

Dense Label Encoding for Boundary Discontinuity Free Rotation Detection

Xue Yang¹, Liping Hou², Yue Zhou¹, Wentao Wang¹, Junchi Yan^{1,*}

¹Shanghai Jiao Tong University, Shanghai, China

²University of Chinese Academy of Sciences, Beijing, China

{yangxue-2019-sjtu, sjtu_zy, wwt117, yanjunchi}@sjtu.edu.cn

houliping17@mails.ucas.ac.cn

Abstract

Rotation detection serves as a fundamental building block in many visual applications involving aerial image, scene text, and face etc. Differing from the dominant regression-based approaches for orientation estimation, this paper explores a relatively less-studied methodology based on classification. The hope is to inherently dismiss the boundary discontinuity issue as encountered by the regression-based detectors. We propose new techniques to push its frontier in two aspects: i) new encoding mechanism: the design of two Densely Coded Labels (DCL) for angle classification, to replace the Sparsely Coded Label (SCL) in existing classification-based detectors, leading to three times training speed increase as empirically observed across benchmarks, further with notable improvement in detection accuracy; ii) loss re-weighting: we propose Angle Distance and Aspect Ratio Sensitive Weighting (ADARSW), which improves the detection accuracy especially for square-like objects, by making DCL-based detectors sensitive to angular distance and object’s aspect ratio. Extensive experiments and visual analysis on large-scale public datasets for aerial images i.e. DOTA, UCAS-AOD, HRSC2016, as well as scene text dataset ICDAR2015 and MLT, show the effectiveness of our approach. The source code is available at https://github.com/Thinklab-SJTU/DCL_RetinaNet_Tensorflow and is also integrated in our open source rotation detection benchmark: <https://github.com/yangxue0827/RotateDetection>.

1. Introduction

Rotation detection has recently attracted increasing attention for their important utility across different scenarios, including aerial images, scene text, and faces etc., which is relatively less studied compared with the vast literature

in horizontal object detectors that do not estimate the exact rotation while only output the horizontal bounding box.

Many mainstream rotation detection algorithms (including aerial image [54, 18, 55, 2, 7, 59, 53, 37, 56, 58, 57], scene text [65, 30, 17, 33, 25, 24] and face [41, 16, 40]) are derived based on the vanilla detection algorithms [12, 39, 38, 6, 26, 27]. Among them, the rotation detection algorithm based on five parameters ($[x, y, h, w, \theta]$) dominates. Similar to the coordinate regression method in horizontal detection, angle parameter is also predicted by regression. Although gratifying results have been achieved, there are still some fundamental flaws in the orientation estimation based on regression. Angle prediction based on regression often introduces boundary discontinuity [59, 37, 56, 57], mainly including periodicity of angle (PoA) and exchangeability of edges (EoE). The main reason for the former is the bounded periodic nature of the angle parameter, while the latter is related to the definition of the bounding box. In general, the root cause is that the ideal predictions are beyond the defined range. Due to the sharp increase in the loss at the boundary, the regression form of the model at the boundary and non-boundary can not be consistent. Therefore, the model has to predict the angle in a more complicated form at the boundary, which increases the burden of the model and also increases the difficulty of prediction at the boundary. This is fatal for the rotation object detection that require high precision, especially for objects with large aspect ratios.

Most existing works aim to eliminate the sudden loss increase by adding constraints on the loss function or changing the way of calculation, such as IoU Smooth L1 Loss [59] and Modulated Loss [37], as shown in Table 1. The advantage is that they can borrow the well developed baselines from horizontal object detectors and reuse the related techniques to boost the detection performance. However, such ad-hoc techniques applied on the regression-based detectors cannot guarantee the full dismiss of boundary discontinuity behavior. To give a more elegant and effective solution, the more recent work called Circular Smooth Label

*Corresponding author is Junchi Yan.

Method	Box Def.	Angle Pred.	PoA	EoE	SLP	Speed	mAP _{50:95}
RetinaNet	Long-Side Def.	Reg.	✓	✓	✓	-	31.49
RetinaNet	OpenCV Def.	Reg.	✓	✓	✗	~1x	34.50
IoU-Smooth L1 Loss [59]	OpenCV Def.	Reg.	✗	✗	✗	~1x	36.23
Modulated Loss [37]	OpenCV Def.	Reg.	✗	✗	✗	~1x	34.61
CSL [56]	Long-Side Def.	Cls.: SCL	✗	✗	✓	~ $\frac{1}{3}$ x	35.04
DCL (BCL)	Long-Side Def.	Cls.: DCL	✗	✗	✗	~1x	36.71

Table 1: Comparison between different solutions for periodicity of angle (PoA), exchangeability of edges (EoE) and square-like problem (SLP) on DOTA val set. The ✓ indicates that the method suffers the corresponding problem.

(CSL) [56, 57] argues to apply angle classification instead of regression to address PoA. Then, CSL-based method combines with the long-side definition (five-parameter with 180° angular range) of bounding box to further tackle the EoE problem. The use of ‘CSL+ 180° ’¹ leads to a natural solution to get rid of the boundary discontinuity issue. As angle classification based detectors are still in its early stage, there are still many limitations e.g. very heavy prediction layer and difficulty in handling square-like objects. The former problem is initially explored by the study [57]. This paper is one of the classification-based endeavors in pushing forward this frontier, with two specific technical advancements as follows.

First, we design two Densely Coded Labels (DCL) in contrast to the Sparsely Coded Label (SCL, including CSL, One-Hot encoding), which has empirically led to notable training time reduction with meanwhile improved detection accuracy. To make DCL as sensitive to angle distance as CSL, we calculate the decimal difference between the predicted angle and the angle label as a angle distance aware weight. However, this weight will reintroduce the PoA problem, and we find that the long-side definition method is not conducive to the training of square-like objects. Based on the findings of the above two issues, we design Angle Distance and Aspect Ratio Sensitive Weighting (ADARSW). ADARSW can eliminate the PoA and can be adaptively adjusted according to the object’s aspect ratio, which can greatly reduce the burden of model training. Combing ‘DCL+ 180° +ADARSW’ as a whole, extensive experiments and visual analysis on different datasets and detectors prove that DCL-based method can be a better baseline choice than the angle regression-based and CSL-based methods. In summary, our work makes the following contributions:

1) To improve the robustness especially for objects with small aspect ratio, we propose Angle Distance and Aspect Ratio Sensitive Weighting (ADARSW), which further improves accuracy by making our proposed DCL-based detector sensitive to angular distance and object’s aspect ratio. In contrast, the existing CSL-based detector suffers from its

¹Unless otherwise specified, the CSL-based method mentioned in this paper is based on the long-side definition method.

