Distilling Object Detectors via Decoupled Features

Jianyuan Guo[1,2], Kai Han[1], Yunhe Wang[1]*, Han Wu[2], Xinghao Chen[1], Chunjing Xu[1], Chang Xu[2]*
[1] Noah's Ark Lab, Huawei Technologies.
[2] School of Computer Science, Faculty of Engineering, University of Sydney.
{jianyuan.guo, kai.han, yunhe.wang}@huawei.com; c.xu@sydney.edu.au

Abstract

Knowledge distillation is a widely used paradigm for inheriting information from a complicated teacher network to a compact student network and maintaining the strong performance. Different from image classification, object detectors are much more sophisticated with multiple loss functions in which features that semantic information rely on are tangled. In this paper, we point out that the information of features derived from regions excluding objects are also essential for distilling the student detector, which is usually ignored in existing approaches. In addition, we elucidate that features from different regions should be assigned with different importance during distillation. To this end, we present a novel distillation algorithm via decoupled features (DeFeat) for learning a better student detector. Specifically, two levels of decoupled features will be processed for embedding useful information into the student, i.e.*, decoupled features from neck and decoupled proposals from classification head. Extensive experiments on various detectors with different backbones show that the proposed DeFeat is able to surpass the state-of-the-art distillation methods for object detection. For example, DeFeat improves ResNet50 based Faster R-CNN from 37.4% to 40.9% mAP, and improves ResNet50 based RetinaNet from 36.5% to 39.7% mAP on COCO benchmark. Code will be released*[1,2].

1. Introduction

As one of the fundamental computer vision tasks, object detection has attracted increasing attention in various real-world applications including autonomous driving and surveillance video analysis. Recent advances of deep learning introduce many convolutional neural network based solutions to object detection. The backbone of a detector is often composed of heavy convolution operations to produce intensive features that is critical to the detection accuracy. But doing so inevitably results in a sharp increase in the cost of computing resource and an apparent decrease in detection speed. Techniques such as quantization [19, 58, 31, 57, 62], pruning [2, 17, 20], network design [55, 49, 15, 18] and knowledge distillation [56, 6] have been developed to overcome this dilemma and achieve an efficient inference on detection task. We are particularly interested in knowledge distillation [24], as it provides an elegant way to learn a compact student network when a performance proven teacher network is available. Classical knowledge distillation methods are firstly developed for the classification task to decide *which* category the image belongs to. The information from soft label outputs [24, 28, 38, 13] or intermediate features [1, 23, 66] of a well-optimized teacher network have been well exploited to learn the student networks, but these methods cannot be directly extended to the detection task which needs to further figure out *where* the objects are.

There are a few attempts investigating knowledge distillation in the object detection task. For example, FGFI [56] asks the student network to imitate the teacher network on

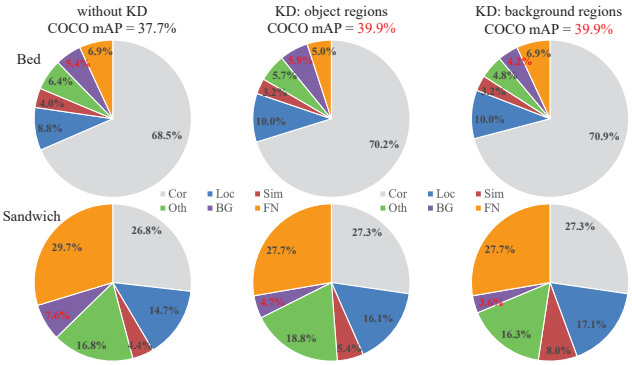


Figure 1. Error analyses of different distillation methods on COCO minival. KD via background regions alleviates the false positive rate and achieves comparable result with KD via object regions. **Cor**: correct class (IoU $>$ 0.5). **Loc**: correct class but misaligned box (0.1 $<$ IoU $<$ 0.5). **Sim**: wrong class but correct supercategory (IoU $>$ 0.1). **Oth**: wrong class (IoU $>$ 0.1). **BG**: background false positives (IoU $<$ 0.1). **FN**: false negatives (remaining errors).

*Corresponding author.
[1]https://github.com/huawei-noah/noah-research/tree/master/DeFeat
[2]https://www.mindspore.cn/resources/hub

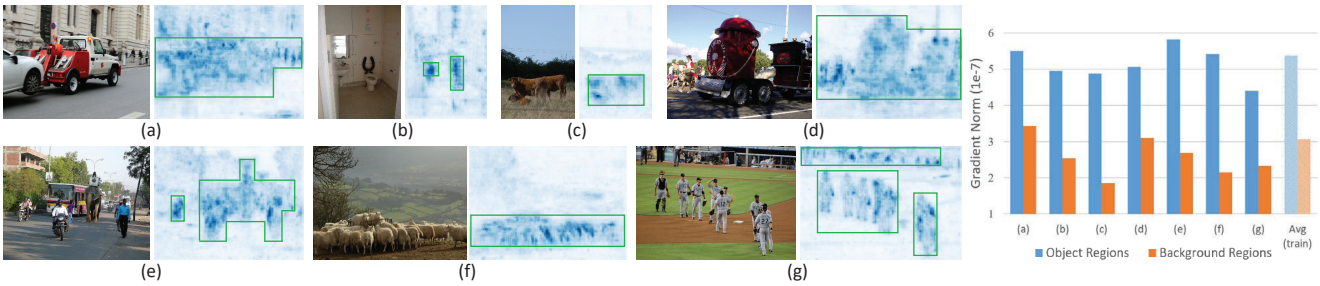


Figure 2. Left: L_2-Norm of the gradient in intermediate neck feature during back propagation, the darkest blue indicates the largest norm value. Images are randomly selected from COCO training set, and object regions are marked with the green box. Right: Average L_2-Norm of object and background regions. Avg (train) indicates the average norm of all images from COCO training set.

the near object anchor locations. TADF [47] distills the student via Gaussian masked object regions in neck features and positive samples in detection head. These works only distilled knowledge from object regions, as background regions were supposed to be not of interests in the detection task. Intuitively, during the distillation, background regions might introduce a large amount of noise and they have rarely been explored. But there lacks a thorough analysis of background regions when conducting the distillation. The hasty decision of throwing away background regions thus might not be wise. Most importantly, background information has already been proven to be helpful for visual recognition [53, 46, 9, 16]. Instead of guessing that background regions are useless or even harmful for distillation, it is time to have a fair and thorough analysis of the background and let the facts speak for themselves.

