directly predict the endpoints of QBB is more flexible and intuitive, it is also limited by the image boundary. In the field of aerial object detection, oversized images are clipped into patches before fed into the network. When an object is truncated in the clipping process, its retention depends on the IOU. Only those objects with small truncated parts are retained. Even so, the approximate OBB of the remaining part will deform, and the missing part of the appearance cannot be predicted in OPLD.

As shown in Fig 5, we find that the RBB center point (blue) of the complete object, the center point (red, green) of the two diagonal lines of QBB, and the center point (black) of the four endpoints of QBB almost coincide. The distance between the four points of the truncated object increases over the growth of the truncation degree. In contrast, RBB's center point can more accurately express the center of the complete object, and the rotated circumscribed rectangle of the incomplete QBB is basically the same as that of the complete QBB.

Therefore, in addition to the four points of QBB, the RBB center point is also added to the keypoint estimation. In the inference stage, a small central region is obtained by taking the predicted center point of the network as a benchmark and scaling the width and height of the proposal. The determination rules for the central region are:

$$\begin{aligned} x_{tl} &= x_{ctr} - \frac{w_p}{2n} \\ y_{tl} &= y_{ctr} - \frac{h_p}{2n} \\ x_{br} &= x_{ctr} + \frac{w_p}{2n} \\ y_{br} &= y_{ctr} + \frac{h_p}{2n} \end{aligned} \tag{6}$$

where (x_{tl}, y_{tl}) denote the coordinates of the top-left corner of the central area, and (x_{br}, y_{br}) denote the coordinates of the bottom-right corner of the area. (x_{ctr}, y_{ctr}) is the center point predicted by the endpoints branch. w_p and h_p denote width and height of the proposal. n is a constant that determines the scale of the central area.

For proposals with an area less than 15625, n = 10, otherwise n = 15. If the center points of the two diagonals of the bounding box are not in the region, it is considered that the box needs to be corrected. The four endpoints of the result box are transformed into its rotated circumscribed rectangle and then back to QBB to obtain the final result.

D. Detection Score Correction

In OPLD, the classification score is derived from the horizontal circumscribed rectangle, and the final result is the QBB

(a) (b)

Fig. 5. Different midpoint of OBB. In (a), the truncated OBB's center-point is far from the center-point of RBB. In (b), the rotated circumscribed rectangle of truncated OBB(white) is almost the same as RBB.

Fig. 6. The inconsistency between localization quality and classification score in OPLD.

inside it, which further increases the gap between classification score and localization quality. It may lead to the situation that detection results with low localization quality but high classification scores filter out the results with low classification confidence at the final NMS, even though these results are more accurately in localization, as shown in Fig. 6.

We use a straightforward weighted calculation to combine the classification score with the localization quality. The response on the heatmap output from the endpoint branch reflects the localization confidence. In the case of higher confidence, the more accurate the endpoint localization, and the more the number of accurate endpoints, the more accurate the composite box is. Therefore, the average responses at the four endpoints are used as the localization quality of OBB. Since the endpoint predictions are category agnostic, the classification scores obtained from the classification branch are calculated together to obtain the final detection score:

$$score_{det} = \alpha * score_{cls} + \beta * mean(score_{endpoint}) \tag{7}$$

where $score_{cls}$ is the confidence derived from the classification branch, $score_{endpoint}$ is the maximum response value on each heat map. To satisfy the definition of probability, $\alpha + \beta = 1$.

IV. Experiments

Our experiment is based on Pytorch, implemented on a server with four blocks of 12GB memory TITAN X (Pascal) GPU, and evaluated on DOTA [14] and HRSC2016 [15]. The ablation study was conducted on DOTA, which is a challenging arbitrary oriented object detection dataset.

A. Datasets and Protocols

1) DOTA: DOTA is one of the largest aerial object detection datasets. It contains 2806 images of 800×800 to 4000×4000, and 15 categories of 188282 examples are annotated by a quadrilateral, including 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), helicopter (HC). The training set, validation set, and test set account for 1/2, 1/6, and 1/3 of the entire data set. The results of DOTA are obtained by submitting predictions to the official DOTA evaluation server.

