G-Rep: Gaussian Representation for Arbitrary-Oriented Object Detection

Liping Hou [1] **Ke Lu** [1 2] **Xue Yang** [3] **Yuqiu Li** [1] **Jian Xue** [1]

Abstract

Arbitrary-oriented object representations contain the oriented bounding box (OBB), quadrilateral bounding box (QBB), and point set (PointSet). Each representation encounters problems that correspond to its characteristics, such as the boundary discontinuity, square-like problem, representation ambiguity, and isolated points, which lead to inaccurate detection. Although many effective strategies have been proposed for various representations, there is still no unified solution. Current detection methods based on Gaussian modeling have demonstrated the possibility of breaking this dilemma; however, they remain limited to OBB. To go further, in this paper, we propose a unified Gaussian representation called G-Rep to construct Gaussian distributions for OBB, QBB, and PointSet, which achieves a unified solution to various representations and problems. Specifically, PointSet or QBB-based objects are converted into Gaussian distributions, and their parameters are optimized using the maximum likelihood estimation algorithm. Then, three optional Gaussian metrics are explored to optimize the regression loss of the detector because of their excellent parameter optimization mechanisms. Furthermore, we also use Gaussian metrics for sampling to align label assignment and regression loss. Experimental results on several public available datasets, DOTA, HRSC2016, UCAS-AOD, and ICDAR2015 show the excellent performance of the proposed method for arbitrary-oriented object detection.

1. Introduction

With the development of deep convolutional neural network (CNN), the object detection, especially arbitrary-orientation object detection (Azimi et al., 2018; Ding et al., 2019; Yang et al., 2019a; 2021b;c), has been developed rapidly and a variety of methods have been proposed.

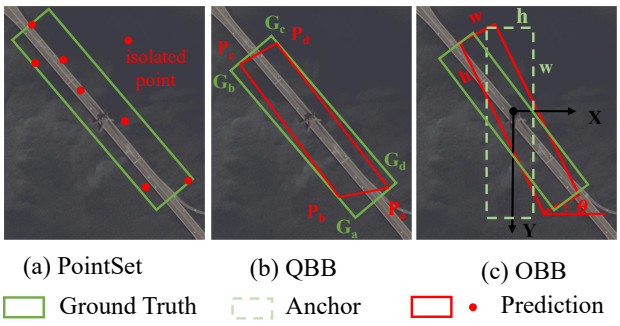


Figure 1. Illustrations of different problems for different representations: (a) Dissociation of PointSet. (b) Representation ambiguity of QBB. (c) Boundary discontinuity of OBB.

Popular representations of the arbitrary-oriented object are divided into an oriented bounding box (OBB), quadrilateral bounding box (QBB), and point set (PointSet). Each representation encounters intrinsic issues because of the properties of its definitions, which are summarized as follows and illustrated in Figure 1. **i)** PointSet uses several individual points to represent the overall arbitrary-orientation object, and the independent optimization between the points makes the trained detector very sensitive to isolated points, particularly for objects with large aspect ratios because a slight deviation causes a sharp drop in the intersection over union (IoU). As shown in Figure 1 (a), although most of the points are predicted correctly, an outlier makes the final prediction fail. Therefore, the joint optimization loss (e.g., IoU loss (Yu et al., 2016; Rezatofighi et al., 2019; Zheng et al., 2020)) based on the point set is more popular than the independent optimization loss (e.g., L_n loss). **ii)** As a special case of PointSet, QBB is defined as the four corners of a quadrilateral bounding box. In addition to the inherent problems of PointSet described above, QBB also suffers from the representation ambiguity problem (Ming et al., 2021a). Quadrilateral detection often sorts the points first (represented by green box Figure 1(b)) to facilitate the point matching between the ground-truth and prediction bounding boxes to calculate the final loss function. Although the red prediction box in Figure 1 (b) does not satisfy the sorting

[1]School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China [2]College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China [3]Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Jian Xue <xuejian@ucas.ac.cn>.

rule and calculates a large loss accordingly using the L_n loss, this prediction is correct according to the IoU-based evaluation metric. **iii)** OBB is the most popular choice for oriented object representation because of its simplicity and intuitiveness. However, the boundary discontinuity and square-like problem are obstacles for high-precision location, which is detailed in (Yang & Yan, 2020; Yang et al., 2021a;c;d). Figure 1 (c) illustrates the boundary problem of OBB representation, considering the OpenCV acute angle definition as an example ($\theta \in [-\pi/2, 0)$) (Yang et al., 2018). The height (h) and width (w) of the box swap at the angle boundary and result in a sudden change in the loss value, coupled with the periodicity of the angle, which makes regression difficult.

To improve detection performance, many researchers have proposed solutions for some issues, which mainly include the boundary discontinuity (Yang et al., 2019a; Yang & Yan, 2020), square-like problem (Yang et al., 2021a; Qian et al., 2021), representation ambiguity (Ming et al., 2021a), and isolated points. Recent methods GWD (Yang et al., 2021c) and KLD (Yang et al., 2021d) have broken the paradigm of existing regular regression frameworks from a unified regression loss perspective. Although promising results have been achieved, the limitation of the OBB representation makes them not truly unified solutions in terms of representation. Additionally, although several distances for the Gaussian distribution have been explored and devised as the regression loss, the metric for dividing positive and negative samples in label assignment is not changed accordingly, and IoU is still used.

In this paper, we aim to develop a fully unified solution to the problems that result from various representations. Specifically, the PointSet representation and its special cases, QBB, are converted into the Gaussian distribution, and evaluate the latter's parameters using the maximum likelihood estimation (MLE) algorithm (Dempster et al., 1977). Furthermore, several evaluation metrics are explored for computing the similarity of the Gaussian distribution, and design Gaussian regression losses based on the above metrics. Accordingly, the selection of positive and negative samples are modified using the Gaussian metric by designing fixed and dynamic label assignment strategies. The entire pipeline is shown in Figure 2. The highlights of this paper are as follows:

1) To uniformly solve the different problems introduced by different representations (OBB, QBB, and PointSet), we propose Gaussian representation (G-Rep) to construct the Gaussian distribution using the MLE algorithm.

2) To achieve an effective and robust measurement for the Gaussian distribution, we explore three statistical distances, Kullback–Leibler divergence (KLD) (Kullback & Leibler, 1951), Bhatacharyya Distance (BD) (Bhattacharyya, 1943), and Wasserstein Distance (WD) (Villani, 2008), and design corresponding regression loss functions using the above metrics.

3) To realize the alignment in the measurement between sample selection and loss regression, we construct fixed and dynamic label assignment strategies based on a Gaussian metric to further boost performance.

4) We conducted extensive on several publicly available datasets, DOTA, HRSC2016, UCAS-AOD, and ICDAR2015, and the results demonstrated the excellent performance of the proposed techniques for arbitrary-oriented object detection. The source code has been open sourced in MMRotate[1](Zhou et al., 2022).

2. Related Work

2.1. Oriented Object Representations

The most popular representation of oriented object detection uses the five-parameter OBB, IoU value between OBB and the ground truth as the metric in label assignment, Smooth l_1 as the regression loss to regress the five parameters (x, y, w, h, θ) (Jiang et al., 2017; Ding et al., 2019; Yang et al., 2021b; 2018; Yang & Yan, 2020; Qian et al., 2021; Yang et al., 2019a), respectively, where (x, y), w, h, and θ denote the center coordinates, width, height, and angle of the box, respectively.

Additionally, some researchers have proposed using an eight-parameter QBB to represent the object and Smooth l_1 as regression loss to as the regression loss to regress the four corner points of QBB (Xu et al., 2020; Ming et al., 2021a; Liu et al., 2019; Feng et al., 2020). The recent anchor-free method CFA (Guo et al., 2021) uses the more flexible PointSet (i.e., point set) to represent the oriented object, inspired by the horizontal object detection method RepPoints (Yang et al., 2019b). Additionally, other complex representations exist, such as polar coordinates (Zhou et al., 2020; Zhao et al., 2021) and middle lines (Wei et al., 2020).

