EXTRACTING AND DISTILLING DIRECTION-ADAPTIVE KNOWLEDGE FOR LIGHTWEIGHT OBJECT DETECTION IN REMOTE SENSING IMAGES

Zhanchao Huang, Wei Li, and Ran Tao

School of Information and Electronics, Beijing Institute of Technology, China

ABSTRACT

Recently, some lightweight convolutional neural network (CNN) models have been proposed for airborne or spaceborne remote sensing object detection (RSOD) tasks. However, these lightweight detectors suffer from performance degradation due to the compromise of limited computing resources on embedded devices. In order to narrow this performance gap, a direction-adaptive knowledge extraction and distillation (DKED) method is proposed. Specifically, a dynamic directional convolution (DDC) is developed to extract the typical arbitrary-oriented features, and a direction-adaptive knowledge distillation (DKD) strategy is designed for guiding the lightweight model to learn the intrinsic knowledge of the RSOD task from the high-performance model. Experiments on public datasets demonstrate that the proposed method can effectively improve the performance of the lightweight RSOD model without additional inference costs.

Index Terms— Dynamic directional convolution, knowledge distillation, lightweight object detection, remote sensing

1. INTRODUCTION

With the richness of remote sensing resources and the rapid development of deep learning technology, many CNN-based RSOD models have been proposed, such as FFA [1], SCRDet [2], ReDet [3]. In order to improve feature extraction and characterization capabilities, these CNN models are designed to be larger and more complex, which lead to higher demands on computing resources. However, for many real applications, especially airborne or spaceborne platforms, computing resources are very limited. The contradiction between complex algorithms and limited computing power restricts the application of RSOD model [4].

In this regard, some lightweight RSOD methods used simpler CNN structures [5] or less convolutional layers [6]. Obviously, such rough model compression methods greatly damage the performance of the models and are only effective for object detection in simple scenes. Some other RSOD models [4] adopted the idea of lightweight the convolution operation, such as the separable convolution proposed by MobileNet [7], or the scheme of feature map shuffling for feature channel interaction proposed by ShuffleNet [8]. Among them, LO-Det [4] analyzed and summarized the efficiency problems existing in the design of the lightweight RSOD model and provided a solution with better performance and efficiency. However, there is still a significant performance gap between these lightweight models and larger models due to their weaker and inefficient feature representation and abstraction.

Therefore, in addition to compressing the model structure, it is also important to refine the feature representation capabilities of the size-constrained model. Knowledge distillation (KD) [9] has achieved more efficient feature refinement by guiding the lightweight model to mimic the feature extraction of the complex model with high-performance. In the object detection task, Chen et al. [10] first proposed to distill the hints features of the larger teacher model when training the lightweight student model. Li et al. [11] designed a supervision of high-level features sampled from the region proposals of the teacher model. Wang et al. [12] developed a fine-gained feature imitation method for the anchor-based localization. Zheng et al. [13] proposed a localization distillation that refined the flexible localization information for the lightweight model. Dai et al. [14] introduced the relation-based knowledge for distillation on object detection tasks.

However, the above-mentioned KD methods designed for the ordinary object detection task do not consider the particularity of object representations in remote sensing images. First, the objects in the remote sensing images are in the top view, and their direction arbitrariness needs to be considered when extracting and distilling features. Second, the rich global structural information of remote sensing images and the feature correlation of objects in the same category should be distilled. Third, objects in the RSOD task are usually represented by oriented bounding boxes (OBBs), the localization distillation needs to be redesigned. Therefore, a direction-adaptive knowledge extraction and distillation (DKED) is proposed. The contributions of this work are summarized as follows:

1) The proposed DKED attempts to narrow the performance gap between the lightweight RSOD model and the

This work was supported by National Key R&D Program of China under Grant No.2021YFB3900502, the National Natural Science Foundation of China under Grant 61922013 and U1833203, the Beijing Natural Science Foundation under Grant L191004 and JQ20021, and by the Aeronautical Science Foundation of China 20200051072001. (Corresponding Author: Wei Li; e-mail: liwei089@ieee.org)

complex model without any additional inference costs.

2) In the proposed DKED, a dynamic directional convolution (DDC) is developed to extract the typical arbitrary-oriented features of remote sensing objects.

3) A direction-adaptive knowledge distillation (DKD) framework consists of feature correlation distillation, feature response distillation, and prediction distillation is designed to enhance and refine the feature representation.

