Adversarial Alignment for Source Free Object Detection

Qiaosong Chu[1,2*], **Shuyan Li**[1,2*], **Guangyi Chen**[3,4], **Kai Li**[5], **Xiu Li**[1,2†]
[1]Tsinghua Shenzhen International Graduate School, Shenzhen, China
[2]Tsinghua University, Beijing, China
[3]Carnegie Mellon University, Pittsburgh PA, USA
[4]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[5]NEC LABORATORIES AMERICA, INC
zqs20@mails.tsinghua.edu.cn, sl2141@cam.ac.uk, {guangyichen1994, li.gml.kai}@gmail.com, li.xiu@sz.tsinghua.edu.cn

Abstract

Source-free object detection (SFOD) aims to transfer a detector pre-trained on a label-rich source domain to an unlabeled target domain without seeing source data. While most existing SFOD methods generate pseudo labels via a source-pretrained model to guide training, these pseudo labels usually contain high noises due to heavy domain discrepancy. In order to obtain better pseudo supervisions, we divide the target domain into source-similar and source-dissimilar parts and align them in the feature space by adversarial learning. Specifically, we design a detection variance-based criterion to divide the target domain. This criterion is motivated by a finding that larger detection variances denote higher recall and larger similarity to the source domain. Then we incorporate an adversarial module into a mean teacher framework to drive the feature spaces of these two subsets indistinguishable. Extensive experiments on multiple cross-domain object detection datasets demonstrate that our proposed method consistently outperforms the compared SFOD methods. Our implementation is available at https://github.com/ChuQiaosong.

Introduction

Despite the promising performance, deep object detection still heavily relies on numerous manually annotated training data. It leads to a significant performance drop in real-world scenarios when the detection system faces a new environment with the domain shift. As collecting labels for all conditions is impractical, it requires detectors to efficiently adapt to new environments without further annotations. To this end, Unsupervised Domain Adaptive (UDA) object detection has gained increasing attention in recent years (He and Zhang 2019; Saito et al. 2019; Wu et al. 2021).

The aforementioned methods are based on the assumption that both source and target domain data are accessible. However, this assumption may not hold in many real-world applications due to infeasible data transmission, computation resource restrictions, or data privacy. It poses a new challenge, which is named source data-free or source-free, i.e. only a well-trained source model is provided during adapting to the target domain without having access to source data (Kundu et al. 2022; Wang et al. 2022a; Yazdanpanah and Moradi 2022; Machireddy et al. 2022).

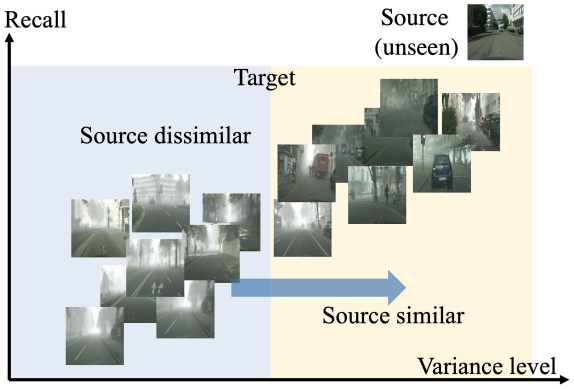

Figure 1: Basic idea of A^{2}SFOD. We propose a variance-based criterion to divide the target domain data into source-similar and source-dissimilar subsets, based on a finding that larger detection variances denote higher recall and larger similarity to the source domain. We can achieve the domain alignment by simulating the source and target domains with the divided source-similar and source-dissimilar subsets.

While the source-free challenge has been well studied for image classification tasks (Xia, Zhao, and Ding 2021; Liang, Hu, and Feng 2020), there are much fewer works that focus on Source-Free Object Detection (SFOD) (Zhang et al. 2021; VS, Oza, and Patel 2022). Due to complex background, obscured objects, and numerous negative samples, directly applying conventional source-free domain adaptation methods to SFOD cannot achieve satisfactory detection accuracy. Therefore, it is desirable to develop effective domain adaptation methods to solve the source-free problem for object detection.

As there is no manually labeled data available during adaptation, most existing SFOD methods train the model by using pseudo-labels generated by a source-pretrained model (Yuan et al. 2022; Zhang et al. 2021). However, the domain shift inevitably introduces high noises in pseudo labels, which deteriorates the detection performance (Deng et al. 2021). Though various data augmentation methods (Li et al. 2021a) have been developed to improve the quality of

*These authors contributed equally.
†*Corresponding author

pseudo labels, the domain discrepancy has not been well narrowed. The source-pretrained knowledge is difficult to adapt to these hard samples far dissimilar from the source data.

To address this problem, we propose an adversarial learning based source free object detection method (A^{2}SFOD). As shown in Figure 1, we aim to drive the source-dissimilar features close to the source-similar ones, such that pseudo labels generated by the source-pretrained model are of high quality over the whole target domain. To this end, we design a detection variance-based criterion to separate the target domain data into source-similar ones and source-dissimilar ones. This criterion is motivated by a finding that larger detection variances denote higher recall and larger similarity to the source domain. Given source-similar and source-dissimilar subsets, we propose to apply adversarial learning to the mean teacher structure to learn a model for feature space alignment. We conduct extensive experiments on five widely used detection datasets to validate the superiority of our method.

The contribution of this paper can be summarized as: 1) we find the detection variance and the similarity to source data are positively correlated, and further propose a criterion to divide the target domain; 2) we present an adversarial alignment method to refine the feature space to obtain pseudo labels of higher quality; 3) we achieve consistent and significant improvement on 5 datasets and 4 settings.

