Cross-Domain Adaptive Teacher for Object Detection

Yu-Jhe Li[1*]   Xiaoliang Dai[2]   Chih-Yao Ma[2]   Yen-Cheng Liu[3]   Kan Chen[4]

Bichen Wu[2]   Zijian He[2]   Kris Kitani[1]   Peter Vajda[2]

[1]Carnegie Mellon University   [2]Meta (Facebook)   [3]Georgia Tech   [4]Waymo

{yujheli,kkitani}@cs.cmu.edu, {xiaoliangdai,cyma,zijian,wbc,vajdap}@fb.com,

ycliu@gatech.edu, kanchen@waymo.com

Abstract

We address the task of domain adaptation in object detection, where there is an obvious domain gap between a domain with annotations (source) and a domain of interest without annotations (target). As a popular semi-supervised learning method, the teacher-student framework (a student model is supervised by the pseudo labels from a teacher model) has also yielded a large accuracy gain in cross-domain object detection. However, it suffers from the domain shift and generates many low-quality pseudo labels (e.g., false positives), which leads to sub-optimal performance. To mitigate this problem, we propose a teacher-student framework named Adaptive Teacher (AT) which leverages domain adversarial learning and weak-strong data augmentation to address the domain gap. Specifically, we employ feature-level adversarial training in the student model, allowing features derived from the source and target domains to share similar distributions. This process ensures the student model produces domain-invariant features. Furthermore, we apply weak-strong augmentation and mutual learning between the teacher model (taking data from the target domain) and the student model (taking data from both domains). This enables the teacher model to learn the knowledge from the student model without being biased to the source domain. We show that AT demonstrates superiority over existing approaches and even Oracle (fully-supervised) models by a large margin. For example, we achieve 50.9% (49.3%) mAP on Foggy Cityscape (Clipart1K), which is 9.2% (5.2%) and 8.2% (11.0%) higher than previous state-of-the-art and Oracle, respectively.

1. Introduction

Developing algorithms that can transfer the knowledge learned from one labeled dataset (*i.e.,* source domain) to another unlabeled dataset (*i.e.,* target domain) becomes in-
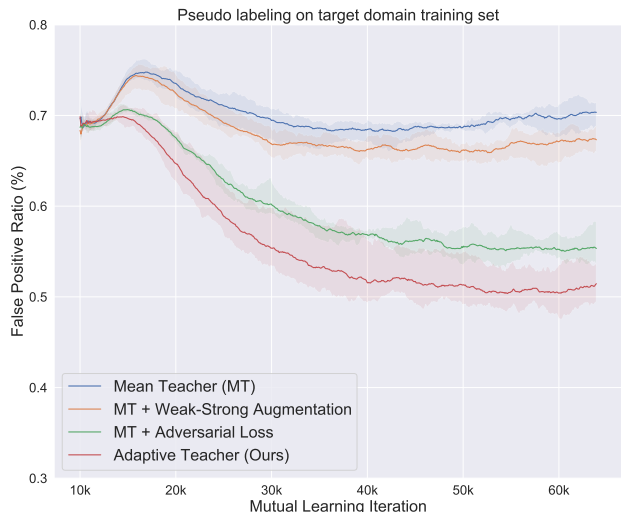

Figure 1. **The effectiveness of domain loss and weak-strong augmentation on pseudo labeling in Mean Teacher (MT) [40]**. The figure shows the false positive ratio on the training set of Clipart1k (target) with PASCAL VOC as source. We run 5 identical experiments for each setting and plot the error bound accordingly. Due to inherent domain shift in the Teacher model, it generates noisy pseudo labels without domain loss. The weak-strong augmentation is able to stabilize pseudo labeling.

creasingly important for object detection. Researchers have proposed various methods, such as domain classifier and adversarial learning [10], to address the task of cross-domain adaptation in object detection [2, 3, 14, 32, 39, 42, 44]. Even though these methods have led to accuracy improvement, solely using adversarial learning on the complex recognition task such as object detection is still limited. Hence, there is generally still a large performance gap from the Oracle model (fully supervision) on the target domain.

To explore the potential of self-training on the unlabeled target domain for improved detection performance, researchers have exploited and extend the teacher-student self-training method from semi-supervised learning to domain adaptation [40]. These approaches are able to learn

without annotations by typically involving a teacher model to generate pseudo labels to update student model. These methods have led to notable accuracy gains in the domain adaptation scenario. For example, MTOR [1] employs the Mean Teacher (MT) [40] as its pipeline to identify relations using region-level, inter-graph, and intra-graph consistency. Unbiased Mean Teacher (UMT) [8] is proposed to augment the teacher-student framework with CycleGAN [43] and achieved further performance improvement.

Despite the accuracy gain, the teacher-student framework still face a major challenge upon the settings of domain adaptation: Unlike semi-supervised learning, the pseudo label generated from the teacher model usually contains a substantial amount of errors and false positives, as shown in Figure 1. This is because the scenario of domain adaptation typically involves a large domain gap between the labeled data (source domain) and unlabeled data (target domain). The teacher model is trained on, biased to, and only able to capture features precisely on the source domain, hence unable to provide high-quality pseudo labels in the target domain. As a result, directly applying the teacher-student framework only leads to sub-optimal adaptation performance.