2.2. Regression Loss in Oriented Object Detection

The mainstream regression loss for OBB and QBB is Smooth l_1, which encounters boundary discontinuity and square-like problems. To solve these problems, SCRDet (Yang et al., 2019a) and RSDet (Qian et al., 2021) adopt the IoU-Smooth l_1 loss and modulated loss to smooth the the boundary loss jump. CSL (Yang & Yan, 2020) and DCL (Yang et al., 2021a) transform angular prediction from regression to classification. RIDet (Ming et al., 2021a) uses the representation invariant loss to optimize bounding box regression. Using the IoU value as the regression loss (Yu et al., 2016) has become a research topic of great interest in oriented object detection, and can avoid partial problems

[1]https://github.com/open-mmlab/mmrotate

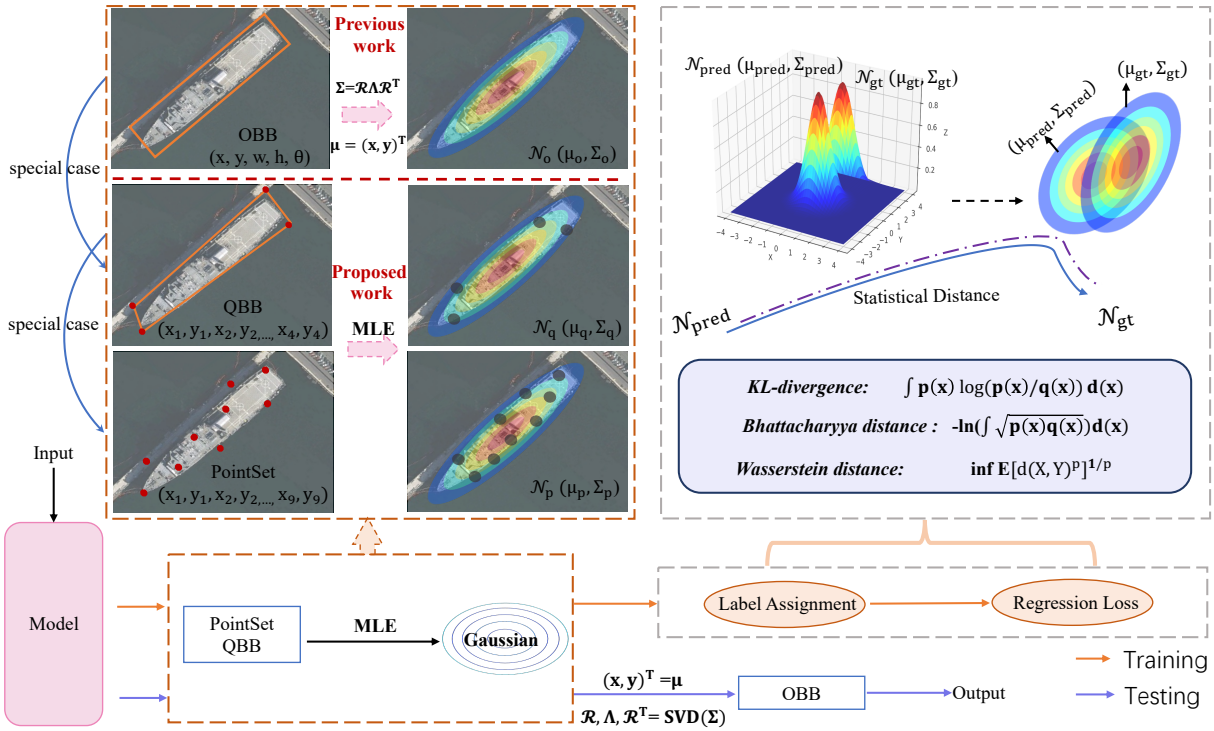


Figure 2. Overview of the main contributions of this paper. Gaussian distributions of QBB and PointSet are constructed, and label assignment strategies and regression losses are designed in an alignment manner on the basis of statistical distances.

caused by the regression of the angle parameter or point ordering. Additionally, many variants of the IoU loss have been developed, for instance, DIoU (Zheng et al., 2020), GIoU (Rezatofighi et al., 2019), and PIoU (Chen et al., 2020). GWD (Yang et al., 2021c) and KLD (Yang et al., 2021d) construct the Gaussian distribution for OBB, which overcomes the dilemma of object detection and demonstrates the possibility of a unified solution for various issues. However, they are limited to OBB representation, and are not truly unified solutions in terms of representation.

2.3. Label Assignment Strategies

Label assignment plays a vital role in detection, and many fixed and dynamic label assignment strategies have been proposed. Classic object detection methods, for instance, Faster RCNN (Ren et al., 2015), and RetinaNet (Lin et al., 2017b), adopt a fixed max IoU strategy, which requires predefined positive and negative thresholds in advance. To overcome the difficulty in setting hyperparameters, ATSS (Zhang et al., 2020) uses statistical characteristics to calculate dynamic IoU thresholds. Furthermore, PAA (Kim & Lee, 2020) adaptively separates anchors into positive and negative samples for a ground-truth bounding box in a probabilistic manner. Additionally, other excellent dynamic label assignment strategies exist, for example, DAL (Ming et al., 2021b) and FreeAnchor (Zhang et al., 2021). Despite this, these methods still rely on the IoU value as the main metric for evaluating the quality of the sample.

3. Proposed Approach

In this paper, our main contribution focuses on three aspects. First, constructing Gaussian distribution for PointSet and QBB representation, overcomes the dilemma that the Gaussian distribution can only be applied to OBB, as shown in previous studies (Yang et al., 2021c;d). Second, we explore new regression loss functions for supervising network learning based on the Gaussian distribution. Third, we align the measurement between label assignment and loss regression based on the Gaussian distribution, and then construct new fixed and dynamic label assignment strategies accordingly.

Figure 2 shows an overview of the proposed method. In the training phase, we convert the predicted bounding box and ground truth into the Gaussian distribution in a manner according to their respective representations. For example, we use the MLE algorithm for the boxes in PointSet or QBB representations according to Equation (2). Then, we adopt the devised fixed or dynamic G-Rep label assignment strategy to select samples based on Gaussian distance metrics (KLD, BD, or WD). Finally, we design regression loss functions based on the Gaussian distance metrics to minimize the distance between two Gaussian distributions. In the

testing phase, we obtain the output in OBB from the trained model of the parameter weight, which we construct based on predictions in PointSet or QBB.

3.1. Object Representation as the Gaussian Distribution

PointSet. RepPoints (Yang et al., 2019b) is a anchor-free method that uses PointSet, and is also a baseline used in this paper. Specifically, the object is represented as a set of adaptive sample points (i.e., point set) and the regression framework adopts deformable convolution for point learning. The PointSet is defined as $P = \{(x_i^p, y_i^p)\}_{i=1}^N$, where (x_i^p, y_i^p) represents the coordinates of the i-th point, and N is the number of all points in a point set, which is set to 9 by default following (Yang et al., 2019b).

QBB/OBB. The baseline with QBB representation is constructed on the anchor-based method Cas-RetinaNet constructed in (Ming et al., 2021a), which contains a backbone network and two detection heads. QBB is defined as the four corner points of the object, that is, $(Q = \{(x_i^q, y_i^q)\}_{i=1}^4)$. Note that the four corner points of QBB must be sorted in advance to match the corners of the given ground truth representation one-to-one for regression in the original QBB baseline. Additionally, from the definitions of the three representations, we can deduce that QBB can be regarded as a special case of PointSet, and OBB can be regarded as a special case of QBB. Therefore, constructing the Gaussian distribution for PointSet is extremely generalized, which is also our major focus in this paper.

