Harmonious Teacher for Cross-domain Object Detection

Jinhong Deng[1] Dongli Xu[2] Wen Li[3*] Lixin Duan[3,4]
[1]University of Electronic Science and Technology of China [2]University of Sydney
[3]Shenzhen Institute for Advanced Study, UESTC [4]Sichuan Provincial People's Hospital, UESTC
{jhdengvision, dongliixu, liwenbnu, lxduan}@gmail.com

Abstract

Self-training approaches recently achieved promising results in cross-domain object detection, where people iteratively generate pseudo labels for unlabeled target domain samples with a model, and select high-confidence samples to refine the model. In this work, we reveal that the consistency of classification and localization predictions are crucial to measure the quality of pseudo labels, and propose a new Harmonious Teacher approach to improve the self-training for cross-domain object detection. In particular, we first propose to enhance the quality of pseudo labels by regularizing the consistency of the classification and localization scores when training the detection model. The consistency losses are defined for both labeled source samples and the unlabeled target samples. Then, we further remold the traditional sample selection method by a sample reweighing strategy based on the consistency of classification and localization scores to improve the ranking of predictions. This allows us to fully exploit all instance predictions from the target domain without abandoning valuable hard examples. Without bells and whistles, our method shows superior performance in various cross-domain scenarios compared with the state-of-the-art baselines, which validates the effectiveness of our Harmonious Teacher. Our codes will be available at https://github.com/kinredon/Harmonious-Teacher.

1. Introduction

Object detection aims to recognize and localize objects in images simultaneously. As one of the fundamental tasks in computer vision, it plays an important role in many downstream vision tasks, including face recognition [33], person re-identification [42], instance segmentation [11], action recognition [8] and so on. With the development of deep convolution neural network (DCNN) [12, 32], we have witnessed a performance breakthrough of object detection in recent years [1, 11, 27, 28, 36]. One important driving force for such an advance is the availability of large-scale annotated training data. However, collecting and annotating those data are often extremely expensive in both time and fund, which even has been a major challenge for many real-world applications, for example, face authentication [39], autonomous driving [4], etc.

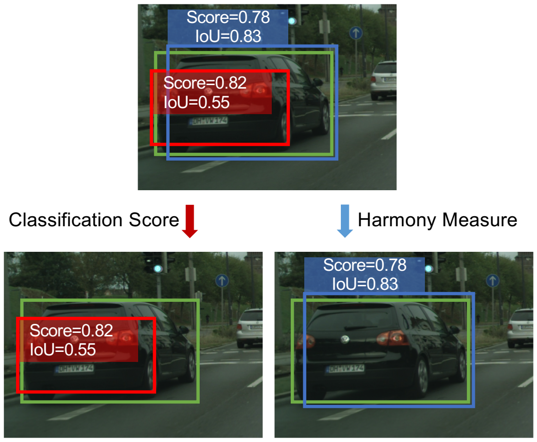

Figure 1. Comparison of pseudo labels selection using classification score and our proposed harmony measure. Object detection models often produce inconsistent predictions, *e.g*., bounding boxes with low classification score but high localization IoU (blue box) or with high classification score but low localization IoU (red box) with ground truth box (green box). Existing self-training methods [6, 13, 22] usually adopt classification scores to rank the predictions and are easily biased to low-quality prediction. In contrast, we use the harmony measure to consider the consistency of the classification and localization scores and prefer the accurate bounding box.

Cross-domain Object Detection (CDOD) [4, 6, 7, 13, 17–19, 22, 30] has been proposed to address this problem, where the goal is to adapt an object detector from a labeled source domain to a novel unlabeled target domain. In this way, great efforts can be saved from annotating training data in the target domain. Recently, researchers have reported that the self-training strategy achieves promising results in the CDOD task. Generally, in the self-training framework, people use an existing object detection model to predict the ob-

[*]The corresponding author

ject category labels and bounding boxes for the target domain images, and select confident predictions as pseudo labels to continuously train the object detection model. These two steps are repeated alternatively for a certain times, and the final model is found to perform quite well in the target domain in many scenarios [6, 13, 22, 26, 43], as the information of the target domain is effectively exploited through training model with pseudo labels.

Nevertheless, existing methods are mainly motivated by self-training classification works, which may have drawbacks for the object detection task. One major issue is that existing methods usually adopt the classification score to select pseudo labels. However, since the classification and localization branches are separately trained, inconsistency between classification and localization scores may happen when predicting the target domain images. For example, the bounding boxes with high classification scores could considerably deviate from the ground-truth position (see example in Fig. 1). Such noise in pseudo labels inevitably introduces bias in the learnt object detection model, leading to degradation in performance. Another issue is the hard thresholding for selecting confident pseudo-labeled instances. Despite how sophisticated it is for determining such a threshold, simply abandoning low-confidence pseudo boxes is definitely undesirable, since valuable hard examples cannot be fully exploited, which is actually crucial for training the object detection model.

In this work, we propose a novel approach called Harmonious Teacher (HT) to improve the self-training framework for the CDOD task. On the one hand, to generate high-quality pseudo labels, we first propose to regularize the consistency of the classification prediction and the localization score when training the detection model. For this purpose, a supervised harmonious loss and an unsupervised harmonious loss are respectively designed for the labeled source domain and the unlabeled target domain. On the other hand, to simultaneously alleviate the damage of low-quality predictions while fully exploiting hard examples, we design a harmony measure to estimate the quality of pseudo-labeled samples based on the consistency of the classification prediction and the localization score. Then, we take all predicted instances into consideration and use the harmony measure to reweigh these instances for self-training, thus avoiding simply abandoning those valuable hard examples.

