Periodically Exchange Teacher-Student for Source-Free Object Detection

Qipeng Liu, Luojun Lin[*], Zhifeng Shen, Zhifeng Yang
College of Computer and Data Science, Fuzhou University
lqpwiki@gmail.com, linluojun2009@126.com, {shen_zhifeng, yzf2001}@outlook.com

Abstract

Source-free object detection (SFOD) aims to adapt the source detector to unlabeled target domain data in the absence of source domain data. Most SFOD methods follow the same self-training paradigm using mean-teacher (MT) framework where the student model is guided by only one single teacher model. However, such paradigm can easily fall into a training instability problem that when the teacher model collapses uncontrollably due to the domain shift, the student model also suffers drastic performance degradation. To address this issue, we propose the Periodically Exchange Teacher-Student (PETS) method, a simple yet novel approach that introduces a multiple-teacher framework consisting of a static teacher, a dynamic teacher, and a student model. During the training phase, we periodically exchange the weights between the static teacher and the student model. Then, we update the dynamic teacher using the moving average of the student model that has already been exchanged by the static teacher. In this way, the dynamic teacher can integrate knowledge from past periods, effectively reducing error accumulation and enabling a more stable training process within the MT-based framework. Further, we develop a consensus mechanism to merge the predictions of two teacher models to provide higher-quality pseudo labels for student model. Extensive experiments on multiple SFOD benchmarks show that the proposed method achieves state-of-the-art performance compared with other related methods, demonstrating the effectiveness and superiority of our method on SFOD task.

1. Introduction

Object detection has achieved significant progress with rapid development of dataset scale and computation capability [31, 22, 4]. However, these detectors are typically trained under an i.i.d assumption that the train and test data are independently and identically distributed, which does not always hold in real-world due to the existence of domain shift between the train and test data. This can cause significant performance degradation when applying a model well-trained on source domain (train data) to the target domain (test data). Unsupervised domain adaptation (UDA), a recent research hotspot, can resolve this dilemma by enabling the model to adapt effectively to the target domain. This is achieved through joint training, leveraging both labeled source domain data and unlabeled target domain data to enhance the model's performance in the target domain.

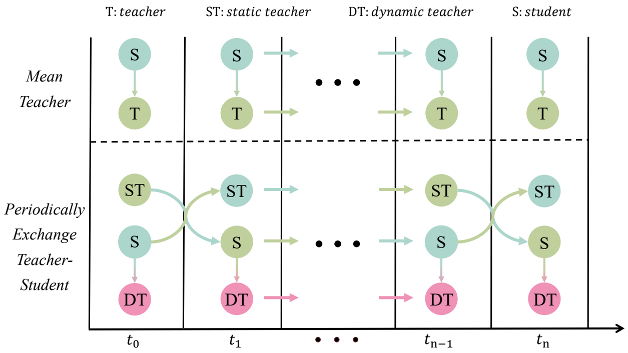
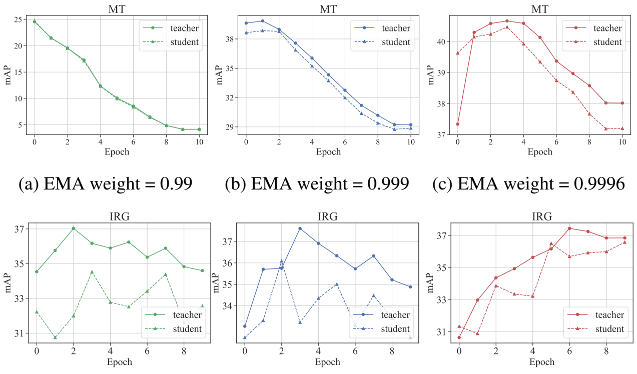


Figure 1: The training paradigms of the *mean-teacher* and the proposed *periodically exchange teacher-student* method. **T** and **S** denote the teacher model and student model, respectively. **ST** represents the static teacher with fixed weights in each period, and **DT** is the dynamic teacher updated by the EMA of the student models. t_i represents the i-th period in whole training stage.

There are many UDA methods developed to address domain shift in image classification tasks [27, 35, 37, 46]. However, these methods cannot meet the growing demand for data privacy protection. Moreover, directly applying these UDA methods to object detection tasks cannot achieve satisfactory performance. In light of the above considerations, source-free object detection (SFOD) has rapidly emerged as an urgent task to attract the attention of researchers. The purpose of SFOD is to achieve effective adaptation of a detector, originally trained on a labeled source domain, to the unlabeled target domain, without accessing any source data during adaptation. Compared with source-free image classification, SFOD is a more challenging task that not only requires regression, i.e., locating the

[*] Corresponding author

bounding box of each object, but also involves classification, i.e., identifying the associated class of each object in diversely-scaled images.



(a) EMA weight = 0.99 (b) EMA weight = 0.999 (c) EMA weight = 0.9996

(d) EMA stepsize = 600 (e) EMA stepsize = 1500 (f) EMA stepsize = 3000

Figure 2: The training curves of different SFOD methods (i.e. conventional MT and IRG [39]) with different EMA hyper-parameters on C2F benchmark [33]. These methods show a consistent phenomenon: when the performance of the teacher model crashes, the student model always follows the downward trend of the teacher model even with different EMA weights or stepsizes.

Most of the existing SFOD studies [24, 43, 25, 8, 40] are based on self-training paradigm using a mean-teacher (MT) framework [36] along with other improved UDA techniques. These MT-based methods involves using a single teacher model to guide the student model, where the teacher model is an exponential moving average (EMA) of the student model at different time steps, and the student model is updated based on the pseudo labels provided by the teacher model. The MT framework assumes that *the teacher model can be improved continuously as the training progresses, and the student model can gradually approach the performance of the teacher model*. However, since the source-pretrained model introduces inherent biases when applied to the target domain, the teacher model, as an EMA of the student model inherited from the source-pretrained model, is susceptible to accumulating errors from the student model. This error accumulation leads to a concerning issue of *training instability* for the teacher model, thereby making the initial assumption no longer holds true. That is, when the single teacher model makes mistakes, the student model tends to replicate the errors without any correction measures. It finally leads to uncontrollable degradation of the detection performance for MT-based SFOD methods.

