Task-Balanced Distillation for Object Detection

Ruining Tang[1], Zhenyu Liu[1], Yangguang Li[2], Yiguo Song[1], Hui Liu[1], Qide Wang[1],
Jing Shao[2], Guifang Duan[1*], Jianrong Tan[1]
[1] State Key Lab of CAD&CG, Zhejiang University, [2] SenseTime Research
{tangruining,liuzy,ygsong,liuhui2017,btwqd,gfduan,egi}@zju.edu.cn
liyangguang@sensetime.com, shaojing@senseauto.com

Abstract

Mainstream object detectors are commonly constituted of two sub-tasks, including classification and regression tasks, implemented by two parallel heads. This classic design paradigm inevitably leads to inconsistent spatial distributions between classification score and localization quality (IOU). Therefore, this paper alleviates this misalignment in the view of knowledge distillation. First, we observe that the massive teacher achieves a higher proportion of harmonious predictions than the lightweight student. Based on this intriguing observation, a novel Harmony Score (HS) is devised to estimate the alignment of classification and regression qualities. HS models the relationship between two sub-tasks and is seen as prior knowledge to promote harmonious predictions for the student. Second, this spatial misalignment will result in inharmonious region selection when distilling features. To alleviate this problem, a novel Task-decoupled Feature Distillation (TFD) is proposed by flexibly balancing the contributions of classification and regression tasks. Eventually, HD and TFD constitute the proposed method, named Task-Balanced Distillation (TBD). Extensive experiments demonstrate the considerable potential and generalization of the proposed method. Specifically, when equipped with TBD, RetinaNet with ResNet-50 achieves 41.0 mAP under the COCO benchmark, outperforming the recent FGD and FRS.

1. Introduction

In recent years, the development of object detectors has drawn wide attention of the computer vision community, especially with the growth of convolutional neural networks (CNNs). As a fundamental pillar of the computer vision task, object detectors have been universally involved in all walks of life, such as autonomous driving, security monitoring, and pedestrian detection. In general, mainstream object detectors [1, 22, 24, 34, 38, 48] can be approximately divided into two-stage detectors [1, 34] and one-stage detectors [22,24,38,48] depending on whether the region proposal network (RPN) is implemented.

To generate both the location coordinates and the corresponding label for an object, modern object detectors typically adopt a multi-task pipeline, which consists of a classification branch and regression branch, implemented by two parallel heads. However, this parallel implementation may lead to inconsistent distributions of classification score and regression quality (IOU). As shown in the top sub-figures of Fig.1, the vanilla RetinaNet outputs inconsistent predictions due to the overlap between the person and motorcycle. Specifically, the green candidate contains a high score but a low IOU, whereas the orange one has an accurate bbox but a low score. When the post-procedure (e.g., Non-Maximum Suppression) is executed, the green one with a larger score will be reserved since the classification score is used as a general criterion for NMS ranking. As a result, the prediction with an accurate bbox (orange one) may be mistakenly filtered. Generally speaking, this incorrect filtering can be attributed to the inconsistent distribution between classification and localization accuracy.

Previous works [9, 10, 17, 19, 38, 40] attempt to overcome this problem in three ways, including recomposing the NMS score via adding an additional head (*i.e.*, IOUNet [17], Centerness [38, 48]), focusing on consistent regions [19, 38], and enhancing the dependency between classification and regression tasks to output more harmonious predictions [9, 40]. Although these studies have made remarkable progress in alleviating the influence of the inconsistent spatial distributions, the motivations and solutions are derived from the detector itself. Different from these methods above, this paper alleviates this inherent problem in the view of knowledge distillation by designing a customized teacher-student training workflow.

To elicit the proposed method, we meticulously compare the behaviour between the teacher and student models in handling this inharmonious distribution. A valu-

[*] Corresponding author

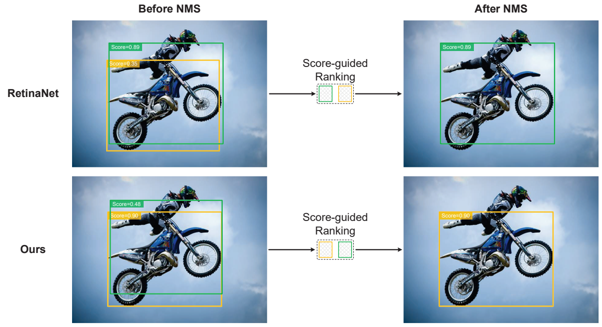


Figure 1. Visualization of the NMS mechanisms between the vanilla RetinaNet (top sub-figures) and the proposed model (bottom sub-figures). For ease of understanding, only two samples are shown here. The vanilla RetinaNet generates inconsistent predictions, leading to inaccurate preservation (green bbox). After equipping with the proposed method, the high-quality candidate (orange bbox) is conclusively preserved.

Table 1. The IOU distributions of easy-classified predictions on COCO *minival* split. Specifically, predictions with scores larger than 0.9 are counted.

Model	$IOU \geq 0.9$	$0.5 \leq IOU < 0.9$	$IOU < 0.5$
Teacher	69.2%	29.72%	1.0%
Student	67.43%	31.68%	0.9%
HD (ours)	70.97%	28.4%	0.58%
Relative gains	**+5.25%**	**-10.35%**	**-35.56%**

able observation is derived as follows. **In general, the teacher is more inclined to generate harmonious predictions than the student.** We count the IOU distributions of easy-classified predictions, as demonstrated in Table 1. Overall, the teacher model performs superiorly at achieving highly consistent predictions (*i.e.*, 69.2 vs. 67.43), while maintaining fewer inharmonious predictions (*i.e.*, 29.72 vs. 31.68). This observation indicates that owing to the disparate distributions of classification and regression qualities, some easy-classified samples may suffer from inaccurate locations for the student (*e.g.*, the motorcycle appeared in Fig. 1). Therefore, one meaningful question is whether the student can generate more harmonious predictions with the assistance of the teacher model.

In addition, we observe that **the inharmonious distributions of two sub-tasks will affect the selection of substantial areas when distilling features.** For transferring the intermediate features from the teacher to the student, a meaningful route [5,18,20,41,43,47,51] is how to screen the significant regions. Previous works [20, 51] attempt to generate spatial masks to denote the meaningful areas by using the predictions of the classification branch. However, considering the inharmonious distributions between classification and localization accuracy [9], purely utilizing the classification information might result in sub-optimal region selection. Therefore, another critical question is whether the classification and localization information can be fully utilized to guide the feature imitation.

A novel **Harmony Distillation (HD)** component is devised to achieve the transformation of harmonious predictions. Firstly, the Harmony Score (HS) is defined to quantitatively describe the deviation of the classification score and the corresponding regression quality. In particular, a large HS implies the classification score is positively correlated with the regression quality and vice versa. Secondly, the HD is derived by aligning the HS between teacher and student models. The proposed HD affords prior knowledge that models the relationship between classification and regression to assist the generation of high-quality predictions for the student. As presented in Tab. 1, the proportion of harmonious predictions is significantly improved, even surpassing that of the teacher.

To achieve the effective feature imitation, a new **Task-decoupled Feature Distillation (TFD)** is devised to integrate the information from both classification and regression tasks. The classification-aware and localization-aware masks are firstly obtained by using the corresponding predictions. Furthermore, instead of combining these masks with a heuristic weight scheme, we propose a Task-collaborative Weight Generation (TWG) module to balance the contributions of classification and regression tasks. Concretely, TWG dynamically assigns the task-aware weights according to both teacher's and student's predictions.

