Supplementary Material for CVPR 2023
Dynamic Coarse-to-Fine Learning for Oriented Tiny Object Detection

Chang Xu[1], Jian Ding[2], Jinwang Wang[1], Wen Yang[1]*, Huai Yu[1], Lei Yu[1]*, Gui-Song Xia[2]
[1] School of Electronic Information, Wuhan University
[2] School of Computer Science, Wuhan University
{xuchangeis,jian.ding,jwwangchn,yangwen,yuhuai,ly.wd,guisong.xia}@whu.edu.cn

This supplementary mainly contains the following contents. First of all, in Sec.1, we perform comparisons with more learning paradigms for oriented object detection, empirically and experimentally. Then, to supplement the DGMM and GJSD's advantages over other counterparts, we provide more experimental and theoretical analyses in Sec. 2. Following this, more visualization results of predictions and sampled positive priors are shown in Sec. 4. Therein, more intuitive results further verify the DCFL's superiority in detecting oriented tiny objects. Finally, we illustrate two typical failure cases of the DCFL and the potential future work in Sec. 7.

1. More Comparisons

We compare with more learning paradigms in this section. Here we summarize previous works of oriented object detection into four categories, their schematic diagrams are shown in Fig. 1. The first one is the fixed paradigm, which lays a solid foundation for effective oriented object detection, representative works include the RetinaNet [11], FCOS [17], and Rotated RPN [12], they statically assign labels between fixed priors and fixed *gts* via a hand-craft heuristic. Hence, their samples for each *gt* remain the same in different iterations, and they are unable to adaptively filter out low-quality positive samples which fall on the background.

The second paradigm explores better alignment between the prior and *gt*, where detectors update the prior during different iterations. Among them, the S^2A-Net [7] introduces an anchor refinement network to generate high-quality anchors, then a fixed label assignment rule is applied, obtaining dynamic samples during the network learning. Although samples can be dynamic at different iterations, the detector cannot adaptively identify positive samples that fall on the object's background since the assignment rule remains fixed.

The third paradigm excavates a better utility of the fixed prior, where the detector's prior location is fixed while the dynamic assignment rule is employed to sample or measure these fixed anchors in a prediction-aware manner. For example, the DAL [13] fixes the position of each anchor and then designs a matching degree to dynamically reweight anchors. Therefore, the importance of high-quality samples can be highlighted when the network gradually converges. However, the position of each feature and prior remains fixed. Although they can achieve a better separation of *pos/neg* samples according to the instance's semantic pattern, most positive samples deviate from the tiny object's main body. That means prior and feature themselves cannot well-match the extreme shapes of oriented tiny objects, no matter how we divide pos/neg samples.

By contrast, our proposed DCFL further releases the flexibility of the network by updating the *prior*, *positive samples*, and *gt* representation during the network training. To begin with, the DCFL enables the detector to dynami-

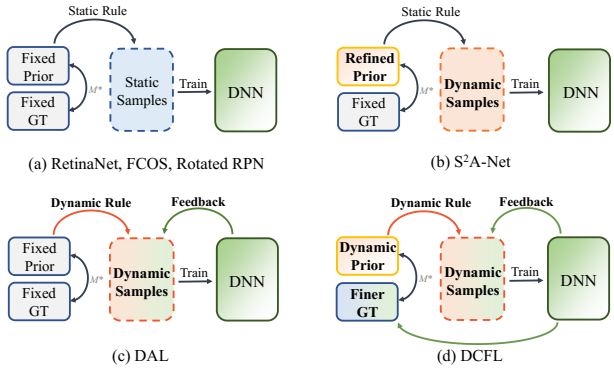

Figure 1. Comparisons of different learning paradigms for oriented object detection. M* means the matching function. (a) RetinaNet, FCOS, and Rotated RPN statically assign labels between fixed priors and fixed *gts*. (b) S^2A-Net statically assigns labels between the learnable anchors and *gts*. (c) DAL dynamically reweights fixed anchors. (d) Our proposed DCFL dynamically updates priors and *gts*, and dynamically assigns labels.

