Adaptive Unbiased Teacher for Cross-Domain Object Detection

Anonymous authors
Paper under double-blind review

Abstract

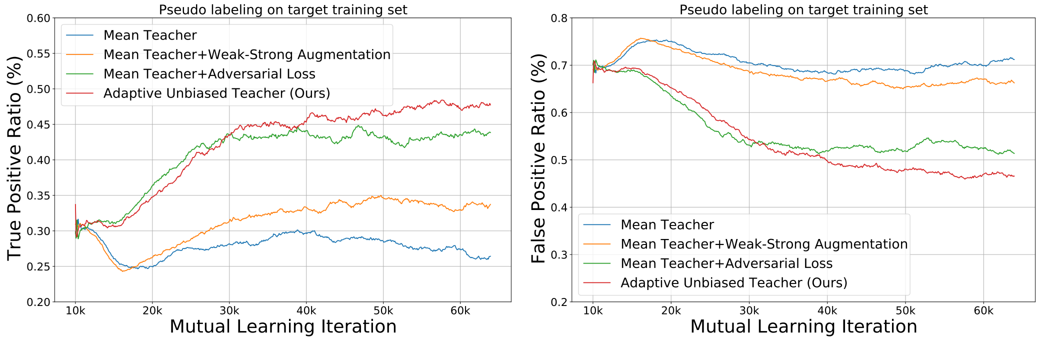
We tackle the problem of domain adaptation in object detection, where the main challenge lies in significant domain shifts between source (one domain with supervision) and target (a domain of interest without supervision). Although the teacher-student framework (a student model learns from pseudo labels generated from a teacher model) has been adopted to enable domain adaptation and yielded accuracy gains on the target domain, the teacher model still generates a large number of low-quality pseudo labels (*e.g.,* false positives) due to its bias toward source domain. This leads to sub-optimal domain adaptation performance. To address this issue, we propose Adaptive Unbiased Teacher (AUT), a teacher-student framework leveraging adversarial learning (on features derived from backbone) and weak-strong data augmentation to address domain shifts. Specifically, we employ feature-level adversarial training, ensuring features extracted from the source and target domains share similar statistics. This enables the student model to capture domain-invariant features. Furthermore, we apply weak-strong augmentation and mutual learning of the teacher for target domain and student model for both domains. This enables the updated teacher model to gradually benefit from the student model without suffering domain shift. We show that AUT demonstrates superiority over all existing approaches and even Oracle (fully-supervised) models by a huge 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 arts and Oracle, respectively.

1 Introduction

As the task of object detection is bringing pervasive impact on various real-world applications, collecting large-scale and diverse datasets with bounding box annotations still remains challenging since the labeling process is usually labor intensive. Thus, 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 increasingly important. Researchers have proposed various methods, such as domain classifier and adversarial learning, to address the task of cross-domain adaptation (Chen et al., 2018a; Zhu et al., 2019; Saito et al., 2019; He & Zhang, 2019; Xu et al., 2020; Chen et al., 2020; Su et al., 2020) (Ganin & Lempitsky, 2015). Even though these methods have led to accuracy improvement on the target domain, there is generally still a large performance gap from the Oracle model (fully supervision).

To further improve the model's performance after domain adaptation, researchers start to exploit and extend teacher-student self-training method from semi-supervised learning to domain adaptation (Tarvainen & Valpola, 2017). These approaches typically involves a teacher model to generate pseudo labels thus to allow a student model to learn without annotations. These methods have led to notable accuracy gains in the domain adaptation scenario. For example, MTOR (Cai et al., 2019) adapts the Mean Teacher (MT) (Tarvainen & Valpola, 2017) to explore object relation in region-level consistency, inter-graph consistency, and intra-graph consistency. Unbiased Mean Teacher (UMT) (Deng et al., 2021) proposed to augment the teacher-student framework with CycleGAN (Zhu et al., 2017) 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



(a) True positive ratio on pseudo labels

(b) False positive ratio on pseudo labels

Figure 1: **The effectiveness of domain loss and weak-strong augmentation on pseudo labeling by Teacher model**. The figure shows the true positive and false positive ratio on the entire training set of Clipart1k (target) with PASCAL VOC as source. Due to inherent domain shift in the Teacher model, the Teacher model without domain loss generates noisy pseudo labels. The weak augmentation is able to stabilize pseudo labeling.

teacher model usually contains a substantial amount of errors and false positives, as shown in Figure 1 (b). This is because the scenario of domain adaptation typically involve 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 on the source domain, hence unable to provide high-quality pseudo labels in the target domain. As a result, direct applying the teacher-student framework only leads to sub-optimal adaptation performance.

To address this problem, we propose a framework named Adaptive Unbiased Teacher (AUT) to mitigate the domain shift and improve the pseudo labeling quality on 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 target domain into the Teacher model. 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 true positive ratio is improved by up to 1.8× compared to the vanilla teacher-student framework. Meanwhile, 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 framework based on Mean Teacher and leverage 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 able to outperformed 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 (fully supervision).

