



HAL
open science

PDF-Distil: including Prediction Disagreements in Feature-based Distillation for object detection

Heng Zhang, Elisa Fromont, Sébastien Lefèvre, Bruno Avignon

► **To cite this version:**

Heng Zhang, Elisa Fromont, Sébastien Lefèvre, Bruno Avignon. PDF-Distil: including Prediction Disagreements in Feature-based Distillation for object detection. BMVC 2021 - 32nd British Machine Vision Conference, Nov 2021, Virtual, United Kingdom. pp.1-13. hal-03487128

HAL Id: hal-03487128

<https://hal.science/hal-03487128v1>

Submitted on 17 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PDF-Distil: including Prediction Disagreements in Feature-based Distillation for object detection

Heng Zhang¹
heng.zhang@irisa.fr

Elisa Fromont¹
elisa.fromont@irisa.fr

Sébastien Lefevre²
sebastien.lefevre@irisa.fr

Bruno Avignon³
bavignon@atermes.fr

¹ Univ Rennes, IUF, Inria,
CNRS, IRISA; F-35000 Rennes

² University Bretagne Sud, IRISA
Vannes, France

³ ATERMES company
Paris area, France

Abstract

Knowledge distillation aims at compressing deep models by transferring the learned knowledge from precise but cumbersome teacher models to compact student models. Due to the extreme imbalance between the foreground and the background of images, when traditional knowledge distillation methods are directly applied to the object detection task, there is a large performance gap between the teacher model and the student model. We tackle this imbalance problem from a sampling perspective, and we propose to include the teacher-student prediction disagreements into a feature-based detection distillation framework. This is done with PDF-Distil by dynamically generating a weighting mask applied to the knowledge distillation loss, based on the disagreements between the predictions of both models. Extensive experiments on PASCAL VOC and MS COCO datasets demonstrate that, compared to state-of-the-art methods, PDF-Distil is able to better reduce the performance gap between the teacher and student models.

1 Introduction

Despite their outstanding performance on different computer vision tasks, deep learning-based techniques still suffer from practical limitations that make them difficult to deploy at a large scale, especially when dealing with real-time embedded applications such as automatic surveillance and autonomous driving. This is due to the fact that state-of-the-art trained deep learning models often have a huge number of parameters that make them both slow at inference and heavy to store. Therefore, model compression techniques such as network pruning [1], parameter quantification [2] and knowledge distillation [3] are suggested to reduce the computational complexity and the storage cost of deep models, while minimizing the performance degradation.

