ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413925

OBJECT-ORIENTED RELATIONAL DISTILLATION FOR OBJECT DETECTION

*Shuyu Miao, Rui Feng**

School of Computer Science, Fudan University
Shanghai Key Laboratory of Intelligent Information Processing

ABSTRACT

Object detection models have achieved increasingly better performance based on more complex architecture designs, but the heavy computation limits their further widespread application on the devices with insufficient computational power. To this end, we propose a novel *Object-Oriented Relational Distillation* (**OORD**) method that drives small detection models to have an effective performance like large detection models with constant efficiency. Here, we introduce to distill relative relation knowledge from teacher/large models to student/small models, which promotes the small models to learn better soft feature representation by the guiding of large models. OORD consists of two parts, i.e., *Object Extraction* (**OE**) and *Relation Distillation* (**RD**). OE extracts foreground features to avoid background feature interference, and RD distills the relative relations between the foreground features through graph convolution. Related experiments conducted on various kinds of detection models show the effectiveness of OORD, which improves the performance of the small model by nearly 10% without additional inference time cost.

Index Terms— Object detection, object-oriented relational distillation, knowledge distillation

1. INTRODUCTION

As a fundamental vision task, object detection has attracted widespread attention and experimentally solved a variety of real-world scenario problems. Although the accuracy of object detection models has been greatly improved with increasingly complex architecture, they are not competent to adopt in edge devices or high real-time scenarios due to heavy computation. To alleviate this problem, some works explored the use of network pruning [1, 2, 3, 4] that pruned redundant connections in large models, and quantification [5, 6, 7, 8] that reduced the model size and parameters. However, these approaches are still far from real-time requirement and require dedicated designs. Another efficient end-to-end approach is *knowledge distillation* (KD) [9], in which the teacher/large models guide the student/small models to learn excellent feature representation to improve their performance. Thus, we can deploy small models with competitive accuracy but less computation to meet these limited scenarios.

Recently, several works have explored the applications of KD in the object detection task. KD firstly and successfully demonstrated for the multi-class object detection problem [10] via an end-to-end trainable framework to learn compact multi-class object detection models. However, full feature-based distillation limited the performance of the student model due to noise background interference. A mimic method [11] distilled the feature after proposal sampling instead of the whole feature, but it was only suitable for the two-stage Region Proposal Networks (RPN)-based approaches and it was a hard feature representation method. Wang et. al. [12] introduced a fine-grained feature imitation method to exploit the cross-location discrepancy of feature response, yet this method was still disturbed by certain background features and only facilitated anchor-based detection models. Two structured distillation schemes [13] pair-wise distillation and holistic distillation were studied to distill structured knowledge from large models to small models, which just focused on segmentation-based anchor-free detection models. In conclusion, although these methods solve the problem to some extent, they have their own defects. The greatest challenges of knowledge distillation in object detection lie in the following issues, i.e., avoiding background distractions (*issue.(1)*), distilling soft feature representation (*issue.(2)*), and widely applicable to various detection models (*issue.(3)*).

To address the above issues, we propose a novel *Object-Oriented Relational Distillation* method, called **OORD**. It aims to distill the soft relative relation among foreground features from teacher/large detection models to student/small models. Our method consists of two parts, i.e., *Object Extraction* (**OE**) and *Relation Distillation* (**RD**). OE generates an object-oriented mask and further extracts foreground features by means of the ground-truth label, which avoids the interference of background features to address the *issue.(1)*. RD distills the relative relation based on foreground features with graph convolution, which guides small models to imitate soft feature representation like large models to attack the *issue.(2)*. OORD is a feature-based distillation method and not dependent on a specific model structure, which can be utilized into

* Corresponding Author. This work was supported (in part) by the Science and Technology Commission of Shanghai Municipality (No. 20511100800, No. 20511101502, and No. 20511101704).

