Checking the Reproducibility and Generalizability of Dynamic Coarse-to-Fine Learning for Oriented Tiny Object Detection

Hsiang-Chen Chiu **Yu-Chien Lin** **Desmond Cheong Zheng Yong**

Abstract

In the domain of computer vision and object detection, the task of accurately identifying extremely small objects with diverse orientations remains a substantial challenge. While recent progress has been made in addressing this issue through adaptive label assignment strategies, the core challenges of dealing with mismatches and imbalances continue to persist. This proposal outlines our intention to conduct an in-depth examination of a research paper authored by a team from Wuhan University, who assert that their method for detecting oriented tiny objects represents the current state of the art.

Our exploration is structured into three key sections. The first segment serves as an introduction, providing an overview of the subject matter and our overarching objectives. The second section delves into the technical aspects of the paper, dissecting the methodologies and innovations presented in the mentioned paper. Finally, the third part outlines our proposed schedule of milestones, including dates and sub-goals to ensure the systematic execution of our research on this topic.

1. Introduction

1.1. Challenges

Detecting extremely small objects with arbitrary orientations presents significant challenges for current detection methods, particularly when it comes to assigning appropriate labels. [2]

While recent oriented object detectors have explored adaptive label assignment, the unique geometric shapes and limited feature information of these small, oriented objects continue to create substantial problems related to mismatch and imbalance.

Specifically, there are issues with misalignment between the expected object positions, the features of positive samples, and the instances themselves.

Additionally, learning to detect objects with extreme shapes is hindered by a lack of proper feature guidance, leading to bias and imbalance in the learning process.

1.2. Method Proposed by Wuhan University

In the accepted paper in IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) 2023, titled as Dynamic Coarse-to-Fine Learning for Oriented Tiny Object Detection, six researchers from Wuhan University proposed a novel approach called DCFL, which combines a dynamic prior with a coarse-to-fine label assigner. [1]

Their method aims to mitigate the issue of mismatch by dynamically modeling the prior, label assignment, and object representation. Furthermore, they use a combination of coarse prior matching and fine posterior constraints to dynamically assign labels, thus offering more suitable and balanced supervision for a wide range of object instances. [1]

Their extensive experiments across six datasets demonstrate significant enhancements over the baseline. Notably, they achieve state-of-the-art performance in one-stage object detectors for the DOTA-v1.5, DOTA-v2.0, and DIORR datasets when training and testing under a single scale.

The codes of their approaches and methods can be accessed at https://github.com/ChaselTsui/mmrotate-dcfl.

1.3. Reproducibility and Generalizability

We are interested in checking the reproducibility and generalizability on the other data of the paper's findings through the Ablation Track of NeurIPS Reproducibility Challenge for several compelling reasons. First and foremost, the paper highlights the significant challenges associated with detecting arbitrarily oriented tiny objects, a problem of great relevance in computer vision and object detection applications. Nonetheless, given the ongoing advancements in this field, ensuring the reproducibility and robustness of the proposed solution is critical.

2. Technical Part

Overview:

In the mentioned work, they model the prior *P*, label assignment, and *gt* representation all in a dynamic manner to alleviate the mismatch issue. To begin with, the dynamic prior [3, 4, 5] is formulated to (˜denotes the dynamic item):

$$\widetilde{D} = DNN_h(DNN_p(P)),$$

DNN$_{\text{P}}$, is a learnable block incorporated within the detection head to update the prior. Then, the matching function is formulated to a coarse-to-fine paradigm:

$$\tilde{G} = M_d \left(M_s \left(\tilde{P}, GT\right), \widetilde{GT} \right),$$

where the M_d, M_s are the matching function for static assigners and prediction-aware mapping function for dynamic assigners. The $\widetilde{GT}$ is the finer representation of an object with the Dynamic Gaussian Mixture Model (DGMM). In essence, the final loss is modeled as:

$$L = \sum_{i=1}^{\tilde{N}_{pos}} L_{pos} \left(\widetilde{D}_i, \tilde{G}_i \right) + \sum_{j=1}^{\tilde{N}_{neg}} L_{neg} \left(\widetilde{D}_j, y_j \right) ,$$

where $\widetilde{N}_{pos}$ and $\widetilde{N}_{neg}$ are the number of positive and negative samples respectively, y_j denotes the negative label.

