Cross-Domain Object Detection via Adaptive Self-Training

Yu-Jhe Li\textsuperscript{1}\textsuperscript{*}  Xiaoliang Dai\textsuperscript{2}  Chih-Yao Ma\textsuperscript{2}  Yen-Cheng Liu\textsuperscript{3}  Kan Chen\textsuperscript{4}  Bichen Wu\textsuperscript{2}  Zijian He\textsuperscript{2}  Kris Kitani\textsuperscript{1}  Peter Vadja\textsuperscript{2}
\textsuperscript{1}Carnegie Mellon University \textsuperscript{2}Meta \textsuperscript{3}Georgia Tech \textsuperscript{4}Waymo
\{yujheli,kkitani\}@cs.cmu.edu, \{xiaoliangdai,cyma,zijian,wbc,vajdap\}@fb.com, ycliu@gatech.edu, kanchen@waymo.com

Abstract

We tackle the problem of domain adaptation in object detection, where there is a significant domain shift between a source (a domain with supervision) and target domain (a domain of interest without supervision). As a widely-adopted domain adaptation method, self-training teacher-student framework (a student model learns from pseudo labels generated from a teacher model) has yielded remarkable accuracy gain on the target domain. However, it still suffers from the large amount of low-quality pseudo labels (e.g., false positives) generated from the teacher due to its bias toward source domain. To address this issue, we propose a self-training framework called Adaptive Unbiased Teacher (AUT) leveraging adversarial learning and weak-strong data augmentation during mutual learning to address domain shift. Specifically, we employ feature-level adversarial training in the student model, 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 between the teacher model on the target domain and the student model on both domains. This enables the 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 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 increasingly important. Researchers have proposed various methods, such as domain classifier and adversarial learning [13], to address the task of cross-domain adaptation [2, 3, 19, 39, 46, 49, 52]. 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 teacher-student self-training method from semi-supervised learning to domain
adaptation [47]. 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] adopts the Mean Teacher (MT) [47] to explore object relation in region-level consistency, inter-graph consistency, and intra-graph consistency. Unbiased Mean Teacher (UMT) [9] proposed to augment the teacher-student framework with CycleGAN [51] 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, direct applying the teacher-student framework only leads to sub-optimal adaptation performance.

To address this problem, we propose a self-training 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, which we refer to as “Weak-Strong augmentation”, following Unbiased Teacher (UT) [29]. This helps teacher model to generate clean and reliable pseudo labels without 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 (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 [16], Fast R-CNN [15], and Faster R-CNN [37] which advances selective search with an accurate and efficient Region Proposal Networks (RPN). Next, a few sub-sequent works [6, 7, 21, 26, 33, 44] 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 [34] jointly predicts bounding boxes and confidences of multiple categories as regression problem. SSD [28] further improves it by utilizing multiple feature maps at different scales. Numerous extensions [12, 27, 35, 36] 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) [30–32]. Another branch is to exploit the domain confusion by learning a domain discriminator [13, 14, 41, 48]. Later, self-ensembling [11] extends Mean Teacher [47] 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 [4, 20, 50]. 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 [2, 3, 19, 39, 46,
The goal of cross-domain object detection is to design domain-invariant detectors by leveraging invariant detectors by leveraging D.

We are given of object detection, we first review the problem formulation.

3.1. Problem Formulation and Overview

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 weak augmentation as input of the Teacher to provide reliable pseudo-labels while images with strong augmentation as inputs of the Student. Specifically, the target samples are augmented with ran-
dom horizontal flip and random crop for weak augmentation in Teacher model and randomly color jittering, grayscale, Gaussian blur, and cutout patches for strong augmentations as perturbations.

Initialization. Initialization is significant for the Teacher model 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 \( D_s = \{(x_s, b_s, c_s)\} \) to optimize our model with the supervised loss \( L_{sup} \). Therefore the loss for training the student model with the labeled source samples can be written as:

\[
L_{sup}(x_s, b_s, c_s) = L_{rpn}^{cls}(x_s, b_s, c_s) + L_{rpn}^{reg}(x_s, b_s, c_s) + L_{roi}^{cls}(x_s, b_s, c_s) + L_{roi}^{reg}(x_s, b_s, c_s),
\]

where RPN loss \( L_{rpn} \) is for the Region Proposal Network (RPN) module which is used for the candidate proposals generation, and ROI loss \( L_{roi} \) is for the prediction branch of Region of Interest (ROI) which performs bounding box regression (reg) and classification (cls).