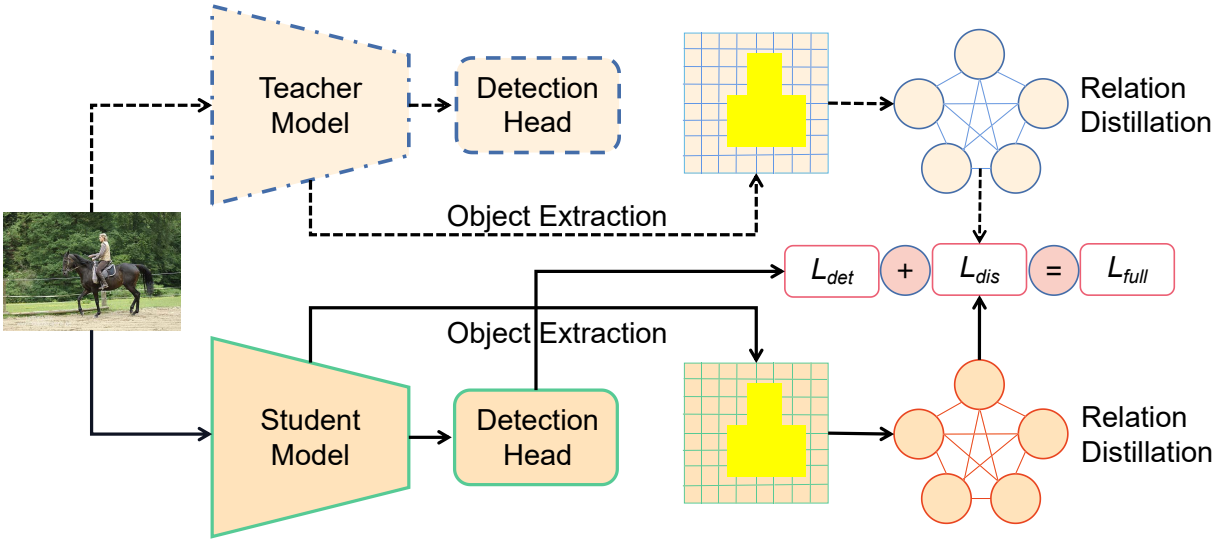

Fig. 1. An overview of our novel *Object-Oriented Relational Distillation* (**OORD**). OORD consists of *Object Extraction* that extracts the foreground features and *Relation Distillation* that distills the relative relation among foreground features. "L_{det}" is the original detection loss, "L_{dis}" is the distillation loss, and "L_{full}" is the full loss in the training.

various kinds of detection models to tackle the *issue.(3)*. The experimental results on widely used PascalVOC and MSCOCO datasets validate that our OORD can achieve a significant performance improvement of student models at the constant inference speed, promoting their applications on the devices with insufficient computational power.

The main contributions of our work can be summarized as: (1) A novel *Object-Oriented Relational Distillation* method, named OORD, is proposed to improve the performance of small detection models comparable to large models; (2) To the best of our knowledge, OORD is the first knowledge distillation method to consider the soft relative relation of features through graph convolution in the object detection task; and (3) Related experiments fully demonstrate the effectiveness of our proposed OORD, which significantly improves the performance of small models by about 10% without inference time cost.

2. PROPOSED METHODS

2.1. Overview

The overview of the Object-Oriented Relational Distillation (OORD) is shown in Figure 1. The teacher/large models with more parameters and computations tend to have better feature representation ability but low speed, while the student/small models with fewer parameters and computations tend to have worse feature representation ability but high speed. The objective of OORD is to supervise the learning of student model by teacher model. Teacher and student models take the same RGB image $\boldsymbol{I} \in \mathbb{R}^{W_0 \times H_0 \times C_0}$ as the input; then $F_t \in \mathbb{R}^{W_t \times H_t \times C_t}$ is yielded via the backbone of teacher model, and $F_s \in \mathbb{R}^{W_s \times H_s \times C_s}$ is generated via the backbone of student model. OORD aims to learn better F_s representation driven by F_t.