To address this problem, we propose a self-training framework named Adaptive Teacher (AT) to mitigate the domain shift and improve the pseudo labeling quality on the target domain leveraging adversarial learning and mutual learning. Our model comprises of two separate modules: target-specific Teacher model and cross-domain Student model. We also apply weak augmentation (only strong augmentation in Student model) and feed images from the target domain into the Teacher model, which we refer to as "Weak-Strong augmentation", following Unbiased Teacher (UT) [22]. This allows the teacher model to generate reliable pseudo labels without being affected by heavy augmentation. In addition, to mitigate the domain bias toward source domain in the Student model, we apply adversarial learning by introducing a discriminator with gradient reverse layer to align the distribution across two domains in the Student model. With all the techniques, we observe the pseudo label quality improved significantly, as shown in Figure 1, where the false positive ratio is suppressed by up to 35%. This further leads to substantial accuracy gain across all the domain adaptation experiments and outperforms all existing methods. We summarize the contributions of this paper as follows:

- We demonstrate the limitation of the teacher-student framework in the domain adaptation scenario: The teacher model is biased toward the source domain and only able to produce low-quality pseudo labels on the target domain.
- We propose a novel framework leveraging adversarial learning augmented mutual learning and weak-strong

augmentation to address domain shift in cross-domain object detection.
- Our method is able to deal with domain shift and outperform all existing SOTA by a large margin. For example, we achieve 50.9% mAP on Foggy Cityscape, which is 9.2% and 8.2% higher than SOTA and Oracle (full supervision).

2. Related Works

Object Detection. Object detection is a task to localize the object and its location given an input image. Recently, deep models have shown to be effective in object detection with anchor-based approaches, *e.g.*, Faster R-CNN [30] which introduces Region Proposal Networks (RPN) to facilitate the proposal generation for region of interests (ROI). Afterwards, several anchor-based works [5, 6, 15, 20, 26, 37] are proposed to improve the performance and efficacy. On the other hand, anchor-free methods are also proposed as one-stage detectors without the step of generating region proposals. YOLO [27] produces bounding boxes and the confidence score jointly for multiple classes as regression task. Several of its variants [28, 29] are also proposed. SSD [21] is also built on top of YOLO yet leverages feature maps generated from different scales of images. For our work, we employ Faster R-CNN as the backbone for detection due to its flexibility.

Domain Adaptation. Given unlabeled data from target domain, unsupervised domain adaptation (UDA) or domain adaptation (DA) aims to learn a model from additional labeled source domain to achieve satisfactory performance on the target domain. Recently, it has demonstrated its effectiveness using deep neural networks. On one hand, some works have developed discrepancy-based methods that learn the representations by minimizing the domain discrepancy, which is also known as Maximum Mean Discrepancy (MMD) [23–25]. Another line of domain adaptation is to map the domain distribution and treat it as a adversarial (min-max) optimization with a domain classifier [10, 11, 34, 41]. Some generative models such as Cycle-GAN [43] can also be seen as image-level domain adaptation. However, compared to these general vision tasks, the problem of object detection is more complicated since it has to predict the bounding box and class label for each object. Comparing with other recognition tasks, we aim at handling such challenging task of cross-domain object detection.

Cross-Domain Object Detection. Recently, more works pay more attention on domain adaptation in the task of object detection and propose various approaches. Some utilize adversarial learning with a gradient reverse layer (GRL) for mapping the feature across different domains
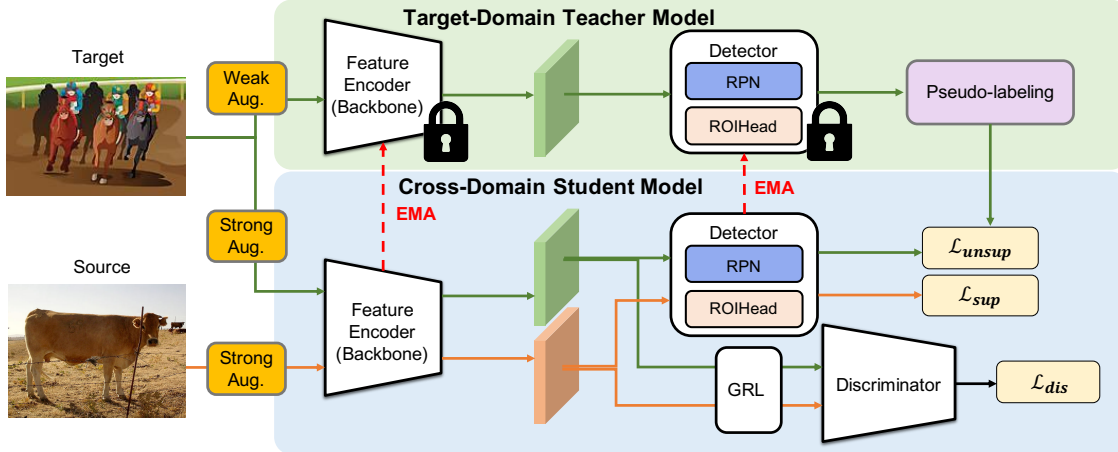
Figure 2. **Overview of our proposed Adaptive Teacher (AT).** Our model consists of two modules: 1) target-specific Teacher model for taking weakly-augmented images from target domain and 2) cross-domain Student model for taking strongly-augmented images from both domains. We train our model using two learning streams: Teacher-Student mutual learning and adversarial learning. The Teacher model generates pseudo-labels to train the Student while the Student updates the Teacher model with exponential moving average (EMA). The discriminator with gradient reverse layer is employed to align the distributions across two domains in Student model.

