OSKDet: Orientation-sensitive Keypoint Localization for Rotated Object Detection

Dongchen Lu, Dongmei Li, Yali Li, Shengjin Wang*
Department of Electronic Engineering, Tsinghua University
ludc19@mails.tsinghua.edu.cn, {lidmei, liyl13, wgsgj}@tsinghua.edu.cn

Abstract

Rotated object detection is a challenging issue in computer vision field. Inadequate rotated representation and the confusion of parametric regression have been the bottleneck for high performance rotated detection. In this paper, we propose an orientation-sensitive keypoint based rotated detector OSKDet. First, we adopt a set of keypoints to represent the target and predict the keypoint heatmap on ROI to get the rotated box. By proposing the orientation-sensitive heatmap, OSKDet could learn the shape and direction of rotated target implicitly and has stronger modeling capabilities for rotated representation, which improves the localization accuracy and acquires high quality detection results. Second, we explore a new unordered keypoint representation paradigm, which could avoid the confusion of keypoint regression caused by rule based ordering. Furthermore, we propose a localization quality uncertainty module to better predict the classification score by the distribution uncertainty of keypoints heatmap. Experimental results on several public benchmarks show the state-of-the-art performance of OSKDet. Specifically, we achieve an AP of 80.91% on DOTA, 89.98% on HRSC2016, 97.27% on UCAS-AOD, and a F-measure of 92.18% on ICDAR2015, 81.43% on ICDAR2017, respectively.

1. Introduction

With the success of deep convolutional neural networks (CNN), object detection has made an unprecedented breakthrough in recent years. General detection models [10, 16, 33] regress the horizontal bounding box of objects. However, real scenes objects may have arbitrary directions, such as cars in drone cameras and texts in streetscape. Only determining the horizontal box is not enough to locate the target accurately. Rotated object detection has a wide range of applications, but still faces great challenges.

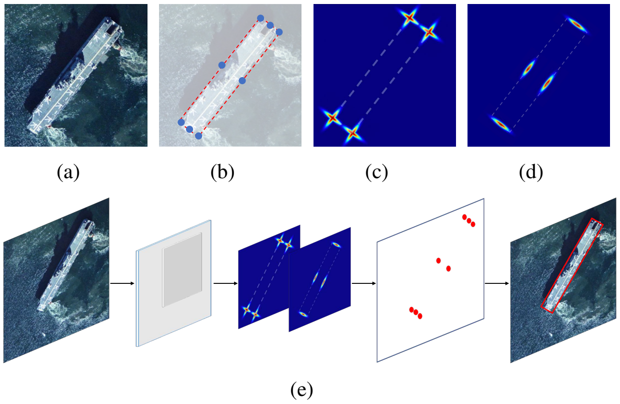


Figure 1. (a) a rotated target represented by 8 keypoints in (b). (c) and (d) display the proposed orientation-sensitive heatmap, we encode the keypoint to cross-star shape in corner and straight shape in edge areas, which represents the outline of target more accurately. (e) OSKDet: keypoint based detector. OSKDet generates the orientation-sensitive keypoint heatmap, which has a strong modeling ability of spatial representation.

Recently, CNN based rotated detection has made a considerable progress. Some works [2, 5, 30, 46, 49] use the angle definition (coordinates of the center point, width, height, and rotated angle) to represent the rotated target. Other works [8, 23, 27, 32, 41] use vertex definition (coordinates of four vertices) to describe the rotated quadrilateral. Mainstream models mainly face the following issues: i) Inadequate target representation. For angle definition, it is difficult for the model to learn the non intuitive angle and small deviations could lead to a sharp decline in IOU, while for the vertex difinition, existing heatmaps, like Gaussian heatmap, could not accurately represent the spatial characteristics of the rotated target. ii) Confusion of parametric order, which is caused by the parametric exchangeability and definition periodicity of rotated targets, and manual labeling errors exacerbate this problem. Similar samples with different optimization directions will cause learning confusion. In fact, these issues could largely affect detection performance.

This study was supported by the state key development program in 14th Five-Year under Grant No. 2021YFF0602103, 2021YFF0602102 and 2021QY1702. We also thank for the research fund under Grant No. 044007016.

*Corresponding author

In this paper, we propose an orientation-sensitive heatmap based rotated detector OSKDet. We encode a set of keypoints to represent rotated target. Considering that rotated target has more obvious features at the vertex and edge areas, we design an orientation-sensitive heatmap, as shown in Fig 1(c) and Fig 1(d), which could better match the target shape, and the model could learn the orientation and shape of rotated target implicitly. OSKDet has stronger modeling capabilities in spatial representation and transformation. Furthermore, we explore an unordered keypoint fusion representation, which could eliminate the keypoint order confusion problem to the greatest extent. We also propose a location quality uncertainty module to better improve the classification accuracy by the generated heatmaps. The proposed three modules can notably improve rotated detection accuracy. Extensive experiments on public benchmarks, including aerial dataset DOTA [40], HRSC2016 [28], UCAS-AOD [54], scene text dataset ICDAR2015 [15] and ICDAR2017MLT [31], show the superiority of OSKDet.

Our main contributions can be summarized as follows:

1) We propose a keypoint heatmap based rotated detector OSKDet. We design an orientation-sensitive heatmap to better represent the rotated target, which could learn the shape and direction of the rotated target, and plays a significant role in improving the localization accuracy.

2) We explore a novel unordered keypoint heatmap fusion method. This new representation could eliminate the learning confusion caused by rule based keypoint ordering.

3) We propose a localization quality uncertainty module, which effectively improves the classification score confidence through the feature fusion of keypoint localization distribution.

2. Related work

2.1. Horizontal object detection

Detection models with deep neural networks achieved superior performance on the public datasets COCO [22] and VOC [7] recently. According to whether there is a series of candidate anchor box, detection models can be divided into anchor based and anchor free methods. While according to the final localization mode, existing models have regression and heatmap methods. Most anchor based detectors localize targets by regression mode. Faster RCNN [36], FPN [21] etc. use fully connected layers to predict the deviation between anchor box and target. YOLO v2 [34] and YOLO v3 [35] adopt fully convolution structure and regress target center point ratio on each grid cell. Recently some anchor free methods detect targets mainly by generating heatmap of keypoints. CenterNet [6], CornerNet [16], ExtremeNet [53] etc. adopt gaussian heatmap to predict the probability of center or corner point in the whole image, and then group the points to form a box. Grid RCNN [29] predicts the keypoint heatmap on ROI and acquires a higher localization accuracy compared to Faster RCNN [36].