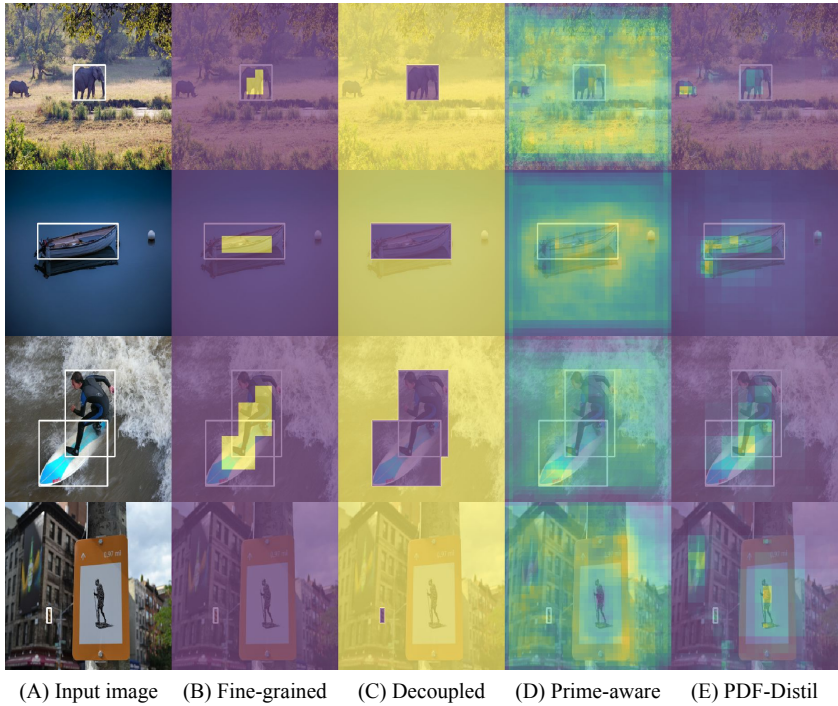


Figure 1: Visualization of sampling strategies from different feature-based detection distillation methods. We plot from left to right: (A) input image with ground truth boxes, (B) Fine-grained [29], (C) Decoupled [9], (D) Prime-aware [36] and (E) our PDF-Distil method.

In this paper, we investigate how knowledge distillation (KD) can be used in the context of object detection. KD utilizes a teacher-student framework to transfer the learned knowledge from a complex model (teacher) to a compact one (student). The concept of KD was firstly introduced in [9], where the KL-divergence between the predicted probability distributions of the teacher model and the student model is treated as a term of the loss function to optimize the student model. The motivation behind this logits-based distillation is that, for a given input image, we expect the image classification prediction of the student model to be as similar as possible to that of the teacher model, so that the student model is able to maintain high compactness and high precision simultaneously.

The internal representations (i.e., the feature maps) in deep neural networks carry rich semantic information, thereby providing better distillation guidance than the probability distributions. Leveraging this, recent studies have implemented feature-based distillation for image classification [23, 31]. However, when directly applying feature-based distillation to object detection, the precision gap between the teacher model and the student model remains significant. As shown in Figure 1 column (A), in the object detection task, the target objects normally only occupy a small part of the images. Therefore, the supervision of the feature distillation is often dominated by the abundant, less informative background. This foreground-background imbalance greatly reduces the efficiency of the knowledge transfer in feature-based distillation for object detection.

We tackle the aforementioned imbalance problem from a sampling perspective, i.e., by

adaptively assigning weights to each sampling location on the feature maps. Some previous works assigned weights according to the foreground-background distinction [8, 29] or the feature-mimicking uncertainty [66], while discarding the initial motivation of KD, which is minimizing the prediction difference between the teacher and the student models. We thus propose to combine feature-based with logits-based distillation, where the former is guided by the latter to more important areas on the feature maps. Specifically, the distillation weight on each feature map location is assigned according to the disagreement degree between the corresponding object detection predictions from the teacher and the student. In this way, the distillation is optimized to focus on areas where the student model makes inaccurate predictions, thereby minimizing the precision performance gap between the two models.

This paper is organized as follows: Section 2 reviews some representative works on object detection and knowledge distillation; Section 3 introduces the implementation details of our proposed method; Section 4 reports experimental results on different public datasets and compares our method with state-of-the-art methods; Section 5 concludes the paper.

2 Related work

Object detection. Object detection is one of the fundamental tasks in computer vision. Modern neural network-based object detection models consist of three sub-networks: the backbone, the neck and the head. Backbone networks are used to extract features from the input images. They are often taken from image classification networks, such as VGG [27], ResNet [8, 60], MobileNet [10, 11, 26] and ShuffleNet [20, 62]. Neck networks realize multi-scale object detection by fusing features at different scales. FPN [24] and PAFPN [19] are, nowadays, the most commonly adopted neck networks. Head networks handle instance classification and bounding-box regression. They can be roughly divided into two types: two-stage and single-stage detectors. Two-stage detectors [0, 21, 23] firstly generate various regions of interest, then refine and classify each region candidate separately; Single-stage detectors [13, 17, 22] directly localize and classify all existing objects on the image. Another criterion divides head networks into anchor-based and anchor-free detectors. Anchor-based detectors [15, 16, 25] resort to numerous predefined anchor boxes to handle objects' scale and shape variations; Anchor-free detectors directly predict objects' key-points [3, 12, 63] or centers-points [28, 37, 69], without the help of anchor boxes.