2. THE PROPOSED DKED FRAMEWORK

The proposed DKDE consists of two parts: the DDC blocks, which enhances the direction feature extraction of remote sensing objects, and the DKD strategy, which distills remote sensing object features from the high-performance model to the lightweight model.

2.1. Dynamic Directional Convolution

In the overhead-view remote sensing images, objects have arbitrary directions. Many studies, such as SCRDet [2], ReDet [3], etc., have shown that this directional feature plays an important role for the RSOD task. Most of the existing methods extract the directional features from the perspective of multi-directional feature fusion [1–3]. However, these methods introduce additional directional feature extraction modules in CNN, which have high computational complexity and slow down the detection speed. While the data argumentation trick is highly random and cannot adapt to a wider variety of object directions. Therefore, dynamic direction convolution is proposed, as shown in Fig. 1, which directly improves the convolution kernel to adaptively extract the features in different directions. The proposed DDC does not require a special feature extraction module design, and can easily replace the convolution in the lightweight model without introducing excessive computational burden.

Without loss of generality, a neuron of the convolutional layer is modeled as:

$$\boldsymbol{y} = f\left(\boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{b}\right), \tag{1}$$

where $f(\cdot)$ is an activation function; $\boldsymbol{x}$ is the input feature vector; $\boldsymbol{y}$ denotes the output feature vector; $\boldsymbol{W}^T$ and $\boldsymbol{b}$ are weight matrix and bias vector, respectively. The DDC by aggregating multiple ($K = 4$) linear function $\tilde{\boldsymbol{W}}^T\boldsymbol{x} + \tilde{\boldsymbol{b}}$ with different directions is represented as:

$$\begin{aligned} &\boldsymbol{y} = f\left(\tilde{\boldsymbol{W}}^T(\boldsymbol{x})\,\boldsymbol{x} + \tilde{\boldsymbol{b}}(\boldsymbol{x})\right) \\ &\tilde{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{k=1}^{K} \alpha_k(\boldsymbol{x})\,\tilde{\boldsymbol{W}}_k,\ \tilde{\boldsymbol{b}}(\boldsymbol{x}) = \sum_{k=1}^{K} \alpha_k(\boldsymbol{x})\,\tilde{\boldsymbol{b}}_k \\ &s.t.\quad 0 \le \alpha_k(\boldsymbol{x}) \le 1,\ \sum_{k=1}^{K} \alpha_k(\boldsymbol{x}) = 1, \end{aligned} \tag{2}$$

where $\alpha_k(\boldsymbol{x})$ denotes the learnable weight for the kth linear function $\tilde{\boldsymbol{W}}_k^T\boldsymbol{x} + \tilde{\boldsymbol{b}}_k$, which is obtained by a lightweight attention block. Define the indices of $\tilde{\boldsymbol{W}}_k$ as (u_k, v_k), $u_k = 0, 1, 2,\ v_k = 0, 1, 2,\ k = 1, 2, 3, 4$, then

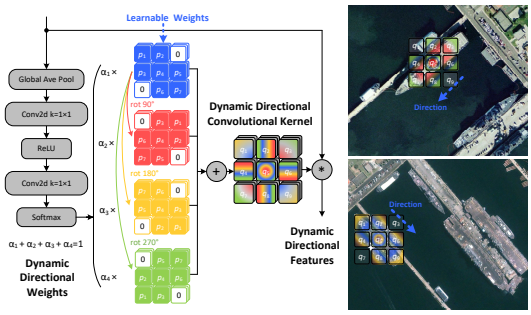

Fig. 1. The principle of the proposed dynamic directional convolution (DDC).

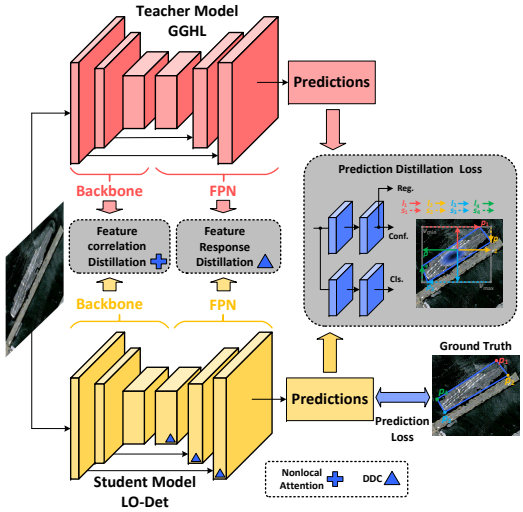

Fig. 2. The framework of the proposed direction-adaptive knowledge distillation (DKD).

