CGP Box: An effective direction representation strategy for oriented object detection in remote sensing images

Qiuyu Guan, Zhenshen Qu, Ming Zeng, Jianxiong Shen & Jingda Du

To cite this article: Qiuyu Guan, Zhenshen Qu, Ming Zeng, Jianxiong Shen & Jingda Du (2021) CGP Box: An effective direction representation strategy for oriented object detection in remote sensing images, International Journal of Remote Sensing, 42:17, 6666-6687, DOI: 10.1080/01431161.2021.1941389

To link to this article: https://doi.org/10.1080/01431161.2021.1941389

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 27 Jul 2021.

Submit your article to this journal

Article views: 463

View related articles

View Crossmark data
CGP Box: An effective direction representation strategy for oriented object detection in remote sensing images

Qiuyu Guan\textsuperscript{a}, Zhenshen Qu\textsuperscript{a}, Ming Zeng\textsuperscript{a}, Jianxiong Shen\textsuperscript{b} and Jingda Du\textsuperscript{a}

\textsuperscript{a}Space Control and Inertial Technology Research Center, Harbin Institute of Technology, Harbin, China; \textsuperscript{b}Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain

\begin{abstract}
In recent years, the emergence of convolutional neural networks (CNN) has greatly promoted the development of the object detection field, and many CNN-based detectors have achieved excellent performance on object detection in remote sensing images. To accurately locate the target, oriented bounding box (OBB) is usually used in remote sensing objects, such as the angle-based OBB, to represent the target. Nevertheless, the critical loss instability caused by the periodicity of the angle is always difficult to solve. In this paper, we propose a novel strategy called the Center-Guide points (CGP) box method that uses the guide points to locate the target, which breaks the limit of the angle-based thinking pattern to solve the critical loss instability problem. To be specific, we define a new guide-points selection rule and prediction structure, which replaces the traditional method of using angle values to indicate the direction. Furthermore, we propose the matching method of centre points and guide points, which is a box decoding method that matches the object and the corresponding guide points. Finally, an attention learning module called the Gaussian Center-Line (GC-L) Attention module based on the Gaussian centre-line is proposed to improve the accuracy of guide points. These strategies are applied to the key point detection framework and tested on three classical-oriented object remote sensing datasets. The results show that our method is effective and competitive.
\end{abstract}

1. Introduction

Object detection is a classical branch in the field of computer vision (Lin et al. 2017a; Ren et al. 2015; Dai et al. 2016; Liu et al. 2016a; Zhou, Wang, and Philipp 2019). Its goal is to analyse the input image and output the location information and classification information of the specified object category. In recent years, due to the wide application of deep learning technology, this field has made significant progress. However, compared with natural scenes, object detection in remote sensing images is still a challenging work. Objects in remote sensing images tend to be variable in orientation, have a large aspect ratio, and sometimes are closely arranged (vehicles in the parking lot, ships in the harbour). Therefore, using Horizontal Bounding Box (HBB) as labels of objects has great disadvantages. As shown in Figure 1(a), a HBB that accurately surrounds the object will...
cover multiple tightly arranged ships. Such object positioning ambiguity will cause the network to be unable to accurately distinguish the boundaries of adjacent objects in the learning process. Usually, the Intersection over Union (IoU) is used in the field of object detection to evaluate the degree of overlap of two bounding boxes, that is, the ratio of the area of intersection to the area of union. Moreover, the detection boxes with large IoU in the same category will be suppressed by the Non-Maximum Suppression (Neubeck and Luc 2006) algorithm, and only the HBB with the highest score will be left in the end. Oriented Bounding Box (OBB) (Figure 1(b)) eliminates the above defects and is more in line with the requirements of remote sensing object detection. Therefore, accurate prediction of OBB direction is an important part of this field.

A branch of (Cheng, Zhou, and Han 2016; Deng et al. 2017; Long et al. 2017) directly predicts the angle value to represent OBB, which is usually expressed as \((x, y, w, h, \theta)\), we call it \(\theta\)-based OBB, where \((x, y)\) is the centre point coordinates of the bounding box, \(w\) and \(h\) are its width and height, respectively, \(\theta\) is the inclination angle. This type of detector performs extremely well in OBB detection and achieves top-notch results. However, the angle value depicted by the radian system will cause ambiguity at the critical point of the period due to the periodicity, resulting in severe fluctuations of loss value and affecting convergence. For example, as shown in Figure 2, if the angle period is \([-\pi/2, \pi/2]\), the loss of the two prediction boxes (Predicted Box 1 and 2) near the critical point (\(\pi/2\)) of the period is very different from the truth boxes \((\Delta\theta_1, \Delta\theta_2)\) in the figure). Therefore, in order to eliminate the ambiguity of object box representation, we propose a novel OBB representation method called Center-Guide points Box (CGP Box). Unlike other OBB representations, CGP Box uses guide points to determine the direction of the object, which avoids periodic ambiguity and can converge to the optimal solution more stably.

