Optimization for Arbitrary-Oriented Object Detection via Representation Invariance Loss

Qi Ming, Lingjuan Miao, Zhiqiang Zhou, *Member, IEEE*, Xue Yang, and Yunpeng Dong

***Abstract*— Arbitrary-oriented objects exist widely in remote sensing images. The mainstream rotation detectors use oriented bounding boxes (OBBs) or quadrilateral bounding boxes (QBBs) to represent the rotating objects. However, these methods suffer from the representation ambiguity for oriented object definition, which leads to suboptimal regression optimization and the inconsistency between the loss metric and the localization accuracy of the predictions. In this letter, we propose a representation invariance loss (RIL) to optimize the bounding box regression for the rotating objects in the remote sensing images. RIL treats multiple representations of an oriented object as multiple equivalent local minima and hence transforms bounding box regression into an adaptive matching process with these local minima. Next, the Hungarian matching algorithm is adopted to obtain the optimal regression strategy. Besides, we propose a normalized rotation loss to alleviate the weak correlation between different variables and their unbalanced loss contribution in OBB representation. Extensive experiments on remote sensing datasets show that our method achieves consistent and substantial improvement. The code and models are available at https://github.com/ming71/RIDet to facilitate future research.**

***Index Terms*— Bounding box regression, convolutional neural networks, oriented object detection, representation ambiguity.**

I. Introduction

ORIENTED object detection has received extensive attention due to the wide range of applications. In recent years, a series of CNN-based rotation detectors has been proposed to achieve oriented object detection in the remote sensing images [1], [2]. Rotation detectors usually adopt the oriented bounding box (OBB) or the quadrilateral bounding box (QBB) to characterize the rotating objects that surround the objects more closely, which is suitable for detecting oriented remote sensing objects. For example, CFC-Net [3] obtain oriented anchors from horizontal anchors to achieve better spatial alignment with oriented objects. Xu *et al.* [2] proposed a novel framework to regress four ratios on each corresponding side to detect multiorientation objects.

However, both these detectors suffer from the representation ambiguity. Representation ambiguity means that an object can be represented in many different forms under a certain bounding box definition. For a given ground-truth (GT) object g, these ambiguous representations constitute the representation space of $\Omega(g)$, and $\Omega(g) = \{g_0, g_1, g_2, \ldots\}$. For example, QBB directly uses four vertices of the quadrilateral to represent the oriented objects. The unordered vertices have $P_4^4 = 24$ permutations, so theoretically, there are 24 quadrilateral representations for one box in Ω. In OBB, the periodicity of angle and the interchangeability between width and height also derive many equivalent representations with angles exceeding the definition boundary. Ideally, all representations in $\Omega(g)$ are equivalent local optimal solutions in the regression optimization process. These diverse representations in $\Omega(g)$ provide more easily searchable local minima for regression optimization theoretically, but the existing methods do not consider all feasible instances, only to learn a given GT representation g_0, and the rest $g_k \in \Omega(g)$ will cause a sharp increase in regression loss. The incorrect loss metric cannot truly reflect the localization quality of the predictions, which makes the regression process hard to converge and further degrades the detection performance. Besides, the contribution of different variables to the regression accuracy is unequal in OBB, and thus, careful weighting is required to balance the loss terms of different variables.

The issues have been partially explored in some previous work [4]–[9]. For instance, GWD [8] converted the deviation between the rotated boxes into a distance between Gaussian distributions and then use the Gaussian Wasserstein distance to optimize the localization of the bounding boxes. However, the proposed loss is only an approximation of the deviation between the bounding boxes, and it cannot always reflect the real position of the box. CSL [5] and DCL [6] transform the angle regression task into an angle classification task to avoid the discontinuity of regression loss. SBD [7] predicts the key edges of the vertices and uses an extra combination learning to determine the vertices of quadrilateral. SCRDet [4] proposed the smoothed L_1 loss based on intersection-over-union (IoU) to avoid the sharp increase in the loss of out-of-bounds angles. RSDet [9] learned the optimal QBB regression strategy among the four-point sorting methods. However, these methods simply attempt to eliminate the problems caused by the nonunique definitions of oriented objects but ignoring that these ambiguous representations are essentially equivalent local optima for regression optimization. Unlike the previous work, we utilize diverse representations to search for optimal regression strategy. In this case, the representation ambiguity