Optimize Student with Pseudo-Labeling. As the labels are not available in the target domain, we adopt the pseudo-labeling method to generate labels to train the Student with images from the target domain. To prevent the consecutively detrimental effect of noisy pseudo-labels, we set a confidence threshold \( \delta \) of predicted bounding boxes to filter predicted bounding boxes of low-confidence, which are likely to be 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 construct an unsupervised loss on Student model as:

\[
L_{unsup}(x_t, \hat{c}_t) = L_{rpn}^{cls}(x_t, \hat{c}_t) + L_{rpn}^{reg}(x_t, \hat{c}_t),
\]

where \( \hat{c}_t \) denotes the pseudo labels generated by the Teacher model on target domain. Note that we do not apply unsupervised losses for the bounding box regression following [29], since the confidence thresholding would not be able to filter the pseudo-labels that are potentially incorrect for bounding box regression.

Update Teacher via Exponential Moving Average. To obtain more stable pseudo-labels from the target images following MT [47], we apply Exponential Moving Average (EMA) to gradually update the Teacher model. The update can be written as:

\[
\theta_t \leftarrow \alpha \theta_t + (1 - \alpha) \theta_s,
\]

where \( \theta_t \) and \( \theta_s \) 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 unless 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, we place a domain discriminator \( D \), after the feature encoder \( E \) (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 \( L_{dis} \) can be formulated as:

\[
L_{dis} = d \log D(E(x)) + (1 - d) \log(1 - D(E(x))),
\]

Furthermore, adversarial learning in the discriminator is achieved using the Gradient Reverse Layer (GRL) [13] 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:

\[
L_{adv} = \min_E \max_D L_{dis}.
\]

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
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 $\rightarrow$ Clipart1k adaptation. The used backbone is ResNet-101 for fair comparison. We compare our method with SCL [42], SWDA [39], DM [25], CRDA [49], HTCN [2], UMT [9], Source (F-RCNN), and Oracle (F-RCNN).

| Method   | bicycle | bird  | car   | cat   | dog    | person  | table  | dog   | hrs | m-bike | pm | plt | sheep | sofa  | train | tv | mAP  |
|----------|---------|-------|-------|-------|--------|---------|--------|-------|-----|--------|----|-----|-------|-------|-------|----|------|
| Source   | 23.0    | 39.6  | 20.1  | 23.6  | 25.7   | 42.6    | 25.6   | 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 |
| 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 |
| 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 |
| 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  | 50.5| 21.3| 36.1| 52.3 | 53.4| 41.8 |
| CRDA     | 28.7    | 55.3  | 31.6  | 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 |
| 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 |
| 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 |
| 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 |
| 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| 56.7| 19.1| 53.2 | 52.9| 45.0 |

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, 9] 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 loss function to update the model. Since Student model is only updated through EMA discussed in the Sec 3.2.

3.4. Full Objective and Inference

The total loss $L$ for training our proposed AUT is summarized as follows:

$$L = L_{sup} + \lambda_{unsup} \cdot L_{unsup} + \lambda_{dis} \cdot L_{adv},$$

where $\lambda_{unsup}$ and $\lambda_{dis}$ are the hyper-parameters used to control the weighting of the corresponding losses. We note that $L_{sup}$ and $L_{unsup}$ are developed to learn the feature encoder and detector in the Student model while $L_{adv}$ 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 [29]). 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 [5], Foggy Cityscapes [40], PASCAL VOC [10], Clipart1k [22], and Watercolor2k [22].

Cityscapes. Cityscapes [5] 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 [40] 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 [10] is a real-world dataset containing 20 categories of common objects with bounding box annotations. Following [39, 42], we employ PASCAL VOC 2007 and 2012 training and validation images (16,551 images in total) for experiments.

Clipart1k. Clipart1k [22] contains 1k clipart images, which shares the same instance categories with PASCAL VOC but exhibits a large domain shift. We follow the practice in [39, 42] and split it into training and test sets, containing 500 images each.