Inspired by the general object detector, most of the work(Xia et al. 2018; Lin et al. 2017; Ding et al. 2019) in the field of remote sensing object detection is based on the R-CNN framework, which is a kind of classic anchor-based network. The anchor-based detector uses the anchor to obtain the region proposal, and then makes the secondary correction on the preselected region to obtain accurate results. In general, the acquisition of anchor requires sufficient prior knowledge. Many works have proposed oriented anchor to apply

Figure 1. Examples of different label method in HRSC2016 dataset.
R-CNN framework to the field of oriented object detection in remote sensing images. However, the introduction of direction information also leads to a larger number of required anchors and more prior knowledge, which ultimately makes network prediction inefficient. Besides, when the network predicts multiple corner points, the output is usually one-dimensional, which will lead to severe fluctuations of loss. For example, the network outputs three bounding boxes \((E_1, E_2, E_3, E_4), (E_1, E_3, E_2, E_4),\) and \((E_4, E_3, E_2, E_1),\) where \(E_i = (x_i, y_i)\) is the coordinate of the corner points. Although they all refer to the same bounding box, the calculated losses are completely different due to different orders. Therefore, we adopt the anchor-free structure as the basic framework of CGP Box, which is a full convolutional network structure with multiple probabilistic heatmaps as the output. Due to the uniqueness of the guide points and centre points in the heatmap, loss fluctuation is avoided.

In this paper, we proposed an anchor-free network that integrates the Center-Guide points box regression strategy (mentioned in Section 3.2). Besides, to more accurately locate the position of the guide points, we also proposed a supervised central attention learning module called the Gaussian Center-Line (GC-L) Attention Module (mentioned in Section 3.4). Due to the lack of computing resources, we chose to verify the validity of each part on the HRSC2016 (Liu et al. 2016b) dataset and finally compared the test set results on the DOTA (Xia et al. 2018) dataset and UCAS-AOD (Zhu et al. 2015) dataset. In summary, the contributions of this paper are as follows:

![Figure 2. Due to the periodicity of angle, the loss values calculated by the prediction box with the same estimation error are greatly different.](image)
(1) We proposed an OBB representation strategy based on Center-Guide points called CGP Box and merged it with the anchor-free network to solve the hidden ambiguities in other representation methods. Experiments on HRSC2016 proved that this method is superior to the θ-based OBB.
(2) We designed the Gaussian Center-Line Attention learning module to further improve the positioning accuracy of the guide points.
(3) We designed the decoding method of CGP Box and verified the experimental results under different parameters. We tested our approach on three public datasets (DOTA, HRSC2016, and UCAS-AOD) and compared them with other the State-of-the-Art (SOTA) algorithms.

2. Related work

Although remote sensing object detection has its particularity, it still belongs to a branch of general object detection. Therefore, at present, most remote sensing object detectors are based on general detectors, such as R-CNN series (Girshick et al. 2014; Girshick 2015; Ren et al. 2015), SSD (Liu et al. 2016a), YOLO series (Van Etten 2018; Redmon et al. 2016; Redmon and Farhadi 2017), Retinanet (Lin et al. 2017a), FCN (Long, Shelhamer, and Darrell 2015), etc.

2.1. Angle-based Oriented Object Detectors

The method of learning the angle value of the object was called the angle-based OBB detector. RRPN (Ma et al. 2018) was the first to propose R-RPN structure based on the rotated anchor, which enabled the network to learn direction information. RRCNN (Lin et al. 2017) proposed RRoI pooling to obtain the features of directed regions based on the R-RPN structure and developed the Non-Maximum Suppression (Neubeck and Luc 2006) algorithm to improve the suppression rate of overlapping targets. RDFPN (Yang et al. 2018) employed the feature pyramid network structure to improve the detection performance on ship targets. Rol-Transformer (Ding et al. 2019) designed Rol Learner to convert horizontal anchor to rotated anchor, effectively reducing the computing pressure brought by the rotated anchor. SCRDet (Yang et al. 2019b) used supervised attention modules to improve detection performance by fusing channel information. $R^3$Det (Yang et al. 2019a) and RSDet (Qian et al. 2019) put forward their solutions to the unstable problem caused by angle value regression.

2.2. Point-based Oriented Object Detectors

The method of finding four-point coordinates of OBB was called the point-based detector. Based on Faster-RCNN, R2CNN (Jiang et al. 2017) located the target by detecting two corner points and the width of the bounding box. Besides, the strategy of predicting OBB corner offset based on HBB was also a classic method, such as TextBoxes++(Liao, Shi, and Bai 2018) based on single-stage SSD detector, FR-O (Xia et al. 2018) and ICN (Azimi et al. 2018) based on RCNN framework. $O^2$-DNET (Wei et al. 2019) treated the oriented object as pairs of middle lines and predicted the intersection point, which was an emerging anchor-
free keypoint method, such as CornerNet (Law and Deng 2018). Our method was also implemented and tested under this framework.

2.3. Segmentation-based Oriented Object Detectors

The method to obtain the OBB based on the instance segmentation image of the target was called the segmentation-based detector. There was not a lot of work on this field, Mou and Zhu (2018) proposed a semantic bounding-aware unified multitask ResFCN to handle vehicle instance segmentation in the aerial images. STN (Jaderberg et al. 2015) and SENET (Hu, Shen, and Sun 2018) introduced spatial and channel attention mechanism to extract key feature respectively. Wang et al. (2019) proposed Mask OBB and used Inception Lateral Connection Network to enhance FPN. Our Gaussian Center-Line Attention module was similar to these methods but very different.

3. Proposed method

3.1. Framework

The basic framework of the method used in this paper is shown in Figure 3, which is an anchor-free key point detection network based on the CenterNet (Zhou, Wang, and Philipp 2019) backbone. We choose the subsampling part of ResNet (He et al. 2016) as the feature extractor of the network. When the size of the feature map is \( \frac{1}{32} \) times the input size, the transpose convolution with deformable convolution (Dai et al. 2017) module is used to carry out the upsampling of the feature map, and finally, the feature map is \( \frac{1}{2} \) times the input size. GC-L Attention Module is a supervised Gaussian Center-Line feature map learning module. The output and input feature maps of the module are combined by element-wise addition to highlight the central feature of the object. The regression branch is composed of a CenterPoint Map, a GuidePoint Map, a Radius Map,
a RatioMap, and an Offset Map. **Channels** in the figure represent the number of channels in the heatmap. The details of each part are explained one by one in the following sections.