Manuscript received August 5, 2021; revised September 10, 2021; accepted September 21, 2021. Date of publication October 1, 2021; date of current version January 7, 2022. *(Corresponding author: Zhiqiang Zhou.)*

Qi Ming, Lingjuan Miao, Zhiqiang Zhou, and Yunpeng Dong are with the School of Automation, Beijing Institute of Technology, Beijing 100081, China (e-mail: chaser.ming@gmail.com; miaolingjuan@bit.edu.cn; zhzhzhou@bit.edu.cn; bitdyp@gmail.com).

Xue Yang is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yangxue-2019-sjtu@sjtu.edu.cn).

Digital Object Identifier 10.1109/LGRS.2021.3115110

not only no longer causes the various representation problems mentioned above but also helps to achieve better performance.

In this letter, we propose a representation invariance loss (RIL) that treats the bounding box regression as an adaptive matching process with the local minima of multiple representations and then searches for the optimal regression strategy dynamically through the Hungarian algorithm. Specifically, for QBB, RIL transforms the point regression problem into a dynamic sequential-free vertices matching process, and then, the Hungarian loss is used to improve convergence and achieve superior performance. For OBB, the RIL allows the predictions with the out-of-bounds angles and adaptively searches for the optimal regression strategy from all ambiguous representations. Besides, a normalized rotation loss is adopted to alleviate the inconsistency between different variables.

Rotation detectors trained with RIL treat the ambiguous representations as the equivalent local minima of the regression loss. These local minima provide better convergence than the unique minima defined by the GT representation. Therefore, during the training process, the regressor can dynamically select the optimal local minima in the representation space as the regression target according to the loss function.

Based on the proposed RIL, we further build two representation-invariant detectors (RIDets): RIDet-O and RIDet-Q, to prove the superiority of the proposed method. In summary, the main contribution of this letter is as follows.

1) We point out representation ambiguity in arbitrary-oriented object detection and analyzed the problems caused by multiple representations from the perspective of regression optimization.
2) The novel RIL is proposed to transform the bounding box regression task into an optimal matching process. RIL not only solves the inconsistency between the loss and localization quality caused by ambiguous representations but also utilizes these local minima to improve network convergence and achieves better performance.
3) RIL can be easily applied to the existing methods without any additional overhead. Extensive experiments on multiple rotation detectors prove the superiority of our approach.

II. Representation-Invariant Detector

We first build a baseline model and then design RIL for OBB and QBB, respectively. Next, RIL is applied to the baseline detector to construct RIDet.

A. Cascaded RetinaNet for Oriented Object Detection

We build the cascaded RetinaNet (Cas-RetinaNet) as the baseline to achieve high-efficiency-oriented object detection in remote sensing images. Cas-RetinaNet uses the oriented anchor refinement module (O-ARM) to obtain high-quality training samples. O-ARM predicts anchor offsets by convolution operations on the input feature maps. Then, the offsets are decoded into refined anchors for later bounding box regression. The overall structure of our model is shown in Fig. 1.