2.2. Rotated object detection

Similar to horizontal object detection, mainstream rotated detection algorithms can be divided into anchor based and anchor free models, in which anchor based models have horizontal and rotated anchor box. According to the definition of rotated target, there are two types of prediction mode: angle representation and vertex representation. Angle definition uses the rotated angle (θ) of the longer border around the horizontal axis as the fifth parameter, and forms the expression paradigm of rotated target (x, y, w, h, θ) together with other parameters. Vertex representation indicates rotated target by marking four vertices of the quadrilateral $(x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3)$. Compared with angle notation, vertex notation can represent any shape quadrilateral, and as mentioned in [32], the four vertices regression has natural consistency, which means it is easier to optimize.

Angle based detector. DRBox [25] and R2PN [49] introduce rotated anchor box based on rotated RPN. RRPN [30] proposes rotated ROI pooling to extract feature more effectively. R2CNN [14] regresses two adjacent vertices and another side length of rotated target. ROI-transformer [5] transforms horizontal proposals to rotated ones through fully connection learning in RPN stage. EAST [52] regresses the distance between each pixel and four sides of rotated box. SCRDet [46] highlights the target features through attention module, and proposes IOU smooth L1 loss to smooth the boundary loss jump. CSL [43] transforms angle prediction from regression to classification, and proposes circular label smooth. PIOU [2] adopts pixel counting method to calculate polygon IOU and makes it approximately derivable. GWD [44] converts the rotated rectangle into a two-dimensional Gaussian distribution and calculates the Wasserstein distance between GT and DT. ReDet [12] propose rotation-equivariant backbone and rotation-invariant ROI align to extract rotation-equivariant features.

Vertex based detector. Textboxes++ [20] proposes irregular long convolution kernel to adapt to targets with large aspect ratio. Gliding Vertex [41] regresses the ratio of the four vertices relative to the four points of horizontal box, and proposes an obliquity factor distinguishing nearly horizontal and other rotated objects. PolarDet [50] adopts several vertices angles and length ratio around the center point in polar coordinates to generate rotated box. SBD [27] proposes sequential-free box discretization parameterizing rotated boxes into key edges to eliminate the sensitive of label sequence. RSDet [32] proposes modulated loss by swapping point set sequence to obtain the best optimization direction. FR-Est [8] regresses 16 keypoints heatmap on the horizontal ROI. Specifically, they propose a new heatmap representation, which use 4 adjacent pixels to represent a point.

3. Proposed methods

Overview. In this section, we first present the main issue that hinders the rotated detection accuracy, and then introduce our method.

Rotated object representation. Mainstream rotated detectors adopt fully connected layers to regress angle or vertex coordinates. As [51] proposed the flattening operation will lose the spatial context information. While the convolution module has the ability to locate target and maintain this context information. Some detectors adopt convolution layers to generate keypoint heatmap and decode them to rotated box. While existing heatmaps, such as Gaussian heatmap, could not accurately represent the spatial layout of rotated target. Rotated target has more complex spatial diversity and transformation features, especially in the border areas. Accurate characterization and extraction of these local features is very important to improve the localization accuracy.

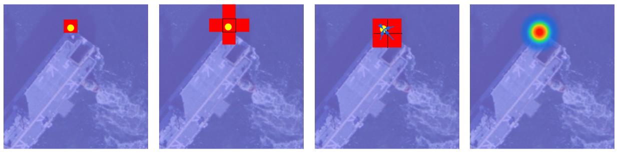
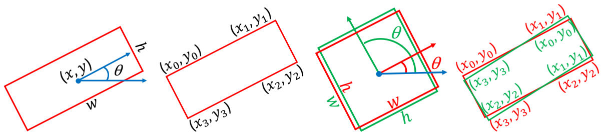

(a) SPH (b) CSH (c) FPH (d) SGH

Figure 2. Different heatmap representation. (a) single point heatmap in [13]. (b) cross-star heatmap in [29]. (c) 4-point heatmap in [8]. (d) standard gaussian heatmap in [16].

Confusion of parametric order. Both two definition methods have learning confusion problems, especially in the definition boundaries. For the angle representation, as shown in Fig 3(c), when the target is near a standard square, a little borderline length variety may lead to the exchange of width and height, and there is a $\pi/2$ jump in angle. Similar phenomena appear in the vertex representation, as illustrated in Fig 3(d), a slight spin may cause the vertex order transform. In the training process, two similar samples with quite different labels will lead to opposite optimization directions.

The proposed method is based on the horizontal anchor box. Similar to Grid RCNN [29] and FR-Est [8], we predict the keypoint heatmap on horizontal ROI. Specifically, we adopt 4 vertices and 4 edge midpoints to represent a complete rotated target. The network architecture is shown in Fig 5. For the feature extraction, we design a dilation convolution concat module (DCB) to adapt to the scale difference of rotated targets. We get fine-grained keypoint heatmap through twice deconvolution. Different from the common Gaussian heatmap, we design a new orientation-sensitive heatmap (OSH) to learn the direction. Our final output is the thermodynamic diagram with four points combined, so as to avoid the sorting of vertices. We also propose a location quality uncertainty module to further enhance the relationship between localization and score. We will introduce each module in detail next.



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

Figure 3. Different rotation representation: (a) angle notation. (b) vertex notation. (c) and (d) show parameters confusion. In (c), the red and green targets are very similar, while different length sequence causes width and height exchange, and their angle has $\pi/2$ gap. In (d), the red and green target has a slight rotation deviation, which leads to the vertex order one bit clockwise move.