### 3.2. Centre-Guide points OBB regression strategy

#### 3.2.1. The composition of Center-Guide points OBB

The Center-Guide points OBB is a direction-representation strategy with no angle value, as shown in Figure 4(a). The blue dotted line box is the ground truth of the object, \( w \) and \( h \) are the long and short sides of the rectangle box, respectively. The red point \( Q \) is the centre point, and the direction guideline (orange dotted line) of the object is drawn through the red centre-point \( Q \). Draw a circle with \( Q \) as the centre and \( r \) as the radius (dotted green circle), intersecting with the orange direction line at point \( G \). The direction of the vector \( \vec{QG} \) is the direction of the object. As shown in Figure 3, the positions of Point \( Q \) and Point \( G \) are obtained from CenterPoint Map \( P_q \) and GuidePoint Map \( P_g \). Let Figure 4(a) serves as the input image for the network in Figure 3, and let \( q, g \in \mathbb{R}^2 \) be the origin coordinates of point

![Figure 4](image-url)

**Figure 4.** Example of the composition of Center-Guide points OBB regression strategy. The truth values of key points are composed by Gaussian kernel.
Q and point G in the input image, respectively. After network down-sampling, the projection coordinates of Q and G on $P_q$ and $P_g$ in Figure 3 are $q'$ and $g'$, which are expressed as:

$$q' = \left\lfloor \frac{q}{T} \right\rfloor$$

$$g' = \left\lfloor \frac{g}{T} \right\rfloor$$

where $\lfloor \cdot \rfloor$ is the symbol of round down. Since the features on the images near the centre point or the guide point are very similar, setting the ground-truth value at only the corresponding position in the heatmap to 1 will cause foreground and background confusion and affect convergence, which is unreasonable. So we make the region near the point as ground truth, as shown in Figure 4(c), the label value at the coordinates of the centre and guide points on the heatmap is set to 1 (the point where the centre of the heatmap is 1), and the Gaussian kernel is used to generate non-zero negative samples around it (the non-1 area of the heatmap). This operation produces a locally high energy region near the positive sample as shown in Figure 4(c), and the rest region of the heatmap is set to 0. The Gaussian kernel is applied as follows:

$$p^q_{xyc} = \exp\left(-\frac{(x - q_x)^2 + (y - q_y)}{2\sigma} \right)$$

$$p^g_{xyc} = \exp\left(-\frac{(x - g_x)^2 + (y - g_y)}{2\sigma} \right)$$

Among them, $\sigma$ is a hyperparameter that changes adaptively with the scale of the object, following CornerNet (Law and Deng 2018). The radius $r$ and the ratio coefficient of the long and short sides $a$ are obtained by regression of Ratio Map $P_r \in \mathcal{R}_T^{\frac{w}{4} \times \frac{h}{4} \times 1}$ and Ratio Map $P_a \in [0, 1]^{\frac{w}{4} \times \frac{h}{4} \times 1}$. The long and short sides of the object box $(w, h)$ are expressed as:

$$h = 2 \times r$$

$$w = a \times h$$

The position of the centre point and direction point will incur quantization loss during downsampling. The loss can be expressed as $O_q = |q' - q|$ and $O_g = |g' - g|$, and the deviations of these two parts share a heatmap Offset Map $P_o \in \mathcal{R}_T^{\frac{w}{4} \times \frac{h}{4} \times 2}$ to correct.

### 3.2.2. Decoding of guide point OBB

In the inference stage, we first get the centre point by extracting $m$ local maximum point integer position coordinates $D_q = \{ (\hat{x}_i^q, \hat{y}_i^q) \}_{i=1}^m$ on the heatmap $P_q$. The corresponding value of each centre point $(\hat{x}_i^q, \hat{y}_i^q)$ on the three regression heatmaps of $P_r$, $P_a$ and $P_o$ are $r_i$, $a_i$ and $\delta x_i^q$. The long side $w_i$ and short side $h_i$ of bounding box can be solved by Eq (3) according to $r_i$ and $a_i$. For the guide point, we take out $2 \times m$ local maximum points coordinates $D_g = \{ (\hat{x}_j^g, \hat{y}_j^g) \}_{j=1}^{2m}$ on the heatmap $P_g$, similar to the extraction method of the centre points. Depending on the position of the guide points, the corresponding offset $\delta x_j^g$ will also be obtained on the heatmap $P_g$. The centre point and guide point obtained
here are integer coordinates that need to be corrected with the offset coordinates. The corrected coordinate values can be calculated according to the following formula:

\[
(x_i, y_i) = (\tilde{x}_i + \delta x_i, \tilde{y}_i + \delta y_i)
\] (4)