in [2, 3, 14, 32, 39, 42, 44]. The annotation-level adaptation [17, 18, 31] or curriculum learning [38] have also been proposed for the task of domain adaptation. Recently, another direction is to utilize Mean Teacher (MT) [40] which is originally proposed for semi-supervised learning on this task. MTOR [1] is proposed on top of MT and train its teacher network with enforcing the region-level, inter-graph, and intra-graph consistency. Similarly, Unbiased Mean Teacher (UMT) [8] has been proposed to reduce the domain shift by augmenting the training samples with Cy-cleGAN [43]. However, the above approaches are likely to suffer the same and inherent issue in Mean Teacher (MT), generating pseudo labels of low quality on target domain.

3. Adaptive Teacher

3.1. Problem Formulation and Overview

Before we demonstrate how our proposed method is able to mitigate the errors of pseudo labels in domain adaptation of object detection, we first review the problem formulation. We are given N_s labeled images $\mathcal{D}_s = \{(X_s, B_s, C_s)\}$ in source domain and N_t unlabeled images $\mathcal{D}_t = \{X_t\}$ in target domain, where $B_s = \{b_s^i\}_{i=1}^{N_s}$ denotes the bounding box annotations and $C_s = \{c_s^i\}_{i=1}^{N_s}$ denotes corresponding class labels for source image $X_s = \{x_s^i\}_{i=1}^{N_s}$. There is no annotations for the target image $X_t = \{x_t^j\}_{j=1}^{N_t}$. The ultimate goal of cross-domain object detection is to design domain-invariant detectors by leveraging $\mathcal{D}_s$ and $\mathcal{D}_t$.

The overview of our framework is presented in Figure 2. Our AT framework consists of two modules: target-specific Teacher model and cross-domain Student model. The Teacher model only takes into the weakly-augmented

images from target domain ($\mathcal{D}_t$) while the Student model takes strongly-augmented images from both domains ($\mathcal{D}_s$ and $\mathcal{D}_t$). We train our model using two training streams which are the Teacher-Student mutual learning and the adversarial learning strategies. To begin with, we train the object detector with the available source labeled data and initialize the feature encoder and detector. At the stage of mutual learning, we duplicate the initialized object detector into two identical detectors, *i.e.*, Teacher and Student models. The Teacher generates pseudo labels to train the Student while the Student updates the knowledge it learned back to the Teacher via exponential moving average (EMA). Iteratively, the pseudo-labels for training the Student are improved. Furthermore, we employ the discriminator and the gradient reverse layer (GRL) [10] for the adaptive learning (Sec. 3.3) to align the distributions across two domains in Student model. This allows the Student model to reduce domain shifts and benefits the Teacher model to generate more accurate pseudo-labels.

3.2. Mutual Learning between Teacher and Student

Following the teacher-student framework initially proposed for semi-supervised object detection, our model is also composed of two architecturally identical models: a Student model and a Teacher model. The Student model is learned by standard gradient updating, and the Teacher model is updated with the exponential moving average (EMA) of the weights from the student model. To generate precise and accurate pseudo labels for target domain images, we feed the images with **weak augmentation** as input to the Teacher to provide reliable pseudo-labels while images with **strong augmentation** as inputs of the Stu-

dent. Specifically, the target samples are augmented with randomly horizontal flipping and cropping as weak augmentation in Teacher model and randomly color jittering, grayscaling, Gaussian blurring, and cutting patches our for strong augmentations as perturbations.

Model Initialization. Initialization is significant for the self-training framework since we rely on the Teacher to generate reliable pseudo-labels for target domain without annotations to optimize the Student model. To achieve this, we first use the available supervised source data $\mathcal{D}_s = \{(X_s, B_s, C_s)\}$ to optimize our model with the supervised loss $\mathcal{L}_{sup}$. Hence, the supervised loss for training and initializing the student model using the labeled source data can be defined as:

$$\mathcal{L}_{sup}(X_s, B_s, C_s) = \mathcal{L}_{cls}^{rpn}(X_s, B_s, C_s) + \mathcal{L}_{reg}^{rpn}(X_s, B_s, C_s) \\ + \mathcal{L}_{cls}^{roi}(X_s, B_s, C_s) + \mathcal{L}_{reg}^{roi}(X_s, B_s, C_s), \tag{1}$$

where RPN loss $\mathcal{L}^{rpn}$ is the loss for learning the Region Proposal Network (RPN), which is designed to generate candidate proposals, and Region of Interest (ROI) loss $\mathcal{L}^{roi}$ is for the prediction branch of ROI. Both of RPN and ROI perform bounding box regression (reg) and classification (cls). We use binary cross-entropy loss for $\mathcal{L}_{cls}^{rpn}$ and $\mathcal{L}_{cls}^{roi}$, and l_1 loss for $\mathcal{L}_{reg}^{rpn}$ and $\mathcal{L}_{reg}^{roi}$.