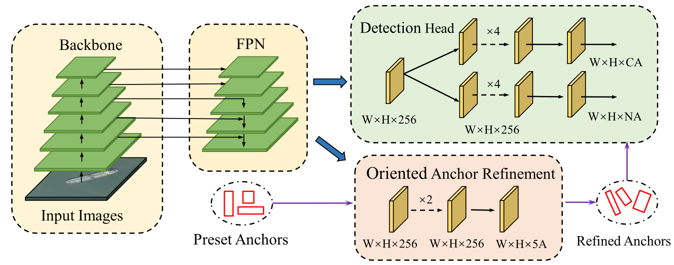


Fig. 1. Overall structure of Cas-RetinaNet. C represents the number of classes and A denotes the number of anchors. N is determined by the parameters of the representation, $N = 5$ for OBB and $N = 8$ for QBB.

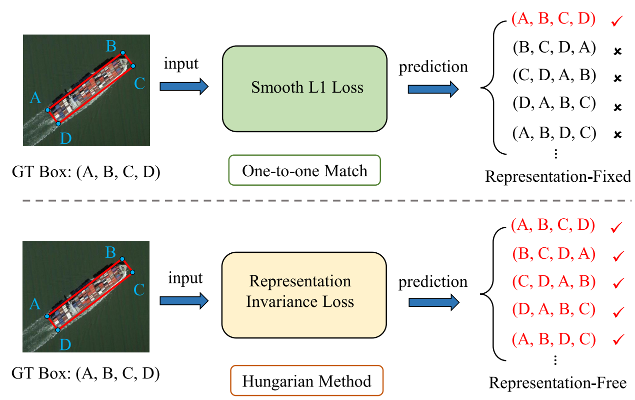


Fig. 2. Comparison of regression results of RIL and smooth-L_1 loss under QBB. RIL treats ambiguity representations as equivalent local minima, allowing the detector to regress and output sequential-free predictions.

The training loss of Cas-RetinaNet is as follows:

$$L = \frac{1}{N}\sum_i \mathrm{FL}(p_i, p_i^*) + \frac{1}{N_{p_1}}\sum_i L_{\mathrm{ref}}(\boldsymbol{t}_i, \boldsymbol{t}_i^*) + \frac{1}{N_{p_2}}\sum_i L_{\mathrm{RI}}(\boldsymbol{b}_i, \boldsymbol{t}_i^*) \quad (1)$$

in which FL denotes the focal loss [10] for classification task and L_{ref} is the smooth-L_1 loss for oriented anchor refinement. p^* and $\boldsymbol{t}^*$ are GT labels for classification, anchor refinement, and box regression, while p, $\boldsymbol{t}$, and $\boldsymbol{b}$ are the corresponding predictions. N is the number of training samples, and N_{p_1} and N_{p_2} denote the number of positives in the O-ARM stage and the detection stage, respectively. L_{RI} is the RIL, which will be elaborated in the following.

B. Quadrilateral Regression as Point Assignment

QBBs are usually convex polygons for remote sensing objects. Therefore, the sequence of vertices is redundant to define a convex quadrilateral. Even so, the current detectors still learn the one-to-one matching between the given GT representation and the predictions, as shown at the top of Fig. 2.

We treat the quadrilateral regression as a dynamic sequential-free points assignment problem, and then, the Hungarian method is adopted to solve the assignment problem. The Hungarian algorithm [11] is a combinatorial optimization algorithm to solve the bipartite graph matching problem in polynomial time. Next, we formulate the regression problem using a bipartite graph. We have a complete bipartite graph

$$G = \left(Q, Q^{*}; E\right) \tag{2}$$

in which Q and Q^{*} denote the sets of four vertices of GT polygon and prediction, respectively. Each edge in E has a nonnegative loss cost, which denotes the deviation of the predicted point from the GT vertex. We aim to find a perfect matching with a minimum total loss.

The RIL for QBB is as follows:

$$L_{\mathrm{RI}}(p, g) = \min_{\pi \in \Pi} \sum_{\substack{v \in p \\ v^{*} \in g_{\pi}}} L_{\mathrm{reg}}\left(v, v^{*}\right) \tag{3}$$

in which Π is permutations of $\{1, 2, 3, 4\}$, it represents all feasible sequence of the quadrilateral. π is a permutation in Π, and g_{π} represents a certain kind of quadrilateral. L_{reg} denotes the smooth-L_1 loss here. With the RIL defined above, the network performs sequential-free automatching among all ambiguous representations to achieve the optimal regression process.