In order to mitigate the training instability problem, a natural solution involves adjusting the EMA hyper-parameters to encourage a more gradual and stable evolution of the teacher model. For example, the recent works [39, 8] have explored the strategy of employing a larger EMA update stepsize, with the aim of slowing down the updating process of the teacher model. Another line of exploration in this direction involves assigning a higher EMA weight to the historical teacher model, amplifying the influence of the past iterations and consequently reducing the updating rate of the teacher model. However, these efforts have yielded limited success. As shown in Figure 2, the efforts to enhance the EMA weights or increase the EMA update stepsize do not completely resolve the issue of training instability problem within the MT-based frameworks. Besides, it is inconvenient to search for an optimal EMA hyper-parameter to properly update the teacher model.

In this paper, we aim to address the instability problem and thus propose a simple yet novel *Periodically Exchange Teacher-Student (PETS)* method to improve the self-training paradigm of the MT framework. As shown in Figure 1, our method is a multiple-teacher framework consisting of a static teacher model, a dynamic teacher model, and a student model. Unlike the previous methods that keep the roles of student and teacher unchanged throughout the training, we periodically exchange the positions between the student model and the static teacher model. Then, the static teacher model freezes its weights until the next exchanging period; while the student model is trained using the supervision signals provided by the two teacher models, and the dynamic teacher model is updated by an EMA of the student per iteration within each period. In this way, the dynamic teacher implicitly reduces error accumulation to improve its performance. Moreover, the exchange between the static teacher and student helps to prevent a rapid decrease in the lower bound of the student model, ultimately improving the robustness of whole models in our method. Besides, we also propose a consensus mechanism to merge the predictions from the static and dynamic teachers, which can provide higher-quality pseudo labels to supervise the student model.

Our method is evaluated on four SFOD benchmarks. The experimental results show that our method achieves competitive results compared with existing SFOD methods, and demonstrate its effectiveness to solve the instability problem of current MT-based frameworks. The main contributions of our method are summarized as follows:

- We highlight the *training instability* issue within the MT framework, where the errors from the teacher model can be replicated by the student model without correction measures. This will result in an uncontrollable degradation of detection performance in MT-based SFOD methods.

- We propose a simple yet novel *Periodically Exchange Teacher-Student (PETS)* method to address the training instability issue for MT framework. Our method consists of a static teacher, a dynamic teacher and a student model. At the end of each period of training, we exchange the weights between the student and the static teacher to reduce error accumulation. Within each period, we train the student model through the two teacher models, and update the dynamic teacher with an EMA of the student model per iteration.

- We design a consensus mechanism to integrate the predictions from the static teacher and the dynamic teacher models. It integrates knowledge from historical iterations to prevent catastrophic forgetting, which can achieve higher-quality pseudo labels to supervise the student model.
- Extensive experiments on multiple SFOD benchmarks show that the proposed method achieves state-of-the-art performance compared with other related methods, demonstrating the effectiveness and superiority of our method on SFOD task.

2. Ralated Works

2.1. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge from a source domain with labeled data to a target domain without labeled data. The current UDA methods can be roughly categorized into three types: domain translation, adversarial learning and pseudo labeling. The domain translation methods aim to transform a target image into a source-like image by using statistic information in the model [45, 15] or employing a translation network [20, 10, 41]. Adversarial learning is also frequently adopted in UDA tasks by employing a domain discriminator [11] or designing adversarial loss functions, in order to narrow the gap between source and target domains in feature space [3, 37, 23, 29]. Unlike previous methods, pseudo labeling, as one of the most popular self-training paradigms [6], has been an effective approach for UDA, which is mainly constructed based on the mean-teacher (MT) framework [36] that exploits the pseudo labels provided by the teacher model to supervise the student model. Most pseudo labeling methods concentrate on designing interaction manners between the student and teacher models [19, 5, 51]. In this paper, we concentrate on source-free object detection and try to improve self-training paradigm for MT-based SFOD framework.

2.2. Source-Free Object Detection

Several UDA approaches have been applied to *Unsupervised Domain Adaptive Object Detection (UDAOD)*, which can also be categorized into adversarial learning [7, 32, 13], domain translation [16, 20] and pseudo labeling [1, 10]. Given that these methods have been introduced briefly in previous section, we only discuss the final one since our work is constructed on the basis of self-training. To obtain more accurate pseudo labels, UMT [10] transforms target domain data into source-like data in order to improve the quality of generated pseudo-labels. SimROD [30] enhances the teacher model by augmenting its capacity for generating higher-quality pseudo boxes.

With the urgent need for data privacy protection, *Source-Free Object Detection (SFOD)* has emerged as a new branch of UDAOD in recent years. Due to the complexity of the object detection task (numerous regions, multi-scale features, and complex network structure) and the challenge of the absent source data, simply applying the existing UDA-Classification or UDAOD methods to SFOD tasks can not get satisfied results [48, 26]. Therefore, SFOD [25] develops a novel framework that uses self-entropy descent to select high-quality pseudo labels for self-training. SOAP [43] devises domain perturbation on the target data to help the model learn domain-invariant features that are invariant to the perturbations. LODS [24] proposes a style enhancement module and graph alignment constraint to help the model learn domain-independent features. A^{2}SFOD [8] divides target images into source-similar and source-dissimilar images and then adopts adversarial alignment between the teacher and student models. IRG [39] designs an instance relation graph network combined with contrastive loss to guide the contrastive representation learning. While the majority of these approaches rely on the MT framework [28], they tend to overlook the issue of training instability arising from a single teacher model. This oversight allows errors to be replicated by the student model, consequently constraining its performance. To tackle this concern, we propose a *Periodically Exchange Teacher-Student* approach that leverages knowledge from historical models to prevent catastrophic forgetting for MT framework.