Knowledge distillation. KD is an effective means to compress deep models. A typical KD framework consists of three components: a teacher model, a student model and a knowledge transfer module. Although the teacher model allows high detection accuracy, it requires enormous parameters and calculations, which is impractical for real-time applications in embedded environments. In the setting of KD, a lighter student model is trained to inherit the knowledge learned by the teacher model. Logits-based and feature-based are two major KD strategies. Logits-based methods [9, 59] assign the output probability from the teacher model as the (soft) target for the training of the student model, which is a straightforward fashion to make the student model learn the class distributions from the teacher model. Alternatively, feature-based methods [29, 61] transfer high-level semantic information by making the student model mimic the intermediate features of the teacher model.

Knowledge distillation for object detection. Feature-based methods are the most commonly adopted KD strategy for object detection [9, 29, 67]. However, due to the extreme

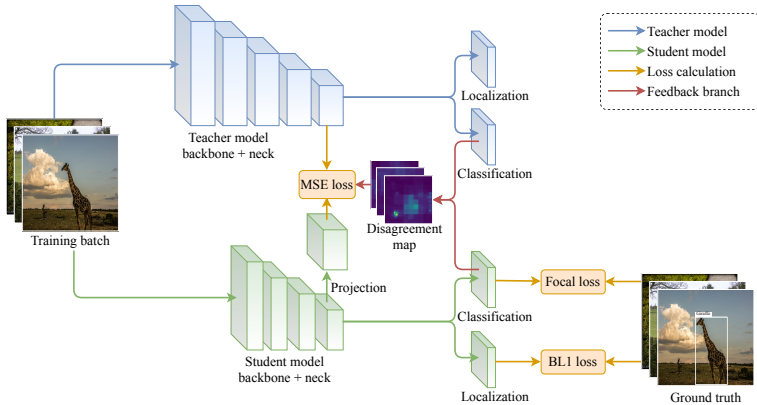


Figure 2: Overview of the proposed PDF-distil. We have added a prediction disagreement aware feedback branch (in red) in a traditional feature-based distillation framework.

foreground-background imbalance in object detection (see Section 1), a direct feature distillation is suboptimal. As shown in Figure 1, various solutions have been proposed to tackle this imbalance problem: Fine-grained [24] (column B) suggested to only perform feature imitation on near object regions; Decoupled [9] (column C) noticed that distillation on background regions reduces false positive detections, and thereby proposed to assign different weights for foreground features and for background features; Prime-aware [27] (column D) realized an adaptive sample weighting by incorporating uncertainty learning into the feature distillation. In their implementation, sample weighting is biased towards “easy” samples, most of which are actually background. Different from the above methods, our proposed weighting mechanism relies on the disagreements between the teacher and student predictions. Our intuition is that regions where the two models make different object detection predictions are actually regions where the student model struggles the most. The column E of Figure 1 shows that our weighting mechanism is biased towards “hard” regions, such as unknown objects (first line), reflection in water (second line), object junctions (third line) and ambiguous objects (fourth line). Our experimental results demonstrate that enhancing distillation on these regions greatly reduces the performance gap between the teacher model and the student model.

3 Proposed approach

3.1 Overview

We illustrate the two involved models for feature-based detection distillation in Figure 2. The student model (presented by the green blocks in the figure) employs a simpler network architecture than the teacher model (blue blocks), namely thinner or shallower backbone and neck networks in the context of object detection. Note that in Figure 2 the multi-scale detection architecture [24, 27] is not presented for the sake of clarity. The yellow blocks in Figure 2 show that the training of the student model is supervised by the normal object detection loss (including the instance classification and the bounding-box localization losses) as well as the knowledge transfer loss, which is defined as the Mean Square Error (MSE)

between the intermediate feature maps of the teacher model and the projected feature maps of the student model. The projection is performed through a 1×1 convolution to map the student hidden layer to the teacher hidden layer.