We first examine the roles of object and background regions in knowledge distillation by comparing two approaches: (i) distilling only via object-region FPN features and (ii) distilling only via background-region FPN features. It was taken for granted that the student would not be enhanced significantly when distilled via the background regions from teacher detector, since the background is less informative and noisy [56]. However, after extensive experiments on various models and datasets, we observe a surprising result that distilling student only via background-region features can also enhance the student remarkably and even, achieve comparable results with that of distillation via object regions (Figure 4). We further explore where the performance improvement comes from by distilling background features. Taking two classes from COCO as an example (see Figure 1), we conduct the error analysis [25] and find that distillation via background regions effectively reduces the number of background false positives.

The above evidence points to the conclusion that background regions can actually be a complementary to that distillation on object regions. Except that, prior literature has shown that there is a strong relationship between objects and background [53, 69]. The object likelihood [53] can be written as $P(O|V_o, V_b)$=$P(O|V_b)\frac{P(V_o|O,V_b)}{P(V_o|V_b)}$ (V_o and V_b are features of object region and background, respectively). All probabilities are related to background information which provides an estimate of the likelihood of finding an object (for example, one is unlikely to find a car in the room). The background-based priors vary from different images [69], thus we need to learn background feature for better prediction. However, the promising expectation above was failed to be justified by previous works [6, 30] that take both object and background regions into account. Although they leveraged both types of regions, the student was not significantly improved compared to those only using object regions, which seems to agree with the phenomenon indicated by [56]. Either the object or background regions can independently benefit the object detection through the distillation, but once they are integrated together, the performance drops unexpectedly. The reason could be that their methods integrate these two types of regions directly. From the gradient point of view, we illustrate the discordance between object and background regions in Figure 2. Images in the left column are randomly selected from COCO training set, and images in the right column are their corresponding gradients of neck features in student detector. We can observe that the magnitude of gradients from object regions are consistently larger than that from background regions. This therefore reminds us of different importance of object regions and background regions during the distillation.

Based on these insightful observations, we propose to decouple the features used for knowledge distillation and highlight their unique importance during the distillation. Two levels of features are included, *i.e.*, FPN features and RoI-aligned features. The FPN features are split into object and background parts using the ground-truth mask, and the mean square error loss is applied between teacher and student. The RoI-aligned features are also decoupled into positive and negative parts using teacher's predicted region proposals. The classification logits generated based on these decoupled RoI-aligned features are distilled using the KL divergence loss. The resulting DeFeat algorithm can be adaptively incorporated into both one-stage and two-stage detectors to improve the detection accuracy.

To validate our method, we conduct extensive experiments on Faster R-CNN [43] and RetinaNet [34] under various scenarios including distillation on shallow student and narrow student on two common detection benchmarks PASCAL VOC [12] and COCO [35]. In particular, our DeFeat improves ResNet50 based FPN from 37.4% to 40.9% mAP, and ResNet50 based RetinaNet from 36.5% to 39.7% mAP on COCO benchmark.

2. Related Work

Object detection is considered as one of the most challenging vision tasks which aims at finding out *what* and *where* the objects are when given an image. In the past few years, noticeable improvements in accuracy have been made in both one-stage [42, 36, 34, 29, 11, 67] and two-stage [43, 21, 33, 26, 27, 5] detectors. Although detectors fitted with very deep backbone [59, 48] have better detection accuracy, they are expensive in terms of computation cost and hard to deploy to mobile devices. There has been an interesting line of research that compresses large detection models by weight quantization [31, 57], representing the parameter weights with fewer bits. Pruning [37, 14, 60, 52, 51] is another line of research that removes unimportant connections from a large pre-trained model to compress detector. Designing a detector coupled with lightweight backbone network [55, 41, 45, 32, 61, 64] is also a trend for faster detection speed. Besides, there is also a line of research that transfers knowledge from a large detector to a smaller detector [6, 56, 30], in which one can boost the performance of a small detector without designing new architectures.

Knowledge distillation (KD) has become one of the most effective techniques to compress large models into smaller and faster ones. KD was first proposed by Buciluǎ *et al*. [4] and popularized by Hinton *et al*. [24] that transfers the dark knowledge from teacher network to student network through the soft outputs. FitNets [44] shows that activations [23] and features of intermediate layers [39] can also be treated as knowledge to guide the student network. Since then, KD has been widely adopted in classification tasks [22, 63, 3, 54, 7, 65, 10]. Recently, there are several works which propose to compress object detector using knowledge distillation. Chen *et al*. [6] distills the student through all components (*i.e*., neck feature, classification head and regression head), but the imitation of entire feature maps and distillation in classification head both ignore the imbalance in foreground and background which could lead to a suboptimal result. Tang *et al*. [50] proposes adaptive distillation loss for one-stage detector to magnify loss on hard samples. Li *et al*. [30] distills the features sampled from region proposals, however, only mimicking above regions could cause misguidance since the proposals can sometimes perform poorly. Wang *et al*. [56] intends to distill the student with fine-grained features from foreground object regions. However, we find that the remaining background features are also critical for distilling a better student detector.