The contributions of this paper are listed as follows:

- We improve the self-training framework for CDOD and reveal that existing methods neglect the inconsistency between classification and localization, which hinders the performance of self-training.
- We propose a simple yet effective approach named Harmonious Teacher (HT). We first propose harmonious model learning to regularize the consistency of the classification and localization predictions for both source and target domains. Then, we design a harmony measure to estimate the quality of predictions and leverage the harmony measure to reweigh all the predictions in self-training without abounding valuable hard examples.
- We have conducted extensive experiments on four widely used CDOD benchmarks. The experimental results show that our method clearly outperforms the state-of-the-art baselines by a large margin. For example, our method reaches 50.4% mAP on Cityscapes→FoggyCityscapes, which exceeds the state-of-the-art method OADA [43] by 5% mAP.

2. Related Work

Object detection. With the prosperity of deep convolution network [12, 32], the performance of object detection has been greatly improved. Object detection can be divided into two directions: single-stage [23,27,36] and two-stage detectors [1, 28]. The single-stage methods directly regress the positions and categories of objects from an image [27]. The two-stage detectors first generate some candidate regions through a region proposal network (RPN) [28] and then refine these candidates to give the final bounding boxes and categories. Due to the independent task of classification and object detection, the detectors usually present misaligned classification and localization accuracy [9,15,21,36,38,45]. This misalignment hurts the Non-Maximum Suppression (NMS) procedure as the NMS usually utilizes the classification score as the metric to rank the proposals, leading to inaccurate proposal suppression, *e.g.*, the proposals with a high localization IoU but a low classification score will be falsely suppressed. To eliminate this issue, previous works attempt to enforce the predictions to be consistent by metric reformulation [15, 36], harmony learning [21, 45] and consistency regularization [9, 38]. However, these works rely on a large number of labeled data to train the model and suffer from performance degradation when deploying the trained model to a novel domain. In this work, we explore the cross-domain generalization of one-stage detector [36] following [14, 18, 19, 25, 35, 43] for considering the great potential in real-world applications.

Cross-domain object detection. Cross-domain object detection (CDOD) aims to adapt the detector trained on a labeled source domain to an unlabeled target domain. The previous works can be categorized into domain alignment and self-training. As one of the main streams, domain alignment minimizes the domain discrepancy to bridge the domain gap by style transfer [6, 17], adversarial training [4, 14, 30, 43, 49, 50] and graph matching [18, 19, 35], *etc.* Although effective, they are challenging to balance the transferability

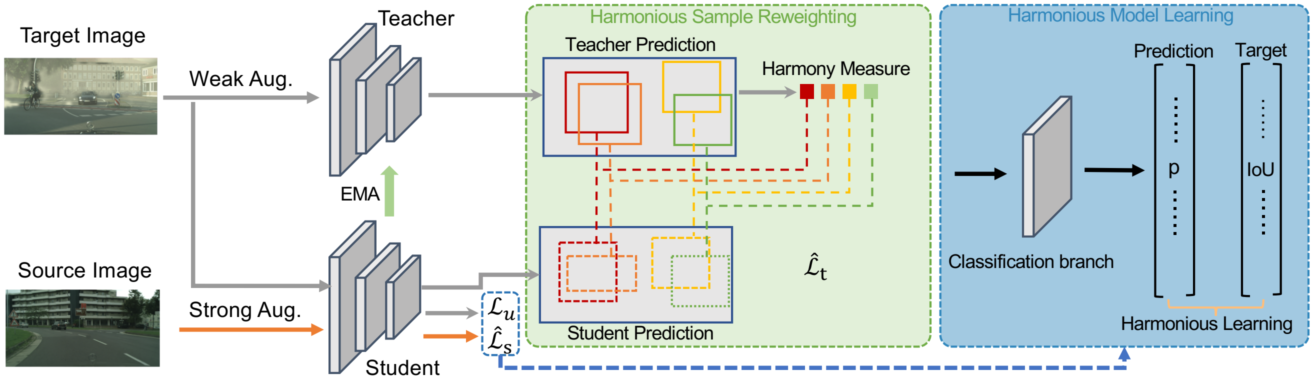

Figure 2. The overview of the proposed Harmonious Teacher framework for CDOD. The source images are fed into the student model and calculate the detection loss $\hat{\mathcal{L}}_s$ with supervised harmonious loss. An unsupervised harmonious loss $\mathcal{L}_u$ is proposed to improve harmony prediction on the target domain. The teacher model adopts the target images with weak augmentation and the outputs of the teacher model are used to formulate harmony measure (HM) to guide the learning of the student by optimizing $\hat{\mathcal{L}}_t$. Our method does not need thresholding for filtering the low-confidence predictions. Instead, we exploit all the predicted instances in the target domain by assigning different weights according to the harmony measure.

and discriminability and do not fully exploit the domain-specific knowledge information in the target domain. However, the self-training methods [3, 6, 13, 22, 25] focus on leveraging the domain-specific information in the target domain and have achieved promising results. UMT [3] feeds the source-like images to the teacher model and generates pseudo labels based on an out-of-distribution detection module. AT [22] combines self-training with adversarial training to improve the quality of pseudo labels. PT [3] develops a probabilistic teacher that leverages uncertainty-guided consistency training to promote classification and localization adaptation. In this work, we reveal that inconsistency between classification and localization will harm self-training approaches. To address this issue, we propose a novel method referred to as Harmonious Teacher (HT) that improves the self-training framework by harmonious model learning and harmonious sample reweighting.