The great scale-difference of rotated target has always been a difficult problem for detection. DCN [3] could effectively improve the feature extraction ability, but it's very time-consuming. We design a dilation convolution concat module (DCB), which acts on the P3-P7 layers of FPN. We extract the features through different expand rates conv kernel, then concat them together, and reduce the dimension through 1×1 conv. With little calculation increases, DCB could effectively improve the detection accuracy, especially for large-scale targets (e.g. harbor, ship).

$$F_{out} = conv(concat(\sum_{i=1}^{3} conv_i(F_{in}))) \tag{1}$$

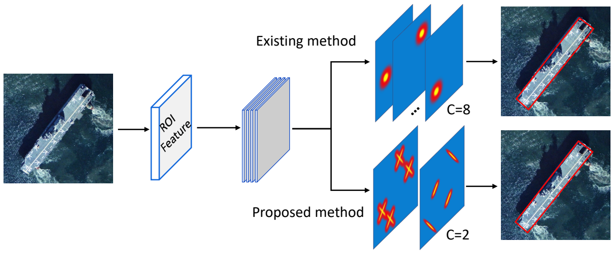


Figure 4. existing method: mainstream works regress the gaussian heatmap of point on each channel. proposed method: OSKDet predict the 4 points orientation-sensitive heatmap on one channel.

3.1. Orientation-sensitive heatmap

[8, 13, 16, 29] propose different heatmap generation methods. Gaussian heatmap adopted by [6, 16, 53] is most widely used. Different encoding methods affect the final detection accuracy greatly. In this section, we propose a novel orientation-sensitive heatmap (OSH), which is different from previous work. Next, we will introduce the generation mode of the new heatmap and the comparison with the standard gaussian heatmap (SGH).

For the p_0 point in Fig 6(a), compared with other regions, the intersection area of p_0p_1 and p_0p_3 have more obvious

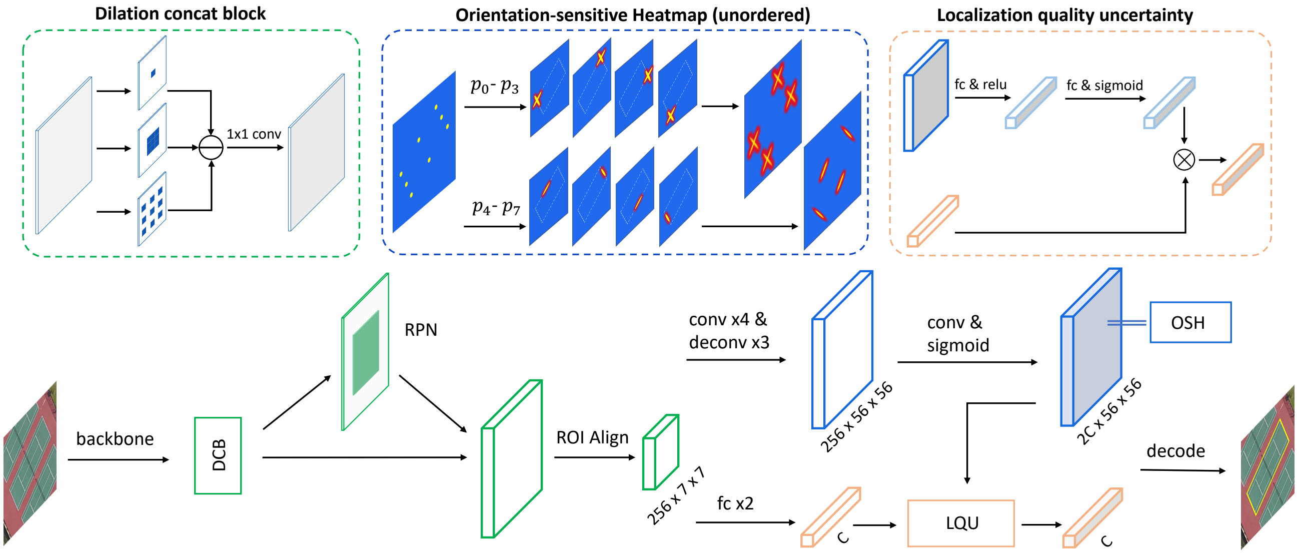
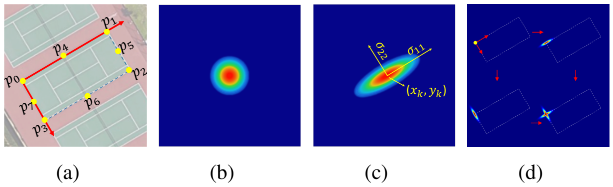


Figure 5. Architecture of the proposed OSKDet. OSKDet contains a dilation convolution concat module (DCM), an orientation-sensitive heatmap (OSH) with unordered fusion representation and a localization quality uncertainty module (LQU). The heatmap encoding process only exists in the training process.

color and texture transformation, and the edge between two vertices has similar feature. We believe that the junction zone information is more important for accurate detection, and hope that the network output has higher response values in these areas. The proposed OSH encodes different response values according to the importance of spatial region. In Fig 6(a), we establish new coordinate axis in the direction of p_0p_1 and p_0p_3 respectively, in which direction the Gaussian distribution has greater variance, as shown in Fig 6(c). OSH has stronger spatial representation and transformation ability.



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

Figure 6. (a) a set of 8 ordered points representing a rotated target, including 4 vertices p_{0-3} and 4 midpoints p_{4-7}. (b) and (c) are the comparison of SGH and proposed OSH. Compared to the SGH with same variance in x and y axis, the OSH assigns different weights according to the importance of the area, which more accurately reflects the contours of vertex and edge. (d) displays the generation of OSH for the vertex p_0.