Firstly, the foreground features F_s^f and F_t^f of F_s and F_t are extracted by means of ground-truth label via proposed *Object Extraction* (OE), which is denoted as $\mathcal{O}(\cdot)$. Secondly, F_t^f distills soft relative relation among features to F_s^f through graph convolution via proposed *Relation Distillation* (RD), which is denoted as $\mathcal{R}(\cdot)$. Finally, F_s^f imitates the feature representation of F_t^f with model training by MSE loss function. Thus, the knowledge distillation loss can be formulated as

$$L_{dis} = \frac{1}{2N} \sum ||\mathcal{R}(\mathcal{O}(F_t)) - \mathcal{R}(\mathcal{O}(F_s))||^2, \qquad (1)$$

where N is the number of feature value of foreground feature. L_{dis} and original detection loss L_{det} are added as the full detection loss function L_{full}.

Proposed OORD is a feature-based distillation method, thus it can be widely used in object detection models with different frameworks, like anchor-based, anchor-free, one-stage, and tow-stage detection models.

2.2. Object Extraction

Wang et. al. [12] pointed out that the profit of distilling the whole feature map was very limited because only a few of features contained the objects and most of features were noise background interference in the detection task. They proposed to distill the cross-location discrepancy of feature response via the Intersection over Union (IoU) threshold between the

predesigned anchor and ground truth box. However, many of the anchor areas cover no objects, which still causes background interference. Object Extraction (OE) aims to extract foreground features that just contain objects for distillation.

Generally, F_s and F_t have the same size due to the same input. If they have different sizes, F_s is firstly resized to the size of F_t, then $\{F_s, F_t\} \in \mathbb{R}^{W\times H\times C}$ can be obtained. For simplicity, we define the Ground Truth (GT) box as $\mathcal{B} := \{top_i, left_i, bottom_i, right_i\}_{i=1:n}$, where $\{top_i, left_i, bottom_i, right_i\}$ is the coordinate value of the i-th GT box and n is the number of the GT boxes. Matrix masks $\{\mathcal{M}_i \in \mathbb{R}^{W\times H}\}_{i=1:n}$ are predefined as 0 corresponding to each GT box. $\mathcal{M}_i$ is updated according to the corresponding GT box by $\mathcal{M}_i[left_i : right_i, top_i : bottom_i] = 1$. The final object-oriented mask $\mathcal{M}$ is obtained by $\mathcal{M} = \sum_1^n \mathcal{M}_i$. F_t^f via $\mathcal{O}(F_t)$ is gained by $F_t * \mathcal{M}$, and F_s^f via $\mathcal{O}(F_s)$ by $F_s * \mathcal{M}$ in the Equ. (1) .

Based on the above operations, the foreground features F_t^f and F_s^f are yielded, which are more suitable for feature distillation instead of the full features F_t and F_s without background effect. In addition, models should focus more on the areas where there are overlapping objects, and OE achieves this purpose by the superposition of matrix masks. Thus, OE can cover the features of foreground objects effectively for the following relative relation distillation.

2.3. Relation Distillation

The teacher and student models tend to have their own feature distribution, and it is not optimal for the student model to learn the features of the teacher model rigidly. One more reasonable way is to learn the soft relative feature relation. Relation Distillation (RD) distills the soft feature relation between teacher and student models via graph convolution.