2.1. Dynamic Prior

Each prior location *p(x, y)* is initialized based on the spatial location of each feature point, which is mapped to the image. In each iteration, the network is forwarded to capture the offset sets for each prior location, represented as Δo. These offsets are used to update the spatial location of the prior:

$$\tilde{s} = s + st \sum_{i=1}^{n} \frac{\Delta o_i}{2n},$$

where *st* is the feature map's stride, *n* is the number of offsets.

Finally, the 2-D Gaussian distribution $N_p \left(\mu_p, \sum_p\right)$ is utilized which is demonstrated conducive to small objects [6, 7] and oriented objects [7, 8] to fit the prior spatial location. Concretely, the dynamic $\tilde{s}$ serves as the Gaussian's mean vector μ_p. We preset one prior which is squareshaped (*w, h, θ*) as that in RetinaNet [9] on each feature point, then compute the co-variance matrix $\sum_p$ [10] by :

$$\sum\nolimits_p = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \dfrac{w^2}{4} & 0 \\ 0 & \dfrac{h^2}{4} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

2.2. Coarse Prior Matching

The paper introduces Cross-FPN-layer Coarse Positive Sample (CPS) candidates to improve sample selection. This method narrows down the sample range compared to using all FPN layers [11, 12] while avoiding the limitations of selecting candidates from just one layer [3, 4, 13] . In CPS, the range of candidates is expanded to include the *gt's* nearby spatial location and adjacent FPN layers, ensuring a more diverse and sufficient candidate pool.

The Jensen-Shannon Divergence (JSD) [14] is used for measuring similarity, addressing issues such as scale invariance and overcoming the limitations of Kullback–Leibler Divergence (KLD) [8]. While JSD usually lacks a closed-form solution for Gaussian distributions, it is subsituted with the Generalized Jensen-Shannon Divergence (GJSD) [15], which offers a closed-form solution. For example, the GJSD between two Gaussian distributions N_p and N_g is defined by:

$$GJSD\left(N_p , N_g \right) = (1 - \alpha)\, KL \left(N_p , N_g \right) + \alpha KL\left(N_p , N_g \right),$$

where KL denotes the KLD and α denotes the parameter [15] that controls the weight of the two distributions in similarity measurement.

K priors with the highest GJSD scores are selected as Coarse Positive Samples (CPS) for each ground truth (*gt*), while the remaining priors are regarded as negative samples. This ranking approach, using GJSD, forms the Cross-FPN-layer CPS and resolves issues related to imbalance caused by MaxIoU matching for outlier angles and scales, which will be analyzed in the next section.

2.3. Finer Dynamic Posterior Matching

Based on the Coarse Positive Sample (CPS) candidates, a dynamic posterior matching rule, M_d, is designed to filter out low-quality samples. M_d comprises two key components: a posterior re-ranking strategy and constraints from a Dynamic Gaussian Mixture Model (DGMM). The sample candidates in CPS are re-ranked based on their predicted scores, further refining positive samples based on their Possibility of becoming True predictions (*PT*) [16]. *PT* for the *i-th* sample, Di, is calculated as a linear combination of the predicted classification score and the location score with the ground truth (*gt*), specifically:

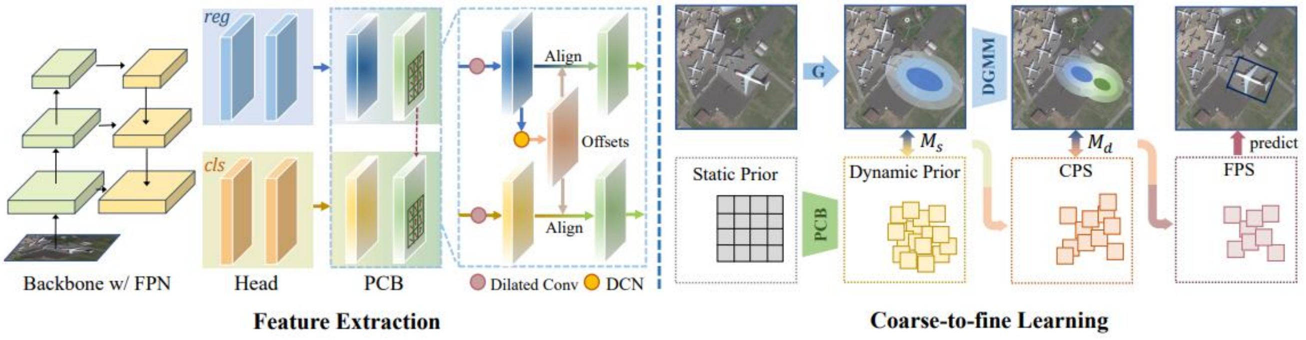

Figure 1: The process of feature extraction and dynamic coarse-to-fine learning. PCB denotes the prior capturing block. [1]

$$PT_i = \frac{1}{2}\, Cls(D_i) + \frac{1}{2}\, IoU(D_i, gt_i)\,,$$

where *CLs* is the predicted classification confident and *IoU* is the rotated *IoU* between the prediction location and its corresponding *gt* location. Candidates with the *Q* highest *PT* values are selected as Medium Positive Sample (MPS) candidates.

Subsequently, samples located too far from the ground truths (*gts*) are filtered out, leading to the extraction of Finer Positive Samples (FPS). In contrast to prior approaches that employ the center probability map [17] or a single Gaussian [10, 18] for instance representation, a finer Dynamic Gaussian Mixture Model (DGMM) composed of two components is utilized. One component is centered on the geometry center, while the other is centered on the semantic center of the object. Specifically, for a given instance, *gti*, the geometry center $(CX_i,\ CY_i)$ serves as the mean vector $\mu_{i,1}$ of the first Gaussian, and the semantic center $(SX_i,\ SY_i)$,which is determined by averaging the location of the samples in the MPS, serves as $\mu_{i,2}$. In other words, the instance is parameterized as follows:

$$DGGM_i\ (s|\ x, y) = \sum_{m=1}^{2} w_{i,m} \sqrt{2\pi|\Sigma_{i,m}|}\ N_{i,m}(\ \mu_{i,1}, \Sigma_{i,m})\,,$$

where $w_{i,m}$ is the weight of each Gaussian, with a total sum of 1, and $\Sigma_{i,m}$ corresponds to the *gt*'s Σ_g. Each sample in MPS possesses a DGMM score *(DGMM (s|MPS))*, and negative masks are applied to samples *with DGMM (s|MP S)* < e^{-g} relative to any *gt*, with the value of *g* being adjustable.

2.4. Our goals

By choosing the Ablation Track, we aim to meticulously analyze and dissect the authors' codebase and methodology through rigorous ablation studies and hyper-parameter explorations on it. This approach will allow us to isolate and evaluate the individual components and design choices of the DCFL method. It will provide insights into how each element contributes to the model's overall performance and behavior, which is crucial for assessing the method's generalizability across different datasets and scenarios.

The paper's claim of achieving state-of-the-art performance on specific datasets is compelling, but it is essential to investigate whether these results hold true in diverse real-world situations. By conducting extensive ablation studies and hyper-parameter explorations, we aim to understand the robustness of the DCFL method, identify its strengths and weaknesses, and provide valuable insights for its potential applications in various contexts beyond the datasets mentioned in the paper. Ultimately, this research will contribute to a better understanding of the model's capabilities and limitations, ensuring that its benefits can be realized in a broader range of practical scenarios.

3. Experiments

3.1 Datasets

DOTA-v2.0 We utilize DOTA-v2.0 to validate and reproduce the experimental results presented in the original paper. DOTA-v2.0 is a dataset featuring aerial images with small objects and consists of 11,268 images across 18 classes. We use the extensive DOTA-v2.0 train set for model training and val set for evaluation; this setting aligns with the methodology outlined in the original paper for ablation study.