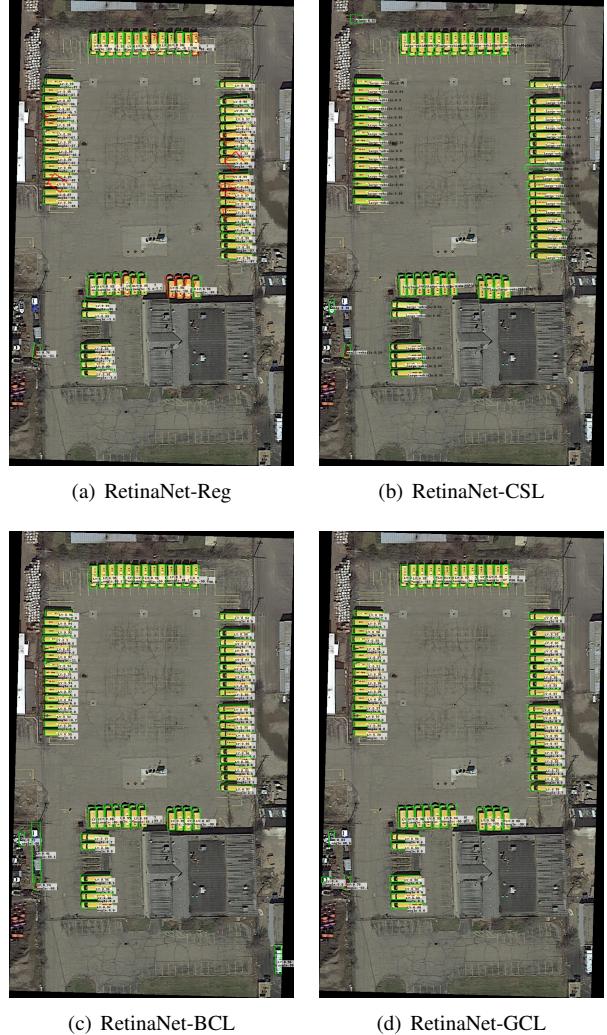


Figure 1: Comparison of four rotation detectors in the boundary case. Red bounding boxes indicate some bad detection cases (zoom in for better view). CSL and DCL based methods (including BCL and GCL) are totally boundary discontinuity free.

long-side definition for detecting square-like objects.

2) We compare the impact of two classic Densely Coded Labels (DCL) by introducing them to the angle classification task for potential speedup, namely Binary Coded Label (BCL) and Gray Coded Label (GCL), which are more compact than existing CSL. Though GCL has been partly studied in [57], while this paper presents a more thorough investigation especially for BCL. We empirically show that DCL, especially BCL can lead to notable training speed boost (about three times) as well as detection accuracy.

3) Extensive experiments and visual analysis on different datasets and detectors prove the efficacy of our techniques. It outperforms state-of-the-art CSL-based detector [56] by:

77.37% vs. 76.17% on DOTA dataset.

2. Related Work

Horizontal Object Detection Object detection is a fundamental task in the field of computer vision, and it has developed rapidly in recent years. Classic convolutional neural networks (CNN) based detectors can mainly be divided into two categories: two-stage object detectors [12, 39, 6, 26] and single-stage object detectors [38, 29, 27]. Two-stage methods are based on region proposal and have achieved promising results on some benchmarks, whereas single-stage approaches simplify detection as a regression problem to maintain a faster speed. Compared to anchor-based methods, many anchor-free based methods have become extremely popular in recent years. CornerNet [20], FCOS [42], CenterNet [8] and ExtremeNet [66] attempt to predict some keypoints of objects such as corners or extreme points, which are then grouped into bounding boxes. What is even more surprising is that DETR [3] has constructed a new object detection paradigm based on transformer [43], which achieves anchor free and non maximum suppression (NMS) free at the same time. However, horizontal object detectors cannot provide accurate orientation and scale information, so they cannot be directly applied to some specific scenes, such as aerial images and scene text.

Rotation Object Detection Rotation detectors are mainly applied in the aerial images and scene text. Recent advances in multi-oriented object detection are mainly driven by adaption of classical object detectors using rotated bounding boxes or quadrangles to represent multi-oriented objects. In the aerial imagery scene, ICN [2], ROI-Transformer [7], CAD-Net[61], SCRDet [59], R³Det [53], and CSL [56] achieve promising performance. Gliding Vertex [51] and RSDet [37] achieve more accurate object detection through quadrilateral regression prediction. RRPN [33], TextBoxes++ [24] and RRD [25] and FOTS [30] are some advanced methods for scene text detection. However, most of the above regression-based arbitrary-oriented methods focus on the prediction of angle using regression yet ignore the boundary discontinuity. Although SCRDet and RSDet solve the boundary discontinuity from the perspective of the loss function, they are not truly boundary discontinuity free methods. CSL-based detector is a new boundary discontinuity free rotation detector, which transforms angular prediction from a regression problem to a classification problem. However, CSL-based method needs to face two obvious shortcomings: heavy prediction layer and unfriendly to square-like objects. In this paper, we aim to solve the above problems from the two perspectives of encoding form and loss function weight.

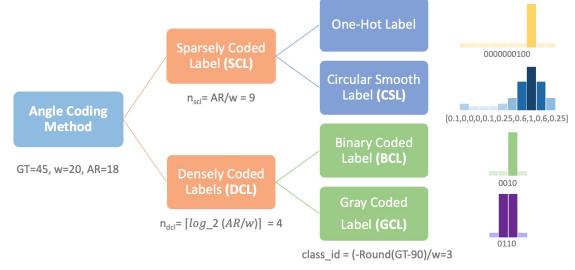


Figure 2: The relationship between the various angle encoding methods.

3. Mitigating Boundary Discontinuity by Classification

The boundary discontinuity [59, 56, 57] usually refers to the sharp loss increase of the regression-based rotation detector at the boundary situation, which makes the model unable to perform regression prediction in the same ideal and simple form at the boundary as at the non-boundary. The reasons for the boundary discontinuity are related to the definition of the object bounding box. The work [56] summarizes several commonly used bounding box definitions and the causes of boundary discontinuity corresponding to these methods. Details are as follows:

1) Five-parameter method with 90° angular range (OpenCV definition method): mainly including periodicity of angular (PoA) and exchangeability of edges (EoE).

2) Five-parameter method with 180° angular range (long-side definition method): mainly suffer from periodicity of angular.

3) Eight-parameter method: sorting of the four corners.

The main cause of boundary discontinuity based on regression methods is that the ideal predictions are beyond the defined range. As a consequence, the detection result at the boundary of the detector that has not solved the boundary discontinuity often shows inaccurate angle prediction, as shown in the red bounding boxes in Figure 1(a). Different solutions are proposed, such as constraining the loss function [37, 59], changing the angle prediction form [56, 57], etc. Although some progress has been made, these methods have their own weaknesses. Many current methods have not notice or eliminate boundary discontinuity from method design. A true boundary discontinuity free detector will provide a more robust high-performance baseline, so designing such detector is a valuable research direction.

