ATTENTION-GUIDED KNOWLEDGE DISTILLATION FOR EFFICIENT SINGLE-STAGE DETECTOR

Tong Wang[1,2], *Yousong Zhu*[1,3], *Chaoyang Zhao*[1], *Xu Zhao*[1], *Jinqiao Wang*[1,2,4] *and Ming Tang*[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing, China
[3] ObjectEye Inc., Beijing, China
[4] NEXWISE Co., Ltd, Guangzhou, China

{tong.wang,yousong.zhu,chaoyang.zhao,xu.zhao,jqwang,tangm}@nlpr.ia.ac.cn

ABSTRACT

Knowledge distillation has been successfully applied in image classification for model acceleration. There are also some works employing this technique to object detection, but they all treat different feature regions equally when performing feature mimic. In this paper, we propose an end-to-end attention-guided knowledge distillation method to train efficient single-stage detectors with much smaller backbones. More specifically, we introduce an attention mechanism to prioritize the transfer of important knowledge by focusing on a sparse set of hard samples, leading to a more thorough distillation process. In addition, the proposed distillation method also provides an easy way to train efficient detectors without tedious ImageNet pre-training procedure. Extensive experiments on PASCAL VOC and CityPersons datasets demonstrate the effectiveness of the proposed approach. We achieve 57.96% and 69.48% mAP on VOC07 with the backbone of 1/8 VGG16 and 1/4 VGG16, greatly outperforming their ImageNet pre-trained counterparts by 11.7% and 7.1% respectively.

Index Terms— Knowledge distillation, model acceleration, object detection

1. INTRODUCTION

Object detection is a fundamental task in computer vision, and is widely applied in intelligent surveillance, autonomous driving, robotics and so on. Thanks to the significant development of convolutional neural networks (CNN) [1, 2, 3], CNN-based object detection pipelines [4, 5, 6] have been proposed successively and made impressive achievements in generic benchmarks. However, the state-of-the-art detectors always rely on deeper and more sophisticated backbone networks, resulting in high computational complexity and huge memory footprint, which greatly hampering their application in resource-constrained devices.

There are already many works devoting to speeding up the CNNs, among which the knowledge distillation is a proven effective method. It tries to transfer the knowledge from a cumbersome teacher model to a small student model to improve performance of the student model and facilitate the deployment. Many researchers devote themselves into this area, trying to design more effective distillation algorithms and apply knowledge distillation to various tasks. Recently, several works have attempted to introduce knowledge distillation to object detection. [7] proposed a weighted cross entropy loss to deal with the misclassification for background samples and exploited hint learning to learn the distribution of neurons in intermediate layers. [8] presented a feature map mimic method by mimicking the features sampled from region of proposals. [9] introduced the quantization operation to discretize the feature map to facilitate the knowledge transfer. [10] combined knowledge distillation with RetinaNet [11]. They proposed ADL to lead the student network to adaptively mimic the teacher's logits.

Although aforementioned methods can effectively improve the performance of small networks, most of them are intended for two-stage detectors, the application of distillation techniques to single-stage detectors has not been well explored yet. Different from two-stage detectors, single-stage detectors capsule all operations into a single network by abandoning the proposal generation and subsequent RoI-wise refinement stage, making the whole process more compact and faster. Despite [11] combines knowledge distillation with RetinaNet, there is still a speed bottleneck due to the relatively large prediction head of RetinaNet. With a view to build extremely fast object detectors, we combine knowledge distillation with classical single-stage detector, Single Shot MultiBox Detector (SSD) [5], making it possible to train small yet accurate detectors.

In addition, it is generally accepted that the quality of feature directly determines the performance of object detectors. We notice that, previous feature mimicking distillation methods for object detection treat different regions of the feature map equally important, which may be sub-optimal for an effective and efficient knowledge transfer process. Our approach comes from the observation that most of the samples are easy to learn during iterative training, and the corresponding features of well-classified samples do not necessarily need to be learned constantly. During training, there always exists some hard samples which can not be handled well by the student itself. Therefore, the student should focus on these samples when learning from the teacher to better cope with them. In other words, the corresponding feature regions of hard samples should be attached more importance during feature mimic process.