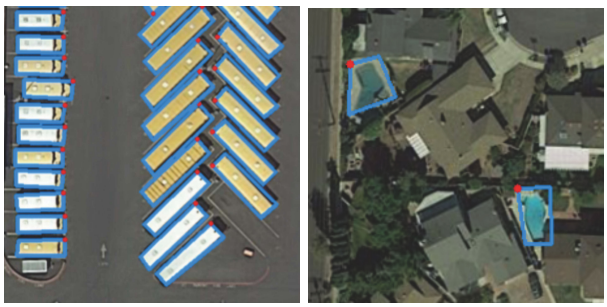
Multi-scale training and testing, data augmentations that are widely used in the DOTA leaderboard are applied when d with state-of-the-art detectors. We resize the original images at two scales (0.5 and 1.0) before dividing the images into patches. After resizing, we divide the resized images into 1024×1024 patches with an overlap of 200 in both the training and inference stage. With all these processes, we obtain about $27,600$ patches to train. Each image is randomly resized to one size of {800, 912, 1024} and rotated an angle from an angle set $\{0°, 90°, 180°, 270°\}$ with a probability of 0.5 in training stage.

2) HRSC2016: HRSC2016 is a dataset for ship detection in remote sensing images. It has 1061 images annotated with RBB, including 436 images for training, 181 images for validation, and 444 for testing. The image size ranges from 300×300 to 1500×900 pixels. These images are not oversized, so we directly predict the four endpoints without the center point used in post-processing. We convert RBB to QBB by taking the point closest to the given bow coordinate as the starting point. We use random rotation as used in DOTA for training, and resize it to (1024, 1333) in both the training and testing stage, where 1024 represents the short side of the image, and 1333 is the longest side of the image. We use the standard VOC-style AP metrics with an IoU threshold of 0.5 to evaluate.

B. Implement Detail

We use ResNet50/101 [42] based FPN [19] as the backbone, which is pretrained on ImageNet [43]. For FPN, we use pyramid levels {P2, P3, P4, P5, P6}, which have strides of {4, 8, 16, 32, 64} respect to the input image. There is no special design adopted in RPN, different anchor aspect ratios {1:2, 1:1, 2:1} are adopted at each level, five anchor scales of $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ are corresponding to five pyramid levels, so we get 15 anchors over the pyramid. We apply RoIAlign to generate RoI features and set the output resolution 7 in the classification branch and 14 in the endpoint branch. For supervision map generation, the radius of the positive region is 3.

We use SGD with a weight decay of 0.0001 and a momentum of 0.9 with a total of 8 images per mini-batch (2 images per GPU). We train 12 epochs in total with an initial learning rate of 0.01 and decrease it by a factor of 0.1 at epoch 9 and 11. In joint loss function, we set $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 15$.



(a) (b)

Fig. 7. Visualization of OPLD results, the red point is the starting point.

TABLE II
COMPARISON OF OPLD WITH RRPN IN SPEED.

method	backbone	image size	inference(FPS)
RRPN	ResNet101-FPN	(1024, 1024)	2.8
OPLD	ResNet50-FPN	(1024, 1024)	7.3
OPLD	ResNet101-FPN	(1024, 1024)	5.2

In the inference stage, RPN produces 2000 RoIs for RoIAlign and classification per image. Thereafter, we apply NMS with a 0.4 IOU threshold to select positive proposals, which send to the endpoint branch for QBB prediction and uses Quadrilateral-NMS with a threshold of 0.2 to produce the final result.

C. Baseline

Our baseline uses ResNet50-FPN without any data augmentation. As mentioned before, the regression-based OBB detection suffers from the regression targets discontinuity problem, which leads to unsatisfactory results in some cases. OPLD changes object localization from regression to keypoint estimation, directly predicting the QBB endpoint for each object.

