Bridging Cross-task Protocol Inconsistency for Distillation in Dense Object Detection

Longrong Yang[1*], Xianpan Zhou[2*], Xuewei Li[1], Liang Qiao[1,3]
Zheyang Li[1,3], Ziwei Yang[3], Gaoang Wang[4], Xi Li[1,5,6†]

[1]College of Computer Science & Technology, Zhejiang University
[2]Polytechnic Institute, Zhejiang University
[3]Hikvision Research Institute   [4]ZJU – UIUC Institute, Zhejiang University
[5]Shanghai Institute for Advanced Study of Zhejiang University
[6]Zhejiang – Singapore Innovation and AI Joint Research Lab, Hangzhou

{longrongyang, zhouxianpan, xueweili, xilizju}@zju.edu.cn
{qiaoliang6, lizheyang, yangziwei5}@hikvision.com, gaoangwang@intl.zju.edu.cn

Abstract

Knowledge distillation (KD) has shown potential for learning compact models in dense object detection. However, the commonly used softmax-based distillation ignores the absolute classification scores for individual categories. Thus, the optimum of the distillation loss does not necessarily lead to the optimal student classification scores for dense object detectors. This cross-task protocol inconsistency is critical, especially for dense object detectors, since the foreground categories are extremely imbalanced. To address the issue of protocol differences between distillation and classification, we propose a novel distillation method with cross-task consistent protocols, tailored for the dense object detection. For classification distillation, we address the cross-task protocol inconsistency problem by formulating the classification logit maps in both teacher and student models as multiple binary-classification maps and applying a binary-classification distillation loss to each map. For localization distillation, we design an IoU-based Localization Distillation Loss that is free from specific network structures and can be compared with existing localization distillation losses. Our proposed method is simple but effective, and experimental results demonstrate its superiority over existing methods. Code is available at https://github.com/TinyTigerPan/BCKD.

1. Introduction

Recent progress in dense object detectors has yielded significant performance improvements in the object detec-
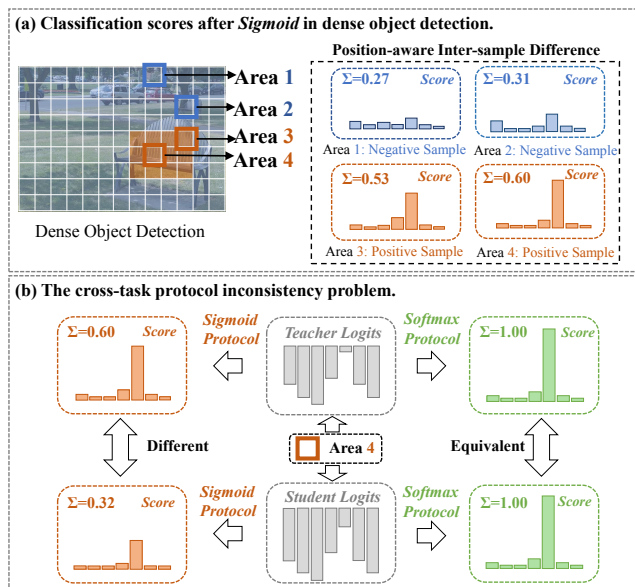

Figure 1. (a) In dense object detection, different samples exhibit inter-sample differences in their classification score sums on various positions on dense maps, which is significantly different from those in image classification. (b) The cross-task protocol inconsistency problem arises in dense object detection due to the mismatch between *Sigmoid* protocol used in this task and *Softmax* protocol used in classification distillation. Specifically, when classification distillation loss equals 0, inconsistencies emerge between the scores of the student and teacher models in dense object detection.

tion task [25, 29, 18, 17, 33]. However, the high computational burden of existing detection methods poses a significant challenge for deployment on resource-constrained devices. To address this problem, knowledge distillation (KD) [13, 5, 11, 35, 36, 4, 15, 22, 47, 6, 40] has emerged as a promising approach to compress models. The KD frame-

*The first two authors contributed equally to this paper.
†Corresponding author.

work involves training a smaller student model by leveraging a larger and more capable teacher model, for enhancing the student model's generalization ability.

Knowledge distillation approaches can be roughly classified into two categories: feature-based distillation methods [27, 37, 23, 1, 23, 4, 36] and logit-based distillation methods [13, 43, 34, 47]. In object detection, existing knowledge distillation methods have focused primarily on feature-based distillation due to the marginal performance gain from original logit-based distillation techniques [14, 32, 38]. However, it is worth exploring logit-based methods as they are usually simpler to use and have the potential to further improve performance when combined with feature-based methods. LD [47] is a representative logit-based distillation technique that transforms bounding boxes into probability distributions to facilitate localization distillation. However, classification distillation in dense object detection remains a challenge.

In this work, we further investigate this problem. Figure 1 (a) demonstrates that dense object detection faces a severe foreground-background imbalance problem when predicting classification scores on dense maps. Consequently, dense object detectors typically use the *Sigmoid* protocol to transfer classification logits to classification scores, which results in the position-aware inter-sample difference: Samples closer to positive sample regions generate higher classification score sums across all categories, indicating inter-sample differences. However, common classification distillation methods [13, 43, 34, 47] directly use the *Softmax* protocol from image classification to transfer classification logits to classification scores. The *Softmax* protocol normalizes classification scores, ignoring the absolute classification scores for individual categories and eliminating the inter-sample difference characteristic of classification scores. Additionally, in distillation, classification scores for each category are jointly optimized with inter-class dependencies, while in dense object detection, they are individually optimized without such dependencies. These differences lead to the **cross-task protocol inconsistency** problem, as shown in Figure 1 (b): when the teacher scores are equal to the student scores after *Softmax*, the classification distillation loss is 0, indicating that the student scores have achieved the optimal solution in the distillation loss. However, after *Sigmoid*, the student scores still differ from the teacher scores, showing lower score sums and incorrect inter-class relationships.