Therefore, in this paper, we propose an effective end-to-end distillation framework for single-stage detector SSD. Specifically, we propose to distill the feature maps before the classification and regression branch, and the ground-truth supervision is normally added to the final prediction layers. Additionally, we design an attention-based distillation mechanism to automatically locate the region of interests that are hard to learn and then adaptively adjust the distillation weights for each region. Thereby, the student network can put more emphasis on the current hard regions, which will accelerate the network convergence and make the knowledge transfer easier and more thorough. To verify the effectiveness of our distillation method, we conduct extensive experiments on varies datasets, including PASCAL VOC [12] and CityPersons [13].

To summarize, the contributions of this work are as follows:

1. We propose a novel knowledge distillation method for single-stage detectors, which can efficiently train smaller student detectors without tedious ImageNet pre-training procedure.

2. We propose an attention mechanism for efficient knowledge transfer and a weighted Euclidean loss for integrating the attention map to knowledge distillation.

3. We conduct comprehensive experiments on PASCAL VOC and CityPersons datasets. The results prove the effectiveness and generality of our method.

2. METHODS

In this section, we first introduce the overall architecture of our distillation method. And then, we introduce the proposed spatial attention mechanism in detail. Finally, we elaborate how the attention maps are combined with the distillation loss and how to optimize the whole network in an end-to-end way.

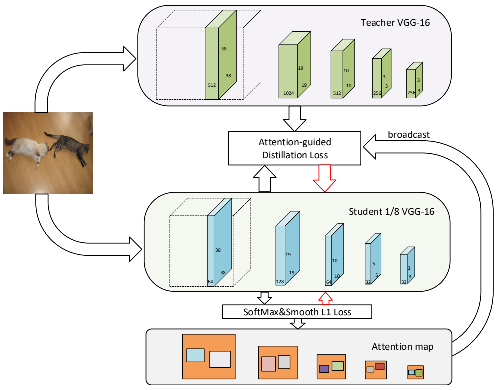

Fig. 1: An overview of the proposed attention-guided knowledge distillation for single-stage detector. The black arrows in the picture show the forward data flow, and the red arrows display the backward gradient flow.

2.1. Architecture

The overall architecture is illustrated in Figure 1. Similar to other distillation methods, it consists of two networks, a well-trained teacher network which has high accuracy on target task and a student network with random initialization parameters. We remove the prediction heads (softmax and bounding box regression layers) of the teacher network, since only the backbone feature maps of the teacher network are useful. The student network has an intact SSD architecture. Both the backbone and prediction heads are preserved. Under our experiment setting, we obtain the small network by cutting off part of convolutional channels from the original network. The 1/n network represents a model that has the same architecture as the original network, but only has 1/n channels for each convolutional layer.

As SSD makes predictions on multiple feature maps, it is intuitive to choose these feature maps as the guided layers. The guided layers are defined as those layers which are responsible for transferring knowledge from the teacher to student. In consideration of the fact that teacher network is usually wider than the student network, an adaptation layer is added to the student's guided layer to deal with the dimension mismatch between teacher's and student's feature maps. More specifically, a convolution layer with 1×1 kernel size followed by ReLU activation function is used as the adaptation layer. The distillation loss combined with attention maps is tailored to optimize the distance between the features of teacher and student. Here, the attention maps are generated according to the classification loss of the student. We will give a comprehensive explanation on how to construct the attention maps later.

2.2. Attention-guided Distillation

Previous works [7, 8] perform knowledge distillation for object detection in a direct way, simply adding an Euclidean loss between the corresponding feature maps of teacher and student. However, for the object detection task, different regions of a feature map are not of equal importance for the student network. For example, most of the samples usually have relatively simple contents and can be well classified by the student. Thus, when learning from the teacher, the student network should focus on the feature regions of those hard samples. The whole procedure can be decomposed into two steps, i.e. distillation weight assignment and attention map construction.