Transformation between PointSet/QBB and G-Rep. Considering (x_i, y_i) as a two-dimensional (2-D) variable $\boldsymbol{x}_i$, its probability density under the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined as

$$\mathcal{N}\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)\right]}{2\pi\sqrt{\det\left(\boldsymbol{\Sigma}\right)}}. \tag{1}$$

The parameters mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of the Gaussian distribution are calculated using the maximum likelihood estimation (MLE) (Richards, 1961) algorithm, which is calculated as

$$\begin{aligned}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})_{\mathrm{MLE}} &= \arg\max \prod_{i=1}^{N} \mathcal{N}\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ &= \arg\max \sum_{i=1}^{N} \log \mathcal{N}\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right),\end{aligned} \tag{2}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T, \tag{3}$$

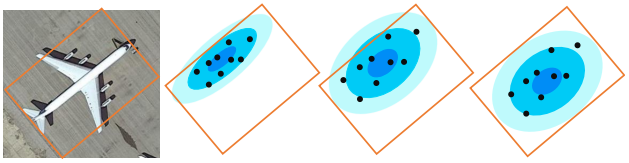


Figure 3. Illustration of the learning process of the Gaussian distribution based on PointSet.

where $\boldsymbol{x}_i = (x_i, y_i)$ denotes the coordinates of the i-th point, and N denotes the number of all points in a 2-D vector. Figure 3 illustrates the Gaussian learning process based on PointSet. The MLE algorithm evaluates the parameters of the Gaussian distribution for the initialized PointSet, which is considered as a 2-D vector. The coordinates of the points are updated according to the gradient feedback of the loss, and the Gaussian parameters are updated correspondingly.

Transformation between OBB and G-Rep. In previous studies (Yang et al., 2021c;d), a 2-D Gaussian distribution for OBB is constructed using a matrix transformation.

The transformation from a Gaussian distribution to an OBB is necessary for calculating the mean average precision (mAP) in the testing process. As shown in Figure 2, prediction (QBB or PointSet) is obtained in the testing process. Because network learning is supervised in the basis of the object Gaussian distribution, the prediction of the network output is also distributed in the form of the object Gaussian distribution. Therefore, it is necessary to convert the prediction (PointSet or QBB) into the Gaussian distribution, whose parameters are calculated using the MLE algorithm in Equation (2). To go further, the OBB of the predicted object is obtained from the Gaussian distribution. For a predicted Gaussian distribution with known parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and the parameters of the corresponding OBB (x, y, w, h, θ) can also be obtained using the singular value decomposition (SVD).

3.2. Gaussian Distance Metric

The keystone of an effective regression loss and label assignment is how to compute the similarity between the predicted Gaussian distribution $\mathcal{N}_p\left(\boldsymbol{x}_p \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\right)$ and the ground-truth Gaussian distribution $\mathcal{N}_g\left(\boldsymbol{x}_g \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$. Next, we will focus on three metrics to calculate the distance between two Gaussian distributions and further analyze their characteristics.

Kullback–Leibler divergence (Kullback & Leibler, 1951). The KLD between two Gaussian distributions is

defined as

$$D_{\mathrm{KL}}\left(\mathcal{N}_g, \mathcal{N}_p\right) = \frac{1}{2}\Big(\operatorname{tr}\left(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_g\right) + \ln\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_g|} - 2 + \left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_g\right)^{\top}\boldsymbol{\Sigma}_p^{-1}\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_g\right)\Big). \tag{4}$$

Although KLD is not strictly a mathematical distance because it is asymmetric, KLD is scale invariant. According to Equation (4) each item of KLD contains shape parameters $\boldsymbol{\Sigma}$ and center parameters $\boldsymbol{\mu}$. All parameters form a chain coupling relationship and influence each other, which is very conducive to high-precision detection, as proved in (Yang et al., 2021d).

Bhatacharyya distance (Bhattacharyya, 1943). The BD between two Gaussian distributions is defined as

$$D_{\mathrm{B}}\left(\mathcal{N}_g, \mathcal{N}_p\right) = \frac{1}{8}\left(\boldsymbol{\mu}_g - \boldsymbol{\mu}_p\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_g - \boldsymbol{\mu}_p\right) + \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}|}{\sqrt{|\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_p|}}\right), \tag{5}$$

where $\boldsymbol{\Sigma} = \frac{1}{2}\left(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_g\right)$. Although BD is symmetric, not an actual distance because it does not satisfy the triangle inequality. Similar with KLD, BD is also scale invariant.

Wasserstein distance (Villani, 2008). The WD between two Gaussian distributions is defined as

$$D_{W}\left(\mathcal{N}_g, \mathcal{N}_p\right) = \left\|\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right\|^2 + \operatorname{Tr}\left(\boldsymbol{\Sigma}_p\right) + \operatorname{Tr}\left(\boldsymbol{\Sigma}_q\right) - 2\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{\Sigma}_q\boldsymbol{\Sigma}_p^{1/2}\right)^{1/2}\right). \tag{6}$$

Different from KLD and BD, WD is an actual mathematical distance, and satisfies the triangle inequality and is symmetric. Note that WD is mainly divided into two parts: the distance between the center points (x, y) and and the coupling terms about h, w, and θ. Although WD can greatly improve the performance of high-precision rotation detection because of the coupling between parts of the parameters, the independent optimization of the center point slightly shifts the detection result (Yang et al., 2021d).

3.3. Regression Loss Based on a Gaussian Metric

The actual value obtained by the distance metric between Gaussian distributions is too large to be the regression loss, which leads to difficult convergence. Therefore, normalization is necessary so that the Gaussian distance can be used as regression loss. The normalized functions are devised by a series of small-scale experiments, in which some empirical functions are tested, and the best functions for different metrics are chosen according to the results. The normalized regression loss functions for KLD, BD, and WD are defined as

$$\mathcal{L}_{\mathrm{KLD}} = 1 - \frac{1}{2 + \sqrt{D_{\mathrm{KL}}\left(\mathcal{N}_g, \mathcal{N}_p\right)}}, \tag{7}$$

$$\mathcal{L}_{\mathrm{BD}} = 1 - \frac{1}{1 + D_{\mathrm{B}}\left(\mathcal{N}_g, \mathcal{N}_p\right)}, \tag{8}$$

$$\mathcal{L}_{\mathrm{WD}} = 1 - \frac{1}{1 + \log(1 + D_{W}\left(\mathcal{N}_g, \mathcal{N}_p\right))}, \tag{9}$$

where $D_{\mathrm{KL}}\left(\mathcal{N}_g, \mathcal{N}_p\right)$, $D_{\mathrm{B}}\left(\mathcal{N}_g, \mathcal{N}_p\right)$, and $D_{W}\left(\mathcal{N}_g, \mathcal{N}_p\right)$ are calculated in Equation (4), Equation (5), and Equation (6), respectively.

Recall that the most common regression loss for object detection is the Smooth l_1 loss. For different representations, the Smooth l_1 loss suffers from different issues, as described in Figure 1. By contrast, all three Gaussian metric-based regression losses avoid issues of the boundary discontinuity, square-like problem, representation ambiguity, and isolated points. Among them, WD is strictly the actual distance because of its triangle equality and symmetry, whereas KLD is asymmetric and BD satisfies the triangle inequality. Although the three proposed regression loss functions have different characteristics, the overall difference in detection performance is slight and all are superior to the traditional Smooth l_1 or IoU loss.

3.4. Label Assignment Based on a Gaussian Metric

Label assignment strategy is another key task for object detection. The most popular label assignment strategy is the IoU-based strategy, which assigns a label by comparing IoU values (proposals and ground truth) with IoU threshold. However, there is a misalignment when the metrics in label assignment and the regression loss are different when the regression loss is based on the Gaussian distribution. Therefore, new label assignment strategies based on the Gaussian distribution are devised.