The main contribution of the proposed approach consists in adding a prediction disagreement aware feedback branch (the **red branch** in Figure 2), in a traditional feature-based detection distillation framework. This feedback branch leverages the prediction difference between the teacher model and the student model to generate a disagreement map, which is used as a weighting mask for the knowledge transfer loss.

3.2 Disagreement mapping

In order to obtain the aforementioned disagreement map, we compute the distance between the respective classification branches of the teacher model and the student model¹. Formally, let P^t and P^s respectively represent the output probability distributions from the classification branches of the teacher and the student, and let N denotes the number of object categories. Assume that there are M classification predictions associated to a specific feature map location. To be more specific, for anchor-based methods, M equals to the number of anchors per location, e.g., $M = 6$ for SSD [17] and $M = 9$ for RetinaNet [18]; for anchor-free methods like FCOS [23], M equals to 1 since each feature map location only produces one bounding-box prediction. The prediction disagreement at each feature location ($D_{h,w}$) is defined as:

$$D_{h,w} = \sum_M \sum_N \mathcal{F}(P_{h,w}^t, P_{h,w}^s) \quad (1)$$

where \mathcal{F} is a given dissimilarity function (in Section 4, we compare KL-divergence, L1 and L2 distances). Let H , W and C denote the height, width and depth of the feature maps, the actual weighting value at each location on the disagreement map ($W_{h,w}$) is assigned as:

$$W_{h,w} = \frac{H \times W \times D_{h,w}}{\sum_H \sum_W D_{h,w}}. \quad (2)$$

Let X^t denote the intermediate feature maps of the teacher model and X^s the projected feature maps of the student model, the weighted knowledge transfer loss L_{kd} is computed as:

$$L_{kd} = \frac{\sum_H \sum_W (W_{h,w} \times \sum_C (X^t - X^s)^2)}{H \times W \times C}. \quad (3)$$

For a better understanding of the proposed weighting strategy, we provide more visualization results in Figure 3. Specifically, the column E corresponds to the presented disagreement map. As is shown, the key difference between the proposed weighting method with previous methods [8, 24, 26], is that ours is capable to adaptively locate challenging areas for the student model to perform object detection, e.g., ambiguous objects (first line), shadow of objects (second line), defocused objects (third line) and human photos (fourth line).

¹Since localization predictions on background areas are meaningless, we do not consider the prediction difference between the teacher model and the student model in the localization branches.



Figure 3: More visualization of sampling strategies from different feature-based detection distillation methods.

4 Experiments

4.1 Experimental setting

Datasets and evaluation metric. Extensive experiments are conducted on PASCAL VOC [1] and MS COCO [2] datasets, containing 20 and 80 object categories respectively. For PASCAL VOC dataset, models are trained on the union of the 2007 trainval set and the 2012 trainval set, and evaluated on the 2007 test set; For MS COCO dataset, we use the 2017 train set for training and the 2017 val set for evaluation. Following the common practice, we adopt the (COCO-style) mean Average Precision (denoted as mAP) as the evaluation metric, which is defined as the average of AP scores across 10 Intersection-over-Union (IoU) thresholds from 0.5 to 0.95. We report using (+...) the absolute mAP improvement from KD for each distilled model. Moreover, the AP scores with the IoU threshold 0.5 and 0.75 (denoted as AP_{50} and AP_{75}) are also listed for comparisons.