Therefore, the corrected points \((x^q_i, y^q_i)\) and \((x^g_j, y^g_j)\) can be obtained. As shown in Figure 5, each centre point generates an valid area, and the guide points that fall within the area (valid points) will match the centre point, and the points outside the area (invalid points) will be discarded, where \(\eta \in (0, 1)\) is a super hyperparameter that controls the size of the valid area. The best match guide point \((x^g_i, y^g_i)\) is:

\[
Dis_{qg} = \sqrt{\left( x_i^q - x_j^g \right)^2 + \left( y_i^q - y_j^g \right)^2} - r_i, \ j \in [0, 2m), i \in [0, m), \ i, j \in \mathbb{N}
\] (5)

\[
(x^g_i, y^g_i) = \begin{cases} 
\arg \min(Dis_{qg}) & \text{if } Dis_{qg} \leq \eta \times r_i \\
\text{None} & \text{Otherwise}
\end{cases}
\] (6)

Finally, the direction angle of the object \(\theta_i\) is determined by the vector formed by the centre point and the guide point.

\[
\theta_i = \arccos \left( \frac{y^g_i - y^q_i}{x^g_i - x^q_i} \right)
\] (7)

where \(\theta_i \in [0, \pi]\). According to the geometric relationship, the bounding box of the object is calculated by \((x^q_i, y^q_i, w_i, h_i, \theta_i)\).
3.3. Angle-based method baseline

The angle-based OBB representation strategy is shown in Figure 4(b). Same as CGP OBB strategy, the orange-dotted line is the direction guide line, and the direction of the object is the angle $\theta$ between the guide line and the X-coordinate axis. So the object can be represented as $(x, y, w, h, \theta)$. In order to make a fair comparison, we also designed the angle-based OBB regression network as shown in Figure 6 as the baseline network. The output head of the network is composed of four parts: a CenterPoint Map, a Size Map, an Angle Map, and an Offset Map. Among them, the backbone, CenterPoint Map and Offset Map are set the same as CGP strategy. The Size Map $P_s \in \mathcal{R}^{\frac{H}{2} \times \frac{W}{2} \times 2}$ is used to regress the length of the long side and the short side of the target $(w, h)$, similar to the Radius Map $P_r$. The angle $\theta$ of the object is regression by Angle Map $P_a \in [0, \pi]^{\frac{H}{2} \times \frac{W}{2} \times 1}$, and the decoding method of bounding box is from OpenCV.

3.4. Gaussian Center-Line Attention Module

For the proposed CGP regression strategy, it is obvious that the positioning accuracy of the guide point and the centre point determines the accuracy of direction prediction. Since both the guide point and the centre point are on the centerline of the target, to further improve the accuracy of the orientation of the guide point, we propose a supervised object centerline prediction module called the Gaussian Center-Line (GC-L) Attention module. As shown in Figure 3, on the previous feature map of the output head, we obtain a Center-Line Map $P_{gcd} \in [0, 1]^{\frac{H}{2} \times \frac{W}{2} \times 1}$ through a simple convolution structure in Figure 7. The production method of CenterLine labels is shown in Figure 8. The longer edges of the two centerlines of the target are used to make a 0–1 binary map label (upper
Figure 7. The detailed structure of Gaussian Center-Line Attention module.

Figure 8. The ground truth heatmap of the GC-L attention module and the Gaussian kernel.
right of the figure). Where the width of the centerline (white line in the figure) on the label map is 1, and the value is 1. The values of the non-label area are all 0. Similar to the centre point, because the features of foreground and background near the centerline are similar and difficult to distinguish, it is easy to cause loss fluctuations in the training. Therefore, the Gaussian kernel is applied to the label to generate a heatmap of Gaussian distribution centred on the centerline (lower right in the figure), where the lighter the colour, the larger the value. The Gaussian kernel is set to Eq (2), and the hyperparameters are the same as the centre point settings. Finally, the centerline heatmap predicted by this module will be element-wise addition with the input feature map before the output head (⊕ in Figure 3), so the features of the centerline area will be enhanced.

3.5. Loss Function

In the prediction of the Centre-points Map, Guide-points Map and GC-L attention Map, the probabilistic heatmap applied by Gaussian kernel is used as the label to guide the learning of the network to the centre region, so as to reduce the influence caused by the similarity between foreground and the nearby background features. However, the foreground (the centre of the distribution) and background (the area around the centre of the distribution) are greatly unbalanced in quantity, so we follow the work of CenterNet (Zhou, Wang, and Philipp 2019) and CornerNet (Law and Deng 2018) and use the following Focal Loss (Lin et al. 2017b) as the loss function:

$$L_{focal}(P, \tilde{P}) = -\frac{1}{N} \sum_{xyc} \left( \frac{(1 - P_{xyc})^\gamma \log(P_{xyc})}{(1 - \tilde{P}_{xyc})^\beta P_{xyc}^\gamma \log(1 - P_{xyc})} \right)$$

if $\tilde{P}_{xyc} = 1$

otherwise

$$L_q = L_{focal}(P_q), \quad L_g = L_{focal}(P_g), \quad L_{gcl} = L_{focal}(P_g^{cl})$$

(9)

where $\tilde{P}_{xyc}$ and $P_{xyc}$ are the label value and predictive value of each probabilistic heatmap, $L_q$ indicates the loss on the CenterPoint Map, $L_g$ indicates the loss on the GuidePoint Map, $L_{gcl}$ indicates the loss on the CenterLine Map. Where $\gamma$ and $\beta$ are the tunable hyperparameters of the focal loss, which are used to balance the positive and negative losses. Following the experience of CenterNet, we set the parameter to $\gamma = 2$ and $\beta = 4$. $N$ is the number of keypoints in the image. For other output heatmaps, we use the L1 loss function for regression:

$$L_{l_i}(P, \tilde{P}) = \frac{1}{N} \sum_{i=1}^{N} |P_i - \tilde{P}_i|$$

(10)

$$L_r = L_{l_i}(P_r), \quad L_a = L_{l_i}(P_a), \quad L_o = L_{l_i}(P_o)$$

(11)

where $\tilde{P}_i$ and $P_i$ are the true value and predicted value of the i-th target on each regression heatmap, respectively. $L_r$ indicates the loss on the Radius Map, $L_a$ indicates the loss on the Ratio Map, $L_o$ indicates the loss on the Offset Map. Finally, the overall loss function is expressed as the weighted sum of the losses of each part:

$$L_{cg} = L_q + L_g + L_{gcl} + \lambda L_r + L_a + L_o$$

(12)
we set $\lambda = 0.5$ in our experiment to simply keep the loss of each part within the same order of magnitude.