3. Preliminary

Let $\mathcal{D}_S = (\mathcal{X}_S, \mathcal{Y}_S)$ represent the labeled data in the source domain, and $\mathcal{D}_T = (\mathcal{X}_T)$ denote the unlabeled data in the target domain, where $\mathcal{X}_S = \{x_s^i\}_{i=1}^{N_S}$ represents the image set of the source domain, $\mathcal{Y}_S = \{y_s^i\}_{i=1}^{N_S}$ represents the corresponding label set containing object locations and category assignments for each image, and $\mathcal{X}_T = \{x_t^i\}_{i=1}^{N_T}$ denotes the image set of the unlabeled target domain. N_s and N_t correspond to the number of labeled source data and unlabeled target data, respectively.

In the setting of SFOD task, a source pre-trained model, denoted as $f_S : \mathcal{X}_S \rightarrow \mathcal{Y}_S$, is initially available to perform adaptation on unlabeled target domain. However, due to the inherent domain gap between the source and target domains, the mapping f_S diminishes performance when directly applied to the target domain. Consequently, the primary objective of SFOD is to acquire a new mapping $f_T : \mathcal{X}_T \rightarrow \mathcal{Y}_T$ by leveraging the source-pretrained model f_S in conjunction with the unlabeled target data $\mathcal{X}_T$ without accessing any source data.

Most previous SFOD methods use Faster-RCNN [31] as

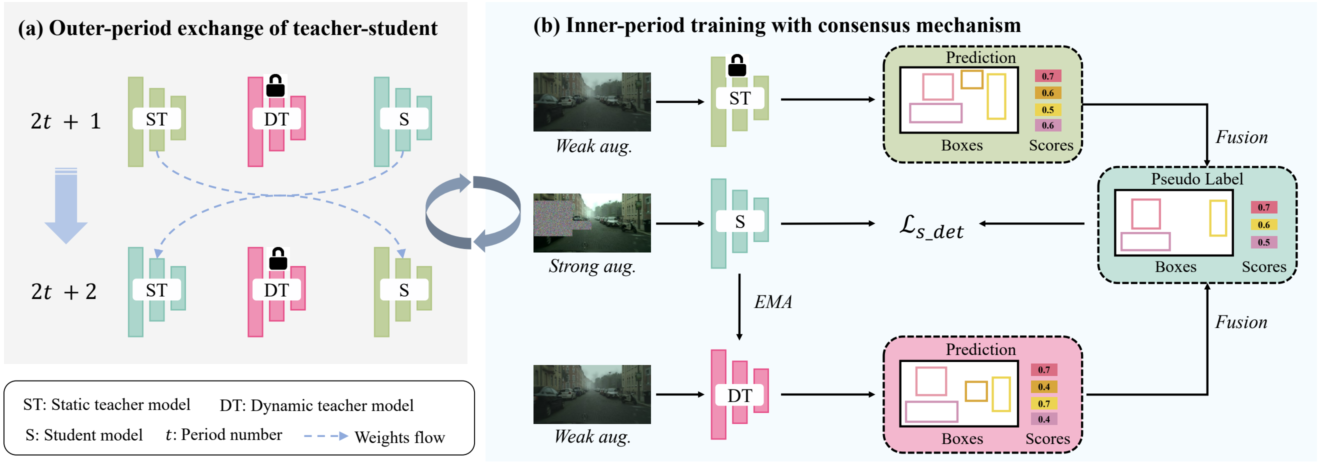


Figure 3: The training pipeline of the proposed *Periodically Exchange Teacher-Student* method, which can be divided into two parts: (a) Outer-period exchange of teacher-student: exchange the weights between the student and static teacher after each period; (b) Inner-period training with consensus mechanism: update the dynamic teacher with an EMA of the student model, and train the student model with a consensus mechanism that fusions the predictions from multiple teachers.

their backbone network. To ensure a fair comparison with previous methods, we also adopt Faster-RCNN as the backbone network here. Therefore, the training goal of f_T is similar to Faster-RCNN, which can be written as:

$$\mathcal{L}_{det} = \mathcal{L}_{cls}^{RPN} + \mathcal{L}_{reg}^{RPN} + \mathcal{L}_{cls}^{ROI} + \mathcal{L}_{reg}^{ROI}, \qquad (1)$$

where $\mathcal{L}_{cls}^{RPN}$ and $\mathcal{L}_{reg}^{RPN}$ represent the losses of foreground prediction and box location from the RPN network, respectively. $\mathcal{L}_{cls}^{ROI}$ and $\mathcal{L}_{reg}^{ROI}$ are the losses of category prediction and box location from the ROI head, respectively.

4. Methodology

4.1. Overview

Our method involves using a static teacher model, a dynamic teacher model, and a student model. The pseudo code of training process can be seen in Algorithm 1. Figure 3 shows the training pipeline of our method, which can be divided into two parts:

1) Outer-period exchange of teacher-student: After each period of training, we exchange the weights between the student model and the static teacher model. In other words, the static teacher and the student reverse their roles per period, as shown in Figure 3(a). Note that the term "period" is synonymous with the concept of an epoch during training.

2) Inner-period training with consensus mechanism: The weights of static teacher model are fixed within each period. The dynamic teacher is updated by the EMA of the student model in each iteration, and the student model is supervised by the pseudo labels merged from the dynamic and static teacher models with consensus mechanism, as shown in Figure 3(b).

Algorithm 1 Python-like code of training process

```
# Outer-period exchange of teacher-student
if epoch % time_period == 0:
   exchange_weight(student, static_teacher)

# Inner-period training with consensus mechanism
for _, images in enumerate(loader):
   # images: [N, C, H, W]
   # N: number of images per mini-batch
   # pre-process images by data augmentation
   img_w = weak_aug(images)
   img_s = strong_aug(img_w)

   # obtain predictions
   pred_s = student(img_s)
   pred_st = static_teacher(img_w)
   pred_dt = dynamic_teacher(img_w)

   # produce pseudo label
   pseudo_labels = consensus(pred_st, pred_dt)

   # compute detection loss
   loss = compute_loss(pred_s, pseudo_labels)

   # update the student by back-propagation
   loss.backward()

   # update the dynamic teacher by EMA
   update_teacher(student, dynamic_teacher)
```

Notations For better understanding our method, we use Θ_S, Θ_{ST} and Θ_{DT} to denote the student model, the static teacher model and the dynamic teacher model, respectively.