In summary, current distillation frameworks for object detection ignore the important roles of the background regions in intermediate features and negative region proposals in classification head. In this work, we identify that the object and background regions in FPN features are both practical for distillation and treating positive and negative proposals equally would withhold the detector of stronger performance. Therefore we first generate a binary mask to decouple the intermediate features and then distill the features accordingly. Meanwhile, we decouple the positive and negative proposals in classification head to further improve the generalization.

3. Distillation via Decoupled Features

Generally, an object detector consists of three or four components: (a) backbone for extracting semantic features; (b) neck for fusing multi-level features; (c) RPN for generating proposals (only in two-stage detectors); and (d) head for object classification and bounding box regression. The purpose of distillation is to imbue the student with dark knowledge inside the teacher, which can be features of intermediate layer or soft predictions of region proposals in classification head. Define $\mathcal{S} \in \mathbb{R}^{H \times W \times C}$ and $\mathcal{T} \in \mathbb{R}^{H \times W \times C}$ as the intermediate features of student and teacher, respectively. The distillation via intermediate features can be formulated as:

$$\mathcal{L}_{fea} = \frac{\gamma}{2N} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} I(\phi(\mathcal{S}_{h,w,c}) - \mathcal{T}_{h,w,c})^2, \quad (1)$$

where $N = HWC$ is the total number of elements, γ is used to control the scale of distillation loss, ϕ denotes the adaptation layer [6] and I denotes the imitation mask, *i.e*., Gaussian mask in [47] and fine-grained mask in [56]. In previous works, only object regions are considered or the entire feature maps are distilled uniformly. Mask in methods that treat all regions uniformly [6, 30] can be seen as an all-one tensor.

Given K region proposals output from RPN, the classification head needs to compute soft labels of all proposals. The distillation via soft predictions can be formulated as:

$$\mathcal{L}_{cls} = \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}_{CE}(y_i^s, Y_i) + \frac{\lambda}{K} \sum_{i=1}^{K} \mathcal{L}_{KL}(y_i^s, y_i^t), \quad (2)$$

where the hyper-parameter λ is used to balance different loss items, $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$ denote the cross entropy loss and the KL divergence loss, respectively. Y_i is the ground truth label of the i-th proposal, and the predictions of student and

teacher detectors are y_i^s and y_i^t, respectively. The overall training targets of the student can be formulated as:

$$\mathcal{L} = \mathcal{L}_{fea} + \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{rpn}, \tag{3}$$

where $\mathcal{L}_{reg}$ is the bounding box regression loss in detector head and $\mathcal{L}_{rpn}$ denotes the RPN loss in two-stage detector.

3.1. Decouple Intermediate Features in Distillation

Previous works either choose partial regions or use all regions but treat each location on intermediate features equally. In particular, FGFI [56] presumed that background regions could introduce a large amount of noise and would impair the performance. However, this intuitive judgment is not consistent with what we have observed in experiments, as shown in Figure 1. Distillation via background only regions still achieves comparable results as distillation via object only regions. We come to a conclusion that the background regions in intermediate features can complement the object regions and further help the training of student detector, but the remaining question is how to appropriately integrate these two types of regions in distillation.

Based on the observations above, we propose to distill the student via decoupled features. Given the intermediate features of size $H \times W$, we first generate a binary mask M according to the ground truth box B:

$$M_{i,j} = \mathbb{1}[(i,j) \in \mathrm{B}], \tag{4}$$

where $M \in \{0,1\}^{H \times W}$, the value of location (i,j) is 1 if it belongs to an object, and 0 otherwise. Specifically, if detectors contain the feature pyramid network (FPN) which can output multi-level features, we will assign each ground truth box to its corresponding level and generate the mask M for each level accordingly. Then we use the generated binary mask to decouple the neck features, as shown in Figure 3. The intermediate feature distillation is formulated as:

$$\begin{aligned}\mathcal{L}_{fea} = &\frac{\alpha_{obj}}{2N_{obj}} \sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} M_{h,w}(\phi(\mathcal{S}_{h,w,c}) - \mathcal{T}_{h,w,c})^2 \\ + &\frac{\alpha_{bg}}{2N_{bg}} \sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} (1 - M_{h,w})(\phi(\mathcal{S}_{h,w,c}) - \mathcal{T}_{h,w,c})^2,\end{aligned} \tag{5}$$

where $N_{obj} = C\sum_{h=1}^{H}\sum_{w=1}^{W} M_{w,h}$ is the number of elements in object regions, $N_{bg} = C\sum_{h=1}^{H}\sum_{w=1}^{W}(1 - M_{w,h})$ is the number of elements in background regions. α_{obj} and α_{bg} are the loss coefficients for object and background regions, respectively. Through the ground-truth based mask, we decouple the features into object and background regions to distill both of them in a balanced manner.

3.2. Decouple Region Proposals in Distillation

Knowledge distillation via the soft predictions has been widely used in classification task, and can be useful for distilling the classification head in detection task. However, different from the classification task that there is no background category during training (*e.g.*, CIFAR and ImageNet), the object and background categories in detection head can have extremely different numbers of proposals. We conduct experiments to explore the separate distillation losses of object (positive) proposals and background (negative) proposals as shown in Figure 6. The distillation loss of positive proposals is consistently larger than that of negative proposals. If they are not properly balanced, the small gradients produced by background proposals can be drowned into the large gradients produced by positive ones, thus limiting further refinement. Besides, Table 5 shows that treating all proposals equally gets worse result compared to using negative only proposals. Hence, we propose to decouple the region proposals into positive ones and negative ones towards the optimal convergence when distilling the classification head. We feed the region proposals produced by teacher detector into both teacher's and student's head to generate the category predictions p^t and p^s as shown in Figure 3. The positive proposals and negative proposals are processed separately in our method. Given the logits z of positive proposals, we soften the predictions by a temperature T_{obj} for teacher and student as following:

$$p^{s,T_{obj}}(c \mid \theta^s) = \frac{exp(z_c^s/T_{obj})}{\sum_{j=1}^{C} exp(z_j^s/T_{obj})}, c \in Y \tag{6}$$