4. Proposed Method

In this section, we first give a retrospection to the recent classification-based rotation detectors namely Circular Smooth Label (CSL) [56], pointing out its limitation in achieving a cost-efficient detector. Then we propose our so-called Densely Coded Label (DCL) technique to improve

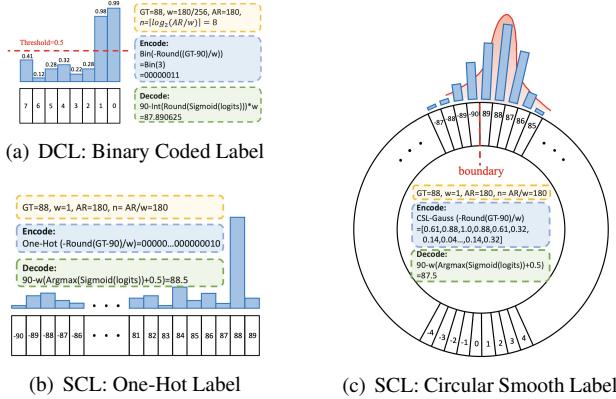


Figure 3: Examples of encoding and decoding process of One-Hot, CSL-Gaussian and BCL for angle prediction.

the efficiency and also develop the Angle Distance and Aspect Ratio Sensitive Weighting technique, to improve its sensitivity to small aspect ratio objects. Figure 2 shows the relationship between various angle encoding methods.

4.1. Rethinking on Sparsely Coded Label Encoding

Instead of using the regression-based loss function, the Circular Smooth Label (CSL) [56] detectors have been recently proposed which transform rotation detection to a classification task such that the boundary issue naturally disappear. The CSL-based detectors adopt the so-called Sparsely Coded Label (SCL) encoding technique [56] to discretize the angle into a finite number of intervals, and then predicts a discrete angle by classification². Equation 1 describes the angle prediction process in CSL:

$$\begin{aligned} \text{Encode: } & \text{CSL}(-\text{Round}((\theta_{gt} - 90)/\omega)) \\ \text{Decode: } & 90 - \omega(\text{Argmax}(\text{Sigmoid}(logits)) + 0.5) \end{aligned} \quad (1)$$

where θ_{gt} presents the angle decimal label, $\omega = AR/C_\theta$ indicates the angle discretization granularity. AR and C_θ represents angle range (the default value is 180) and the number of angle categories, respectively.

Figure 3(b) and Figure 3(c) show examples of encoding and decoding process of One-Hot and CSL-Gaussian for angle prediction, both of which are embodiments of the SCL encoding. Although the combined techniques ‘CSL+180°’ can well eliminate the impact of boundary discontinuity, it also brings two thorny problems that hurt the efficiency and efficacy: i) a very heavy prediction layer and ii) unfriendliness to objects with small aspect ratio.

Thick prediction layer. Equation 2 compares the pre-

²CSL can only solve the PoA, and the EoE problem can be solved by the 180° angular definition method (recall the discussion in Section ??).

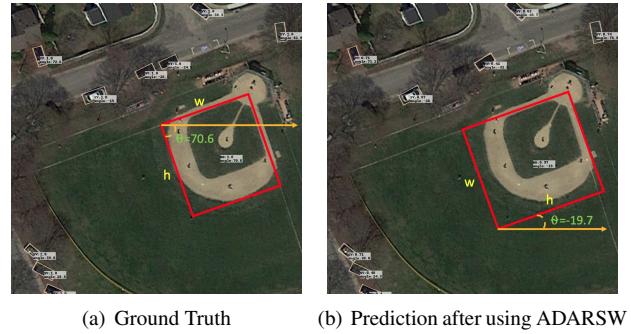


Figure 4: Illustration for the limitation of the long-edge definition method on the square-like objects. High IoU, but large training loss due to angle difference.

Base Model	ω	GFlops	Δ GFlops	Params (M)	Δ Params	Training Time
RetinaNet-Reg	-	139.35	-	36.97	-	-
RetinaNet-CSL	1	254.96	+82.96%	45.63	+23.42%	~3x
RetinaNet-BCL	1	143.87	+3.24%	37.31	+0.92%	~1x
RetinaNet-GCL	1	143.87	+3.24%	37.31	+0.92%	~1x

Table 2: Comparison of GFlops and Param over rotation detectors, under the same setting and hyperparameters.

diction layer thickness of three angle prediction methods:

$$\begin{aligned} \text{Th}_{\text{reg.}} &= A \\ \text{Th}_{\text{onehot}} &= \text{Th}_{\text{cls}} = A \times AR/\omega \end{aligned} \quad (2)$$

where A indicates the number of anchors.

Taking $A = 21$, $AR = 180$, $w = 1$ as an example, the thickness of the prediction layer required by CSL and One-Hot is 3,780, while the thickness of regression-based approach is only 21. From the perspective of GFlops and Param, detectors based on CSL have increased by about 82.96% and 23.42% respectively. In addition, the training time of RetinaNet-CSL is three times longer than regression-based detector. The detailed statistical results are shown in Table 2.

Unfriendliness to small aspect ratio objects. Five-parameter method with 180° angular range is a widely used rectangular definition (long-side definition method, $[x, y, h, w, \theta]$) without EoE problem. The θ is determined by the long side (h) of the rectangle and x-axis. However, this definition method is not suitable for square-like box and will suffer a special problem, as shown in Figure 4. Figure 4(a)-4(b) are ground truth and candidate prediction bounding box with an aspect ratio close to 1, and their angles are 70.6° and -19.7°, respectively. By calculating the Intersection-over-Union (IoU) and regression (e.g. smooth 11) or classification (e.g. CSL) loss of these two boxes, we find that the IoU between them is close to 1, but a relatively large loss value is produced. This loss value mainly comes from the angle parameter. Therefore, the prediction results

Decimal Number	0	1	2	3	4	5	6	7
Binary Coded Label	000	001	010	011	100	101	110	111
Gray Coded Label	000	001	011	010	110	111	101	100

Table 3: The three-digit binary code and gray code corresponding to the decimal number.

in Figure 4(b) are not allowed by the model, which is too harsh and increases the model’s difficulty in predicting objects with small aspect ratio. In fact, this phenomenon becomes less noticeable as the aspect ratio increases. For the definition of square-like objects, using the OpenCV definition method with a period of 90° can effectively avoid this problem, but it will introduce EoE problems.

4.2. Densely Coded Label

The introduction of excessive amount of parameters and the unfriendliness to small aspect ratio objects seriously hurts the applicability of classification based rotation detectors. In this section, we will solve the above problems from the two perspectives of encoding form and loss function weight.

Binary Coded Label (BCL) [14] and Gray Coded Label (GCL) [10, 57] are two Densely Coded Label (DCL) methods commonly used in the field of electronic communication. Their advantage is that they can represent a larger range of values with less coding length. Thus, they can effectively solve the problem of excessively long coding length in CSL and One-Hot based methods. The prediction layer thickness of BCL and GCL based methods are calculated as follows:

$$\underbrace{\text{Th}_{\text{bcl}} = \text{Th}_{\text{gcl}}}_{\text{TH}_{\text{dcl}}} = A \times \lceil \log_2(AR/\omega) \rceil \quad (3)$$

Under the same setting of $A = 21$, $AR = 180$, $w = 1$, the coding length of DCL is only 168. According to Table 2, GFlops and Param only increase by **3.24%** and **0.92%**. The training time is almost the same as the regression-based method. The performance of DCL-based method is not drop but rises, and the specific performance comparison can quickly refer to Figure 1(b)-1(d) and Table 4.