Network architectures. We evaluate our proposed method by implementing object detectors using different combinations of backbone, neck and head networks. To be more specific, in terms of the backbone network, a deeper or wider version of ResNet [3] or ShuffleNetV2 [4] is adopted for teacher models, and their shallower or thinner version is used for student models; For the neck network, teacher models are equipped with the more complex PAFPN [5], while student models employ the simpler FPN [6]; As for the head network, we use RetinaNet [7] as a representative for anchor-based methods and FCOS [8] as a

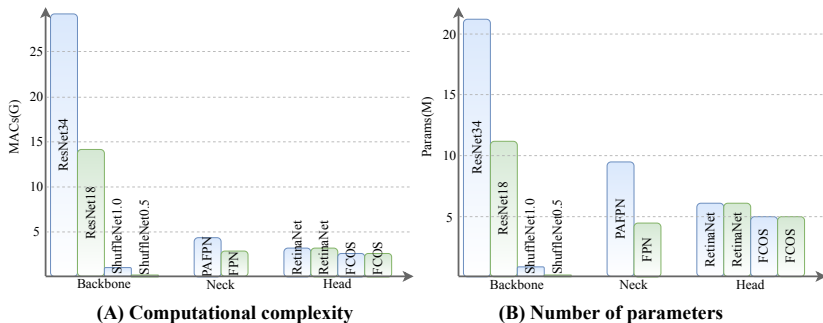


Figure 4: Comparisons on computational complexity and number of parameters for each network component from teacher models (blue bars) and student models (green bars).

representative for anchor-free methods. Both detection heads are optimized by the Mutual Guidance label assignment strategy [8]. To compare the computational cost of both teacher and student models, we summarize in Figure 4 the amount of Multiply–Accumulate operations (denoted as *MACs*) as well as the amount of learnable parameters (denoted as *Params*) for each network component. It can be observed that student models require much less computing resources than teacher models. Since we implement the same head network for each teacher and the corresponding student, the computational complexity and the number of parameters remain unchanged for this component.

Implementation details. For each detector, the backbone network is pre-trained on the ImageNet-1k dataset [9], while the neck and the head networks are randomly initialized. We adopt single-scale training and evaluation, where the input image resolution is fixed to 320×320 for all experiments. Several data augmentation strategies are applied, such as random image flipping, shifting, cropping, padding, noising and mixup [63]. Note that the mentioned data augmentation strategies are applied to all the compared methods, including our competitors. We use the Stochastic Gradient Descent (SGD) optimizer with 32 images per mini-batch and with an initial learning rate of $1e-2$. The warm-up strategy [9] is applied to stabilize the training at the beginning, followed by the cosine annealing strategy [19] for learning rate decay. Models are trained for 70 and 140 epochs for PASCAL VOC and MS COCO, respectively. We use Balanced L1 loss [20] and Generalized IoU loss [24] to optimize the localization branch of RetinaNet [15] and FCOS [28], respectively. Focal loss [25] is adopted for the training of the classification branch for both head networks.

4.2 Ablation study

Ablation experiments are conducted on PASCAL VOC to explore the relationship between the teacher-student prediction disagreements and the knowledge transfer effects. In Table 1, we consider eight different feature sampling strategies for detection distillation: 1) the baseline setting where all samples are treated equally (equivalent to Fitnet [27]); 2-5) hard sampling strategies where the distillation is only conducted on 25% or 50% of feature areas with the most similar or the most different teacher-student predictions; 6-8) the proposed adaptive sampling approach with respectively KL-divergence, L1 distance or L2 distance