3. Harmonious Teacher

In this section, we present our new self-training approach called Harmonious Teacher (HT). In the following, we first briefly introduce the mean teacher based self-training framework for CDOD in Sec. 3.1, and analyze the problems in Sec. 3.2. Then, we elaborate on the design of our Harmonious Teacher model in Sec. 3.3 and 3.4.

3.1. Self-training Framework in Object Detection

Existing self-training methods [3, 13, 22, 43] generally follow the popular mean teacher (MT) framework [34], where a teacher model is used to produce pseudo labels to supervise the student model. In particular, the teacher and student models are two detection models with identity architecture. The student model is trained with both the labeled source samples and pseudo-labeled target samples, where the labels of the pseudo-labeled target samples are generated by the teacher model. On the other hand, the teacher model is updated by using the exponential moving average (EMA) of the weights of the student model.

Usually, when training the MT models, the labeled source samples are directly fed into the student model to calculate the detection loss by leveraging the ground truth of the source domain. For the target samples, two different types of augmentation are used: weak augmentation and strong augmentation. The teacher model adopts weakly augmented images to generate reliable pseudo labels, which are used to guide the learning of the student model with the input of strongly augmented images. The overall optimization objective can be written as follows:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_t, \tag{1}$$

where $\mathcal{L}_s$ and $\mathcal{L}_t$ are the losses for the source domain and the target domain pseudo-labeled samples, respectively.

In particular, $\mathcal{L}_s$ usually is the object detection loss, which often consists of a classification loss and a localization loss. For example, in FCOS [36], the loss is formulated as follows:

$$\mathcal{L}_s = \mathcal{L}_{cls} + \mathcal{L}_{reg}, \tag{2}$$

where $\mathcal{L}_{cls}$ is the focal loss for classification, and $\mathcal{L}_{reg}$ is the localization loss which contains an IoU loss for boxes regression and a binary cross entropy loss for localization quality estimation branch (*e.g.*, centerness or IoU branch).

Accordingly, $\mathcal{L}_t$ can also be similarly defined with pseudo labels (bounding boxes and their corresponding category labels), *i.e.*,

$$\mathcal{L}_t = \mathcal{L}^t_{cls} + \mathcal{L}^t_{reg}, \tag{3}$$

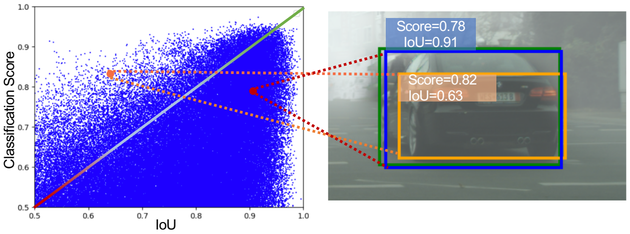

Figure 3. **Left:** The distribution of classification score and IoU with ground truth of predictions from the source model on the validation set of Foggy Cityscapes. **Right**: Illustration of bounding boxes with inconsistent prediction between classification and localization.

where $\mathcal{L}_{cls}^{t}$ and $\mathcal{L}_{reg}^{t}$ are the classification and localization loss using the pseudo labels.

3.2. Analysis on the Self-training Framework

Generating the pseudo labels for target images is the most crucial part for self-training. To generate the pseudo labels for the target domain, existing approaches [6,22] rank the predicted bounding boxes according to their classification scores, and then select high-confidence samples based on a certain threshold for training the student model[1].

However, due to the classification and localization branches being independently trained (see Eq. (2)), the inconsistency between classification and localization scores often occurs in CDOD (see Fig. 1). In other words, the bounding boxes with high classification scores may deviate from the ground truth bounding box, and vice versa. Existing self-training CDOD works usually deploy a simple quality estimation strategy (*e.g.*, classification score) to rank the predictions [6, 22]. While this might be natural for the classification task, due to the inconsistency issue, it may introduce significant noise for the object detection task.

We further illustrate this problem in Fig. 3, where we take the Foggy Cityscapes dataset as an example and plot the distribution of classification score and intersection-of-union (IoU) score with ground truth of the predictions from a pre-trained source model. Inaccurate bounding boxes with high classification scores (points at the upper-left part) are often ranked ahead of good bounding boxes (points at the right part). This will inevitably introduce bias to the model training when taking those inaccurate bounding boxes as pseudo labels.

Moreover, hard examples are also crucial for training object detection models, which are often predicted with low classification scores and fairly good bounding boxes (*i.e.*, the points at the lower-right part of Fig. 3). However, to prevent the model degradation caused by the low-quality instances, existing self-training works [3, 6, 22, 24, 48] often manually set a fixed threshold in a subtle way to select only high-confidence samples. Despite how subtle it is to determine a suitable threshold, hard examples are unfortunately ignored, which could be valuable to the model training.

Our contributions: To address these issues, on the one hand, we first propose to improve the prediction quality by regularizing the consistency of the classification prediction and the localization score when training the detection model. Thus, noise predictions with inconsistent classification and localization scores can be reduced. On the other hand, we design a harmony measure to estimate the quality of predictions and use it to reweigh the pseudo-labeled samples in self-training. In this way, all pseudo-labeled samples can contribute to the model training based on their prediction qualities, and the hard threshold is not needed anymore.