4.2. Outer-period Exchange of Teacher-Student

The training process can be divided into multiple independent time periods (i.e., epochs). Each period is represented as t. At the $2t+2$ period, the weights of the student model are swapped by that of the static teacher model at the $2t+1$ period. Conversely, the weights of the static teacher model at the $2t+2$ period are exchanged by that of the student model at previous period. The exchange process can be written as:

$$\Theta_S^{2t+1} \longrightarrow \Theta_{ST}^{2t+2}, \quad \Theta_{ST}^{2t+1} \longrightarrow \Theta_S^{2t+2}, \tag{2}$$

where Θ_{ST}^{2t+2} and Θ_S^{2t+2} denote the static teacher model and the student model at the $2t+2$ period, respectively. This exchange strategy keeps periodically recycling during the whole training process.

The exchange strategy benefits each model from the following perspectives: 1) Student model: The static teacher model serves as a performance lower bound for the student model. If the student model crashes into a collapse issue guided by the declined dynamic teacher, the exchange can ensure that the student model reverts to previous period, effectively mitigating its downward trend. In essence, the exchange helps prevent a rapid decrease in the performance lower bound of the student model, thus improving its robustness. 2) Static teacher model: The exchange strategy ensures periodic updating to the static teacher model's knowledge, which is executed at a notably slow rate to enable a more stable model. 3) Dynamic teacher model: The dynamic teacher model is a temporal ensemble of the student model exchanged by the past student model. In practice, the updating rate of the dynamic teacher model is implicitly reduced. Thus, it has a better ability to resist noise compared to the conventional mean-teacher framework [36]. In summary, our *periodically exchange teacher-student* strategy can enable the student and teacher models to mutually prevent catastrophic forgetting and uncontrollable collapse, thus improving the detection performance.

4.3. Inner-period Training with Consensus Mechanism

During each period, the static teacher maintains fixed weights until iterating to next period. Simultaneously, the dynamic teacher model is updated by the temporal ensembling of the student model, and the student model is updated by pseudo labels as supervision signals, where the pseudo labels are generated by combining the predictions of the dynamic and static teachers through the consensus mechanism. This procedure is illustrated in Figure 3(b). The following sections delve into the details of the consensus mechanism, the learning process of the student model, and the updating of the dynamic teacher model.

4.3.1 Consensus Mechanism

Our framework incorporates two distinct teachers: the static teacher and the dynamic teacher. A notable advantage of our approach is the ability to leverage predictions from both teachers to enhance the quality of pseudo labels. To this end, we design a consensus mechanism that includes two main steps: filtering and fusion.

Filtering Since the output of teacher models contains inevitable noise (low-confidence predictions), we set a category confidence threshold $\delta = 0.5$ to pre-filter low-confidence predictions. This can prevent the subsequent fusion process suffering from the interference of noisy labels.

Fusion For a weakly-augmented target image $x_t \in \mathcal{X}_T$, the predictions of the static teacher and dynamic teacher are represented as $Y_{ST} = \{(b_{ST}^i, c_{ST}^i, y_{ST}^i)\}_{i=0}^n$ and $Y_{DT} = \{(b_{DT}^j, c_{DT}^j, y_{DT}^j)\}_{j=0}^m$, where b, c, y represent the bounding box coordinates, classification confidence and category label of each predicted object, and n, m denote the number of predicted objects of the static teacher and the dynamic teacher, respectively. Then, we select the objects with identical category and a higher intersection over union (IOU) between the predicted boxes of the static teacher and the dynamic teacher. The selection criterion can be represented as $IOU(b_{ST}^i, b_{DT}^j) \geq \eta \;\; \& \;\; y_{ST}^i = y_{DT}^j$, where η is the threshold of judging whether the predicted box belongs to the same object. We usually set $\eta = 0.5$. Lastly, we employ the weighted boxes fusion (WBF) strategy [34] to merge the selected boxes derived from both the static teacher and dynamic teacher models. The process can be formulated as:

$$\begin{aligned} \widetilde{b} &= \frac{1}{C}\Big(\sum_{i=1}^{N} c_{ST}^i * b_{ST}^i + \sum_{j=1}^{M} c_{DT}^j * b_{DT}^j\Big), \\ \widetilde{c} &= \frac{\beta}{N}\sum_{i=1}^{N} c_{ST}^i + \frac{1-\beta}{M}\sum_{j=1}^{M} c_{DT}^j, \end{aligned} \tag{3}$$

where N, M are the number of boxes belonging to the same object predicted by the static teacher and the dynamic teacher, respectively. C is the sum of $\sum_{i=1}^{N} c_{ST}^i$ and $\sum_{j=1}^{M} c_{DT}^j$. β controls the fusion magnitude between the static teacher and dynamic teacher, which is ranged in $[0, 1]$ and set to 0.5 in this paper. We ultimately obtain pseudo label $\widetilde{Y} = \{(\widetilde{b}, \widetilde{c}, \widetilde{y})\}$ for the unlabeled target image x_t, where $\widetilde{b}$ and $\widetilde{c}$ denote the coordinates and confidence of the fused bounding box, respectively, and $\widetilde{y}$ is equivalent to y_{ST}^i. The fused pseudo labels exhibit greater resistance to confirmation bias compared to those single-teacher framework.

4.3.2 Student Learning

Given an unlabeled target image x_t, its pseudo label can be represented as $\widetilde{Y} = \{(\widetilde{b}, \widetilde{y})\}$ that can be used as the supervision signal of the student model. Following Equation 1, the training loss of the student model Θ_S can be defined as:

$$\begin{aligned} \mathcal{L}_{s\_det} = \sum_{\bar{x}_t \in \mathcal{X}_T} & \mathcal{L}_{cls}^{RPN}(\Theta_S(\bar{x}_t), \widetilde{y}) + \mathcal{L}_{reg}^{RPN}(\Theta_S(\bar{x}_t), \widetilde{b}) + \\ & \mathcal{L}_{cls}^{ROI}(\Theta_S(\bar{x}_t), \widetilde{y}) + \mathcal{L}_{reg}^{ROI}(\Theta_S(\bar{x}_t), \widetilde{b}), \end{aligned} \tag{4}$$

where $\bar{x}_t$ denotes the strongly-augmented version of the target image x_t. Since the proposed consensus mechanism can provide more precise bounding boxes compared with previous studies [10], we use both the category prediction loss and box location loss to train the student model.