Note that the training process is constantly changing, so the optimal regression strategy is not fixed, and it may switch between different sequences in Π during training. RIL can adaptively select the most suitable representation according to the current optimization situation and thus can achieve better convergence and performance.

The Hungarian method, which is representation invariant, no longer focuses on the sequence of points and adaptively searches for the nearby local minimum in all possible permutations to achieve better convergence quality. Moreover, each four-point set determines at most one circumscribed convex quadrilateral. Therefore, in the inference stage, the network allows multiple predictions corresponding to ambiguous representations of the GT objects, as shown in Fig. 2.

C. Out-of-Bounds Matching for OBB

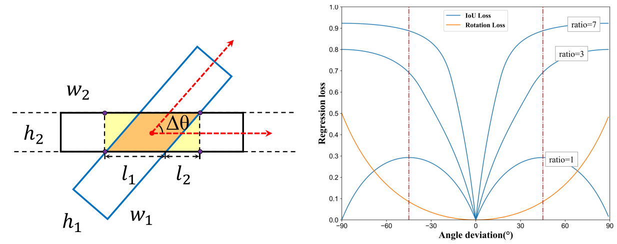
For the OBB, the boundary of the angle definition constrains the search space of regression loss, and thus, the ambiguous representations cannot be effectively utilized to optimize the regression process. Given a GT box g denoted as (cx, cy, w, h, θ), its representation space is $\Omega(g) = \{g_0, g_1, g_2, \ldots, g_i\}$, in which the instances are defined as

$$g_i = \begin{cases} \left(cx, cy, w, h, \theta + \dfrac{i\pi}{2}\right), & i \in \{2k | k \in \mathbb{Z}\} \\ \left(cx, cy, h, w, \theta + \dfrac{i\pi}{2}\right), & i \in \{2k+1 | k \in \mathbb{Z}\}. \end{cases} \tag{4}$$

Then, a Hungarian method-based out-of-bounds matching strategy is proposed to solve the boundary constraints. The RIL for OBB is as follows:

$$L_{\mathrm{RI}}(p, g) = \min_{g_i \in \Omega(g)} L_{\mathrm{nrl}}(p, g_i) \tag{5}$$

where p is the prediction and L_{nrl} denotes the normalized rotation loss for box regression, which will be elaborated next. RIL guides the prediction to match the optimal instance in $\Omega(g)$ to search for the nearby local optimum. Note that a predicted OBB represents the unique object without causing ambiguity, and thus, the predictions with out-of-bounds angles are allowed.



(a) (b)

Fig. 3. Illustration of normalized angle mapping loss in OBB. (a) Definition of normalized angle mapping loss and (b) its superiority for measuring angular deviation.

Since the periodicity of the angle leads to an infinite number of possible representations, it is impossible to exhaust all of them for the best matching. Therefore, we further propose a normalized angle mapping loss to reduce the dimensionality of $\Omega(g)$.

Given the anchor $(x_1, y_1, w_1, h_1, \theta_1)$ and GT box $(x_1, y_1, w_2, h_2, \theta_2)$ with overlapping center points as shown in Fig. 3, the orientations of the two bounding boxes form an acute angle $\Delta\theta$. This angle can be approximated by the projection of its area and the direction of the reference axis. Taking the orientation of the GT as the reference line, l_1 and l_2 are the projections of the overlapping parallelograms, and the intersection area S_0 and projected area S_1 are as follows:

$$\begin{aligned} &l_1 = \frac{h_1}{\sin\Delta\theta}, \quad l_2 = \frac{h_2}{\tan\Delta\theta} \\ &S_0 = l_1 \cdot h_2, \quad S_1 = (l_1 + l_2) \cdot h_2. \end{aligned} \tag{6}$$