2.2.1. *Distillation Weight Assignment*

The first thing we need to do is to assign a weight to each sample so that the student network can distinguish which samples are more important and pay more attention on them. It is widely accepted that those samples which are hard to learn contribute more to the training. Since the network can't handle these hard samples well, they should be emphasized during training. Inspired by focal loss [11], the classification loss can be viewed as an indicator to measure whether current sample is well learned by the network. Small classification loss indicates that the features are strong enough to make a satisfying prediction, while large loss value shows that the corresponding features are not that discriminative to make a good prediction. Hence, more attention should be paid to these samples with large loss values.

$$w_k = min(w_{max}, \alpha \times (1 - e^{-l_k})^{\beta} \times l_k) \tag{1}$$

We use equation (1) to assign each sample a weight value according to its classification loss. l_k is the classification loss of sample S_k. The weight w_k increases monotonously with the classification loss l_k. Samples with large classification loss values will have large weights during distillation. Additionally, we add an upper bound w_{max} to avoid extremely oversized weights during the early training stage, which will make the training unstable and lead to network divergence. α, β are two introduced hyper-parameters used to control the weight of each sample. Figure 2 shows how the weight-loss curve changes when we fix one of the hyper-parameter and change the other. We can see that hyper-parameter α is equivalent to a scaling factor. And β mainly influences the shape of the curve. Tunning these two parameters together, we are able to adjust the relative weight gap between the easy and hard samples. And thus, the student network can better distinguish easy and hard samples.

2.2.2. *Attention Map Construction*

During distillation, we emphasize those hard samples by guiding the student to focus on the corresponding feature map area of those samples. We accomplish this by integrating the attention mechanism with knowledge distillation. The feature regions corresponding to those hard samples should be given a greater weight during the distillation process. As we have already assigned each sample a weight value, the next step is to construct a spatial attention map according to these weights. Here we adopt a simple strategy. For pixel in position i, j, the weight will be set to 0 if it is not covered by of any samples. If it is situated in the overlap area of different samples, the maximal weight of these samples will be selected as its final weight. The detailed definition can be expressed as equation (2).

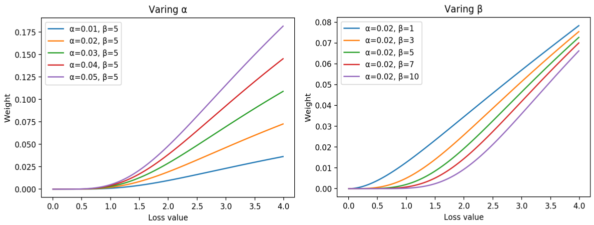

Fig. 2: Weight-loss curves when hyper-parameter α and β are set to different values.

$$a_{i,j} = \begin{cases} 0, & \text{if } G_{i,j} = \emptyset \\ max(w_{g_k}), & \text{if } G_{i,j} \neq \emptyset, g_k \in G_{i,j} \end{cases} \tag{2}$$

For certain feature map, $a_{i,j}$ means the value of attention map A in position i, j. Set S represents the set of all the samples which are selected to compute the loss. For pixel i, j, we define a set $G_{i,j}$, which is a subset of S. The element of $G_{i,j}$ is the sample in S whose corresponding feature region covers pixel i, j. We use g_k to denote the element of $G_{i,j}$. w_{g_k} is the weight calculated by equation (1) of sample g_k. After the operations above, a spatial attention map with $1 \times 1 \times H \times W$ dimension is constructed. However, the dimension of each guided layer is $1 \times C \times H \times W$. We need to broadcast the spatial attention map across C channels to make it match the dimension of guider layer.

In summary, the attention map is constructed in two steps. First, we calculate a distillation weight for each sample based on its classification loss using equation (1). Second, a spatial attention map with dimension $1 \times 1 \times H \times W$ for each guided layer is generated according to each sample's distillation weight as showed in equation (2). And after the broadcast operation, the final spatial attention map with dimension $1 \times C \times H \times W$ is constructed.

2.3. Network Optimization

During training, the parameters of teacher network are frozen. The gradient back propagation is only performed on the student network. In addition, it is also necessary to add ground truth supervision to the student network, since the final goal is to predict the location and category of the object. However, due to the limited capacity and learning ability of the small student network, it is difficult to learn discriminative features for accurate prediction only under the supervision of the ground truth. While the knowledge learned by the teacher network can be effectively transfered to the student network by the distillation loss. Hence, with the aid of teacher network, student network can learn the deep semantic patterns hidden in the data.