After getting the proposal ROI from RPN, we map the GT keypoints to the ROI space. In order to include as many keypoints as possible, we adopt the ROI expansion idea of Grid RCNN [29], and expand the ROI from the center to $(1+r)$ times size (r is set to 0.25). Assuming that the original ROI region is (x, y, w, h), where (x, y) is the left-top point of the ROI, we obtain the expanded ROI (x', y', w', h') through Eq 2.

$$\begin{aligned} x' = x - \tfrac{r}{2}w, \quad & y' = y - \tfrac{r}{2}h \\ w' = (1+r)w, \quad & h' = (1+r)h \end{aligned} \tag{2}$$

Define the ground truth keypoints $P = \{p_k\}_{k\in[0,K-1]}$, $p_k = (\hat{x_k}, \hat{y_k})$ and mapped keypoints $Q = \{q_k\}_{k\in[0,K-1]}$, $q_k = (x_k, y_k, \theta_k)$. we adopt a total of $K = 8$ keypoints. After three deconvolution, we will get a heatmap of $M \times M$ size (default $M = 56$). Through Eq 3, we calculate the mapped coordinates on the heatmap. For the mapped points beyond ROI boundary, the generated heatmap is all 0.

$$x_k = (\hat{x_k} - x')\frac{M}{w'}, \quad y_k = (\hat{y_k} - y')\frac{M}{h'} \tag{3}$$

For the vertex q_{0-3}, we calculate the rotated angle θ in the direction of two adjacent sides (e.g. for q_0, we calculate the angle of q_0q_1 and q_0q_3 directions).

$$\theta_k = \left[\mathrm{argtan}\left(\tfrac{y_{(k+1)\%4} - y_k}{x_{(k+1)\%4} - x_k}\right), \mathrm{argtan}\left(\tfrac{y_{(k-1)\%4} - y_k}{x_{(k-1)\%4} - x_k}\right)\right] \tag{4}$$

For the midpoint q_{4-7}, we only calculate the rotated angle formed by the vertices on both sides (e.g. for q_4, we calculate the angle of q_0q_1 direction).

$$\theta_k = \mathrm{argtan}\left(\frac{y_{(k-3)\%4} - y_{k-4}}{x_{(k-3)\%4} - x_{k-4}}\right) \tag{5}$$

Suppose the mean and covariance matrix of standard gaussian distribution be $\boldsymbol{\mu_k} = (x_k, y_k)^T$, $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$. The rotation matrix of cartesian coordinate about the origin is $\boldsymbol{R_k} = \begin{bmatrix} \cos\theta_k & -\sin\theta_k \\ \sin\theta_k & \cos\theta_k \end{bmatrix}$. The covariance matrix after rotation transformation is

$$\boldsymbol{\Sigma_k} = \boldsymbol{R_k^T} \boldsymbol{\Sigma} \boldsymbol{R_k} \tag{6}$$

The final OSH is generated by Eq 7. For the vertex q_{0-3}, we generate heatmap in two directions. Fig 6(d) displays the OSH generation process.

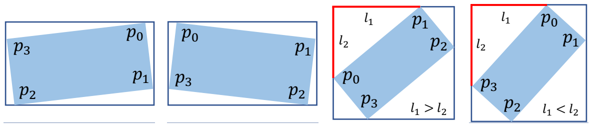
$$\boldsymbol{H(g)_k} = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma_k}|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{g}-\boldsymbol{\mu_k}\right)^T \boldsymbol{\Sigma_k^{-1}}\left(\boldsymbol{g}-\boldsymbol{\mu_k}\right)\right] \tag{7}$$

We adopt the binary cross entropy loss function to optimize the regression error of the heatmap, as illustrated by Eq 8. The $\delta(\cdot)$ is sigmoid function.

$$\begin{aligned} Loss = \tfrac{1}{K\times M\times M} \sum_{k=0}^{K-1} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} -(\hat{h_{kij}}) \log\left(\delta\left(h_{kij}\right)\right) \\ -(1-\hat{h_{kij}}) \log\left(1-\delta\left(h_{kij}\right)\right) \end{aligned} \tag{8}$$

3.2. Unordered keypoints representation

The root of keypoint learning confusion lies in the point order transform on the sorting interface. Most works adopt an angle to sort keypoints, e.g. in Gliding Vertex [41], they define the vertex with the smallest ordinate as the first point, which means they choose 0°-sorting as the cut-off point. The confusion near 0° is largest, as a slight angle spin around 0° will cause the point sets order jump, as shown in Fig 7(a) and Fig 7(b). Some works select the point closest to the top-left point of the smallest horizontal box as the first point, that is, 45°-sorting. However, when the target angle is around 45°, the first point will also jump, as shown in Fig 7(c) and Fig 7(d). In the vicinity of these areas, the mutations of GT labels will lead to the confusion of model learning and the training is hard to converge.



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

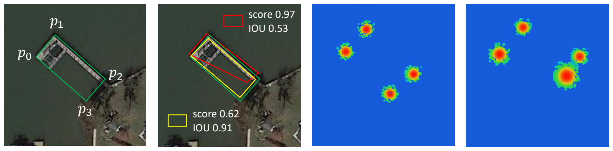
Figure 7. (a) and (b) are two similar samples close to 0°, (c) and (d) are also two similar samples around 45°. Their angle flips at the sorting boundary, causing the jump of keypoint order.

As any rule-based sorting can lead to confusion, the basic solution is to drop the sorting. We design an unordered keypoint heatmap representation. After generating the heatmap of each point, we combine them into one heatmap by taking the maximum value, as illustrated by Eq 9. That is, we generate a heatmap of 4 points on one feature map, which can effectively solve the confusion caused by sorting. During the inference process, the output heatmap can be decoded into up to 4 points per channel. We get each peak value by 3×3 max pooling.

$$H = Maximum(H_i, H_{i+1}, H_{i+2}, H_{i+3}), i \in \{0, 4\} \tag{9}$$

3.3. Localization quality uncertainty

In most detectors, the localization regression and classification score are usually trained independently, which results in the classification score could not directly reflect the quality of the regressed box. In the inference stage, a low quality box with high score may rank in the front, which reflects the gap between classification and regression. Inspired by GFL v2 [18], we adopt a classification and IOU joint representation score to replace the classification score as the final output.



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

Figure 8. (a) Compared with points p_0-p_2, there is no obvious feature of point p_3 in the harbour sample. (b) In the NMS stage, the high score with low quality box may kill the other high quality boxes. (c) and (d) show the heatmap distribution of yellow box and red box in (b). Obviously, the variance of p_3 in (d) is larger than other 3 points, which reflects the ambiguity of p_3 localization.