Fixed G-Rep Label Assignment. The range of the IoU value is $[0, 1]$ and according to the definition of IoU, and the threshold values are selected empirically in the range $[0.3, 0.7]$. However, this strategy is clearly not applicable to the Gaussian distribution distance calculated by the three metrics in Section 3.2, whose value ranges are not closed intervals. Along with the concept of G-Rep regression loss design, normalized functions for each distance evaluation metric are adopted. Additionally, the key formula for normalization function design maps the evaluation values to $[0, 1]$. In this paper, functions are explored for normalizing the obtained distance metrics (KLD, BD, and WD) in the label assignment process, which are listed as

$$\mathcal{S}_{\mathrm{KLD}} = \frac{1}{2 + D_{\mathrm{KL}}\left(\mathcal{N}_g, \mathcal{N}_p\right)}, \tag{10}$$

$$\mathcal{S}_{\mathrm{BD}} = \frac{1}{1 + D_{\mathrm{B}}^2\left(\mathcal{N}_g, \mathcal{N}_p\right)}, \tag{11}$$

$$\mathcal{S}_{\mathrm{WD}} = \frac{1}{2 + D_W\left(\mathcal{N}_g, \mathcal{N}_p\right)}, \tag{12}$$

where $\mathcal{S}$ denotes the normalized metric for evaluating the similarity between the Gaussian distribution of a proposal and the ground truth. The optimal hyperparameters require a empirical and experimental optimization. A series of experiments are conducted with unified hyperparameter positive and negative thresholds $[0.4, 0.3]$ and results are listed in Table 3. The results (rows 3–5) demonstrate that optimal hyperparameter thresholds are difficult to set for different distance metrics; hence, dynamic label assignment strategies based on a Gaussian are further explored.

Dynamic G-Rep Label Assignment. Dynamic G-Rep label assignment strategies are devised based on the three distance metrics in Section 3.2 to avoid the difficulty of selecting the optimal hyperparameters. Inspired by ATSS (Zhang et al., 2020), the threshold for selecting positive and negative samples is calculated dynamically according to the statistical characteristics of all the normalized distance (calculated in Equation (10), Equation (11), and Equation (12)). For the i-th ground truth, the dynamic threshold $\mathcal{T}$ for is calculated as

$$\mathcal{T}_i = \frac{1}{J}\sum_{j=1}^{J}\mathcal{I}_{i,j} + \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\mathcal{I}_{i,j} - \frac{1}{J}\sum_{j=1}^{J}\mathcal{I}_{i,j})^2}, \tag{13}$$

where J is the number of candidate samples, and $\mathcal{I}_{i,j}$ is the normalized KLD, BD, or WD between the i-th ground truth and the j-th proposal, which is calculated in Equation (10), Equation (11), or Equation (12). Next, positive samples are selected using the general assignment strategy, that is, candidates are selected whose similar values are greater than or equal to the threshold $\mathcal{T}_i$. Furthermore, motivated by PAA (Kim & Lee, 2020), another approach to using Gaussian distributions in the label assignment process is adopted in this paper. Specifically, each sample is assigned a score for quality evaluation. The Gaussian mixture model (GMM) with two components is adopted to model the score distribution of samples, and the parameters of the GMM are calculated using the expectation-maximization (Dempster et al., 1977) algorithm. Finally, samples are assigned labels according to their probabilities. PAA and ATSS are modified and combined to construct the robust dynamic label assignment strategy PATSS in the experiments.

4. Experiments

4.1. Datasets and Implementation Details

Experiments are conducted on the public aerial image datasets DOTA (Xia et al., 2018), HRSC2016 (Liu et al., 2017), UCAS-AOD (Zhu et al., 2015), and scene text dataset ICDAR2015 (Karatzas et al., 2015) to verify the superiority of the proposed method.

Table 1. Ablation study of G-Rep for PointSet on DOTA and HRSC2016. $\mathcal{S}^{\mathrm{la}}$ and $\mathcal{L}^{\mathrm{reg}}$ represent the label assignment strategy and regression loss function, respectively.

DATASET	REP.	$\mathcal{S}^{\mathrm{LA}}$	$\mathcal{L}^{\mathrm{REG}}$	MAP(%)
DOTA	POINTSET	IOU (MAX)	GIOU	63.97
	G-REP	IOU (MAX)	$\mathcal{L}_{\mathrm{KLD}}$	64.63 **(+0.66)**
		$\mathcal{S}_{\mathrm{KLD}}$ (MAX)	$\mathcal{L}_{\mathrm{KLD}}$	65.07 **(+1.10)**
	POINTSET	IOU (ATSS)	GIOU	68.88
	G-REP	$\mathcal{S}_{\mathrm{KLD}}$ (ATSS)	$\mathcal{L}_{\mathrm{KLD}}$	70.45 **(+1.57)**
		$\mathcal{S}_{\mathrm{KLD}}$ (PATSS)	$\mathcal{L}_{\mathrm{KLD}}$	72.08 **(+3.20)**
HRSC2016	POINTSET	IOU (ATSS)	GIOU	78.07
	G-REP	$\mathcal{S}_{\mathrm{KLD}}$ (ATSS)	$\mathcal{L}_{\mathrm{KLD}}$	88.06 **(+9.99)**
		$\mathcal{S}_{\mathrm{KLD}}$ (PATSS)	$\mathcal{L}_{\mathrm{KLD}}$	89.15 **(+11.07)**

DOTA (Xia et al., 2018) is a public oriented object detection benchmark dataset for aerial images, and contains 15 categories: plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). In the experiments, DOTA's official training and validation dataset sets are selected as the training set, whose images are split into blocks of 1,024×1,024 pixels with an overlap of 200 pixels and scaled to 1,333×1,024 for training. HRSC2016 (Liu et al., 2017) is a public aerial image dataset for ship detection. The number of images in the training, variation, and testing sets are 436,181, and 444, respectively. The public aerial image dataset UCAS-AOD (Zhu et al., 2015) is another multi-class oriented object detection dataset, and contains two categories: car and airplane. ICDAR2015 (Karatzas et al., 2015) is commonly used for oriented scene text detection and spotting. This dataset contains 1,000 training images and 500 testing images.

Experiments are conducted using the MMDetection framework (Chen et al., 2019) (under the Apache-2.0 License) on 2 Titan V GPUs with 11GB memory. Two baselines of RepPoints with PointSet and Cas-RetinaNet with QBB are conducted. The optimizers of the RepPoints and Cas-RetinaNet frameworks are SGD with a 0.01 initial learning rate and Adam with a 0.0001 initial learning rate, respectively. The framework for the RepPoints baseline is trained for 12 and 36 epochs on DOTA and HRSC2016 for the ablation experiments, and 40, 80, 120, and 240 epochs on DOTA, HRSC2016, UCAS-AOD, and ICDAR2015 with data augmentation, respectively. The framework for the Cas-RetinaNet baseline is trained for 12, 100, 36, and 100 epochs on DOTA, HRSC2016, UCAS-AOD, and ICDAR2015, respectively. Mean Average Precision (mAP) is adopted as evaluation metric for testing results on DOTA, HRSC2016 and UCAS-AOD. Precision, recall, and F-measure (i.e., F1) are adopted on ICDAR2015 following official criteria.

Table 2. Performance comparison of PointSet and G-Rep for large aspect ratio objects. 'mAR' represents the mean aspect ratio (the ratio of long side to short side) of all targets in one category.

REPRESENTATION	BR	SV	LV	SH	HC	MAP(%)
	MAR=2.93	MAR=1.72	MAR=3.45	MAR=2.40	MAR=2.34	
POINTSET	46.87	77.10	71.65	83.71	32.93	62.45
G-REP	50.82 **(+3.95)**	79.33 **(+2.23)**	75.07 **(+3.51)**	87.32 **(+3.61)**	50.63 **(+17.70)**	68.63 **(+6.18)**

Table 3. Comparison of the three Gaussian distances as metrics for label assignment and regression loss on HRSC2016.