2 Related Works

Object Detection. Recent years have witnessed remarkable progress in object detection with deep learning such as a series of two-stage paradigm: R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), and Faster R-CNN Ren et al. (2015) which advances selective search with an accurate and efficient Region Proposal Networks (RPN). Next, a few sub-sequent works (Dai et al., 2016; 2017; Hu et al., 2018; Lin et al., 2017a; Peng et al., 2018; Singh et al., 2018) strive to improve the accuracy and speed of two-stage detectors. Another line of works builds detectors in one-stage manner by skipping region proposal stage. YOLO (Redmon et al., 2016) jointly predicts bounding boxes and confidences of multiple categories as regression problem. SSD (Liu et al., 2016) further improves it by utilizing multiple feature maps at different scales. Numerous extensions (Fu et al., 2017; Lin et al., 2017b; Redmon & Farhadi, 2017; 2018) to the one-stage scheme have been proposed. In this work, we adopt Faster R-CNN as the detection backbone for its robustness and flexibility.

Unsupervised Domain Adaptation. As for the literature on domain adaptation, while it is quite vast, the most relevant category to our work is unsupervised domain adaptation in deep architectures. Recent works have involved discrepancy-based methods that guide the feature learning by minimizing the domain discrepancy with Maximum Mean Discrepancy (MMD) (Long et al., 2015; 2017; 2016). Another branch is to exploit the domain confusion by learning a domain discriminator (Ganin & Lempitsky, 2015; Ganin et al., 2016; Tzeng et al., 2017; Sankaranarayanan et al., 2018). Later, self-ensembling (French et al., 2017) extends Mean Teacher (Tarvainen & Valpola, 2017) for domain adaptation and establishes new records on several cross-domain recognition benchmarks. All of the aforementioned works focus on the domain adaptation for recognition, and recently much attention has been paid to domain adaptation in other tasks, e.g., semantic segmentation (Chen et al., 2018b; Hoffman et al., 2016; Zhang et al., 2018). However, the output of object detection is richer and more complex, consisting of both the class labels and the bounding box locations. Comparing with the above vision tasks, we aim at handling the more challenging task of cross-domain object detection.

Cross-domain Object Detection. Recently, many works have been proposed to address cross-domain object detection with different techniques. Several approaches utilize adversarial learning (discriminator) with a gradient reverse layer to obtain domain-invariant feature in (Chen et al., 2018a; Zhu et al., 2019; Saito et al., 2019; He & Zhang, 2019; Xu et al., 2020; Chen et al., 2020; Su et al., 2020). The annotation-level adaptation (Khodabandeh et al., 2019; Kim et al., 2019a; RoyChowdhury et al., 2019) has also been proposed for the task. Recently, another direction is to utilize Mean Teacher (MT) (Tarvainen & Valpola, 2017) which is originally proposed for semi-supervised learning on this task. MTOR (Cai et al., 2019) performs MT to explore object relation in region-level consistency, inter-graph consistency and intra-graph consistency. Similarly, Unbiased Mean Teacher (UMT) (Deng et al., 2021) has been proposed to reduce the domain shift by augmenting the training samples with CycleGAN (Zhu et al., 2017). However, the above approaches are likely to suffer the inherent issue in Mean Teacher (MT), generating pseudo labels of low quality on target domain.

3 Adaptive Unbiased Teacher

3.1 Problem Formulation and Overview

For the problem of unsupervised domain adaptation in object detection, 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 $\mathcal{D}_t$.

We show an overview of our framework in Figure 2. Our Adaptive Unbiased Teacher 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. To begin with, we simply train the object detector using the available source data to initialize the feature encoder and detector. At the stage of mutual learning (Sec. 3.2), we duplicate the initialized detector into two models (Teacher and Student models). Then 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, for the adaptive learning (Sec. 3.3), the discriminator and the Gradient Reverse Layer (GRL) (Ganin

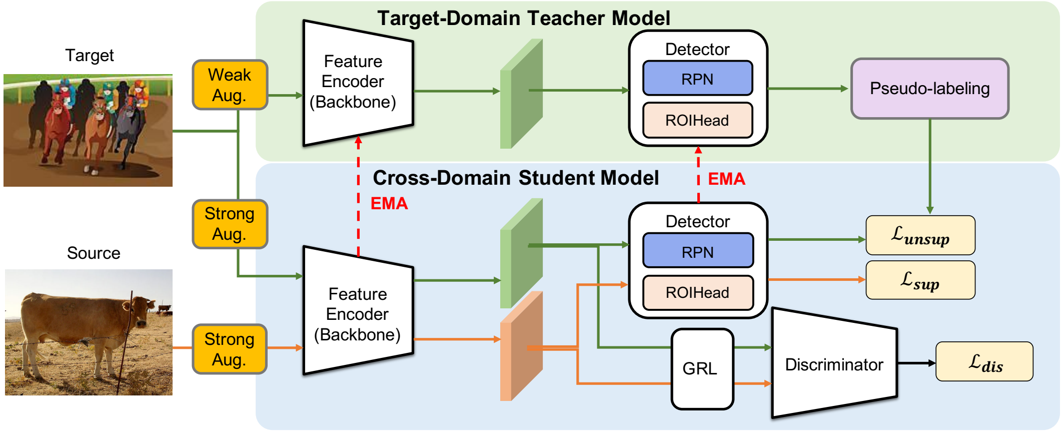


Figure 2: **Overview of our proposed Adaptive Unbiased Teacher (AUT).** Our model consists of two modules: 1) target-specific Teacher model for taking weakly-augmented images from target domain and 2) cross-domain Student model for taking strongly-augmented images from both domains. We train our model using two training streams, which are the Teacher-Student mutual learning and the adversarial learning. The Teacher generates pseudo-labels to train the Student while the Student updates the Teacher via exponential moving average (EMA). The discriminator with GRL is employed to align the distributions across two domains in Student model.