3.3. Harmonious Model Learning

We first consider how to improve the mean teacher model to produce high-quality predictions that the classification score and the localization score are more consistent. In the following, we discuss how to achieve this in both a supervised manner (for the source domain) and an unsupervised manner (for the target domain), respectively.

Supervised Harmonious Loss: Motivated by [21, 45], we enforce the classification branch to predict a score that is consistent with the localization quality. In particular, we leverage the IoU between the predicted box and its ground truth box (referred to as GT-IoU) to represent the localization quality and use it as the learning target of the classification branch. Let us denote by y as the learning target where y is set to GT-IoU for its ground-truth classes, otherwise 0. The loss is formulated as follows:

$$\mathcal{L}_h(y,p) = \begin{cases} -y\left(y\log(p) + (1-y)\log(1-p)\right) & y>0 \\ -\alpha p^{\gamma}\log(1-p) & y=0 \end{cases} \tag{4}$$

where p is the predicted probability from the classification branch. α is a weight coefficient and γ is a focusing parameter. To this end, the supervised harmonious loss that is used to replace the classification loss in Eq. (3) can be written as:

$$\hat{\mathcal{L}}_{cls} = \sum_{i}\sum_{c}\mathcal{L}_h\left(y_{i,c}, p_{i,c}\right), \tag{5}$$

where $y_{i,c}$ and $p_{i,c}$ denote GT-IoU and the predicted classification score for the class c at the location i in the feature map, respectively. As shown in the right part of Fig. 2, the one-hot style target at the ground-truth class label position is replaced by the GT-IoU. In this way, the original classification score evolves to be a consistent prediction that can represent the quality of predictions for both classification and localization.

Unsupervised Harmonious Loss: Although we have enforced the model to give consistent predictions to measure the quality of the model predictions on the source domain,

[1] Some works may firstly exploit Non-Maximum Suppression (NMS) to filter redundant bounding boxes before the ranking process

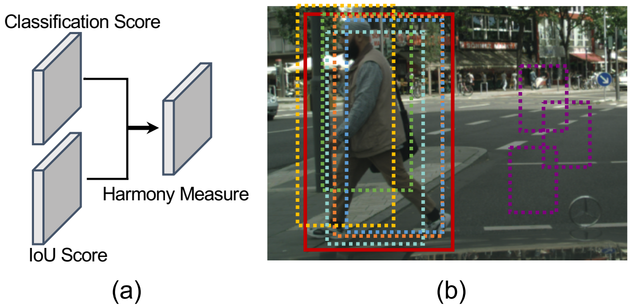

Figure 4. (a) The harmony measure is obtained from the classification branch and the IoU branch, and can be used to evaluate the quality of pseudo labels. (b) Positive boxes often closely surround the ground truth box (solid red line) with a large overlap, while the negative boxes (dotted purple line) generally deviate from each other with only a small overlap.

however, we cannot guarantee this property can generalize well on the target domain because of the domain gap between the source and target domain. We expect the teacher model also predict consistent predictions on the target domain. Because of the lack of annotations, we cannot directly obtain IoUs between the predicted boxes and the ground truth boxes. We propose a heuristic approach to find a proper IoU for the target samples. Specifically, for each box predicted by the student model, we calculate the IoU with other predicted boxes and pick the maximum IoU $\hat{u}$ as a substitute of GT-IoU. We remold the Eq. (5) to formulate the unsupervised harmonious loss:

$$\mathcal{L}_u = \sum_i \sum_c \mathcal{L}_h \left(\hat{u}_{i,c}, q_{i,c} \right), \tag{6}$$

The insight behind this is positive boxes often closely surround the ground truth box with a large overlap, while the negative boxes generally deviate from each other with only a small overlap as shown in Fig. 4 (b). While the negative boxes usually are not grouped together like positive bounding boxes. Thus, the maximum IoU with other boxes can be used as a substitute for GT-IoU. By conducting the unsupervised harmonious model learning, the model can produce more harmonious predictions on the target domain.

3.4. Harmonious Sample Reweighting

After having learned a harmonious model, we are ready to generate pseudo labels for the target domain images. Intuitively, for an instance predicted in the target image, the more consistent between classification score and IoU score is, the more reliable the instance could be as an object. We therefore design a harmony measure to characterize the quality of pseudo labels. Furthermore, with this harmony measure, we are able to fully exploit all the predicted instances in the target domain by assigning different weights according to their harmony measure. Namely, we do not need to manually set a fixed threshold to filter out low-confidence instances as in existing works [3, 6, 22, 24, 48], and the hard examples can be retained to contribute to the model training for improving the detection performance on the target domain.

Harmony Measure: Through harmonious model learning, we enforce the model predictions to be consistent between classification and localization scores. When inferencing the images in the target domain, the deviation between the predictions of classification and localization branches could measure the quality of pseudo labels. To this end, we propose a harmony measure that explicitly encodes the classification prediction and localization score into a unified metric to estimate the quality of pseudo labels as shown in Fig. 4 (a). We define the harmony measure as follows:

$$h = p^{\beta} u^{(1-\beta)}, \tag{7}$$

where p denotes the classification score in harmonious model learning. u is the localization score (*i.e.*, IoU) predicted from the IoU branch. β is a trade-off parameter that balances the contribution between the harmonious model prediction score and localization score to harmony measure. The range of harmony measure is $0-1$. This harmony measure has two essential properties. First, it considers the joint quality from classification and localization branches, $h=1$ is achieved when $p=u=1$. Second, it stands for the harmony between scores from classification and localization branches, where $h=0$ means they are totally adverse qualities for classification and localization.