Related Work

Domain Adaptive Object Detection (DAOD) (Cai et al. 2019; Chen et al. 2018; Xu et al. 2020b; Gu, Sun, and Xu 2020; Zhang, Ma, and Wang 2021; Csaba et al. 2021) aims to address the domain shift problems in object detection task. Existing DAOD methods can be divided into two categories: feature alignment methods and self-training methods. The former ones aim to align source domain and target domain by learning a domain-agnostic feature space (Chen et al. 2018; Saito et al. 2019; Chen et al. 2020; Li et al. 2021b; Zheng et al. 2020; Wang et al. 2022c). For example DA-Faster (Chen et al. 2018) first proposed to tackle domain shift on image-level and instance-level to learn a domain-invariant region proposal network (RPN). SWDA (Saito et al. 2019) attempted to align distributions of foreground objects rather than the whole image based on strong local alignment and weak global alignment. The latter ones train the model recursively by using self-training to gradually increase the accuracy of generated pseudo labels on the target domain (Inoue et al. 2018; Khodabandeh et al. 2019; Kim et al. 2019; RoyChowdhury et al. 2019). These methods vary by different strategies to refine pseudo labels and update models. For example, WST (Kim et al. 2019) proposed weak self-training for stable training and designed adversarial background score regularization to tackle domain shifts. NL (Khodabandeh et al. 2019) formulated domain adaption as training with noisy labels and refined pseudo labels by using a classification module.

In real-world scenarios, the source data is usually inaccessible due to data privacy, leading to the SFOD problem (Lee et al. 2022; Zong et al. 2022; Ding et al. 2022; Kothandaraman et al. 2022; Wang et al. 2022b). Due to complex background and negative examples, SFOD is far more challenging than conventional source-free image classification (Agarwal et al. 2022; Ambekar et al. 2022; Bohdal et al. 2022; Xia, Zhao, and Ding 2021). SFOD-Mosaic (Li et al. 2021a) first formulated the SFOD problem and proposed to search for a fairly good confidence threshold and enabled self-training via generated pseudo labels. SOAP (Xiong et al. 2021) then added a domain-specific perturbation on the target domain and optimized the source-pretrained model via self-supervised learning. HCL (Huang et al. 2021) exploited historical source hypothesis to make up for the lack of source data. S&M (Yuan et al. 2022) proposed a Simulation-and-Mining framework that modeled the SFOD task into an unsupervised false negatives mining problem. Recently, more new methods have been developed (Li et al. 2022; Liang et al. 2022). However, these methods cannot well narrow the gap between the source domain and the target domain. All in all, SFOD is far from being fully explored and more effective SFOD methods are desired to develop.

Method

Source-free object detection (SFOD) aims to adapt a detector pre-trained on the source domain to an unlabeled target domain. In this process, the data in the source domain is untouched. Specifically, given an unlabeled target dataset $\{X_i^t\}_{i=1}^N$ (N is the number of images) and a detector F with source pre-trained parameters θ_s (e.g. a Faster-RCNN (Ren et al. 2015) model), we need to update the parameters to θ_t for the target domain. In this paper, we propose a method for this task, called A^{2}SFOD, whose overall framework is shown in Figure 2. In A^{2}SFOD, we first divide the target data into two subsets, source-similar and source-dissimilar, via the variance of the predictions from the source-pretrained detector F_{θ_s}. Then we align these two subsets via adversarial learning and finetune the detector via mean-teacher learning. In the following sections, we will detail each stage.

Target Self-Division

Aligning source and target domain is a widely-used method for conventional domain adaptation tasks (Kang et al. 2019; Zhu et al. 2019; Wang et al. 2019). Given the source and target data, we can intuitively achieve the alignment in the data space (Chen et al. 2020) or feature space (Saito et al. 2019). However, the lack of source data poses a new challenge for domain alignment.

Although we have no access to source data, the source-pretrained model does convey rich information about the source domain. Accordingly, we propose to self-divide the target data into two subsets by the pre-trained model to explicitly simulate the source and target domain. To achieve this goal, we design a detection variance-based division criterion, where the detection variance is calculated based on predictions yielded by the source-pretrained model on target data. It is motivated by a finding that larger detection variances denote higher similarity to the source data. Specifically, the pre-trained model tends to yield more predictions

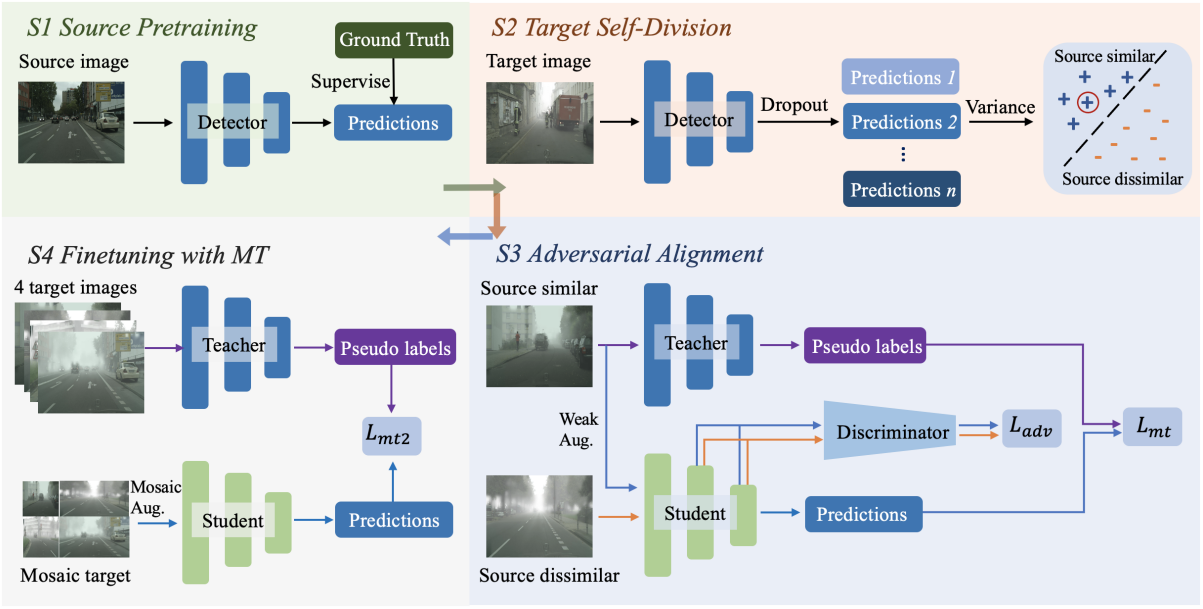

Figure 2: Framework of A^{2}SFOD. It contains four stages, source pre-training, target self-division, adversarial alignment, and fine-tuning with MT, where the first stage is unseen during adaptation.