& Lempitsky, 2015) are employed to align the distributions across two domains in Student model via adversarial learning process. 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 also consists of two models with identical architecture: a Student model and a Teacher model. The Student model is trained using standard gradient back-propagation algorithm, and the Teacher model is updated with the exponential moving average (EMA) weights of the student model. To generate precise and accurate pseudo labels for target domain images, we feed the images with **weakly augmentation** as input of the Teacher to provide reliable pseudo-labels while images with **strongly augmentation** as inputs of the Student.

Initialization. It is significant to have a good initialization for both Student and Teacher models since we rely on the Teacher to generate pseudo-labels for target domain to train the Student. 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 $\mathcal{L}_{sup}$. Therefore the loss for training the student model with the labeled source samples can be written as:

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

where RPN loss $\mathcal{L}^{rpn}$ is for the Region Proposal Network (RPN) module which is used for the candidate proposals generation, and ROI loss $\mathcal{L}^{roi}$ is for the prediction branch which performs bounding box regression and classification.

Optimize Student with Pseudo-Labeling. To address the lack of ground-truth labels for data in target domain, we adapt the pseudo-labeling method to generate labels for training the Student with images from target domain. To prevent the consecutively detrimental effect of noisy pseudo-labels, we set a confidence threshold δ of predicted bounding boxes to filter predicted bounding boxes of low-confidence, which are possibly the false positive samples. In addition, we remove duplicated boxes prediction by applying class-wise non-maximum suppression (NMS). After obtaining the pseudo-labels from Teacher model on the images of target domain, we can construct an unsupervised loss on Student model 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 $\hat{\mathcal{C}}_t$ denotes the pseudo labels generated by the Teacher model on target domain. Not that, following (Liu et al., 2021), we do not apply unsupervised losses for the bounding box regression since the confidence thresholding would not be able to filter the pseudo-labels that are potentially incorrect for bounding box regression. Meanwhile, the target samples are augmented with random horizontal flip and crop for weak augmentation in Teacher model and randomly color jittering, grayscale, Gaussian blur, and cutout patches for strong augmentations as perturbations.

Update Teacher via Exponential Moving Average. To obtain more stable pseudo-labels from the target images following Mean Teacher (MT) (Tarvainen & Valpola, 2017), we apply Exponential Moving Average (EMA) to gradually update the Teacher model. The update can be written 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

However, 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 unless the Teacher model would generate noisy labels on target images and make the learning process collapse. Since we found that adversarial loss is able to reduce the false positive ratio on Teacher model in Figure 1, we introduce adversarial learning into the framework for aligning the distributions across two domains.

Since only Student model takes images from both domains, the adversarial loss is applicable on Student model. To achieve adversarial learning, we place a domain discriminator, denoting D, after the feature encoder, denoting E (shown in Figure 2) on the Student model. The main objective of the discriminator is to discriminate whether the feature $E(X)$ is from the source or the target domain. Through this discriminator, the probability of each sample belonging to the target domain is obtained as $D(E(X))$. We then apply a binary cross-entropy loss to $D(E(X))$ based on the domain label d of the input image, where images from the source distributions are given the label $d = 0$ and the target images receive label $d = 1$. The discriminator loss $\mathcal{L}_{dis}$ can be formulated as:

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

Furthermore, adversarial learning in the discriminator is achieved using the Gradient Reverse Layer (GRL) (Ganin & Lempitsky, 2015) to produce reverse gradient on the feature encoder to learn the domain-invariant feature $E(X)$. That is, GRL is placed in between the discriminator and the detection network. During back propagation, GRL negates the gradients that flow through and the feature encoder E receives gradients that force it to update in an opposite direction which maximizes the discriminator loss. This allows E to produce features that fools the discriminator D while D tries to distinguish the domain of the features. Hence, the min-max loss function of the adaptive detection model is defined as the following:

$$\mathcal{L}_{adv} = \min_E \max_D \mathcal{L}_{dis}. \tag{5}$$

With the above domain 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 Unbiased 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 (Cai et al., 2019; Deng et al., 2021) 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.