<i>Models</i>		<i>mAP</i>	<i>AP50</i>	<i>AP75</i>
Teacher	ResNet34-PAFPN-RetinaNet	60.0	82.5	65.3
Student	ResNet18-FPN-RetinaNet	56.5	80.1	61.5
	1) All samples equally	57.8 (+1.3)	81.4	62.9
	2) 25% most similar predictions	57.5 (+1.0)	81.0	62.6
	3) 50% most similar predictions	57.8 (+1.3)	81.3	62.9
	4) 50% most different predictions	59.2 (+2.7)	82.3	64.6
	5) 25% most different predictions	59.3 (+2.8)	82.3	64.8
	6) PDF-Distil (KL-divergence)	59.4 (+2.9)	82.5	64.7
	7) PDF-Distil (L1 distance)	59.5 (+3.0)	82.7	65.3
	8) PDF-Distil (L2 distance)	59.8 (+3.3)	83.0	65.4
Teacher	ShuffleNet1.0-PAFPN-RetinaNet	51.6	75.9	55.0
Student	ShuffleNet0.5-FPN-RetinaNet	41.3	65.5	43.1
	1) All samples equally	43.0 (+1.7)	66.8	44.6
	2) 25% most similar predictions	42.4 (+1.1)	66.2	44.4
	3) 50% most similar predictions	42.6 (+1.3)	66.3	44.5
	4) 50% most different predictions	44.1 (+2.8)	68.2	46.2
	5) 25% most different predictions	44.7 (+3.4)	68.9	46.3
	6) PDF-Distil (KL-divergence)	45.0 (+3.7)	69.5	47.2
	7) PDF-Distil (L1 distance)	45.1 (+3.8)	69.5	47.6
	8) PDF-Distil (L2 distance)	45.4 (+4.1)	69.6	47.7
Teacher	ResNet34-PAFPN-FCOS	58.6	83.1	63.9
Student	ResNet18-FPN-FCOS	54.9	80.7	59.1
	1) All samples equally	56.7 (+1.8)	81.8	61.3
	2) 25% most similar predictions	56.4 (+1.5)	81.4	60.6
	3) 50% most similar predictions	56.6 (+1.7)	81.7	60.8
	4) 50% most different predictions	57.8 (+2.9)	82.4	62.4
	5) 25% most different predictions	58.0 (+3.1)	82.6	62.9
	6) PDF-Distil (KL-divergence)	58.1 (+3.2)	82.6	63.2
	7) PDF-Distil (L1 distance)	58.2 (+3.3)	82.6	63.2
	8) PDF-Distil (L2 distance)	58.3 (+3.4)	82.9	63.3

Table 1: Ablation studies on PASCAL VOC. We compare eight different feature sampling strategies for detection distillation, and the proposed PDF-Distil with L2 distance as the dissimilarity function achieves the best result.

as the dissimilarity function in Equation 1. The results are summarized in Table 1. When comparing the distillation results of the four hard sampling strategies (i.e., 2-5), we can conclude that feature samples with different teacher-student predictions are much more effective than those with similar predictions. This finding validates our initial hypothesis that the disagreements between the teacher-student object detection predictions can be regarded as an indicator of the importance for feature-based distillation. Moreover, regardless of the specific dissimilarity function, the adaptive sampling strategies (6-8) outperform the hard sampling strategies (2-5), indicating the effectiveness of the proposed dynamic weighting mechanism. As for the selection of the dissimilarity function, L2 distance (i.e., 8) demonstrates a constant advantage for all backbone-neck-head combinations. Therefore, we choose L2 distance as the dissimilarity function for the following experiments.

4.3 Comparison with state-of-the-art

As shown in Tables 2 and 3, we further compare our method with SOTA detection distillation methods on PASCAL VOC and MS COCO datasets. The results show that for either backbone-neck-head combinations and on both datasets, our method outperforms all existing KD methods. In particular, our method brings more than 3% (respectively 2%) of absolute precision improvements in comparison to student models without KD on PASCAL VOC (resp. MS COCO), and about 1% of absolute improvements to all previous detection distillation methods. Moreover, we report on the test set of each dataset the absolute difference between the detection predictions of the teacher model and the student model (denoted as D_{pred}), and we notice that our method effectively reduces the teacher-student prediction difference. Figure 5 illustrates the detection results on a few exemplar images treated by the teacher model, Fitnet [24], Fine-grained [24], Decoupled [8], Prime-aware [36] and our method. As is shown, our method gives detection results more similar to the teacher model than the other SOTA methods that miss some objects (dog, bicycle, potted plant, chair in the four examples, respectively).