In addition to classification, localization is another crucial aspect of the object detection task. Although the localization distillation loss in LD [47] has demonstrated effectiveness, it requires the use of a Discrete Position-probability Prediction Head, such as the Generalized Focal Loss Head [17], for accurately predicting the localization probability distribution of each sample. Unfortunately, current object detectors [25, 29, 18] commonly use a Continuous Box-Offset Prediction Head, which means that the use of LD [47] would require specific training of teacher models to incorporate the Discrete Position-probability Prediction Head. This constraint limits the applicability of LD [47].

To address these issues outlined above, this paper proposes two novel distillation losses, Binary Classification Distillation Loss and IoU-based Localization Distillation Loss, tailored for classification and localization in dense object detectors. For classification, we convert cross-task **inconsistent** protocols into cross-task **consistent** protocols. Specifically, we treat the classification logit maps used in dense object detectors as K (*i.e.*, the number of categories) binary-classification maps. Then, we use the *Sigmoid* protocol to obtain scores and apply a binary cross entropy loss to distill each binary-classification map from teacher to student models, effectively solving the cross-task protocol inconsistency problem. For localization, we convert the special-structure-**dependent** localization distillation loss into a special-structure-**free** localization distillation loss. Specifically, we directly compute the Intersection over Unions (IoUs) between predicted bounding boxes generated by the teacher and student models and employ the IoU loss to minimize the difference between the IoU values and 1 (*i.e.*, the maximal IoU). Our approach is evaluated on widely used COCO [19] dataset, and our experimental results demonstrate that our method outperforms existing logit-based distillation methods and further boosts the existing feature-based distillation methods. Our contributions are summarized as follows:

(i) We identify the cross-task protocol inconsistency problem as the primary obstacle in utilizing original classification distillation techniques for dense object detection. The proposed Binary Classification Distillation Loss greatly enhances the performance gains obtained through classification distillation in dense object detection. We show that transferring semantic knowledge (*i.e.*, classification) alone can be effective in dense object detection, beyond common views in previous work.

(ii) We propose the IoU-based Localization Distillation Loss to distill the localization knowledge from teacher models to student models, which eliminates the need for specific training of teacher models.

(iii) Our proposed method is simple but effective, as demonstrated by our experiments. Besides, our method exhibits flexibility in integrating with existing state-of-the-art methods, resulting in a consistent performance increase.

2. Related Works

2.1. Object Detection

Object detection is a fundamental and challenging task in computer vision, involving the classification and localiza-

tion of objects within a given image. The literature on this topic can be broadly classified into two categories: region-based object detectors and dense object detectors. Region-based object detectors, including Faster-RCNN [26], Cascade R-CNN [3], and Fast R-CNN [10], utilize a Region Proposal Network (RPN) to generate Regions of Interest (RoIs), which are then refined through classification and regression heads to produce the final detection. In contrast, dense object detectors, such as YOLO [25], FCOS [29], RetinaNet [18], and GFL [17], directly predict objects from feature maps, offering advantages in terms of computational efficiency and ease of deployment when compared to region-based object detectors.

Most dense object detectors generate predictions of various sizes and proportions by utilizing dense proposals (such as anchor [18] and point [29]) at all positions on the image. Thus, they face the challenge of a severe imbalance between positive and negative samples, which can lead to poor performance. To address this, some works [20, 42] have explored complex re-sampling schemes for hard example mining. Besides, RetinaNet [18] uses the focal loss to prioritize the training of difficult samples. Additionally, different label assignment strategies, such as ATSS [41] and OTA [9], have been proposed to further improve performance. Through collective efforts, dense object detectors have achieved high accuracy and fast inference times. Recent research has also focused on improving the performance of compact real-time models through model compression techniques. For example, successful approaches include RTMDet [21] and YOLOv7 [30].

2.2. Knowledge Distillation

Knowledge Distillation (KD) is a model compression method that enables training of compact student models with guidance from more powerful teacher models. First introduced by Hinton et al. [13], KD has since been extensively studied in subsequent works [27, 37, 1, 24, 28, 43, 16, 48, 49, 31, 7, 45, 44, 46]. In classification, KD methods are typically classified into two categories: feature-based methods [27, 37, 23, 1] and logits-based methods [13, 43, 34]. Feature-based methods transfer knowledge by mimicking intermediate features from a teacher's hint layer, while logits-based methods by mimicking the logit outputs from the teacher's classifier. In object detection, KD was initially applied in [5], and many subsequent works have been proposed [5, 11, 35, 36, 4, 15, 22, 47, 6, 40] to improve student performance. Feature-based distillation remains the mainstream approach. For example, FGD [35] separates foreground and background and recovers missing information by rebuilding relationships among different pixels. PKD [4] relaxes constraints on the magnitude of features by mimicking the Pearson Correlation Coefficient. MGD [36] randomly masks some pixels of the student's feature and

leverages a simple generative block to force it to imitate the teacher's feature. DIC [12] explores the classifier-to-detector knowledge transfer. TLLM [50] explores "undistillable classes", focusing on scenarios where a significant disparity exists between teacher and student. Regarding logit-based distillation methods, LD [47] treats bounding box regression as probability distribution estimation, and argues that distilling localization knowledge is more effective than semantic knowledge in dense object detection.

In previous works, logit-based distillation methods in image classification are directly utilized to distill the semantic knowledge from teacher models to student models, and they commonly find that the semantic knowledge transfer seldom works for dense object detection. In this work, we argue that these approaches overlook the differences between object detection and image classification tasks, which leads to insufficient performance gains. To address this issue, we propose a novel classification distillation method tailored for dense object detection in this paper.