4.3.3 Dynamic Teacher Updating

Throughout each period, the static teacher model maintains fixed weights across various iterations, whereas the dynamic teacher model adjusts its weights in each iteration. We follow the conventional MT framework that uses the exponential moving average (EMA) strategy to update the dynamic teacher model Θ_{DT}. This can be formulated as:

$$\Theta_{DT} \leftarrow \alpha\Theta'_{DT} + (1-\alpha)\Theta_S, \tag{5}$$

where Θ_{DT} represents the dynamic teacher in current iteration, while Θ'_{DT} pertains to the dynamic teacher in previous iteration. The hyper-parameter α controls the update rate of the dynamic teacher, with a higher value leading to a slower update rate. In this study, we empirically set α to 0.999.

5. Experiments

We conduct comprehensive experiments to evaluate the effectiveness of our method on multiple standard SFOD benchmarks. Then, we perform ablation studies by using different exchange strategies to stress the effectiveness of the proposed periodic exchange strategy. Finally, we analyze the promising results of our method through detailed visualization and component analysis.

5.1. Experimental Setup

Task Settings. Following the existing works [25, 8], we validate our method on the four popular SFOD tasks which represent different types of domain shift, including 1) Cityscapes-to-Foggy-Cityscapes (C2F): Adaptation from normal to foggy weather. 2) Cityscapes-to-BDD100k (C2B): Adaptation from small to large-scale dataset. 3) KITTI-to-Cityscapes-Car (K2C): Adaptation across different cameras. 4) Sim10k-to-Cityscapes-Car (S2C): Adaptation from synthetic to real images. The A-to-B represents the adaption of the model pre-trained on the source domain A to the target domain B.

Datasets. There are five datasets used in the aforementioned tasks: 1) *Cityscapes* [9] is a street view dataset containing 5,000 images with instance-level pixel annotation from different cities in different seasons, where 2,925 training images and 500 validation images are used in the following experiments. 2) *Foggy Cityscapes* [33] is also a street view dataset similar to Cityscapes, but its images are processed by three levels (0.005, 0.01, 0.02) of artificial simulation of extreme foggy scenes. 3) *KITTI* [12] is a widely used benchmark dataset for autonomous driving which contains many images from different real-world street scenes. There are only 7,481 training images used in the experiments. 4) *SIM10k* [18] is a synthetic dataset consisting of 10,000 city scenery images of cars. 5) *BDD100k* [47] is a large-scale open source video dataset for autonomous driving, including 100k images from different times, different weather conditions and driving scenarios.

5.2. Implementation Details

Our method is implemented based on PyTorch platform using detectron2 framework [42]. Following the previous study [25], we use Faster-RCNN [31] with the backbone of VGG16 pre-trained on the ImageNet as the base detection model in our method. All images are scaled by resizing the shorter edge of the image to 600 pixels before training. The data augmentation strategy includes random erasing, random horizontal flip, and color transformation. We adopt the SGD as the optimizer with an initial learning rate of 8e-4, a decay rate of 0.1. The batch size is set to 8.

The training process of our method consists of two stages: warm-up and adaptation. In the warm-up stage, the learning rate increases gradually from 0 to 8e-4. The static teacher model freezes its weights and the dynamic teacher model keeps updating during the first two epochs. In the fine-tuning stage, the weights of the student model and the static teacher model are exchanged per epoch, and the EMA rate of the dynamic teacher model is set to 0.999. During evaluation process, we reserve the dynamic teacher model for inference and choose the mean average precision (mAP) with an IOU threshold of 0.5 as the evaluation measure.

5.3. Comparison with Existing SOTA Methods

UDAOD and SFOD have a similar task setting. Therefore, we compare our method with existing UDAOD and SFOD methods. Table 1-4 show the comparison results, where “Source only” and “Oracle” represent the models which are only trained in source domain or target domain data, respectively. They represent the upper and lower performance bounds of the SFOD task.

C2F: Adaptation from Normal to Foggy Weather. In real-world application scenarios, e.g., automated driving, object detectors tend to encounter various complex weather conditions. To study the domain shift caused by weather conditions, we perform the adaptation from normal weather to foggy weather. For fair comparison, our experiments are conducted in two manners: 1) All levels: Using all target data with three foggy levels for training. 2) Single level: Using partial target data with a foggy level at 0.02 for training. The results are shown in Table 1. Our method achieves an

Methods		Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP
	Source only (Single level)	23.4	23.8	29.7	8.1	12.9	5.0	18.3	24.5	18.2
	Source only (All levels)	35.1	39.4	47.0	10.7	32.5	10.1	30.0	36.9	30.7
UDAOD	MAF [13]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
	SW-Faster [32]	32.3	42.2	47.3	23.7	41.3	27.8	28.3	35.4	34.8
	iFAN [52]	32.6	40.0	48.5	27.9	45.5	31.7	22.8	33.0	35.3
	CR-DA-DET [44]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
	AT-Faster [14]	34.6	47.0	50.0	23.7	43.3	**38.7**	33.4	38.8	38.7
SFOD	SED(Mosaic) [25]	33.2	40.7	44.5	25.5	39.0	22.2	28.4	34.1	33.5
	HCL [17]	26.9	46.0	41.3	**33.0**	25.0	28.1	35.9	40.7	34.6
	A^{2}SFOD [8]	32.3	44.1	44.6	28.1	34.3	29.0	31.8	38.9	35.4
	SOAP [43]	35.9	45.0	48.4	23.9	37.2	24.3	31.8	37.9	35.5
	LODS [24]	34.0	45.7	48.8	27.3	39.7	19.6	33.2	37.8	35.8
	IRG [39]	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1
	Ours (Single level)	42.0	48.7	56.3	19.3	39.3	5.5	34.2	41.6	35.9
	Ours (All levels)	**46.1**	**52.8**	**63.4**	21.8	**46.7**	5.5	**37.4**	**48.4**	**40.3**
	Oracle	51.3	57.5	70.2	30.9	60.5	26.9	40.0	50.4	48.5

Table 1: Results of adaptation from normal to foggy weather (C2F). "Source only" and "Oracle" refer to the models trained by only using labeled source domain data and labeled target domain data, respectively.