HRSC2016 We use the HRSC2016 dataset to check the generalizability of this paper. HRSC2016, which stands for High-Resolution Ship Collections, comprises 1,061 images featuring approximately 2,976 instances of ships. Our motivation for testing on this dataset is driven by the primary aim of DCFL in oriented tiny object detection. We aim to investigate its effectiveness in detecting objects at high resolutions.

Measurement	mAP	
	Paper	Ours
KLD	57.8	50.6
GWD	58.6	57.7
GJSD	**59.2**	**59.4**

Table 1: Comparisons of different similarity measurements in constructing CPS using DOTA-v2.0

K	16			
Q	12	10	8	6
mAP	**59.2**	58.6	59.0	57.8

Table 2: Results of ablation study from the paper on effects of parameter K, Q using DOTA-v2.0

K	16				
Q	14	12	10	8	6
mAP	**59.9**	59.4	59.6	58.6	59.4

Table 3: Our reproduced results of ablation study on effects of parameter K, Q using DOTA-v2.0

	g	1.2	1.0	0.8	0.6	0.4
mAP	Paper	57.9	58.2	**59.2**	-	59.0
	Ours	58.5	59.2	58.7	59.4	**59.7**

Table 4: Results of ablation study on effects of parameter g using DOTA-v2.0

3.2 Experimental Setups

In the original paper, experiments were conducted using a single RTX 3090 GPU. In contrast, we utilize a single RTX 3060 GPU in our experiments. This difference in GPU selection has contributed to some variations in our testing results.

3.3 Results on DOTA-v2.0 Dataset

In accordance with the ablation study settings presented in the original paper, we employed the DOTA-v2.0 dataset to conduct the following experiments aimed at reproducing their results.

Comparisons of Different CPS. We replicated the experimental framework as delineated in the original paper's ablation study, leveraging the DOTA-v2.0 dataset to evaluate various similarity measurements employed in constructing Cross-FPN-layer Coarse Positive Sample (CPS). As detailed in Table 1, our results are directly compared with those from the original paper. Notably, the Kullback-Leibler Divergence (KLD) showed a decrease in performance, with our results at 50.6 mAP compared to the paper's 57.8, while our Gaussian Wasserstein Distance (GWD) performance was closer to the original, recording 57.7 mAP. Remarkably, the Generalized Jaccard Similarity Distance (GJSD) demonstrated an enhanced performance, yielding a 59.4 mAP which surpasses the paper's reported results. This underscores the robustness of GJSD, aligning with the paper's assertions on the superiority of this measurement in CPS construction.

Effects of Parameters K, Q. In our replication of the ablation study from the original paper, we examined the effects of parameters K and Q on the performance of the model. Our results are shown in Table 3, while the results from the original study are shown in Table 2. While the paper reported the highest mAP of 59.2 with a combination of K=16 and Q=12, our experiments reveal that increasing Q to 14 — beyond the highest value considered in the paper—leads to our best performance, achieving an mAP of 59.9. In this context, K determines the number of priors where the top GJSD scores are selected as Coarse Positive Sample (CPS), while Q represents the quantity of candidates chosen as Medium Positive Samples (MPS) based on the highest PT scores. Overall, our experiments confirm the robustness of the K and Q parameters, reinforcing the original study's conclusions and providing additional data points that contribute to a more comprehensive understanding of the parameter space.

Effects of Parameters g. Our results of using different parameter g compared with the original paper are shown in Table 4. At 'g' = 0.6, where the original study's results are absent, our experiments filled this gap, registering an improved mAP of 59.4. Remarkably, setting 'g' to 0.4, we attain our highest mAP of 59.7, surpassing all other configurations. In light of our experimental outcomes, it appears that varying the parameter 'g' does not significantly sway the model's performance, which aligns with the original paper's assertion regarding the robustness of the parameter selection process.

3.4 Main Results on HRSC2016 Dataset

While the original paper focused on the DCFL's performance with tiny oriented objects, our study aims to evaluate its generalizability to larger objects. For this purpose, we employed the HRSC2016 dataset, which contains images of a single class, 'ship', representative of larger object scales. This choice allows us to scrutinize the DCFL's adaptability and performance consistency when confronted with larger-scale objects.