Harmonious Weighting: With this harmony measure, we are able to fully exploit all the predicted instances in the target domain by assigning different weights to an unsupervised loss for all predictions according to their harmony measure. In other words, we do not need to manually set a fixed threshold to filter low-confidence instances, which may be the hard examples and are useful information for improving self-training performance. Specifically, we apply Quality Focal Loss [21] instead of the cross entropy loss for unsupervised classification loss due to we use the soft labels with a continuous value. It can be formulated as follows:

$$\hat{\mathcal{L}}_{cls}^{t}(\hat{y}, p) = -|\hat{y} - p|^{\eta}((1-\hat{y})\log(1-p) + \hat{y}\log(p)), \tag{8}$$

where η is the suppression factor. p and $\hat{y}$ are class probabilities predicted from the student and teacher models, respectively. In summary, the harmonious weighting (HW) loss can be represented as follows:

$$\hat{\mathcal{L}}_t = \sum_i e^{(1-h_i)} (\sum_c \hat{\mathcal{L}}_{cls}^{t} \left(\hat{y}_{i,c}, p_{i,c} \right) + \mathcal{L}_{reg}^{i}), \tag{9}$$

where h_i is the harmony measure at the position i in the feature map. $\hat{y}_{i,c}$ and $p_{i,c}$ denote the predicted classification scores from the teacher and student models for the class c

Table 1. Quantitative results on adaptation from Cityscapes to Foggy Cityscapes with VGG16 backbone network.

Method	Reference	Detector	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
SWDA [30]	CVPR'19	Faster RCNN	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
CFDA [47]	CVPR'20	Faster RCNN	43.2	37.4	52.1	34.7	34.0	46.9	29.9	30.8	38.6
HTCN [2]	CVPR'20	Faster RCNN	33.2	47.5	47.9	31.6	47.4	40.9	32.3	37.1	39.8
UMT [6]	CVPR'21	Faster RCNN	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.4	41.7
MeGA [37]	CVPR'21	Faster RCNN	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
ICCR-VDD [40]	ICCV'21	Faster RCNN	33.4	44.0	51.7	**33.9**	52.0	34.7	34.2	36.8	40.0
TIA [46]	CVPR'22	Faster RCNN	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
TDD [13]	CVPR'22	Faster RCNN	39.6	47.5	55.7	33.8	47.6	42.1	37.0	41.4	43.1
MGA [49]	CVPR'22	Faster RCNN	45.7	47.5	60.6	31.0	52.9	44.5	29.0	38.0	43.6
PT [3]	ICML'22	Faster RCNN	40.2	48.8	59.7	30.7	51.8	30.6	35.4	44.5	42.7
EPM [14]	ECCV'20	FCOS	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
SCAN [18]	AAAI'22	FCOS	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1
KTNet [35]	ICCV'21	FCOS	46.4	43.2	60.6	25.8	41.2	40.4	30.7	38.8	40.9
SSAL [25]	NeurIPS'21	FCOS	45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
SIGMA [19]	CVPR'22	FCOS	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
OADA [43]	ECCV'22	FCOS	47.8	46.5	62.9	32.1	48.5	**50.9**	34.3	39.8	45.4
HT	-	FCOS	**52.1**	**55.8**	**67.5**	32.7	**55.9**	49.1	**40.1**	**50.3**	**50.4**

at the location i in the feature map, respectively. Our harmonious weighting loss has two advantages. First, it does not introduce any extra hyper-parameters. Second, it fully leverages all the predicted instances to provide richer information to the student model.

Overall Optimization Objective: We illustrate the overall architecture of our Harmonious Teacher in Fig. 2. The model is trained jointly by optimizing all losses in an end-to-end manner. The overall optimization objective of our Harmonious Teacher can be written as follows:

$$\mathcal{L} = \hat{\mathcal{L}}_s + \lambda \mathcal{L}_u + \lambda_1 \hat{\mathcal{L}}_t, \qquad (10)$$

where the $\hat{\mathcal{L}}_s$ is the detection loss where we replace the $\mathcal{L}_{cls}$ in Eq. (2) with supervised harmonious loss $\hat{\mathcal{L}}_{cls}$. $\mathcal{L}_u$ is unsupervised harmonious loss which enhances the consistent predictions on the target domain. $\hat{\mathcal{L}}_t$ is the harmonious weighting loss on the unlabeled target domain to make all the predictions that can reasonably contribute to the unsupervised loss. λ and λ_1 are the trade-off parameters.

4. Experiments

4.1. Dataset

We conduct extensive experiments on four widely used adaptation scenarios following the standard CDOD setting in previous works [3, 4, 6].

Cityscapes→Foggy Cityscapes: Cityscapes [5] is collected from street scenes in different cities and captured with an on-board camera under a clean weather condition. It contains a train set ($2,975$ images) and a validation set (500 images) with eight categories of annotated bounding boxes. Following previous works [3, 4, 6], we convert the instance mask to bounding boxes for training object detection model. Foggy Cityscapes [31] is a foggy version of Cityscapes and is rendered from Cityscapes by using the depth information to synthesize foggy weather. It utilizes the same annotations as Cityscapes. In this scenario, we explore the domain discrepancy under inverse weather adaptation. Note that we choose the worst foggy level (*i.e.*, 0.02) from the Foggy Cityscapes as the target domain.