For objects with special head features, such as vehicles and airplanes, each point's location can be easily found, as shown in Fig. 7.(a). For objects lacking head features, such as playgrounds and swimming pools, following the QBB starting point determination rule used in DOTA, OPLD can also learn the starting point expression derived from spatial features, as shown in Fig. 7.(b). Whether an object is horizontal or inclined at 45 degrees, our method can detect it well.

To avoid the heavy calculation burden brought by Skew IoU and NMS, we predicted the horizontal circumscribed rectangle of the object in the RPN stage and then predicted the OBB through the subsequent branches. Compared with the same two-stage method but using oriented RoIs, as shown in Table II, we have an advantage in speed.

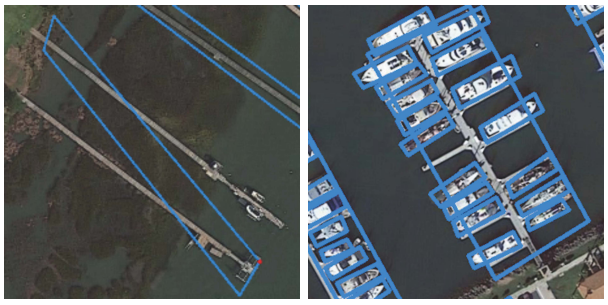
However, the horizontal external rectangle of oriented objects always contains some background, which also inevitably brings it into ROI features. The use of these rectangles also brings two typical failure cases. The first one occurs when a part with the same appearance enters RoI, an endpoint may be detected as a corresponding one of another object, as shown in Fig. 8.(a). This is consistent with the problems faced by

TABLE III
DIFFERENT PARAMETERS ON DOTA-V1.0 DATASET. ALL RESULT IS BASED ON RESNET50-FPN WITHOUT DATA AUGMENTATION.

Parameter	Value	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
(a) r	0	71.22	57.50	46.76	63.41	58.38	52.91	68.04	90.33	76.11	78.11	30.38	60.05	53.24	55.83	33.68	59.73
($\lambda_3 = 15$)	1	78.0	72.7	48.5	63.5	67.7	64.9	74.9	90.1	78.5	83.6	40.3	63.5	59.9	60.3	42.1	65.9
	2	79.42	73.76	48.87	65.40	69.76	69.39	76.52	90.41	84.31	83.63	48.61	62.55	64.27	66.20	46.30	68.62
	3	79.97	73.98	52.78	66.28	71.47	71.31	77.04	90.73	83.61	83.40	49.22	62.90	65.73	66.44	52.23	69.81
	4	80.28	73.90	49.45	65.70	72.56	72.09	83.45	90.85	79.96	84.24	41.52	59.91	64.55	67.06	51.70	69.15
(b) λ_3	1	72.93	61.16	46.67	59.07	59.90	55.38	70.73	90.50	77.41	83.22	32.92	62.95	53.83	58.15	36.00	61.39
($r = 2$)	5	78.55	69.27	48.78	66.55	68.15	66.00	75.62	90.20	81.42	83.45	43.85	62.42	62.66	60.73	49.00	67.11
	10	79.04	75.35	49.51	64.65	69.79	68.90	76.62	90.38	79.77	83.89	46.22	60.81	63.21	67.24	50.41	68.39
	15	79.42	73.76	48.87	65.40	69.76	69.39	76.52	90.41	84.31	83.63	48.61	62.55	64.27	66.20	46.30	68.62
	20	79.36	74.17	50.79	67.91	70.35	69.85	76.71	90.13	79.12	78.41	48.30	61.46	64.62	65.93	50.62	68.52

TABLE IV
ABLATION EXPERIMENTS ON DOTA-V1.0 DATASET.