of hard samples for source-similar images while directly ignore these hard samples for source-dissimilar images. These predictions of hard samples usually have high uncertainty, and hence have larger variances during adaptation.

We formulate the calculation of the detection variance as:

$$v_i = E[(F_{\theta_s}(X_i) - E[F_{\theta_s}(X_i)])^2], \tag{1}$$

where $F_{\theta_s}(X_i)$ represents the predictions of image X_i via the source-pretrained model. As such calculation is intractable in practice, we instead approximate this calculation with Monte-Carlo sampling. Inspired by (Gal and Ghahramani 2016), we formulate the sampling function with dropout, which is a widely-used stochastic regularization tool in deep learning (Blundell et al. 2015). This approximation is easy to perform via M stochastic forward passes without changing the detection model.

As the corresponding outputs $F_{\theta_s}(X_i) = (\boldsymbol{b_i}, \boldsymbol{c_i})$ are composed of localization coordinates and classification scores, we formulate the detection variance as the product of two terms, box localization variance v_{bi} and classification score variance v_{ci}. Given a prediction with N_i boxes and K classes, we have $\{\boldsymbol{b_{ij}} = (x_{ij}^1, y_{ij}^1, x_{ij}^2, y_{ij}^2)\}_{j=1}^{N_i}$ and $\{\boldsymbol{c_{ij}} = (c_{ij}^1, c_{ij}^2, ..., c_{ij}^K)\}_{j=1}^{N_i}$. We can formulate v_{bi} and v_{ci} as follows:

$$v_{bi} = \frac{1}{MN_i} \sum_{j=1}^{N_i} \sum_{m=1}^{M} ||\boldsymbol{b_{ij}^m} - \overline{\boldsymbol{b_{ij}}}||^2, \tag{2}$$

$$v_{ci} = \frac{1}{MN_i} \sum_{j=1}^{N_i} \sum_{m=1}^{M} ||\boldsymbol{c_{ij}^m} - \overline{\boldsymbol{c_{ij}}}||^2, \tag{3}$$

where $\boldsymbol{b_{ij}^m}$, $\boldsymbol{c_{ij}^m}$ denote the localization coordinates and classification scores of the m-th forward pass of the j-th bounding box in X_i respectively, and $\overline{\boldsymbol{b_{ij}}}$, $\overline{\boldsymbol{c_{ij}}}$ denote the corresponding average value of total M forward passes. Then we have the detection variance of X_i as $v_i = v_{bi}v_{ci}$. We order these images according to their variances from small to large, and use r_i to denote the ranking of X_i. We define the variance level of the i-th image as $vl_i = \frac{r_i}{N}$. We consider X_i as source-similar if $vl_i \geq \sigma$ and source-dissimilar otherwise, where $\sigma \in (0, 1)$ is a pre-defined threshold. In this way, we divide the target domain data into source-similar and source-dissimilar subsets for the preparation of adversarial alignment.

Adversarial Alignment

In this subsection, we introduce how to achieve domain alignment with the divided source similar and source dissimilar subsets. As shown in Stage 3 in Figure 2, we incorporate an adversarial training procedure into a mean-teacher architecture to achieve domain alignment in the feature space. Specifically, we build a teacher model F_{tea} and a student model F_{stu}, which apply the same network architecture with pretrained model F and are initialized with parameters θ_s. The parameters of the student model are quickly updated for domain alignment while the parameters of the teacher model are slowly updated. There are two loss functions in our adversarial alignment process to learn the student model: mean teacher loss for model adaptation and adversarial loss for domain alignment.

The goal of the mean teacher loss is to use the pseudo labels generated with the pretrained teacher model to supervise the training of the student model. First, we feed the teacher model with source similar data X_i^s and feed the student model with the data augmentation version $Aug(X_i^s)$. Then the mean teacher loss is formulated with the outputs of

both teacher and student models as:

$$L_{mt} = L_{cls}(F_{stu}(Aug(X_i^s), F_{tea}(X_i^s))) + L_{reg}(F_{stu}(Aug(X_i^s), F_{tea}(X_i^s))), \quad (4)$$

where L_{reg} is the location regression loss which calculates the L1-smooth distance for predicted and supervised bounding box and L_{cls} is the cross-entropy loss for classification supervision. Both L_{reg} and L_{cls} are widely-used losses in the field of object detection, such as Faster-RCNN (Ren et al. 2015).

To align the feature spaces of source similar and source dissimilar subsets, we further conduct adversarial learning with the outputs of the student model. We take the student model as the "generator" and build extra "discriminators" to play a min-max adversarial game, where the "generator" tries to fool the "discriminators" by generating features that can't be distinguished as source similar or source dissimilar. To capture both local and global information, we follow SW-Faster (Saito et al. 2019) to build a local discriminator D_l and a global discriminator D_g, where D_l focus on the foreground objects and D_g focus on the background.