REPRESENTATION	$\mathcal{S}^{\text{LA}}$	$\mathcal{L}^{\text{REG}}$	MAP(%)
POINTSET	IOU (ATSS)	GIOU	78.07
G-REP	$\mathcal{S}_{\text{KLD}}$ (MAX)	$\mathcal{L}_{\text{KLD}}$	73.44
	$\mathcal{S}_{\text{BD}}$ (MAX)	$\mathcal{L}_{\text{BD}}$	46.71
	$\mathcal{S}_{\text{WD}}$ (MAX)	$\mathcal{L}_{\text{WD}}$	84.39
	$\mathcal{S}_{\text{KLD}}$ (ATSS)	$\mathcal{L}_{\text{KLD}}$	88.06
	$\mathcal{S}_{\text{BD}}$ (ATSS)	$\mathcal{L}_{\text{BD}}$	85.32
	$\mathcal{S}_{\text{WD}}$ (ATSS)	$\mathcal{L}_{\text{WD}}$	88.56
	$\mathcal{S}_{\text{KLD}}$ (ATSS)	$\mathcal{L}_{\text{BD}}$	88.90
	$\mathcal{S}_{\text{KLD}}$ (ATSS)	$\mathcal{L}_{\text{WD}}$	88.80
	$\mathcal{S}_{\text{BD}}$ (ATSS)	$\mathcal{L}_{\text{KLD}}$	85.32
	$\mathcal{S}_{\text{BD}}$ (ATSS)	$\mathcal{L}_{\text{WD}}$	85.28

Table 4. Ablation study of G-Rep for QBB representations on various datasets. '*' denotes that dynamic ATSS-based strategies are adopted. The regression loss of G-Rep is the $\mathcal{L}_{\text{KLD}}$.

DATA	REPRESENTATION	MAP(%)	GAIN ↑
DOTA	POINTSET*	68.88	–
	G-REP* (POINTSET)	70.45	**+1.57**
	QBB	63.05	–
	G-REP (QBB)	67.92	**+4.87**
HRSC2016	POINTSET*	78.07	–
	G-REP* (POINTSET)	88.06	**+9.99**
	QBB	87.70	–
	G-REP (QBB)	88.01	**+0.31**
UCAS-AOD	POINTSET*	90.15	–
	G-REP* (POINTSET)	90.20	**+0.05**
	QBB	88.50	–
	G-REP (QBB)	88.82	**+0.32**
ICDAR2015	POINTSET*	76.20	–
	G-REP* (POINTSET)	81.30	**+5.10**
	QBB	75.10	–
	G-REP (QBB)	75.83	**+0.73**

4.2. Ablation Study

Table 1 compares the performance of the G-Rep and PointSet on the DOTA dataset. "IoU (Max)" and "IoU (ATSS)" represent the fixed predefined threshold strategy and dynamic ATSS (Zhang et al., 2020) label assignment strategy based on the IoU metric, respectively. The baseline method of PointSet is RepPoints (Yang et al., 2019b) with ResNet50-FPN (He et al., 2016; Lin et al., 2017a). Note that the IoU for the Gaussian distribution is calculated as the IoU between the box transformed by the Gaussian distribution and ground truth. The predefined positive and negative thresholds in the fixed strategy for IoU and $\mathcal{S}_{\text{KLD}}$ are both $[0.3, 0.4]$. The superiority of G-Rep is reflected in two aspects: regression loss and label assignment.

Analysis of regression loss based on G-Rep. Even if only the GIoU is replaced by $\mathcal{L}_{\text{KLD}}$, the performance of G-Rep is better than that of PointSet (**64.63%** vs. **63.97%**). The dynamic label assignment strategies avoided the influence of unsuitable hyperparameters for a fair comparison of the GIoU and Gaussian regression loss. Additionally, the superiority of G-Rep is clearly demonstrated when the dynamic label assignment strategies are used. $\mathcal{L}_{\text{KLD}}$ still surpassed the GIoU with the same dynamic label assignment strategy ATSS on DOTA (**70.45%** vs. **68.88%**).

Analysis of label assignment based on G-Rep. The label assignment strategy is another important step for high-detection performance. For the $\mathcal{L}_{\text{KLD}}$ loss, Table 1 shows the detection results of different label assignment strategies. Using KLD performed better than using IoU as a metric of the label assignment, which demonstrates the effectiveness of aligning the label assignment and regression loss metrics. The optimal fixed negative and positive thresholds for selecting samples are difficult to select, whereas dynamic label assignment strategies avoided this issue. PATSS denotes the combination of the ATSS (Zhang et al., 2020) and PAA (Kim & Lee, 2020) strategies. The mAP further reached **70.45%** and **72.08%** under the more robust dynamic selection strategies ATSS and PATSS, respectively. Without additional features, the combination of the dynamic label assignment strategy and regression loss increased the mAP by **8.11%** mAP compared with the baseline method.

Analysis of the advantages for an object with a large aspect ratio. Outlier results in more serious location errors for objects with large aspect ratios than square objects. Table 2 shows that G-Rep is more effective for large aspect ratio objects than PointSet, with the mAP increasing by **6.18%** mAP for the five typical categories with narrow objects on DOTA because G-Rep is not sensitive to isolated points.

Comparison of different Gaussian distance metrics. Table 3 compares the performances when different evolution metrics, KLD, WD, and BD, are used in fixed and dynamic label assignment strategies and regression loss. The experimental results demonstrated that the overall performances

Table 5. Comparison of various detectors of mAP values on the OBB-based task of the DOTA-v1.0. MS indicates that multi-scale training. R-101 denotes ResNet-101 (likewise for R-50, R-152). RX-101, H-104 and Swin-T denotes ResNeXt101 (Xie et al., 2017), Hourglass-104 (Newell et al., 2016) and Swin Transformer (Liu et al., 2021).

METHOD	BACKBONE	MS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	MAP(%)
two-stage:																		
GSDET (2021)	R-101		81.12	76.78	40.78	75.89	64.50	58.37	74.21	89.92	79.40	78.83	64.54	63.67	66.04	58.01	52.13	68.28
SCRDET (2019A)	R-101	✓	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	86.86	65.02	66.68	66.25	68.24	65.21	72.61
SARD (2019)	R-101		89.93	84.11	54.19	72.04	68.41	61.18	66.00	90.82	87.79	86.59	65.65	64.04	66.68	68.84	68.03	72.95
FADET (2019)	R-101	✓	**90.21**	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.40	84.64	53.40	65.42	74.17	69.69	64.86	73.28
MFIAR-NET(2020A)	R-152	✓	89.62	84.03	52.41	70.30	70.13	67.64	77.81	90.85	85.40	86.22	63.21	64.14	68.31	70.21	62.11	73.49
GLIDING VERTEX (2020)	R-101		89.64	**85.00**	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	**70.91**	72.94	70.86	57.32	75.02
CENTERMAP (2020)	R-101	✓	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
CSL (FPN-BASED) (2020)	R-152	✓	90.25	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
RSDET (2021)	R-152	✓	89.93	84.45	53.77	74.35	71.52	78.31	78.12	91.14	87.35	86.93	65.64	65.17	75.35	79.74	63.31	76.34
OPLD (2020)	R-101	✓	89.37	85.82	54.10	79.58	75.00	75.13	86.92	90.88	86.42	86.62	62.46	68.41	73.98	68.11	63.69	76.43
SCRDET++ (2020B)	R-101	✓	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
one-stage:																		
P-RSDET (2020)	R-101		89.02	73.65	47.33	72.03	70.58	73.71	72.76	90.82	80.12	81.32	59.45	57.87	60.79	65.21	52.59	69.82
O^2-DET (2020)	H-104		89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
R^3DET (2021B)	R-152	✓	89.24	80.81	51.11	65.62	70.67	76.03	78.32	90.83	84.89	84.42	65.10	57.18	68.10	68.98	60.88	72.81
BBAVECTORS (2021)	R-101	✓	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	73.32
DRN (2020)	H-104	✓	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
GWD (2021C)	R-152		88.88	80.47	52.94	63.85	76.95	70.28	83.56	88.54	83.51	84.94	61.24	65.13	65.45	71.69	73.90	74.09
CFA (2021)	R-101		89.26	81.72	51.81	67.17	79.99	78.25	84.46	90.77	83.40	85.54	54.86	67.75	73.04	70.24	64.96	75.05
KLD (2021D)	R-50		88.91	83.71	50.10	68.75	78.20	76.05	84.58	89.41	86.15	85.28	63.15	60.90	75.06	71.51	67.45	75.28
S^2A-NET (2021)	R-101		88.70	81.41	54.28	59.75	78.04	80.54	88.04	90.69	84.75	86.22	65.03	65.81	76.16	73.37	58.86	76.11
POLARDET (2021)	R-101	✓	89.65	87.07	48.14	70.97	78.53	80.34	87.45	90.76	85.63	86.87	61.64	70.32	71.92	73.09	67.15	76.64
DAL (S^2A-NET) (2021B)	R-50	✓	89.69	83.11	55.03	71.00	78.30	81.90	88.46	90.89	84.97	87.46	64.41	65.65	76.86	72.09	64.35	76.95
DCL (R^3DET) (2021A)	R-152	✓	89.26	83.60	53.54	72.76	79.04	82.56	87.31	90.67	86.59	86.98	67.49	66.88	73.29	70.56	69.99	77.37
RIDET (2021A)	R-50		89.31	80.77	54.07	76.38	79.81	81.99	89.13	90.72	83.58	87.22	64.42	67.56	**78.08**	79.17	62.07	77.62
QBB (BASELINE)	R-50		77.52	57.38	37.20	65.97	56.29	69.99	70.04	90.31	81.14	55.34	57.98	49.88	56.01	62.32	58.37	63.05
POINTSET (BASELINE)	R-50		87.48	82.53	45.07	65.16	78.12	58.72	75.44	90.78	82.54	85.98	60.77	67.68	60.93	70.36	44.41	70.39
G-REP (QBB)	R-101		88.89	74.62	43.92	70.24	67.26	67.26	79.80	90.87	84.46	78.47	54.59	62.60	66.67	67.98	52.16	70.59
G-REP (POINTSET)	R-50		87.76	81.29	52.64	70.53	80.34	80.56	87.47	90.74	82.91	85.01	61.48	68.51	67.53	73.02	63.54	75.56
G-REP (POINTSET)	RX-101	✓	88.98	79.21	57.57	74.35	**81.30**	85.23	88.30	90.69	85.38	85.25	63.65	68.82	77.87	78.76	71.74	78.47
G-REP*(POINTSET)	SWIN-T	✓	88.15	81.64	**61.30**	**79.50**	80.94	**85.68**	**88.37**	**90.90**	85.47	**87.77**	**71.01**	67.42	77.19	**81.23**	**75.83**	**80.16**