$$p^{t,T_{obj}}(c \mid \theta^t) = \frac{exp(z_c^t/T_{obj})}{\sum_{j=1}^{C} exp(z_j^t/T_{obj})}, c \in Y \tag{7}$$

where θ^s and θ^t denote the parameters of the student and the teacher, respectively. $Y = \{1, 2, ..., C\}$ are the classes of detection benchmark. For proposals belonging to background regions, we soften the predictions by a temperature T_{bg} for teacher and student similar to equations above. To distill the student with knowledge from teacher detectors, we use the Kullback Leibler (KL) divergence written as:

$$\begin{aligned}\mathcal{L}_{cls} = &\frac{\beta_{obj}}{K_{obj}} \sum_{i=1}^{K} b_i \mathcal{L}_{KL}(p_i^{s,T_{obj}}, p_i^{t,T_{obj}}) \\ + &\frac{\beta_{bg}}{K_{bg}} \sum_{i=1}^{K} (1 - b_i)\mathcal{L}_{KL}(p_i^{s,T_{bg}}, p_i^{t,T_{bg}})\end{aligned} \tag{8}$$

$$\mathcal{L}_{KL}(p^{s,T}, p^{t,T}) = T^2 \sum_{c=1}^{C} p^{t,T}(c \mid \theta^t) \log \frac{p^{t,T}(c \mid \theta^t)}{p^{s,T}(c \mid \theta^s)} \tag{9}$$

where $b_i \in \{0,1\}$ is the binary label of i-th proposal with respect to ground truth object. β_{obj} and β_{bg} are the coefficients of positive and negative samples, respectively. $K_{obj} = \sum_i b_i$ and $K_{bg} = \sum_i (1 - b_i)$ are the numbers of positive and negative proposals, respectively. And we multiply the distillation loss by T^2 to ensure the scale of gradient magnitudes.

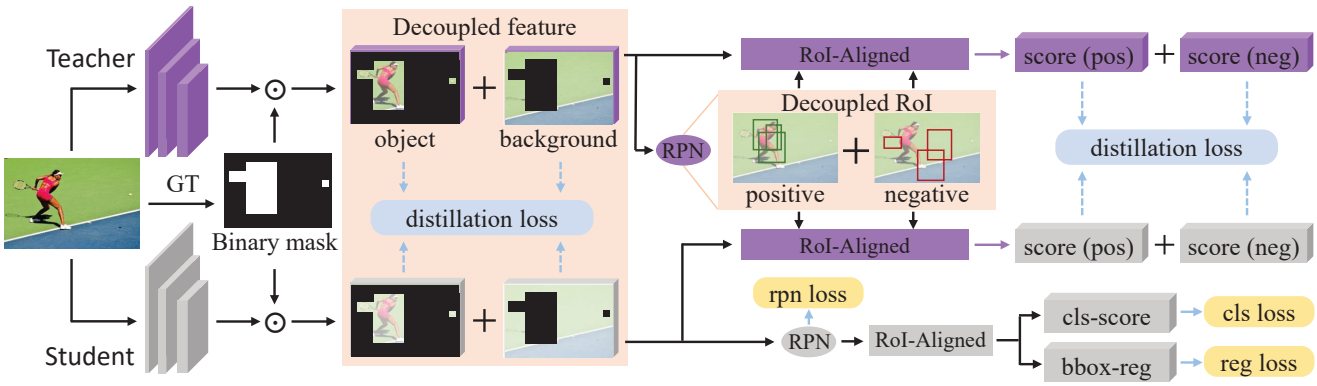


Figure 3. Overview of the proposed distillation via decoupled features (DeFeat) framework. We decouple the regions in intermediate FPN features and the region proposals from RPN to distill the student detector. The terms "score (pos)" and "score (neg)" indicate the classification scores on positive and negative proposals, respectively.

Table 1. Distillation results of both two-stage detector FPN and one-stage detector RetinaNet on COCO benchmark. "ResNet152-R50-FPN" indicates that the teacher detector is ResNet152 based FPN, while the student detector is ResNet50 based FPN.

Method	Distillation	mAP	AP_S	AP_M	AP_L
Teacher	ResNet152-FPN	41.3	24.4	45.3	54.0
Student	ResNet50-FPN	37.4	21.8	41.0	47.8
KD: all-neck	ResNet152-R50-FPN	40.1 (▲2.7)	22.8	44.0	53.3
KD: decoupled-neck	ResNet152-R50-FPN	40.4 (▲3.0)	23.4	44.4	53.1
KD: decoupled-neck + all-cls	ResNet152-R50-FPN	40.5 (▲3.1)	23.6	44.6	53.1
KD: decoupled-neck + decoupled-cls	ResNet152-R50-FPN	40.8 (▲3.4)	23.5	44.8	53.3
KD: backbone + decoupled-neck + decoupled-cls	ResNet152-R50-FPN	**40.9** (▲3.5)	**23.6**	**44.8**	**53.5**
Teacher	ResNet152-RetinaNet	40.5	24.1	44.7	53.4
Student	ResNet50-RetinaNet	36.5	20.9	40.2	47.0
KD: all-neck	ResNet152-R50-RetinaNet	39.1 (▲2.6)	22.1	43.1	52.3
KD: decoupled-neck	ResNet152-R50-RetinaNet	39.5 (▲3.0)	23.3	43.4	52.7
KD: backbone + decoupled-neck	ResNet152-R50-RetinaNet	**39.7** (▲3.2)	**23.4**	**43.6**	**52.9**

4. Experiments

4.1. Datasets and Metrics

COCO [35] is a challenging benchmark in object detection which contains 80 object classes. Our training set is the union of 80k training images and 35k subset of validation images (trainval35k), and the validation set is the remaining 5k validation images (minival). We consider Average Precision as evaluation metric, *i.e.*, mAP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L. The last three measure performance with respect to objects with different scales.