The adversarial losses with global and local discriminators can be formulated as:

$$L_{local} = \frac{1}{WHN_s}\sum_{i=1}^{N_s}\sum_{w=1}^{W}\sum_{h=1}^{H} D_l(F_l(X_i^s))_{wh}^2 + \frac{1}{WHN_d}\sum_{j=1}^{N_d}\sum_{w=1}^{W}\sum_{h=1}^{H}(1 - D_l(F_l(X_j^d))_{wh}^2, \quad (5)$$

$$L_{global} = -\frac{1}{N_s}\sum_{i=1}^{N_s}(1 - D_g(F_g(X_i^s)))^{\gamma} log(D_g(F_g(X_i^s))) - \frac{1}{N_d}\sum_{j=1}^{N_d} D_g(F_g(X_j^d))^{\gamma} log(1 - D_g(F_g(X_j^d))), \quad (6)$$

where X_i^s, X_j^d denote the i-th source similar image and j-th source dissimilar image; F_l, F_g denote different layers on the backbone that capture the local feature (feature maps) and global feature (a feature vector) respectively; W, H denote width and height of local feature map on the backbone; N_s, N_d denote the total number of source similar images and source dissimilar images, and γ is a Focal loss (Lin et al. 2017) parameter which controls the model to focus on hard-to-classify examples but not the easy ones. Compared with the local adversarial loss, the global one applies the Focal loss to focus on the hard examples and ignore easy-to-classify examples to achieve a weak alignment, without hurting the performance of the local model (Saito et al. 2019).

Finally, we update the student model by fusing the mean teacher loss and adversarial losses as

$$\max_{D}\min_{F} L_{mt} - \lambda L_{adv}, \quad (7)$$

where $L_{adv} = L_{local} + L_{global}$. With this loss function, these embeddings are encouraged to have similar distributions for both source-similar and source-dissimilar images (achieved by adversarial loss L_{adv}) and have a great discriminative ability for the target domain detection (achieved by mean teacher loss L_{mt}).

Fine-Tuning with Mean Teacher

After adversarial alignment, we can get pseudo labels of high quality over the whole target data. Hence, we fine-tune the detector by using both source-similar and source-dissimilar data to make full use of the information of the whole target domain. Both teacher and student models are initialized with the parameters of the student model learned in stage 3. As the detection error mainly comes from false negative objects (Li et al. 2021a), we employ mosaic augmentation (VS, Oza, and Patel 2022; Wang et al. 2021) to simulate the false negatives to better detect small-scale and obscured objects.

As shown in Figure 2 Stage 4, we feed the teacher model with four independent target images $\{X_{i1}^t, X_{i2}^t, X_{i3}^t, X_{i4}^t\}$ and generate independent predictions $\{Y_{i1}^t, Y_{i2}^t, Y_{i3}^t, Y_{i4}^t\}$. Then we resize and mosaic these four images with data augmentation into a combined image X_{im}, with the same size as the original input X_{i1}^t. Likewise, we obtain the mosaic pseudo label Y_{im}. We formulate the loss during the fine-tuning stage as follows:

$$L_{mt2} = L_{cls}(F_{stu}(X_{im}), Y_{im}) + L_{reg}(F_{stu}(X_{im}), Y_{im}), \quad (8)$$

where L_{reg} and L_{cls} are basic location regression and classification losses respectively, which are the same as the ones in Equ. (4).

Experiments

We have conducted extensive experiments to evaluate our method, including comparisons with other SFOD methods, detailed ablation studies, and analysis.

Datasets

We evaluated our method on five popular object detection datasets. The detailed information of these datasets is summarized in the following: **(1)Cityscapes** (Cordts et al. 2016) collects different scenes from various cities on the street, which contains 2,975 training images and 500 validation images. We utilized the rectangle of the instance mask to obtain bounding boxes following previous work. **(2)Foggy-Cityscapes** (Sakaridis et al. 2018) is constructed from Cityscapes by simulating three levels of foggy weather. It contains the same amount of images as the Cityscapes. We inherited the annotations of Cityscapes. **(3)KITTI** (Geiger, Lenz, and Urtasun 2012) is a dataset containing 7,481 training images for autonomous driving different from Cityscapes. **(4)Sim10k** (Johnson-Roberson et al. 2017) is a simulation dataset generated from a popular computer game Grand Theft Auto V. It contains 10,000 images of the synthetic driving scene with 58,071 bounding boxes of the car. **(5)BDD100k** (Yu et al. 2018) is a large dataset including 100k images with six types of weather, six different scenes, and three categories for the time of day. Following (Xu et al. 2020a), we applied the daytime subset of the dataset in our experiment, where 36,728 images were used for training and the other 5,258 images for validation.

Methods	truck	car	rider	person	train	motor	bicycle	bus	mAP
Source only	10.7	30.2	30.8	23.6	9.2	16.3	24.7	19.7	20.6
DA-Faster (Chen et al. 2018)	19.5	43.5	36.5	28.7	12.6	24.8	29.1	33.1	28.5
SW-Faster (Saito et al. 2019)	23.7	47.3	42.2	32.3	27.8	28.3	35.4	41.3	34.8
MAF (He and Zhang 2019)	23.8	43.9	39.5	28.2	33.3	29.2	33.9	39.9	34.0
CR-DA-DET (Xu et al. 2020a)	27.2	49.2	43.8	32.9	36.4	30.3	34.6	45.1	37.4
AT-Faster (He and Zhang 2020)	23.7	50.0	47.0	34.6	38.7	33.4	38.8	43.3	38.7
HCL (Huang et al. 2021)	26.9	46.0	41.3	33.0	25.0	28.1	35.9	40.7	34.6
SFOD-Mosaic (Li et al. 2021a)	25.5	44.5	40.7	**33.2**	22.2	28.4	34.1	**39.0**	33.5
A^2SFOD(Ours)	**28.1**	**44.6**	**44.1**	32.3	**29.0**	**31.8**	**38.9**	34.3	**35.4**
Oracle	38.1	49.8	53.1	33.1	37.4	41.1	57.4	48.2	44.8