The proposed **Task-Balanced Distillation (TBD)** consists of the above HD and TFD, jointly considering the properties of classification and localization. To evaluate the effectiveness of the proposed method, we conduct experiments on the Pascal VOC [7], COCO [25], TJU-DHD [29], and Cityscapes [3] benchmarks. Abundant experimental results demonstrate the effectiveness and generalization of the proposed method. For instance, when equipped with the proposed TBD, RetinaNet-R50 achieves 41.0 mAP, surpassing the baseline by a large margin (*i.e.*, 3.6 mAP), even outperforming the current SOTA methods such as FRS [51] and FGD [43].

To sum up, the contributions of this paper are summarized as follows:

- A new Harmony Score (HS) is firstly defined to capture the relationship between classification and regression qualities. Then a novel Harmony Distillation (HD) is proposed to assist the generation of harmonious predictions for the student.

- A novel Task-decoupled Feature Distillation (TFD) is devised to mimic the intermediate features. The classification and regression masks are synthetically combined by balancing the contributions of these two tasks.

- Abundant experiments among various datasets and detectors are conducted. In addition, we achieve effective distillation between homogeneous (CNN to CNN) and heterogeneous (Transformer to CNN) backbones. The proposed method is easily plugged in and achieves SOTA performance.

2. Related Works

2.1. Object Detection

With the wide application of deep learning technology, the architecture of object detectors has progressively moved towards an end-to-end pipeline. Inherited from the ideology of pioneers, Faster RCNN [34] innovatively applies a two-stage scheme to detect objects. The two-stage detectors primarily consist of two core components. The first component completes the generation of potential candidates, whereas the second one further enables precise classification and regression based on these candidates. This architecture is popularized by its variants [1, 4, 8, 13, 31]. On the contrary, well-known one-stage detectors [9, 22, 24, 26, 27, 33, 37, 38] vastly simplified this two-step paradigm, and they directly make predictions based on the learned features.

2.2. Harmonious Predictions

The content of the inharmonious prediction is originally derived from reference [17], which means a predicted bounding box with misaligned classification and localization accuracy. This misalignment makes the Non-Maximum Suppression (NMS) procedure unreliable since the NMS only uses the classification score as the metric to rank the proposals, resulting in inaccurate suppression. To alleviate this problem, previous works [9, 10, 17, 22, 38, 40] attempt to make the predictions more harmonious. The route to tackling this issue can be divided into three categories, including reformulating the ranking metric [10, 17, 38], focusing on harmonious regions [22, 38], and enhancing the dependency between classification and localization tasks [9, 40]. IOUNet [17] utilizes an extra head to predict the localization-aware score and reformulate the NMS score to pay more attention to the localization task. This paradigm has been popularized by subsequent works such as FCOS [38] and DIR [10]. In addition, FCOS [38] proposes a centering sampling strategy based on the observation that the center region of GT usually has high classification and regression accuracy. Unlike these methods, GFL [22] incorporates the IOU into the classification label. TOOD [9] proposes a novel T-head with task alignment learning (TAL) to enhance the interaction between classification and regression tasks. HarmonicDet [40] improves the prediction consistency from the perspective of loss function excogitation.

Unlike the previous works, the proposed method generates harmonious predictions from the perspective of knowledge distillation. We first define the Harmony Score (HS) to capture the harmonious relationship between classification score and localization IOU. Then the HS of the teacher model is viewed as prior knowledge and ultimately transferred to the student with the supervision of the proposed HD.

2.3. Detection-oriented Knowledge Distillation

Knowledge Distillation (KD) is initially proposed in [15] to compress cumbersome models and has achieved remarkable progress in image classification. Existing KD-based methods can be broadly divided into response-based methods [15, 23, 28, 44], feature-based methods [14, 35, 45], and relation-based methods [30, 39]. After that, KD has been increasingly applied to the object detection task, achieving significant improvements. Compared with the image classification task, since the image used for object detection normally contains a mass of background pixels, one of the technical routes of the detection-based knowledge distillation is to select suitable distillation regions. Mimicking [21] mimics the feature divergence between teacher and student proposals. FGFI [41] claims that the regions near the ground truths should be regarded as the crucial distillation areas. GID [5] defines GIs based on the predictions and proposes distillation in an instance-wise manner. DeFeat [11] confirms that both foreground and background areas are valuable and then proposes a spatial-decoupled distillation to achieve feature imitation. PFI [20] and FRS [51] propose a prediction-guided distillation by utilizing the classification score. FGD [43] further decouples the feature imitation at the spatial and channel dimensions.

The primary difference between the proposed TFD and the above works is listed as follows. We revisit the selection of valuable feature areas from the perspective of inharmonious task distributions. In particular, both the classification-aware and localization-aware regions are regarded as valuable areas. Moreover, the TWG module is proposed to dynamically assign weights to balance the contributions of these two tasks.

3. Proposed Method

This section systematically expounds on the overall architecture of the proposed TBD. As demonstrated in Fig.2, the proposed TBD consists of Harmony Distillation (HD) and Task-decoupled Feature Distillation (TFD), elaborated in 3.1 and 3.2, respectively.

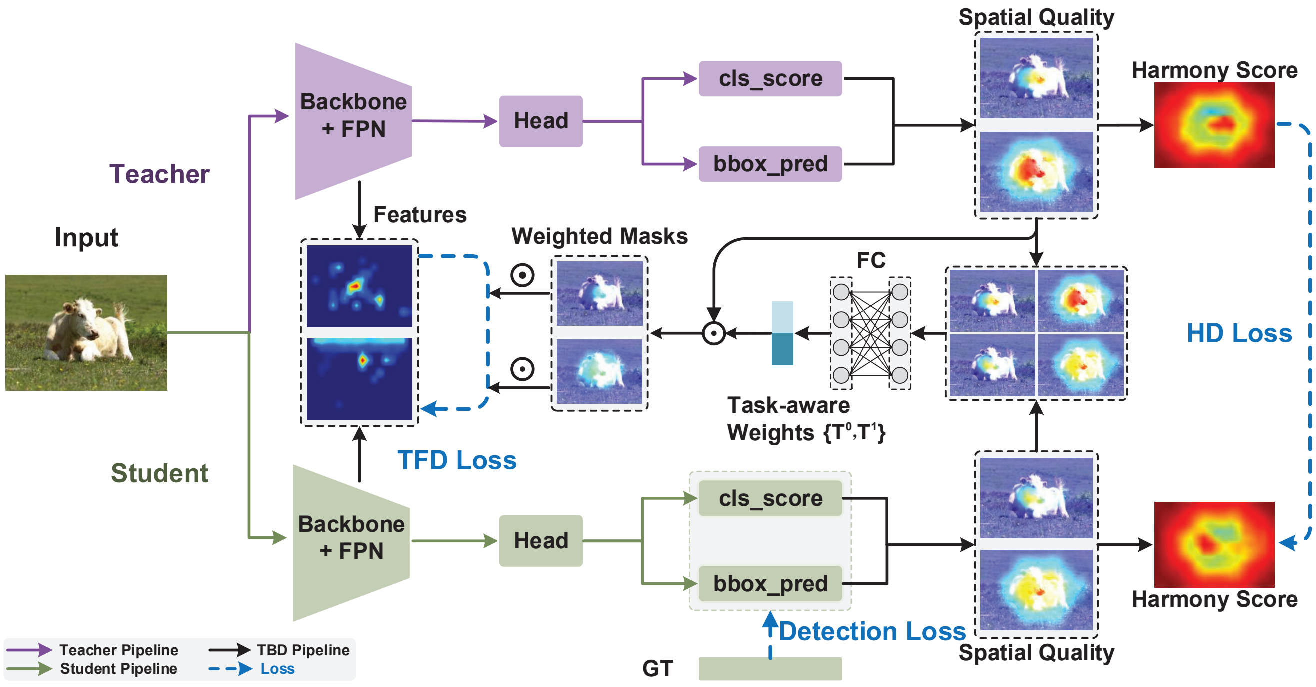


Figure 2. The whole architecture of the proposed TBD. For simplicity, only single-level feature and prediction are shown here.