Table 6. Comparison of the mAP of various rotation methods on HRSC2016 dataset.

METHOD	MAP(%)
ROI-TRANSFORMER (DING ET AL., 2019)	86.20
RSDET (QIAN ET AL., 2021)	86.50
GLIDING VERTEX (XU ET AL., 2020)	88.20
BBAVECTORS (YI ET AL., 2021)	88.60
R^3DET (YANG ET AL., 2021B)	89.26
DCL (YANG ET AL., 2021A)	89.46
G-REP (QBB)	88.02
G-REP (POINTSET)	**89.46**

Table 7. Comparison of the AP with state-of-the-art methods on UCAS-AOD dataset.

METHOD	CAR	AIRPLANE	MAP(%)
RETINANET (2017B)	84.64	90.51	87.57
FASTER-RCNN (2015)	86.87	89.86	88.36
ROI-TRANSFORMER (2019)	88.02	90.02	89.02
RIDET-Q (2021A)	88.50	89.96	89.23
RIDET-O (2021A)	88.88	90.35	89.62
DAL (2021B)	89.25	90.49	89.87
G-REP (QBB)	87.35	90.30	88.82
G-REP (POINTSET)	**89.64**	**90.67**	**90.16**

of the G-Rep loss functions surpassed that of the GIoU loss. There are tolerable performance differences between BD and the other two losses, and a slight difference (within 0.5%) between KLD and WD. To further explore whether KLD and WD are more suitable as the regression loss than BD, the label assignment metrics are unified as KLD for the ablation of the BD loss. In fact, all three G-Rep losses outperformed the baseline (RepPoints). There are slight differences between them in detection performance.

Ablation study on various datasets. Table 4 shows the experimental results of G-Rep using two baselines on various datasets. The QBB baseline adopted the anchor-based method Cas-RetinaNet (Ming et al., 2021a) (i.e., the cascaded RetinaNet (Lin et al., 2017b)). G-Rep resulted in varying degrees of improvement on the anchor-based baseline with QBB and anchor-free baseline with PointSet on various datasets. G-Rep is dramatically effective for objects with large aspect ratios (e.g. ships of HRSC2016, text of ICDAR2015). In contrast, the isotropic Gaussian distribution of the square-like objects has only a slight boost to the baseline due to the unpredictable angle.

4.3. Comparison with the State-of-the-Art

The performance of the proposed method is compared with that of other state-of-the-art detection methods on the DOTA

dataset, which is a benchmark aerial image dataset for multi-class oriented object detection. As shown in Table 5, the G-Rep based on anchor-free method RepPoints achieves state-of-the-art performance. R-50/101, RX-101, H-104, and Swin-T denote ResNet-50/101 (He et al., 2016), ResNeXt-101 (Xie et al., 2017), Hourglass-104 (Newell et al., 2016), and Swin-Transformer(Liu et al., 2021),respectively. G-Rep is a pioneering new paradigm in oriented object detection. To further verify the effectiveness of the proposed method and compare it with other state-of-the-art detectors, a series of experiments are conducted on HRSC2016 and UCAS-AOD, and the results are shown in Table 6 and Table 7, respectively. The results demonstrate that the baseline models that used G-Rep achieved state-of-the-art performance.

5. Conclusions

The main contribution of this study is that G-Rep is proposed to construct the Gaussian distribution on PointSet and QBB, which overcame the limitation of Gaussian applications for current object detection methods and truly achieved a unified solution. Thus, a series of detection challenges that resulted from object representations are alleviated. Additionally, label assignment strategies for the Gaussian distribution are designed, and the metrics are aligned between label assignment and regression loss. More importantly, G-Rep overcomes the current dilemma, and provides inspiration for exploring other forms of label assignment strategies and regression loss functions.

References

Azimi, S. M., Vig, E., Bahmanyar, R., Körner, M., and Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, pp. 150–165. Springer, 2018.

Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Chen, Z., Chen, K., Lin, W., See, J., Yu, H., Ke, Y., and Yang, C. Piou loss: Towards accurate oriented object detection in complex environments. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2020.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Proceedings of the Royal Statistical Society*, 39(1):1–22, 1977.

Ding, J., Xue, N., Long, Y., Xia, G.-S., and Lu, Q. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2849–2858, 2019.

Feng, P., Lin, Y., Guan, J., He, G., Shi, H., and Chambers, J. Toso: Student'st distribution aided one-stage orientation target detection in remote sensing images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4057–4061. IEEE, 2020.

Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., and Ye, Q. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8792–8801, 2021.

Han, J., Ding, J., Li, J., and Xia, G.-S. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., and Luo, Z. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition*, pp. 1156–1160. IEEE, 2015.

Kim, K. and Lee, H. S. Probabilistic anchor assignment with iou prediction for object detection. In *European Conference on Computer Vision*, pp. 355–371. Springer, 2020.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Li, C., Xu, C., Cui, Z., Wang, D., Zhang, T., and Yang, J. Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE International Conference on Image Processing*, pp. 3886–3890. IEEE, 2019.

Li, W., Wei, W., and Zhang, L. Gsdet: Object detection in aerial images based on scale reasoning. *IEEE Transactions on Image Processing*, 30:4599–4609, 2021.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017a.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017b.

Liu, Y., Zhang, S., Jin, L., Xie, L., Wu, Y., and Wang, Z. Omnidirectional scene text detection with sequential-free box discretization. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2019.

Liu, Z., Yuan, L., Weng, L., and Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, volume 2, pp. 324–331, 2017.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ming, Q., Miao, L., Zhou, Z., Yang, X., and Dong, Y. Optimization for arbitrary-oriented object detection via representation invariance loss. *IEEE Geoscience and Remote Sensing Letters*, 2021a.

Ming, Q., Zhou, Z., Miao, L., Zhang, H., and Li, L. Dynamic anchor learning for arbitrary-oriented object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2355–2363, 2021b.

Newell, A., Yang, K., and Deng, J. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pp. 483–499. Springer, 2016.

Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., and Xu, C. Dynamic refinement network for oriented and densely packed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11207–11216, 2020.

Qian, W., Yang, X., Peng, S., Yan, J., and Guo, Y. Learning modulated loss for rotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2458–2466, 2021.