Graph Convolutional Network (GCN) [14] enables to learn unstructured topological relationships, which can be defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. $\mathcal{V} = \{v_i\}_{i=1:M}$ and $\mathcal{E} = \{e_{a,b}\}_{\{a,b\}=1:M}$ are the graph nodes and edges, and M is the number of graph nodes and $e_{a,b}$ means the graph edge between node v_a and v_b. $\mathcal{G}$ takes a feature matrix $\mathcal{F}$ and an adjacency matrix $\mathcal{A}$ as the input. $\mathcal{F} \in \mathbb{R}^{M\times D}$ denotes the feature description of the graph, which is stacked by every node feature with D feature dimension. $\mathcal{A} \in \mathbb{R}^{M\times M}$ indicates the representative description of the graph structure in matrix form, where $\mathcal{A}_{a,b}$ corresponds to the graph edge $e_{a,b}$. The transmission of information in the graph can be expressed as

$$\mathcal{Z} = \alpha(\tilde{D}^{-\frac{1}{2}}\tilde{\mathcal{A}}\tilde{D}^{-\frac{1}{2}}\mathcal{F}\mathcal{W}), \tag{2}$$

where $\tilde{\mathcal{A}} = \mathcal{A} + I$; I is the identity matrix; $\tilde{D}$ is the diagonal node degree matrix of $\tilde{\mathcal{A}}$; $\mathcal{W}$ is a learnable weight matrix; and α is an activation function.

RD distills the feature relation of F_t^f to F_s^f, where F_t^f and F_s^f aims to construct the relative feature relation in the spatial dimension. Here, the graph nodes are the features in the $\{H, W\}$ dimension, and the node feature representation is in the channel $\{C\}$ dimension. The feature matrix $\mathcal{F}_t$ and $\mathcal{F}_s$ of the teacher and student models are generated with the same size of $HW \times C$. To describe the graph structure, the graph edge $\mathcal{A}_{i\to j}$ between the node i and node j is calculated by the cosine distance between node features with the same size of $HW \times HW$ as

$$\mathcal{A}_{i\to j} = \frac{\sum_{k=1}^{C}(\mathcal{F}_{i,k} \times \mathcal{F}_{j,k})}{\sqrt{\sum_{k=1}^{C}(\mathcal{F}_{i,k})^2} \times \sqrt{\sum_{k=1}^{C}(\mathcal{F}_{j,k})^2}}. \tag{3}$$

The adjacency matrix $\mathcal{A}_t$ and $\mathcal{A}_s$ of the teacher and student models are derived by Equ. (3). The final relational output $\mathcal{Z}_t$ and $\mathcal{Z}_s$ of the teacher and student models are modeled by the Equ. (2). $\mathcal{Z}_t$ and $\mathcal{Z}_s$ are drawn closer through the Equ. (1) to distill the relation knowledge.

3. EXPERIMENTS

3.1. Datasets and Settings

Two widely used datasets *PascalVOC* [15] and *MSCOCO* [16] are used in our experiments. PascalVOC07 is a challenging dataset in daily life with 20 classes, in which we train the model on 5k training images and 12k annotated objects, and test on the PascalVOC07 `test` set. MSCOCO is a larger dataset in daily life with 80 classes, in which we train the model on 80k training images and 35k validation images, and test on the MSCOCO `test-dev` set. The Average Precision (AP) and mean Average Precision (mAP) are adopted as the evaluation metrics to validate the model performance.

To demonstrate the effectiveness of our proposed OORD, we conduct experiments on different teacher and student detection models on anchor-based, anchor-free, one-stage, and two-stage models. Teacher models are pretrained and fix parameters during training. The objective of the relation distillation is to drive the student model to learn better feature representation by aid of teacher model like classical knowledge distillation [9]. The experimental settings of OORD are similar as the original student models to keep fair. Note that all our experiments are implemented on Pytorch framework.

3.2. Overall Performance

We evaluate the performance of our proposed OORD on *PascalVOC* and *MSCOCO* on anchor-based, anchor-free, one-stage, and two-stage detection models.

Results on PascalVOC. Table 1 shows the experimental results of proposed OORD on the *PascalVOC*. We verify the performance of OORD on Faster RCNN [17], which is a classical two-stage or called anchor-based model. Faster RCNN with VGG16 and ResNet101 as the backbone are the teacher models, and Faster RCNN with VGG11 and ResNet50 are the student models, respectively.