3.1. Harmony Distillation

In this subsection, we progressively expound on the derivation of HD. As discussed above, the divergence between classification score and localization quality will lead to incorrect NMS suppression. Therefore, the primary point is to quantify this divergence, called Harmony Score (HS). In general, the derivation of HS consists of two steps: converting the prediction into the classification and localization probabilities and then devising the expression of HS based on the task probabilities.

For each predicted bounding box, P_{cls} and P_{reg} are used to denote the outputs of classification and regression branches, respectively. Concretely, P_{cls} is a score vector of C dimensions, where C represents the number of classes. Similarly, P_{reg} encodes the regularized offsets from the anchor (anchor box or anchor point) to the actual prediction.

The defining principle of task-specific probability comprises two parts. First, the probability amplitude should be normalized to $[0, 1]$. Second, a large probability signifies a precise prediction. For the classification task, the probability p_c is graciously generated by succinctly using the normalized maximum activation value:

$$p_c = \texttt{softmax}\left(\max_{1\le k\le C} P^k_{cls}\right) \tag{1}$$

where P^k_{cls} is the k-th element of classification prediction, and softmax is a spatial-wise softmax function to normalize the reserved classification score.

For the regression task, the implementation is similar to the above one. Concretely, the normalized prediction is firstly converted to the actual bounding box. Then we evaluate the IOU scores among each bounding box and the overall ground truths (GTs). For each predicted box, only the largest IOU score is preserved as the regression probability p_r.

$$p_r = \max_{1\le g\le G} IOU\left(\texttt{decode}(P_{reg}), GT_g\right) \tag{2}$$

where G denotes the number of GTs in each image, and g is defined as the corresponding index. decode represents the transformation function to obtain the actual prediction.

According to Equations 1 and 2, given a predicted bounding box, the classification and localization qualities are straightforwardly expressed by a binary tuple (p_c, p_r). Based on this definition, it is unequivocal to derive the formulation of HS. Theoretically, the definition of HS should satisfy two requirements. First, the formulation is monotonically decreasing. For instance, a diminutive divergence between classification and localization probabilities indicates encouraging consistency; therefore, a high HS should be achieved. Second, the definition should be bounded, which is conducive to avoiding the needless learning dilemma. Based on the above guidelines, HS is arranged through the activation function tanh.

$$\Delta p = |p_r - p_c| \tag{3}$$

$$HS = 1 - \texttt{tanh}\left(\Delta p\right) = 2 \times \frac{e^{-\Delta p}}{e^{-\Delta p} + e^{\Delta p}} \tag{4}$$

Since the teacher model performs better than the student in handing inharmonious predictions, a natural thought is to transfer the HS of the teacher as new knowledge to guide the student's learning phase. For distinction, the superscript t, s are used to denote the HS of teacher and student, respectively. In addition, $L1$ Loss is conducted to implement the knowledge transformation. The expression of HD is demonstrated as follows:

$$\mathcal{L}_{HD} = \sum_{l=1}^{L} \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} \left| HS_{i,j,l}^{t} - HS_{i,j,l}^{s} \right| \tag{5}$$

where l denotes the l-th FPN layer, and i, j are the spatial positions. W and H correspond to the spatial width and height of the prediction.

Moreover, we notice that Equation 5 is calculated with equal contribution of foreground and background predictions. To highlight the contributions of foreground predictions, an IOU-guided harmony loss is established. The p_r^t generated by the teacher model is employed as a spatial mask to up-weight the significant foreground locations. In addition, a dynamic modulation factor $\sqrt{1 + |p_c^t - p_c^s|}$ is introduced to magnify the loss of unfaithful predictions that have large performance gaps with the teacher. Considering the above two points, the spatial mask $\Psi_{i,j,l}$ and Equation 5 can be amended as:

$$\Psi_{i,j,l} = p_{r,i,j,l}^{t} \times \sqrt{1 + \left| p_{c,i,j,l}^{t} - p_{c,i,j,l}^{s} \right|} \tag{6}$$

$$\mathcal{L}_{HD} = \sum_{l=1}^{L} \frac{\sum_{i=1}^{W} \sum_{j=1}^{H} \Psi_{i,j,l} \left| HS_{i,j,l}^{t} - HS_{i,j,l}^{s} \right|}{\sum_{m=1}^{W} \sum_{n=1}^{H} \Psi_{m,n,l}} \tag{7}$$

3.2. Task-decoupled Feature Distillation

The proposed Task-decoupled Feature Distillation (TFD) is constructed based on FPN features $F = \{F_1, F_2, ..., F_L\}$. The reasons can be explained on two sides: for one thing, distilling FPN features can facilitate the imitation of both backbone and FPN features. For another, since most recent studies perform distillation on FPN features, it is natural to select FPN features to accomplish a fair comparison. The classic feature mimicking can be denoted as:

$$\mathcal{L}_{FPN} = \sum_{l=1}^{L} \sum_{i=1}^{W} \sum_{j=1}^{H} (F_{i,j,l}^{t} - \phi(F_{i,j,l}^{s}))^2 \tag{8}$$

where $\phi(\cdot)$ denotes the adaptive layer to align the teacher and student features. As can be seen from the definition in Equation 8, the loss will be dominated by the background regions since the background pixels are far more than the foreground ones in the object detection task. Therefore, how to determine the distillation area is a valuable topic.

Unlike the previous works [20, 41, 43, 51], we revisit the selection of crucial areas from the perspective of task-aware spatial distributions. The theoretical reasons for combining classification-aware and localization-aware regions are listed here. On the one hand, the FPN features are the pillar of subsequent classification and regression heads, so combining them is a natural choice. On the other hand, since the previous work [9] reveals the inharmonious distributions between two sub-tasks, only applying the classification mask [51] might miss some localization-aware regions. Based on the above analysis, the proposed TFD completely utilizes the prediction probabilities (p_c^t, p_r^t) of the teacher to generate task-aware masks. The mathematics formula is shown as follows:

$$\begin{aligned} \mathcal{L}_{TFD} = & \sum_{l=1}^{L} \frac{\omega_c \cdot \sum_{i=1}^{W} \sum_{j=1}^{H} p_{c,i,j,l}^{t} (F_{i,j,l}^{t} - \phi(F_{i,j,l}^{s}))^2}{\sum_{m=1}^{W} \sum_{n=1}^{H} p_{c,m,n,l}^{t}} \\ & + \sum_{l=1}^{L} \frac{\omega_r \cdot \sum_{i=1}^{W} \sum_{j=1}^{H} p_{r,i,j,l}^{t} (F_{i,j,l}^{t} - \phi(F_{i,j,l}^{s}))^2}{\sum_{m=1}^{W} \sum_{n=1}^{H} p_{r,m,n,l}^{t}} \end{aligned} \tag{9}$$

where ω_c and ω_r are hyper-parameters to control the weights of classification-aware and localization-aware losses. However, fixed weights applied in Equation 9 may suffer from some limitations. For instance, fixed weights are unenviable to adapt to the dynamic inputs comprehensively. In addition, extra hyperparametric optimization overhead is introduced compared with these methods that only utilize the classification mask. Therefore, we propose a Task-collaborative Weight Generation (TWG) module, to dynamically assign weights to overcome these limitations. Motivated by SENet [16], TWG only consists of two Fully-Connected (FC) layers and one softmax layer to generate the task-aware weights. Theoretically, the learned weights should be jointly determined by the teacher's prediction and the current learning state of the student. Based on this point, when implementing TWG, the prediction masks $(p_c^t, p_r^t, p_c^s, p_r^s)$ are firstly concatenated at the channel dimension:

$$\mathcal{P} = \texttt{concat}\left(p_c^t, p_r^t, p_c^s, p_r^s\right) \tag{10}$$