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

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

Richards, F. S. A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(2):469–475, 1961.

Song, Q., Yang, F., Yang, L., Liu, C., Hu, M., and Xia, L. Learning point-guided localization for detection in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 1084–1094, 2020.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wang, J., Yang, W., Li, H.-C., Zhang, H., and Xia, G.-S. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4307–4323, 2020.

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

Wei, H., Zhang, Y., Chang, Z., Li, H., Wang, H., and Sun, X. Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:268–279, 2020.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., and Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Yang, F., Li, W., Hu, H., Li, W., and Wang, P. Multi-scale feature integrated attention-based rotation network for object detection in vhr aerial images. *Sensors*, 20(6): 1686, 2020a.

Yang, X. and Yan, J. Arbitrary-oriented object detection with circular smooth label. In *European Conference on Computer Vision*, pp. 677–694. Springer, 2020.

Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., and Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sensing*, 10(1):132, 2018.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., and Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8232–8241, 2019a.

Yang, X., Yan, J., Yang, X., Tang, J., Liao, W., and He, T. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*, 2020b.

Yang, X., Hou, L., Zhou, Y., Wang, W., and Yan, J. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15819–15829, 2021a.

Yang, X., Yan, J., Feng, Z., and He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3163–3171, 2021b.

Yang, X., Yan, J., Qi, M., Wang, W., Xiaopeng, Z., and Qi, T. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International Conference on Machine Learning*, 2021c.

Yang, X., Yang, X., Yang, J., Ming, Q., Wang, W., Tian, Q., and Yan, J. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 2021d.

Yang, Z., Liu, S., Hu, H., Wang, L., and Lin, S. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9657–9666, 2019b.

Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., and Metaxas, D. Oriented object detection in aerial images with box boundary-aware vectors. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 2150–2159, 2021.

Yu, J., Jiang, Y., Wang, Z., Cao, Z., and Huang, T. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520, 2016.

Zhang, S., Chi, C., Yao, Y., Lei, Z., and Li, S. Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9756–9765, 2020. doi: 10.1109/CVPR42600.2020.00978.

Zhang, X., Wan, F., Liu, C., Ji, X., and Ye, Q. Learning to match anchors for visual object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3050494.

Zhao, P., Qu, Z., Bu, Y., Tan, W., and Guan, Q. Polardet: A fast, more precise detector for rotated target in aerial images. *International Journal of Remote Sensing*, 42(15): 5821–5851, 2021.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12993–13000, 2020.

Zhou, L., Wei, H., Li, H., Zhang, Y., Sun, X., and Zhao, W. Objects detection for remote sensing images based on polar coordinates. *IEEE ACCESS*, 2020.

Zhou, Y., Yang, X., Zhang, G., Wang, J., Liu, Y., Hou, L., Jiang, X., Liu, X., Yan, J., Lyu, C., Zhang, W., and Chen, K. Mmrotate: A rotated object detection benchmark using pytorch. *arXiv preprint arXiv:2204.13317*, 2022.

Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., and Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing*, pp. 3735–3739. IEEE, 2015.

A. Normalized Function Design

Table 8 lists the experimental results of the normalized function design. The design principle of the normalized functions for label assignment is mapping the calculated values of Gaussian statistical distance to the values in $[0, 1]$. The design principle of the normalized functions for regression loss is transforming the calculated values to the values in a suitable range (e.g. $[0, 10]$). In addition, the normalized regression loss functions should have a reasonable gradient descent. Some generic functions, such as $\log(*)$ and e^*, are selected in the experiments. Finally, we chose the appropriate normalized functions according to the best results of the experiments.

Table 8. Experiment results of normalized function design of $\mathcal{S}^{\text{la}}$ and $\mathcal{L}^{\text{reg}}$ on HRSC2016.

METRIC	FUN. OF $\mathcal{S}^{\text{LA}}$	FUN. OF $\mathcal{L}^{\text{REG}}$	MAP(%)
KLD	$\frac{1}{2+D_{\text{KL}}}$	$1-\frac{1}{2+\sqrt{D_{\text{KL}}}}$	**88.06**
	$\frac{1}{2+D_{\text{KL}}}$	$1-e^{-\sqrt{D_{\text{KL}}}}$	87.32
	$\frac{1}{2+D_{\text{KL}}}$	$1-e^{-D_{\text{KL}}^2}$	50.73
	$\frac{1}{2+\sqrt{D_{\text{KL}}}}$	$1-\frac{1}{2+\sqrt{D_{\text{KL}}}}$	35.46
BD	$\frac{1}{1+D_{\text{BD}}^2}$	$1-\frac{1}{1+D_{\text{BD}}}$	**85.32**
	$\frac{1}{1+D_{\text{BD}}^2}$	$\log(1+D_{\text{BD}})$	85.02
	$\frac{1}{1+D_{\text{BD}}^2}$	$5\times D_{\text{BD}}$	60.68
	$\frac{1}{1+D_{\text{BD}}}$	$1-\frac{1}{1+D_{\text{BD}}}$	85.12
WD	$\frac{1}{2+D_{\text{WD}}}$	$1-\frac{1}{1+\log(1+D_{\text{WD}})}$	**88.56**
	$\frac{1}{2+D_{\text{WD}}}$	$1-\frac{1}{2+\sqrt{D_{\text{WD}}}}$	87.04
	$\frac{1}{2+D_{\text{WD}}}$	$1-e^{-\sqrt{D_{\text{WD}}}}$	88.24
	$\frac{1}{2+\sqrt{D_{\text{WD}}}}$	$1-\frac{1}{1+\log(1+D_{\text{WD}})}$	87.54

B. Overview of Two Baselines

RepPoints. RepPoints baseline is constructed with a backbone network, an initial detection head and a refined detection head. The backbone network utilizes ResNet and Feature Pyramid Network (FPN) (He et al., 2016; Lin et al., 2017a) to extract multi-scale features, and two detection heads consist of non-shared classification and regression sub-networks based on deformable convolution. The object is represented as PointSet, which is defined as:

$$R = \{(x_i, y_i)\}_{i=1}^{K}, \tag{14}$$

where (x_i, y_i) is the coordinates of the i-th feature point and K is the point number of a point set, which is set to 9 by default. In the refined detection head, the model predicts offset $(\Delta x_i, \Delta y_i)$ for refinement, and the new refined predicted point set can be simply expressed as:

$$R' = \{(x_i + \Delta x_i, y_i + \Delta y_i)\}_{i=1}^{K}, \tag{15}$$

Finally, a function is used to convert a predicted point set to a horizontal bounding box, which is selected from three converting functions: min-max, partial min-max, and moment-based function, as mentioned in (Yang et al., 2019b).

Cas-RetinaNet. The baseline method of QBB representation is Cas-RetinaNet, which is contructed by the cascaded RetinaNet (Lin et al., 2017b) with two detection steps. Specifically, another detection step is added to refine the predicted boxes and improve the accuracy of localization.