3. Methodology

3.1. Overview

A dense object detector can be represented as the combination of a feature extractor $f(\cdot)$ and a detection head $h(\cdot)$. Given an input image I, the detector first extracts features $F = f(I)$, and then generates the final prediction $P = h(F)$. The prediction P typically comprises classification logits $l \in \mathbb{R}^{n \times K}$ and localization offsets $o \in \mathbb{R}^{n \times 4}$, where n is the number of anchors or points in dense object detection, and K is the number of foreground categories. In existing knowledge distillation (KD) methods for dense object detection, knowledge is transferred from a frozen large teacher detector T_{det} to a small student detector S_{det}. For feature-based methods, the distillation loss is defined as $\mathcal{L}_{dis} = \text{loss}(F_t, F_s)$, where F_t and F_s indicate the features of T_{det} and S_{det}, respectively. For logits-based methods, the distillation loss is defined as $\mathcal{L}_{dis} = \text{loss}(P_t, P_s)$, where P_t and P_s indicate the predictions of T_{det} and S_{det}, respectively, and loss denotes the distillation loss function.

In this work, we propose two distillation losses tailored for classification and localization in dense object detection, as illustrated in Figure 2. We observe the cross-task protocol inconsistency problem between dense object detection and classification distillation loss, which impedes the effectiveness of the classification distillation in dense object detection. To address this problem, we introduce a novel Binary Classification Distillation Loss that converts the inconsistent cross-task protocol distillation into the consistent cross-task protocol distillation. Moreover, we find that existing localization distillation methods rely on the Discrete Position-probability Prediction Head, such as the Generalized Focal Loss Head [17], which requires specific training
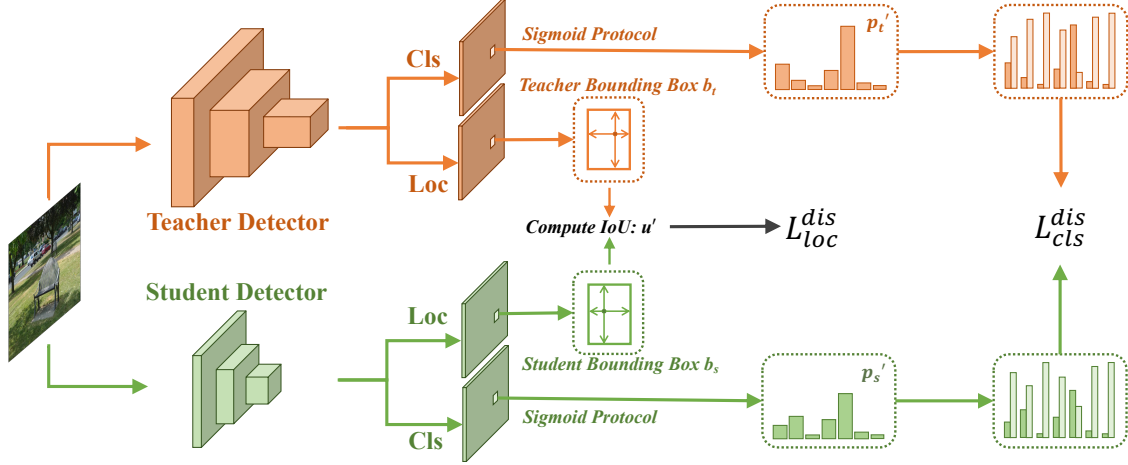
Figure 2. Distillation pipeline of our method. We leverage two novel distillation losses tailored for the object detection task. (i) Binary Classification Distillation Loss $\mathcal{L}_{cls}^{dis}$, which represents classification logit maps as multiple binary-classification maps and distills classification knowledge through a distillation loss similar to binary cross entropy. (ii) IoU-based Localization Distillation Loss $\mathcal{L}_{loc}^{dis}$, which transfers localization knowledge from teacher models to student models by computing the IoUs between predicted bounding boxes from both models and using the IoU loss. Best viewed in color.

of teacher models. To overcome this limitation, we propose a special structure-free IoU distillation loss that enables the distillation of localization knowledge from teacher models to student models.

3.2. Binary Classification Distillation Loss

Protocol in Dense Object Detectors: Dense object detectors aim to predict the corresponding classification score and bounding box for each sample point in dense maps generated from the entire image. However, as the background pixels occupy a significant portion of the image, foreground and background samples are severely imbalanced in dense object detectors. Specifically, during training, the majority of the samples are background samples. When using the *Softmax* protocol for transferring classification logits to classification scores, which assigns a sample to $K+1$ probabilities (where K is the number of foreground categories and an additional probability indicates the background), it may not be effective due to its tendency to assign higher probabilities to the majority class, *i.e.*, the background. Consequently, dense object detectors such as YOLO [25], FCOS [29], RetinaNet [18], and GFL [17] commonly use the *Sigmoid* protocol for transferring classification logits to classification scores. By modeling the multi-classification problem as multiple binary-classification problems, this approach can more effectively handle the foreground-background class imbalance issue.

Specifically, dense object detectors produce classification maps of varying sizes, with a size of $H \times W \times K$, where H, W and K represent the height, width and number of classes, respectively. Existing methods assign labels to each point on the classification map, with positive samples la-

beled as a one-hot tensor and negative samples labeled as a fully-zero tensor. Let x be a sample, and $l \in \mathbb{R}^{n \times K}$ denote its classification logits. To obtain classification scores for each point, existing methods use the *Sigmoid* protocol, *i.e.*, $p = Prot_{Sig}(l)$. We also have a label tensor y for x. Therefore, we can compute the binary cross entropy loss between the classification scores and labels:

$$\mathcal{L}_{cls}(x) = \sum_{i=1}^{n} \sum_{j=1}^{K} \mathcal{L}_{CE}(p_{i,j}, y_{i,j}), \qquad (1)$$

where $\mathcal{L}_{CE}(p_{i,j}, y_{i,j})$ is the binary cross entropy loss for the i-th position and j-th class, defined as:

$$\mathcal{L}_{CE}(p_{i,j}, y_{i,j}) = \begin{cases} -log(p_{i,j}) & y_{.,j} = 1, \\ -log(1 - p_{i,j}) & y_{.,j} = 0. \end{cases} \qquad (2)$$

Protocol in Common Classification Distillation: Common classification distillation methods [14, 32, 38, 47] are usually developed for the class-balanced scenario in image classification. The *Softmax* protocol plays a crucial role in establishing strong inter-class relationships, providing strong discriminative ability for identifying different categories in image classification. Therefore, the *Softmax* protocol is typically used in classification distillation.