The results of employing DCFL on HRSC2016 are shown in Table 5. DCFL achieved a mAP of 85.1 on the HRSC2016 dataset, a significant improvement over the

	mAP
atss_le135	67.1
retinanet_le135	61.8
DCFL	**85.1**

Table 5: Results of DCFL on HRSC2016 dataset, comparing with 2 baselines

Measurement	mAP
KLD	86.4
GWD	**87.3**
GJSD	85.1

Table 6: Comparisons of different similarity measurements in constructing CPS using HRSC2016

K	16				
Q	14	12	10	8	6
mAP	87.9	85.1	88.0	87.3	**88.6**

Table 7: Results of ablation study on effects of parameter K, Q using HRSC2016

g	1.2	1.0	0.8	0.6	0.4
mAP	86.3	**88.2**	87.1	85.1	88.0

Table 8: Results of ablation study on effects of parameter g using HRSC2016

Backbone	LSKNet[20]	ResNet50	ResNet101	ResNet152
mAP	54.6	59.4	60.6	**61.4**

Table 9: Results of backbone replacement using DOTA-v2.0

baseline models atss_1e135 and retinanet_1e135, which scored 67.1 and 61.8, respectively. The results provide insights into the model's versatility across varying object sizes and demonstrates the generalizability of DCFL on different datasets.

3.5 Ablation Studies on HRSC2016 Dataset

In parallel to our experiments on the DOTA-v2.0 dataset, we conducted a series of ablation studies on the HRSC2016 dataset.

Comparisons of Different CPS. Table 6 presents the results of our ablation studies on the HRSC2016 dataset, where we investigated the effectiveness of various similarity measurements in constructing CPS. A noteworthy divergence from the results obtained with the DOTA-v2.0 dataset is observed. While GJSD yielded the best performance in the DOTA-v2.0, it was not the top-performing metric on the HRSC2016 dataset. Instead, the GWD achieved the highest mAP with a score of 87.3, followed by the KLD with 86.4, and GJSD with 85.1. We think that the possible reason for this inconsistency may be tied to the distinctive characteristics of the HRSC2016 dataset. Unlike DOTA-v2.0, which features a multitude of tiny objects, the HRSC2016 dataset is comprised exclusively of larger-scale, single-class objects — ships. This may suggest that different similarity measurements may interact with object scale and class characteristics in varied ways. The superior performance of GWD in this context seems to underscore its capacity for capturing the larger object scales more effectively than GJSD, which might be better suited for datasets with smaller and more varied objects, as in DOTA-v2.0.

Effects of Parameters. Table 7 shows the results obtained when selecting different values for parameters Q and K on the HRSC2016 dataset. Meanwhile, Table 8 presents the results regarding the effectiveness of parameter g. Despite the changes in parameter settings, the results demonstrate that the model retains a high mAP across a range of values for K and Q, as well as for the parameter g. For instance, the highest mAP recorded for K and Q was 88.6 with K=16 and Q=6, and for the g parameter, the highest was 88.2 at g=1.0. These results are not entirely consistent with the DOTA-v2.0 dataset, and generally support the claim of robustness in choosing different parameters as asserted in the paper.

3.6 DCFL Backbone Replacement Experiments

In order to assess the flexibility of the DCFL model, we conducted experiments to replace its backbone with various architectures. We tested three variants of ResNet and the LSKNet [20], which was introduced at ICCV 2023, focusing on their performance on the DOTA-v2.0 dataset. The results of these experiments are presented in Table 9.

The experiments revealed that ResNet152 achieved the best result, with an mAP of 61.4. However, a noteworthy decline in performance was recorded when the backbone of the DCFL model was replaced with LSKNet, which resulted in an mAP of 54.6. We think the possible reason may be that the LSKNet already yield a very good result on tiny object detection, so in the coarse-to-fine learning pipeline, each stage of selecting K priors and Q candidates may cause information loss. Hence, the LSKNet architecture, while highly effective on its own, may not work well with the DCFL framework.