Algorithm 1-2 describe the pseudo codes that generate all n -bit gray and binary coded labels. BCL processes the angle by binarization to obtain a string of codes represented by multiple ‘0’ and ‘1’. Although the coding length is greatly reduced, there may be huge changes in the coding results between adjacent values, that is, there is no classification tolerance mentioned in the CSL. For example, the three-bit binary coding results of the values ‘3’ and ‘4’ are ‘011’ and ‘100’, respectively. It can be seen that all three positions have changed, resulting in a very large difference in the loss value of the two angle predictions. GCL can solve this problem [57]. In the encoding of a group of num-

Algorithm 1 Binary Coded Label (BCL)

Input: angle range AR , discretization granularity ω .
Output: A list L containing all binary coded labels;
 $L = []$, $n = \lceil \log_2(AR/\omega) \rceil$ # coding length;
for i in AR **do**
 $bcl = Bin(i, n)$ # generate n -bit binary code;
 $L.append(bcl)$;
end for
Return L ;

Algorithm 2 Gray Coded Label (GCL)

Input: coding length $n = \lceil \log_2(AR/\omega) \rceil$ (AR means angle range, ω represents angle discretization granularity).
Output: A list L containing all gray coded labels.
if $n = 1$ **then**
 Return ['0', '1'];
else
 $L1 = GCL(n - 1)$ # recursive call;
 $L2, L3, L4 = L1.reverse(), [], []$;
 for $s1$ in $L1$ **do**
 $L3.append('0' + s1)$;
 end for
 for $s2$ in $L2$ **do**
 $L3.append('1' + s2)$;
 end for
 $L = concat(L3, L4)$;
end if
Return L ;

bers, if any two adjacent codes differ only by one binary number, then this kind of encoding is called Gray Code. Due to only one digit is different between the maximum number and the minimum number, it is also called Cyclic Code. The coding results of ‘3’ and ‘4’ in the GCL method are ‘010’ and ‘110’. Table 3 compares the coding results of BCL and GCL. The shortcomings of GCL are also obvious. Although the encoding forms between adjacent angles are not much different, which makes GCL also have a certain classification tolerance, the encoding differences of angles with large differences are not very significant, such as ‘1 (001)’ and ‘6 (101)’. In summary, these two methods are agnostic or partially agnostic to the angle distance.

In the DCL-based method, only the number of categories is a power of 2 to ensure that each coding corresponds to a valid angle. For example, if the 180 degree range is divided into $2^8 = 256$ categories, then the range of each division interval is $\omega = 180/256 = 0.703125^\circ$. According to the $Max(error) = \omega/2$ and $E(error) = \omega/4$ proposed in work [56], the maximum and expected accuracy error are only 0.3515625° and 0.17578125° , respectively whose influence on final detecton accuracy can be negligible. How-

ever, the above condition is not necessary. We find that even with some redundant invalid codes, there is no significant drop in final performance. Equation 4 specifies the encoding and decoding process of DCL (take BCL as an example):

$$\begin{aligned} \text{Encode: } & \text{Bin}(-\text{Round}((\theta_{gt} - 90)/\omega)) \\ \text{Decode: } & 90 - \omega \text{Int}(\text{Round}(\text{Sigmoid}(logits))) \end{aligned} \quad (4)$$

Figure 3(a) gives an example which also takes BCL as an embodiment. In the decoding process, the threshold for converting the predicted logits into binary coding is 0.5.

4.3. Angle Distance and Aspect Ratio Sensitive Weighting

To make the model sensitive to the distance of the angle, we calculate the decimal difference between the predicted angle and the angle label as an angle distance aware weight. The specific formula is designated as follows:

$$\begin{aligned} W(\Delta\theta) &= \log(|\Delta\theta| + 1) = \log(|\theta_{gt} - \theta_{pred}| + 1) \\ \theta_{pred} &= \text{Decode}_{dcl}(logits) \end{aligned} \quad (5)$$

where $\Delta\theta$ denotes the decimal difference between the predicted angle (θ_{pred}) and the angle label (θ_{gt}). $logits$ represents the prediction vector of the angle.

However, the above-mentioned angle distance aware weight reintroduces the PoA problem. Take $\theta_{gt} = -90$, $\theta_{pred} = 89$ as an example, although the angle of the two bounding boxes are very close, a very large weight is calculated. Therefore, we consider adding a periodic trigonometric function to solve this problem. As discussed in Section 4.1, the square-like object is not suitable to be defined by the long-side definition method. We propose an Angle Distance and Aspect Ratio Sensitive Weighting as ADARSW, as shown in Equation 6:

$$W_{ADARSW}(\Delta\theta) = |\sin(\alpha(\Delta\theta))| = |\sin(\alpha(\theta_{gt} - \theta_{pred}))|$$

$$\alpha = \begin{cases} 1, & (h_{gt}/w_{gt}) > r \\ 2, & \text{otherwise} \end{cases} \quad (6)$$

where h_{gt} and w_{gt} are the long and short sides of ground truth. r is the aspect ratio threshold, the default value is 1.5.

When the object has a certain aspect ratio, the period of $|\sin(\alpha(\theta_{gt} - \theta_{pred}))|$ is set to 180° ($\alpha = 1$), and when the object is square-like, the period becomes 90° ($\alpha = 2$). Thus, the model can solve the PoA and can flexibly adjust the training strategy for different aspect ratio objects. The DCL-based angle classification loss is as follows:

$$\begin{aligned} L_{dcl}(\theta_{gt}, logits) &= \text{FL}(\text{Encode}_{dcl}(\theta_{gt}), logits) \\ &\times W_{ADARSW}(\Delta\theta) \end{aligned} \quad (7)$$

where FL indicates focal loss [27].

4.4. Loss Function

For RetinaNet-based rotation detection, we use five parameters (x, y, h, w, θ) to represent arbitrary-oriented rectangle. Ranging in $[-\pi/2, \pi/2]$, the θ is determined by the long side (h) of the rectangle and x-axis. For DCL based method, it calls for an additional angle classification prediction layer. The other four parameters are predicted by regression, the regression formula is as follows:

$$\begin{aligned} t_x &= (x - x_a)/w_a, t_y = (y - y_a)/h_a \\ t_w &= \log(w/w_a), t_h = \log(h/h_a) \\ t'_x &= (x' - x_a)/w_a, t'_y = (y' - y_a)/h_a \\ t'_w &= \log(w'/w_a), t'_h = \log(h'/h_a) \end{aligned} \quad (8)$$

where x, y, h, w denote the box's center coordinates, height and width respectively. Variables x, x_a, x' are for the ground-truth box, anchor box, and predicted box, respectively (likewise for y, w, h).