Then, the concatenated $\mathcal{P}$ is compressed by the average pooling operator. Two lightweight FC layers are subsequently added to generate task-aware weights. Eventually, the softmax function outputs the normalized weights to guarantee that the sum of these weights is 1. Note that we accomplish the implementation of TWG using the most straightforward way to avoid falling into the cumbersome network construction. Consequently, the learned weights can be mathematically expressed as:

$$T^0, T^1 = \texttt{softmax}\left(\texttt{FC}\left(\texttt{FC}\left(\texttt{AvgPool}\left(\mathcal{P}\right)\right)\right)\right) \tag{11}$$

Thus, given the task-aware weights $\{T^0, T^1\}$, Equation 9 can be rewritten as follows:

$$\mathcal{L}_{TFD-c} = \sum_{l=1}^{L} \frac{T_l^0 \cdot \sum_{i=1}^{W}\sum_{j=1}^{H} p^t_{c,i,j,l}(F^t_{i,j,l} - \phi(F^s_{i,j,l}))^2}{\sum_{m=1}^{W}\sum_{n=1}^{H} p^t_{c,m,n,l}}$$
$$\mathcal{L}_{TFD-r} = \sum_{l=1}^{L} \frac{T_l^1 \cdot \sum_{i=1}^{W}\sum_{j=1}^{H} p^t_{r,i,j,l}(F^t_{i,j,l} - \phi(F^s_{i,j,l}))^2}{\sum_{m=1}^{W}\sum_{n=1}^{H} p^t_{r,m,n,l}} \quad (12)$$

$$\mathcal{L}_{TFD} = \mathcal{L}_{TFD-c} + \mathcal{L}_{TFD-r} \quad (13)$$

3.3. Overall Loss

To sum up, the proposed model is trained in an end-to-end manner, and the whole loss includes the original detector loss and the customized distillation loss, demonstrated as follows:

$$\mathcal{L} = \mathcal{L}_{detector} + \alpha \cdot \mathcal{L}_{HD} + \beta \cdot \mathcal{L}_{TFD} \quad (14)$$

where $\mathcal{L}_{detector}$ is the original detector loss for the student model. α and β are hyper-parameters introduced in HD and TFD to balance the distillation loss.

4. Experiments

4.1. Datasets and Experimental Settings

Datasets. To verify the effectiveness and generalization of the proposed TBD, we conduct abundant experiments on four common datasets, including COCO [25], Pascal VOC [7], Cityscapes [3] and TJU-DHD [29]. Concretely, the main results are firstly reported on COCO, and then other datasets are selected to evaluate the generalization of the proposed method. Following the most common settings for COCO, the distillation models are trained on the 118k training split and evaluated on the *minival* split. In addition, we also report distillation results trained on COCO *mini-train* split [36], which is a curated mini-training set containing about 25k images. For simplicity, we use miniCOCO to describe this dataset. For Pascal VOC, the union of VOC2007 $trainval$ and VOC2012 $trainval$ is chosen for training, whereas the VOC2007 $test$ split is selected for evaluation. For Cityscapes, we use 5000 fine-labeled images for training and testing. For the TJU-DHD dataset, the traffic split is used as the benchmark, which is a diverse high-resolution dataset covering five common categories for generic object detection.

Settings. All the experiments are conducted based on mmdetection [2] toolbox. The hyper-parameters α, β in Equation 14 are set as $\{\alpha = 5.0, \beta = 0.01\}$. The impact of the hyper-parameters is discussed in Table 12. All the student models are pre-trained on ImageNet [6]. The inheriting strategy [18, 43] is not applied unless otherwise indicated. In addition, 1x and 2x indicate that the models are trained with 12 and 24 epochs, respectively. The initial learning rate is fixed as 0.01 and 0.02 for one-stage and two-stage detectors, respectively. Moreover, the batch size is set as 16 for all datasets, and the warm-up strategy is applied in the first 500 iterations to make the training procedure more stable. The remaining hyper-parameters are consistent with those in mmdetection.

4.2. Main Results

4.2.1 Comparison with SOTA methods

In this part, we compare the proposed TBD with existing state-of-the-art (SOTA) detection-based distillation methods. The main comparison experiments are implemented based on two well-known detectors, including RetinaNet [24]and Faster RCNN [34]. To verify the superiority of the proposed TBD, seven recent SOTA models [5, 12, 20, 41, 43, 47, 51] are used for comparisons, as shown in Table 2. In addition, to compare with a recent SOTA, named LD [50], supplementary experiments are implemented based on GFL [22] benchmark. Moreover, referring to the comparison settings of previous works [5, 43, 51], all the distillation models are conducted with various ResNet/ResNeXt models (*e.g.*, R101-R50, X101-R50). Complementary to the basic configurations, the distillation experiments of using other backbones are shown in Table 4.

As presented in Table 2, by equipping the proposed TBD, RetinaNet with the ResNet50 student can even exceed the teacher model by a large margin (40.0 vs. 38.9). When an enormous teacher (e.g., ResNeXt101-64x4d) is applied, the lightweight ResNet50 student can still achieve comparable performance. Additionally, according to the results, we can discover that no matter under the guidance of ResNet101 or ResNeXt101, the proposed TBD significantly outperforms the previous SOTA methods. Concretely, with the ResNeXt101 teacher and ResNet50 student, TBD outperforms FRS [51] and FGD [43] by 0.9 and 0.6, respectively. In addition, when compared with the existing methods on Faster RCNN, consistent gains are achieved, which indicates the effectiveness of the proposed model. Moreover, according to the experimental results shown in Table 3, the proposed TBD is also compatible with GFL, exceeding the recent LD [50] from 0.6 to 1.7 AP.

4.2.2 Results of TBD using various teacher-student configurations

This part shows the results of using diverse teacher-student configurations on COCO, including CNN-CNN, Transformer-Transformer, and Transformer-CNN. All the experiments are based on the RetinaNet. The correlative

Table 2. Comparison results of proposed TBD and existing SOTA methods on COCO *minival*. The symbol - means the results are not available in the original papers. T and S represent the teacher and student models.