Pascal VOC [12] contains 20 object classes. Our training set is the union of VOC 2007 trainval (5K) and VOC 2012 trainval (11K), and the validation set is the VOC 2007 test (4.9K). We report the mAP scores using IoU at 0.5.

4.2. Implementation Details

All experiments are performed on 8 Tesla V100 GPUs. Our implementation is based on mmdetection [8] with Pytorch framework [40]. We use SGD optimizer with a batch size of 4 images per GPU, all models are trained for 12 epochs, known as 1× schedule. The input image is resized such that its shorter side has 800 pixels on COCO and 600 pixels on VOC. The initial learning rate is set as 0.04 and 0.02 for FPN [33] and RetinaNet [34], respectively. And the learning rate is divided by 10 at the 8-th and 11-th epochs. We set momentum as 0.9 and weight decay as 0.0001.

4.3. Main Results

We first verify the effectiveness of our proposed DeFeat on typical two-stage detection framework FPN [33] on COCO [35] benchmark, as shown in Table 1. ResNet50 based FPN is chosen as the student detector and ResNet152 based FPN as the teacher detector. "All-neck" indicates that the student is distilled via treating all regions in FPN features equally as illustrated in Equation 1. "Decoupled-neck" means the student is distilled via decoupled FPN features as illustrated in Equation 5. "All-cls" indicates that all region proposals are treated equally in distillation. "Decoupled-cls" denotes that region proposals are decou-

Table 2. Comparison with state-of-the-art methods on COCO.

Method	Distillation	mAP	AP_S	AP_M	AP_L
Teacher	R152-FPN	41.3	24.4	45.3	54.0
Student	R50-FPN	37.4	21.8	41.0	47.8
FGFI	R152-R50-FPN	39.9	22.9	43.6	52.8
TADF	R152-R50-FPN	40.1	23.0	43.6	53.0
DeFeat	R152-R50-FPN	**40.9**	**23.6**	**44.8**	**53.5**
Teacher	R50-FPN	37.4	21.8	41.0	47.8
Student	R50(1/4)-FPN	29.1	16.2	31.1	38.5
FGFI	R50-R50(1/4)-FPN	31.8	17.1	34.2	43.0
DeFeat	R50-R50(1/4)-FPN	**33.0**	**18.2**	**35.5**	**44.0**
Teacher	R152-RetinaNet	40.5	24.1	44.7	53.4
Student	R50-RetinaNet	36.5	20.9	40.2	47.0
FGFI	R152-R50-RetinaNet	38.9	21.9	42.5	52.2
DeFeat	R152-R50-RetinaNet	**39.7**	**23.4**	**43.6**	**52.9**

Table 3. Comparison with state-of-the-art methods on VOC.

Method	Distillation	mAP
Teacher	R152-FPN	82.69
Student	R50-FPN	80.53
FGFI [56]	R152-R50-FPN	81.57
TADF [47]	R152-R50-FPN	81.71
DeFeat	R152-R50-FPN	**82.28**
Teacher	R101-FPN	82.13
Student	R50-FPN	80.53
FGFI [56]	R101-R50-FPN	81.02
FGFI + PAD [68]	R101-R50-FPN	81.25
Mimic [30]	R101-R50-FPN	80.90
Mimic + PAD [68]	R101-R50-FPN	81.11
DeFeat	R101-R50-FPN	**81.47**

pled into positive (object) and negative (background) as illustrated in Equation 8. "Backbone" means the backbone features are also distilled. Directly distilling all regions in FPN features achieves 40.1% mAP, and the decoupled FPN features can further improve student detector by 0.3% mAP. Student distilled via decoupled FPN features and RPN proposals achieves a higher result, and our decoupled proposals can boost the result from 40.5% to 40.8% mAP. Further adopting the backbone features in distillation will achieve 40.9% mAP on COCO benchmark, bringing 3.5% gains compared to the student baseline model. In addition, we also conduct the experiments on typical one-stage detection framework RetinaNet [34], our ResNet152-R50-RetinaNet improves the baseline counterpart from 36.5% to 39.7% mAP on COCO. These results clearly elucidate the versatility and generality of our proposed DeFeat in both one-stage and two-stage detectors.

4.4. Comparison with State-of-the-art Methods

Comparison of the results obtained with other state-of-the-art distillation methods on COCO [35] benchmark and Pascal VOC [12] benchmark are shown in Table 2 and Table 3, respectively. Mimic [30] uses all regions in neck features to distill the student detector. FGFI [56] distills the

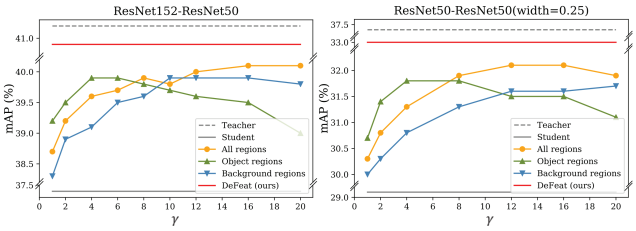


Figure 4. Comparisons of different distillation regions and various distillation loss coefficients on COCO. Left: ResNet152 based FPN teacher distills a shallower ResNet50-FPN student. Right: ResNet50-FPN teacher distills a narrower Quartered-ResNet50-FPN student (number of backbone channels is quartered).

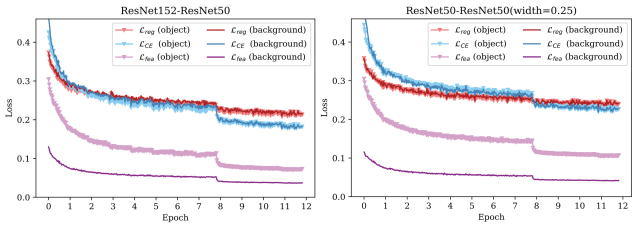


Figure 5. Training loss of distillation via neck features on COCO. Legend "object" denotes using object only regions and "background" denotes using background only regions.