4. Experiments

4.1. Datasets and implementation details

At present, there are not many public remote sensing image datasets labelled by oriented bounding box, among which DOTA, HRSC2016, and UCAS-AOD are the most widely used datasets in this field.

5. DOTA dataset

The DOTA (Xia et al. 2018) dataset contains a total of 15 categories of object, which are ‘plane,’ ‘baseball-diamond (BD),’ ‘bridge,’ ‘ground-track-field (GTF),’ ‘small-vehicle (SV),’ ‘large-vehicle (LV),’ ‘ship,’ ‘tennis-court (TC),’ ‘basketball-court (BC),’ ‘storage-tank (ST),’ ‘soccer-ball-field (SBF),’ ‘roundabout (RA),’ ‘harbor,’ ‘swimming-pool (SP),’ and ‘helicopter (HC).’ The dataset includes 1411 training set images and 458 validation set images. The entire dataset of DOTA contains a total of 188,282 target instances. These target instances are abundant in terms of scale, direction, and aspect ratio. The datasets are labelled with both OBB and HBB, and OBB is the only one selected for our experiment. In this dataset, ground sample distance (GSD) in metres is given, which ranges from 0.096 to 4.496 m, with a mean of 0.395 m, a standard deviation of 0.371 m. Although the number of images in the dataset is relatively small compared with the natural scene object detection dataset, the pixel size of the images in the dataset is very large, with many images even larger than 10 million pixels. The detection is usually carried out by spliting large image into small image. The DOTA dataset also provides an official server to test the results and is higher in data size dimensions, fairness, and authority than other datasets. The dataset is available on the official website.¹

6. HRSC2016 dataset

HRSC2016 (Liu et al. 2016b) is also a challenging remote sensing oriented object detection dataset for ships. The images of this dataset are from Google Earth. The data set contains 1061 pictures and the image size ranges from $300 \times 300$ to $1500 \times 900$. Among them, there are 436 pictures in the training set, 181 in the validation set, and 444 in the test set. Due to the small amount of data, the train set and the valid set are usually trained together, and the results are verified on the test set. The data set does not provide official verification services and no GSD data is provided. HRSC2016 is also an open source dataset and can be found on the Kaggle competition website.²

7. UCAS-AOD dataset

UCAS-AOD (Zhu et al. 2015) Orientation is an oriented object detection remote sensing dataset labelled by the Pattern Recognition and Intelligent System Development Laboratory of the University of Chinese Academy of Sciences. The dataset contains 1510
images of 2 categories (Plane and car) and gsd is not provided. The image size is fixed at 1250 ×750. Among them, there are 1000 images in the plane category, 7482 target instances, 510 images in the car category, and 7114 target instances. The official website\(^3\) of UCAS Pattern Recognition and Intelligent System Development Laboratory provides the download link for the dataset.

8. Evaluation Metric

The evaluation protocol follows the PASCAL VOC benchmark (Everingham et al. 2010), which uses mean Average Precision (mAP) as metric. Since the IoU calculation method of the oriented object was different from the general object detection method, we used the code provided by the official DOTA dataset to calculate. The metric of mAP on DOTA and UCAS-AOD was the VOC2012, in which the results on DOTA dataset were calculated by the official server, and the results of our study and the state-of-the-art detectors on UCAS-AOD dataset all adopted the same AP calculation method (VOC2012). On HRSC2016 dataset, most detectors used the VOC2007 metric to calculate mAP. Therefore, we also adopted this metric and compared it with the other SOTA detectors which using the same mAP metric.

9. Data preprocessing

Apart from the simple random rotation augmentation of all the input data during the training phase, we did not do any other augmentation operation. For the DOTA dataset, we processed it following the official data split tool and default parameters (size: 1024 × 1024 pixels, gap: 200 pixels). In our experimental preprocessing work, incomplete targets will appear due to image cropping. The incomplete targets with an area ratio of less than 0.4 to the original target will be removed. This processing not only ensures that the centre point will not out of the valid range, but also as far as possible ensures that the target has main features for detection.

10. Implementation details

The code was implemented by Pytorch 1.0 and experimented on 2 × NVIDIA GeForce GTX 2080 Ti. ResNet18 was used as the backbone for all experiments, and the batch size was set to 16. For parameter Settings, we followed the work of CenterNet, the initial learning rate was set to 1.25 × 10\(^{-4}\), which decreased by 10 times at the 100 epoch and the 150 epoch, respectively. After data statistics and experimental comparison, the maximum object number was usually set that \(m = 50\) but set \(m = 2000\) on the DOTA dataset due to its larger target density, and the parameters of the valid area of the guide point were selected as the optimal solution \(\eta = 0.4\). Finally, the loss function was optimized by Adam (Kingma and Jimmy 2014) optimizer.
Limited by equipment performance and computing power, we chose to do ablation experiments on the HRSC2016 dataset. The image input size of the experiment was selected as \(1024 \times 1024\). The simple experimental results are shown in Table 1. We chose the network which mentioned in Section 3.3 as the baseline. The results showed that compared to the angle-based method, CGP-based method improved mAP by 0.41\%. At the same time, the GC-L Attention module had an additional 0.1\% improvement in results. It could also be seen from the PR curve in Figure 9 that, on HRSC2016 dataset, the network performance with the application of CGP strategy was better, and the addition of GC-L achieved a small improvement.