Detector	Model	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Reference
RetinaNet	T:ResNet101	38.9	58.0	41.5	21.0	42.8	52.4	ICCV2017
	S:ResNet50	37.4	56.7	39.6	20.0	40.7	49.7	ICCV2017
	FGFI [41]	38.6(+1.2)	58.7	41.3	21.4	42.5	51.5	CVPR2019
	GID [5]	39.1(+1.7)	59.0	42.3	22.8	43.1	52.3	CVPR2021
	FRS [51]	39.7(+2.3)	58.6	42.4	21.8	43.5	52.4	NIPS2021
	[20]	39.6(+2.2)	-	-	21.4	44.0	52.5	AAAI2022
	FGD [43]	39.6(+2.2)	-	-	**22.9**	43.7	53.6	CVPR2022
	TBD (ours)	**40.0(+2.6)**	**59.1**	**42.8**	22.2	**44.1**	**54.0**	NaN
	T:ResNeXt101	41.0	60.9	44.0	23.9	45.2	54.0	ICCV2017
	S:ResNet50	37.4	56.7	39.6	20.0	40.7	49.7	ICCV2017
	FKD [47]	39.6(+2.2)	58.8	42.1	22.7	43.3	52.5	ICLR2021
	DICOD [12]	37.9(+0.5)	-	-	20.5	41.3	50.5	NIPS2021
	FRS [51]	40.1(+2.7)	59.5	42.5	21.9	43.7	54.3	NIPS2021
	FGD [43]	40.4(+3.0)	-	-	23.4	44.7	54.1	CVPR2022
	TBD (ours)	**41.0(+3.6)**	**60.4**	**43.8**	**23.9**	**45.1**	**54.7**	NaN
	T:ResNet101	38.9	58.0	41.5	21.0	42.8	52.4	ICCV2017
	S:ResNet18	33.2	51.5	35.1	17.3	35.4	44.7	ICCV2017
	FKD [47]	35.9(+2.7)	54.4	38.0	17.9	39.1	49.4	ICLR2021
	FGD [43]	35.9(+2.7)	53.9	38.6	18.1	39.2	49.5	CVPR2022
	TBD (ours)	**37.1(+3.9)**	**55.5**	**39.8**	**19.4**	**40.4**	**51.6**	NaN
Faster RCNN	T:ResNet101	39.8	60.1	43.3	22.5	43.6	52.8	NIPS2015
	S:ResNet50	38.4	59.0	42.0	21.5	42.1	50.3	NIPS2015
	FGFI [41]	39.3(+0.9)	59.8	42.9	22.5	42.3	52.2	CVPR2019
	GID [5]	40.2(+1.8)	60.7	43.8	22.7	44.0	53.2	CVPR2021
	FGD [43]	40.4(+2.0)	-	-	22.8	44.5	53.5	CVPR2022
	TBD (ours)	**40.6(+2.2)**	**61.0**	**44.2**	**23.5**	**44.7**	**53.5**	NaN
	T:ResNet101	39.8	60.1	43.3	22.5	43.6	52.8	NIPS2015
	S:ResNet18	34.5	54.6	37.2	19.2	36.8	45.2	NIPS2015
	FKD [47]	37.0(+2.5)	57.2	39.7	**19.9**	39.7	**50.3**	ICLR2021
	FGD [43]	37.0(+2.5)	57.1	40.0	18.9	40.6	**50.3**	CVPR2022
	TBD (ours)	**37.3(+2.8)**	**57.3**	**40.1**	19.7	**40.8**	50.0	NaN

Table 3. Quantitative results of the proposed TBD and LD [50] on COCO2017 *minival*. The teacher model is ResNet101, and S represents the student model.

Detector	Model	mAP	AP_S	AP_M	AP_L
GFL	S:ResNet18	35.7	19.4	38.8	47.5
	LD [50]	37.5(+1.8)	20.2	41.2	49.4
	TBD (ours)	**39.2(+3.5)**	**22.5**	**43.0**	**51.9**
	S:ResNet34	38.9	21.5	42.8	51.4
	LD [50]	41.0(+2.1)	23.2	45.0	54.2
	TBD (ours)	**41.6(+2.7)**	**24.4**	**45.7**	**54.2**
	S:ResNet50	40.2	23.3	44.0	52.2
	LD [50]	42.1(+1.9)	24.5	46.2	54.8
	TBD (ours)	**43.4(+3.2)**	**25.9**	**47.6**	**55.6**

results are presented in Table 4. Overall, consistent proceeds are obtained, declaring that the proposed TBD is feasible, efficient, and stable. Specifically, using CNN teachers, the ResNet18 student with TBD can effectively obtain performance gains from 2.8 to 4.3. When the heterogeneous teacher-student pair is applied (*e.g.,* from PVT [42] to ResNet), the ResNet18 student overcomes the architecture divergence between teacher and student and finally achieves 4.3 mAP gains. Moreover, the improvements of experiments based on PVT [42] and RegNet [32] verify the superiority of the proposed method, as well.

4.2.3 Results of TBD on various detectors

In this piece, we implement the proposed TBD on six prevalent detectors, including two-stage detector Faster RCNN [34], Dynamic RCNN [46], and one-stage detector FreeAnchor [49], RetinaNet [24], GFL [22] and FSAF [52]. Here ResNet50 is adopted as the teacher while the lightweight ResNet18 is set as the student. All the models are trained with the 1x learning paradigm. When the proposed TBD is

Table 4. Results of applying TBD among diverse backbones on COCO. All the experiments are based on RetinaNet with the 1x training schedule.

Student	Teacher	Distillation Type	mAP
ResNet18	-	Baseline	31.9
	ResNet50	CNN-CNN	34.7 (+2.8)
	ResNet101	CNN-CNN	35.3 (+3.4)
	PVTb0 [42]	Trans-CNN	35.1 (+3.2)
	PVTb1 [42]	Trans-CNN	36.2 (+4.3)
ResNet50	-	Baseline	36.5
	PVTb1 [42]	Trans-CNN	39.5 (+3.0)
	PVTb2 [42]	Trans-CNN	40.5 (+4.0)
RegNetX 800MF [32]	-	Baseline	35.6
	RegNetX 3.2GF [32]	CNN-CNN	38.2 (+2.6)
PVTb0	-	Baseline	37.1
	PVTb2	Trans-Trans	39.7 (+2.6)

Table 5. Results of applying TBD on diverse detectors based on COCO.

Detector	Distill	mAP	AP_s	AP_m	AP_l
Faster RCNN	×	33.2	18.2	35.9	43.2
	√	35.4 (+2.2)	19.6	38.8	46.4
Dynamic RCNN	×	34.9	18.3	37.2	47.7
	√	36.9 (+2.0)	19.6	39.7	49.4
RetinaNet	×	31.9	16.4	34.6	43.4
	√	34.7 (+2.8)	17.9	38.0	47.6
GFL	×	35.7	19.4	38.8	47.5
	√	38.2 (+2.5)	20.6	41.7	50.2
FSAF	×	32.4	17.1	35.5	42.3
	√	35.1 (+2.7)	17.1	38.1	47.3
FreeAnchor	×	34.0	18.1	36.3	46.5
	√	37.2 (+3.2)	19.2	40.2	50.5

adaptive to the two-stage detector, the classification and localization masks are generated on Region Proposal Network (RPN). Table 5 summarizes the detailed results. Overall, the consistent improvements indicate that the proposed TBD is compatible with mainstream detectors.

4.2.4 Results of TBD on other datasets

The above experiments are completely implemented based on MS COCO. In this piece, we evaluate our TBD on other datasets. Concretely, the widely used Pascal VOC, miniCOCO, TJU-DHD, and Cityscapes are introduced to evaluate the performance of TBD on small datasets. Analogously, all the models are trained with RetinaNet-R18 baseline with ResNet50 as the teacher. The complete results are shown in Table 6. For small datasets, we notice that remarkable improvements are achieved with the assistance of the proposed TBD. For example, TBD dramatically improves the vanilla student model by 5.7 mAP on miniCOCO. The progress can be extended to 9.2 using the teacher training on the complete COCO train split. Furthermore, the consistent gains in Table 6 manifest that the proposed TBD performs magnificently among multifarious datasets.

Table 6. Results of applying TBD on other datasets. The last row of data in miniCOCO is obtained using the teacher model trained on full COCO split.