Then, the IoU mapping of the angle is as

$$L_{\theta} = \frac{S_0}{S_1} - 0.5 = \frac{1}{1 + \alpha \cdot \cos\Delta\theta} - 0.5 \tag{7}$$

in which $\alpha = \min(h_1/h_2, h_2/h_1)$ and $L_{\theta} \in [0, 0.5]$. Only when the heights of the anchor and the GT box are equal, the angle deviation is equal to 0 and L_{θ} reaches the minimum value of 0. With this metric, the representation space of g can be simplified into $\Omega(g) = \{g_0, g_1\}$. In addition, as shown in the right of Fig. 3, L_{θ} is not sensitive to the aspect ratio of the bounding box itself, and it can converge steadily to objects of different shapes.

Besides, the IoU deviations caused by the same offset of different variables in (cx, cy, w, h, θ) are also different. There will be inconsistencies between the loss function and the detection metric, which has also been discussed in some previous work [8], [9]. Therefore, it is necessary to normalize the loss of variable offsets, just like L_{θ}.

We further constrain the center coordinates and shape of the OBB to balance the loss contribution between different variables. The center loss L_c is as follows:

$$L_c = \frac{\|\boldsymbol{c_g} - \boldsymbol{c_a}\|_2}{0.5 \cdot \|\boldsymbol{d}\|_2} \tag{8}$$

in which $\boldsymbol{c_g}$ and $\boldsymbol{c_a}$ are the center coordinates of GT and anchor, respectively, and $\boldsymbol{d} = (w, h)$ denotes the shape of the

GT box. During the label assignment phase, only the anchors whose center points fall within the GT box may be selected as positive samples for regression. In this way, L_c is normalized in [0, 1]. Our normalized center loss is no longer sensitive to the shape of GT box, and thus, it is conducive to the detection performance of small objects.

Moreover, we utilize the IoU of the horizontal box as the scale-independent shape loss, which is defined as follows:

$$L_s = \min_{g_i \in \Omega(g)} \text{hIoU}(p, g_i) \tag{9}$$

in which hIoU calculates the IoU with two overlapping centers without angle consideration, and it is defined as follows:

$$\text{hIoU}(b_1, b_2) = \frac{\min(w_1, w_2) \cdot \min(h_1, h_2)}{\max(w_1, w_2) \cdot \max(h_1, h_2)}. \tag{10}$$

Obviously, the range of the shape metric is also in [0, 1]. The shape loss only evaluates the similarity of the shape and does not specify the width and height, which can speed up the convergence.

Then, the normalized rotation loss for box regression is as follows:

$$L_{\text{nrl}} = L_\theta + L_c + L_s. \tag{11}$$

This novel RIL effectively uses the entire representation space to optimize the regression process. Simultaneously, the normalized rotation loss can better alleviate the inconsistency of the contribution of different variables to the loss.

III. Experiments

A. Datasets

We conduct extensive experiments on three datasets to prove the effectiveness of the proposed method, including HRSC2016 [12] and DOTA [13]. HRSC2016 is a challenging ship detection dataset, which contains 1061 images including 70 sea images and 991 sea-land images. The image sizes range from 300 × 300 to 1500 × 900. DOTA is the largest public dataset with OBB annotations for object detection in remote sensing imagery. There are 15 categories in total, and the entire dataset contains 188 282 instances. Since the images in DOTA are too large, we crop images into 800 × 800 patches with the stride set to 200.

B. Implementation Details

All images are resized to 416 × 416 or 800 × 800 for training and testing. Horizontal anchors are preset with aspect ratios of {0.5, 1, 2} and scales of $\{2^0, 2^{1/3}, 2^{2/3}\}$. The IoU thresholds used to select training samples are 0.4 in O-RAM and 0.5 in the detection stage. We use the Adam optimizer for training with the initial learning rate set to $1e^{-4}$. We train the models in 100 epochs for HRSC2016 and 24 epochs for DOTA. All models are trained on RTX 2080 Ti GPUs with batch size set to 8. Random flip, rotation, and HSV color space transformation are adopted for data augmentation.