<i>Models</i>		<i>mAP</i>	<i>AP50</i>	<i>AP75</i>	<i>D_{pred}</i>
Teacher	ResNet34-PAFPN-RetinaNet	60.0	82.5	65.3	-
Student	ResNet18-FPN-RetinaNet	56.5	80.1	61.5	2.96E-4
	w/ Fitnet [24]	57.8 (+1.3)	81.4	62.9	2.54E-4
	w/ Fine-grained [24]	58.6 (+2.1)	81.6	64.4	2.66E-4
	w/ Decoupled [8]	58.4 (+1.9)	81.8	63.5	2.43E-4
	w/ Prime-aware [36]	58.6 (+2.1)	81.9	63.7	2.44E-4
	w/ PDF-Distil (L2 distance)	59.8 (+3.3)	83.0	65.4	2.20E-4
Teacher	ShuffleNet1.0-PAFPN-RetinaNet	51.6	75.9	55.0	-
Student	ShuffleNet0.5-FPN-RetinaNet	41.3	65.5	43.1	4.34E-4
	w/ Fitnet [24]	43.0 (+1.7)	66.8	44.6	3.97E-4
	w/ Fine-grained [24]	44.8 (+3.5)	69.1	47.3	4.01E-4
	w/ Decoupled [8]	43.2 (+1.9)	67.1	45	3.79E-4
	w/ Prime-aware [36]	43.1 (+1.8)	66.9	44.7	3.79E-4
	w/ PDF-Distil (L2 distance)	45.4 (+4.1)	69.6	47.7	3.62E-4
Teacher	ResNet34-PAFPN-FCOS	58.6	83.1	63.9	-
Student	ResNet18-FPN-FCOS	54.9	80.7	59.1	1.56E-3
	w/ Fitnet [24]	56.7 (+1.8)	81.8	61.3	1.35E-3
	w/ Fine-grained [24]	57.0 (+2.1)	81.5	61.4	1.56E-3
	w/ Decoupled [8]	57.1 (+2.2)	82.2	62.0	1.33E-3
	w/ Prime-aware [36]	57.3 (+2.4)	82.2	61.9	1.30E-3
	w/ PDF-Distil (L2 distance)	58.3 (+3.4)	82.9	63.3	1.24E-3
Teacher	ShuffleNet1.0-PAFPN-FCOS	50.0	76.2	52.4	-
Student	ShuffleNet0.5-FPN-FCOS	39.4	66.0	39.3	2.44E-3
	w/ Fitnet [24]	41.4 (+2.0)	67.9	41.4	2.04E-3
	w/ Fine-grained [24]	42.4 (+3.0)	68.9	43.0	2.32E-3
	w/ Decoupled [8]	41.4 (+2.0)	67.5	41.4	2.04E-3
	w/ Prime-aware [36]	42.0 (+2.6)	68.1	43.0	2.05E-3
	w/ PDF-Distil (L2 distance)	43.2 (+3.8)	69.6	44.4	1.96E-3

Table 2: Comparisons with SOTA detection distillation methods on PASCAL VOC.

<i>Models</i>		<i>mAP</i>	<i>AP50</i>	<i>AP75</i>	<i>D_{pred}</i>
Teacher	ResNet34-PAFPN-RetinaNet	38.7	56.2	41.4	-
Student	ResNet18-FPN-RetinaNet	35.0	52.2	37.2	1.75E-4
	w/ Fitnet [24]	35.6 (+0.6)	52.7	37.8	1.62E-4
	w/ Fine-grained [24]	36.0 (+1.0)	53.0	38.3	1.58E-4
	w/ Decoupled [8]	35.9 (+0.9)	53.2	37.7	1.53E-4
	w/ Prime-aware [67]	35.6 (+0.6)	52.8	37.7	1.57E-4
	w/ PDF-Distil (L2 distance)	36.9 (+1.9)	54.2	39.1	1.46E-4
Teacher	ShuffleNet1.0-PAFPN-RetinaNet	28.9	44.4	30.4	-
Student	ShuffleNet0.5-FPN-RetinaNet	21.3	35.2	22.1	2.65E-4
	w/ Fitnet [24]	22.0 (+0.7)	35.7	23.1	2.54E-4
	w/ Fine-grained [24]	22.7 (+1.4)	36.0	24.0	2.59E-4
	w/ Decoupled [8]	22.4 (+1.1)	36.2	23.5	2.50E-4
	w/ Prime-aware [67]	22.4 (+1.1)	36.3	23.3	2.47E-4
	w/ PDF-Distil (L2 distance)	23.6 (+2.3)	37.5	24.7	2.36E-4

Table 3: Comparisons with SOTA detection distillation methods on MS COCO.