$$\begin{bmatrix} u_k \\ v_k \end{bmatrix} = \begin{bmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \times \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \begin{bmatrix} 2t_u \\ 2t_v \end{bmatrix}, \tag{3}$$

where $\beta = \frac{(k-1)\pi}{2}$ denotes the rotated angle of the weight matrix. When $k = 1, 2, 3, 4,\ t_u = 0, 0, 1, 1,\ t_v = 0, 1, 1, 0$, respectively. Define the element at (u_k, v_k) of the weight matrix $\tilde{\boldsymbol{W}}_k$ as $\tilde{W}_k^{u,v}$, $\tilde{W}_1^{0,0} = \tilde{W}_1^{2,2} = 0$. Since the size of the convolution kernel is very small (usually 3×3), compared to the scheme of fusing the larger-size multi-directional feature maps, the cost for the kernel aggregating and the attention in the proposed DDC is very cheap. In addition, like ordinary convolution, the proposed DDC also has the functions of dilated convolution and grouped convolution. This work chooses the state-of-the-art lightweight RSOD model LO-Det [4] as the baseline, and replaces the 3×3 convolution in its CSA-DRF module with the proposed DDC. The number of channels, the dilated rates, and the number of groups are the same as those of the original 3×3 convolutions in LO-Det.

2.2. Direction-adaptive Knowledge Distillation

The proposed DKED framework shown in Fig. 2 consists of three distillation modules located at different locations

of the CNN-based detector for guiding the student model to learn different knowledge from the teacher model.

1) Feature correlation distillation (FCD). Remote sensing images have rich global structural information, and the objects in the same category have similar characteristics, which are ignored by existing KD methods [10, 13]. Therefore, a distillation method improved from non-local feature correlation [15] is employed to distill the global correlation features on the backbone of teacher and student detectors. Define the output feature tensor of backbone as $\boldsymbol{Y} \in \mathbb{R}^{W\times H\times C}$, which consists of the feature vectors $\boldsymbol{y}$ of all neurons at the output convolutional layer of the backbone. W and H represent the width and height of the feature map, C represent the length (channel) of the feature vector $\boldsymbol{y}$. Define the vectorized $\boldsymbol{Y}$ as $vec(\boldsymbol{Y}) \in \mathbb{R}^{WHC\times 1}$, the non-local correlation feature vector $\boldsymbol{z}$ is represented as

$$\boldsymbol{z} = g\left(vec(\boldsymbol{Y}), vec(\boldsymbol{Y})^T\right) h\left(vec(\boldsymbol{Y})\right), \tag{4}$$

where $g(\cdot)$ denotes the correlation function that computes the affinity of each position of the two feature vectors. Here, the dot-product form is used in $g(\cdot)$, that is,

$$g\left(vec(\boldsymbol{Y}), vec(\boldsymbol{Y})^T\right) = vec(\boldsymbol{Y})^T vec(\boldsymbol{Y}). \tag{5}$$

where $h(\cdot)$ denotes a learnable transformation on the input, which is implemented by a 1×1 convolution with nonlinear activation. So, the loss function of feature correlation distillation is represented as

$$L_{fcd} = \|\boldsymbol{z}_T - \boldsymbol{z}_S\|_2^2 \tag{6}$$

where $\boldsymbol{z}_T \in \mathbb{R}^{WHC\times 1}$ and $\boldsymbol{z}_S \in \mathbb{R}^{WHC\times 1}$ denote the non-local correlation feature vectors of the teacher model and student model, respectively.

2) Direction-adaptive feature response distillation (DFRD). Considering the arbitrary oriented features, the proposed DDC is used to extract the multi-scale direction-adaptive features on the feature pyramid networks (FPNs) of teacher model and student model, respectively. Second, a DFRD contrastive loss function is designed to guide the student model to learn the similar features response values to those of the teacher model. Define the direction-adaptive feature tensors of teacher model and student model as $\boldsymbol{Y}'_T \in \mathbb{R}^{W\times H\times C}$ and $\boldsymbol{Y}'_S \in \mathbb{R}^{W\times H\times C}$. Note that both the $\boldsymbol{Y}'$ here and the $\boldsymbol{Y}$ in FCD denote the feature tensors, but they are located in different components of the CNN model and play different roles. The loss of DFRD is

$$L_{drfd} = \left\|\boldsymbol{G} \odot \left(\boldsymbol{Y}'_T - \boldsymbol{Y}'_S\right)\right\|_2^2 \tag{7}$$

where $\odot$ denotes the Hadamard product. $\boldsymbol{G}$ denotes a tensor consisting of prior Gaussian heatmaps generated from the training label [16].