*Corresponding Authors

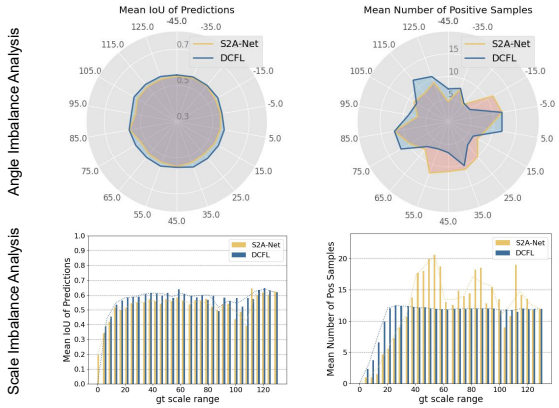

Figure 2. Statistical analysis of imbalance issues. The first and second columns show quality and quantity imbalance respectively.

Method	Circle	RBox	Single-G	Shrink-G	DGMM
mAP	64.36	66.91	66.97	67.81	**68.41**

Table 1. Comparison of different instance representations for posterior constraint.

cally update the prior location and feature sampling region that better fits the objects' characteristics. Simultaneously, the DCFL ensures a dynamic label assignment by the dynamic *gt* representation, getting rid of the samples that are away from the instance's main body.

In this part, we perform more statistical experiments of the S^2A-Net [7] since it is one of the state-of-the-art methods that shows competitive performance on the DOTA-v2.0. As shown in Fig. 2, it is easy to conclude that the S^2A-Net performs much better than the baseline RetinaNet [11], compared to Fig. 6 in the main paper. Nevertheless, the DCFL still exhibits better sample quality and sample quantity compared to the S^2A-Net. Concretely, as implied in the left part of Fig. 2, the sample quality of the DCFL is slightly higher than the S^2A-Net across almost all angles and scales. While the right part of Fig. 2 indicates that there still exist severe imbalance issues in the S^2A-Net which guides the network optimization focus on larger objects, although it can compensate more positive samples for the network training.

2. More Analyses

2.1. Dynamic Gaussian Mixture Model

We perform more detailed ablations to verify the advantages of the Dynamic Gaussian Mixture Model (DGMM). The advantages of utilizing the DGMM as a posterior constraint can be summarized into *higher accuracy*, *faster convergence*, and *finer representation* compared to other counterparts.

In Tab. 1, we report the performance of different instance representations for the posterior constraint. In this table, the "Circle" means eradicating positive samples outside the *gt* box's minimum circumcircle, the "RBox" means directly filtering out positive samples outside the *gt*'s rectangle bounding box, the "Single-G" means modeling the box into single Gaussian distribution and filtering out samples outside *gt* boxes' minimum circumscribed ellipse, the "Shrink-G" means shrinking the region of the "Single-G" via the same threshold e^{-g} as the DGMM. All models are trained on the DOTA-v1.0 `train set` and evaluated on the `val set`. We can see that the proposed DGMM gets the highest mAP among all choices, we attribute the performance improvement of the DGMM to its ability to dynamically capture the high response area of a specific *gt*, which better fits an object's semantic pattern. By contrast, other counterparts mainly rely upon the strong heuristic that objects are located at the center region of a given rectangle bounding box, yielding sub-optimal accuracy.

Besides the enhanced final accuracy, we also observe that the DGMM can facilitate the model's convergence. We show a comparison of epoch-aware mAP between a competitive counterpart "Shrink-G" and the proposed DGMM in Fig. 3. We can find that the DGMM significantly facilitates the model's convergence at early training epochs, where there is a boost of more than 5 mAP points. This can also be attributed to the DGMM's ability to retain high-quality samples. More precisely, the samples falling on the low response region can be filtered out by the DGMM, ensuring the stability and consistency of positive samples and leading to better convergence.

At last, apart from the accuracy improvement brings to the detector, the DGMM itself is a finer representation of the objects, which can fit the main body of the objects compared to the rectangle box. Please refer to Sec. 4 for more visualization results.

2.2. Generalized Jensen-Shannon Divergence

We present the theoretical analyses of the properties of Generalized Jensen-Shannon Divergence (GJSD), namely the scale-invariance, symmetry, and capability of measuring non-overlapping boxes.

First of all, note that the GJSD score in the main paper is calculated by simply normalizing the GJSD:

$$\text{score} = \frac{1}{1 + \text{GJSD}}, \tag{1}$$

since positive samples are obtained through a ranking manner, the introduced constant 1 in the denominator will not affect the performance. Then, we normalize the Wasserstein distance [15] and Kullback-Leibler divergence [9] in the same way to compare their difference.