Methods		Truck	Car	Rider	Person	Motor	Bicycle	Bus	mAP
	Source only	9.9	51.5	17.8	28.7	7.5	10.8	7.6	19.1
UDAOD	DA-Faster [7]	14.3	44.6	26.5	29.4	15.8	20.6	16.8	24.0
	SW-Faster [32]	15.2	45.7	29.5	30.2	17.1	21.2	18.4	25.3
	CR-DA-DET [44]	19.5	46.3	31.3	31.4	17.3	23.8	18.9	26.9
SFOD	SED [25]	20.4	48.8	32.4	31.0	15.0	24.3	21.3	27.6
	SED(Mosaic) [25]	20.6	50.4	32.6	32.4	18.9	25.0	23.4	29.0
	A^{2}SFOD [8]	**26.6**	50.2	**36.3**	33.2	**22.5**	**28.2**	**24.4**	**31.6**
	Ours	19.3	**62.4**	34.5	**42.6**	17.0	26.3	16.9	31.3
	Oracle	47.7	72.1	38.4	50.0	25.5	32.3	42.8	44.1

Table 2: Results of adaptation from small-scale to large-scale dataset (C2B).

Methods	mAP	Methods	mAP
Source only	36.3	MeGA-CDA [38]	43.0
DA-Faster [7]	38.5	NL [19]	43.0
SW-Faster [32]	37.9	SAPNet [21]	43.4
MAF [13]	41.0	SGA-S [49]	43.5
AT-Faster [14]	42.1	CST-DA [50]	43.6
SOAP [43]	42.7	A^{2}SFOD [8]	44.9
SFOD [25]	43.6	IRG [39]	45.7
LODS [24]	43.9	Ours	**47.0**
SED(Mosaic) [25]	44.6	Oracle	68.9

Table 3: Results of adaptation across cameras (K2C).

Methods	mAP	Methods	mAP
Source only	40.5	NL [19]	43.0
MAF [13]	41.1	UMT [10]	43.1
AT-Faster [14]	42.1	MeGA-CDA [38]	44.8
HTCN [2]	42.5	CR-DA-DET [44]	46.1
SED [25]	42.3	A^{2}SFOD [8]	44.0
SED(Mosaic) [25]	43.1	Ours	**57.8**
IRG [39]	43.2	Oracle	68.9

Table 4: Results of adaptation from synthetic to real scenes (S2C).

mAP score of 40.3%, which outperforms both the UDAOD and SFOD methods on this benchmark.

C2B: Adaptation from Small-scale to Large-scale Dataset. Annotating a large number of data for detection task can be very expensive and time-consuming. Therefore, the most economical way is to transfer knowledge from small-scale labeled datasets to large-scale unlabeled datasets. However, different datasets exhibit varying degrees of domain shifts. To validate the effectiveness of our method on such task, we transfer the source-pretrained model from Cityscapes (source domain) to BDD100k (target domain). Following the setting of previous studies [25, 8], we keep 8 categories in BDD100k that are the same as Cityscapes. Since the detection performance of the category "train" is always close to 0, we only report the mAP score of 7 categories in Table 2. The results show that our method achieves very competitive performance with the

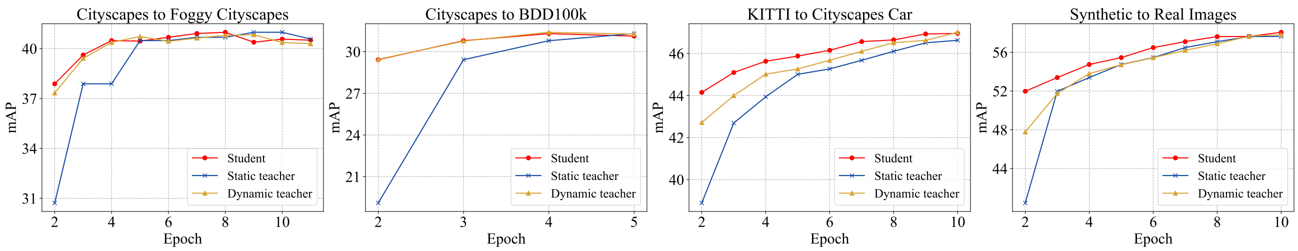


Figure 4: The training curves of each model within the multi-teacher framework during the whole training process.

Foggy level	Method	DT	ST	mAP
All levels	Source only	-	-	30.7
	Single-teacher	-	✓	36.6
	Single-teacher	✓	-	38.0
	Ours	✓	✓	**40.3**
Single level	Source only	-	-	18.2
	Single-teacher	-	✓	27.2
	Single-teacher	✓	-	32.9
	Ours	✓	✓	**35.9**

Table 5: Results of single-teacher and multi-teacher methods on C2F benchmark. **DT** and **ST** represent the dynamic teacher and static teacher, respectively.

Weights flowing strategy	K2C	C2F	Avg
Baseline	43.8	36.6	40.2
S $\longrightarrow$ ST	46.8	39.6	43.2
DT $\longrightarrow$ S	44.1	37.8	41.0
DT $\longrightarrow$ ST	46.4	38.9	42.7
S $\longleftrightarrow$ ST (Ours)	**47.0**	**40.3**	**43.7**

Table 6: Results of different exchange strategies on K2C and C2F benchmarks. "Baseline" means training the proposed multi-teacher framework without any weights flowing.

latest state-of-the-art SFOD method on this benchmark.

K2C: Adaptation across Various Cameras. Due to different camera settings (e.g., angle, resolution, quality, and type), domain shifts always occur in cross-camera images. To explore our method on cross-camera images, we adapt the model trained on KITTI to SIM10k, a dataset with images taken from real-world but different photographic equipment. Following previous studies, we only evaluate the performance on "Car" category. The results are reported in Table 3, where we can see that our method obtains state-of-the-art performance on this benchmark.