3) Prediction distillation (PD). Define the predicted locations of teacher model and student model as $\boldsymbol{r}_T$ and $\boldsymbol{r}_S$, which are encoded by the four-vertex-coordinates of predicted OBBs. Define the ground truth of OBBs' vertex coordinates as $\boldsymbol{r}_{GT}$. Tensors $\boldsymbol{r}_T$, $\boldsymbol{r}_S$, and $\boldsymbol{r}_{GT} \in \mathbb{R}^{8\times M}$, where M represents the number of objects. Then, the OBB localization distillation loss is

$$\begin{aligned} L_{loc} &= JOL_{obb}\left(\boldsymbol{r}_S, \boldsymbol{r}_{GT}\right) + \lambda \times L_{obb}\left(\boldsymbol{r}_T, \boldsymbol{r}_S, \boldsymbol{r}_{GT}\right) \\ L_{obb}\left(\boldsymbol{r}_T, \boldsymbol{r}_S, \boldsymbol{r}_{GT}\right) &= \begin{cases} JOL_{obb}\left(\boldsymbol{r}_S, \boldsymbol{r}_{GT}\right), \\ \quad \text{if } JOL_{obb}\left(\boldsymbol{r}_S, \boldsymbol{r}_{GT}\right) + \delta \\ \quad > JOL_{obb}\left(\boldsymbol{r}_T, \boldsymbol{r}_{GT}\right) \\ 0, \quad otherwise \end{cases}, \end{aligned} \tag{8}$$

where $JOL_{obb}(\cdot)$ represents the OBB localization sub-loss of the joint-optimization loss (JOL) presented by the GGHL [16], which is a state-of-the-art remote sensing object detection model. More specifically, JOL_{obb} measures the object localization loss by calculating the sum of Intersection over Union (IoU) error and corresponding coordinate error of each predicted bounding box and ground truth bounding box. For its detailed form, see paper [16]. λ and δ are hyperparameters, which are set to 0.6 and 0.5. This OBB localization distillation combined loss encourages the student to be close to or better than teacher in terms of OBB localization. Define the classification output of teacher model and student model as $\boldsymbol{c}_T$ and $\boldsymbol{c}_S$; define the classification ground truth of training samples as $\boldsymbol{c}_{GT}$. Tensors $\boldsymbol{c}_T$, $\boldsymbol{c}_S$, and $\boldsymbol{c}_{GT} \in \mathbb{R}^{Cls\times M}$, where Cls represents the number of objects. Then, the classification distillation loss is

$$L_{cls} = \lambda \times JOL_{cls}\left(\boldsymbol{c}_S, \boldsymbol{c}_{GT}\right) + (1-\lambda) \times JOL_{cls}\left(\boldsymbol{c}_T, \boldsymbol{c}_S\right), \tag{9}$$

where $JOL_{cls}(\cdot)$ represents the classification sub-loss of the JOL presented by the GGHL [16]. It measures the classification loss by calculating the sum of the improved binary cross entropy (BCE) between the prediction and the ground truth of each category of each object. Therefore, the total loss of the proposed DKED is

$$L_{DKED} = \xi \times (L_{fcd} + L_{dfrd}) + L_{loc} + L_{cls}. \tag{10}$$

The weight ξ of the knowledge distillation loss with indirect supervision is set to be lower than the weight of the object detection loss with direct supervision. The optimal hyperparameter $\xi = 0.1$ is set through experiments, and the detailed experimental results are listed in Table 4.

3. EXPERIMENTS AND DISCUSSIONS

In this section, experiments on public datasets are conducted to verify the effectiveness of the proposed DKED.