Datasets	Distill	mAP	AP_s	AP_m	AP_l
miniCOCO [36]	×	19.8	9.2	21.4	26.9
	√	25.5 (+5.7)	12.2	27.5	34.2
	√	29.0 (+9.2)	13.1	31.4	39.9
Pascal VOC [7]	×	48.8	17.5	32.0	54.6
	√	52.7 (+3.9)	18.6	35.9	58.4
Cityscapes [3]	×	30.1	10.6	31.6	47.7
	√	33.9 (+3.8)	12.5	34.5	54.6
TJU-DHD [29]	×	50.4	19.2	47.1	65.9
	√	52.5 (+2.1)	21.7	48.8	68.7

Table 7. The individual results of HD and TFD on COCO.

Model	HD	TFD	mAP	AP_S	AP_M	AP_L
T: R50	-	-	36.5	20.4	40.3	48.1
S: R18	-	-	31.9	16.4	34.6	43.4
	√		33.2 (+1.3)	17.2	36.2	44.2
		√	34.4 (+2.5)	17.6	37.8	46.7
	√	√	**34.7 (+2.8)**	**17.9**	**38.0**	**47.6**

4.3. Ablation Study

In this part, we conduct abundant experiments based on RetinaNet to demonstrate the effectiveness of each component and explore some implementation details of TBD. The analytical experiments are implemented on the ResNet18 with ResNet50 as the teacher. The whole experiments are trained with the 1x learning schedule.

4.3.1 Ablation study of each component

As presented in Table 7, the vanilla student model achieves 31.9 mAP. When the proposed method is applied, both HD and TFD can consistently promote student performance. Concretely, HD obtains 1.3 gains while TFD harvests 2.5 improvements. In addition, the combination of HD and TFD brings the maximum promotion (*i.e.*, 2.8 mAP).

4.3.2 Ablation study of HD

Definition of HS. In this piece, we dive deeper into the definition of HS. Two variants of HS are customized, named HS_{exp} and HS_{log}, respectively. In addition, L1 and L2 Loss are also introduced to evaluate the design of the distillation loss function. The expressions of HS_{exp} and HS_{log}

Table 8. The comparison of disparate definitions of HS. The experiments are conducted with RetinaNet on COCO2017. The teacher model is ResNet50, while the student is ResNet18.

HS	Loss	mAP	AP_S	AP_M	AP_L
-	-	31.9	16.4	34.6	43.4
HS_{exp}	L2	33.0 (+1.1)	16.8	35.7	44.6
HS_{exp}	L1	33.0 (+1.1)	16.8	35.8	44.3
HS_{log}	L2	32.8 (+0.9)	16.9	35.5	44.2
HS_{log}	L1	**33.3 (+1.4)**	16.9	36.2	**45.2**
HS_{tanh}	L2	33.0 (+1.1)	17.0	36.0	44.4
HS_{tanh}	L1	33.2 (+1.3)	**17.2**	**36.2**	44.2

Table 9. Comparisons between HD and other response-based distillation methods. `cls` and `loc` means the classification and localization task.

Model	Distillation	mAP	Knowledge
	Baseline	31.9	None
	KD [15]	32.4 (+0.5)	`cls` logits
T: R50	FRS [51]	33.0 (+1.1)	`cls` logits
S: R18	RM [20]	33.3 (+1.4)	Anchor rank
	HD	33.2 (+1.3)	Relationship between
	HD†	**33.9 (+2.0)**	`cls` and `loc`

† means we retrain the proposed HD with the configuration of RM. The values of KD and FRS are our reproduction results.

Table 10. Quantitative results of the proposed TFD. *cls* and *reg* represent using the classification-aware and regression-aware masks, respectively. *fixed* means the weights are optimized as the fixed hyper-parameters. In contrast, *dynamic* denotes the weights are generated by the proposed TWG.

Mask	Weight	mAP	AP_S	AP_M	AP_L
-	-	31.9	16.4	34.6	43.4
whole	*fixed*	33.3 (+1.4)	17.6	36.3	44.7
cls	*fixed*	34.0 (+2.1)	17.4	37.4	46.7
reg	*fixed*	34.0 (+2.1)	17.4	37.5	46.5
cls + reg	*fixed*	34.1 (+2.2)	17.3	37.6	46.4
cls + reg	*dynamic*	**34.4 (+2.5)**	**17.6**	**37.8**	**46.7**

Table 11. Comparison results of the proposed TFD and other feature imitation methods.

Model	Distillation	mAP	Key Region
	Baseline	31.9	*cls*
	FitNet [35]	33.3 (+1.4)	*cls*
T: R50	FRS [51]	34.0 (+2.1)	*cls*
S: R18	PFI [20]	34.2 (+2.3)	*cls*
	TFD (ours)	34.4 (+2.5)	*cls + loc*
	TFD† (ours)	**35.1 (+3.2)**	*cls + loc*

† means we retrain the proposed TFD with the configuration of PFI. The values of FitNet and FRS are our reproduction results.

are listed as follows.

$$
\begin{aligned}
HS_{exp} &= e^{-|p_c - p_r|} \\
HS_{log} &= \frac{1}{log(e + |p_c - p_r|)}
\end{aligned}
\tag{15}
$$

To make it convenient to distinguish, we use HS_{tanh} to represent the HS definition in Equation 4. The overall experiment results are shown in Table 8. We can discover that all the implementations of HS can consistently boost the baseline performance. In addition, HS_{log} and HS_{tanh} achieve superior performance compared with other definitions, which can improve the student model by 1.4 and 1.3. Moreover, we notice that HS_{tanh} slightly outperforms HS_{log} when HD is combined with TFD, so we prefer HS_{tanh} as the ultimate representation.

Relationship with other response-based methods. Table 9 compares the proposed HD with other response-based distillation methods such as KD [15], FRS [51], and RM [20]. Although the conventional soft label distillation [15] obtains remarkable improvements in the image classification task, the promotion is unsatisfactory when applied to the object detection task. FRS [51] ameliorates the traditional KD by introducing a feature richness mask. Unlike these methods using classification logits, the proposed HD captures the relationship between classification and localization tasks as the prior knowledge and outperforms KD and FRS by 0.8 and 0.2. When compared with RM, our proposed HD† also shows distinct advantages, demonstrating the tremendous potential of the proposed HD.

4.3.3 Ablation study of TFD

Impact of decoupling task-aware masks. This part verifies the effectiveness of decoupling classification-aware and localization-aware regions. Concretely, we delicately excogitate several comparison experiments for quantitative verification. As shown in Table 10, compared with distilling on the whole feature map, utilizing both the classification mask p_c^t and localization mask p_r^t can obviously promote effective distillation. Technically, p_c^t or p_r^t only capture the corresponding task information, which might neglect potential clues about another one since the spatial distributions of p_c^t and p_r^t might be disparate. In addition, we observe that the key to integrating the classification-aware and localization-aware regions is how to balance the contribution of each task. Obviously, the information on input characteristics and current training status cannot be considered comprehensively by using a fixed weight scheme, leading to inconspicuous improvement. When equipped with the proposed TWG, the performance of TFD obtains a noticeable