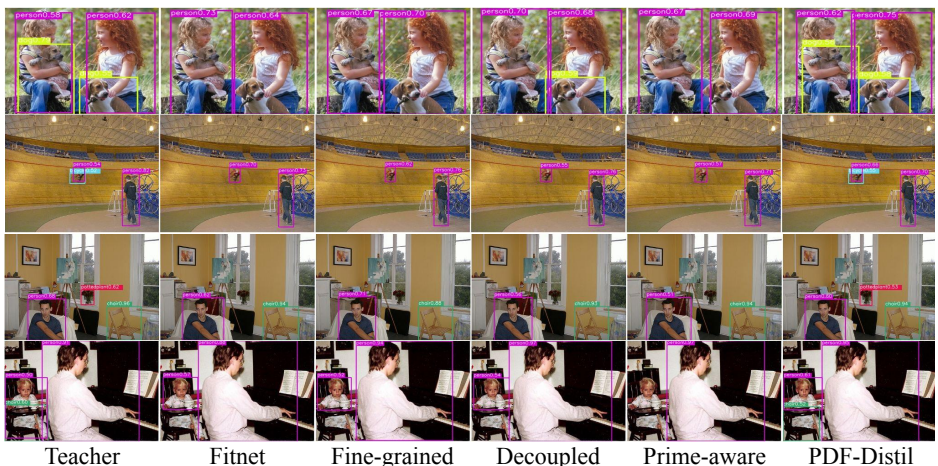


Figure 5: Visualization of some detection results from teacher model and student models distilled by Fitnet, Fine-grained, Decoupled, Prime-aware and our PDF-Distil method.

5 Conclusion

We address the foreground-background imbalance problem which happens when distilling knowledge from a teacher model to a student model in the context of object detection. To do so, we leverage the teacher-student prediction disagreements (i.e. logits-level information) to guide the knowledge distillation in a feature-based distillation framework. Our experiments demonstrate that the proposed method helps to reduce the performance gap between the teacher and the student models compared to all related state-of-the-art methods. Future studies could investigate how to include predictions from localization branches into the disagreement mapping to further improve the distillation.

References

- [1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018.
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.
- [3] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019.
- [4] 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*, 88(2):303–338, 2010.
- [5] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [6] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. *arXiv preprint arXiv:2103.14475*, 2021.
- [7] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [9] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, pages 1314–1324, 2019.
- [11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [12] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018.
- [13] 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 *ECCV*, pages 740–755, 2014.
- [14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

- [15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.
- [16] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018.
- [17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.
- [18] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *ICCV*, pages 2736–2744, 2017.
- [19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [20] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, pages 116–131, 2018.
- [21] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: Towards balanced learning for object detection. In *CVPR*, pages 821–830, 2019.
- [22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.
- [23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.
- [24] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019.
- [25] Adriana Romero, Samira Ebrahimi Kahou, Polytechnique Montréal, Y. Bengio, Université De Montréal, Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.
- [26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.
- [27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [28] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019.
- [29] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, pages 4933–4942, 2019.

- [30] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017.
- [31] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.
- [32] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Localize to classify and classify to localize: Mutual guidance in object detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [33] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.
- [34] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018.
- [35] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018.
- [36] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. Prime-aware adaptive distillation. In *ECCV*, pages 658–674, 2020.
- [37] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.
- [38] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, pages 850–859, 2019.
- [39] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *CVPR*, pages 840–849, 2019.