Table 1: Adaptation from Normal to Foggy Weather: Cityscapes → Foggy-Cityscapes

Methods	AP of car
Source only	35.7
DA-Faster (Chen et al. 2018)	38.5
SW-Faster (Saito et al. 2019)	37.9
MAF (He and Zhang 2019)	41.0
AT-Faster (He and Zhang 2020)	42.1
Noise Labeling (Khodabandeh et al. 2019)	43.0
SOAP (Xiong et al. 2021)	42.7
SFOD-Mosaic (Li et al. 2021a)	44.6
A^2SFOD(Ours)	**44.9**
Oracle	58.5

Table 2: Adaptation to a New Sense: KITTI → Cityscapes

Methods	AP of car
Source only	33.7
DA-Faster (Chen et al. 2018)	38.5
SW-Faster (Saito et al. 2019)	40.1
MAF (He and Zhang 2019)	41.1
AT-Faster (He and Zhang 2020)	42.1
HTCN (Chen et al. 2020)	42.5
Noise Labeling (Khodabandeh et al. 2019)	43.0
SFOD-Mosaic (Li et al. 2021a)	43.1
A^2SFOD(Ours)	**44.0**
Oracle	58.5

Table 3: Adaptation from Synthetic to Real Images: Sim10k → Cityscapes

Implementation Details

We followed the setting in (Chen et al. 2018) that adopted Faster-RCNN (Ren et al. 2015) with VGG-16 pretrained on ImageNet (Russakovsky et al. 2015) for our detection model. In all experiments, the shorter side of each input image was resized to 600. The detector was trained with Stochastic Gradient Descent (SGD) with a learning rate of 0.001. To stabilize the training of Mean Teacher (Tarvainen and Valpola 2017), we only updated the teacher model every 2500 iterations using exponential moving average (EMA) weights of the student model. In the pseudo label generation process, we filtered out the bounding boxes whose classification scores were lower than 0.7 to control the quality of pseudo labels. In the testing phase, we evaluated the adaptation performance by reporting mean average precision (mAP) with an IoU threshold of 0.5. Following (Saito et al. 2019), we set $\lambda = 0.1$ for Sim10k → Cityscapes in Equ. (7) and $\lambda = 1$ for other tasks. We set the threshold parameter $\sigma = 0.7$ as it is empirically found to result in the best performance. All experiments were implemented with PyTorch 1.7.1.

Comparisons with Other SFOD Methods

In this subsection, we evaluated the transferability of our method in 4 aspects, including from a normal environment to foggy weather, from training dataset to unseen new scenes, from synthetic to real images, and from a data-limited source to a large-scale target. For a fair comparison, we strictly followed the experiment setting of SFOD-Mosaic (Li et al. 2021a) and applied the similar source-only model, even if a more complex source-only model results in better performance. Specifically, we mainly compared our method A^2SFOD with multiple recent methods such as DA-Faster (Chen et al. 2018), SW-Faster (Saito et al. 2019), DA-Detection (Hsu et al. 2020), MAF (He and Zhang 2019), AT-Faster (He and Zhang 2020), and Noise Labeling (Khodabandeh et al. 2019); the baseline method SFOD-Mosaic (Li et al. 2021a); the "Source only" method trained with only source training data as the lower bound; and the "Oracle" method trained using labeled target data as the upper bound. Generally speaking, we achieved significant performance improvement over other methods in all settings.

Adaptation from Normal to Foggy Weather Weather condition shift is very common in real-world applications, such as autonomous driving, which requires the strong transferability of models in different weathers, especially for the obscure objects caused by extreme weather. Thus, we employed Cityscapes as the source domain and Foggy-Cityscapes (a dataset in the foggy weather) as the target domain to benchmark the methods. The results of A^2SFOD and other methods are summarized in Table 1. Compared with the baseline method SFOD-Mosaic (Li et al. 2021a), we achieved a 1.9% mAP score improvement on average. For a closer look at different classes, A^2SFOD obtained great suc-

Methods	truck	car	rider	person	motor	bicycle	bus	mAP
Source only	14.0	40.7	24.4	22.4	14.5	20.5	16.1	22.6
DA-Faster (Chen et al. 2018)	14.3	44.6	26.5	29.4	15.8	20.6	16.8	24.0
SW-Faster (Saito et al. 2019)	15.2	45.7	29.5	30.2	17.1	21.2	18.4	25.3
CR-DA-DET (Xu et al. 2020a)	19.5	46.3	31.3	31.4	17.3	23.8	18.9	26.9
SFOD-Mosaic (Li et al. 2021a)	20.6	**50.4**	32.6	32.4	18.9	25.0	23.4	29.0
A^2SFOD(Ours)	**26.6**	50.2	**36.3**	**33.2**	**22.5**	**28.2**	**24.4**	**31.6**
Oracle	53.4	53.5	42.8	41.9	37.3	38.8	58.1	47.1

Table 4: Adaptation to Large-Scale Dataset: Cityscapes → BDD100k

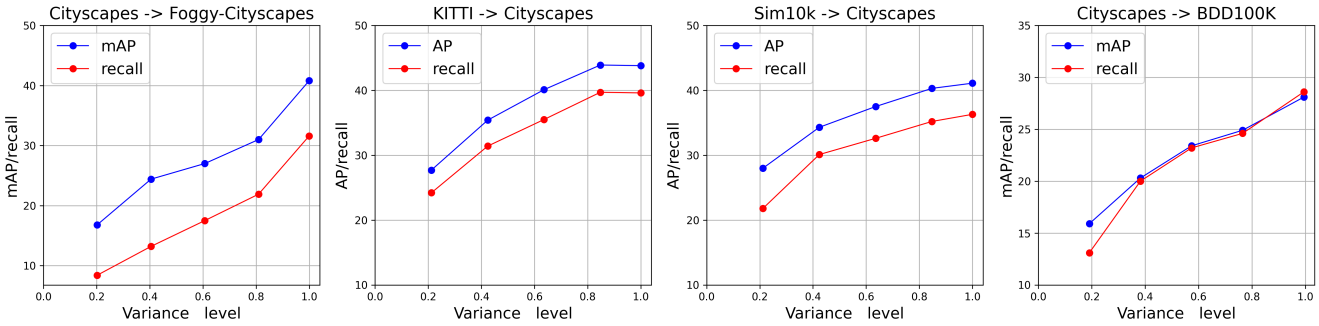

Figure 3: mAP/recall-variance relation curves in four adaption tasks. We compute the variance of each image and split data into image groups by the level of their variance. We then measure the mAP/AP and the recall of these image groups. The recall is computed under the prediction confidence = 0.5.