Method	Config	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
(a) Baseline	$r = 3$, $\lambda_3 = 15$	79.97	73.98	52.78	66.28	71.47	71.31	77.04	90.73	83.61	83.40	49.22	62.90	65.73	66.44	52.23	69.81
(b) Post-Processing	center	80.22	74.11	52.60	66.86	72.48	72.76	77.61	90.75	83.68	83.65	51.83	62.95	66.26	65.94	54.57	70.42(↑ 0.61)
	all	79.99	74.09	52.08	66.77	72.25	72.46	77.35	90.55	83.58	83.49	50.64	63.16	65.33	67.24	54.27	70.21(↑ 0.40)
(c) Score-Weighting	$(0.9, 0.1)$	81.25	75.38	53.01	66.42	72.43	73.98	84.48	90.89	83.13	83.53	51.01	62.50	65.91	66.70	53.73	70.96(↑ 1.15)
(cls, loc)	$(0.8, 0.2)$	81.22	75.42	52.68	66.56	72.60	74.26	84.60	90.89	83.21	83.33	51.71	62.54	65.65	66.54	54.08	71.02(↑ 1.21)
	$(0, 1)$	78.58	53.24	39.21	33.66	57.27	57.22	73.67	90.30	62.23	76.14	20.52	43.66	46.25	43.37	13.91	52.62(↓ 17.19)
(d) Both	center+$(0.8, 0.2)$	81.29	75.28	52.55	66.73	73.17	74.93	85.03	90.89	83.30	83.34	53.46	62.59	66.01	66.29	55.66	71.37(↑ 1.56)



(a) (b)

Fig. 8. Some typical failure predictions of our method.(a) occurs when an object with the same appearance is mixed in the RoI feature. (b) occurs when there are other small objects near the large object

mainstream methods based on keypoint estimation. The use of RPN has helped us alleviate this problem, but it has not been entirely resolved. The other occurs when there are many objects near a large object. We map the endpoints to the heatmap and determine the positive region by a radius. The appearance of adjacent objects may also be in this area, which makes the detector unable to learn the endpoint position of the large target correctly, as shown in Fig. 8.(b).

D. Comparison with Different Parameters

In this section, we will do a series of comparative experiments based on ResNet50-FPN without data augmentation to analyze the influence of different parameter settings.

1) Radius of positive region: The training of the endpoint branch in OPLD is guided by the supervision map discussed in Sec. III-B. Taking the point mapped from the image as the center and r is the radius, a circle is determined as a positive region. The impact of the radius is summarized in Table. III.(a). It is too strict for OPLD that only predicting the mapping point, so a radius of the positive region is necessary. With the increase of r, the performance of OPLD increases until the radius is 4.

2) Weight of localization: When using joint loss as training objectiveness, the weight of each loss determines the importance of different tasks. Detection is a combination of classification and location, so we set λ_1 and λ_2 to 1 by default and explore different λ_3 in Table. III.(b). It is necessary to assign a high weight to the localization task, but when it is high enough, the performance is not so sensitive to this parameter.

E. Ablation Study

Ablation experiments were also performed. All models are based on ResNet50-FPN without data augmentation to ensure accuracy. As shown in Table IV, Our baseline gets 69.81 mAP for OBB task, and achieves 71.37 mAP by using proposed center point post-processing(CPP) and score-weighting(SW).

1) CPP: In the experiment, we found that the center point post-processing(CPP) not only solve the object truncation problem but also correct some low-quality detection results. To verify the effectiveness of CPP, we also experimented with all the detection results conversion processing. Table. IV(b) shows the performance improvement of the different post-processing methods: the total post-processing improved AP50 by 0.40, while CPP improved by 0.61. The additional improvement indicates that our central point judgment condition is meaningful.

2) Score Correlation: Only taking the classification score as the final score cannot adequately reflect the localization quality. We use the average value of each endpoint's response on the heatmap as the localization quality of the OBB and get the detection score by weighting the classification score to ensure that both classification and localization are taken into

TABLE V
QUANTITATIVE COMPARISON OF OBB TASK ON DOTA-V1.0 DATASET.