Optimize Student with Target Pseudo-Labels As the labels are not available in the target domain, we adopt the pseudo-labeling method to produce dummy labels on the images from target domain to train the Student. To filter out the noisy pseudo-labels, we set a confidence threshold δ on the predicted bounding boxes from the teacher model to remove the false positives. In addition, we exclude duplicated boxes prediction by non-maximum suppression (NMS) for each class. Hence, after obtaining the pseudo-labels from Teacher model on the images of target domain, we can update the Student with the loss as:

$$\mathcal{L}_{unsup}(X_t, \hat{C}_t) = \mathcal{L}_{cls}^{rpn}(X_t, \hat{C}_t) + \mathcal{L}_{cls}^{roi}(X_t, \hat{C}_t), \tag{2}$$

where $\hat{C}_t$ denotes the pseudo labels generated by the Teacher model on target domain. Here, unsupervised losses are not applied in the task of bounding box regression since the confidence score of predicted bounding boxes on the unlabeled data can only represent the confidence of the categories for each object instead of the locations for the produced bounding boxes.

Temporally Update Teacher from Student. To obtain strong pseudo-labels from the target images following MT [40], we apply Exponential Moving Average (EMA) to update Teacher model by temporally copying the weights of the student model. The update formula can be defined as:

$$\theta_t \leftarrow \alpha\theta_t + (1 - \alpha)\theta_s, \tag{3}$$

where θ_t and θ_s denote the network parameters of Teacher and Student, respectively.

3.3. Adversarial Learning to Bridge Domain Bias

Since annotations are only available on source data, both of the Teacher and the Student can be easily biased towards the source domain during the mutual learning process. To be particular, the pseudo labels generated on target images from Teacher model are basically derived using the knowledge of the model trained with labels from source domain. As a result, we need to bridge the domain bias across source and target domains otherwise the Teacher model would generate noisy labels on target images and make the learning process collapse. Thus, we introduce adversarial learning into the framework for aligning the distributions across two domains. This leads to substantial false positive ratio reduction (20% on MT+adversarial loss) in pseudo label generation, as shown in Figure 1.

Since Student model takes images from both domains, the adversarial loss is applicable on Student model to align two distribution. To achieve adversarial learning, a domain discriminator D is placed after the feature encoder E (shown in Figure 2) on the Student model. The discriminator is aimed at discriminate where the derived feature $E(X)$ is from (source or target). Then we can define the probability of each input sample that belonging to the target domain as $D(E(X))$ and the probability of it belonging to source domain as $1 - D(E(X))$. We can update the domain discriminator D using binary cross-entropy loss given the domain label d for each of the input image. Specifically, images from the source domain are labeled as $d = 0$ and images from target domain are labeled as $d = 1$. The discriminator loss $\mathcal{L}_{dis}$ can be formulated as:

$$\mathcal{L}_{dis} = -d \log D(E(X)) - (1 - d) \log(1 - D(E(X))), \tag{4}$$

On the other hand, the feature encoder E is encouraged to produce features that confuse the discriminator D while the discriminator D aim to distinguish which domain the derived features are from. Hence, such adversarial optimization objective function can be defined as the following:

$$\mathcal{L}_{adv} = \max_E \min_D \mathcal{L}_{dis}. \tag{5}$$

Fortunately, to simply the min-max optimization, we can append an additional Gradient Reverse Layer (GRL) [10] between the feature encoder and the discriminator to produce reverse gradient. During the gradient calculation, the

Table 1. The results of cross-domain object detection on the Clipart1k test set for **PASCAL VOC → Clipart1k** adaptation. The average precision (AP, in %) on all classes is reported. The used backbone is ResNet-101 for fair comparison. We compare our method with SCL [35], SWDA [32], DM [19], CRDA [42], HTCN [2], UMT [8], Source (F-RCNN), and Oracle (F-RCNN).

Method	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	m-bike	prsn	plnt	sheep	sofa	train	tv	mAP
Source	23.0	39.6	20.1	23.6	25.7	42.6	25.2	0.9	41.2	25.6	23.7	11.2	28.2	49.5	45.2	46.9	9.1	22.3	38.9	31.5	28.8 (-16.2)
SCL	**44.7**	50.0	33.6	27.4	42.2	55.6	38.3	**19.2**	37.9	**69.0**	30.1	26.3	34.4	67.3	61.0	47.9	21.4	26.3	50.1	47.3	41.5 (-3.5)
SWDA	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1 (-6.9)
DM	25.8	**63.2**	24.5	42.4	47.9	43.1	37.5	9.1	47.0	46.7	26.8	24.9	48.1	**78.7**	63.0	45.0	21.3	36.1	52.3	53.4	41.8 (-3.2)
CRDA	28.7	55.3	31.8	26.0	40.1	**63.6**	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3 (-6.7)
HTCN	33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	51.8	40.3 (-4.7)
UMT	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	67.6	21.4	**29.5**	48.2	75.9	70.5	56.7	25.9	28.9	39.4	43.6	44.1 (-0.9)
AT	33.8	60.9	**38.6**	**49.4**	**52.4**	53.9	**56.7**	7.5	**52.8**	63.5	**34.0**	25.0	**62.2**	72.1	**77.2**	**57.7**	**27.2**	**52.0**	**55.7**	**54.1**	**49.3** (+4.3)
Oracle	33.3	47.6	43.1	38.0	24.5	82.0	57.4	22.9	48.4	49.2	37.9	46.4	41.1	54.0	73.7	39.5	36.7	19.1	53.2	52.9	45.0

Table 2. The results of cross-domain object detection on the Watercolor2k test set for **PASCAL VOC → Watercolor2k** adaptation. The average precision (AP, in %) on all classes is reported. The used backbone is ResNet-101 for fair comparison.