Specifically, for a sample x, let l^t and l^s denote the classification logits from the teacher and student models, respectively. Existing methods use the *Softmax* protocol to obtain classification scores, *i.e.*, $p^t = Prot_{Smax}(l^t)$ and $p^s = Prot_{Smax}(l^s)$. The classification distillation loss is computed between p^t and p^s to encourage the student model to mimic the output of the teacher model. Specifically, this loss is typically defined as the Kullback-Leibler

(KL) divergence between teacher scores and student scores:

$$\mathcal{L}_{cls}^{kl}(x) = \mathcal{L}_{kl}(p^s, p^t), \tag{3}$$

where $\mathcal{L}_{cls}^{kl}(\cdot)$ denotes the classification distillation loss, and $\mathcal{L}_{kl}(\cdot, \cdot)$ denotes the Kullback-Leibler (KL) divergence.

Analysis of Cross-task Protocol Inconsistency: Existing distillation methods [47] in object detection typically apply the classification distillation loss used in image classification directly to dense object detection, leading to cross-task protocol inconsistency. Specifically, we firstly present the *Softmax* protocol and the *Sigmoid* protocol below:

$$Prot_{Smax}(l^t) = \frac{e^{l^t}}{\sum_{i=1}^{K} e^{l_i^t}}, Prot_{Sig}(l^t) = \frac{1}{1 + e^{-l^t}}, \tag{4}$$

where l^t is the logits of the teacher model. When n is a constant tensor with the same shape with l^t and $l^s = l^t + n$ (l^s is the logits of the student model), we have:

$$\begin{aligned} Prot_{Smax}(l^s) &= \frac{e^{l^t + n}}{\sum_{i=1}^{K} e^{l_i^t + n}} = \frac{e^{l^t} \cdot e^n}{\sum_{i=1}^{K}(e^{l_i^t} \cdot e^n)} \\ &= \frac{e^{l^t}}{\sum_{i=1}^{K} e^{l_i^t}} = Prot_{Smax}(l^t). \end{aligned} \tag{5}$$

Thus, the distillation loss is equal to zero, and there is no further transfer of localization knowledge from the teacher to the student model. However, $Prot_{Sig}(l_s) \neq Prot_{Sig}(l_t)$, resulting in a significant gap between the classification scores of the teacher and student models during inference. Typically, the scores obtained by the student model are lower than those of the teacher model and may have incorrect inter-class relationships. As a result, the student model cannot inherit the correct prediction ability from the teacher model.

Bridge Cross-task Protocol Inconsistency: To bridge cross-task protocol inconsistency, we propose a straightforward but effective solution. Specifically, we treat classification logit maps as multiple binary-classification maps during distillation. To achieve this, we compute $p^{t'} = Prot_{Sig}(l^t)$ and $p^{s'} = Prot_{Sig}(l^s)$, resulting in the binary-classification scores $p^{t'}$ and $p^{s'}$ with a size of $n \times K$. The classification distillation loss can then be calculated based on these binary-classification scores:

$$\begin{aligned} \mathcal{L}_{BCE}(p_{i,j}^{s}{}', p_{i,j}^{t}{}') &= \\ -((1 - p_{i,j}^{t}{}') \cdot log(1 - p_{i,j}^{s}{}') &+ p_{i,j}^{t}{}' \cdot log(p_{i,j}^{s}{}')), \\ \mathcal{L}_{cls}^{dis}(x) = \sum_{i=1}^{n} \sum_{j=1}^{K} &\mathcal{L}_{BCE}(p_{i,j}^{s}{}', p_{i,j}^{t}{}'), \end{aligned} \tag{6}$$

where $\mathcal{L}_{cls}^{dis}(\cdot)$ denotes the classification distillation loss, and $\mathcal{L}_{BCE}(\cdot, \cdot)$ denotes the binary cross entropy loss, $p_{i,j}^{s}{}', p_{i,j}^{s}{}'$

denotes the i-th position and j-th class of $p^{s'}$, $p^{t'}$, respectively.

Besides, we propose a loss weighting strategy for models to focus on distilling important samples, inspired by the Focal Loss [18]. Specifically, we compute the importance weighting w of the sample x as follows:

$$w = \left| p^{t'} - p^{s'} \right|, \tag{7}$$

where $w \in \mathbb{R}^{n \times K}$. Each element in w weighted to the classification distillation loss of sample x. Thus, the classification distillation loss in this paper is formulated as:

$$\mathcal{L}_{cls}^{dis}(x) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{i,j} \cdot \mathcal{L}_{BCE}(p_{i,j}^{s}{}', p_{i,j}^{t}{}'). \tag{8}$$

3.3. IoU-based Localization Distillation Loss

In addition to classification, another crucial aspect of object detection is localization. LD [47] transforms the bounding box into a probability distribution to tackle the localization distillation problem. In LD [47], a Discrete Position-Probability Prediction Head, such as the Generalized Focal Loss Head [17], is essential for precisely predicting the localization probability distribution of each sample. Regrettably, this type of head is not commonly employed in current object detectors [25, 29, 18] due to their complexity, especially in inference, resulting in the need for specific training of teacher models. To address this issue, we propose an innovative structure-free localization distillation loss, motivated by the Interaction-over-Union (IoU) loss widely used in dense object detectors, to replace the existing ones.