Cityscapes→BDD100K: BDD100K [44] is a large-scale driving dataset consisting of 100k images. Following [20, 41], we extract the daytime subset of BDD100K as the target domain, resulting in $36,728$ training images and $5,258$ validation images. The Cityscapes and BDD100K have different scene layouts and share seven categories.

Sim10K→Cityscapes: Sim10K [16] is a synthesized dataset based on the computer game Grand Theft Auto V (GTA V) yielding the domain gap with the real-world scene (*i.e.*, Cityscapes). It consists of 10k images with $58,071$ bounding boxes of the car. We use all the images of Sim10K as the source domain and report the performance of the car category following the existing works [14, 19, 43].

KITTI→Cityscapes: KITTI [10] is collected from a different camera (vehicle-mounted) with Cityscapes (on-board camera), resulting in cross-camera domain shift. This dataset includes $7,481$ labeled images with the car category. We use all the images of KITTI as the source domain and report the performance of the car category.

4.2. Implemental Details

Following [14, 18, 19], we take FCOS [36] with the IoU branch as the base detector for experiments. The VGG16 [32] pre-trained on ImageNet [29] is adopted as the backbone. We use batch-size 8 for both source and target images and 4 RTX 3090 GPUs. The learning rate is set to 0.005 without any decay. We pre-train the model 10k iterations with labeled source samples to initialize the teacher and student models and continuously train the model 30k iterations for both the source and target domain. The shorter side of the images is resized at most 800 following [14].

Table 2. Quantitative results on adaptation from Cityscapes to BDD100K with the VGG16 backbone network.

Methods	Reference	Detector	person	rider	car	truck	bus	mcycle	bicycle	mAP
DA-Faster [4]	CVPR'18	Faster RCNN	28.9	27.4	44.2	19.1	18.0	14.2	22.4	24.9
SWDA [30]	CVPR'19	Faster RCNN	29.5	29.9	44.8	20.2	20.7	15.2	23.1	26.2
SCDA [50]	CVPR'19	Faster RCNN	29.3	29.2	44.4	20.3	19.6	14.8	23.2	25.8
ECR [41]	CVPR'20	Faster RCNN	32.8	29.3	45.8	22.7	20.6	14.9	25.5	27.4
SED [20]	AAAI'21	Faster RCNN	32.4	32.6	50.4	20.6	23.4	18.9	25.0	29.0
TDD [13]	CVPR'22	Faster RCNN	39.6	38.9	53.9	24.1	25.5	24.5	28.8	33.6
PT [3]	ICML'22	Faster RCNN	40.5	39.9	52.7	25.8	33.8	23.0	28.8	34.9
Source Only [36]	-	FCOS	36.9	22.4	49.7	16.1	16.3	13.0	22.1	25.2
EPM [14]	ECCV'20	FCOS	39.6	26.8	55.8	18.8	19.1	14.5	20.1	27.8
SIGMA [19]	CVPR'22	FCOS	46.9	29.6	**64.1**	20.2	23.6	17.9	26.3	32.7
HT (Ours)	-	FCOS	**53.4**	**40.4**	63.5	**27.4**	30.6	**28.2**	**38.0**	**40.2**

Table 3. Quantitative results on adaptation from Sim10K to Cityscapes with the VGG16 backbone network.

Method	Reference	Detector	AP of car
SCDA [50]	CVPR'19	Faster RCNN	43.0
HTCN [2]	CVPR'20	Faster RCNN	42.5
UMT [6]	CVPR'21	Faster RCNN	43.1
SED [20]	AAAI'21	Faster RCNN	42.5
TDD [13]	CVPR'22	Faster RCNN	53.4
MGA [49]	CVPR'22	Faster RCNN	54.6
PT [3]	ICML'22	Faster RCNN	55.1
Source Only	-	FCOS	39.8
EPM [14]	ECCV'20	FCOS	49.0
KTNet [35]	ICCV'21	FCOS	50.7
SSAL [25]	NeurIPS'21	FCOS	51.8
SCAN [18]	AAAI'22	FCOS	52.6
SIGMA [19]	CVPR'22	FCOS	53.7
OADA [43]	ECCV'22	FCOS	59.2
Ours	-	FCOS	**65.5**

Table 4. Quantitative results on adaptation from KITTI to Cityscapes with the VGG16 backbone network.

Method	Reference	Detector	AP of car
SCDA [50]	CVPR'19	Faster RCNN	42.5
HTCN [2]	CVPR'20	Faster RCNN	42.1
SED [20]	AAAI'21	Faster RCNN	43.7
TDD [13]	CVPR'22	Faster RCNN	47.4
MGA [49]	CVPR'22	Faster RCNN	48.5
PT [3]	ICML'22	Faster RCNN	60.2
Source Only	-	FCOS	34.4
EPM [14]	ECCV'20	FCOS	43.2
KTNet [35]	ICCV'21	FCOS	45.6
SSAL [25]	NeurIPS'21	FCOS	45.6
SCAN [18]	AAAI'22	FCOS	45.8
SIGMA [19]	CVPR'22	FCOS	45.8
OADA [43]	ECCV'22	FCOS	47.8
Ours	-	FCOS	**60.3**

The data augmentation is the same with [24]. The weights λ and λ_1 are set to 1.0. The parameter β is set to 0.5. For α and γ in Eq. (4) are set to 0.75 and 2 for all the experiments, respectively. And the η in Eq. (8) is set to 2.0. For the ablation studies that need to select pseudo labels, we set the selective ratio $\rho\%$ to 1% following [48]. We provide more implemental details in Supplementary.