In order to further strengthen the relationship between localization and classification, we propose a localization quality uncertainty module (LQU). Similar to GFL v2 [18], the output heatmap is also essentially a probability distribution, representing the confidence at each point. As shown in Fig 8, the harbor has no distinct features at point p_3. In Fig 8(d), compared with the other 3 points, the variance of p_3 heatmap is greater and blurred, which means the localization is not clear enough. We hope to model this uncertainty and guide the classification score by the localization quality. LQU transforms the output heatmap to a C-dimensional vector by 2 fully connected layers. After the sigmoid function activation, we multiply the C-dimensional vector with the original classification score as the final score. LQU could significantly improve the detection accuracy with little calculation.

4. Experiments

4.1. Dataset and implementation details

To compare the performance of OSKDet with the state-of-the-art methods, we conduct evaluations on several public datasets, including aerial detection dataset DOTA [40], HRSC2016 [28] and UCAS-AOD [54], text detection dataset ICDAR2015 [15] and ICDAR2017MLT [31].

DOTA [40] contains 2806 aerial images, ranging in size from 800 × 800 to 4000 × 4000. Train, validation and test set are split to 3:1:2. Same as other models, we use train set and validation set for training. We adopt the same processing strategy as [5, 41], which cropped the image to 1024 × 1024.

HRSC2016 [28] is a high resolution ship recognition dataset which contains 1061 images, and the image size ranges from 300 × 300 to 1500 × 900.

UCAS-AOD [54] contains 1510 aerial images of about 659 × 1280 pixels, with 2 categories (car and plane) of 14596 instances.

ICDAR2015 [15] and **ICDAR2017MLT** [31] are two incidental scene Text dataset. Both of them contain natural scene text images with location annotations. IC15 has 1500 images and IC17 has 18000 images.

OSKDet is implemented on pytorch. We use 3 RTX 2080ti GPU, 2 images per GPU in our experiment. We adopt the FPN [21] based ResNet101 as our backbone and train 12 epochs by using momentum gradient descent optimization. The weight decay is set to 0.0001 and momentum is set to 0.9. The initial learning rate was 0.0075 and divided by 10 at (8, 11) epochs. For aerial set, we resize the image to 1×, 0.5× and 1.5× scales and use random rotation (0°, 90°, 270°). For text set, we resize the longer side to {800, 1000, 1200}. We also adopt class balancing for DOTA.

4.2. Aerial detection

DOTA. Tab 4 shows our testing results on the DOTA v1.0 OBB task. Compared with the state-of-the-art methods, OSKDet shows superior performance. With ResNet101-FPN as backbone, OSKDet achieves 76.37% AP in single scale and 80.91% AP in multi-scale, which outperforms all previous works. Compared with the angle based detectors [12, 43, 45], OSKDet improves the state-of-the-art method by 0.81% AP. Compared with other vertex based detectors [32, 41, 50], OSKDet improves the state-of-the-art method by 4.27% AP. Furthermore, OSKDet has achieved the best performance in multiple categories, especially in irregularly shaped categories such as baseball diamond and swimming pool, which illustrates the great advantages of OSKDet in spatial representation and transformation. Through accurate characterization and feature extraction of keypoints, OSKDet is more effective in detecting complex and irregular shape targets.

HRSC2016 and UCAS-AOD. Tab 1 depicts the comparison of different methods on HRSC2016 [28] and UCAS-AOD [54]. OSKDet achieves 89.98% AP and 97.27% AP respectively in the two challenging datasets and outperforms all other methods, which demonstrates the proposed OSKDet brings consistent gain on different aerial datasets.

Ablation study. We perform all ablation experiments on DOTA dataset. Unless otherwise specified, all test benchmarks are VOC2007 [7] AP.5 IOU metric. Tab 2 shows the comprehensive ablation experiment results.

Method	AP
RoI-Trans [5]	86.20
RSDet [32]	86.50
Gliding Vertex [41]	88.20
BBAVectors [48]	88.60
CSL [43]	89.62
OSKDet(ours)	**89.98**

Method	Plane	Car	AP
S2ARN [1]	97.60	92.20	94.90
FADet [17]	98.69	92.72	95.71
R3Det [42]	98.20	94.14	96.17
SCRDet++ [45]	98.93	94.97	96.95
PolarDet [50]	**99.08**	94.96	97.02
OSKDet(ours)	99.06	**95.52**	**97.27**

Table 1. Comparison with state-of-the-art methods on HRSC2016 (left) and UCAS-AOD (right).

Method	MS	DCB	OSH	UKR	LQU	AP.5	FPS
OSKDet						73.38	-
		✓				73.96	10.07
		✓	✓			75.34	-
		✓		✓		74.47	-
		✓		✓	✓	75.10	9.86
		✓	✓	✓	✓	76.37	-
	✓	✓	✓	✓	✓	**80.91**	-

Table 2. Ablation experiment of different strategies on DOTA. MS, DCB, OSH, UKR and LQU mean multi-scale training and testing, dilation concat block orientation-sensitive heatmap, unordered keypoint representation, and localization quality uncertainty module, respectively.

Orientation-sensitive heatmap. We compare the proposed method with different heatmap formats, including, single point heatmap [13] (SPH), cross-star heatmap [29] (CSH), 4 point heatmap [8] (FPH) standard gaussian heatmap [16] (SGH) and orientation-sensitive heatmap (OSH). We implement the four different format heatmaps on OSKDet. Tab 3 displays the result. The OSH surpasses all baselines with an increment of 0.89%-1.76%. Compared with the SGH, the OSH improves 0.89% and 1.16% AP without/ with FPN respectively, which proves the stronger spatial modeling capabilities of OSH will bring localization accuracy gains.

Method	FPN	AP.5
SGH [16]		69.62
OSH(Ours)		70.51(+0.89)
SPH [13]	✓	72.86
CSH [29]	✓	73.15
FPH [29]	✓	73.04
SGH [16]	✓	73.46
OSH(Ours)	✓	**74.62(+1.16)**

Table 3. Comparison of different prediction format results

To further verify the effect of OSH and find best parameters, we try different gaussian kernel ratio and kernel size combination experiments. As shown in Fig 9, σ represents the standard deviation in the vertical direction σ_{22}, and r represents the scale factor between the axis direction and the vertical direction, that is, $\sigma_{11} = r * \sigma_{22}$ ($r = 1$ means SGH). We get the best effect when $r = 3$ and $\sigma = 0.8$. The result in Fig 9 shows that r has a greater influence on the results than σ. The direction and shape of the Gaussian kernel are more important than the size of the Gaussian kernel, which is why OSH is effective.