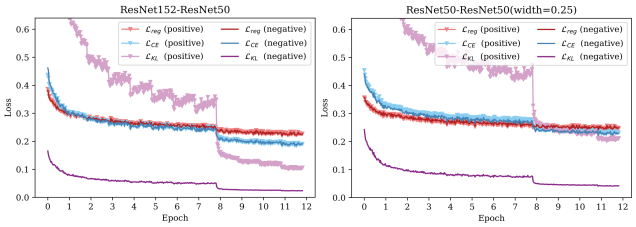


Figure 6. Training loss of distillation via region proposals on COCO. Legend "positive" denotes using positive only proposals and "negative" denotes using negative only proposals.

student detector via partial fine-grained regions in neck features. TADF [47] distills the student detector via Gaussian masked regions in neck features and positive only region proposals in both classification and regression head. PAD [68] proposes to estimate the weight of each region proposal during distillation. For fair comparison, we re-implement their methods, and most of them are slightly higher than the results in original paper. Our proposed DeFeat can be easily applied to the two most mainstream frameworks and consistently improve the performances of student detectors under various circumstances, *e.g.*, different backbones, shallower student and narrower student. FGFI achieves 39.9% mAP and TADF obtains 40.1% mAP on COCO benchmark. However, these two methods both ignore the important roles of background regions in neck features. Our distillation via decoupled features outperforms the FGFI by 1.0% mAP, and surpasses the TADF by 0.8% mAP, which indicates the effectiveness of the proposed method. Further-

more, our ResNet152-R50-FPN boosts the baseline model from 80.5% to 82.3% mAP, and ResNet101-R50-FPN improves the baseline model from 80.5% to 81.5% mAP on Pascal VOC benchmark, which outperforms other distillation methods apparently. To be specific, PAD [68] uses a stronger baseline implemented on detectron2 (ms-train, 17 epoch, and 1200×800 size). We report the results on mmdetection (ss-train, 12 epoch, and 1000×600 size) for fair comparisons with FGFI and TADF here.

4.5. Ablation Study

Impact of background regions in neck features. We investigate the object and background regions in neck features under two circumstances: (1) ResNet152 based FPN as teacher detector and ResNet50 based FPN as student detector; (2) ResNet50 based FPN as teacher detector and Quartered-ResNet50 based FPN (number of backbone channel is quartered, 64.28% Top-1 accuracy on ImageNet) as student detector. Figure 4 shows the corresponding results on COCO minival. We combine distillation loss based on FG-Mask [56] with original detection loss to guide the learning of student detector. The orange line indicates that the student detector is distilled via all regions in teacher's neck features. The green line indicates that the student is distilled via object only regions. The blue line denotes that the student is distilled via background only regions. γ is the coefficient in Equation 1. When the coefficient of distillation loss is small, learning from the object only (fine-grained) regions achieves better results. While the coefficient increases, performance improvement brought by background (non fine-grained) regions increases and then surpasses the result of mimicking object only regions. And the best results of mimicking object only regions and mimicking background only regions are comparable. Thus, we can come to the conclusion that both object regions and background regions in neck features are critical and have detrimental effects on the distillation of student detector. And we set α_{obj}=4 and α_{bg}=16 in Equation 5 accordingly.

In addition, we find that the distillation loss of background regions is much smaller than that of object regions during training, which indicates that background regions have smaller gradients compared to object regions. Figure 5 shows the training losses (*i.e.*, regression loss, classification loss and distillation loss) of mimicking object only and background only regions. Besides, the values of classification and regression losses are about four times larger than that of distillation via object regions, which should be the reason for why "Object regions" gets the best performance at $\gamma = 4$ in Figure 4. This can also explain that background regions need larger loss weight in training phase.

Evaluation of different region selection masks. Here we evaluate several region selection masks, namely FG-Mask from [56], Gaussian-Mask from [47], GT-Mask that directly leverages the ground truth boxes and Random-Mask which is generated randomly. Table 4 depicts the corresponding results. We can find that simply choosing scaled ground truth boxes as imitation regions achieves similar result compared to using fine-grained regions. And the distillation via object regions selected by Gaussian-Mask [47] obtains worse result. R152-R50-FPN achieves 40.1% mAP by treating all regions equally, decoupled regions can further boost the result by 0.3% mAP, while Random-Mask leads to a slight decrease of performance.

Table 4. Comparison of various region selection masks on COCO.

Model	Region	Mask	mAP
R152-R50-FPN	obj	FG-Mask	39.9
R152-R50-FPN	obj	Gaussian-Mask	39.8
R152-R50-FPN	obj	GT-Mask	39.9
R152-R50-FPN	obj + bg	FG-Mask	40.4
R152-R50-FPN	obj + bg	GT-Mask	40.4
R152-R50-FPN	obj + bg	Random-Mask	40.0
R152-R50-FPN	obj + bg	w/o Mask	40.1
R101-R50-FPN	obj	FG-Mask	38.9
R101-R50-FPN	obj	Gaussian-Mask	38.7
R101-R50-FPN	obj	GT-Mask	38.8
R50-R50(1/4)-FPN	obj	FG-Mask	31.8
R50-R50(1/4)-FPN	obj	Gaussian-Mask	31.5
R50-R50(1/4)-FPN	obj	GT-Mask	31.7

Table 5. Ablation study on the effects of positive and negative region proposals for R152-R50-FPN on COCO.