cess in the long-tail classes such as truck and train, while got limited improvement in the main classes such as car and bus. It is because our method which aligns the source-dissimilar images to source-similar ones encourages the model to give more confident detection. For the easy main classes, it may bring more false detection, while for long-tail classes, it effectively reduces the missing detection.

Adaptation to Unseen New Scenes Besides, intelligent systems are always required to robustly adapt from a training environment to unseen new environments. To evaluate the transferability of our method to unseen new scenes, we measure the detection performance of methods that are trained on KITTI and tested on Cityscapes with camera setup differences. The experiment results are shown in Table 2.

We found that the performance improvement was not as significant as in other settings like weather changing. It may be because the variance of the target dataset, Cityscapes, is limited, and the difference between KITTI-similar and KITTI-dissimilar sets is not significant.

Adaptation from Synthetic to Real Images In autonomous driving, labeling real-world data is expensive and time-consuming. One potential solution is to simulate the real world and train the model with synthetic data. However, there is a large domain gap between the synthetic data and the real environment. It motivates us to transfer the knowledge learned from synthetic data to real images. In this experiment, we selected a synthetic dataset, Sim10k, as the source domain and Cityscapes as the target domain. As shown in Table 3, our method consistently outperforms the baseline SFOD-Mosaic (Li et al. 2021a) and other methods.

Adaptation to Large-Scale Dataset Despite easily collecting large amounts of image data, the data annotation is expensive and labor intensive. Therefore, the transferability from limited labeled data to an unlabeled large-scale target dataset matters. In this experiment, we used Cityscapes as the source domain and BDD100k as the target domain to train the model and evaluated the detection results on only 7 common categories of the two datasets including "truck", "car", "rider", "person", "motor", "bicycle", and "bus". In Table 4, A^2SFOD obtained 31.6% mAP, which achieves $+2.6\%$ mAP improvement than the existing best result.

Ablation Studies and Further Analysis

In this subsection, we conducted ablation studies to investigate the effectiveness of each component and gave more analysis. We are to answer the following questions.

Q: Why the detection variance of the pretrained model can be regarded as a criterion for target division. A: It is because of a finding that the images with larger detection variances are more similar to the data in the source domain. Target division aims to divide the target data into a source-similar set and a source-dissimilar set for aligning the target data and untouched source data. In this paper, we proposed to use detection variance as the criterion for target division since we found the larger detection variance denotes higher recall and more similar to the source data. The pretrained model tends to give more predictions on the source-similar images. It causes a higher recall since more

Methods	truck	car	rider	person	train	motor	bicycle	bus	mAP
Baseline	20.0	43.7	37.8	26.6	13.0	24.8	37.1	36.4	29.9
Baseline + MT	23.7	44.0	42.9	**32.5**	12.9	29.7	38.1	**37.0**	32.6
Baseline + MT + TSD (Our A^{2}SFOD)	**28.1**	**44.6**	**44.1**	32.3	**29.0**	**31.8**	**38.9**	34.3	**35.4**

Table 5: Ablation study with regarding to different components of A^{2}SFOD.

predictions mean less missed detection. On the other hand, more predictions indicate that some predictions are uncertain, which have a large variance with different dropout samplings. We designed two experiments to verify our findings.

To verify the relation between the detection variance and the similarity to source data, we designed a qualitative experiment to show the source images and target images at different variance levels. While it is difficult to quantitatively identify this relation due to the lack of the similarity metric, we can clearly observe that the images with larger variance are more similar to the source data from Figure 4. As we set the threshold parameter $\sigma = 0.7$, (a), (b) and (c) are considered to be source-dissimilar images and (d), (e) represent source-similar images.

To verify whether the detection variance and recall are positively correlated, we provided a quantitative evaluation. Specifically, given a pre-trained model, we first sorted the testing data by the variance value and split them into several groups. Then we tested the model on the groups with different variance levels and calculated the recall score. As shown in Figure 3, we conducted experiments on four settings including $Cityscapes \rightarrow Foggy-Cityscapes$, $KITTI \rightarrow Cityscapes-Car$, $Sim10k \rightarrow Cityscapes-Car$, and $Cityscapes \rightarrow BDD100k$, where $A \rightarrow B$ denotes that we pretrained the model on A and tested the model on B. In all settings, we found the variance level and recall score have highly positive correlations, with correlation coefficients $R^2 = 0.962, 0.933, 0.848, 0.927$ respectively.

Besides, we also provided the mAP performance in the blue curves of Figure 3, which shows that we can obtain better detection results in images with larger variances. Generally speaking, we can obtain better detection results with the source pretrained model on the source similar images, which also demonstrates the positive relation between large variance and source similarity.

(a) vl = 0.2 (b) vl = 0.4 (c) vl = 0.6

(d) vl = 0.8 (e) vl = 1 (f) source

Figure 4: Examples of images with different variance level from 0.2 to 1.0 and images of source domain.

Q: Can we directly apply the recall as the criterion? A: No. Recall is a metric that evaluates the detection of a set of images. However, the criterion is a metric for each image to divide the source-similar and source-dissimilar sets. When considering recall in a single image, the number of objects is variant and may mislead the division.