As demonstrated in the previous work [14], the Generalized Jensen Shannon Divergence inherits the scale in-

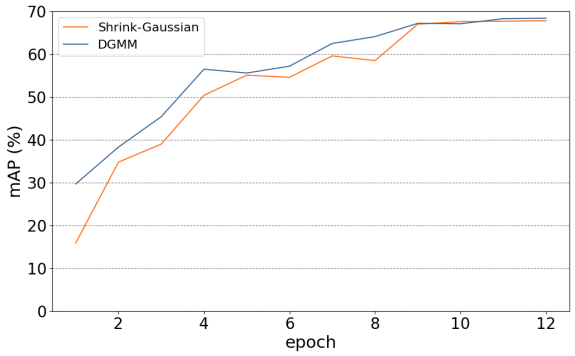

Figure 3. A comparison of the convergence between different instance representations.

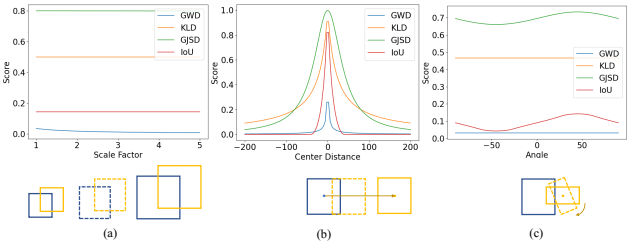

Figure 4. Properties of the GJSD.

Dataset	10-50 pixels	50-300 pixels	$>$ 300 pixels
DOTA-v1.0 [18]	57%	41%	2%
DOTA-v1.5 [3]	79%	20%	1%
DOTA-v2.0 [3]	77%	22%	1%

Table 2. The absolute object size distribution of the DOTA series.

variance property of the Jensen-Shannon Divergence [5, 8]. Specifically, suppose we have two Gaussian distributions $\mathcal{N}_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $\mathcal{N}_g(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. We then multiply the box's center point and side length by a full-rank scale factor $\mathbf{S} = k\mathbf{I}$ ($\mathbf{I}$ denotes identity matrix) [20], getting $\mathcal{N}_{p'}(\mathbf{S}\boldsymbol{\mu}_p, \mathbf{S}\boldsymbol{\Sigma}_p\mathbf{S}^\top)$ and $\mathcal{N}_{g'}(\mathbf{S}\boldsymbol{\mu}_g, \mathbf{S}\boldsymbol{\Sigma}_g\mathbf{S}^\top)$, the scale invariance of GJSD means that $\mathrm{GJSD}(\mathcal{N}_p||\mathcal{N}_g) = \mathrm{GJSD}(\mathcal{N}_{p'}||\mathcal{N}_{g'})$. Then, we experimentally investigate the scale invariance property in Fig. 4 (a). It can be seen that when we multiply the center points and the side length of the bounding boxes by a scale factor, the GJSD score remains constant, and the scale invariance is thus verified.

The symmetry of GJSD means that given two distributions $\mathcal{N}_g(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, we have $\mathrm{GJSD}(\mathcal{N}_p||\mathcal{N}_g) = \mathrm{GJSD}(\mathcal{N}_g||\mathcal{N}_p)$. By contrast, the Kullback-Leibler divergence is asymmetric.

Moreover, the GJSD can measure the similarity of non-overlapping boxes. As shown in Fig. 4 (b), we keep the blue box fixed and move away the yellow box away from the blue box's center, we can see that the GJSD score keeps changing even if boxes do not overlap. However, the IoU metric remains constant when boxes do not overlap.

In addition, we experimentally observe that the GJSD can distinguish some cases that cannot be handled by the KLD [20] and GWD [19], which might serve as one of the reasons for its superiority over the KLD and GWD in label assignment. Specifically, as shown in Fig 4 (c), we keep the square blue box $((cx_b, cy_b, w_b, h_b, \theta_b)$ is $(50, 50, 40, 40, 0))$ fixed, and keep the center and side length of the yellow box fixed $((cx_y, cy_y, w_y, h_y, \theta_y)$ is $(70, 70, 40, 40, angle))$, then we rotate the yellow box. From Fig 4 (c), we can see that the GJSD successfully distinguishes different angles, while other distribution distances (GWD, KLD) all fail.