Watercolor2k. Watercolor2k [22] contains 2k watercolor images, which consists of 2,000 images from 6 classes...
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 used backbone is VGG-16 for fair comparison.

| Method   | bus  | bicycle | car  | motorcycle | 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 [42] | 41.8 | 36.2    | 44.8 | 33.6       | 31.6   | 44.0  | 40.7  | 30.4  | 37.9 (-48)   |
| DA-Faster [3] | 35.3 | 27.1    | 40.5 | 20.0       | 25.0   | 31.0  | 20.2  | 22.1  | 27.6 (-15.1) |
| SCDA [52] | 39.0 | 33.6    | 48.5 | 28.0       | 33.5   | 38.0  | 23.3  | 26.5  | 33.8 (-4.8)  |
| SWDA [39] | 36.2 | 35.3    | 43.5 | 30.0       | 29.9   | 42.3  | 32.6  | 24.5  | 34.3 (-8.4)  |
| DM [25]  | 38.4 | 32.2    | 44.3 | 28.4       | 30.8   | 40.5  | 34.5  | 27.2  | 34.6 (-8.1)  |
| MITOR [1] | 38.6 | 35.6    | 44.9 | 28.3       | 30.6   | 41.4  | 40.6  | 21.9  | 35.1 (-7.6)  |
| MAF [19] | 39.9 | 33.9    | 43.9 | 29.2       | 28.2   | 39.5  | 33.3  | 23.8  | 34.0 (-8.7)  |
| SWDA [39] | 36.2 | 35.3    | 43.5 | 30.0       | 29.9   | 42.3  | 32.6  | 24.5  | 34.3 (-8.4)  |
| DM [25]  | 38.4 | 32.2    | 44.3 | 28.4       | 30.8   | 40.5  | 34.5  | 27.2  | 34.6 (-8.1)  |
| MITOR [1] | 38.6 | 35.6    | 44.9 | 28.3       | 30.6   | 41.4  | 40.6  | 21.9  | 35.1 (-7.6)  |
| MAF [19] | 39.9 | 33.9    | 43.9 | 29.2       | 28.2   | 39.5  | 33.3  | 23.8  | 34.0 (-8.7)  |
| SWDA [39] | 36.2 | 35.3    | 43.5 | 30.0       | 29.9   | 42.3  | 32.6  | 24.5  | 34.3 (-8.4)  |
| DM [25]  | 38.4 | 32.2    | 44.3 | 28.4       | 30.8   | 40.5  | 34.5  | 27.2  | 34.6 (-8.1)  |
| MITOR [1] | 38.6 | 35.6    | 44.9 | 28.3       | 30.6   | 41.4  | 40.6  | 21.9  | 35.1 (-7.6)  |
| MAF [19] | 39.9 | 33.9    | 43.9 | 29.2       | 28.2   | 39.5  | 33.3  | 23.8  | 34.0 (-8.7)  |
| SWDA [39] | 36.2 | 35.3    | 43.5 | 30.0       | 29.9   | 42.3  | 32.6  | 24.5  | 34.3 (-8.4)  |
| DM [25]  | 38.4 | 32.2    | 44.3 | 28.4       | 30.8   | 40.5  | 34.5  | 27.2  | 34.6 (-8.1)  |
| MITOR [1] | 38.6 | 35.6    | 44.9 | 28.3       | 30.6   | 41.4  | 40.6  | 21.9  | 35.1 (-7.6)  |
| MAF [19] | 39.9 | 33.9    | 43.9 | 29.2       | 28.2   | 39.5  | 33.3  | 23.8  | 34.0 (-8.7)  |
| Source (F-RCNN) | 20.1 | 31.9    | 39.6 | 16.9       | 29.0   | 37.2  | 5.2   | 8.1   | 23.5 (-19.2) |
| Oracle (F-RCNN) | 50.3 | 40.7    | 61.3 | 32.5       | 43.1   | 49.8  | 35.1  | 28.6  | 42.7  |

in common with the PASCAL VOC dataset. Following the practice in [39, 42], we split it into training set and test sets, containing 1000 images each.

4.2. Implementation Details

Following [3] and [39], we take the Faster RCNN [37] as the base detection model in our Adaptive Unbiased Teacher. The ResNet-101 [18] or VGG16 [43] model pre-trained on ImageNet [8] is used as the backbone. Following the implementation of Faster RCNN with ROI-alignment [17], 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 [3, 39] 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 [18] 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 [43] is used following previous settings.

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.
Table 4. The average precision (AP, in %) on all classes from different methods for domain generalization on unseen target dataset, which leverages the labeled source data and another domain without supervision. 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 |
| AUT             | 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 [47] + 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 [47] + \( L_{dis} \) | 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 [47] | 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

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 [1] and UMT [9] 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.