TABLE I
Evaluation of Different Components of RIL for OBB

Backbone	matching	L_θ	L_c	mAP
ResNet-50				84.1
ResNet-50	✓			85.0$^{(+0.9)}$
ResNet-50	✓	✓		85.8$^{(+1.7)}$
ResNet-50	✓	✓	✓	**86.3**$^{(+2.2)}$

TABLE II
Performance Improvement of RIL on Different Models and Datasets

Different Models	Size	Form	Dataset	with RIL	mAP
RetinaNet [10]	512	OBB	HRSC	×	82.2
				✓	**84.5** (+2.3)
Cas-RetinaNet [10]	416	QBB	HRSC	×	84.3
				✓	**86.3** (+2.0)
CFC-Net [3]	416	OBB	HRSC	×	85.2
				✓	**87.3** (+1.9)
DAL [14]	384	OBB	HRSC	×	87.1
				✓	**88.5** (+1.4)
S^2A-Net [15]	416	OBB	HRSC	×	88.3
				✓	**89.1** (+0.8)
Cas-RetinaNet [10]	800	QBB	DOTA	×	70.5
				✓	**72.8** (+2.3)
Cas-RetinaNet [10]	800	OBB	DOTA	×	72.9
				✓	**74.7** (+2.2)
S^2A-Net [15]	800	OBB	DOTA	×	76.4
				✓	**77.6** (+1.2)

C. Ablation Study

1) Evaluation of Normalized Rotation Loss for OBB: Experiments conducted on the HRSC2016 datasets confirmed the effectiveness of RIL and normalized rotation loss. The regression head outputs the oriented rectangular bounding box in the experiments in RIDet-O here. Note that we do not use complex data augmentation. The experimental results are shown in Table I.

RIDet-O achieves the mAP of 86.3% on the HRSC2016 dataset, which is improved by 2.2% compared to the baseline. The optimal matching strategy allows the out-of-bounds predictions to search for the optimal local minima for regression, which is conducive to network convergence. Thus, it improves the mAP by 0.9%. The normalized rotation loss further achieves the increase of 0.8%, which balances the contribution of different variables and alleviates the inconsistency between loss and detection metric. Finally, selecting the anchors with the center point inside the GT box for training improves by 0.5. These high-quality positives not only provide better priori about the position of the center points but also help constrain the range of center loss.

2) Evaluation on Different Models: We further conduct experiments on different detectors to verify the generalization of RIL. Experimental results in Table II show that RIL improves the mAP by 2.3% on RetinaNet with OBB and 2.0% on Cas-RetinaNet with QBB on HRSC2016. It proves that the strategy of transforming the bounding box regression task

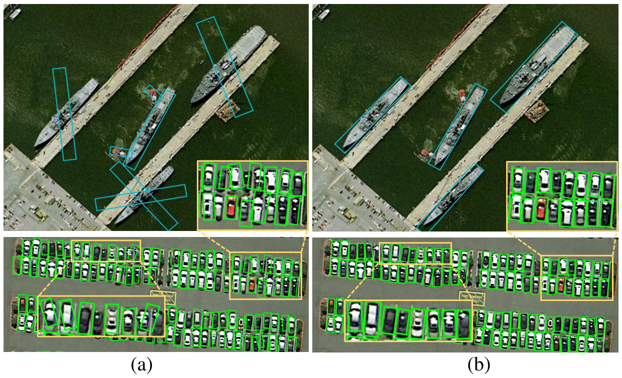


Fig. 4. Comparison of detection results between detectors with (a) smmoth-L_1 loss and (b) RIL. (Top) Detections under OBB representation. (Bottom) Results under the QBB representation.