Q: What are the effects of each component in the method? A: Both target self-division (TSD) based alignment and Mean Teacher (MT) based fine-tuning are critical for our method. The main differences between A^{2}SFOD and other methods are the target self-division (TSD) based alignment and Mean Teacher (MT) based fine-tuning. To explore the effectiveness of each component, we implemented an ablation study experiment on Cityscapes $\rightarrow$ Foggy-Cityscapes. The experiment results are summarized in Table 5. We remove the Mean-Teacher based fine-tuning and target self-division from A^{2}SFOD as our baseline. When adding the MT to the baseline, we can achieve +2.7% mAP improvement. When applying TSD-based alignment, we can further achieve the 35.4% mAP score with +2.8% improvement. These significant improvements demonstrate that both components are critical for our method.

Conclusion

In this work, we have proposed an adversarial learning based method A^{2}SFOD to enable better object detection performance in source-free circumstances. The basic idea is to self-divide the target dataset into source-similar and source-dissimilar parts and align them in the feature space. The source-pretrained model can generate high-quality pseudo supervisions for the aligned target domain. To achieve this, we developed a detection variance-based self-division criterion. We evaluated our proposed method on five cross-domain object detection datasets. Experimental results show our superiority over compared SFOD methods as well as the effectiveness of each component. In the future, we will explore to integrate our proposed A^{2}SFOD to various detection backbones.

Acknowledgments

This research was partly supported by the National Key R&D Program of China (Grant No. 2020AAA0108303) & NSFC 41876098, and by Shenzhen Science and Technology Project (Grant No. JCYJ20200109143041798) & Shenzhen Stable Supporting Program (WDZC20200820200655001) & Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No. ZDSYS20210623092001004).

References

Agarwal, P.; Paudel, D. P.; Zaech, J.; and Gool, L. V. 2022. Unsupervised Robust Domain Adaptation without Source Data. In *WACV*, 2805–2814.

Ambekar, S.; Ankit, A.; van der Mast, D.; Alence, M.; and Tafuro, M. 2022. SKDCGN: Source-free Knowledge Distillation of Counterfactual Generative Networks using cGANs. *CoRR*, abs/2208.04226.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Networks. In *ICML*, 1613–1622.

Bohdal, O.; Li, D.; Hu, S. X.; and Hospedales, T. M. 2022. Feed-Forward Source-Free Latent Domain Adaptation via Cross-Attention. *CoRR*, abs/2207.07624.

Cai, Q.; Pan, Y.; Ngo, C.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring Object Relation in Mean Teacher for Cross-Domain Detection. In *CVPR*, 11457–11466.

Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; and Dou, Q. 2020. Harmonizing Transferability and Discriminability for Adapting Object Detectors. In *CVPR*, 8866–8875.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Gool, L. V. 2018. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In *CVPR*, 3339–3348.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 3213–3223.

Csaba, B.; Qi, X.; Chaudhry, A.; Dokania, P. K.; and Torr, P. H. S. 2021. Multilevel Knowledge Transfer for Cross-Domain Object Detection. *CoRR*, abs/2108.00977.

Deng, J.; Li, W.; Chen, Y.; and Duan, L. 2021. Unbiased Mean Teacher for Cross-Domain Object Detection. In *CVPR*, 4091–4101.

Ding, Y.; Sheng, L.; Liang, J.; Zheng, A.; and He, R. 2022. ProxyMix: Proxy-based Mixup Training with Label Refinery for Source-Free Domain Adaptation. *CoRR*, abs/2205.14566.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M.; and Weinberger, K. Q., eds., *ICML*, 1050–1059.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 3354–3361.

Gu, X.; Sun, J.; and Xu, Z. 2020. Spherical Space Domain Adaptation With Robust Pseudo-Label Loss. In *CVPR*, 9098–9107.

He, Z.; and Zhang, L. 2019. Multi-Adversarial Faster-RCNN for Unrestricted Object Detection. In *ICCV*, 6667–6676.

He, Z.; and Zhang, L. 2020. Domain Adaptive Object Detection via Asymmetric Tri-Way Faster-RCNN. In *ECCV*, 309–324.

Hsu, H.; Yao, C.; Tsai, Y.; Hung, W.; Tseng, H.; Singh, M. K.; and Yang, M. 2020. Progressive Domain Adaptation for Object Detection. In *WACV*, 738–746.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021. Model Adaptation: Historical Contrastive Learning for Unsupervised Domain Adaptation without Source Data. In *NIPS*, 3635–3649.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation. In *CVPR*, 5001–5009.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S. N.; Rosaen, K.; and Vasudevan, R. 2017. Driving in the Matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*, 746–753.

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive Adaptation Network for Unsupervised Domain Adaptation. In *CVPR*, 4893–4902.

Khodabandeh, M.; Vahdat, A.; Ranjbar, M.; and Macready, W. G. 2019. A Robust Learning Approach to Domain Adaptive Object Detection. In *ICCV*, 480–490.

Kim, S.; Choi, J.; Kim, T.; and Kim, C. 2019. Self-Training and Adversarial Background Regularization for Unsupervised Domain Adaptive One-Stage Object Detection. In *ICCV*, 6091–6100.

Kothandaraman, D.; Shekhar, S.; Sancheti, A.; Ghuhan, M.; Shukla, T.; and Manocha, D. 2022. DistillAdapt: Source-Free Active Visual Domain Adaptation. *CoRR*, abs/2205.12840.

Kundu, J. N.; Kulkarni, A. R.; Bhambri, S.; Mehta, D.; Kulkarni, S. A.; Jampani, V.; and Radhakrishnan, V. B. 2022. Balancing Discriminability and Transferability for Source-Free Domain Adaptation. In *ICML*, 11710–11728.