Some visual images of the detection results are shown in Figure 10. Figure 10(a) is the detection result of the angle-based method, Figure 10(b) is the detection result of the CGP-based method, and Figure 10(c) is the visual heatmap of GC-L. We also compared the effect of the decoding parameter \(\eta\) on the detection results, as shown in Table 2. The results showed the best performance when \(\eta=0.4\).

### Table 1. Comparison of results under different decoding parameters \(\eta\) on HRSC2016 dataset.

<table>
<thead>
<tr>
<th>(\eta)</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>81.544</td>
<td>90.439</td>
<td>90.459</td>
<td>90.461</td>
<td>90.452</td>
</tr>
</tbody>
</table>

### Table 2. Comparison of results of each components on HRSC2016 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Baseline</th>
<th>CGP strategy</th>
<th>CGP strategy + GC-L Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>89.92</td>
<td>90.33</td>
<td>90.43</td>
</tr>
</tbody>
</table>
Comparison with the State-of-the-Art methods

We compared our method with the State-of-the-Art methods on three datasets. Our method had excellent performance on HRSC2016, and had achieved competitive results on other datasets.

11. Results on HRSC2016

As shown in Table 3, we conducted experiments based on ResNet18 and ResNet50 on the HRSC2016 test dataset. Compared with other algorithms, our method had the SOTA performance (90.46% mAP) only using ResNet18. For a fair comparison, we chose the same input size (800 × 800) and backbone (ResNet50) as other methods and achieved 90.43% mAP.

In the case of real-time demand, the combination of ResNet18 and (800 × 800) could be used, which could achieve 90.2% mAP at 58FPS. For accuracy, a combination of ResNet50 and (1024 × 1024) could be selected, which had the best detection performance (90.52% mAP) and competitive performance on inference speed (23FPS). The Visualization results on the test dataset are shown in Figure 11.
Results on DOTA

Due to the limitation of computing power, it was difficult for our equipment to support more complex backbone on our method for sufficient training and testing on the DOTA dataset. Therefore, we only tested the network performance which ResNet18 as the backbone. Part of the results are shown in Figure 12 and Table 4. Under the condition of the ResNet18 backbone, our method not only achieved excellent performance in the categories of BD, LV, SP and RA, but also achieved competitive results on the whole (71.35\% mAP). At the same time, we welcomed other researchers to use more complex backbone to test the performance of our detectors, such as ResNet50, ResNet101, ResNet152, etc.

Table 3. Comparison with state-of-the-art methods on the HRSC2016 dataset.

<table>
<thead>
<tr>
<th>method</th>
<th>Backbone</th>
<th>Input Size</th>
<th>mAP</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2CNN (Jiang et al. 2017)</td>
<td>VGG16</td>
<td>800 × 800</td>
<td>73.07</td>
<td>2</td>
</tr>
<tr>
<td>RRPN (Ma et al. 2018)</td>
<td>ResNet101</td>
<td>800 × 800</td>
<td>79.08</td>
<td>3.5</td>
</tr>
<tr>
<td>R2PN(Zhang et al. 2018)</td>
<td>VGG16</td>
<td>–</td>
<td>79.6</td>
<td>–</td>
</tr>
<tr>
<td>RetinaNet-H(Yang et al. 2019a)</td>
<td>ResNet101</td>
<td>800 × 800</td>
<td>82.89</td>
<td>14</td>
</tr>
<tr>
<td>RRD(Liao et al. 2018)</td>
<td>VGG16</td>
<td>384 × 384</td>
<td>84.3</td>
<td>–</td>
</tr>
<tr>
<td>RoiTransformer(Ding et al. 2019)</td>
<td>ResNet101</td>
<td>512 × 800</td>
<td>86.2</td>
<td>6</td>
</tr>
<tr>
<td>Gliding Vertex(Chen, G.S. Xia, and X.Bai 2020)</td>
<td>ResNet101</td>
<td>–</td>
<td>88.20</td>
<td>–</td>
</tr>
<tr>
<td>RetinaNet-R(Yang et al. 2019a)</td>
<td>ResNet101</td>
<td>800 × 800</td>
<td>89.18</td>
<td>10</td>
</tr>
<tr>
<td>R3Det(Yang et al. 2019a)</td>
<td>ResNet101</td>
<td>800 × 800</td>
<td>89.26</td>
<td>12</td>
</tr>
<tr>
<td>R3Det-DCL(Yang et al. 2020)</td>
<td>ResNet101</td>
<td>–</td>
<td>89.46</td>
<td>–</td>
</tr>
<tr>
<td>PolarDet(Zhao et al. 2020)</td>
<td>ResNet50</td>
<td>800 × 800</td>
<td>90.13</td>
<td>32</td>
</tr>
<tr>
<td>Ours</td>
<td>ResNet18</td>
<td>800 × 800</td>
<td>90.20</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>800 × 800</td>
<td>90.43</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>ResNet18</td>
<td>1024 × 1024</td>
<td>90.46</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>1024 × 1024</td>
<td>90.52</td>
<td>23</td>
</tr>
</tbody>
</table>

Figure 11. Visualization of detection results on HRSC2016 dataset.