4. Conclusions

Our study provides a detailed account of the experimental validation and ablation studies conducted to assess the performance and generalizability of the Dynamic Coarse-to-Fine Learning (DCFL) method for oriented tiny object detection. The research utilizes two datasets: DOTA-v2.0, with a focus on aerial images of small objects, and HRSC2016, which features high-resolution images of ships..

For the DOTA-v2.0 dataset, the study compares different Cross-FPN-layer Coarse Positive Sample (CPS) constructions and examines the influence of various parameters (K, Q, and g) on the mean Average Precision (mAP). The results show that while some performance metrics like Kullback-Leibler Divergence (KLD) underperform compared to the paper's results, others like the Generalized Jaccard Similarity Distance (GJSD) exceed expectations. This indicates the method's potential robustness and adaptability. Adjusting parameters has shown to improve the mAP, suggesting that fine-tuning can enhance DCFL's performance.

For the HRSC2016 dataset, the study seeks to understand DCFL's applicability to larger objects. The results indicate a significant improvement in performance over baseline models, demonstrating DCFL's versatility across varying object sizes and suggesting its generalizability to different datasets.

The ablation studies on the HRSC2016 dataset show that while some metrics like GJSD perform well on DOTA-v2.0, others like the Gaussian Wasserstein Distance (GWD) are more effective for HRSC2016, highlighting the importance of context and dataset characteristics in performance.

Further experiments involving the replacement of DCFL's backbone with different architectures reveal that while some replacements like ResNet152 enhance performance, others like LSKNet may not be as compatible with the DCFL framework, indicating the nuanced relationship between architecture and method.

Overall, our study suggests that while DCFL has strong potential and adaptability, its performance can vary significantly based on the dataset, parameters, and architecture used, necessitating careful consideration and adjustment for optimal results.

5. References

[1] Xu, Chang and Ding, Jian and Wang, Jinwang and Yang, Wen and Yu, Huai and Yu, Lei and Xia, Gui-Song. Dynamic Coarse-To-Fine Learning for Oriented Tiny Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[2] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In IEEE Conference on Computer Vision and Pattern Recognition, pages 9759–9768, 2020 FirstName Alpher, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? *Journal of Foo,* 14(1):234–778, 2004.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-toend object detection with transformers. In European Conference on Computer Vision, pages 213–229. Springer, 2020.

[4] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In European Conference on Computer Vision, pages 355–371. Springer, 2020.

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

[6] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Rfla: Gaussian receptive field based label assignment for tiny object detection. In European Conference on Computer Vision, pages 526–543. Springer, 2022.

[7] 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.

[8] 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 kullbackleibler divergence. Advances in Neural Information Processing Systems, 34, 2021.

[9] 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.

[10] Xue Yang, Gefan Zhang, Xiaojiang Yang, Yue Zhou, Wentao Wang, Jin Tang, Tao He, and Junchi Yan. Detecting rotated objects as gaussian distributions and its 3-d generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[11] Mohsen Zand, Ali Etemad, and Michael Greenspan. Objectbox: From centers to boxes for anchor-free object detection. European Conference on Computer Vision, 2022.

[12] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In IEEE Conference on Computer Vision and Pattern Recognition, pages 840–849, 2019.

[13] Zhonghua Li, Biao Hou, Zitong Wu, Licheng Jiao, Bo Ren, and Chen Yang. Fcosr: A simple anchor-free rotated detector

for aerial object detection. arXiv preprint arXiv:2111.10780, 2021.

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

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

[16] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[17] 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, 59(5):4307–4323, 2021.

[18] Zhanchao Huang, Wei Li, Xiang-Gen Xia, and Ran Tao. A general gaussian heatmap label assignment for arbitraryoriented object detection. IEEE Transactions on Image Processing, 31:1895–1910, 2022.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, pages 8024–8035, 2019.

[20] LI, Yuxuan, et al. Large Selective Kernel Network for Remote Sensing Object Detection. *arXiv preprint arXiv:2303.09030*, 2023.