The overall training loss consists of two parts: the attention-guided distillation loss and detection loss. In particular, the former is the weighted Euclidean loss and the latter is the standard detection loss used in SSD (i.e. classification loss and bounding box regression loss). We optimize the following function:

$$L_{total} = L_{det} + \lambda_1 L_{dis} \tag{3}$$

The detection loss is defined as:

$$L_{det} = L_{cls} + \lambda_2 L_{reg} \tag{4}$$

where L_{cls} is the classification loss and L_{reg} is bounding box regression loss. λ_1, λ_2 are loss weight balance parameters(set to 1 by default).

The attention-guided distillation loss can be formulated as:

$$L_{dis} = \frac{1}{2N}\sum_{m=1}^{M}\frac{1}{C_m \times H_m \times W_m}||A_m \times (T_m - R(S_m))||^2 \tag{5}$$

Given that SSD detects objects at multiple CNN layers (corresponding to multiple feature maps), we also generate multiple attention maps and perform hierarchical distillation strategy respectively. In equation (5), m is the index of the distilled feature maps, and M is the total number of prediction layers. R is the adaptation layer, which is used to increase the dimension of the student feature map. A_m is the generated attention map for m_{th} distilled feature map. After obtaining the difference between the teacher feature map T_m and corresponding adapted student feature map $R(S_m)$, the Hadamard product of attention map A_m and the difference is calculated. Next, we calculate the Euclidean norm and normalize it with the product of the dimension of feature map.

3. EXPERIMENTS

3.1. Experiments on PASCAL VOC

PASCAL VOC is a common object detection dataset with 20 object categories. In our experiments, we use VOC2012 trainval and VOC2007 trainval(16551 images) for training, and test on VOC2007 test (4952 images). The input size is set to 300×300.

3.1.1. Ablation Studies

We choose the 1/8 VGG16, 1/4 VGG16 and 1/2 MobileNet as the students, and their corresponding compact networks as their teachers as showed in Table 1. For comparison, we train the students fine-tuned from ImageNet pre-trained model. Because of the limited capacity of such small network, it is not easy to get good performance in this conventional way. To reveal the effectiveness of our attention-guided distillation method, we make a baseline by directly distilling the whole feature map without attention. More specifically, knowledge distillation is performed by simply adding an Euclidean loss between corresponding feature maps of the student and teacher. The student model is trained on 4 GPUs for 240k iterations. During training, we use SGD with initial learning rate of 0.008, momentum of 0.9, weight decay of 0.0005 , and batch size of 32 on each device. The learning rate is divided by 10 at 180k and again at 220k. We adopt a linear warm up learning rate strategy for first 5k iterations starting from a learning rate of 0.0008. The hyper-parameters w_{max}, α, β are set to 15, 0.05, 2, respectively. Table 1 shows the results of the experiments above.

Table 1: The mAP(%) of different student networks on VOC2007 test dataset with different distillation strategies. ImageNet indicates that the student is fine-tuned from ImageNet pre-trained models. SimpleDis means directly distilling the whole feature map without attention. AttentionDis implies the proposed attention-guided knowledge distillation.

	Teacher&mAP(%)	Methods	Student mAP(%)
1/8 VGG16	VGG16(77.85)	ImageNet	46.23
		SimpleDis	53.96(+7.73)
		AttentionDis	**57.96(+11.73)**
1/4 VGG16	VGG16(77.85)	ImageNet	62.38
		SimpleDis	65.62(+3.24)
		AttentionDis	**69.48(+7.1)**
1/2 MobileNet	MobileNet(72.00)	ImageNet	58.22
		SimpleDis	61.32(+3.1)
		AttentionDis	**63.92(+5.7)**

Table 2: The influence of different hyper-parameters to the student performance.

w_{max}	α	β	Student mAP(%)
15	0.05	1	57.81
		2	57.96
		3	55.22
15	0.03	2	57.75
	0.05		57.96
	0.1		57.90
10	0.05	2	57.94
15			57.96
20			57.96

As shown in Table 1, for 1/8 VGG16, directly distilling the whole feature map can obtain 53.96% mAP, outperforming its counterpart fine-tuned from ImageNet pre-trained model by a large margin, nearly 8 percent. Our attention-guided distillation method further improves the accuracy by 4 percent compared to distillation without attention. The 1/8 VGG16 can finally reach 57.96% mAP. Similar results can be observed on the student 1/4 VGG16 and 1/2 MobileNet. From the results above, we can clearly know that the knowledge distillation is beneficial for training small object detectors, and can significantly improve the performance. With our attention-guided distillation method, the student net can adaptively learn important parts from the teacher. Thus the distillation process becomes oriented and efficient.