12. Results on DOTA

Due to the limitation of computing power, it was difficult for our equipment to support more complex backbone on our method for sufficient training and testing on the DOTA dataset. Therefore, we only tested the network performance which ResNet18 as the backbone. Part of the results are shown in Figure 12 and Table 4. Under the condition of the ResNet18 backbone, our method not only achieved excellent performance in the categories of BD, LV, SP and RA, but also achieved competitive results on the whole (71.35\% mAP). At the same time, we welcomed other researchers to use more complex backbone to test the performance of our detectors, such as ResNet50, ResNet101, ResNet152, etc.
Table 4. Comparison with state-of-the-art detectors on DOTA dataset.

<table>
<thead>
<tr>
<th>method</th>
<th>Backbone</th>
<th>Plane</th>
<th>BD</th>
<th>Bridge</th>
<th>GTF</th>
<th>SV</th>
<th>LV</th>
<th>Ship</th>
<th>TC</th>
<th>BC</th>
<th>ST</th>
<th>SBF</th>
<th>RA</th>
<th>Harbour</th>
<th>SP</th>
<th>HC</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR-O</td>
<td>ResNet50</td>
<td>79.42</td>
<td>77.13</td>
<td>17.7</td>
<td>64.05</td>
<td>35.3</td>
<td>38.02</td>
<td>37.16</td>
<td>89.41</td>
<td>69.64</td>
<td>59.28</td>
<td>50.3</td>
<td>52.91</td>
<td>47.89</td>
<td>46.3</td>
<td>54.13</td>
<td></td>
</tr>
<tr>
<td>RRPN</td>
<td>ResNet101</td>
<td>80.94</td>
<td>65.75</td>
<td>35.34</td>
<td>67.44</td>
<td>59.92</td>
<td>50.91</td>
<td>55.81</td>
<td>90.67</td>
<td>66.92</td>
<td>72.39</td>
<td>55.06</td>
<td>52.23</td>
<td>55.14</td>
<td>53.35</td>
<td>48.22</td>
<td>61.01</td>
</tr>
<tr>
<td>R2CNN</td>
<td>ResNet101</td>
<td>88.52</td>
<td>71.2</td>
<td>31.66</td>
<td>59.3</td>
<td>51.85</td>
<td>56.19</td>
<td>57.25</td>
<td>90.81</td>
<td>72.84</td>
<td>67.38</td>
<td>56.69</td>
<td>52.84</td>
<td>53.08</td>
<td>51.94</td>
<td>53.58</td>
<td>60.67</td>
</tr>
<tr>
<td>R-DFPN</td>
<td>ResNet101</td>
<td>80.92</td>
<td>65.82</td>
<td>33.77</td>
<td>58.94</td>
<td>55.77</td>
<td>50.94</td>
<td>54.78</td>
<td>90.33</td>
<td>66.34</td>
<td>68.66</td>
<td>48.73</td>
<td>51.76</td>
<td>55.1</td>
<td>51.32</td>
<td>55.88</td>
<td>57.94</td>
</tr>
<tr>
<td>ICN</td>
<td>ResNet101</td>
<td>81.36</td>
<td>74.3</td>
<td>47.7</td>
<td>70.32</td>
<td>64.89</td>
<td>67.82</td>
<td>69.98</td>
<td>90.76</td>
<td>79.06</td>
<td>78.2</td>
<td>53.64</td>
<td>62.9</td>
<td>67.02</td>
<td>64.17</td>
<td>50.23</td>
<td>68.16</td>
</tr>
<tr>
<td>RoITransformer</td>
<td>ResNet101</td>
<td>88.64</td>
<td>78.52</td>
<td>43.44</td>
<td>75.92</td>
<td>68.81</td>
<td>73.68</td>
<td>83.59</td>
<td>90.74</td>
<td>77.27</td>
<td>81.46</td>
<td>58.39</td>
<td>53.54</td>
<td>62.83</td>
<td>58.93</td>
<td>47.67</td>
<td>69.56</td>
</tr>
<tr>
<td>O^2-DNet</td>
<td>Hourglass104</td>
<td>89.31</td>
<td>82.14</td>
<td>47.33</td>
<td>61.21</td>
<td>71.32</td>
<td>74.03</td>
<td>78.62</td>
<td>90.76</td>
<td>82.23</td>
<td>81.36</td>
<td>60.93</td>
<td>60.17</td>
<td>58.21</td>
<td>66.98</td>
<td>61.03</td>
<td>71.04</td>
</tr>
<tr>
<td>R^3Det</td>
<td>ResNet101</td>
<td>89.54</td>
<td>81.99</td>
<td>48.46</td>
<td>62.52</td>
<td>70.48</td>
<td>74.29</td>
<td>77.54</td>
<td>90.80</td>
<td>81.39</td>
<td>83.54</td>
<td>61.97</td>
<td>59.82</td>
<td>65.44</td>
<td>67.46</td>
<td>60.05</td>
<td>71.69</td>
</tr>
<tr>
<td>SCRDet</td>
<td>ResNet101</td>
<td>89.98</td>
<td>80.65</td>
<td>52.09</td>
<td>68.36</td>
<td>68.36</td>
<td>60.32</td>
<td>72.41</td>
<td>90.85</td>
<td>87.94</td>
<td>86.86</td>
<td>65.02</td>
<td>66.68</td>
<td>66.25</td>
<td>68.24</td>
<td>65.21</td>
<td>72.61</td>
</tr>
<tr>
<td>Ours</td>
<td>ResNet18</td>
<td>89.00</td>
<td><strong>82.83</strong></td>
<td>34.12</td>
<td>63.91</td>
<td>74.04</td>
<td><strong>74.70</strong></td>
<td>79.35</td>
<td>90.81</td>
<td>80.19</td>
<td>84.74</td>
<td>55.77</td>
<td><strong>67.51</strong></td>
<td>64.72</td>
<td><strong>73.15</strong></td>
<td>55.47</td>
<td>71.35</td>
</tr>
</tbody>
</table>
13. Results on UCAS-AOD