3. Detailed Data Descriptions

Experiments are performed on five datasets, including four oriented object detection datasets: DOTA-v1.0 [18], DOTA-v1.5, DOTA-v2.0 [3], DIOR-R [2] and one horizontal small object detection dataset: VisDrone2019 [4]. These datasets all contain a considerable ratio of tiny objects, we report their detailed information in this section.

DOTA-v1.0 [18] is a large-scale dataset dedicated to object detection in aerial images. DOTA-v1.0 contains 2806 images ranging from 800×800 to 4000×4000 pixels, where 1/2, 1/6, and 1/3 of the images are officially divided as the training set, validation set, and testing set, respectively. There are total of 15 common categories and 188, 282 instances in this dataset. As shown in Tab. 2, 57% objects are smaller than 50 pixels, which indicates that there are many tiny objects in this dataset.

DOTA-v1.5 [3] is the later version of the DOTA-v1.0 and they share the same images along with the image set splits. However, DOTA-v1.5 additionally provides the annotations of tiny objects compared to DOTA-v1.0. Moreover, DOTA-v1.5 provides a new category named "container crane". In total, there are 403, 318 instances, where most of them are tiny objects as shown in Tab. 2.

DOTA-v2.0 [3] is the latest version of the DOTA series. To date, it is the largest dataset for Object Detection in Aerial Images (ODAI), there are 18 common categories, 11,268 images (in the range from 800×800 to $20,000 \times 20,000$ pixels), and 1,793,658 instances in DOTA-v2.0, which is much larger than the previous two versions. Concretely, 1,830 images, 593 images, 2,792 images, and 6,053 images are officially chosen as the training set, validation set, test-dev set, and test-challenge set, respectively. Also, it is worth noting that there are a large proportion of tiny objects, where 77% objects are smaller than 50 pixels. Considering that it is the largest dataset for ODAI and most of the objects in it are of tiny size, it is employed in the main experiments in this study.

DIOR-R [2] is the oriented version of the DIOR [10].

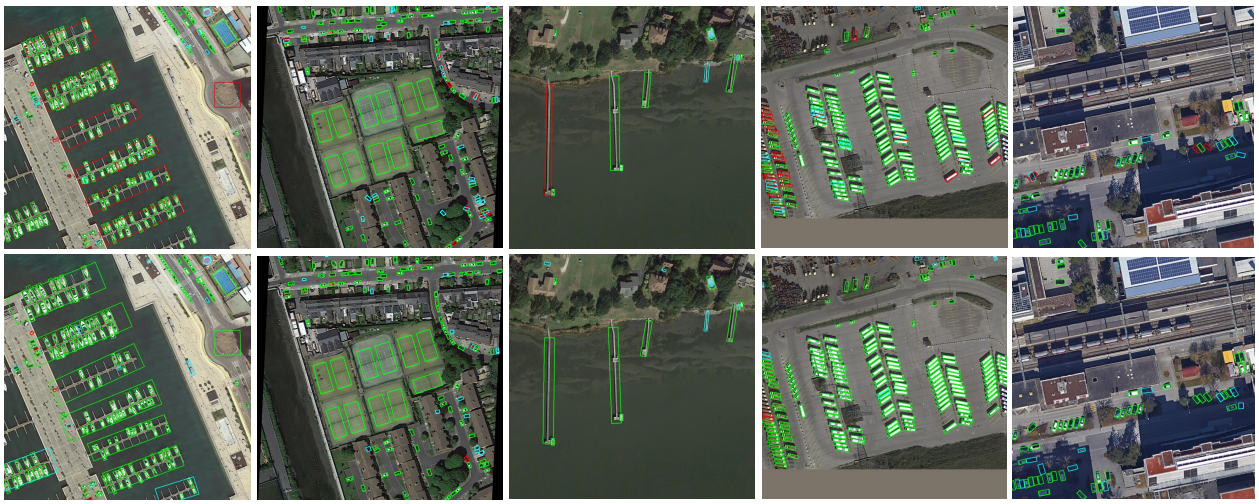

Figure 5. More visual results of predictions. The first row is the result of the RetinaNet-OBB while the second row is the result of the DCFL. TP, FN, and FP predictions are marked in green, red, and blue respectively.

Therefore, DIOR-R and DIOR contain the same images but different annotations, where objects in the DIOR-R are annotated with the oriented bounding box. DIOR-R has 23,463 images of 800×800 pixels and 192, 518 instances, covering 20 common object categories. DIOR-R also contains many tiny objects, where the windmill, bridge, and vehicle are three classes of the smallest sizes [10]. Hence, we report the mAP and classwise AP of the windmill, bridge, and vehicle of DIOR-R.