Method	bicycle	bird	car	cat	dog	person	mAP
Source	84.2	44.5	53.0	24.9	18.8	56.3	46.9 (-3.7)
SCL [35]	82.2	55.1	51.8	39.6	38.4	64.0	55.2 (+4.8)
SWDA [32]	82.3	55.9	46.5	32.7	35.5	66.7	53.3 (+2.7)
UMT [8]	88.2	55.3	51.7	**39.8**	**43.6**	69.9	58.1 (+7.5)
AT	**93.6**	**56.1**	**58.9**	37.3	39.6	**73.8**	**59.9** (+9.3)
Oracle	51.8	49.7	42.5	38.7	52.1	68.6	50.6

GRL negates the gradients that pass back and the gradients of feature encoder E is calculated in an opposite direction. This helps to maximize the discriminator loss for learning E while we only need to minimize the objective $\mathcal{L}_{dis}$. With the above discriminator loss, our Student model resolves the domain bias in visual features and helps Teacher to generate precise pseudo labels after several EMA updates.

We would like to note that, the design of adversarial learning in the Student model of our Adaptive Teacher is reasonable for two reasons. First, since we only feed images from target domain into Teacher model to avoid domain bias on Teacher model, the process of aligning two domains could be preferable in Student model which takes images across two domains. Feeding images from source domain like [1, 8] may bring more bias toward source domain to both Teacher and Student models. Second, adversarial learning is a min-max learning problem and requires loss function to update the model. Since Student model is updated via objective losses, applying adversarial loss to the Student model is a simple and suitable way in the standard learning of the Mean Teacher.

3.4. Full Objective and Inference

The total loss $\mathcal{L}$ for training our proposed AUT is summarized as follows:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{unsup}} \cdot \mathcal{L}_{\text{unsup}} + \lambda_{\text{dis}} \cdot \mathcal{L}_{\text{dis}}, \qquad (6)$$

where λ_{unsup} and λ_{dis} are the hyper-parameters used to control the weighting of the corresponding losses. We note that $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{unsup}}$ are developed to learn the feature encoder and detector in the Student model while $\mathcal{L}_{\text{dis}}$ is introduced to update the feature encoder and discriminator. The Teacher model is only updated through EMA discussed in the Sec 3.2.

4. Experiment

4.1. Datasets

We conduct our experiments on five public datasets, including Cityscapes [4]), Foggy Cityscapes [33], PASCAL VOC [9], Clipart1k [16], and Watercolor2k [16].

Cityscapes. Cityscapes [4] is collected by capturing images from outdoor street scenes in normal weather conditions from 50 cities, which has diverse scenes. It contains 2,975 images for training and 500 images for validation with dense pixel-level labels. The annotations of bounding boxes are converted from instance segmentation labels.

Foggy Cityscapes. Foggy Cityscapes [33] is synthesized from the images in the Cityscapes. Therefore, it has the same train/test split as Cityscapes. It simulates the condition of foggy weather according to depth information provided in Cityscapes and generates three levels of foggy weather.

PASCAL VOC. PASCAL VOC [9] contains 20 categories of common objects from real world with bounding box and class annotations. Following [32, 35], the dataset is combined from PASCAL VOC 2007 and 2012 with total 16,551 images.

Clipart1k. Clipart1k [16] contains clipart images and shares the same 20 classes with PASCAL VOC. Yet, it exhibits a large domain shift from PASCAL VOC. We follow the practice in [32, 35] and split it into training and test sets, containing 500 images each.

Watercolor2k. Watercolor2k [16] contains watercolor style images, which consists of images from 6 classes and shares with the same classes in PASCAL VOC dataset. Following [32, 35], the dataset is splitted halfly into training set and testing sets and each contains 1000 images.

Table 3. The results and comparison on cross-domain object detection on the Foggy Cityscapes test set for **Cityscapes → Foggy Cityscapes** adaptation. The average precision (AP, %) on all classes is presented. The used backbone is VGG-16 for fair comparison.

Method	bus	bicycle	car	mcycle	person	rider	train	truck	mAP
Source (F-RCNN)	20.1	31.9	39.6	16.9	29.0	37.2	5.2	8.1	23.5 (-19.2)
SCL [35]	41.8	36.2	44.8	33.6	31.6	44.0	40.7	30.4	37.9 (-4.8)
DA-Faster [3]	35.3	27.1	40.5	20.0	25.0	31.0	20.2	22.1	27.6 (-15.1)
SCDA [44]	39.0	33.6	48.5	28.0	33.5	38.0	23.3	26.5	33.8 (-8.9)
SWDA [32]	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3 (-8.4)
DM [19]	38.4	32.2	44.3	28.4	30.8	40.5	34.5	27.2	34.6 (-8.1)
MTOR [1]	38.6	35.6	44.0	28.3	30.6	41.4	40.6	21.9	35.1 (-7.6)
MAF [14]	39.9	33.9	43.9	29.2	28.2	39.5	33.3	23.8	34.0 (-8.7)
iFAN [45]	45.5	33.0	48.5	22.8	32.6	40.0	31.7	27.9	35.3 (-7.4)
CRDA [42]	45.1	34.6	49.2	30.3	32.9	43.8	36.4	27.2	37.4 (-5.3)
HTCN [2]	47.4	37.1	47.9	32.3	33.2	47.5	40.9	31.6	39.8 (-2.9)
UMT [8]	**56.5**	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7 (-1.0)
AT	56.3	**51.9**	**64.2**	**38.5**	**45.5**	**55.1**	**54.3**	**35.0**	**50.9** (+8.2)
Oracle (F-RCNN)	50.3	40.7	61.3	32.5	43.1	49.8	35.1	28.6	42.7