The multi-task loss is used which is defined as follows:

$$\begin{aligned} L &= \frac{\lambda_1}{N} \sum_{n=1}^N obj_n \sum_{j \in \{x, y, h, w\}} L_{reg}(v'_{nj}, v_{nj}) \\ &+ \frac{\lambda_2}{N} \sum_{n=1}^N obj_n L_{dcl}(\theta_{gt}, logits) + \frac{\lambda_3}{N} \sum_{n=1}^N L_{cls}(p_n, t_n) \end{aligned} \quad (9)$$

where N indicates the number of anchors, obj_n is a binary value ($obj_n = 1$ for foreground and $obj_n = 0$ for background, no regression for background). v'_{nj} denotes the predicted offset vectors, v_{nj} is the targets vector of ground truth. t_n represents the label of object, p_n is the probability distribution of various classes calculated by sigmoid function. The hyper-parameter $\lambda_1, \lambda_2, \lambda_3$ control the trade-off and are set to $\{1, 0.5, 0.1\}$ by default. The classification loss L_{cls} is focal loss [27]. The regression loss L_{reg} is smooth L1 loss as used in [12].

5. Experiments

We use Tensorflow [1] to implement the proposed methods on a server with GeForce RTX 2080 Ti and 11G memory. The experiments in this article are initialized by ResNet50 [13] by default unless otherwise specified. We perform experiments on both aerial benchmarks and scene text benchmarks to verify the generality of our techniques. Weight decay and momentum are set 0.0001 and 0.9, respectively. We employ MomentumOptimizer over 4 GPUs with a total of 4 images per minibatch (1 images per GPU).

5.1. Datasets and Protocols

DOTA [49] is comprised of 2,806 large aerial images from different sensors and platforms. Objects in DOTA

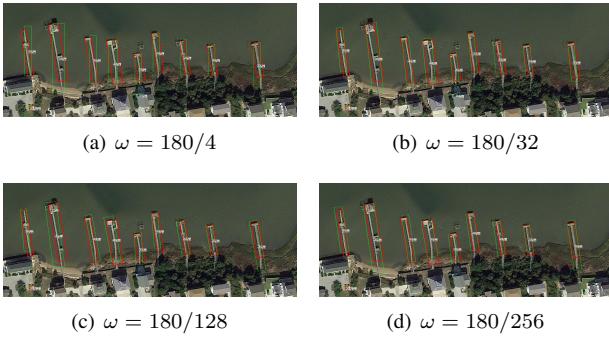


Figure 5: Visualization of detection results under different angle discretization granularity ω . The red and green bounding box indicate ground truth and prediction.

exhibit a wide variety of scales, orientations, and shapes. These images are then annotated by experts using 15 object categories. The short names for categories are defined as (abbreviation-full name): PL-Plane, BD-Baseball diamond, BR-Bridge, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, SH-Ship, TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, HA-Harbor, SP-Swimming pool, and HC-Helicopter. The fully annotated DOTA benchmark contains 188,282 instances, each of which is labeled by an arbitrary quadrilateral. Half of the original images are randomly selected as the training set, 1/6 as the validation set, and 1/3 as the testing set. We divide the images into 600×600 subimages with an overlap of 150 pixels and scale it to 800×800 . With all these processes, we obtain about 20,000 training and 7,000 validation patches.

UCAS-AOD [67] contains 1,510 aerial images of approximately $659 \times 1,280$ pixels, with two categories of 14,596 instances in total. In line with [2, 49], we randomly select 1,110 for training and 400 for testing.

ICDAR2015 [19] is the Challenge 4 of ICDAR 2015 Robust Reading Competition, which is commonly used for oriented scene text detection and spotting. This dataset includes 1,000 training images and 500 testing images.

ICDAR 2017 MLT [34] is a multi-lingual text dataset, which includes 7,200 training images, 1,800 validation images and 9,000 testing images. The dataset is composed of complete scene images in 9 languages, and text regions in this dataset can be in arbitrary orientations, being more diverse and challenging.

HRSC2016 [32] contains images from two scenarios including ships on sea and ships close inshore. All images are collected from six famous harbors. The training, validation and test set include 436, 181 and 444 images, respectively.

All the used datasets are trained by 20 epochs in total, and learning rate is reduced tenfold at 12 epochs and 16

Method	BR	SV	LV	SH	HA	5-mAP ₅₀	mAP ₅₀
RetinaNet-Reg	38.31	60.48	49.77	68.29	51.28	53.63	64.17
RetinaNet-CSL	40.55	66.77	51.50	73.60	53.76	57.24 (+3.61)	65.69 (+1.52)
RetinaNet-BCL	41.58	67.98	57.34	74.66	54.28	59.17 (+5.54)	66.53 (+2.36)
RetinaNet-GCL	42.55	68.38	56.40	73.53	54.36	59.04 (+5.41)	66.27 (2.10)

Table 4: Ablation study of four orientation detectors on DOTA test dataset. 5-mAP₅₀ means the performance of the five categories listed. The number in parentheses indicates the performance gain compared to the RetinaNet-Reg.

Method	ω	BR	SV	LV	SH	HA	5-mAP ₅₀	mAP ₅₀	mAP ₇₅	mAP ₅₀₋₉₅
Reg	-	34.52	51.42	50.32	73.37	55.93	53.12	62.21	26.07	31.49
CSL	180/1	35.99	53.42	61.08	81.81	62.14	58.87	64.40	32.58	35.04
	180/4	30.74	40.54	50.98	72.07	59.54	50.77	62.38	24.88	31.01
	180/8	36.65	52.58	60.46	82.24	61.60	58.71	66.17	33.14	35.77
	180/32	39.83	54.41	60.62	80.81	60.32	59.20	65.93	35.66	36.71
	180/64	38.22	54.70	60.16	80.75	60.11	58.79	65.00	34.31	36.00
	180/128	36.70	53.73	61.35	82.52	58.42	58.56	65.14	34.28	35.69
	180/180	37.42	53.72	58.70	80.73	63.31	58.78	65.83	33.94	36.35
BCL	180/256	37.66	53.83	60.66	80.43	60.74	58.66	64.97	33.52	35.21
	180/512	37.93	53.85	58.52	80.04	60.87	58.24	64.88	33.09	34.99
	180/4	30.90	41.20	48.30	72.93	60.16	50.70	62.98	23.83	30.81
	180/8	36.88	51.10	59.81	82.40	61.57	58.35	65.23	33.92	35.29
	180/32	38.04	54.77	60.88	82.75	61.24	59.54	65.11	34.67	36.15
	180/64	38.05	54.36	60.59	81.84	60.39	59.05	64.78	33.23	35.67
	180/128	37.74	54.36	59.43	81.15	60.51	58.64	66.13	33.65	36.34
GCL	180/256	35.81	53.78	58.35	81.45	59.84	57.85	64.87	33.77	35.97
	180/512	37.99	54.23	61.61	80.84	62.13	59.36	64.34	34.08	35.92

Table 5: Comparison of detection results under different angle discretization granularities denoted by ω .