Teacher (R152-FPN)			41.3
Student (R50-FPN)			37.4
Proposal	β_{obj}	T_{obj}	mAP
positive	1	1	35.2
positive	0.1	1	37.7
positive	0.1	3	37.9
positive	0.05	3	38.1

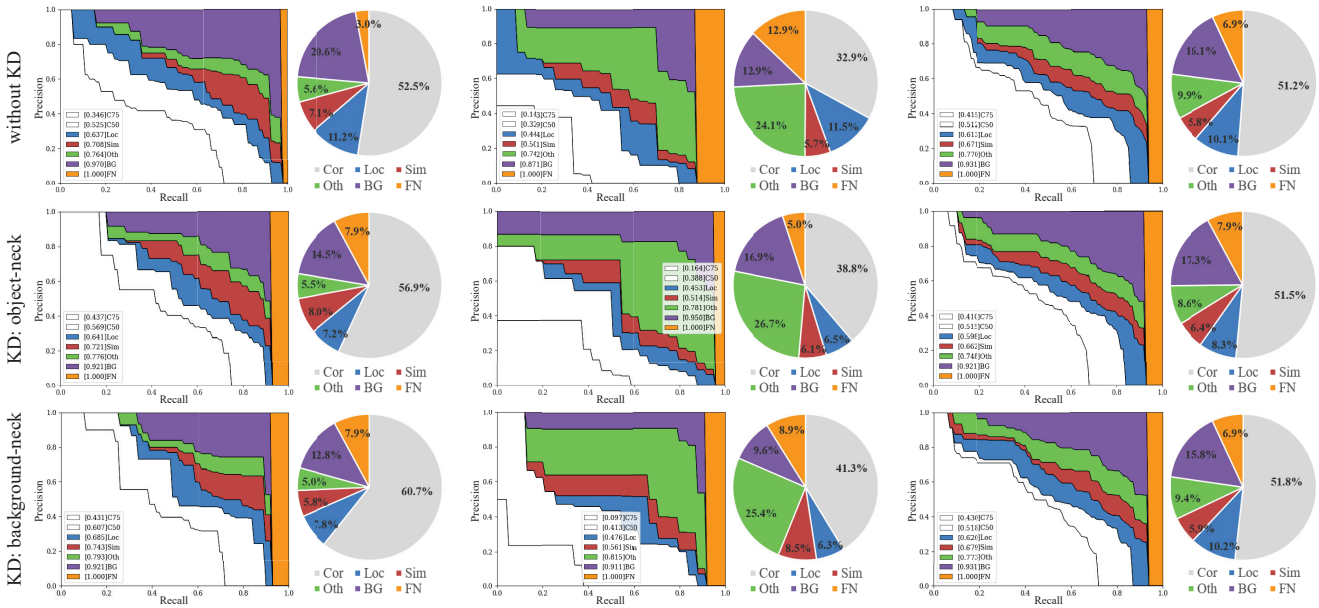
Proposal	β_{bg}	T_{bg}	mAP
negative	4	1	38.3
negative	2	1	38.6
negative	1	1	38.4
negative	1	2	38.2
sub-neg.	1	1	38.2
negative	0.1	1	37.4

Proposal	β_{obj}	β_{bg}	T_{obj}	T_{bg}	λ	mAP
positive + sub-neg.	-	-	-	-	1	37.4
positive + negative	-	-	-	-	1	38.2
positive + negative	-	-	-	-	0.1	38.1
positive + negative	0.05	2	1	1	-	38.6
positive + negative	0.05	2	3	1	-	38.9

Table 6. Ablation study on shared proposals on COCO.

Model	Proposal	mAP
R152-R50-FPN, decoupled-cls	Teacher	38.9
R152-R50-FPN, decoupled-cls	Student	38.7
R50-R50(1/4)-FPN, decoupled-cls	Teacher	31.5
R50-R50(1/4)-FPN, decoupled-cls	Student	31.2

Impact of positive and negative proposals in classification head. The evaluation of distillation on proposals are shown in Table 5, coefficients β_{obj}, β_{bg}, T_{obj}, T_{bg} and λ



(a) Apple-large (b) Backpack-large (c) Orange-medium

Figure 7. Precision-Recall curves and error analyses of different distillation methods on COCO benchmark. For each case, top figures correspond to raw student model, middle figures correspond to distillation via object only regions, bottom figures correspond to distillation via background only regions.

are illustrated in Equation 8 and Equation 2. We can find that using positive only proposals in distillation can slightly boost the performance of student detector, but demands a smaller coefficient. Using negative only proposals in distillation achieves better performance compared to using positive only proposals. One main reason is that the numbers of positive and negative proposals are imbalanced, and the difficulty in optimizing these two types of proposals can be different. Figure 6 also indicates that the distillation loss of negative proposals drops faster than that of positive proposals. The "sub-negative" in Table 5 denotes that we randomly select samples with the same number of positive proposals from negative proposals, which decreases the mAP by 0.2%, demonstrating that the distillation result is associated with the number of proposals. Our decoupled distillation improves the performance of previous method that treats all proposals equally from 38.2% to 38.9% mAP, which demonstrates the effectiveness of DeFeat.

Comparison of the shared proposals from teacher and student. Given a teacher detector and a student detector, the region proposals output by two models are inevitably different and consequently the student cannot be distilled directly. We analyze the performances of sharing teacher proposals with student and sharing student proposals with teacher. As can be seen in Table 6, feeding the proposals from teacher into distillation performs better than feeding the proposals from student detector. One main reason is that due to the large amount of possible region proposals, the proposals from teacher detector would be more accurate and contain more intensive information for distillation.

Performance gain from object and background regions. Figure 7 presents analyses on three randomly selected classes. Distillation via object regions and background regions both improve the number of correct detection significantly. Object regions can bring stronger localization ability (Loc) to the student, while background regions can effectively reduce the false positive rate (BG).

5. Conclusion

In this paper, we propose a simple yet efficient distillation method via decoupled features for object detection. We analyze and demonstrate the important roles of background regions during the distillation process. Based on ample observations, we introduce the DeFeat method in which the features are split into object and background parts at FPN level and RoI-aligned feature level, and distillation is applied on these two parts separately. DeFeat is general and can be easily used for both one-stage and two-stage detection frameworks. Extensive experiments validate the effectiveness of DeFeat by consistently outperforming other distillation techniques.