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
R-DFPN [34]	ResNet101-FPN	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
RRPN [9]	ResNet101	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 [33]	ResNet101-FPN	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 [10]	ResNet101-FPN	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
CAD-Net [44]	ResNet101-FPN	87.8	82.4	49.4	73.5	71.1	63.5	76.7	90.9	79.2	73.3	48.4	60.9	62.0	67.0	62.2	69.9
SCRDet [11]	ResNet101-FPN	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
SARD [45]	ResNet101-FPN	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
FADet [46]	ResNet101-FPN	90.21	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.40	84.68	53.40	65.42	74.17	69.69	64.86	73.28
Gliding Vertex [12]	ResNet101-FPN	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
Mask-obb [47]	ResNeXt101 [48]-FPN	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
FFA [49]	ResNet101-FPN	90.1	82.7	54.2	75.2	71.0	79.9	83.5	90.7	83.9	84.6	61.2	68.0	70.7	76.0	63.7	75.7
APE [36]	ResNeXt101-FPN	89.96	83.64	53.42	76.03	74.01	77.16	79.45	90.83	87.15	84.51	67.72	60.33	74.61	71.84	65.55	75.75
Ours	ResNet101-FPN	89.37	85.82	54.10	79.58	75.00	75.13	86.92	90.88	86.42	86.62	62.46	68.41	73.98	68.11	63.69	76.43

TABLE VI
QUANTITATIVE COMPARISON OF HBB TASK ON DOTA-V1.0 DATASET.

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
R-FCN [50]	ResNet101	81.01	58.96	31.64	58.97	49.77	45.04	49.29	68.99	52.07	67.42	41.83	51.44	45.15	53.30	33.89	52.58
FMSSD [51]	VGG16 [52]	89.11	81.51	48.22	67.94	69.23	73.56	76.87	90.71	82.67	73.33	52.65	67.52	72.37	80.57	60.15	72.43
ICN [33]	ResNet101-FPN	90.00	77.70	53.40	73.30	73.50	65.00	78.20	90.80	79.10	84.80	57.20	62.10	73.50	70.20	58.10	72.50
IoU-Adaptive [53]	ResNet101-FPN	88.62	80.22	53.18	66.97	76.30	72.59	84.07	90.66	80.95	76.24	57.12	66.65	74.08	66.36	56.85	72.72
SCRDet [11]	ResNet101-FPN	90.18	81.88	55.30	73.29	72.09	77.65	78.06	90.91	82.44	86.39	64.53	63.45	75.77	78.21	60.11	75.35
FADet [46]	ResNet101-FPN	90.15	78.60	51.92	75.23	73.60	71.27	81.41	90.85	83.94	84.77	58.91	65.65	76.92	79.36	68.17	75.38
Mask-obb [47]	ResNeXt101-FPN	89.69	87.07	58.51	72.04	78.21	71.47	85.20	89.55	84.71	86.76	54.38	70.21	78.98	77.46	70.40	76.98
Ours	ResNet101-FPN	89.44	86.16	59.69	79.62	75.36	79.38	86.94	90.88	86.63	86.70	65.32	68.53	77.59	77.17	65.91	78.35

TABLE VII
EVALUATION RESULTS OF THE OBB TASK ON THE HRSC2016 DATASET.

Method	RC2 [54]	R^2PN [55]	RRD [56]	RoI Trans [10]	Gliding Vertex [12]	Ours
mAP	75.7	79.6	84.3	86.2	88.20	88.44

account. Table. IV.(c) shows the influence of different weight coefficients on the detector.

The additional localization score makes the accuracy increase significantly, and slight weight changes make no difference. If we only consider the localization quality, there will be a significant decrease in detection performance because the endpoints score does not contain the classification information at all.

F. Comparison with State-of-the-Art Detectors

We compare our proposed OPLD with the state-of-the-art algorithms on two datasets DOTA [14] and HRSC2016 [15].

1) DOTA: We compare OPLD with the state-of-the-art methods on OBB and HBB tasks of DOTA dataset in Table V and Table VI. To ensure the fairness of comparison, we adopt ResNet101-FPN as the backbone and compare it with the similar backbone method. All data augmentations that we used are explained in IV-A1. Our OPLD achieved 76.43 and 78.35 mAP in OBB and HBB tasks, respectively. When comparing category by category, OPLD is outstanding in bridge, ship, ground-track field, and harbor. The first two categories have a high aspect ratio, which makes the training of regression difficult to converge. In the proposal generation stage, predicting the horizontal circumscribed rectangle instead of the OBB greatly alleviates this problem. The harbor is usually not a regular rectangle, so the predicted quadrilateral boundary box can better fit it. However, the performance on the soccer-ball field and helicopter is obviously worse than the leading method. We found that many soccer-ball fields are classified as ground track field, and helicopters are classified as ships. The classifier of OPLD is not strong enough to distinguish them. We thought that the weight in joint loss affects the performance of the classifier. Besides, It is worth noting that some of the compared methods, such as FADet and SCRDet, use attention mechanism [57]–[59], and some of them, such as mask OBB, use Inception module [60], [61]. These methods will generally bring improvement, but we did not use them. Some visualization results on OBB tasks can be found in Fig. 9.