Since the DOTA [40] official evaluation system only provides the test results for AP.5 metric. In order to test the performance of the model under different IOU metrics, we

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP.5
FR-O [40]	ResNet101	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.6	58.96	49.40	52.52	46.69	44.80	46.30	52.93
IENet [23]	ResNet101	80.20	64.54	39.82	32.07	49.71	65.01	52.58	81.45	44.66	78.51	46.54	56.73	64.40	64.24	36.75	57.14
R2CNN [14]	ResNet101	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 [30]	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
RADet [19]	ResNext101	79.45	76.99	48.05	65.83	65.46	74.4	68.86	89.7	78.14	74.97	49.92	64.63	66.14	71.58	62.16	69.09
ROI-Trans [5]	ResNet101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet [46]	ResNet101	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
FADet [17]	ResNet101	90.21	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.4	84.68	53.4	65.42	74.17	69.69	64.86	73.28
RSDet [32]	ResNet152	90.10	82.00	53.80	68.50	70.20	78.70	73.60	91.20	87.10	84.70	64.30	68.20	66.10	69.30	63.70	74.10
FR-Est [8]	ResNet101	89.63	81.17	50.44	70.19	73.52	77.98	86.44	90.82	84.13	83.56	60.64	66.59	70.59	66.72	60.55	74.20
Gliding Vertex [41]	ResNet101	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 [37]	ResNext101	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 [9]	ResNet101	90.10	82.70	54.20	75.20	71.00	79.90	83.50	90.70	83.90	84.60	61.20	68.00	70.70	76.00	63.70	75.70
APE [55]	ResNext101	89.96	83.62	53.42	76.03	74.01	77.16	79.45	90.83	87.15	84.51	67.7	60.33	74.61	71.84	65.55	75.75
CenterMap [38]	ResNet101	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
CSL [43]	ResNet152	**90.25**	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.66	68.04	73.83	71.10	68.93	76.17
PolarDet [50]	ResNet101	89.65	87.07	48.14	70.97	78.53	80.34	87.45	90.76	85.63	86.87	61.64	70.32	71.92	73.09	67.15	76.64
SCRDet++ [45]	ResNet101	90.05	84.39	55.44	73.99	77.54	71.11	86.05	90.67	87.32	87.08	69.62	68.90	73.74	71.29	65.08	76.81
GWD [44]	ResNet152	89.28	83.70	59.26	79.85	76.42	83.87	86.53	89.06	85.53	86.50	**73.04**	67.56	76.92	77.09	71.58	79.08
S2aNet [11]	ResNet101	88.89	83.60	57.74	**81.95**	**79.94**	83.19	**89.11**	90.78	84.87	**87.81**	70.30	68.25	78.30	77.01	69.58	79.42
ReDet [12]	ReResNet50	88.81	82.48	60.83	80.82	78.34	**86.06**	88.31	**90.87**	88.77	87.03	68.65	66.90	**79.26**	79.71	**74.67**	80.10
OSKDet (Ours)	ResNet101	90.07	**87.08**	54.16	75.61	72.64	76.86	87.63	90.77	79.10	86.88	59.88	**71.28**	75.16	71.71	66.67	76.37
OSKDet-MS (Ours)	ResNet101	90.03	86.94	**61.24**	81.48	79.63	85.72	88.52	90.84	**89.26**	87.55	68.38	71.24	78.89	**79.95**	73.97	**80.91**

Table 4. Comparison of different method results on DOTA-v1.0 OBB task (MS means multi-scale training and testing)

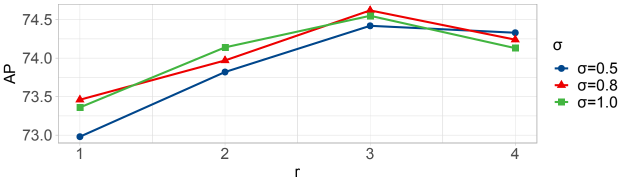


Figure 9. Comparison of different σ and r results on OSH. The two parameters represent the direction of rotated object. More accurate characterization will improve the localization accuracy, which demonstrates the superiority of OSH.

use the training set to train and validation set to evaluate. We reimplement [41] and other format heatmap in OSKDet. OSH achieves 42.67% mAP, which excels other format predictions with considerable performance gains, specifically 2.53%, 2.23%, 2.14% and 1.87% mAP for SPH, CSH, FPH and SGH. As illustrated in Fig 10, OSH surpasses all other methods, especially under the high IOU metrics. Under AP.9 IOU threshold, OSH surpasses other methods by 4.11%-6.10%, which demonstrates that OSKDet has a huge advantage in obtaining high quality detections.

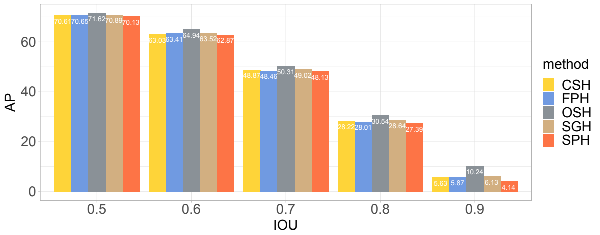
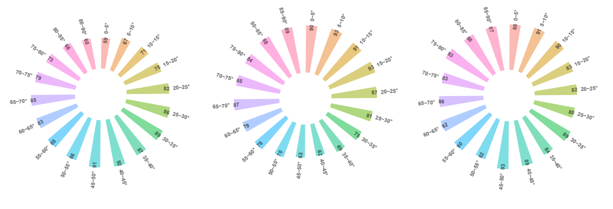


Figure 10. Different prediction format results under AP.5-.95

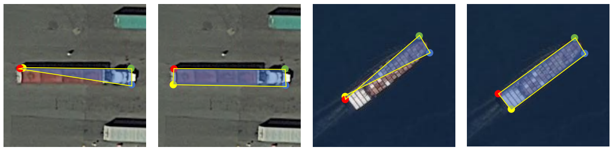
Unordered keypoint representation. We compare different keypoint sorting methods, including 0°-sorting, 45°-sorting, and unordered representation. To further explore the impact of each sorting methods on fine-grained detection, we classify 0-90° into 18 categories using 5° as an interval and calculate the precision in each angle interval. Experiment results in Fig 11 show that our unordered representation method performs better than other methods. In Fig 11, the precision of angle-based sorting decreases dramatically near the cut-off point, while our method maintains high accuracy in all angle range. The unordered representation effectively eliminates the confusion caused by keypoints sorting, which improves the average precision by 3.63%-5.48%.