3.1. Experimental Conditions

1) Experimental platforms. In order to evaluate the performance of the proposed DKED comprehensively, a variety of experimental platforms are used, including: a) a computer with an NVIDIA GeForce RTX 3090 GPU (24GB); b) an

Table 1. Ablation experiments on the HRSC2016 dataset

Modules	Teacher GGHL [16]	Student LO-Det [4]	DDC	FCD	DFRD	PD	mAP (%)	Speed 1 (fps)	Speed 2 (fps)	Speed 3 (fps)
	✓						87.30	42.39	-	-
Selected Module(s)		✓					80.05	62.12	7.34	23.51
		✓	✓				82.35	62.18	7.37	23.50
		✓		✓		✓	82.69	62.20	7.37	23.50
		✓	✓		✓	✓	84.21	62.16	7.39	23.51
		✓	✓	✓	✓	✓	85.17	62.16	7.38	23.51

Note: Speed 1 is the speed on RTX 3090 GPU, speed 2 is the speed on NVIDIA Jetson TX2, Speed 3 is the speed on NVIDIA Jetson AGX Xavier. DDC: Dynamic Directional Convolution; FCD: Feature correlation distillation; DFRD: Direction-adaptive feature response distillation; PD: Prediction distillation.

Table 2. Comparative results on the HRSC2016 dataset

Methods	Backbone	mAP	Speed (fps)
R^2CNN [19]	ResNet101	73.12	6.70
ROI Trans. [20]	ResNet101	85.99	7.83
Gliding Vertex [21]	ResNet101	87.33	10.52
R^3Det [22]	ResNet101	89.26	12.00
LO-Det [4]	MobileNetv2	80.05	62.12
DKED	MobileNetv2	85.17	**62.16**
GGHL [16]	DarkNet53	87.30	42.39
GGHL-DKED	DarkNet53	**93.45**	42.39

Note: Bold font indicates the best results.

NVIDIA Jetson TX2 embedded device; c) an NVIDIA Jetson AGX Xavier embedded device.

2) Datasets. The widely-used public RSOD datasets HRSC2016 [17] and DOTA [18] are used. HRSC2016 [17] is a ship detection dataset with a total of 1061 images, in which the training set, validation set, and testing set include 436, 181, and 444 images, respectively. DOTA [18] dataset contains 2806 aerial images from 800 × 800 pixels to 4000 × 4000 pixels, in which more than 188,000 objects falling into 15 categories are annotated.

3) Baseline models. In this work, the high-performance RSOD model GGHL [16] is used as the teacher model and the lightweight model LO-Det [4] is used as the student model.

3.2. Ablation Experiments and Discussions

Ablation results of each component of the proposed DKED on the HRSC2016 [17] dataset are listed in Table 1. The proposed DDC improves the mean Average Precision (mAP) of LO-Det [4] by 2.30% with a faster detection speed. The distillation strategy of FCD+PD and DDC+DRFD+PD increases the mAP of LO-Det by 2.64% and 4.16%, respectively. The proposed DKED reduces the performance gap between the teacher model and the lightweight student model from 7.25% to 2.13%, which makes LO-Det gain a 5.12% mAP increase in total without any speed loss. The ablation experiments verify the effectiveness of each component of the proposed DKED from quantitative perspectives.

Table 2 and Table 3 list the comparative experimental results of the proposed DKED and high-performance models on the HRSC2016 dataset and DOTA dataset, respectively. Experimental results show that the mAP of the proposed DKED is close to that of the state-of-the-art models while its detection speed is much faster than that of the high-performance

Table 3. Comparative results on the DOTA dataset

Methods	Backbone	mAP	Speed (fps)
ROI Trans. [20]	ResNet101	67.74	7.80
SCRDet [2]	ResNet101	72.61	9.51
Gliding Vertex [21]	ResNet101	75.02	13.10
R^3Det [22]	ResNet101	76.47	10.53
BBAVectors [23]	ResNet101	72.32	18.37
LO-Det [4]	MobileNetv2	66.17	62.00
DKED	MobileNetv2	71.39	**62.08**
GGHL [16]	DarkNet53	76.95	42.39
GGHL-DKED	DarkNet53	**77.43**	42.39

Note: Bold font indicates the best results.

models. In addition, Table 4 discusses the influence of different hyperparameter settings on the model performance. The experimental results show that the model achieves the optimal performance when $\lambda = 0.6$, $\delta = 0.5$, and $\xi = 0.1$. The visual experimental results of the proposed DKED on the HRSC2016 dataset and DOTA dataset are shown in Fig. 3. In summary, extensive experiments on public datasets have verified the effectiveness of the proposed DKED from improving the performance of the lightweight RSOD model designed for embedded platforms.