Acknowledgement

Chang Xu was supported in part by the Australian Research Council under Projects DE180101438 and DP210101859. And we sincerely thank all reviewers and ACs for their valuable comments.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019. 1

[2] Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems*, pages 2270–2278, 2016. 1

[3] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. *arXiv preprint:1804.03235*, 2018. 3

[4] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 3

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[6] 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*, pages 742–751, 2017. 1, 2, 3

[7] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Learning student networks via feature embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 3

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint:1906.07155*, 2019. 5

[9] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019. 2

[10] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *Advances in Neural Information Processing Systems*, 2020. 3

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Object detection with keypoint triplets. *arXiv preprint: 1904.08189*, 2019. 3

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, pages 303–338, 2010. 3, 5, 6

[13] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019. 1

[14] Sanjukta Ghosh, Shashi KK Srinivasa, Peter Amon, Andreas Hutter, and André Kaup. Deep network pruning for object detection. In *2019 IEEE International Conference on Image Processing*, pages 3915–3919, 2019. 3

[15] Jianyuan Guo, Kai Han, Yunhe Wang, Chao Zhang, Zhaohui Yang, Han Wu, Xinghao Chen, and Chang Xu. Hit-detector: Hierarchical trinity architecture search for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11405–11414, 2020. 1

[16] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[17] Yiwen Guo, Anbang Yao, and Yurong Chen. Dynamic network surgery for efficient dnns. In *Advances in neural information processing systems*, pages 1379–1387, 2016. 1

[18] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[19] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint:1510.00149*, 2015. 1

[20] 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*, pages 1135–1143, 2015. 1

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3

[22] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1921–1930, 2019. 3

[23] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3779–3787, 2019. 1, 3

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint:1503.02531*, 2015. 1, 3

[25] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *Proceedings of the European Conference on Computer Vision*, pages 340–353, 2012. 2

[26] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[27] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[28] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in neural information processing systems*, pages 2760–2769, 2018. 1

[29] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*, 2018. 3

[30] 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*, pages 6356–6364, 2017. 2, 3, 6

[31] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2810–2819, 2019. 1, 3

[32] Yuxi Li, Jiuwei Li, Weiyao Lin, and Jianguo Li. Tiny-dsod: Lightweight object detection for resource-restricted usages. *arXiv preprint:1807.11013*, 2018. 3

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3, 5

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3, 5, 6

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014. 3, 5, 6

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 2016. 3

[37] Marc Masana, Joost van de Weijer, and Andrew D Bagdanov. On-the-fly network pruning for object detection. *arXiv preprint arXiv:1605.03477*, 2016. 3

[38] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. *arXiv preprint:1902.03393*, 2019. 1

[39] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision*, pages 268–284, 2018. 3

[40] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[41] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection on mobile devices. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6718–6727, 2019. 3

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3

[44] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint:1412.6550*, 2014. 3

[45] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3

[46] Farah Sarwar, Anthony Griffin, Priyadharsini Periasamy, Kurt Portas, and Jim Law. Detecting and counting sheep with a convolutional neural network. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 2

[47] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint:2006.13108*, 2020. 2, 3, 6, 7

[48] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint:1602.07261*, 2016. 3

[49] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020. 1

[50] Shitao Tang, Litong Feng, Wenqi Shao, Zhanghui Kuang, Wei Zhang, and Yimin Chen. Learning efficient detector with semi-supervised adaptive distillation. *arXiv preprint:1901.00366*, 2019. 3

[51] Yehui Tang, Yunhe Wang, Yixing Xu, Yiping Deng, Chao Xu, Dacheng Tao, and Chang Xu. Manifold regularized dynamic network pruning. *arXiv preprint arXiv:2103.05861*, 2021. 3

[52] Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chunjing Xu, Chao Xu, and Chang Xu. Scop: Scientific control for reliable neural network pruning. *arXiv preprint arXiv:2010.10732*, 2020. 3

[53] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, pages 169–191, 2003. 2

[54] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 3

[55] Robert J Wang, Xiang Li, and Charles X Ling. Pelee: A real-time object detection system on mobile devices. In *Advances in Neural Information Processing Systems*, pages 1963–1972, 2018. 1, 3

[56] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 1, 2, 3, 4, 6, 7

[57] Yi Wei, Xinyu Pan, Hongwei Qin, Wanli Ouyang, and Junjie Yan. Quantization mimic: Towards very tiny cnn for object detection. In *Proceedings of the European Conference on Computer Vision*, pages 267–283, 2018. 1, 3

[58] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016. 1

[59] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

[60] Zihao Xie, Li Zhu, Lin Zhao, Bo Tao, Liman Liu, and Wenbing Tao. Localization-aware channel pruning for object detection. *Neurocomputing*, 2020. 3

[61] Zhaohui Yang, Yunhe Wang, Xinghao Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, and Chang Xu. Cars: Continuous evolution for efficient neural architecture search. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1829–1838, 2020. 3

[62] Zhaohui Yang, Yunhe Wang, Kai Han, Chunjing Xu, Chao Xu, Dacheng Tao, and Chang Xu. Searching for low-bit weights in quantized neural networks. *arXiv preprint arXiv:2009.08695*, 2020. 1

[63] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 3

[64] Shan You, Tao Huang, Mingmin Yang, Fei Wang, Chen Qian, and Changshui Zhang. Greedynas: Towards fast one-shot nas with greedy supernet. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2020. 3

[65] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. 3

[66] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint:1612.03928*, 2016. 1

[67] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[68] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. Prime-aware adaptive distillation. *arXiv preprint:2008.01458*, 2020. 6, 7

[69] Zhuotun Zhu, Lingxi Xie, and Alan L Yuille. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596*, 2016. 2