LD [47] discretizes the continuous regression range into a uniform discrete variable $[e_1, e_2, ..., e_n]^T$ with n intervals. To predict the n logits corresponding to each regression interval of each edge e, denoted by z_T and z_S for the teacher and student, respectively, a Discrete Position-Probability Prediction Head (*e.g.*, the Generalized Focal Loss Head) is needed. The generalized *Softmax* function is then employed to transform z_T and z_S into the probability distribution p_T and p_S, respectively. Finally, the Kullback-Leibler Divergence is used to minimize the distance between p_T and p_S. Although effective, this approach requires the use of a specific head, namely the Generalized Focal Loss Head, to predict discrete logits for all possible positions of each edge. Instead, these detectors typically predict continuous bounding box offsets that are more convenient for obtaining the predicted bounding box in inference. Therefore, the applicability of LD [47] is limited.

In this work, our objective is to transfer localization knowledge from teacher models to student models without relying on complex transformations of bounding box predictions. To achieve this, we leverage the most fundamental location relationship between two bounding boxes,

Method	Schedule	mAP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
GFocal-Res101(Teacher)	2x	44.9	63.1	49.0	28.0	49.1	57.2
GFocal-Res50(Student)	1x	40.1	58.2	43.1	23.3	44.4	52.5
LD [47]	1x	42.1(+2.0)	60.3(+2.1)	45.6(+2.5)	24.5(+1.2)	46.2(+1.8)	54.8(+2.3)
Ours	1x	**43.2(+3.1)**	**61.6(+3.4)**	**46.9(+3.8)**	**25.7(+2.4)**	**47.3(+2.9)**	55.9(+3.4)
LD [47] + Ours	1x	**43.2(+3.1)**	61.4(+3.2)	46.7(+3.6)	25.1(+1.8)	**47.3(+2.9)**	**56.1(+3.6)**
GFocal-Res101(Teacher)	2x	44.9	63.1	49.0	28.0	49.1	57.2
GFocal-Res34(Student)	1x	38.9	56.6	42.2	21.5	42.8	51.4
LD [47]	1x	41.0(+2.1)	58.6(+2.0)	44.6(+2.4)	23.2(+1.7)	45.0(+2.2)	54.2(+2.8)
Ours	1x	42.0(+3.1)	60.0(+3.4)	45.6(+3.4)	24.1(+2.6)	46.3(+3.5)	54.1(+2.7)
LD [47] + Ours	1x	**42.3(+3.4)**	**60.2(+3.6)**	**46.0(+3.8)**	**24.4(+2.9)**	**46.4(+3.6)**	**54.8(+3.4)**
GFocal-Res101(Teacher)	2x	44.9	63.1	49.0	28.0	49.1	57.2
GFocal-Res18(Student)	1x	35.8	53.1	38.2	18.9	38.9	47.9
LD [47]	1x	37.5(+1.7)	54.7(+1.6)	40.4(+2.2)	20.2(+1.3)	41.2(+2.3)	49.4(+1.5)
Ours	1x	38.6(+2.8)	56.4(+3.3)	41.7(+3.5)	21.4(+2.5)	42.0(+3.1)	50.0(+2.1)
LD [47] + Ours	1x	**38.9(+3.1)**	**56.6(+3.5)**	**42.0(+3.8)**	**22.2(+3.3)**	**42.5(+3.6)**	**50.8(+2.9)**

Table 1. Quantitative results of the proposed method and existing logits-based distillation methods for lightweight detectors. All results are evaluated on MS COCO *val2017*. Boldface indicates the best results.

Intersection over Union (IoU), as the distillation target. Specifically, we obtain localization maps from both the teacher and student models, and for a given input sample x, we denote the corresponding localization predictions from the teacher and student models in i-th position as o_i^t and o_i^s, respectively. We then obtain the bounding box for x by using the anchor position and localization prediction, where A_i denotes the i-th anchor. The bounding box for the teacher model and student model are obtained as $b_i^t = Decoder(A_i, o_i^t)$ and $b_i^s = Decoder(A_i, o_i^s)$, respectively. We compute the IoU between b_i^t and b_i^s, denoted as u_i'. In addition, we introduce a loss weighting strategy for models to focus on distilling important samples in the above section, which we also use for the localization distillation. Therefore, the localization distillation loss can be computed as:

$$\mathcal{L}_{loc}^{dis}(x) = \sum_{i=1}^{n} max(w_{\cdot,j}) \cdot (1 - u_i'). \quad (9)$$

The localization distillation loss is straightaway but comparable to existing localization distillation losses.

3.4. Total Distillation Loss

In this work, we introduce two novel distillation losses, namely Binary Classification Distillation Loss and IoU-based Localization Distillation Loss, for improving the performance of both classification and localization tasks. The proposed classification distillation loss is specifically designed for the classification task, whereas the IoU loss is developed for the localization task. The combined distillation loss is formulated as follows:

$$\mathcal{L}_{total}^{dis}(x) = \alpha_1 \cdot \mathcal{L}_{cls}^{dis}(x) + \alpha_2 \cdot \mathcal{L}_{loc}^{dis}(x), \quad (10)$$

where α_1 and α_2 are two hyper-parameters, denoting the loss weightings for the classification distillation loss and the localization distillation loss, respectively.

4. Experimental and Results

4.1. Datasets and Evaluation Metrics

To verify the effectiveness of the proposed method, we conducted experiments on the popular MS COCO dataset [19], which contains about 118k images in the *train* set, 5k in the *val* set, and 20k in the *test-dev* set spanning 80 categories. We choose the *train* set for training and the *val* set for testing. We report the detection mean average precision (mAP) as an evaluation metric, meanwhile under the different thresholds (*e.g.* AP$_{50}$) and scales (*e.g.* AP$_S$).

4.2. Main Results

In this paper, we rethink the limitations of the original Knowledge Distillation (KD) approach in dense object detection, and propose two novel distillation losses, namely the Binary Classification Distillation Loss and the IoU-based Localization Distillation Loss, to address the shortcomings of KD in the context of both classification (**Cls**) and localization (**Loc**) in detectors. Our proposed approach achieves notable performance improvements over the baseline method, without any additional costs.