(a) 0° sorting (b) 45° sorting (c) unordered

Figure 11. recision of differnet sorting methods on validation set. In (a) and (b), the precision near cut-off point drops greatly.

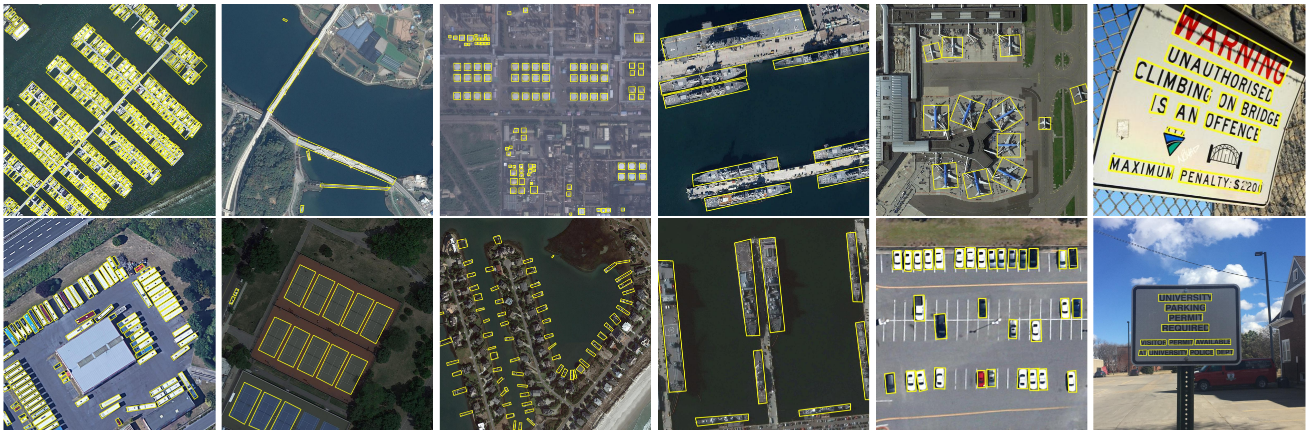
Fig 12 shows two cases of keypoints order confusion. In previous work, the keypoints order of similar samples may be different in the training stage, which may cause different channel heatmaps predict the same point in the inference phase, like the red and yellow point in Fig 12(a) and Fig 12(c). An unordered representation by heatmap fusion will eliminate this issue greatly.



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

Figure 12. (a) and (c): prediction of 0°-sorting and 45°-sorting. (b) and (d): prediction of unordered representation.

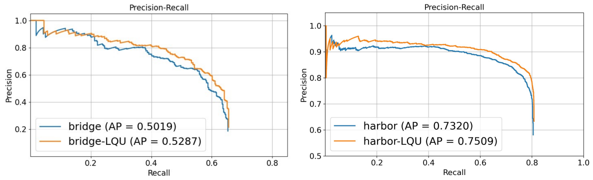
Localization quality uncertainty. LQU module effectively improves the detection accuracy, we improve 0.63%


WARNING
UNAUTHORISED CLIMBING ON BRIDGE IS AN OFFENCE
MAXIMUM PENALTY: $2200
UNIVERSITY PARKING PERMIT REQUIRED
VISITOR PERMIT AVAILABLE AT UNIVERSITY POLICE DEPT

(a) (b) (c) (d) (e) (f)

Figure 13. Visualization of OSKDet detections. (a)(b)(c): DOTA (d): HRSC2016 (e): UCAS-AOD (f): ICDAR.

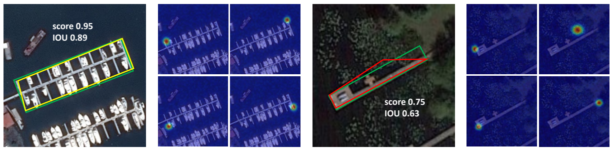
.5AP metric on the DOTA test set. We selected two types of objects, bridge and harbor, which have the characteristics of large aspect ratio and insignificant keypoint features, to verify the effectiveness of LQU. As illustrated in Fig 14, the accuracy curve of LQU is always above the baseline model. LQU improves the .5AP metric by 2.68% and 1.89% for two categories, respectively.



(a) AP-bridge (b) AP-harbor

Figure 14. The effects of LQU on validation set.

Fig 15 shows the relationship of LQU score and heatmap distribution. For points with unclear location, the heatmap usually has a large Gaussian kernel and divergent shape, like the top-left point in Fig 15(c) and Fig 15(d). The predicted score of LQU also faithfully reflects the true IOU.



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

Figure 15. (a): a high score box with high quality IOU, and in (b), the heatmap of 4 points are all clear and convergent. (c): a lower score box with low quality IOU, and in (d), the top-left heatmap distribution is divergent and ambiguous. In (a) and (c), green box is Ground truth, whild red and yellow boxes are detection results.

4.3. Text detection

The text dataset has a greater challenge due to a large number of irregular quadrilaterals and large aspect ratio long text. Tab 5 shows our detection results on ICDAR2015 [15] and ICDAR2017MLT [31]. Without other techniques designed for text targets, OSKDet achieves 92.18% F-measure on IC15 and 81.43% F-measure on IC17 with 8.74 FPS, which surpasses all of other state-of-the-art methods. Compared with other algorithms, our model has great advantages in detection accuracy and speed. OSKDet could easily extend to other datasets.