Table 1: The average precision (AP, in %) on all classes from different methods for cross-domain object detection on the Clipart1k test set for **PASCAL VOC** → **Clipart1k** adaptation. The backbone is ResNet-101. We compare our method with SCL (Shen et al., 2019), SWDA (Saito et al., 2019), DM (Kim et al., 2019b), CRDA (Xu et al., 2020),HTCN (Chen et al., 2020), UMT (Deng et al., 2021), 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)
AUT	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

3.4 Full Objective and Inference

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

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{unsup}} \cdot \mathcal{L}_{\text{unsup}} + \lambda_{\text{dis}} \cdot \mathcal{L}_{\text{adv}}, \tag{6}$$

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 $\mathcal{L}_{\text{adv}}$ is introduced to update the feature encoder and discriminator. The Teacher model is only updated through EMA discussed in the Sec 3.2.

With the interaction between the Teacher and the Student, both models can evolve jointly and continuously to improve detection accuracy. With the improvement on detection accuracy, this also means that the Teacher generates more accurate and stable pseudo-label for target domain. In another perspective, we can also regard the Teacher as the temporal ensemble of the cross-domain Student models in different time steps, which aligns the observation that the accuracy of the Teacher on target domain is consistently higher than the Student (noted in (Liu et al., 2021)). As the result, during the inference stage we only keep the target-specific Teacher model for evaluating on the target testing dataset.

4 Experiment

4.1 Datasets

We conduct our experiments on five public datasets, including Cityscapes (Cordts et al., 2016)), Foggy Cityscapes (Sakaridis et al., 2018), PASCAL VOC (Everingham et al., 2010), Clipart1k (Inoue et al., 2018), and Watercolor2k (Inoue et al., 2018).

Cityscapes. Cityscapes (Cordts et al., 2016) focuses on capturing high variability of outdoor street scenes in common weather conditions from different cities. It contains 2,975 training images and 500 validation images with dense pixel-level labels. We transform the instance segmentation annotations into bounding boxes for our experiments.

Foggy Cityscapes. Foggy Cityscapes (Sakaridis et al., 2018) is built upon the images in the Cityscapes. This dataset simulates the foggy weather using depth maps provided in Cityscapes with three levels of foggy weather, and thus is suitable to conduct weather adaptation experiments.

PASCAL VOC. PASCAL VOC (Everingham et al., 2010) is a real-world dataset containing 20 categories of common objects with bounding box annotations. Following (Saito et al., 2019; Shen et al., 2019), we employ PASCAL VOC 2007 and 2012 training and validation images (16,551 images in total) for experiments.

Clipart1k. Clipart1k (Inoue et al., 2018) contains 1k clipart images, which shares the same instance categories with PASCAL VOC but exhibits a large domain shift. We follow the practice in Saito et al. (2019); Shen et al. (2019) and split it into training and test sets, containing 500 images each.

Watercolor2k. Watercolor2k (Inoue et al., 2018) contains 2k watercolor images, which consists of 2,000 images from 6 classes in common with the PASCAL VOC dataset. Following the practice

Table 2: The average precision (AP, in %) on all classes from different methods for cross-domain object detection on the Watercolor2k test set for **PASCAL VOC** → **Watercolor2k** adaptation. The backbone is ResNet-101.

Method	bicycle	bird	car	cat	dog	person	mAP
Source (F-RCNN)	84.2	44.5	53.0	24.9	18.8	56.3	46.9 (-3.7)
SCL (Shen et al., 2019)	82.2	55.1	51.8	39.6	38.4	64.0	55.2 (+4.8)
SWDA (Saito et al., 2019)	82.3	55.9	46.5	32.7	35.5	66.7	53.3 (+2.7)
UMT (Deng et al., 2021)	88.2	55.3	51.7	**39.8**	**43.6**	69.9	58.1 (+7.5)
AUT (Ours)	**93.6**	**56.1**	**58.9**	37.3	39.6	**73.8**	**59.9** (+9.3)
Oracle (F-RCNN)	51.8	49.7	42.5	38.7	52.1	68.6	50.6

in Saito et al. (2019); Shen et al. (2019), we split it into training set and test sets, containing 1000 images each.

4.2 Implementation Details

Following Chen et al. (2018a) and Saito et al. (2019), we take the Faster RCNN (Ren et al., 2015) as the base detection model in our Adaptive Unbiased Teacher. The ResNet-101 (He et al., 2016) or VGG16 (Simonyan & Zisserman, 2014) model pre-trained on ImageNet (Deng et al., 2009) is used as the backbone. Following the implementation of Faster RCNN with ROI-alignment (He et al., 2017), we rescale all images by setting the shorter side of the image to 600 while keeping the image aspect ratios. For the hyperparameter, we set the $\lambda_{unsup} = 1.0$ and $\lambda_{dis} = 0.1$ for all the experiments. We set the confidence threshold as $\delta = 0.8$. During the initialization stage described in Sec. 3.2, we train the Faster RCNN 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 adaptive unbiased teacher 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). For the data augmentation. We apply random horizontal flip for weak augmentation and randomly add color jittering, grayscale, Gaussian blur, and cutout patches 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 (Chen et al., 2018a; Saito et al., 2019) 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 addressing the large domain gap. In this setting, we test our model with domain shift between the real image domain and the artistic image domain. We utilize Pascal VOC as the real source domain and the Clipart1k or Watercolor2k as the target domain. The backbone of ResNet-101 (He et al., 2016) is used following existing settings.