Our proposed approach yields notable object detection performance improvements, as shown in Table 1. Specifically, we achieve mAP score improvements of +2.8, +3.1, and +3.1 when using GFocal-Res18, GFocal-Res34, and GFocal-Res50 as student models, respectively, significantly outperforming the state-of-the-art method LD [47]. More-

Method	Schedule	mAP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
RetinaNet-ResX101(Teacher)	1x	41.0	60.9	43.9	23.9	45.2	54.0
RetinaNet-Res50(Student)	1x	36.5	55.4	39.1	20.4	40.3	48.1
Ours	1x	39.2(+2.7)	58.7(+3.3)	42.0(+2.9)	22.4(+2.0)	43.1(+2.8)	52.1(+4.0)
MGD [36]	1x	39.6(+3.1)	59.0(+3.6)	42.4(+3.3)	**22.7(+2.3)**	43.9(+3.6)	53.0(+4.9)
PKD [4]	1x	39.7(+3.2)	59.0(+3.6)	42.4(+3.3)	22.5(+2.1)	44.2(+3.9)	53.7(+5.6)
MGD [36] + Ours	1x	**40.1(+3.6)**	59.5(+4.1)	**43.0(+3.9)**	22.3(+1.9)	**44.3(+4.0)**	53.3(+5.2)
PKD [4] + Ours	1x	**40.1(+3.6)**	**59.6(+4.2)**	42.8(+3.7)	22.3(+1.9)	**44.3(+4.0)**	**53.8(+5.7)**
FCOS-Res101(Teacher)	2x	40.8	60.0	44.0	24.2	44.3	52.4
FCOS-Res50(Student)	1x	36.6	56.0	38.8	21.0	40.6	47.0
Ours	1x	39.2(+2.6)	58.8(+2.8)	42.0(+3.3)	22.7(+1.7)	43.2(+2.6)	50.3(+3.3)
MGD [36]	1x	39.6(+3.0)	59.0(+3.0)	42.3(+3.5)	23.1(+2.1)	43.7(+3.1)	51.1(+4.1)
PKD [4]	1x	39.9(+3.3)	59.3(+3.3)	42.6(+3.8)	22.9(+1.9)	44.3(+3.7)	**51.4(+4.4)**
MGD [36] + Ours	1x	40.0(+3.4)	59.3(+3.3)	42.9(+4.1)	23.4(+2.4)	44.1(+3.5)	51.1(+4.1)
PKD [4] + Ours	1x	**40.2(+3.6)**	**59.5(+3.5)**	**43.0(+4.2)**	**23.7(+2.7)**	**44.5(+3.9)**	**51.4(+4.4)**

Table 2. Quantitative results of the proposed method combined with existing feature-based methods on different dense object detectors. All results are evaluated on MS COCO *val2017*. Boldface indicates the best results.

over, we achieve further mAP score improvements of +0.3 when combining LD [47] with our proposed method in GFocal-Res18 and GFocal-Res34.

Feature-based distillation methods such as MGD [36] and PKD [4] have shown powerful performance improvements. Fortunately, our method can be easily combined with these approaches to further enhance detector performance. As illustrated in Table 2, our method achieves mAP score improvements of +0.4 and +0.5 over PKD and MGD, respectively, when using RetinaNet as the basic dense object detector. Moreover, our proposed method is highly flexible and can be used with various dense object detectors. In the case of FCOS, our approach leads to significant performance improvements. Similar to the results with RetinaNet, our method yields mAP score improvements of +0.3 and +0.4 over PKD and MGD, respectively.

4.3. Ablation Analysis

Sensitivity Study of Different Losses. To demonstrate the effectiveness of our proposed Binary Classification Distillation Loss ($\mathcal{L}_{cls}^{dis}(x)$) and IoU-based Localization Distillation Loss ($\mathcal{L}_{loc}^{dis}(x)$), we conduct experiments on the GFocal student model. As shown in Table 3, both $\mathcal{L}_{cls}^{dis}(x)$ and $\mathcal{L}_{loc}^{dis}(x)$ contribute to improved detector performance, particularly in AP$_{50}$ and AP$_{75}$, which more impacts classification and localization performance, respectively. Furthermore, the combination of the two losses leads to significant performance improvements compared to the baseline.

Sensitivity Study of Different Hyper-parameters. Our proposed method employs two hyper-parameters, α_1 and α_2, to balance the Binary Classification Distillation Loss and IoU-based Localization Distillation Loss. As shown in Table 4 and Table 5, the experiments demonstrate that our

Method				GFocal Res101-Res50		
	Cls	Loc		mAP	AP$_{50}$	AP$_{75}$
Baseline				40.1	58.2	43.1
LD [47]	✓			40.4	58.9	43.4
		✓		41.8	59.5	45.4
	✓	✓		42.1	60.3	45.6
Ours	✓			42.0	60.9	45.6
		✓		42.3	60.0	45.9
	✓	✓		**43.2**	**61.6**	**46.9**

Table 3. Ablation study of distillation losses on different branch in detectors. *Cls* and *Loc* indicates distillation on classification and localization in detector head, respectively. which are represented as $\mathcal{L}_{cls}^{dis}(x)$ and $\mathcal{L}_{loc}^{dis}(x)$ in our proposed method. Boldface indicates the best results.

method is insensitive to the hyper-parameters and various values of α_1 and α_2 can lead to similar significant improvements in performance. Besides, we can achieve the best quantitative results when setting $\alpha_1 = 1.0$ and $\alpha_2 = 4.0$.

Visualization. In order to demonstrate the effectiveness of our proposed method in reducing classification errors, we compared the performance of the teacher detector and the student detector by forwarding the same image to both and recording the L1 error of summation of the classification score after *Sigmoid*. Figure 3 shows that our proposed method significantly reduces the ambiguity in classifying teachers and students in almost all locations at all FPN levels, thus validating the effectiveness of our method.