Method	ADARSW	PL	BD	GTF	TC	BC	ST	SBF	RA	SP	HC	10-mAP ₅₀	mAP ₅₀
BCL	88.63	71.62	65.18	90.70	76.32	78.47	52.26	60.25	66.61	49.15	69.92	66.53	
	✓	88.92	72.11	66.32	90.79	79.03	54.11	63.18	67.86	60.04	72.22	67.39	
		88.52	73.58	64.38	90.80	77.66	76.38	50.84	59.46	65.83	48.42	69.59	66.27
GCL	✓	88.96	75.20	65.24	90.78	79.13	77.95	55.60	61.90	66.18	56.27	71.72	67.02

Table 6: The verification of ADARSW on small aspect ratio objects in the DOTA dataset.

epochs, respectively. The initial learning rates for RetinaNet is 5e-4. The number of image iterations per epoch for DOTA, UCAS-AOD, HRSC2016, ICDAR2015, and MLT are 27k, 5k, 10k, 10k and 10k respectively, and doubled if data augmentation and multi-scale training are used.

5.2. Ablation Study

Comparison of four object orientation detectors. Table 4 compares the performance of a regression based detector: RetinaNet-Reg and three classification based detectors: RetinaNet-CSL, RetinaNet-BCL and RetinaNet-GCL. We mainly focus on comparing five categories with large aspect ratios and more boundary conditions. It can be clearly seen that classification based detector outperforms those based on regression, with about 1.5%-2.3% and 3.6%-5.5% gain in overall performance (mAP₅₀) and five categories performance (5-mAP₅₀). More importantly, the performance of the DCL-based detector is nearly three times faster than the CSL-based detector, and the performance can still be further improved by about 2% and 0.8% in 5-mAP₅₀ and mAP₅₀. Figure 1 shows the visual qualitative comparison of the four methods under boundary conditions. We can fully draw the

Method	ICDAR2015						UCAS-AOD						MLT					
	Recall	Precision	Hmean	car(07/12)	plane(07/12)	mAP ₅₀ (07)	mAP ₅₀ (12)	Recall	Precision	Hmean								
RetinaNet-Reg	81.49	83.29	82.38	87.28/90.79	90.42/97.52	88.85	94.16	55.70	75.24	64.01								
RetinaNet-CSL	80.50	87.40	83.81 (+1.43)	88.09/ 92.93	90.38/97.22	89.23 (+0.38)	95.07 (+0.91)	58.32	73.62	65.08 (+1.07)								
RetinaNet-BCL	81.61	84.79	83.17 (+0.79)	88.15 /92.35	90.57 / 97.86	89.36 (+0.51)	95.10 (+0.94)	58.91	73.14	65.26 (+1.25)								

Table 7: Comparison between classification-based and regression-based methods on more datasets. 2007 or 2012 in bracket means using the 2007 or 2012 evaluation metric. ResNet101, data augmentation, multi-scale training and testing are used.

Method	Backbone	MS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP ₅₀
FR-O [49]	ResNet101		79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
R-DFPN [54]	ResNet101		80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
R ² CNN [17]	ResNet101		80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN [33]	ResNet101		88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
ICN [2]	ResNet101	✓	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
RADet [23]	ResNeXt101 [50]	✓	79.45	76.99	48.05	65.83	65.46	74.40	68.86	89.70	78.14	74.97	49.92	64.63	66.14	71.58	62.16	69.09
RoI-Transformer [7]	ResNet101	✓	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
CAD-Net [61]	ResNet101		87.8	82.4	49.4	73.5	71.1	63.5	76.7	90.9	79.2	73.3	48.4	60.9	62.0	67.0	62.2	69.9
AODD [69]	DPN [4]	✓	89.99	81.25	44.50	73.20	68.90	60.33	66.86	90.89	80.99	86.23	64.98	63.88	65.24	68.36	62.13	71.18
Cascade-FF [15]	ResNet152		89.9	80.4	51.7	77.4	68.2	75.2	75.6	90.8	78.8	84.4	62.3	64.6	57.7	69.4	50.1	71.8
SCRDet [59]	ResNet101	✓	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
SARD [46]	ResNet101		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
GLS-Net [21]	ResNet101		88.65	77.40	51.20	71.03	73.30	72.16	84.68	90.87	80.43	85.38	58.33	62.27	67.58	70.69	60.42	72.96
FADet [22]	ResNet101	✓	90.21	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.40	84.68	53.40	65.42	74.17	69.69	64.86	73.28
MFiar-Net [52]	ResNet152	✓	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 [51]	ResNet101		89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
Mask OBB [44]	ResNeXt101	✓	89.56	85.95	54.21	72.90	76.52	74.16	85.63	89.85	83.81	86.48	54.89	69.64	73.94	69.06	63.32	75.33
FFA [11]	ResNet101	✓	90.1	82.7	54.2	75.2	71.0	79.9	83.5	90.7	83.9	84.6	61.2	68.0	70.7	76.0	63.7	75.7
APE [68]	ResNeXt101		89.96	83.62	53.42	76.03	74.01	77.16	79.45	90.83	87.15	84.51	67.72	60.33	74.61	71.84	65.55	75.75
CenterMap OBB [45]	ResNet101	✓	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
FPN-CSL [56]	ResNet152	✓	90.25	85.53	55.64	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
SCRDet++ [58]	ResNet101	✓	90.05	84.39	55.44	73.99	77.54	71.11	86.05	90.67	87.32	87.08	69.62	68.90	73.74	71.29	65.08	76.81
IENet [28]	ResNet101	✓	80.20	64.54	39.82	32.07	49.71	65.01	52.58	81.45	44.66	78.51	46.54	56.73	64.40	64.24	36.75	57.14
TOSO [9]	ResNet101	✓	80.17	65.59	39.82	39.95	49.71	65.01	53.58	81.45	44.66	78.51	48.53	56.73	64.40	64.24	36.75	57.92
PlIoU [5]	DLA-34 [5]		80.9	69.7	24.1	60.2	38.3	64.4	64.8	90.9	77.2	70.4	46.5	37.1	57.1	61.9	64.0	60.5
P-RSDet [64]	ResNet101	✓	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 -DNet [47]	Hourglass104 [35]	✓	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
BBAVectors [60]	ResNet101	✓	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	72.32
DRN [36]	Hourglass104	✓	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
R ³ Det [53]	ResNet152		89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74
RSDet [37]	ResNet152		90.1	82.0	53.8	68.5	70.2	78.7	73.6	91.2	87.1	84.7	64.3	68.2	66.1	69.3	63.7	74.1
PolarDet [63]	ResNet101	✓	89.65	87.07	48.14	70.97	78.53	80.34	87.45	90.76	85.63	86.87	61.64	70.32	71.92	73.09	67.15	76.64
RetinaNet-DCL (Ours)	ResNet152	✓	89.10	84.13	50.15	73.57	71.48	58.13	78.00	90.89	86.64	86.78	67.97	67.25	65.63	74.06	67.05	74.06
R ³ D-DCL (Ours)	ResNet152		89.78	83.95	52.63	69.70	76.84	81.26	87.30	90.81	84.67	85.27	63.50	64.16	68.96	68.79	65.45	75.54
R ³ D-DCL (Ours)	ResNet101	✓	89.14	83.93	53.05	72.55	78.13	81.97	86.94	90.36	85.98	86.94	66.19	65.66	73.72	71.53	68.69	76.97
R ³ D-DCL (Ours)	ResNet152	✓	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

Table 8: Detection accuracy on different objects (AP₅₀) and overall performance (mAP₅₀) evaluation on DOTA. MS indicates that multi-scale training or testing is used.

conclusion that the orientation estimation based on classification is a boundary discontinuity free method.