Task	Method	Recall	Precision	F-measure	FPS
ICDAR2015	EAST [52]	73.50	83.60	78.20	6.52
	PixelLink [4]	82.00	85.50	83.70	7.30
	PSENet [39]	85.22	89.30	87.21	2.33
	FOTS [26]	85.17	91.00	87.99	7.50
	Textfusenet [47]	89.70	94.70	92.10	4.10
	OSKDet(ours)	89.35	95.21	**9**2.18	8.74
ICDAR2017	PSENet [39]	68.35	76.97	72.40	-
	SBD [27]	70.10	83.60	76.30	-
	PMTD [24]	76.25	84.42	80.13	-
	OSKDet(ours)	76.02	87.66	**81.43**	8.74

Table 5. Comparison of different methods on ICDAR

5. Conclusion

This paper proposes a rotated object detection model OSKDet. By the proposed orientation-sensitive heatmap with unordered keypoint representation, OSKDet can extract spatial feature effectively and eliminate the confusion of keypoint order greatly. The proposed localization quality uncertainty module further improves the detection accuracy. Experimental results on several public datasets show the state-of-the-art performance of OSKDet.

References

[1] Songze Bao, Xing Zhong, Ruifei Zhu, Xiaonan Zhang, Zhuqiang Li, and Mengyang Li. Single shot anchor refinement network for oriented object detection in optical remote sensing imagery. *IEEE Access*, 7:87150–87161, 2019. 6

[2] Zhiming Chen, Kean Chen, Weiyao Lin, John See, Hui Yu, Yan Ke, and Cong Yang. Piou loss: Towards accurate oriented object detection in complex environments. In *ECCV (5)*, pages 195–211, 2020. 1, 2

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017. 3

[4] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *AAAI*, pages 6773–6780, 2018. 8

[5] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2849–2858, 2019. 1, 2, 6, 7

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6569–6578, 2019. 2, 3

[7] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 2, 6

[8] Kun Fu, Zhonghan Chang, Yue Zhang, and Xian Sun. Point-based estimator for arbitrary-oriented object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99), 2020. 1, 2, 3, 6, 7

[9] Kun Fu, Zhonghan Chang, Yue Zhang, Guangluan Xu, Keshu Zhang, and Xian Sun. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *Isprs Journal of Photogrammetry and Remote Sensing*, 161:294–308, 2020. 7

[10] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 1

[11] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 7

[12] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 2, 6, 7

[13] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2020. 3, 6

[14] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R 2 cnn: Rotational region cnn for arbitrarily-oriented scene text detection. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3610–3615, 2018. 2, 7

[15] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160, 2015. 2, 5, 6, 8

[16] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *International Journal of Computer Vision*, 128(3):642–656, 2020. 1, 2, 3, 6

[17] Chengzheng Li, Chunyan Xu, Zhen Cui, Dan Wang, Tong Zhang, and Jian Yang. Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3886–3890, 2019. 6, 7

[18] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11641, 2021. 5

[19] Yangyang Li, Qin Huang, Xuan Pei, Licheng Jiao, and Ronghua Shang. Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing*, 12(3):389, 2020. 7

[20] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27(8):3676–3690, 2018. 2

[21] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 2, 6

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 2

[23] Youtian Lin, Pengming Feng, and Jian Guan. Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection. *arXiv preprint arXiv:1912.00969*, 2019. 1, 7

[24] Jingchao Liu, Xuebo Liu, Jie Sheng, Ding Liang, Xin Li, and Qingjie Liu. Pyramid mask text detector. *arXiv preprint arXiv:1903.11800*, 2019. 8

[25] Lei Liu, Zongxu Pan, and Bin Lei. Learning a rotation invariant detector with rotatable bounding box. *arXiv preprint arXiv:1711.09405*, 2017. 2

[26] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5676–5685, 2018. 8

[27] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3052–3058, 2019. 1, 2, 8

[28] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *6th International Conference on Pattern Recognition Applications and Methods*, pages 324–331, 2017. 2, 5, 6

[29] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *2019 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 7363–7372, 2019. 2, 3, 4, 6

[30] 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*, 20(11):3111–3122, 2018. 1, 2, 7

[31] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, Wafa Khlif, Muhammad Muzzamil Luqman, Jean-Christophe Burie, Cheng lin Liu, and Jean-Marc Ogier. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 1454–1459, 2017. 2, 5, 6, 8

[32] Wen Qian, Xue Yang, Silong Peng, Yue Guo, and Chijun Yan. Learning modulated loss for rotated object detection. *arXiv preprint arXiv:1911.08299*, 2019. 1, 2, 6, 7

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 1

[34] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. 2

[35] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2

[37] Jinwang Wang, Jian Ding, Haowen Guo, Wensheng Cheng, Ting Pan, and Wen Yang. Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing*, 11(24):2930, 2019. 7

[38] Jinwang Wang, Wen Yang, Heng-Chao Li, Haijian Zhang, and Gui-Song Xia. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–17, 2020. 7

[39] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9336–9345, 2019. 8

[40] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 2, 5, 6, 7

[41] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 5, 6, 7

[42] 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. 6

[43] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *ECCV (8)*, pages 677–694, 2020. 2, 6, 7

[44] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. *arXiv preprint arXiv:2101.11952*, 2021. 2, 7

[45] Xue Yang, Junchi Yan, Xiaokang Yang, Jin Tang, Wenlong Liao, and Tao He. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*, 2020. 6, 7

[46] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8232–8241, 2019. 1, 2, 7

[47] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du. Textfusenet: Scene text detection with richer fused features. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 1, pages 516–522, 2020. 8

[48] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris N. Metaxas. Oriented object detection in aerial images with box boundary-aware vectors. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2150–2159, 2020. 6

[49] 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*, 15(11):1745–1749, 2018. 1, 2

[50] Pengbo Zhao, Zhenshen Qu, Yingjia Bu, Wenming Tan, Ye Ren, and Shiliang Pu. Polardet: A fast, more precise detector for rotated target in aerial images. *arXiv preprint arXiv:2010.08720*, 2020. 2, 6, 7

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 3

[52] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2017. 2, 8

[53] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 850–859, 2019. 2, 3

[54] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3735–3739, 2015. 2, 5, 6

[55] Yixing Zhu, Jun Du, and Xueqing Wu. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7247–7257, 2020. 7