2) HRSC2016: The results on HRSC2016 are shown in Table VII. We use ResNet50-FPN as the backbone to maintaining the fairness of comparison. This dataset gives the OBB annotations in the form of RBB and the position of the bow. It should be noted that the bow position is visually determined and is usually not at the endpoint of the bounding box. In the process of converting RBB to QBB, we find that both the left and right bow endpoints may be the starting point when we choose the point closest to the bow as the starting point, which is very detrimental to our method. We use OpenCV's order, and the results are much better, but still not good enough. OPLD is not much improved compared to other methods, but considering the aforementioned problems, this result is still

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/JSTARS.2020.3036685, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

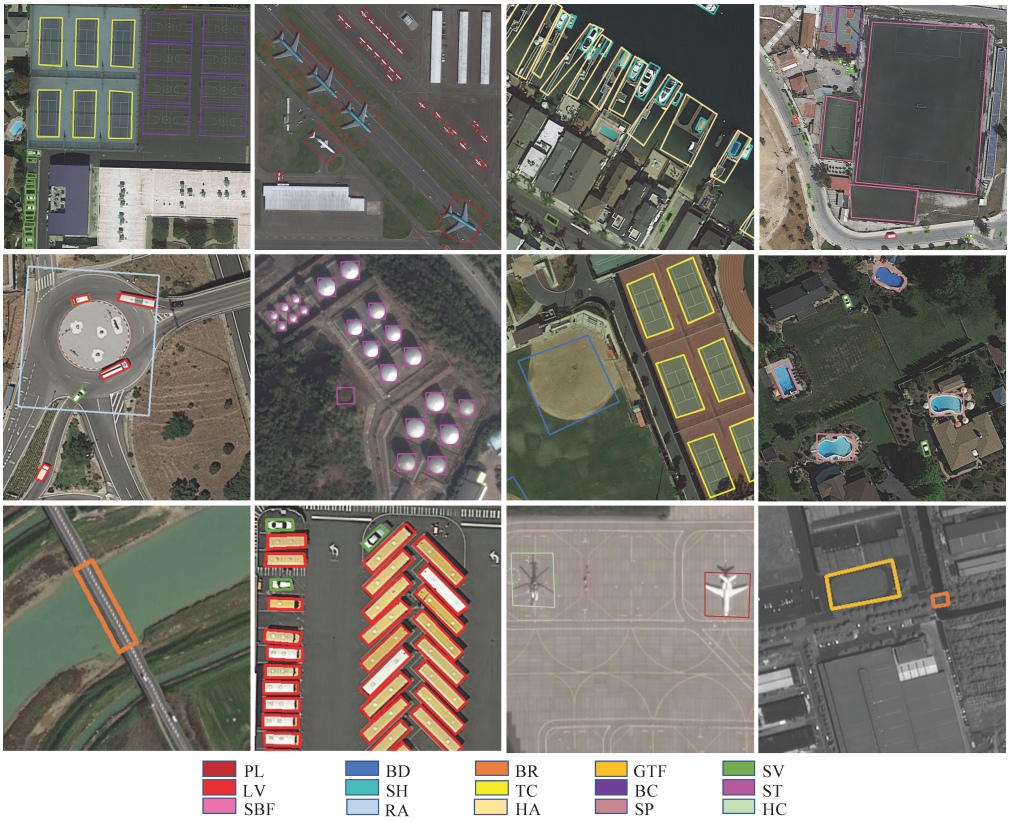


Fig. 9. Visualization of detection results using OPLD on OBB task of DOTA. The threshold for visualization is 0.4. The corresponding colors of different types of objects are shown in the figure.