Weather Adaptation. In this setting, we test our model with domain shift between the image in normal weather and the image with adverse weather (foggy). The training set of the Cityscapes dataset is used as the source domain while the Foggy Cityscapes dataset is used as the target domain. We take labeled Cityscapes train set images and unlabeled Foggy Cityscapes train set images in our experiment and report the evaluated results on the validation set of Foggy Cityscapes. Although there exists a one-to-one correspondence between images in Cityscapes and Foggy Cityscapes datasets, we do not leverage such information in all of the experiments. The backbone of VGG16 (Simonyan & Zisserman, 2014) is used following previous settings.

Table 3: The average precision (AP, in %) on all classes from different methods for cross-domain object detection on the Foggy Cityscapes test set for **Cityscapes** → **Foggy Cityscapes** adaptation. The backbone is VGG-16.

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 (Shen et al., 2019)	41.8	36.2	44.8	33.6	31.6	44.0	40.7	30.4	37.9 (-4.8)
DA-Faster (Chen et al., 2018a)	35.3	27.1	40.5	20.0	25.0	31.0	20.2	22.1	27.6 (-15.1)
SCDA (Zhu et al., 2019)	39.0	33.6	48.5	28.0	33.5	38.0	23.3	26.5	33.8 (-8.9)
SWDA (Saito et al., 2019)	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3 (-8.4)
DM (Kim et al., 2019b)	38.4	32.2	44.3	28.4	30.8	40.5	34.5	27.2	34.6 (-8.1)
MTOR (Cai et al., 2019)	38.6	35.6	44.0	28.3	30.6	41.4	40.6	21.9	35.1 (-7.6)
MAF (He & Zhang, 2019)	39.9	33.9	43.9	29.2	28.2	39.5	33.3	23.8	34.0 (-8.7)
iFAN (Zhuang et al., 2020)	45.5	33.0	48.5	22.8	32.6	40.0	31.7	27.9	35.3 (-7.4)
CRDA (Xu et al., 2020)	45.1	34.6	49.2	30.3	32.9	43.8	36.4	27.2	37.4 (-5.3)
HTCN (Chen et al., 2020)	47.4	37.1	47.9	32.3	33.2	47.5	40.9	31.6	39.8 (-2.9)
UMT (Deng et al., 2021)	**56.5**	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7 (-1.0)
AUT (Ours)	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

4.4 Results and Comparisons

In this section, we report the performance of our Adaptive Unbiased 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)" by training a Faster RCNN model using only source images as the lower bound adaptation. On the other hand, we also include an oracle model denoted "Oracle (F-RCNN)" by training a Faster RCNN model using the same images with target domain but with the ground truth annotations, which can be viewed as a reference for the upper bound adaptation performance.

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. 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.

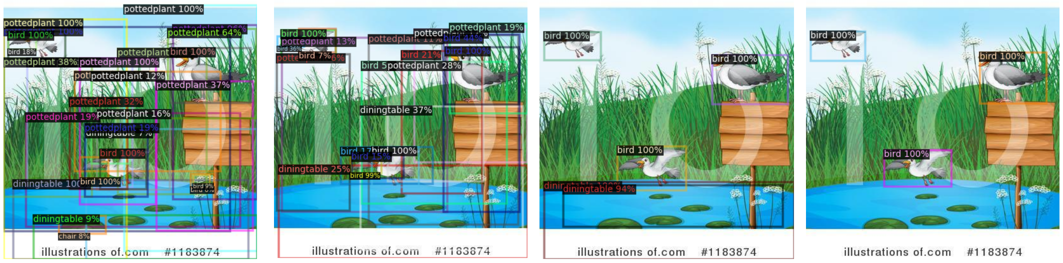
Weather Adaptation. The results of the setting: normal weather to adverse weather adaptation is presented in Table 3. We also report the performance gap between the oracle model (fully supervision) and each of the competitors. When comparing to the state of the arts, we can see that, first, our model also outperforms all of the state-of-the-art approaches by a large margin (more than 9%). Among these methods, MTOR (Cai et al., 2019) and UMT (Deng et al., 2021) 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 adaptive unbiased teacher. Second, the performance of our model also exceeds the oracle model by a large margin, showing that the clear weather images with high visibility are useful for boosting the limitation of the object detection in the adverse foggy weather with low visibility, without requiring any annotations on those low visibility images.

4.5 Ablation Studies

We further conduct ablation studies on each of important componenets in Table 4 and also present the qualitative studies of pseudo labels in Figure 3.

Table 4: The ablation studies on each of the losses and modules. We report mean average precision (mAP, in %) on each of the experimental datasets.

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



(a) Mean Teacher (MT) (b) AUT w/o $\mathcal{L}_{dis}$ (c) AUT w/o Aug. (d) AUT (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.

Adversarial loss $\mathcal{L}_{dis}$. To further analyze the importance of adversarial learning in our Adaptive Unbiased Teacher, we exclude the loss $\mathcal{L}_{dis}$ in discriminator and report the performance on three experimental settings. 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.