C. Transformation between OBB and G-Rep

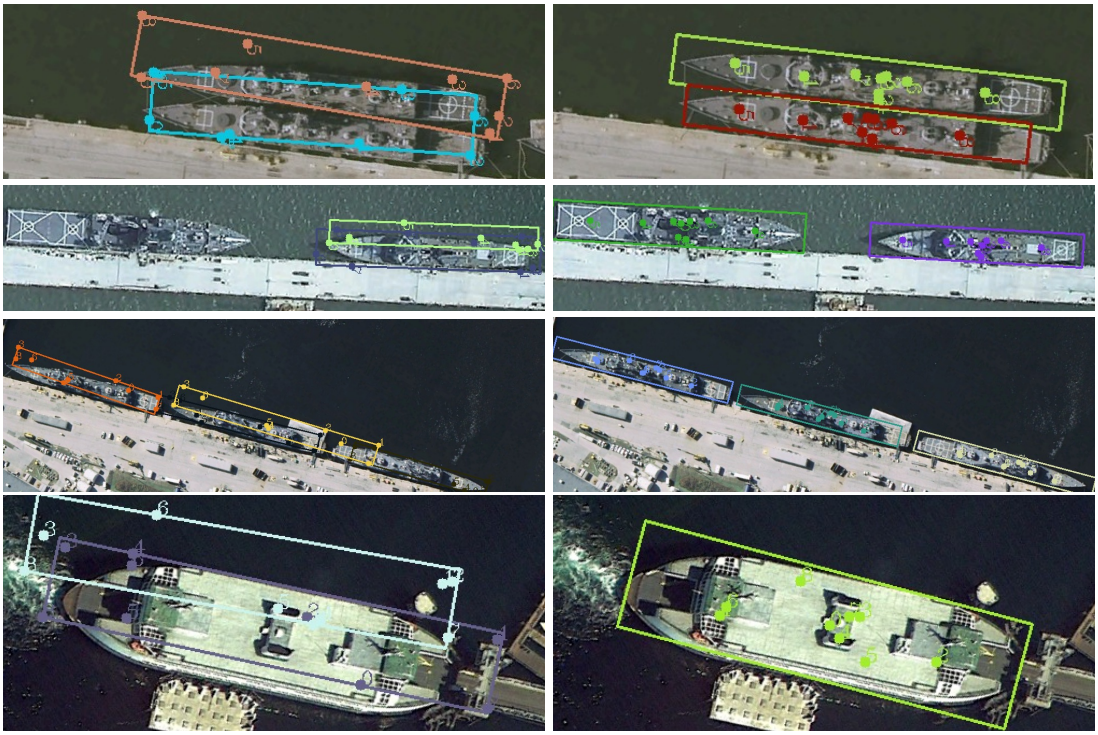
For an OBB (x, y, w, h, θ), the parameters of the Gaussian distribution $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are calculated as

$$\begin{aligned}
\boldsymbol{\mu} &= [x, y]^\top \\
\boldsymbol{\Sigma} &= \mathbf{R}_\theta \boldsymbol{\Lambda} \mathbf{R}_\theta{}^\top \\
&= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta & (\lambda_1 - \lambda_2)\cos\theta\sin\theta \\ (\lambda_1 - \lambda_2)\cos\theta\sin\theta & \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta \end{bmatrix},
\end{aligned} \tag{16}$$

where $\boldsymbol{x}$ denotes the vector representation of the coordinates (x, y). $\mathbf{R}_\theta$ and $\boldsymbol{\Lambda}$ are the rotation matrix and diagonal matrix of the eigenvalues, respectively. The eigenvalues are calculated as $\lambda_1 = \frac{w^2}{4}$ and $\lambda_2 = \frac{h^2}{4}$.

D. Additional Visualization Results

Figure 4 visualizes the detection results of PointSet and G-Rep. The points of PointSet are distributed at the boundary of objects, while the points of the G-Rep are distributed in the interior of the object. Therefore, G-Rep is superior in terms of accurate localization and is not sensitive to outlier. More visualization examples of the experimental results on DOTA and UCAS-AOD datasets are shown in Figure 5 and Figure 6. These visualization results of different kinds of objects on different datasets sufficiently demonstrate the superiority of G-Rep.



PointSet **G-Rep (Ours)**

Figure 4. Comparison of the visual results of PointSet and G-Rep on HRSC2016 dataset.

E. Discussion of Limitation

Although the proposed method provides a uniform representation of the Gaussian distribution for various representations of the input, an obvious limitation is that the format of the output can only be OBB. The output points are dispersed as the Gaussian distribution inside the object, and OBB is transformed from the Gaussian distribution of the output points.

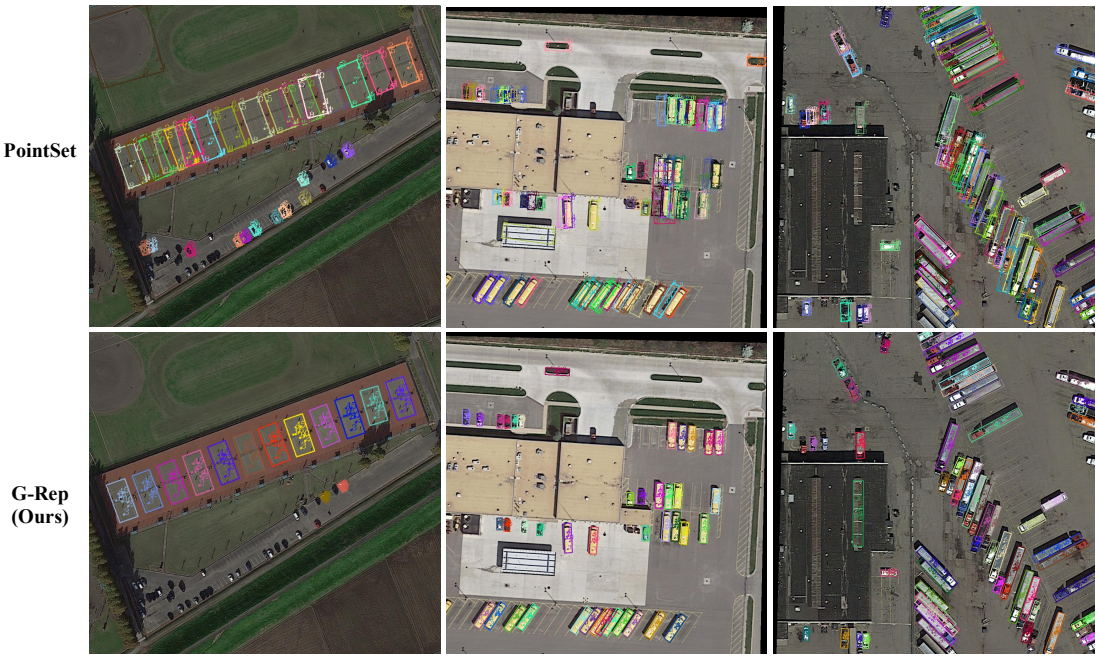


Figure 5. Comparison of the visualization results of PointSet and G-Rep on DOTA dataset.

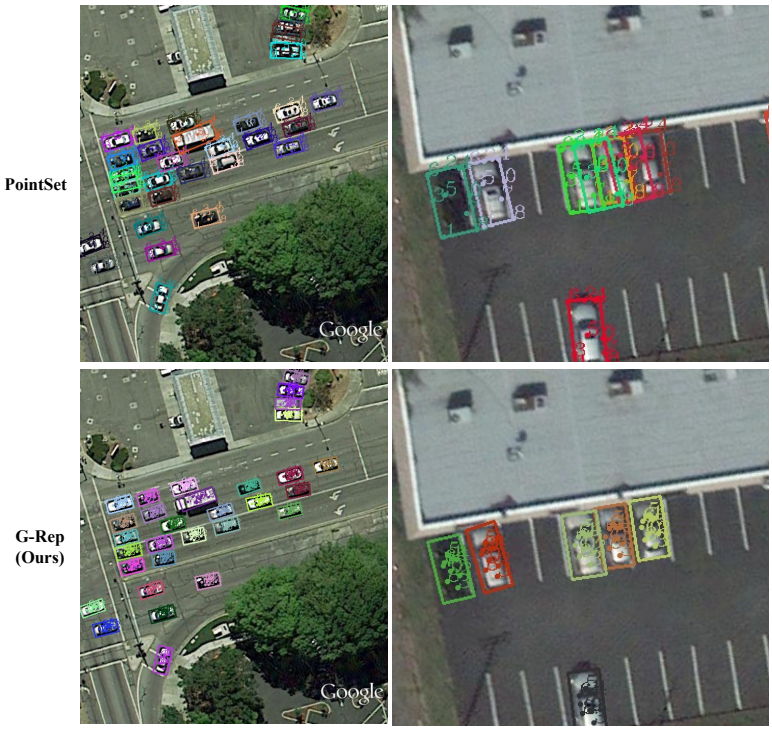


Figure 6. Comparison of the visualization results of PointSet and G-Rep on UCAS-AOD dataset.

Additionally, the angle prediction of the square-like object transformed by the isotropic Gaussian is inaccurate. This paper provides a robust object representation for arbitrary-oriented object detection, which is a generic image analysis method with no significant potential negative societal impact.