4.2. Implementation Details

Following [3] and [32], we employ Faster RCNN [30] as the base detection model in our Adaptive Teachera and implement it using Detectron2. Either of the network ResNet-101 [13] or VGG16 [36] pre-trained on ImageNet [7] is used as the backbone according to the settings. Following the implementation of Faster RCNN with ROI-alignment [12], we scale all images by resizing the shorter side of the image to 600 while maintaining the image ratios. For the hyperparameter, we set the $\lambda_{unsup} = 1.0$ and $\lambda_{dis} = 0.1$ for all the experiments for simplicity. We set the confidence threshold as $\delta = 0.8$. For the initialization stage of training the framework described in Sec. 3.2, we train the AT using the source labels for 10k iterations. Then we copy the weights to both Teacher and Student models in the beginning of mutual learning and train the AT for 50k iterations. We set the learning rate as 0.04 during the entire training stage without applying any learning rate decay. We optimize the network using Stochastic Gradient Descent (SGD). The used data augmentation methods include random horizontal flip for weak augmentation , and randomly color jittering, grayscaling, Gaussian blurring, and cuting out patches are utilized for strong augmentations. The weight smooth coefficient parameter of the exponential moving average (EMA) for the teacher model is set to 0.9996. Each experiment is conducted on 8 Nvidia GPU V100 with the batch size of 16 and implemented in PyTorch.

4.3. Experimental Settings and Evaluation

We report the average precision (AP) of each class as well as the mean AP over all classes for object detection following existing works [3, 32] for all of the experimental settings, which are described as follows:

Real to Artistic Adaptation. To begin with, we would like to benchmark the effectiveness of our model for ad-

dressing the large domain gap. In this setting, we test our model with the effect of domain shift between the real images and the artistic images. We use Pascal VOC as the source dataset and the Clipart1k or Watercolor2k as the target dataset. The backbone of ResNet-101 [13] is used following the settings in existing works.

Adverse Weather Adaptation. For this setting, we evaluate our model on the domain shift between the image in normal weather and the image with adverse weather (foggy). The data from Cityscapes dataset is served as the source domain while the data from Foggy Cityscapes dataset is served as the target domain. The model is trained with the labeled images from Cityscapes and unlabeled images from Foggy Cityscapes in the experiments. The testing results on the validation set of Foggy Cityscapes are reported. Even though there are one-to-one mappings of images between Cityscapes and Foggy Cityscapes datasets, we do not use such information in all of the experiments. The backbone of VGG16 [36] is used following previous settings.

4.4. Results and Comparisons

In this section, we report the performance of our Adaptive Teacher and other state-of-the-art approaches in Table 1 and Table 3. We additionally report the source-only model denoted "Source (F-RCNN)" as training the base Faster RCNN model with only source images as the lower bound benchmark. On the other hand, we also include an oracle model denoted "Oracle (F-RCNN)" as training a base Faster RCNN model using the images from target domain and ground truth annotations, which can be viewed as the upper bound benchmark.

Real to Artistic Adaptation. The results of the setting: real to artistic adaptation on Clipart1k is presented in Table 1 and the one on Watercolor2k is presented in Table 2.

Table 4. The results of domain generalization on unseen target dataset, which leverages the labeled source data and another domain without supervision. The average precision (AP, %) is reported. The backbone is ResNet-101 for fair comparison. "WS Aug." indicates weak-strong augmentation.

Method	PASCAL VOC (sup.) & Watercolor2k (unsup.) → Clipart1k							PASCAL VOC (sup.) & Clipart1k (unsup.) → Watercolor2k						
	bicycle	bird	car	cat	dog	person	mAP	bicycle	bird	car	cat	dog	person	mAP
AT	78.6	30.1	40.3	10.9	32.6	72.8	44.5 (+1.1)	91.2	55.2	60.4	37.0	39.6	69.8	56.7 (-6.1)
MT [40] + WS Aug.	68.2	25.6	35.2	2.9	25.5	64.5	37.4 (-6.0)	82.1	49.0	55.6	29.5	25.4	66.2	50.4 (-0.2)
MT [40] + $\mathcal{L}_{dis}$	73.2	29.7	38.8	9.0	28.6	69.2	41.3 (-2.1)	84.3	51.2	58.7	34.2	24.3	62.4	51.1 (+0.5)
MT [40]	64.8	23.4	34.6	3.1	22.0	61.4	34.2 (-9.2)	80.5	43.4	53.0	27.6	19.5	55.6	47.6 (-3.0)
Oracle	47.6	39.1	51.4	20.1	38.4	69.7	43.4	51.8	49.7	42.5	38.7	52.1	68.6	50.6