Augmentation pipeline. We also benchmarked the effectiveness of weak-strong augmentation in our Adaptive Unbiased Teacher, and around 4% to 5% performance drop is observed when it is excluded (Table 4). 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 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 without supervision on target domain outperform the Oracle model with fully supervision.

REFERENCES

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11457–11466, 2019.

Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8869–8878, 2020.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3339–3348, 2018a.

Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7892–7901, 2018b.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3213–3223, 2016.

Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Adv. Neural Inform. Process. Syst.*, pp. 379–387, 2016.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Int. Conf. Comput. Vis.*, pp. 764–773, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 248–255. Ieee, 2009.

Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4091–4101, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. pp. 1180–1189. PMLR, 2015.

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.

Ross Girshick. Fast r-cnn. In *Int. Conf. Comput. Vis.*, pp. 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 580–587, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pp. 2961–2969, 2017.

Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Int. Conf. Comput. Vis.*, pp. 6668–6677, 2019.

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3588–3597, 2018.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5001–5009, 2018.

Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Int. Conf. Comput. Vis.*, pp. 480–490, 2019.

Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Int. Conf. Comput. Vis.*, pp. 6092–6101, 2019a.

Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12456–12465, 2019b.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, pp. 2980–2988, 2017b.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Eur. Conf. Comput. Vis.*, pp. 21–37. Springer, 2016.

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 *Int. Conf. Learn. Represent.*, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. pp. 97–105. PMLR, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Adv. Neural Inform. Process. Syst.*, 2016.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. pp. 2208–2217. PMLR, 2017.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6181–6189, 2018.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7263–7271, 2017.

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

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, volume 28, pp. 91–99, 2015.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 780–790, 2019.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6956–6965, 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.*, 126(9):973–992, 2018.

Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8503–8512, 2018.

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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2014.

Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. R-fcn-3000 at 30fps: Decoupling detection and classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1081–1090, 2018.

Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *Eur. Conf. Comput. Vis.*, pp. 403–419. Springer, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Adv. Neural Inform. Process. Syst.*, 2017.

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*, pp. 7167–7176, 2017.

Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11724–11733, 2020.

Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6810–6818, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, pp. 2223–2232, 2017.

Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 687–696, 2019.

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

A APPENDIX

A.1 QUALITATIVE COMPARISON

To further justify the effectiveness of our propose adaptive unbiased teacher (AUT), we present the qualitative results of object detection on the testing set in Figure 4, Figure 5, and Figure 6. We compare our AUT with Source (F-RCNN) model and Oracle (F-RCNN) model. For the setting of weather adaptation, we can observe that there are many missing detections on source model when the fog is serious. Yet, our model is able to achieve the comparable detection result than Oracle model, which demonstrates our model is able to handle domain bias successfully without supervision on target domain. For the setting of real to artistic adaptation, both the Source (F-RCNN) model and Oracle (F-RCNN) model shows missing or incorrect detections in the two given samples. Our AUT instead produce more correct detected bounding boxes than both of the models.