Fig. 3. The visual experimental results of the proposed DKED on (a) the HRSC2016 dataset and (b) the DOTA dataset.

Table 4. Experiments with different values of λ, δ, and ξ on the HRSC2016 dataset

λ ($\delta = 0.5, \xi = 0.1$)	mAP	δ ($\lambda = 0.6, \xi = 0.1$)	mAP	ξ ($\lambda = 0.6, \delta = 0.5$)	mAP
0.2	83.84	0.3	83.42	0.1	**85.17**
0.4	84.25	0.5	**85.17**	0.3	85.01
0.6	**85.17**	0.7	84.11	0.5	84.13

Note: Bold indicates the best result. When evaluating one variable, the other variables are fixed to take the optimal value.

4. CONCLUSIONS

In this paper, a novel knowledge distillation method DKED has been proposed to improve the performance of lightweight RSOD models. In the proposed DKED, the developed DDC enhances the arbitrary-oriented feature extraction of the lightweight model, and the designed DKD guides the lightweight model to learn the intrinsic knowledge from the high-performance model. The experiments on public datasets have demonstrated that each component proposed in DKED is valid, and the detection performance of lightweight model can be greatly improved without any inference costs by using the proposed DKED.

5. REFERENCES

[1] 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*, vol. 161, pp. 294–308, 2020.

[2] 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 IEEE International Conference on Computer Vision*, Seoul, South Korea, Oct. 2019, pp. 8232–8241.

[3] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia, "ReDet: A rotation-equivariant detector for aerial object detection.," *arXiv preprint arXiv:2103.07733*, 2021.

[4] Zhanchao Huang, Wei Li, Xiang-Gen Xia, Hao Wang, Feiran Jie, and Ran Tao, "LO-Det: Lightweight oriented object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2021.

[5] Peng Ding, Ye Zhang, Wei-Jian Deng, Ping Jia, and Arjan Kuijper, "A Light and Faster Regional Convolutional Neural Network for Object Detection in Optical Remote Sensing Images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 141, pp. 208–218, 2018.

[6] Nan Wang, Bo Li, Xingxing Wei, Yonghua Wang, and Huanqian Yan, "Ship detection in spaceborne infrared image based on lightweight CNN and multisource feature cascade decision," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4324–4339, 2021.

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018, pp. 4510–4520.

[8] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in *Proceedings of the European Conference on Computer Vision*, Munich, Germany, Sept. 2018, pp. 122–138.

[9] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[10] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker, "Learning efficient object detection models with knowledge distillation," in *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, vol. 30, pp. 742–751.

[11] Quanquan Li, Shengying Jin, and Junjie Yan, "Mimicking very efficient network for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7341–7349.

[12] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng, "Distilling object detectors with fine-grained feature imitation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4933–4942.

[13] Zhaohui Zheng, Rongguang Ye, Ping Wang, Jun Wang, Dongwei Ren, and Wangmeng Zuo, "Localization distillation for object detection.," *arXiv: Computer Vision and Pattern Recognition*, 2021.

[14] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou, "General instance distillation for object detection.," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7842–7851.

[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local Neural Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018, pp. 7794–7803.

[16] Zhanchao Huang, Wei Li, Xiang-Gen Xia, and Ran Tao, "A general gaussian heatmap labeling for arbitrary-oriented object detection," *arXiv preprint arXiv:2109.12848*, 2021.

[17] Zikun Liu, Liu Yuan, Lubin Weng, and Yang Yiping, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *6th International Conference on Pattern Recognition Applications and Methods*, 2017, pp. 324–331.

[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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018, pp. 3974–3983.

[19] 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.

[20] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu, "Learning RoI transformer for oriented object detection in aerial images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Rcognition*, Long Beach, CA, USA, June 2019, pp. 2849–2858.

[21] 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*, pp. 1–1, Feb. 2020.

[22] Xue Yang, Qingqing Liu, Junchi Yan, and Ang Li, "R3Det: Refined single-stage detector with feature refinement for rotating object.," *arXiv preprint arXiv:1908.05612*, 2019.

[23] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris N. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *2020 IEEE/CVF Winter Conference on Applications of Computer Vision*, Dec. 2020, pp. 2150–2159.