Lee, J.; Jung, D.; Yim, J.; and Yoon, S. 2022. Confidence Score for Source-Free Unsupervised Domain Adaptation. In *ICML*, 12365–12377.

Li, S.; Ye, M.; Zhu, X.; Zhou, L.; and Xiong, L. 2022. Source-Free Object Detection by Learning to Overlook Domain Style. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*.

Li, X.; Chen, W.; Xie, D.; Yang, S.; Yuan, P.; Pu, S.; and Zhuang, Y. 2021a. A Free Lunch for Unsupervised Domain Adaptive Object Detection without Source Data. In *AAAI*, 8474–8481.

Li, Y.; Dai, X.; Ma, C.; Liu, Y.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2021b. Cross-Domain Object Detection via Adaptive Self-Training. *CoRR*, abs/2111.13216.

Liang, J.; Hu, D.; and Feng, J. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *ICML*, 6028–6039.

Liang, J.; Hu, D.; Wang, Y.; He, R.; and Feng, J. 2022. Source Data-Absent Unsupervised Domain Adaptation Through Hypothesis Transfer and Labeling Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Machireddy, A.; Krishnan, R.; Ahuja, N.; and Tickoo, O. 2022. Continual Active Adaptation to Evolving Distributional Shifts. In *CVPR*, 3444–3450.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 91–99.

RoyChowdhury, A.; Chakrabarty, P.; Singh, A.; Jin, S.; Jiang, H.; Cao, L.; and Learned-Miller, E. G. 2019. Automatic Adaptation of Object Detectors to New Domains Using Self-Training. In *CVPR*, 780–790.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3): 211–252.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-Weak Distribution Alignment for Adaptive Object Detection. In *CVPR*, 6956–6965.

Sakaridis, C.; Dai, D.; Hecker, S.; and Gool, L. V. 2018. Model Adaptation with Synthetic and Real Data for Semantic Dense Foggy Scene Understanding. In *ECCV*, 707–724.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 1195–1204.

VS, V.; Oza, P.; and Patel, V. M. 2022. Instance Relation Graph Guided Source-Free Domain Adaptive Object Detection. *CoRR*, abs/2203.15793.

Wang, F.; Han, Z.; Gong, Y.; and Yin, Y. 2022a. Exploring Domain-Invariant Parameters for Source Free Domain Adaptation. In *CVPR*, 7151–7160.

Wang, F.; Han, Z.; Zhang, Z.; and Yin, Y. 2022b. Active Source Free Domain Adaptation. *CoRR*, abs/2205.10711.

Wang, W.; Liao, S.; Zhao, F.; Kang, C.; and Shao, L. 2021. DomainMix: Learning Generalizable Person Re-Identification Without Human Annotations. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*.

Wang, W.; Zhang, J.; Zhai, W.; Cao, Y.; and Tao, D. 2022c. Robust Object Detection via Adversarial Novel Style Exploration. *TIP*, 31: 1949–1962.

Wang, X.; Cai, Z.; Gao, D.; and Vasconcelos, N. 2019. Towards Universal Object Detection by Domain Attention. In *CVPR*, 7289–7298.

Wu, A.; Liu, R.; Han, Y.; Zhu, L.; and Yang, Y. 2021. Vector-Decomposed Disentanglement for Domain-Invariant Object Detection. In *ICCV*, 9322–9331.

Xia, H.; Zhao, H.; and Ding, Z. 2021. Adaptive Adversarial Network for Source-free Domain Adaptation. In *ICCV*, 8990–8999.

Xiong, L.; Ye, M.; Zhang, D.; Gan, Y.; Li, X.; and Zhu, Y. 2021. Source data-free domain adaptation of object detector through domain-specific perturbation. *IJIS*, 3746–3766.

Xu, C.; Zhao, X.; Jin, X.; and Wei, X. 2020a. Exploring Categorical Regularization for Domain Adaptive Object Detection. In *CVPR*, 11721–11730.

Xu, M.; Wang, H.; Ni, B.; Tian, Q.; and Zhang, W. 2020b. Cross-Domain Detection via Graph-Induced Prototype Alignment. In *CVPR*, 12352–12361.

Yazdanpanah, M.; and Moradi, P. 2022. Visual Domain Bridge: A Source-Free Domain Adaptation for Cross-Domain Few-Shot Learning. In *CVPR*, 2868–2877.

Yu, F.; Xian, W.; Chen, Y.; Liu, F.; Liao, M.; Madhavan, V.; and Darrell, T. 2018. BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. *CoRR*, abs/1805.04687.

Yuan, P.; Chen, W.; Yang, S.; Xuan, Y.; Xie, D.; Zhuang, Y.; and Pu, S. 2022. Simulation-and-Mining: Towards Accurate Source-Free Unsupervised Domain Adaptive Object Detection. In *ICASSP*, 3843–3847.

Zhang, D.; Ye, M.; Xiong, L.; Li, S.; and Li, X. 2021. Source-Style Transferred Mean Teacher for Source-data Free Object Detection. In *MMAsia*, 4:1–4:8.

Zhang, T.; Ma, W.; and Wang, G. 2021. Six-Channel Image Representation for Cross-Domain Object Detection. In *ICIG*, 171–184.

Zheng, Y.; Huang, D.; Liu, S.; and Wang, Y. 2020. Cross-domain Object Detection through Coarse-to-Fine Feature Adaptation. In *CVPR*, 13763–13772.

Zhu, X.; Pang, J.; Yang, C.; Shi, J.; and Lin, D. 2019. Adapting Object Detectors via Selective Cross-Domain Alignment. In *CVPR*, 687–696.

Zong, Z.; He, J.; Zhang, L.; and Huan, H. 2022. Domain Gap Estimation for Source Free Unsupervised Domain Adaptation with Many Classifiers. *CoRR*, abs/2207.05785.