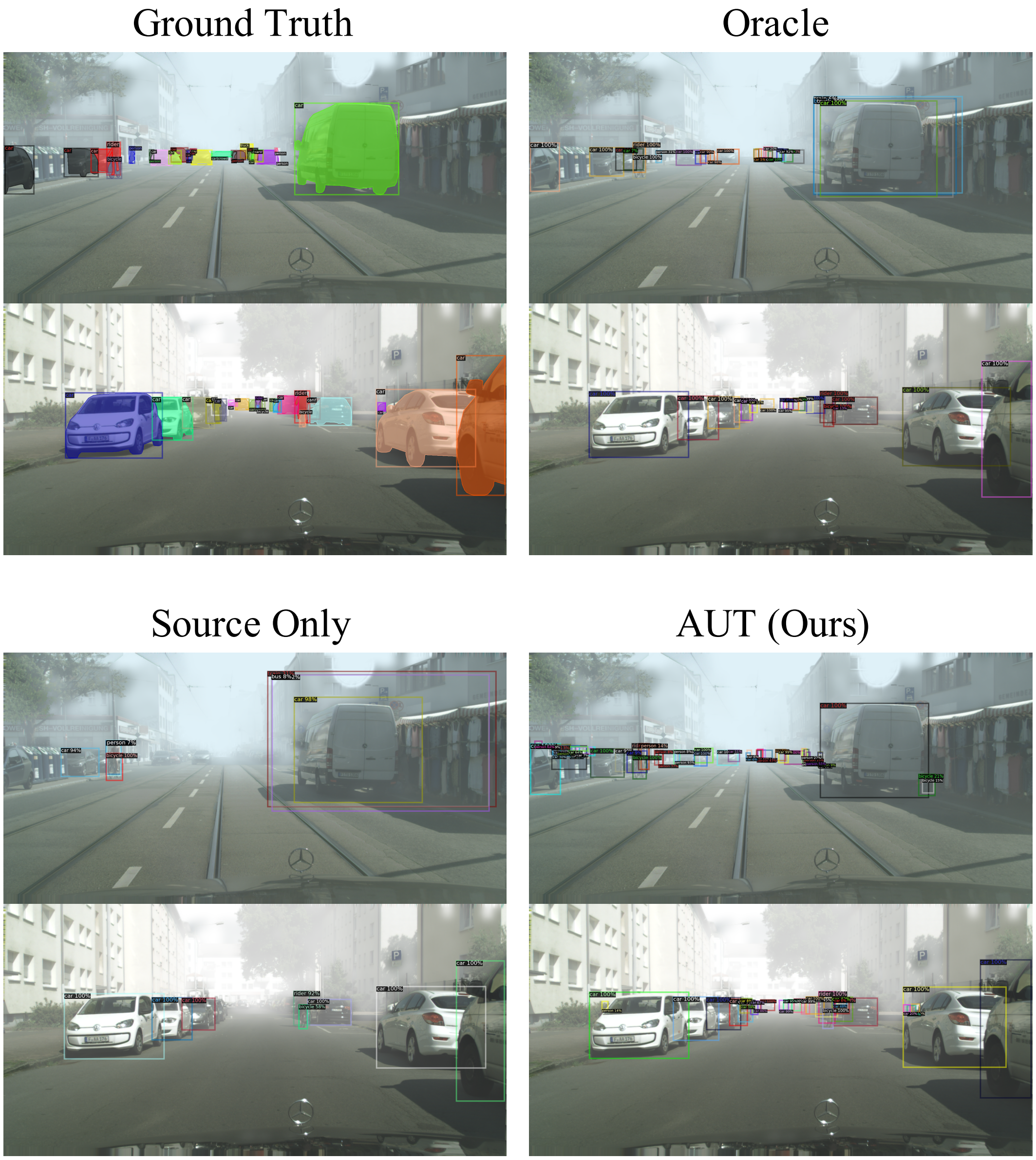


Figure 4: **Qualitative results and comparisons on Foggy Cityscapes**.

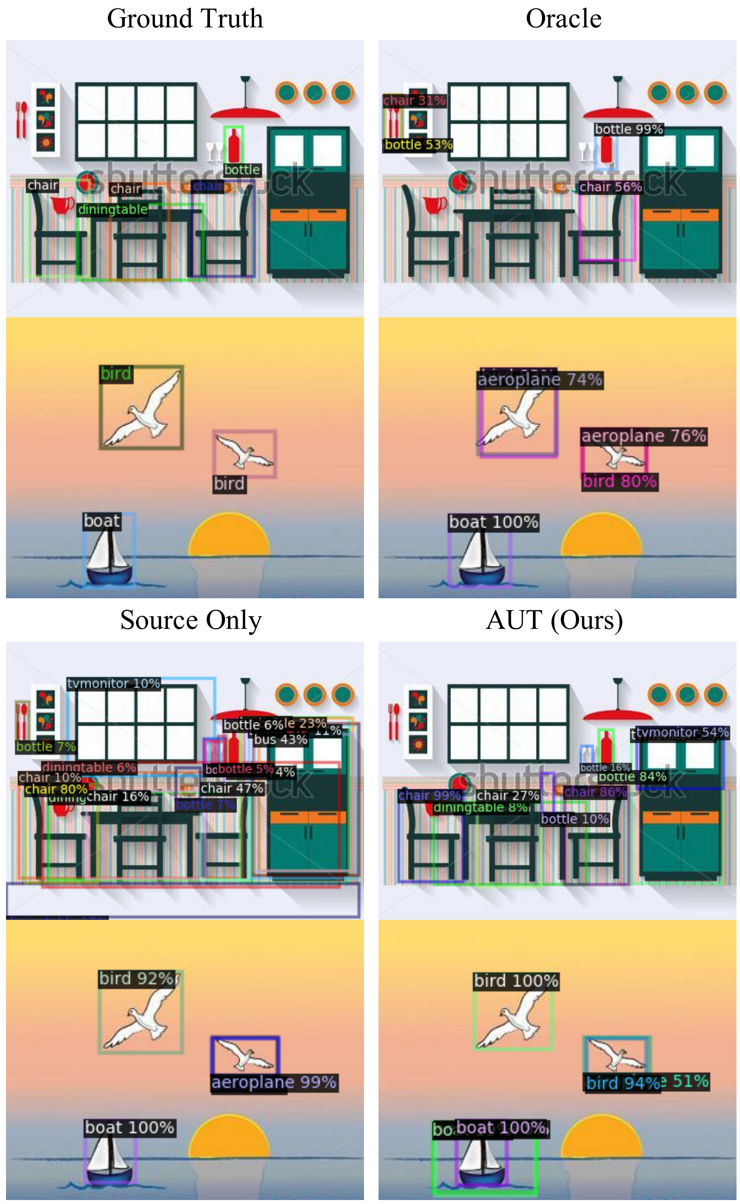


Figure 5: **Qualitative results and comparisons on Clipart1k**.