The detection results of UCAS-AOD are shown in Table 5. Here, we used the ResNet50 backbone for training and testing, and achieved 95.08\% mAP. There was not much difference between this result and the optimal detector, indicating that our method has good universality in remote sensing images.

Table 5. Comparison with state-of-the-art detectors on UCAS-AOD dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Plane</th>
<th>Car</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv2 [Redmon and Farhadi 2017]</td>
<td>96.60</td>
<td>79.20</td>
<td>87.90</td>
</tr>
<tr>
<td>R-DFPN [Yang et al. 2018]</td>
<td>98.90</td>
<td>82.50</td>
<td>89.20</td>
</tr>
<tr>
<td>DRBox [Liu, Pan, and Lei 2017]</td>
<td>94.90</td>
<td>85.00</td>
<td>89.95</td>
</tr>
<tr>
<td>O2-DNet [Wei et al. 2019]</td>
<td>93.21</td>
<td>86.72</td>
<td>89.96</td>
</tr>
<tr>
<td>S2ARN [Bao et al. 2019]</td>
<td>97.60</td>
<td>92.20</td>
<td>94.90</td>
</tr>
<tr>
<td>RetinaNet-H [Yang et al. 2019a]</td>
<td>97.34</td>
<td>93.60</td>
<td>95.47</td>
</tr>
<tr>
<td>ICN [Azimi et al. 2018]</td>
<td>—</td>
<td>—</td>
<td>95.67</td>
</tr>
<tr>
<td>FADet [Li et al. 2019]</td>
<td>98.69</td>
<td>92.72</td>
<td>95.71</td>
</tr>
<tr>
<td>Ours</td>
<td>97.88</td>
<td>92.28</td>
<td>95.08</td>
</tr>
</tbody>
</table>
14. Discussion of Limitations

In the experiments, we found that the setting of batch size on the DOTA dataset had a great influence on the result, and the best result could be achieved when the batch size was greater than 32. Due to device limitations, it was difficult for us to use more complex base networks under the condition of batch size = 32, such as ResNet50 and ResNet101. However, no such phenomenon was found in the HRSC2016 or the USAC-AOD dataset. We believed that the reason was that DOTA dataset has more abundant features and target categories, and training under a small batch was easy to overfit some features and categories. Therefore, we will open-source the code, and welcome interested workers to continue our work.

15. Conclusion

Oriented object detection in remote sensing image is always an important research field. In the process of direction prediction, there is ambiguity in many methods to represent OBB. In this paper, we propose a novel OBB representation called CGP Box. We apply CGP Box to an anchor-free key point detection framework and compare it with the traditional representation method based on angle value. The experiment proves that the CGP Box-based method is better than the angle-based representation method. To further improve the accuracy of Guide Point positioning, we propose Gaussian Center-Line Attention Learning Module. Our experiments on three classic remote sensing public datasets (DOTA, HRSC2016, UCAS-AOD) show that CGP Box performs competitively even with a low-complexity backbone, and we would like to welcome other workers to optimize CGP performance with a more complex backbone in the future.

List of acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>Oriented Bounding Box</td>
</tr>
<tr>
<td>OBB</td>
<td>Oriented Bounding Box</td>
</tr>
<tr>
<td>HBB</td>
<td>Horizontal Bounding Box</td>
</tr>
<tr>
<td>CGP</td>
<td>Center-Guide Points</td>
</tr>
<tr>
<td>GC-L</td>
<td>Gaussian Center-Line</td>
</tr>
<tr>
<td>IoU</td>
<td>Intersection over Union</td>
</tr>
<tr>
<td>BD</td>
<td>baseball-diamond</td>
</tr>
<tr>
<td>GTF</td>
<td>ground-track-_eld</td>
</tr>
<tr>
<td>SV</td>
<td>small-vehicle</td>
</tr>
<tr>
<td>LV</td>
<td>large-vehicle</td>
</tr>
<tr>
<td>TC</td>
<td>tennis-court</td>
</tr>
<tr>
<td>BC</td>
<td>basketball-court</td>
</tr>
<tr>
<td>ST</td>
<td>storage-tank</td>
</tr>
<tr>
<td>SBF</td>
<td>soccer-ball-_eld</td>
</tr>
<tr>
<td>RA</td>
<td>roundabout</td>
</tr>
<tr>
<td>SP</td>
<td>swimming-pool</td>
</tr>
<tr>
<td>HC</td>
<td>helicopter</td>
</tr>
<tr>
<td>SOTA</td>
<td>State-of-the-Art</td>
</tr>
<tr>
<td>mAP</td>
<td>mean Average Precision</td>
</tr>
<tr>
<td>GSD</td>
<td>Ground Sample Distance</td>
</tr>
</tbody>
</table>
Notes


Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Qiuyu Guan http://orcid.org/0000-0003-0263-5600

References