Self-KD. We have demonstrated the effectiveness of our proposed method for knowledge transfer from a strong
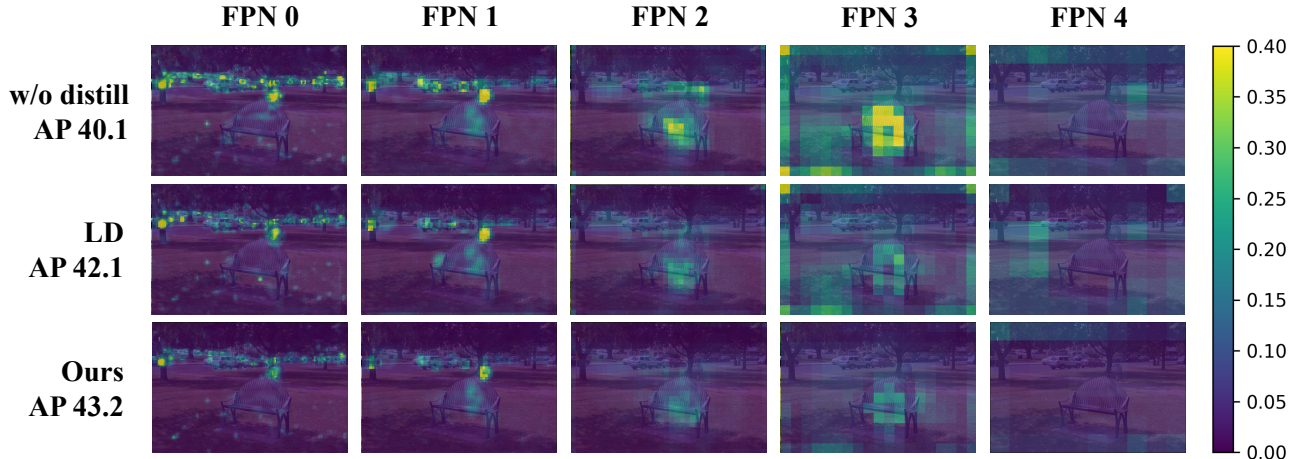
Figure 3. Visualization of L1 error summation of the classification score after *Sigmoid* between the teacher (GFocal-Res101) and the student (GFocal-Res50) at different levels of the Feature Pyramid Network (FPN). We can observe that our proposed method achieves a significant reduction in errors for almost all locations compared to the state-of-the-art method LD [47]. To better observe subtle differences, we bound the margin of error between 0 and 0.4. Darker is better. Best viewed in color.

α_1	0	0.25	0.5	1.0	1.5	2.0	3.0
mAP	40.1	41.5	41.9	**42.0**	41.9	41.5	41.2
AP_{50}	58.2	60.1	60.8	**60.9**	60.6	60.6	60.2
AP_{75}	43.1	44.9	45.4	**45.6**	45.4	44.7	44.5

Table 4. Ablation study of hyper-parameter α_1 on GFocal Res101-Res50. To show the sensitivity of $\mathcal{L}_{cls}^{dis}(x)$, we fix $\alpha_2 = 0$. Boldface indicates the best results.

α_2	0	0.5	1.0	1.5	2.0	4.0	5.0
mAP	40.1	41.3	41.6	41.8	42.2	**42.3**	42.1
AP_{50}	58.2	59.4	59.5	59.8	**60.1**	60.0	59.8
AP_{75}	43.1	44.7	44.9	45.4	**45.9**	**45.9**	45.8

Table 5. Ablation study of hyper-parameter α_2 on GFocal Res101-Res50. To show the sensitivity of $\mathcal{L}_{loc}^{dis}(x)$, we fix $\alpha_1 = 0$. Boldface indicates the best results.

Method	Self-KD	mAP	AP_{50}	AP_{75}
GFocal-Res50		40.1	58.2	43.1
	✓	**40.9**	**59.1**	**44.2**
GFocal-Res34		38.9	56.6	42.2
	✓	**39.4**	**57.2**	**42.6**
GFocal-Res18		35.8	53.1	38.2
	✓	**36.2**	**53.5**	**38.9**

Table 6. Quantitative results of proposed method under the self-KD strategy. Boldface indicates the best results.
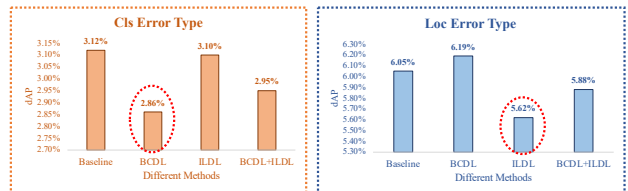

Figure 4. Error analysis conducted using the TIDE toolbox [2]. The decrease in average precision (dAP) resulting from two types of errors (*i.e.*, Cls, Loc) [2] is reported. The student model without any distillation losses is denoted as "Baseline", while the use of Binary Classification Distillation Loss and the application of IoU-based Localization Distillation Loss are denoted as "BCDL" and "ILDL", respectively.

teacher to a compact student in Table 1. However, in cases where a stronger teacher is not available, self-KD [8, 39] can still be employed for classification tasks. We apply $S_{det} = T_{det}$ to the dense object detection task with our method, where S_{det} is the student detector and T_{det} is the teacher detector. Table 6 shows that our proposed method can still yield performance gains under the self-KD strategy.

Error Analysis. The TIDE toolbox [2] is used to analyze the distribution of error types, as presented in Figure 4. The **Cls** error type indicated correctly localized but misclassified predictions, and the **Loc** error type indicated correctly classified but incorrectly localized predictions. The results showed two key findings: (i) The Binary Classification Distillation Loss effectively reduced **Cls** errors but did not contribute to reducing **Loc** errors. (ii) The IoU-based Local-

ization Distillation Loss effectively reduced **Loc** errors but did not contribute to reducing **Cls** errors. These results provide further evidence of the efficacy of Binary Classification Distillation Loss and IoU-based Localization Distillation Loss in enhancing classification and localization performance, respectively.