Table 3: Performance of the student when it is trained under different teacher's supervision (VOC dataset).

Student	FLOPs(B)	Teacher&mAP(%)	Student mAP(%)

1/2 MobileNet	0.30	MobileNet(72.00)
		VGG16(77.85)
			63.92
			64.40

1/4 ResNet18	0.34	ResNet18(72.12)
		VGG16(77.85)
			57.25
			57.69

Another advantage of this method lies in its robustness to hyper-parameter chosen, which will save a lot of time to tune these parameters. At the beginning of training, the network is unstable and thus extremely large loss values may occur with a very high probability. Based on equation (1), large loss value leads to a large distillation weight. Such large weight will incur network divergence. The upper bound w_{max} is introduced to reduce the potential divergence risk. Besides, α is designed to control the overall magnitude of the weight value. And the duty of hyper-parameter β is to regulate the gap between easily classified samples and those hard ones. We design control experiments to investigate the influence of these hyper-parameters to the student's performance. The teacher net is VGG16 with 77.85% mAP. And we choose 1/8 VGG16 as the student. As the results showed in Table 2, our method is not sensitive to the hyper-parameter changes of α and w_{max}. It works well in a fairly wide range of hyper-parameter variations, which provides us great conveniences in practical use. And we also observe that a large β causes obvious performance degradation since an oversize β suppresses much to those samples with small loss values, which will ignore too much useful information.

3.1.2. *Experiments of Different Teacher-student Pairs*

Our distillation method can efficiently transfer the knowledge when the teacher and student share the same backbone architecture, but not limited to this scenario. It also works well when the teacher and student have different backbone architectures. We can see from Table 3 that when the student net is trained under the supervision of a better teacher, its performance will get slightly increase. For student 1/2 MobileNet, when the teacher net is VGG16 with 77.85% mAP, the student net obtains 64.40% mAP. It is marginally higher (0.5%) then the mAP of the student 1/2 MobileNet supervised by the MobileNet (72.0%). By comparing the performance of 1/4 ResNet18 under different teachers' supervision in Table 3, we can draw the same conclusion. In Table 4, when the teacher is VGG16 (77.85% mAP), although 1/2 MobileNet has less FLOPs, it still greatly outperforms the other student 1/8 VGG16. This is because the special network structure of MobileNet endows itself the talent to obtain high performance with little computation cost. When we change another teacher, we can still observe the same phenomenon.

Table 4: Performance of different students when they are supervised by the same teacher (VOC dataset).

Teacher&mAP(%)	Student&FlOPs(B)	Student mAP(%)
VGG16(77.85)	1/8 VGG16(0.65) 1/2 MobileNet(0.30)	57.96 64.40
ResNet18(72.12)	1/4 ResNet18(0.34) 1/2 MobileNet(0.30)	57.25 64.34

Table 5: Comparison with state-of-the-art KD method. (VOC dataset). †: The performance of VGG16 on VOC dataset is not reported. We use their official open source code to conduct experiments.

Method	Teacher&mAP(%)	Student mAP(%)
1/16 ResNet-18 (Yi Wei *et al.*)[9]	ResNet-18(72.9)	47.0
AlexNet (Chen Guobin *et al.*)[7]	VGG16(70.4)	60.1
1/8 VGG16 (Wang Tao† *et al.*)[14]	VGG16(75.5)	50.8
1/8 VGG16 (Zhu Yousong *et al.*)[15]	VGG16(77.85)	56.88
1/8 VGG16-ours	VGG16(77.85)	**57.96**
1/4 VGG16 (Li Quanquan *et al.*)[8]	ResNet-50(78.78)	48.70
1/4 VGG16 (Wang Tao† *et al.*)[14]	VGG16(75.5)	60.3
1/4 VGG16-ours	VGG16(77.85)	**69.48**

3.1.3. *Comparison with State-of-the-art*

Up to now, there have been several works that apply knowledge distillation to two-stage object detection pipelines. We compare our method with other state-of-the-art knowledge distillation methods and list the results in Table 5. We can clearly know that our method yields a higher precision when the student is supervised by the same teacher. And even our 1/8 VGG16 detector outperforms the 1/4 VGG16 detector in [8] by nearly 11%. Under the condition of approximately equal backbone computation cost, our distillation method can surpass other distillation approaches, which strongly shows the superiority of our method.