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

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

4.5. Experiments on domain generalization

As we observe that our AUT outperforms all of the Oracle models on the three benchmark domain adaptation datasets, 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 AUT with the baseline model MT [47]:

- 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 AUT.
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.

**Adversarial loss $\mathcal{L}_{\text{dis}}$.** To further analyze the importance of adversarial learning in our Adaptive Unbiased Teacher, we exclude the loss $\mathcal{L}_{\text{dis}}$ 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 $\mathcal{L}_{\text{dis}}$ 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 $\lambda_{\text{dis}}$ of the adversarial loss $\mathcal{L}_{\text{dis}}$ 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 Unbiased 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.

**$\mathcal{L}_{\text{unsup}}$ & 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 $\mathcal{L}_{\text{dis}}$. 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 a 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 ex-
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 or 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, 2, 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, 2, 6

[4] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In CVPR, pages 7892–7901, 2018. 2

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, pages 3213–3223, 2016. 5

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. arXiv preprint arXiv:1706.05208, 2017. 2

[12] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659, 2017. 2

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. pages 1180–1189. PMLR, 2015. 1, 2, 3, 4

[14] 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

[15] Ross Girshick. Fast r-cnn. In ICCV, pages 1440–1448, 2015. 2

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, pages 580–587, 2014. 2

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

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

[19] Zhenwei He and Lei Zhang. Multi-adversarial faster-r-cnn for unrestricted object detection. In ICCV, pages 6668–6677, 2019. 1, 2, 6

[20] 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. 2

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

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

[23] 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

[24] 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

[25] 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

[26] 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

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In ICCV, pages 2980–2988, 2017. 2

[28] 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

[29] 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, 4, 5

[30] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. pages 97–105. PMLR, 2015. 2

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In NeurIPS, 2016. 2
[32] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. pages 2208–2217. PMLR, 2017. 2
[33] 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
[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In CVPR, pages 7263–7271, 2017. 2
[35] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. 2
[36] 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
[37] 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
[38] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In CVPR, pages 6956–6965, 2019. 1, 2, 5, 6
[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2014. 6
[40] 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
[41] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scf: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. arXiv preprint arXiv:1911.02559, 2019. 5, 6
[42] 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, 2, 6
[43] 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
[44] 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
[45] 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
[46] 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
A. Broader Impact

We now elaborate the potential impact of our addressed task of cross-domain object detection. In the context of object detection, collecting annotations for target environment can be extremely expensive. 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. As stated in the main paper, 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. However, such method for domain adaptation (DA) can still have limitation when the domain shift is significant between the target and the source datasets. In other words, labeling annotations would be necessary for training a reliable model with supervision.

B. More ablation studies

We also analyze the weight \( \lambda_{\text{unsup}} \) of the unsupervised loss \( L_{\text{unsup}} \) in Figure 5 and confidence threshold \( \delta \) for the teacher model in Figure 6. We run 5 identical experiments for each setting and plot the error bound in these figures.

Weight of unsupervised loss \( \lambda_{\text{unsup}} \) We can see that from the Figure 5, first, setting weight \( \lambda_{\text{unsup}} \) with too large number such as 4.0 would lead to performance drop. This infers that the performance of the model relies on the supervision in source domain. Second, we see that setting with a reasonable weight between 0.25 and 2.0, the model will be able to have satisfactory performance. Third, while we set the \( \lambda_{\text{unsup}} \) as 1.0 in main paper as the default setting for simplicity of implementation, we found that \( \lambda_{\text{unsup}} = 0.25 \) has the best performance (round 52% on Clipart). We note that, the performance of our model can be improved with parameter sweep. However, to show that our model still outperforms state-of-the-arts by a large margin in the main paper without such parameter tuning, we only report the number in the default setting.

Confidence threshold \( \delta \) From the Figure 6, we can observe that setting confidence threshold \( \delta \) with smaller threshold such as 0.5 and 0.6 would lead to drastically performance drop. This shows that we are able to prevent the consecutively detrimental effect of noisy pseudo-labels by setting a confidence threshold \( \delta \) of predicted bounding boxes to filter predicted bounding boxes of low-confidence, which are likely to be false positive samples. With a reasonable weight between 0.8 and 0.9, the model will be able to have satisfactory performance.

C. Qualitative result and 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 7, Figure 8, and Figure 9. 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 artis-
tic 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 7. Qualitative results and comparisons on Foggy Cityscapes.
Figure 8. Qualitative results and comparisons on Clipart1k.
Figure 9. Qualitative results and comparisons on Watercolor2k.