Angle discretization granularity. In general, the smaller ω , the higher theoretical upper bound of the model's performance. However, the decrease of ω will lead to an increase in the number of angle categories, which poses a challenge to the angle classification performance of the model. Therefore, we need to explore the impact of ω on the detection performance under different IoU thresholds, and find a suitable range of ω . In order to get the performance indicators under different IoU threshold, we conduct experiments on the DOTA validation set, and the number of image iterations per epoch is 20k. According to Table 5, when the number of angle categories is between 32 and 128, the performance of the model reaches its peak. If the number of categories is too small, the theoretical accuracy loss is too large, resulting in a sharp drop in performance;

if the number of categories is too large, the angle classification network of the model cannot be effectively processed and the performance will decrease slightly. Figure 5 shows the comparison of angle estimates under different ω .

Redundant invalid coding. To make each code have a corresponding different angle value, the number of categories must be a power of 2 in the DCL-based method. However, this is not required. When we only set 180 categories, about 76 codings are invalid, but BCL-based method can still achieve good performance, at 36.35% as shown in Table 5. We also artificially increase the length based on the theoretical shortest code length to increase the proportion of invalid codes, and the performance is only slightly reduced.

Angle Distance and Aspect Ratio Sensitive Weighting. We mainly focus on comparing ten categories with small aspect ratio in Table 6 to verify the effectiveness of ADARSW. Under the same environment and hyperparameters, we add

Method	Backbone	mAP (07)	mAP (12)
R ² CNN [17]	ResNet101	73.07	79.73
RC1 & RC2 [32]	VGG16	75.7	—
RRPN [33]	ResNet101	79.08	85.64
R ² PN [62]	VGG16	79.6	—
RetinaNet-H [53]	ResNet101	82.89	89.27
RRD [25]	VGG16	84.3	—
RoI-Transformer [7]	ResNet101	86.20	—
Gliding Vertex [51]	ResNet101	88.20	—
BBAVectors [60]	ResNet101	88.6	—
DRN [36]	Hourglass104	—	92.70
CenterMap OBB [45]	ResNet50	—	92.8
SBD [31]	ResNet50	—	93.70
RetinaNet-R [53]	ResNet101	89.18	95.21
R ³ Det [53]	ResNet101	89.26	96.01
R ³ Det-DCL (Ours)	ResNet101	89.46	96.41

Table 9: Detection accuracy on HRSC2016. 07 (12) means using the 2007(2012) evaluation metric.

ADARSW to the BCL and GCL based methods, which increase by 2.3% and 2.13% in ten categories performance (10-mAP₅₀). The overall performance has also increased to 67.39% and 67.02%. After ADARSW is used, the model can predict the bounding box as shown in Figure 4(b).

Comparison on more datasets and detectors. Table 7 further verifies the performance advantage of DCL based methods than CSL and regression based method on more datasets, including the text dataset ICDAR2015, MLT, and another remote sensing dataset UCAS-AOD. It is worth noting that the comparison results are based on large backbone, data augmentation, and multi-scale training and testing. We can still draw the conclusions: classification is better than regression for orientation estimation; DCL outperforms CSL in most cases. In order to verify the portability of DCL, we also conduct experiments on R³Det. As shown in Table 8, DCL can still make R³Det get 1.8% improvement under the use of large backbone and data augmentation.

Visual analysis of angular features. To further analyze the angle classification ability of the model, we use the principal component analysis (PCA) [48] to visualize each positive angle feature vector. We show the visualization results when the number of angle categories are 4 and 8, as shown in Figure 6. This proves that it is feasible to use classification for orientation estimation, even if only the simplest cross-entropy loss function is used.

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

We choose DOTA as the main comparison dataset due to the complexity of the aerial image and the large number of small, cluttered and rotated objects. As shown in Table 8, through data augmentation, multi-scale training and testing commonly used by other advanced methods, RetinaNet-DCL-ResNet152 and R³Det-DCL-ResNet152 can achieve

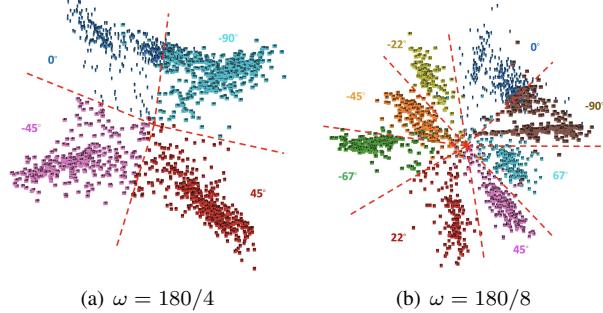


Figure 6: Angular feature visualization of the RetinaNet-DCL. The red dotted lines divide the different categories.

competitive performance, about 74.06% and 77.37%, respectively.

The HRSC2016 contains lots of large aspect ratio ship instances with arbitrary orientation, which poses a huge challenge to the positioning accuracy of the detector. Experimental results show that our model achieves state-of-the-art performances, about 89.46% (96.41%).

6. Conclusion

This paper develops the line of research in classification based methodology for rotation detection in two folds: i) for the prediction layer, two Densely Coded Labels (DCL) techniques are devised by shortening the code length to achieve a more light-weighted prediction layer. They both accelerate the training speed of the recently proposed Sparsely Coded Label model in orientation classification based detectors notably. ii) We further propose the technique called Angle Distance and Aspect Ratio Sensitive Weighting (ADARSW), which further improves the performance by making DCL-based detector sensitive to angular distance and object's aspect ratio. Extensive experiments on different detectors and datasets show competitive performance regarding with both accuracy and efficiency.