3.2. Experiments on CityPersons

To validate the generality of our method, we conduct experiments on CityPersons dataset. CityPersons is a set of high quality annotations on top of the Cityscapes dataset. The train set contains 2975 images and the val set has 500 images. Our experiments are conducted on the "reasonable" setup (pedestrian scale [50, ∞], occlusion ratio[0, 0.35]). The log missrate(MR) which is averaged over the FPPI(false positives per image) range of [10^{-2}, 10^{0}] is used for evaluation.

We first train a SSD512 with a VGG16 backbone as the teacher. It is worth noticing that the original image in CityPersons has a size of 1024×2048. Directly resize the image to 512×512 when test will severely distort the objects. As a result, it will bring catastrophic damage to the final performance. To tackle this issue, we cut the original image into three 1024×1024 parts along the long side during test. These three parts start from the 0, 512, 1024 pixel of the original image, respectively. We let the detector make prediction on these 3 sub-images and merge the predictions as final results.

Table 6: The MR(%) of the student on the CityPersons dataset. The lower, the better.

Student	Teacher&MR(%)	Method	Student MR(%)
1/8 VGG16	VGG16(34.06)	SimpleDis	53.62
		AttentionDis	**51.15**
1/4 VGG16		SimpleDis	39.66
		AttentionDis	**35.82**

Similar to the experiments on PASCAL VOC dataset, we choose 1/8 VGG16 and 1/4 VGG16 as the students. We use SGD algorithm with initial learning rate of 0.006, momentum of 0.9, weight decay of 0.0005, and batch size of 8 on each device. We adopt a linear warm up learning rate for first 5k iterations starting from an initial learning rate of 0.0008. The hyper-parameters w_{max}, α, β are set to 15, 0.05, 2, respectively. When we choose the 1/4 VGG16 as the student, we train it on 4GPUs for 180k iterations. The learning rate is divided by 10 at 150k and again at 170k. And when the student is 1/8 VGG16, as such small network is hard to converge, we extend the training iterations to 350k. And the learning rate is divided by 10 at 320k and 340k.

Table 6 gives the final results. Noticing that in these two groups of experiments, our method displays consistent improvement over SimpleDis. 1/8 VGG16 distilled by our method shows 2.5% increase. And the 1/4 student trained by our method obtains a 3.8% improvement over SimpleDis. Our 1/4 VGG16 student detector achieves a remarkable 35.82% MR, even on par with the teacher. These results strongly prove the effectiveness of our method and demonstrate our attention-guided distillation approach can be well generalized to other datasets.

4. CONCLUSION

In this paper, we propose a novel approach for training efficient single-stage detectors from scratch. By integrating the attention mechanism to knowledge distillation, the student can distinguish which region is more important and thus pay more attention to these import regions, resulting in a high-efficiency learning process. Experiment results indicate that our attention-guided knowledge distillation can bring consistent accuracy improvement to the student detectors over various datasets and network backbones.

Acknowledgement: This work was supported by the Research and Development Projects in the Key Areas of Guangdong Province (No.2019B010153001) and National Natural Science Foundation of China under Grants No.61772527, No.61976210, No.61876086, No.62002357, No.61806200, No.51975044 and No.62002356.

5. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. 2012.

[2] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] 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 28*, pp. 91–99. 2015.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, 2016, pp. 21–37.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] 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*, pp. 742–751. 2017.

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

[9] Yi Wei, Xinyu Pan, Hongwei Qin, and Junjie Yan, "Quantization mimic: Towards very tiny CNN for object detection," in *European Conference on Computer Vision (ECCV)*, September 2018.

[10] Shitao Tang, Litong Feng, Wenqi Shao, Zhanghui Kuang, Wenjun Zhang, and Yimin Chen, "Learning efficient detector with semi-supervised adaptive distillation," in *British Machine Vision Conference (BMVC)*, September 2019.

[11] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. p.303–338, 2010.

[13] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

[15] Yousong Zhu, Chaoyang Zhao, Chenxia Han, Jinqiao Wang, and Hanqing Lu, "Mask guided knowledge distillation for single shot detector," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019.