4.3. Comparison with State-of-the-arts

Cityscapes→Foggy Cityscapes. The adaptation results of Cityscapes to Foggy Cityscapes are shown in Table 1. The proposed HT significantly outperforms all state-of-the-art works by an absolute margin of 6.8% mAP for two-stage Faster RCNN detector MGA [49] and 5.0% mAP for one-stage FCOS detector OADA [43], respectively. The compelling results clearly demonstrate that HT can provide accurate guidance to the student model by harmonious model learning and harmonious sample weighting.

Cityscapes→BDD100K. Table 2 shows the results of adaptation from Cityscapes to BDD100K. Our HT achieves 40.2% mAP, outperforming all the baselines by a notable margin of 5.3% mAP over top-performing two-stage adaptive detector PT [3] and 7.5% mAP over one-stage CDOD approach SIGMA [19]. This clearly verifies the robustness of HT in different cross-domain scenarios.

Sim10K→Cityscapes. Collecting and labeling a large-scale dataset is a challenge for object detection. An alternative way is to capture the data from the synthetic platform. However, there is a domain discrepancy between the synthetic and real samples. We study the synthetic to real adaptation scenario, and the results of Sim10K to Cityscapes are shown in Table 3. Our HT achieves 65.5% mAP, which exceeds all the other works by a large margin. This further verifies the effectiveness of our methods for improving self-training methods by considering the inconsistency issue between classification and localization.

KITTI→Cityscapes. The cross-camera scenario widely exists in real-world applications. The results of adaptation from KITTI to Cityscapes are shown in Table 4. We also achieve state-of-the-art results over the compared baselines. This again demonstrates the effectiveness of our method.

4.4. Further Empirical Analysis

Ablation Studies: We conduct ablation studies by adding each component of our method. The results are shown in Table 5. First, we can observe the effectiveness of harmonious model learning. In particular, SHL improves 1.8% mAP compared with the baseline that utilizes the original

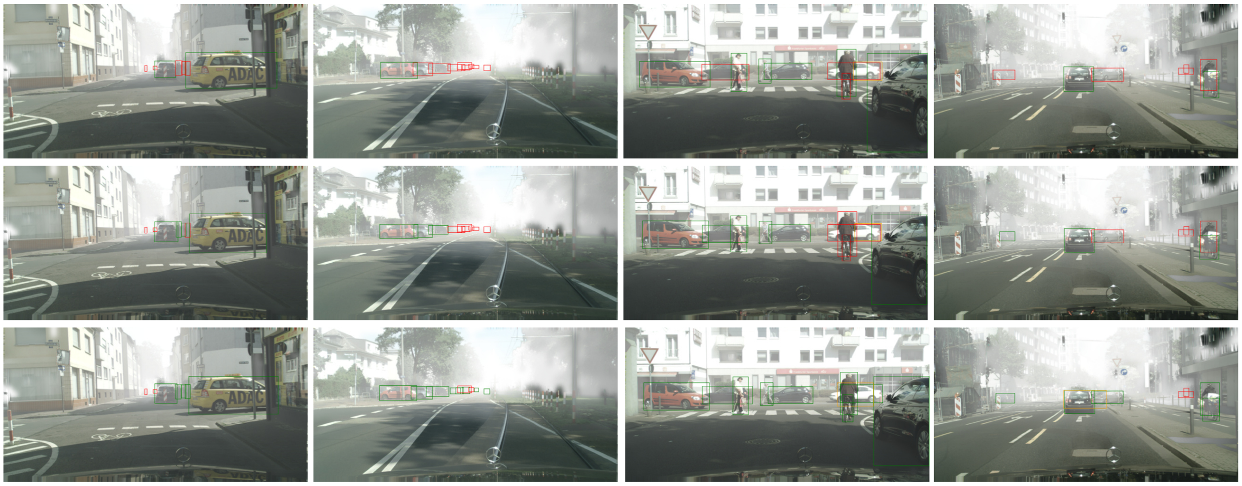

Figure 5. Qualitative results on the target domain of Cityscapes to Foggy Cityscapes for Source Only [36] (top row), SIGMA [19] (middle row) and Ours (bottom row). Green, red and orange boxes indicate true positive (TP), false negative (FN) and false positive (FP), respectively. We set the score threshold to 0.7 for better visualization. Best appreciated when viewed in color and zoomed up.

Table 5. Ablation studies of HT on Cityscapes → FoggyCityscapes. SHL and UHL denote supervised harmonious and unsupervised harmonious loss. HM-Rank indicates that we use the HM to select pseudo labels. HW is the harmonious weighting.

Method	SHL	UHL	HM-Rank	HW	mAP (%)
Baseline	-	-	-	-	37.3
Proposed	✓				39.1
	✓	✓			40.5
	✓	✓	✓		45.2
	✓	✓		✓	50.4

Table 6. The effect of β on Cityscapes → FoggyCityscapes.