We compare our method with several state-of-the-art approaches and report the performance gap between the oracle model (fully supervision) and each of the competitors. We observed that, first, our model achieves state-of-the-art performance at 49.3% mAP and outperforms the recent competitor UMT by 5.2% and other methods by a large margin. We note that, UMT using Mean Teacher already had significant performance improvement with augmented-styled training images. Yet, due to the inherent issue with the quality of pseudo labels in Mean Teacher on target domain, their model may also suffer large domain shift between real and artistic images when generate pseudo labels. On the other hand, our model mitigates the domain gap and achieve largely improved performance. Second, our model is the only one exceeding the oracle model on Clipart1k dataset, showing that the mutual learning adopted form Mean Teacher plus adversarial learning is capable to bridge the domain gap. Similar observations can be found on experiments conducted on Watercolor2k.

Adverse Weather Adaptation. The results of the setting: normal weather to adverse weather adaptation is presented in Table 3. We also report the performance gap between the oracle model (fully supervision) and each of the competitors. When comparing to the state of the arts, we can see that, first, our model also outperforms all of the state-of-the-art approaches by a large margin (more than 9%). Among these methods, MTOR [1] and UMT [8] are the two methods adopting Mean Teacher in their model. However, due to the problems discussed earlier regarding the augmentation in Teacher model and bias to source domain, both of their model suffer from generating noisy labels and lead to performance gap between our AT. Second, the performance of our model is able to exceed the oracle model by a large margin, which indicates that training the images with only the annotations from clear weather (high visibility) are sufficient to have satisfactory performance of the object detection on the adverse foggy weather (low visibility).

4.5. Experiments on domain generalization

As we observe that our AT outperforms all of the Oracle models on the three benchmark domain adaptation datasets,

Table 5. The ablation studies on our AT. We report mean average precision (mAP, %) on each of the experimental settings while "WS Aug." indicates weak-strong augmentation.

Source:	PASCAL VOC	PASCAL VOC	Cityscapes
Target:	Clipart1k	Watercolor2k	Foggy Cityscapes
AT	49.3	59.9	50.9
AT w/o $\mathcal{L}_{dis}$	40.6 (-8.7)	55.5 (-4.4)	48.7 (-2.2)
AT w/o WS Aug.	45.3 (-4.0)	55.1 (-4.8)	45.9 (-5.0)
AT w/o $\mathcal{L}_{unsup}$ & EMA	31.6 (-17.7)	50.2 (-9.7)	36.0 (-14.9)

we are more interested in the ability of generalization of our model on the unseen domains. We define such problem as domain generalization: Instead of focusing on the model's accuracy on the target domain, we further generalize the model to a completely unseen domain and evaluate it's generalization capability. In this section, we further conduct two experimental settings and compare our AT with the baseline model MT [40]:

- Train: PASCAL VOC (supervised) & Watercolor2k (unsupervised) → Test: Clipart1k

- Train: PASCAL VOC (supervised) & Clipart1k (unsupervised) → Test: Watercolor2k

In each of the setting, we train the model on the source real dataset with labels and another artistic dataset without labels. We then inference the model on the target dataset which is unseen during the training. We only train on the overlapped classes (6 classes) between the Clipart1k and Watercolor2k, and presented the results in the Table 4. From the table, we can observe that our model achieves superior performance comparing with the Oracle model and MT. This shows that our model is able to generalize to unseen domain without observing any target images. In addition, each of the ablation studies on either adversarial loss or augmentation on MT also show the importance of their roles in our proposed AT.

4.6. Ablation studies

We further conduct ablation studies on each of important components in Table 5 and also present the qualitative studies of pseudo labels in Figure 3.
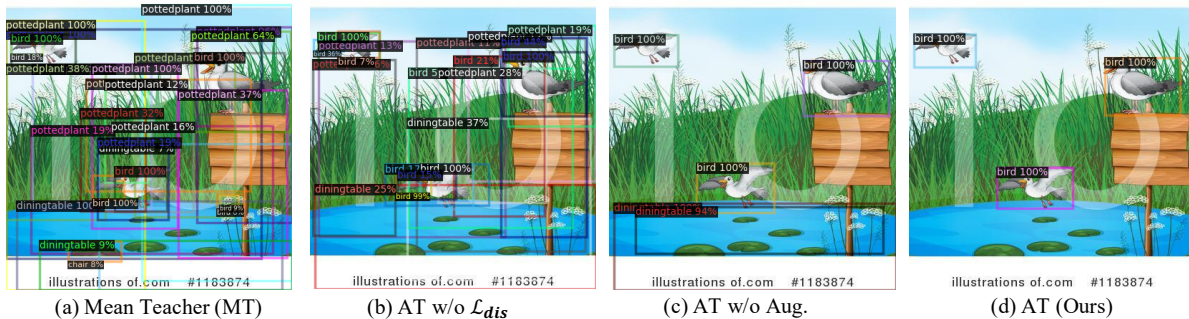
(a) Mean Teacher (MT)	(b) AT w/o $\mathcal{L}_{dis}$	(c) AT w/o Aug.	(d) AT (Ours)