TABLE III
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE HRSC2016 DATASET

Methods	Backbone	Size	mAP
Two-stage:			
RoI Trans. [16]	ResNet101	512×800	86.20
Gliding Vertex [2]	ResNet101	512×800	88.20
DCL [6]	ResNet101	800×800	89.46
Single-stage:			
RetinaNet [10]	ResNet50	416×416	80.81
RSDet [9]	ResNet50	800×800	86.50
R^3Det [17]	ResNet101	800×800	89.26
RIDet-O (Ours)	ResNet18	416×416	89.25
RIDet-Q (Ours)	ResNet101	800×800	89.10
RIDet-O (Ours)	ResNet101	800×800	**89.63**

into the optimal matching process is beneficial in different representations.

S^2A-Net [15] is an advanced rotation detector that achieves 76.4% mAP on the DOTA. Integrating RIL to S^2A-Net improves performance by 1.2% to achieve the mAP of 77.6%, which further proves the generalization and superiority of our method. DAL [14] and CFC-Net [3] are also currently advanced detectors that achieve accurate remote sensing object detection. However, these detectors will still be confused by ambiguous representations of GT. After training with RIL, as shown in Table II, the detection performance is further improved by 1.9% and 1.4% on DAL and CFC-Net on HRSC2016, respectively. It indicates that our method can obtain consistent performance gains by solving representation ambiguity.

The visualization results in Fig. 4 show that the convergence of the angle regression is slow with the normal smooth-L_1 loss, which only learns a specific representation. RIL treats all feasible representations as equivalent local optimums. Therefore, better performance can be achieved under the same number of iterations.

D. Comparisons With Other Methods

We compared the proposed method with other state-of-the-art detectors on the HRSC2016 dataset. As shown in Table III, RIDet outperforms other single-stage detectors and even some recent advanced two-stage detectors such as DCL [6] and RoI transformer [16] in remote sensing object detection. RIDet-Q and RIDet-O achieve the mAP of 89.10% and 89.63%, respectively. Especially when the images are resized to 416 × 416 with lightweight ResNet-18 as the backbone, RIDet-O still achieves the competitive performance of 89.25%.

IV. CONCLUSION

In this letter, the RIL is proposed to solve the representation ambiguity and accelerate network convergence in oriented object detection in remote sensing images. RIL treats multiple representations as equivalent local minima and then transforms the bounding box regression task into the optimal matching between predictions and these local minima. Our method can be embedded into existing detectors to achieve performance improvement without introducing additional inference overhead. Extensive experiments on multiple datasets proved the superiority of our method.

REFERENCES

[1] Q. Ming, L. Miao, Z. Zhou, J. Song, and X. Yang, "Sparse label assignment for oriented object detection in aerial images," *Remote Sens.*, vol. 13, no. 14, p. 2664, Jul. 2021.

[2] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[3] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical featurecapturing network for arbitrary-oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 16, 2021, doi: 10.1109/TGRS.2021.3095186.

[4] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8232–8241.

[5] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 677–694.

[6] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15819–15829.

[7] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, "Omnidirectional scene text detection with sequential-free box discretization," in *Proc. IJCAI*, 2019, pp. 1–8.

[8] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian wasserstein distance loss," 2021, *arXiv:2101.11952*. [Online]. Available: http://arxiv.org/abs/2101.11952

[9] W. Qian, X. Yang, S. Peng, Y. Guo, and J. Yan, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–11.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[11] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[12] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, vol. 2, 2017, pp. 324–331.

[13] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[14] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2355–2363.

[15] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, early access, Mar. 12, 2021, doi: 10.1109/TGRS.2021.3062048.

[16] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2849–2858.

[17] X. Yang, Q. Liu, J. Yan, A. Li, Z. Zhang, and G. Yu, "R^3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–9.