S2C: Adaptation from Synthetic to Real Scenarios. Synthetic images provide an alternative to address the challenges of data collection and manual labeling. However, there is a substantial domain gap between synthetic data and real data. To study the adaptation from synthetic to real scenes, we use the model pre-trained on the entire Sim10k dataset as the source model. The training set of Cityscapes is used as target data by reserving car images and discarding other categories. Results in Table 4 show that our method outperforms the existing SFOD approach by a large margin of +13.8%, which demonstrates the superiority of our method on this benchmark.

5.4. Ablation study

Single-teacher VS. Multi-teacher. We investigate the necessity of multi-teacher framework by comparing it with single-teacher method on C2F benchmark. The single-teacher methods employ either a static teacher or a dynamic teacher to guide the student learning process, which no longer involves using the exchange strategy and consensus mechanism. As shown in Table 5, our multi-teacher framework achieves the best performance compared to the single-teacher frameworks on both foggy levels. The success can be attributed to the superiority of exchange strategy and consensus mechanism in multi-teacher framework.

Weights Flowing Strategy. To verify the effectiveness of the proposed method, we also explore the performance of other weights flowing strategies. The comparison results are shown in Table 6, where $A \to B$ represents the single-direction weights flowing strategy that model B copies the weights of model A, while model A retains its weights, and $A \leftrightarrow B$ denotes our double-direction weights flowing strategy. We can see that all weights flowing strategies show the superiority to the baseline model that does not involve any weights swapping. Moreover, the proposed double-direction weights flowing strategy outperforms other single-direction strategies on both K2C and C2F benchmarks. This again demonstrates the superiority of our method.

5.5. Result Analysis

Training Stability. The training curves of each model within our multi-teacher framework on the four benchmarks are shown in Figure 4. Compared with the training curves of the conventional MT framework (see Figure 2), the performance of the student, static teacher and dynamic teacher models is stably improved and gradually converges to a consistent point as the training progresses. We can see that the training instability problem of conventional MT framework is effectively alleviated by our method.

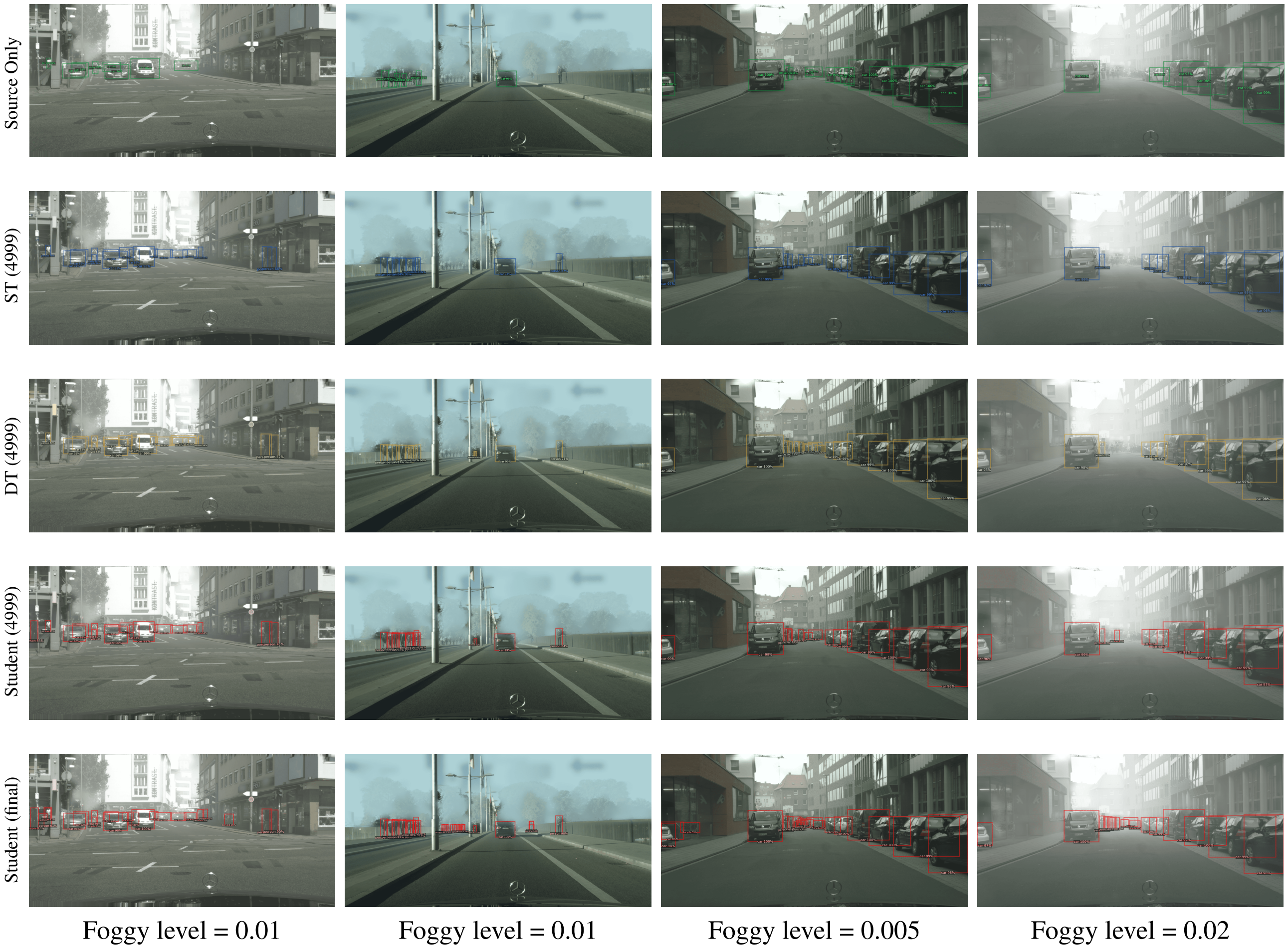


Figure 5: Detection results of different foggy-level images predicted by the dynamic teacher, static teacher, student models trained in different times. “DT (4999)”, “ST (4999)”, and “Student (4999)” represent the dynamic teacher, static teacher, and student model in the 4999-th iteration, respectively. “Student (final)” represents the student model saved at the end of training.