References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016. 6
- [2] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, pages 150–165. Springer, 2018. 1, 3, 7, 8

- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 3
- [4] Yunpeng Chen, Jianan Li, Huixin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in neural information processing systems*, pages 4467–4475, 2017. 8
- [5] Zhiming Chen, Kean Chen, Weiyao Lin, John See, Hui Yu, Yan Ke, and Cong Yang. Piou loss: Towards accurate oriented object detection in complex environments. *Proceedings of the European Conference on Computer Vision*, 2020. 8
- [6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016. 1, 3
- [7] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 1, 3, 8, 9
- [8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 3
- [9] Pengming Feng, Youtian Lin, Jian Guan, Guangjun He, Huifeng Shi, and Jonathon Chambers. 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*, pages 4057–4061. IEEE, 2020. 8
- [10] Gray Frank. Pulse code communication, Mar. 17 1953. US Patent 2,632,058. 5
- [11] Kun Fu, Zhonghan Chang, Yue Zhang, Guangluan Xu, Keshu Zhang, and Xian Sun. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161:294–308, 2020. 8
- [12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 1, 3, 6
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [14] FG Heath. Origins of the binary code. *Scientific American*, 227(2):76–83, 1972. 5
- [15] Liping Hou, Ke Lu, Jian Xue, and Li Hao. Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images. In *2020 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2020. 8
- [16] Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao. High-performance rotation invariant multiview face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 29(4):671–686, 2007. 1
- [17] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017. 1, 8, 9
- [18] Jiao Jiao, Yue Zhang, Hao Sun, Xue Yang, Xun Gao, Wen Hong, Kun Fu, and Xian Sun. A densely connected end-to-end neural network for multiscale and multiscale sar ship detection. *IEEE Access*, 6:20881–20892, 2018. 1
- [19] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition*, pages 1156–1160. IEEE, 2015. 7
- [20] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*, pages 734–750, 2018. 3
- [21] Chengyuan Li, Bin Luo, Hailong Hong, Xin Su, Yajun Wang, Jun Liu, Chenjie Wang, Jing Zhang, and Linhai Wei. Object detection based on global-local saliency constraint in aerial images. *Remote Sensing*, 12(9):1435, 2020. 8
- [22] Chengzheng Li, Chunyan Xu, Zhen Cui, Dan Wang, Tong Zhang, and Jian Yang. Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE International Conference on Image Processing*, pages 3886–3890. IEEE, 2019. 8
- [23] Yangyang Li, Qin Huang, Xuan Pei, Licheng Jiao, and Ronghua Shang. Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing*, 12(3):389, 2020. 8
- [24] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27(8):3676–3690, 2018. 1, 3
- [25] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5909–5918, 2018. 1, 3, 9
- [26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 1, 3
- [27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 1, 3, 6
- [28] Youtian Lin, Pengming Feng, and Jian Guan. Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection. *arXiv preprint arXiv:1912.00969*, 2019. 8
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37. Springer, 2016. 3

- [30] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5676–5685, 2018. 1, 3
- [31] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. *arXiv preprint arXiv:1906.02371*, 2019. 9
- [32] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. 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, pages 324–331, 2017. 7, 9
- [33] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018. 1, 3, 8, 9
- [34] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition*, volume 1, pages 1454–1459. IEEE, 2017. 7
- [35] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 483–499. Springer, 2016. 8
- [36] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11207–11216, 2020. 8, 9
- [37] Wen Qian, Xue Yang, Silong Peng, Yue Guo, and Junchi Yan. Learning modulated loss for rotated object detection. *arXiv preprint arXiv:1911.08299*, 2019. 1, 2, 3, 8
- [38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 1, 3
- [39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1, 3
- [40] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Rotation invariant neural network-based face detection. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*, pages 38–44. IEEE, 1998. 1
- [41] Xuepeng Shi, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Real-time rotation-invariant face detection with progressive calibration networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2295–2303, 2018. 1
- [42] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9627–9636, 2019. 3
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3
- [44] Jinwang Wang, Jian Ding, Haowen Guo, Wensheng Cheng, Ting Pan, and Wen Yang. Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing*, 11(24):2930, 2019. 8
- [45] Jinwang Wang, Wen Yang, Heng-Chao Li, Haijian Zhang, and Gui-Song Xia. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 8, 9
- [46] Yashan Wang, Yue Zhang, Yi Zhang, Liangjin Zhao, Xian Sun, and Zhi Guo. Sard: Towards scale-aware rotated object detection in aerial imagery. *IEEE Access*, 7:173855–173865, 2019. 8
- [47] Haoran Wei, Yue Zhang, Zhonghan Chang, Hao Li, Hongqi Wang, and Xian Sun. Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:268–279, 2020. 8
- [48] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 9
- [49] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Beßongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 6, 7, 8
- [50] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 8
- [51] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3, 8, 9
- [52] Feng Yang, Wentong Li, Haiwei Hu, Wanyi Li, and Peng Wang. Multi-scale feature integrated attention-based rotation network for object detection in vhr aerial images. *Sensors*, 20(6):1686, 2020. 8
- [53] Xue Yang, Qingqing Liu, Junchi Yan, Ang Li, Zhiqiang Zhang, and Gang Yu. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*, 2019. 1, 3, 8, 9
- [54] Xue Yang, Hao Sun, Kun Fu, Jirui Yang, Xian Sun, Menglong Yan, and Zhi Guo. 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. 1, 8
- [55] Xue Yang, Hao Sun, Xian Sun, Menglong Yan, Zhi Guo, and Kun Fu. Position detection and direction prediction for

- arbitrary-oriented ships via multitask rotation region convolutional neural network. *IEEE Access*, 6:50839–50849, 2018. 1
- [56] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *Proceedings of the European Conference on Computer Vision*, pages 677–694. Springer, 2020. 1, 2, 3, 4, 5, 8
- [57] Xue Yang, Junchi Yan, and He Tao. On the arbitrary-oriented object detection: Classification based approaches revisited. *arXiv preprint arXiv:2003.05597*, 2020. 1, 2, 3, 5
- [58] Xue Yang, Junchi Yan, Xiaokang Yang, Jin Tang, Wenlong Liao, and Tao He. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*, 2020. 1, 8
- [59] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8232–8241, 2019. 1, 2, 3, 8
- [60] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. Oriented object detection in aerial images with box boundary-aware vectors. *arXiv preprint arXiv:2008.07043*, 2020. 8, 9
- [61] Gongjie Zhang, Shijian Lu, and Wei Zhang. Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024, 2019. 3, 8
- [62] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1745–1749, 2018. 9
- [63] Pengbo Zhao, Zhenshen Qu, Yingjia Bu, Wenming Tan, Ye Ren, and Shiliang Pu. Polardet: A fast, more precise detector for rotated target in aerial images. *arXiv preprint arXiv:2010.08720*, 2020. 8
- [64] Lin Zhou, Haoran Wei, Hao Li, Yue Zhang, Xian Sun, and Wenzhe Zhao. Objects detection for remote sensing images based on polar coordinates. *arXiv preprint arXiv:2001.02988*, 2020. 8
- [65] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 1
- [66] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019. 3
- [67] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing*, pages 3735–3739. IEEE, 2015. 7
- [68] Yixing Zhu, Jun Du, and Xueqing Wu. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 8
- [69] Fuhao Zou, Wei Xiao, Wanting Ji, Kunkun He, Zhixiang Yang, Jingkuan Song, Helen Zhou, and Kai Li. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. *Neural Computing and Applications*, pages 1–14, 2020. 8