β	0	0.25	0.5	0.75	1.0
mAP	48.7	49.0	50.4	50.0	49.8

classification score to select pseudo labels. And the UHL further improve the mAP to 40.5%. These improvements support the claim that harmonious predictions better evaluate the quality of pseudo labels than the classification score. We also show the effect of the harmony measure. The harmony measure can be used in two manners. One is to provide a more accurate ranking of predictions to help select the pseudo labels. We can observe that using the harmony measure to rank the candidate pseudo labels improves the performance by 4.9% in terms of mAP. On the other hand, we integrate the harmony measure into the unsupervised optimization objective to fully exploit all the predicted instances by harmonious sample reweighting. The HW achieves 50.4%, which is a large improvement compared with HM-Rank. This shows that our threshold-free approach has significant advantages and provides effective supervision in the target domain.

Qualitative Results: Fig. 5 illustrates the qualitative results of our method on adaptation scenarios of Cityscapes to Foggy Cityscapes with different approaches. The proposed HT produces more accurate predictions than both Source Only [36] and the state-of-the-art SIGMA [19] model, which shows that HT can significantly improve the detection ability on the target domain, *i.e*., reducing the false negative (FN) and false positive (FP) and detecting more true positive (TP) objects.

The Effect of Balance Factor β: Table 6 shows the results of different β in Eq. (7) that balances the contribution of harmonious prediction and localization IoU score on Cityscapes →Foggy Cityscapes. The results show that only relying on the prediction from the classification or localization branches will hurt the performance. Moreover, the $\beta = 0.5$ achieve the best results.

5. Conclusion

In this work, we study cross-domain object detection (CDOD), which aims to adapt a source detector from a labeled source domain to an unlabeled target domain and reveal that previous self-training methods overlook the inconsistency between classification and localization. Therefore, we propose Harmonious Teacher to improve the self-training for CDOD. We first propose a harmonious model learning for both labeled source and unlabeled target domains to make the model predicts consistent predictions. Then, we design a harmony measure to evaluate the quality of pseudo labels and use it to reweigh the unsupervised loss to alleviate the damage of low-quality predictions and exploit hard examples. The results demonstrate the effectiveness of our method.

Acknowledgement: This work is supported by the Major Project for New Generation of AI under Grant No. 2018AAA0100400, the National Natural Science Foundation of China (No. 62176047, 82121003), Shenzhen Fundamental Research Program (No. JCYJ20220530164812027), Sichuan Science and Technology Program (No. 2021YFS0374, 2022YFS0600), and the Fundamental Research Funds for the Central Universities Grant (No. ZYGX2021YGLH208).

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 1, 2

[2] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8869–8878, 2020. 6, 7

[3] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *ICML*, pages 3040–3055. PMLR, 2022. 3, 4, 5, 6, 7

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2, 6, 7

[5] 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 *CVPR*, pages 3213–3223, 2016. 6

[6] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 1, 2, 3, 4, 5, 6, 7

[7] Jinhong Deng, Xiaoyue Zhang, Wen Li, and Lixin Duan. Cross-domain detection transformer based on spatial-aware and semantic-aware token alignment. *arXiv preprint arXiv:2206.00222*, 2022. 1

[8] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, pages 2969–2978, 2022. 1

[9] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, pages 3490–3499. IEEE Computer Society, 2021. 2

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 6

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2

[13] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *CVPR*, pages 9570–9580, 2022. 1, 2, 3, 6, 7

[14] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748. Springer, 2020. 2, 6, 7

[15] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018. 2

[16] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*, pages 746–753. IEEE, 2017. 6

[17] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, pages 12456–12465, 2019. 1, 2

[18] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, 2022. 1, 2, 6, 7

[19] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 1, 2, 6, 7, 8

[20] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *AAAI*, pages 8474–8481, 2021. 6, 7

[21] 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. *NeurIPS*, 33:21002–21012, 2020. 2, 4, 5

[22] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022. 1, 2, 3, 4, 5

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 2

[24] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. 4, 5, 7

[25] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. *NeurIPS*, 34, 2021. 2, 3, 6, 7

[26] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *ICCV*, pages 3570–3579, 2021. 2

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

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 1, 2

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

[30] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 2, 6, 7

[31] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 6

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 6

[33] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 1

[34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 3

[35] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *ICCV*, pages 9133–9142, 2021. 2, 6, 7

[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 1, 2, 3, 6, 7, 8

[37] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, pages 4516–4526, 2021. 6

[38] Keyang Wang and Lei Zhang. Reconcile prediction consistency for balanced object detection. In *ICCV*, pages 3631–3640, 2021. 2

[39] Ge Wen, Huaguan Chen, Deng Cai, and Xiaofei He. Improving face recognition with domain adaptation. *Neurocomputing*, 287:45–51, 2018. 1

[40] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, pages 9342–9351, 2021. 6

[41] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11724–11733, 2020. 6, 7

[42] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *T-PAMI*, 2021. 1

[43] Jayeon Yoo, Inseop Chung, and Nojun Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *ECCV*, 2022. 2, 3, 6, 7

[44] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, June 2020. 6

[45] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, pages 8514–8523, 2021. 2, 4

[46] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, 2022. 6

[47] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13766–13775, 2020. 6

[48] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *ECCV*, 2022. 4, 5, 7

[49] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *CVPR*, pages 9581–9590, 2022. 2, 6, 7

[50] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 2, 7