VisDrone2019 [4] is an UAV dataset for object detection. It is annotated by horizontal bounding boxes, covering 10,209 images with 10 categories. Captured in different places at different heights, objects in VisDrone2019 have large-scale variance and complex backgrounds, where many objects also exhibit extremely tiny scales.

4. More Visual Results

4.1. Detection Results

We provide more visualization of detection results in this section. In Fig. 5, the extensive visualization results can further support the superiority of the DCFL in detecting oriented tiny objects. Especially, we can see that the DCFL can alleviate both false positive and false negative predictions, for example, the oriented small vehicles and ships. This mainly results from the improvement of sample quantity and sample quality, as analyzed in Fig. 2. In other words, the balanced training of diverse instances and high-quality samples provided by the coarse-to-fine assigner significantly optimize the learning of extreme-shaped objects.

4.2. Positive Samples

We provide the visualization of more sampled positive priors in this section. As shown in Fig. 6, it is clear that the generated prior positions are arbitrarily distributed, getting rid of the fixed feature stride constraint, and providing more possibility for tiny object feature sampling. It can also be seen that the sampled positive priors better match the instance's main body, thus, the learning of the model is guided to focus on the strong semantic region.

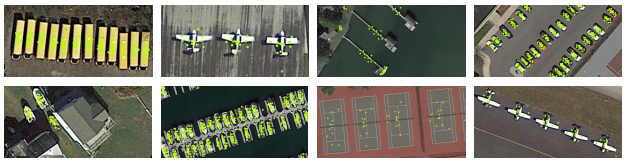

Figure 6. Visualization of sampled positive priors.

5. Discussions

5.1. Failure Cases

We illustrate these two failure cases in Fig. 7. The first one is that the performance of the DCFL is not satisfactory in the densely arranged scene. When objects are densely arranged, there may exist spatial feature aliasing [6] issues, leading to sub-optimal feature extraction and object location. Moreover, the post-processing NMS is a local optimal algorithm, which is not density-aware. Hence, some redundant predictions cannot be successfully deleted in this process. The second failure case is that the DCFL cannot easily handle weak objects. Compared to normal tiny objects, weak objects are even more lacking in appearance information, making the network hard to extract discriminative features.

(a) dense arrangement (b) weak objects

Figure 7. Failure cases.

5.2. Future Work

First of all, the proposed learning scheme could be further exploited to adapt to multi-stage object detectors. For example, the region proposal network [16] can be modified as the coarse matching process and the R-CNN [16] can serve as the finer matching process. One could even design a cascaded framework [1] to make the learning of objects finer and finer.

In addition, as implied in the failure cases, future work can be focused on resolving densely arranged objects and weak tiny objects. For example, we can combine the one-to-one assignment which is NMS-free and the many-to-one assignment which holds state-of-the-art performance to improve the detection performance in the densely arranged scenario. Besides, we can embed super-resolution strategies into the end-to-end object detection pipeline to enhance weak objects' features, improving weak objects' detection performance.

References

[1] Zhaowei Cai and Nuno Vas. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. 5

[2] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 3

[3] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Micheal Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page in press, 2021. 3

[4] Dawei Du, Pengfei Zhu, Longyin Wen, and et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *IEEE International Conference on Computer Vision Workshops*, pages 213–226, 2019. 3, 4

[5] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory (TIT)*, 49(7):1858–1860, 2003. 3

[6] Zonghao Guo, Chang Liu, Xiaosong Zhang, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8801, 2021. 4

[7] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 1, 2

[8] Takafumi Kanamori. Scale-invariant divergences for density functions. *Entropy*, 16(5):2611–2628, 2014. 3

[9] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 2

[10] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. 3, 4

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 1, 2

[12] 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

[13] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. Dynamic anchor learning for arbitrary-oriented object detection. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 2355–2363, 2021. 1

[14] Frank Nielsen. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy*, 22(2):221, 2020. 2

[15] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and Its Applications*, 48:257–263, 1982. 2

[16] 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. 5

[17] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision*, pages 9627–9636, 2019. 1

[18] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 3

[19] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International Conference on Machine Learning*, volume 139, pages 11830–11841, 2021. 3

[20] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision

bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 34, 2021. 3