Fig. 10. Visualization of detection results using OPLD on OBB task of HRSC2016. Ships of different appearances are counted as one category.

acceptable. Some visualization results on the HRSC2016 can be found in Fig. 10.

V. Conclusion

In this paper, we analyze the regression targets discontinuity problem in the regression-based detection in remote sensing images and propose OPLD, which transforms localization from a regression problem to a keypoint estimation problem. Besides, we also used center point post-processing to solve the problem of limited expression of QBB, and found that it can slightly improve the detection results of poor localization quality in experiments; for the problems of low correlation between the quality of localization and detection caused by using classification score as the final score in mainstream methods, we use a simple score weighting but got a significant improvement. The experimental results based on DOTA and HRSC2016, state-of-the-art performance on DOTA prove the effectiveness of our method in oriented aerial object detection.

References

[1] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,

pp. 7263–7271, July 2017.

[2] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, Oct 2017.

[3] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6154–6162, June 2018.

[4] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 9627–9636, October 2019.

[5] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *The European Conference on Computer Vision (ECCV)*, pp. 734–750, September 2018.

[6] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 6569–6578, October 2019.

[7] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 850–859, June 2019.

[8] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, December 2015.

[9] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[10] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning roi transformer for detecting oriented objects in aerial images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 8232–8241, October 2019.

[12] Y. Xu, M. Fu, Q. Wang, Y. Wang, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2020.

[13] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid r-cnn," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7363–7372, June 2019.

[14] G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3974–3983, June 2018.

[15] Z. Liu, Y. Liu, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *International Conference on Pattern Recognition Applications and Methods*, 2017.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[17] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 761–769, June 2016.

[18] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 845–853, June 2016.

[19] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, July 2017.

[20] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection – snip," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3578–3587, June 2018.

[21] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, Oct 2017.

[22] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 364–373, June 2019.

[23] L. Yang, Q. Song, Z. Wang, M. Hu, C. Liu, X. Xin, W. Jia, and S. Xu, "Renovating parsing r-cnn for accurate multiple human parsing," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[24] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 6054–6063, October 2019.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "Ssd: Single shot multibox detector," in *The European Conference on Computer Vision (ECCV)*, pp. 21–37, December 2015.

[26] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd : Deconvolutional single shot detector," *CoRR*, 2017.

[27] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4203–4212, June 2018.

[28] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2965–2974, June 2019.

[29] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 9657–9666, October 2019.

[30] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2337–2348, 2018.

[31] Z. Tian, W. Wang, R. Zhan, Z. He, J. Zhang, and Z. Zhuang, "Cascaded detection framework based on a novel backbone network and feature fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3480–3491, 2019.

[32] S. Jiang, W. Yao, M. S. Wong, G. Li, Z. Hong, T. Kuc, and X. Tong, "An optimized deep neural network detecting small and narrow rectangular objects in google earth images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1068–1081, 2020.

[33] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Asian Conference on Computer Vision(ACCV)*, pp. 150–165, 2018.

[34] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sensing*, vol. 10, no. 1, pp. 132–, 2018.

[35] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252–3261, 2018.

[36] Y. Zhu, J. Du, and X. Wu, "Adaptive period embedding for representing oriented objects in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–11, 2020.

[37] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: An efficient and accurate scene text detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5551–5560, July 2017.

[38] L. Tychsen Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6877–6885, June 2018.

[39] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *The European Conference on Computer Vision (ECCV)*, pp. 784–799, September 2018.

[40] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms – improving object detection with one line of code," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 5561–5569, Oct 2017.

[41] B. Zhu, Q. Song, L. Yang, Z. Wang, C. Liu, and M. Hu, "Cpm r-cnn: Calibrating point-guided misalignment in object detection," *arXiv:2003.03570*, 2020.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[44] G. Zhang, S. Lu, and W. Zhang, "Cad-net: A context-aware detection network for objects in remote sensing imagery," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 10015–10024, 2019.