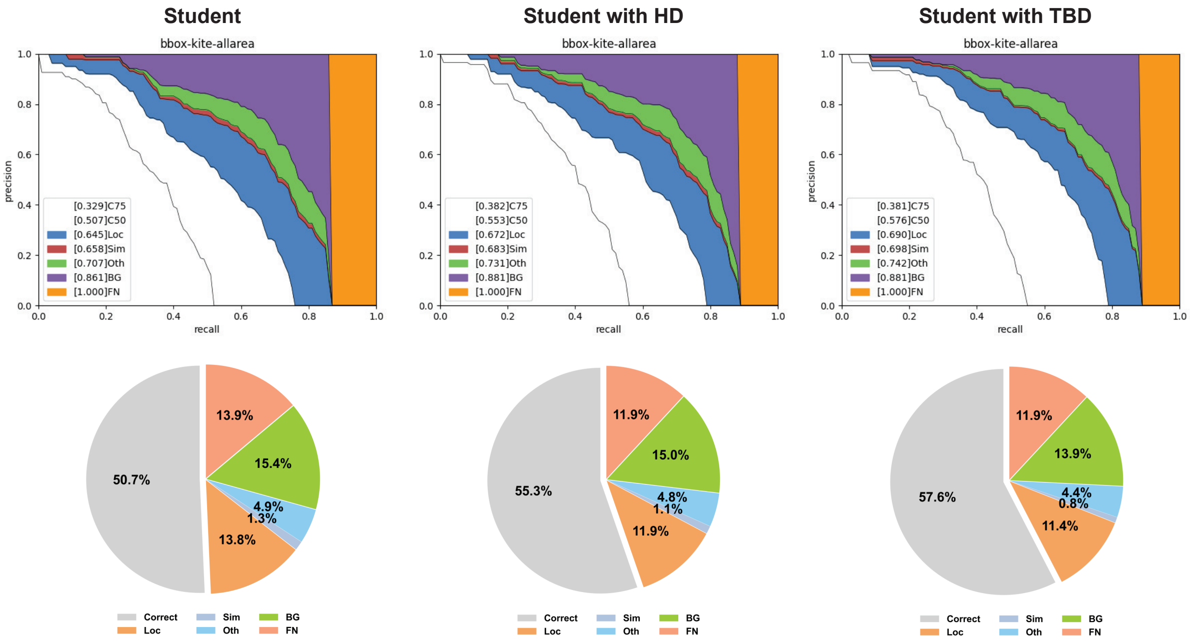


Figure 3. PR curves and error analyses among diverse models. 'Correct': predictions with correct label and $IOU > 0.5$. 'Loc': predictions with correct label but $0.1 < IOU < 0.5$. 'Sim': predictions with an incorrect label but accurate supercategory. 'Oth': predictions with an incorrect label. "BG": false positives predicted on background regions. 'FN': false negatives.

Table 12. The impact of applying different α and β.

α	10.0	7.5	5.0	2.5	β	0.015	0.01	0.005	0.0025
mAP	34.6	34.6	**34.7**	34.6	mAP	34.6	**34.7**	34.5	34.5

promotion (*i.e.*, 34.4 vs. 34.0).

Relationship with other feature-based methods. Similarly, we compare the proposed TFD with FitNet [35] and other prediction-guided feature imitations [20, 51], and the overall results are presented in Table 11. The proposed TBD surpasses the FitNet by a large margin (1.1 AP). Compared with the recent SOTA models such as FRS and PFI, the proposed TFD still shows its superiority. In particular, FRS and PFI only utilize the information of the classification branch to generate the feature mask, thus resulting in suboptimal results (34.0 and 34.2). On the contrary, the proposed TFD combines the superiority of classification and regression tasks and consistently outperforms them by 0.4 and 0.9.

4.3.4 Ablation study of hyper-parameters

Compared with other detection-based KD methods such as FGD [43], the number of hyper-parameters introduced in this paper is significantly reduced (2 vs. 5) so that it is not difficult to prune them. Concretely, the hyper-parameter analysis is conducted based on RetinaNet-R18 student with knowledge distilling from RetinaNet-R50 teacher. The detailed comparison results of various values of α, β are shown in Table 12. Obviously, the performance of the proposed TBD is sightly affected by these two hyper-parameters with only 0.2 mAP fluctuation. Therefore, $\alpha = 5.0, \beta = 0.01$ are set as the default configurations.

Table 13. Comparison results of the proposed TBD and related detection methods. The experiments are constructed on the detectors with ResNet50. Besides, the detectors with ResNeXt101 is served as the teacher model. †† means the inhering strategy [43] is applied to help the convergence of the student model.

Baseline	Method	mAP
Faster RCNN ResNet50	HarmonicDet [40]	39.2
	TBD (ours)	40.3
	TBD†† (ours)	**40.6**
RetinaNet ResNet50	HarmonicDet [40]	37.6
	TBD (ours)	40.0
	TBD†† (ours)	**40.3**

4.3.5 Comparison between TBD and related detection-based methods

As mentioned above, the generation of harmonious predictions can be promoted by a series of methods based on the detector itself, such as devising a customized training strategy [40]. In this part, we compared our TBD with the recent solution named HarmonicDet [40] to show the superiority of the KD-based method. The results are summarized in Table 13. According to the results, the proposed TBD outperforms HarmonicDet by 1.1 and 2.4 on Faster RCNN and RetinaNet, indicating the impressive potential of applying knowledge distillation to alleviate the inherent detection problem. In addition, after using the initialization strategy proposed in [18], the proposed TBD allows for faster convergence and achieves more satisfactory performance.

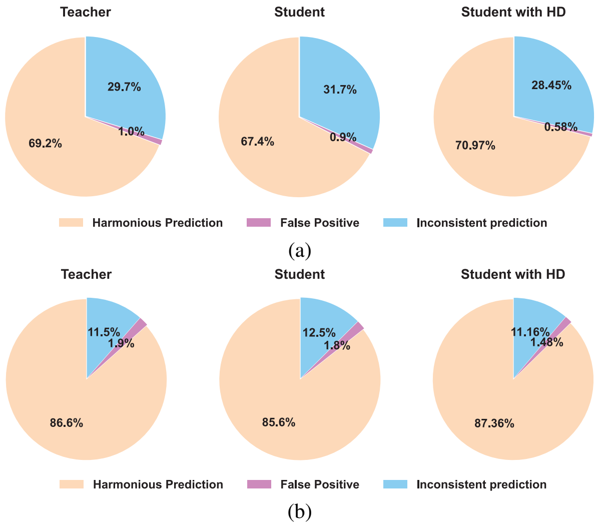
4.4. Analysis and Visualization

4.4.1 Error Analysis

We use the official COCO toolbox [25] to conduct error analysis between RetinaNet-R18 and RetinaNet-R18 with the proposed TBD. Note that the RetinaNet-R50 is chosen as the cumbersome teacher. As presented in Fig. 3, after equipping with the proposed HD, the localization error (Loc) is significantly decreased (*i.e.*, from 13.8 to 11.9). When incorporated with TFD, both the localization and classification errors are further declined. Ultimately, the proportion of correct predictions is boosted from 50.7 to 57.6, verifying the effectiveness of the proposed TBD.

4.4.2 Visualization

Visualization of Harmony Distillation. We compare the proportions of harmonious predictions of easy-classified samples between the original student model and the model with the proposed HD. The corresponding results of the teacher model are also shown here for reference. The overall results of these models are depicted in Fig. 4, where the predictions with classification scores larger than 0.9 and 0.8 are selected, respectively. As can be seen from the left and middle sub-figures of Fig. 4a and 4b, the teacher model is more inclined to generate high-quality predictions (69.2 vs. 67.4, 86.6 vs. 85.6), illustrating that the teacher model has the capacity to transfer knowledge to the lightweight student. Furthermore, the student model's proportions of harmonious predictions tremendously increased from 67.4 to 70.97 and 85.6 to 87.36, even exceeding the teacher model. Besides, we also observe that the amount of False Positives (FPs) is partly reduced. We explain that some FPs with IOU closer to 0.5 can be turned into True Positives (TPs) with the help of HD.