Model	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
VGG16	70.4	70.9	78.0	67.8	55.1	53.2	79.6	85.5	83.7	48.7	78.0	63.5	80.2	82.0	74.5	77.2	43.0	73.7	65.8	76.0	72.5
VGG11	59.6	67.3	71.4	56.6	44.3	39.3	68.8	78.4	66.6	37.7	63.2	51.6	58.3	76.4	70.0	71.9	32.2	58.1	57.8	62.9	60.0
VGG11-I	67.6	72.5	73.8	62.8	53.1	49.2	80.5	82.7	76.8	44.8	73.5	64.3	72.6	81.1	75.3	76.3	40.2	66.3	61.8	73.4	70.6
VGG11-O	68.9	73.9	75.9	63.2	53.4	51.2	83.1	84.0	78.1	45.8	73.7	66.1	73.8	81.5	77.2	77.1	40.8	68.0	62.8	74.7	71.8
	+9.3	**+6.6**	**+4.5**	**+6.6**	**+9.1**	**+11.9**	**+14.3**	**+5.6**	**+11.5**	**+8.1**	**+10.5**	**+14.5**	**+15.5**	**+5.1**	**+7.2**	**+5.2**	**+8.6**	**+9.9**	**+5.0**	**+11.8**	**+11.8**
Res101	74.4	77.8	78.9	77.5	63.2	62.6	79.2	84.4	85.6	54.5	81.5	68.7	85.7	84.6	77.8	78.6	47.1	76.3	74.9	78.8	71.2
Res50	69.1	68.9	79.0	67.0	54.1	51.2	78.6	84.5	81.7	49.7	74.0	62.6	77.2	80.0	72.5	77.2	40.0	71.7	65.5	75.0	71.0
Res50-I	72.0	71.5	80.6	71.1	57.0	52.4	82.1	90.0	82.7	51.6	74.5	66.2	82.3	82.3	75.7	78.3	43.5	79.6	69.1	77.3	72.1
Res50-O	73.1	72.9	80.8	74.1	59.7	52.9	82.5	91.3	83.7	51.9	75.5	67.8	82.9	83.2	76.8	78.5	45.0	79.4	71.2	77.7	72.2
	+4.0	**+4.0**	**+1.8**	**+7.1**	**+5.6**	**+1.7**	**+3.9**	**+6.8**	**+2.0**	**+2.2**	**+1.5**	**+5.2**	**+5.7**	**+3.2**	**+4.3**	**+1.3**	**+5.0**	**+7.7**	**+5.7**	**+2.7**	**+1.2**

Table 1. Experiments on *PascalVOC07* dataset with Faster RCNN model. VGG16 and ResNet101 are the teacher models, and VGG11 and ResNet50 are the student models. VGG11-I is the results with the paper [12], and VGG11-O is the results with our proposed OORD.

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
FCOS	ResNeXt-101	42.7	62.2	46.1	26.0	45.6	52.6	7.7
FCOS	MobileNetV2	31.4	49.2	33.3	17.1	33.5	38.8	22.3
FCOS-S	MobileNetV2	34.1	52.2	36.4	19.0	36.2	42.0	22.3
FCOS-O	MobileNetV2	**35.6**	**54.2**	**37.0**	**19.9**	**37.8**	**43.8**	22.3

Table 2. Experiments and inference speed on the COCO test-dev with FCOS model. FCOS-S is the results with the paper [13], and FCOS-O is the results with our proposed OORD.

Method	Model	mAP
1	Faster RCNN + VGG16	70.4
2	Faster RCNN + VGG11	59.6
3	Faster RCNN + VGG11 + OE	66.2
4	Faster RCNN + VGG11 + RD	63.4
5	Faster RCNN + VGG11 + OE+ RD	**68.9**

Table 3. Ability study experiments on *PascalVOC07* dataset with Faster RCNN model. OE is our proposed Object Extraction (OE), and RD is our proposed Relation Distillation (RD).