Visualization. We conduct an analysis by visualizing the detection results of the static teacher, dynamic teacher, and student models. This visualization is performed by inputting several images with varying foggy degrees from the Foggy Cityscape dataset [33]. The detection results of the three models for these images are shown in Figure 5. It is evident that the two teacher models yield varying detection results for each image, implying the potential complementarity of their predictive results. This observation prompts us to make a consensus on the divergent predictions of the two teacher models to enhance the quality of pseudo labels. The effectiveness of the consensus mechanism is further proven by the detection results of the student model obtained at the final iteration, which has shown superior recall and accuracy compared to the student model at the intermediate $(4,999$-th) iteration.

6. Conclusion

In this paper, we present a simple yet novel *Periodically Exchange Teacher-Student* method to tackle the training instability problem ignored by current MT-based SFOD methods. Our method employs a static teacher model, a dynamic teacher model, and a student model. At the end of each training period, we exchange the weights between the static teacher and student models. Within each period, the static teacher maintains its weights, while the student model is trained using pseudo labels generated by both teachers. Meanwhile, the dynamic teacher is continually updated using the EMA of the student model per iteration throughout the whole training phase. The extensive experimental results demonstrate the effectiveness of our method. Our method provides a new insight for MT-based self-training methods.

Acknowledgements

This work was supported by the Fujian Provincial Natural Science Foundation (No. 2022J05135), the University-Industry Project of Fujian Provincial Department of Science and Technology (No. 2020H6005), and the National Natural Science Foundation of China (No. U21A20471).

References

[1] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, pages 11457–11466, 2019.

[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.

[3] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *CVPR*, pages 7181–7190, 2022.

[4] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.

[5] Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, and Yueting Zhuang. Self-supervised noisy label learning for source-free unsupervised domain adaptation. In *IROS*, pages 10185–10192, 2022.

[6] Weijie Chen, Shiliang Pu, Di Xie, Shicai Yang, Yilu Guo, and Luojun Lin. Unsupervised image classification for deep representation learning. In *ECCVw*, pages 430–446. Springer, 2020.

[7] 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.

[8] Qiaosong Chu, Shuyan Li, Guangyi Chen, Kai Li, and Xiu Li. Adversarial alignment for source free object detection. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *AAAI*, pages 452–460, 2023.

[9] 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.

[10] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, 2016.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Robotics Res.*, 32(11):1231–1237, 2013.

[13] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019.

[14] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *ECCV 2020*, pages 309–324. Springer, 2020.

[15] Jin Hong, Yu-Dong Zhang, and Weitian Chen. Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation. *Knowl. Based Syst.*, 250:109155, 2022.

[16] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *WACV*, pages 749–757, 2020.

[17] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *NIPS*, 34:3635–3649, 2021.

[18] 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, 2017.

[19] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, pages 480–490, 2019.

[20] 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.

[21] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, pages 481–497. Springer, 2020.

[22] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

[23] Jingjing Li, Zhekai Du, Lei Zhu, Zhengming Ding, Ke Lu, and Heng Tao Shen. Divergence-agnostic unsupervised domain adaptation by adversarial attacks. *IEEE Trans. Patt. Anal. Mach. Intell.*, 44(11):8196–8211, 2021.

[24] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *CVPR*, pages 8014–8023, 2022.

[25] 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*, volume 35, pages 8474–8481, 2021.

[26] Zhaoyang Li, Long Zhao, Weijie Chen, Shicai Yang, Di Xie, and Shiliang Pu. Target-aware auto-augmentation for unsupervised domain adaptive object detection. In *ICASSP*, pages 3848–3852, 2022.

[27] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pages 6028–6039, 2020.

[28] Luojun Lin, Zhifeng Yang, Qipeng Liu, Yuanlong Yu, and Qifeng Lin. Run and chase: Towards accurate source-free domain adaptive object detection. In *ICME*, 2023.

[29] Rang Meng, Weijie Chen, Shicai Yang, Jie Song, Luojun Lin, Di Xie, Shiliang Pu, Xinchao Wang, Mingli Song, and Yueting Zhuang. Slimmable domain adaptation. In *CVPR*, pages 7141–7150, 2022.

[30] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A sim-

ple adaptation method for robust object detection. In *ICCV*, pages 3570–3579, 2021.

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

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

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

[34] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image Vis. Comput.*, 107:104117, 2021.

[35] Tao Sun, Cheng Lu, and Haibin Ling. Prior knowledge guided unsupervised domain adaptation. In *ECCV*, pages 639–655, 2022.

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

[37] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.

[38] 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.

[39] Vibashan VS, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. In *CVPR*, pages 3520–3530, 2023.

[40] Vibashan Vs, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mixture of teacher experts for source-free domain adaptive object detection. In *ICIP*, pages 3606–3610, 2022.

[41] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *IEEE Trans. Image Process.*, 30:4046–4056, 2021.

[42] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[43] Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, Xue Li, and Yingying Zhu. Source data-free domain adaptation of object detector through domain-specific perturbation. *Int. J. Intell. Syst.*, 36(8):3746–3766, 2021.

[44] 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.

[45] Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Anal.*, 79:102457, 2022.

[46] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *WACV*, pages 520–530, 2023.

[47] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[48] Peng Yuan, Weijie Chen, Shicai Yang, Yunyi Xuan, Di Xie, Yueting Zhuang, and Shiliang Pu. Simulation-and-mining: Towards accurate source-free unsupervised domain adaptive object detection. In *ICASSP*, pages 3843–3847, 2022.

[49] Chong Zhang, Zongxian Li, Jingjing Liu, Peixi Peng, Qixiang Ye, Shijian Lu, Tiejun Huang, and Yonghong Tian. Self-guided adaptation: Progressive representation alignment for domain adaptive object detection. *IEEE Trans. Multim.*, 24:2246–2258, 2021.

[50] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, pages 86–102, 2020.

[51] Lihua Zhou, Siying Xiao, Mao Ye, Xiatian Zhu, and Shuaifeng Li. Adaptive mutual learning for unsupervised domain adaptation. *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[52] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew Scott. ifan: Image-instance full alignment networks for adaptive object detection. In *AAAI*, volume 34, pages 13122–13129, 2020.