[45] Y. Wang, Y. Zhang, Y. Zhang, L. Zhao, X. Sun, and Z. Guo, "Sard: Towards scale-aware rotated object detection in aerial imagery," *IEEE Access*, vol. 7, pp. 173855–173865, 2019.

[46] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, "Feature-attentioned object detection in remote sensing imagery," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3886–3890, 2019.
[47] J. Wang, J. Ding, H. Guo, W. Cheng, and W. Yang, "Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images," *Remote Sensing*, vol. 11, no. 24, p. 2930, 2019.
[48] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500, July 2017.
[49] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *Isprs Journal of Photogrammetry and Remote Sensing*, p. S0924271618301096, 2018.
[50] Dai, L. Jifeng, H. Yi, S. Kaiming, and Jian, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems 29*, pp. 379–387, 2016.
[51] P. Wang, X. Sun, W. Diao, and K. Fu, "Fmssd: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3377–3390, 2020.
[52] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.
[53] J. Yan, H. Wang, M. Yan, W. Diao, X. Sun, and H. Li, "Iou-adaptive deformable r-cnn: Make full use of iou for multi-class object detection in remote sensing imagery," *Remote Sensing*, vol. 11, no. 3, 2019.
[54] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based cnn for ship detection," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 900–904, 2017.
[55] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 11, pp. 1745–1749, 2018.
[56] L. Minghui, Z. Zhen, S. Baoguang, X. G. Song, and B. Xiang, "Rotation-sensitive regression for oriented scene text detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 15, pp. 1745–1749, 2018.
[57] L. Yang, Q. Song, Y. Wu, and M. Hu, "Attention inspiring receptive-fields network for learning invariant representations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1744–1755, 2019.
[58] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, June 2018.
[59] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, June 2018.
[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015.
[61] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, June 2016.

Qing Song received her Ph.D. degree from Tianjin University, Tianjin, China, in 2006.

She is currently a scientific researcher at the Beijing University of Posts and Telecommunications (BUPT), where she is engaged in the field of computer vision technology. She is the founder of Pattern Recognition and Intelligent Vision Laboratory (PRIV) and led the PRIV team to the championship of COCO2018 and COCO2019-DensePose Challenge, and won two championships in the CVPR LIP international competition in 2020. She is in charge of many national, provincial and ministerial projects and enterprise cooperation projects. She has published more than 70 academic papers in international journals and conferences.

Fan Yang received his B.S. degree in 2019 from Beijing University of Posts and Telecommunications, Beijing, China, where he is currently working toward the School of Automation.

His research interests include computer vision, image processing, deep learning.

Lu Yang is a doctoral student of automated institute from Beijing University of Posts and Telecommunications (BUPT), Beijing, China. And he received a bachelor's degree from BUPT at 2012. He is engaged in research work at Pattern Recognition and Intelligent Vision laboratory from 2012. His research interests cover the fields of artificial intelligence, computer vision, machine learning.

Chun Liu received her Ph.D. degree from University of Kassel, Germany, in 2014. She is now a lecturer at Beijing University of Posts and Telecommunications (BUPT). Her main research interests are primarily in Intelligent Computation and Optimization, especially evolutionary algorithms in solving optimization problems, e.g., planning and scheduling.

Mengjie Hu received the Ph.D. degree from Beihang University, Beijing, China, in 2017. She is now a lecturer at the Beijing University of Posts and Telecommunications. Her current research interests are primarily in computer vision and machine learning, especially object detection, visual tracking and visual geometry.

Lurui Xia received his Ph.D. degree from the National University of Defense Technology in 2010 and worked in the electronic science and technology postdoctoral mobile station of Chinese Academy of Sciences for two years. He is currently a scientific researcher in the Aerospace Engineering University, where he is mainly engaged in the research of space AI science and technology. He led the construction of the space AI technology innovation laboratory. He is in charge of more than 10 scientific research projects. He has obtained 7 invention patents and 10 software copyrights. He has published more than 20 academic papers.