It can be seen from the table that for the VGG-style model, the mAP of student model with the VGG11 backbone is improved by 9.3% from 59.6% to 68.9%. For the ResNet-style model, the mAP of the student model with ResNet50 as the backbone is boosted from 69.1% to 73.1% by 4.0% improvement. The experimental results validate that our method is effective for both the two-stage detection model and the anchor-based detection model. Note that the inference speed of the models with VGG11 backbone is same because knowledge distillation has no effect on the inference time.

Results on *MSCOCO*. Table 2 presents the experimental results on *MSCOCO*. We validate the effectiveness of OORD on FCOS [18], which is a classical one-stage or anchor-free detection model. FCOS with ResNeXt-101 backbone is the teacher model, and FCOS with MobileNetV2 backbone is the student model.

It can be viewed from the table that with our distillation method FCOS improves the performance of AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L by 4.2%, 5.0%, 3.7%, 2.8%, 4.3%, and 5.0%, respectively, which outperforms that of [13]. More importantly, our OORD does not bring any additional speed cost with the same speed of 22.3 FPS.

3.3. Ablation Study

To verify the reasonability and reliability of our proposed O-ORD, we perform the related ablation studies shown as Table 3. Related experiments are conducted on *PascalVOC07* with Faster RCNN. The teacher model is based on VGG16, and the student model is based on VGG11.

It can be found from the table that when we distill the model just with proposed Object Extraction, the performance is improved by 6.6%. It proves that extracting foreground features for distillation is important without the interfere of background feature. When the model is distilled just with proposed Relation Distillation, the mAP is boosted by 3.8%. It verifies that distilling the relative relation between features facilitates the feature learning of student model. The model performance with full OORD is improved most, which shows the effectiveness of OORD.

4. CONCLUSION

In this paper, we propose a novel *Object-Oriented Relational Distillation* (**OORD**) method for object detection to promote student/small detection models to yield better performance like teacher/large models without any additional inference time cost. OORD consists of *Object Extraction* (**OE**) and *Relation Distillation* (**RD**), which distills the relative relation between just foreground features containing objects far away from background feature interfere. OORD is suitable for all kinds of detection models, like one-stage, two-stage, anchor-based, and anchor-free models. We hope that our O-ORD can shew new light on the development of object detection on the devices with insufficient computing power.

5. REFERENCES

[1] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang, "Soft filter pruning for accelerating deep convolutional neural networks," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 7 2018, pp. 2234–2240, International Joint Conferences on Artificial Intelligence Organization.

[2] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell, "Rethinking the value of network pruning," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. 2019, OpenReview.net.

[3] Song Han, J. Pool, John Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 2015.

[4] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li, "Learning structured sparsity in deep neural networks," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, Eds., 2016, pp. 2074–2082.

[5] Song Han, Jeff Pool, John Tran, and William Dally, "Learning both weights and connections for efficient neural network," in *Advances in Neural Information Processing Systems 28 (NeurIPS)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 1135–1143.

[6] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European Conference on Computer Vision (ECCV)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., 2016.

[7] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan, "Deep learning with limited numerical precision," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML)*, 2015, ICML'15, p. 17371746.

[8] Shuchang Zhou, Zekun Ni, Xinyu Zhou, He Wen, Yuxin Wu, and Yuheng Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," 06 2016.

[9] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[10] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker, "Learning efficient object detection models with knowledge distillation," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, pp. 742–751, Curran Associates, Inc.

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

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

[13] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, pp. 1–1, 2020.

[14] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[15] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *In International Journal of Computer Vision (IJCV)*, , no. 2, pp. 303–338, 2010.

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, p. 740755.

[17] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *In Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 2015, pp. 1–10.

[18] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "FCOS: Fully convolutional one-stage object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.