(a)

(b)

Figure 4. The proportions of harmonious and inconsistent predictions between vanilla student and student with proposed HD. The predictions with the score larger than 0.9 and 0.8 are counted in (a) and (b), respectively.

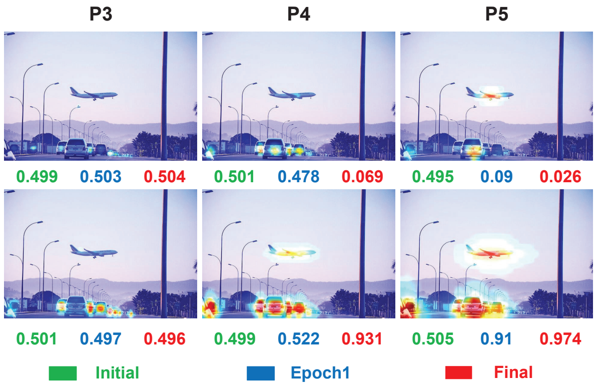


Figure 5. Visualization of the classification-aware and localization-aware masks with the corresponding learned weights. The top row sub-figures are the classification masks, and the bottom are the regression masks.

Visualization of task-ware masks and the learned weights. In this part, we provide several visualization results of the proposed TFD. Concretely, we visualize the task-aware masks and the learned weights in Fig. 5. Two meaningful observations can be discovered. For one thing, the distributions of classification and localization masks are disparate. The classification task concentrates on significant parts of the instance, while the regression task encodes rich information between foreground and background. For another, the learned weights behave diversely at different FPN levels and training stages. The task-aware weights tend to be evenly distributed in the initial phase with random initialization. Owing to the proposed TWG, these weights are

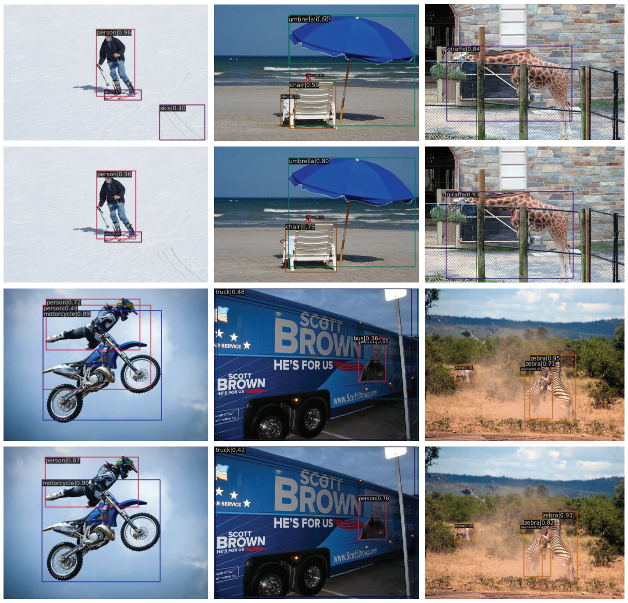


Figure 6. Qualitative comparisons between vanilla RetinaNet-R50 (the top sub-figures) and RetinaNet-R50 with the proposed TBD (the bottom sub-figures).

rapidly modulated by the teacher's predictions and the student's current learning state.

Visualization of detection results. Qualitative comparisons between the vanilla student and student with TBD are demonstrated in Fig. 6. Compared with the vanilla student, the proposed TBD achieves more credible predictions, such as accurate bounding boxes and fewer duplicates, indicating the effectiveness of our method.

5. Conclusion

This paper thoroughly investigates the impact of the inharmonious distributions between classification and regression tasks on distilling object detectors. To alleviate this limitation, we propose a novel Task-Balanced Distillation (TBD), composed of Harmony Distillation (HD) and Task-decoupled Distillation (TFD). HD enhances the harmonious predictions for the student by aligning the Harmony Score (HS) between the teacher and student to make the NMS more credible. In addition, TFD dynamically combines the classification-aware and localization-aware regions as the meaningful regions for distilling features. Extensive experiments among various datasets and detectors verify the effectiveness and generalization of the proposed method.

References

[1] 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*, pages 6154–6162, 2018. 1, 3

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 6, 8

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016. 3

[5] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021. 2, 3, 6, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[7] 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. 2, 6, 8

[8] Zhongjie Fan and Qiong Liu. Adaptive region-aware feature enhancement for object detection. *Pattern Recognition*, 124:108437, 2022. 3

[9] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021. 1, 2, 3, 5

[10] Yan Gao, Qimeng Wang, Xu Tang, Haochen Wang, Fei Ding, Jing Li, and Yao Hu. Decoupled iou regression for object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5628–5636, 2021. 1, 3

[11] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2154–2164, June 2021. 3

[12] Shuxuan Guo, Jose M Alvarez, and Mathieu Salzmann. Distilling image classifiers in object detectors. *Advances in Neural Information Processing Systems*, 34, 2021. 6, 7

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[14] 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*, volume 33, pages 3779–3787, 2019. 3

[15] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3, 9

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[17] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018. 1, 3

[18] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 6, 11

[19] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 1

[20] 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. *arXiv preprint arXiv:2112.04840*, 2021. 2, 3, 5, 6, 7, 9, 10

[21] 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. 3

[22] 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. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 1, 3, 6, 7

[23] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017. 3

[24] 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*, pages 2980–2988, 2017. 1, 3, 6, 7

[25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6, 11

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

[27] Shuyu Miao, Shanshan Du, Rui Feng, Yuejie Zhang, Huayu Li, Tianbi Liu, Lin Zheng, and Weiguo Fan. Balanced single-shot object detection using cross-context attention-guided network. *Pattern Recognition*, 122:108258, 2022. 3

[28] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020. 3

[29] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing*, 30:207–219, 2020. 2, 6, 8

[30] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 3

[31] Junran Peng, Haoquan Wang, Shaolong Yue, and Zhaoxiang Zhang. Context-aware co-supervision for accurate object detection. *Pattern Recognition*, 121:108199, 2022. 3

[32] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 7, 8

[33] 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*, pages 779–788, 2016. 3

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 3, 6, 7

[35] 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. 3, 9, 10

[36] Nermin Samet, Samet Hicsonmez, and Emre Akbas. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *European Conference on Computer Vision*, pages 406–423. Springer, 2020. 6, 8

[37] Hu Su, Yonghao He, Rui Jiang, Jiabin Zhang, Wei Zou, and Bin Fan. Dsla: Dynamic smooth label assignment for efficient anchor-free object detection. *Pattern Recognition*, page 108868, 2022. 3

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1, 3

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

[40] Keyang Wang and Lei Zhang. Reconcile prediction consistency for balanced object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3631–3640, 2021. 1, 3, 10, 11

[41] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 2, 3, 5, 6, 7

[42] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, pages 1–10, 2022. 7, 8

[43] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 2, 3, 5, 6, 7, 10

[44] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. 3

[45] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 3

[46] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In *European conference on computer vision*, pages 260–275. Springer, 2020. 7

[47] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 2, 6, 7

[48] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 1

[49] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. *Advances in neural information processing systems*, 32, 2019. 7

[50] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9407–9416, 2022. 6, 7

[51] Du Zhixing, Rui Zhang, Ming Chang, Shaoli Liu, Tianshi Chen, Yunji Chen, et al. Distilling object detectors with feature richness. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 5, 6, 7, 9, 10

[52] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 840–849, 2019. 7