5. Conclusion

Our study reveals the cross-task protocol inconsistency is the reason behind the inefficiency of original classification distillation in dense object detection. To solve this problem, we present a novel Binary Classification Distillation Loss. Besides, we design an IoU-based Localization Distillation Loss for eliminating the need for specific structure. Experimental results demonstrate the effectiveness of our proposed method, especially in improving classification distillation performance. We expect that our work will provide valuable insights and encourage further research into logit-based distillation methods.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proc. CVPR*, pages 9163–9171, 2019. 2, 3

[2] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *Proc. ECCV*, pages 558–573, 2020. 8

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proc. CVPR*, pages 6154–6162, 2018. 3

[4] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *arXiv preprint arXiv:2207.02039*, 2022. 1, 2, 3, 7

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Proc. NeurIPS*, pages 16468–16480, 2017. 1, 3

[6] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proc. CVPR*, pages 7842–7851, 2021. 1, 3

[7] Yongjian Fu, Songyuan Li, Hanbin Zhao, Wenfu Wang, Weihao Fang, Yueting Zhuang, Zhijie Pan, and Xi Li. Elastic knowledge distillation by learning from recollection. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 3

[8] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proc. ICML*, pages 1607–1616. PMLR, 2018. 8

[9] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proc. CVPR*, pages 303–312, 2021. 3

[10] Ross Girshick. Fast r-cnn. In *Proc. ICCV*, pages 1440–1448, 2015. 3

[11] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proc. CVPR*, pages 2154–2164, 2021. 1, 3

[12] Shuxuan Guo, Jose M Alvarez, and Mathieu Salzmann. Distilling image classifiers in object detectors. *Advances in Neural Information Processing Systems*, 34:1036–1047, 2021. 3

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3

[14] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. In *Proc. NeurIPS*, pages 16468–16480, 2021. 2, 4

[15] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proc. AAAI*, volume 36, pages 1306–1313, 2022. 1, 3

[16] Xuewei Li, Songyuan Li, Bourahla Omar, Fei Wu, and Xi Li. Reskd: Residual-guided knowledge distillation. *IEEE Trans. Image Process.*, 30:4735–4746, 2021. 3

[17] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Proc. NeurIPS*, pages 21002–21012, 2020. 1, 2, 3, 4, 5

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. ICCV*, pages 2980–2988, 2017. 1, 2, 3, 4, 5

[19] 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 *Proc. ECCV*, pages 740–755. Springer, 2014. 2, 6

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. ECCV*, pages 21–37. Springer, 2016. 3

[21] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 3

[22] Chuong H Nguyen, Thuy C Nguyen, Tuan N Tang, and Nam LH Phan. Improving object detection by label assignment distillation. In *"Proc. WACV*, pages 1005–1014, 2022. 1, 3

[23] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proc. CVPR*, pages 3967–3976, 2019. 2, 3

[24] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *Proc. ICML*, pages 5142–5151. PMLR, 2019. 3

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. CVPR*, pages 779–788, 2016. 1, 2, 3, 4, 5

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, pages 6906–6919, 2015. 3

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2, 3

[28] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? In *Proc. NeurIPS*, pages 6906–6919, 2021. 3

[29] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proc. ICCV*, pages 9627–9636, 2019. 1, 2, 3, 4, 5

[30] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 3

[31] Hui Wang, Hanbin Zhao, Xi Li, and Xu Tan. Progressive blockwise knowledge distillation for neural network acceleration. In *IJCAI*, pages 2769–2775, 2018. 3

[32] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proc. CVPR*, pages 4933–4942, 2019. 2, 4

[33] Shaokai Wu and Fengyu Yang. Boosting detection in crowd analysis via underutilized output features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15609–15618, 2023. 1

[34] Zhendong Yang, Zhe Li, Yuan Gong, Tianke Zhang, Shanshan Lao, Chun Yuan, and Yu Li. Rethinking knowledge distillation via cross-entropy. *arXiv preprint arXiv:2208.10139*, 2022. 2, 3

[35] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proc. CVPR*, pages 4643–4652, 2022. 1, 3

[36] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *Proc. ECCV*, pages 53–69. Springer, 2022. 1, 2, 3, 7

[37] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proc. CVPR*, pages 4133–4141, 2017. 2, 3

[38] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *Proc. ICLR*, 2021. 2, 4

[39] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proc. ICCV*, pages 3713–3722, 2019. 8

[40] Peizhen Zhang, Zijian Kang, Tong Yang, Xiangyu Zhang, Nanning Zheng, and Jian Sun. Lgd: label-guided self-distillation for object detection. In *Proc. AAAI*, volume 36, pages 3309–3317, 2022. 1, 3

[41] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proc. CVPR*, pages 9759–9768, 2020. 3

[42] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proc. CVPR*, pages 4203–4212, 2018. 3

[43] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proc. CVPR*, pages 11953–11962, 2022. 2, 3

[44] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[45] Hanbin Zhao, Hui Wang, Yongjian Fu, Fei Wu, and Xi Li. Memory-efficient class-incremental learning for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5966–5977, 2021. 3

[46] Hanbin Zhao, Fengyu Yang, Xinghe Fu, and Xi Li. Rbc: Rectifying the biased context in continual semantic segmentation. In *European Conference on Computer Vision*, pages 55–72. Springer, 2022. 3

[47] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proc. CVPR*, pages 9407–9416, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[48] Dawei Zhou, Tongliang Liu, Bo Han, Nannan Wang, Chunlei Peng, and Xinbo Gao. Towards defending against adversarial examples via attack-invariant features. In *International Conference on Machine Learning*, pages 12835–12845. PMLR, 2021. 3

[49] Dawei Zhou, Nannan Wang, Bo Han, and Tongliang Liu. Modeling adversarial noise for adversarial training. In *International Conference on Machine Learning*, pages 27353–27366. PMLR, 2022. 3

[50] Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. In *Advances in Neural Information Processing Systems*, 2022. 3