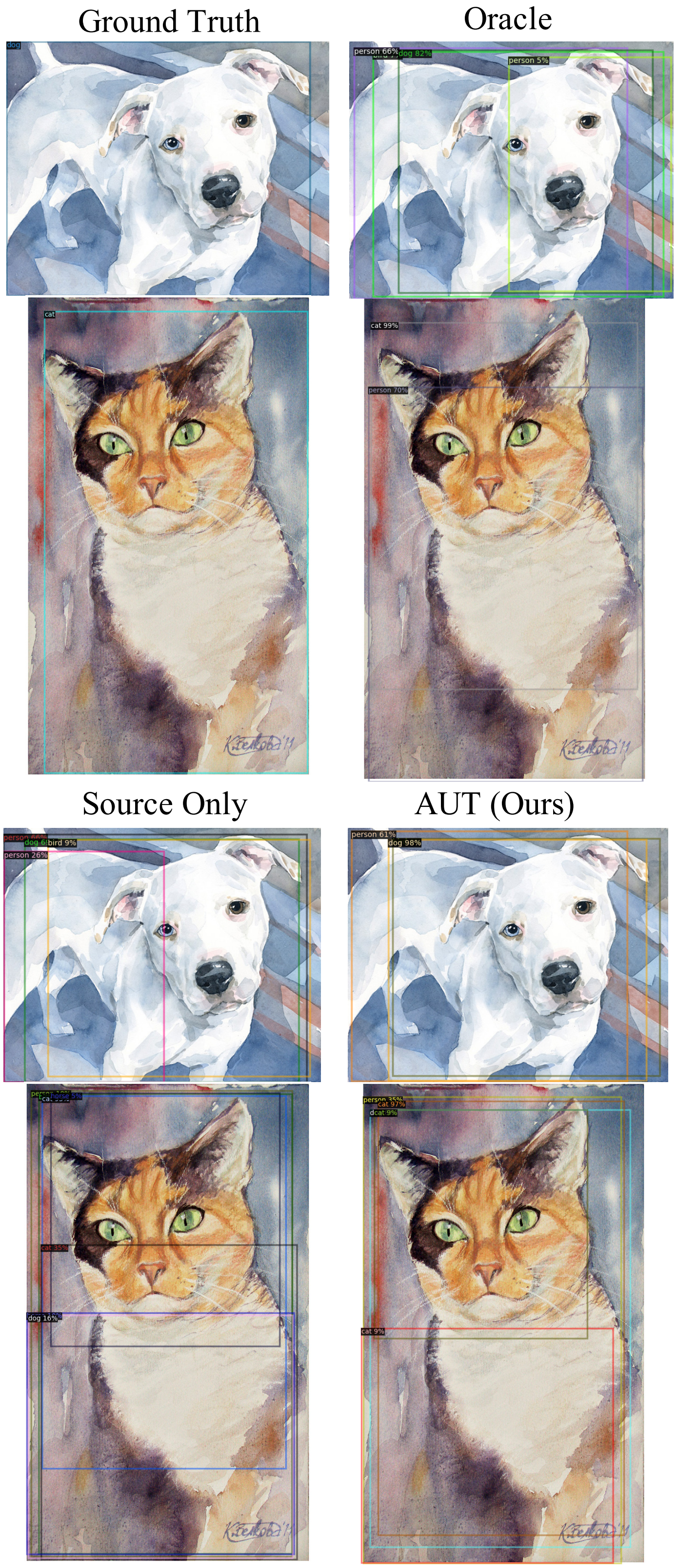


Figure 6: **Qualitative results and comparisons on Watercolor2k**.

A.2 More ablation studies

We also analyze the weight λ_{dis} of the adversarial loss $\mathcal{L}_{dis}$ in in Figure 7. We can see that, first, increasing weights can lead to improved performance, which supports the effectiveness of the discriminator in our model. Second, we see that with a sufficient weight of discriminator loss, the model will be able to have satisfactory performance. Third, the performance with the adversarial loss $\mathcal{L}_{dis}$ shows to be stable without decreasing drastically as the one without the loss.

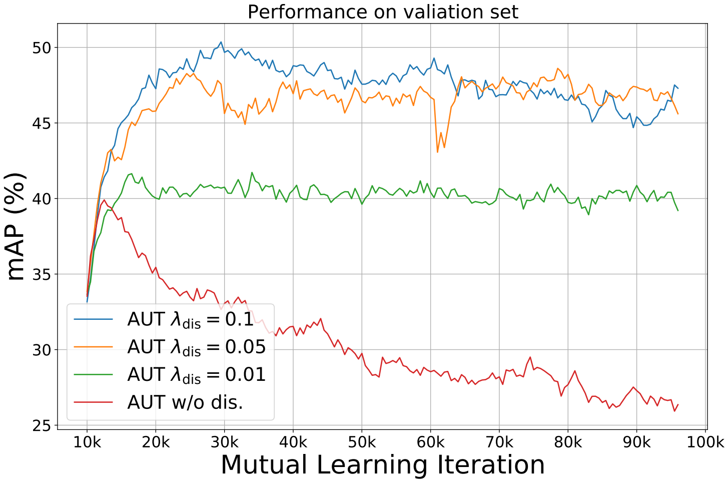


Figure 7: **Mutual Learning curve on Clipart1k dataset**. Increasing weights of λ_{dis} can achieve improved performance and stable learning curve.