Figure 3. **Qualitative ablation studies on pseudo labels generated on the image from the training set of Clipart1k**. This figure show the importance of adversarial loss $\mathcal{L}_{dis}$ and weak-strong augmentation on pseudo labeling. Note that the thresholding is not applied here.
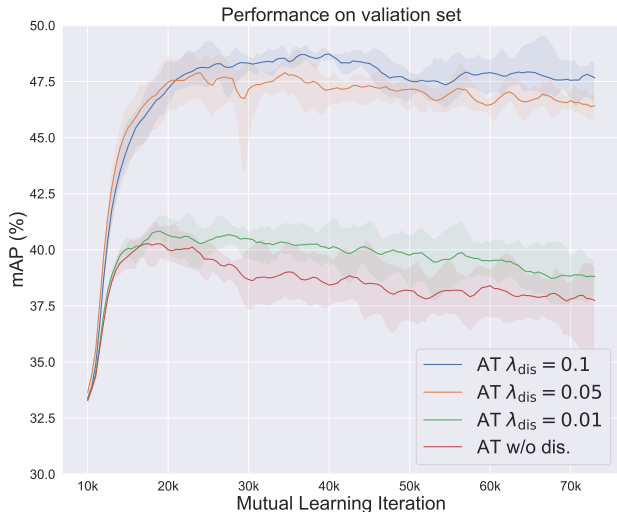

Figure 4. **Mutual Learning curve on Clipart1k dataset**. We run 5 identical experiments for each settings and plot the error bound accordingly in the figure. The results show that increasing weights of λ_{dis} can achieve improved performance and stable learning curve.

Adversarial loss $\mathcal{L}_{dis}$. To further analyze the importance of adversarial learning in our Adaptive Teacher, we exclude the loss $\mathcal{L}_{dis}$ in discriminator and report the performance on three experimental settings in Table 5. It can be observed that the 8.7% and 4.4% performance drop appears on Clipart1k and Watercolor2k in the scenario with larger domain gap (real to artistic adaptation). Yet, in another scenario with smaller domain gap (weather adaptation), only 2.2% performance drop is observed. We can also observe that $\mathcal{L}_{dis}$ is able to largely reduce the ratio of false positives in pseudo labels generated by the Teacher model in Figure 3. On the other hand, we also analyze the weight λ_{dis} of the adversarial loss $\mathcal{L}_{dis}$ in in Figure 4. Some phenomenons can be observed in this figure in two folds. We can see that, first, increasing weights can lead to improved performance, which supports the effectiveness of the discriminator in our model. Second, without applying the adversarial loss, the performance of the model keeps dropping due to the error

propagation coming from the noisy pseudo labels.

Augmentation pipeline. We also benchmarked the effectiveness of weak-strong (WS) augmentation in our Adaptive Teacher, and around 4% to 5% performance drop is observed when it is excluded (Table 5). This demonstrates that the simple modification on the training pipeline (weak and strong augmentation for Teacher and Student, respectively) is vital. We can also observe that such augmentation pipeline is able to reduce the ratio of false positives in pseudo labels generated by the Teacher model in Figure 3.

$\mathcal{L}_{unsup}$ **& EMA.** Similarly, we ablated the importance of utilizing Mean Teacher in Table 5 as previous works (*i.e.,* excluding the mutual learning and the Teacher model from our model) and report the performance of the Student model for cross-domain training with only strong augmentation and adversarial loss $\mathcal{L}_{dis}$. We can see that there is a significant performance drop, thus the performance gain mainly came from the mutual learning with pseudo labels on target domain.

5. Conclusion

In this paper, we proposed an novel framework to address the task of cross-domain object detection. With the introduced target-domain Teacher model and cross-domain Student model, the framework is able to generate correct pseudo labels on the target domain via mutual learning. Our design of training pipeline with proper augmentation strategies and adversarial learning also resolve the bias toward source domain in both Teacher and Student model. The experiments on two benchmarks confirmed the effectiveness and superiority of our model for cross-domain object detection. The extensive experiments of ablation studies also demonstrated our proposed model trained without seeing both labels nor images on target domain outperform the Oracle model which is trained with fully supervision.

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, 3, 5, 6, 7

[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. 1, 3, 5, 6

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 3, 6

[4] 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. 5

[5] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NeurIPS*, pages 379–387, 2016. 2

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6

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

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 5

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1180–1189. PMLR, 2015. 1, 2, 3, 4

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[14] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019. 1, 3, 6

[15] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018. 2

[16] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 5

[17] 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. 3

[18] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6092–6101, 2019. 3

[19] 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. 5, 6

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 2

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

[22] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. 2

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 97–105. PMLR, 2015. 2

[24] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, 2016. 2

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2208–2217. PMLR, 2017. 2

[26] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *CVPR*, pages 6181–6189, 2018. 2

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

[28] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017. 2

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, volume 28, pages 91–99, 2015. 2, 6

[31] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *CVPR*, pages 780–790, 2019. 3

[32] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 3, 5, 6

[33] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 5

[34] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*, pages 8503–8512, 2018. 2

[35] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 5, 6

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014. 6

[37] Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. R-fcn-3000 at 30fps: Decoupling detection and classification. In *CVPR*, pages 1081–1090, 2018. 2

[38] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 2021. 3

[39] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *ECCV*, pages 403–419. Springer, 2020. 1, 3

[40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1, 2, 3, 4, 7

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[42] 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. 1, 3, 5, 6

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2, 3

[44] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 1, 3, 6

[45] 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. 6