Unbiased Mean Teacher for Cross-domain Object Detection

Jinhong Deng[1]   Wen Li[1*]   Yuhua Chen[2]   Lixin Duan[1]

[1]University of Electronic Science and Technology of China [2]Computer Vision Lab, ETH Zurich

{jhdeng1997, liwenbnu, lxduan}@gmail.com, yuhua.chen@vision.ee.ethz.ch

Abstract

Cross-domain object detection is challenging, because object detection model is often vulnerable to data variance, especially to the considerable domain shift between two distinctive domains. In this paper, we propose a new Unbiased Mean Teacher (UMT) model for cross-domain object detection. We reveal that there often exists a considerable model bias for the simple mean teacher (MT) model in cross-domain scenarios, and eliminate the model bias with several simple yet highly effective strategies. In particular, for the teacher model, we propose a cross-domain distillation method for MT to maximally exploit the expertise of the teacher model. Moreover, for the student model, we alleviate its bias by augmenting training samples with pixel-level adaptation. Finally, for the teaching process, we employ an out-of-distribution estimation strategy to select samples that most fit the current model to further enhance the cross-domain distillation process. By tackling the model bias issue with these strategies, our UMT model achieves mAPs of 44.1%, 58.1%, 41.7%, and 43.1% on benchmark datasets Clipart1k, Watercolor2k, Foggy Cityscapes, and Cityscapes, respectively, which outperforms the existing state-of-the-art results in notable margins. Our implementation is available at https://github.com/kinredon/umt.

1. Introduction

In recent years, deep domain adaptation has gained increasing attention in computer vision community, due to that the supreme performances of deep models are normally restricted only in the domain of training data. When those trained models are applied to new environments, significant performance drops have often been observed in many computer vision tasks [5, 6, 11, 37, 54].

In this work, we are specifically interested in the cross-domain object detection problem, because the strong demands from real-world scenarios. For instance, in autonomous driving, robust object detection is needed in different weathers and lighting conditions. Collecting annotations for all conditions can be extremely costly, and therefore models that can adapt to new environments without labeled data are highly desirable.

Designing such domain adaptive detection models can be challenging. Compared to the image classification task, the output of object detection is richer and more complex, consisting of both the class labels and the bounding box locations. The two outputs are intrinsically coupled, making it more vulnerable towards data variance like scene changes, weather conditions, camera diversity, *etc*. Various approaches have been proposed to address these issues, including instance and image-level adversarial training [6], strong and weak adversarial training [37], graph-based consistency [1], *etc*.

In this paper, we propose a new approach called Unbiased Mean Teacher (UMT) for cross-domain object detection. We build our approach based on the Mean Teacher (MT) model [46], which is originally proposed for semi-supervised learning. By enforcing the consistency over perturbed unlabeled samples between the teacher and student models via distillation, it naturally gains improved robustness against data variance to some extent, and thus being used as our starting point for cross-domain object detection.

However, solely using MT for cross-domain object detection often fails to produce promising results (see Section 4 for detailed experimental study). We conduct a further experimental investigation into this issue and observe that there exists an essential model bias issue for MT in cross-domain scenarios. Specifically, in the presence of a large domain gap, the MT model can be easily biased towards the source domain, as the supervision is mainly from the source domain.

To overcome such model bias occurred in the mean teacher model for cross-domain object detection, we design the unbiased mean teacher model with three strategies. Firstly, observing the biased teacher model often produces more precise predictions for the source images, we design a cross-domain distillation approach by using the source-like target images translated with CycleGAN [55] as the input for teacher model, and original target images as the input for

---

student model. This significantly improves the effectiveness of distillation and reduces the influence of the model bias. Then, to further remedy the model bias of student model, we use the target-like source images as additional labeled data for training. Finally, for the teaching process, we also employ an out-of-distribution estimation strategy to identify the target samples that most fit the current MT model to enhance the cross-domain distillation process.

The contributions of this work mainly can be summarized in three aspects: 1) **A new observation:** We reveal the essential model bias issue in the MT model for cross-domain object detection. 2) **A new model:** We propose a new domain adaptation framework called Unbiased Mean Teacher (UMT) for object detection, which addresses the model bias with several simple yet effective strategies. 3) **A new benchmark:** Our new UMT model achieves state-of-the-art performances on multiple datasets, setting up a new benchmark for cross-domain object detection research. Our UMT model achieves mAPs of 44.1%, 58.1%, 41.7%, and 43.1% on benchmark datasets Clipart1k, Watercolor2k, Foggy Cityscapes, and Cityscapes, respectively.

2. Related Works

Unsupervised Domain Adaptation: Unsupervised domain adaptation methods are designed to adapt a model from the labeled source domain to an unlabeled target domain. Many previous works aim to minimize the distance metric such as maximum mean discrepancy(MMD) [3, 30, 31, 32, 45]. Alternatively, adversarial training with domain classifier is also commonly used to learn domain-invariant representation [10, 11, 39, 48]. More similar to our work, French *et al.* proposed a model based on the Mean Teacher model [46], and achieved state-of-the-art results on various benchmarks. Their model includes student and teacher models. The student model is trained using gradient descent while the weights of the teacher network are an exponential moving average of those of the student, and the inconsistency in predictions between the two models is penalized to encourage model robustness. The aforementioned domain adaptation models have focused on the task image classification, while in this work we study a more challenging object detection problem.

Object Detection: Powered by the strong representation ability of deep convolution neuron network(CNN) models [16, 24, 43], object detection has made significant progress in recent years. Many DCNN-based methods have been proposed [2, 12, 13, 15, 35, 33, 29, 34, 41, 27, 49], achieving remarkable performance in benchmarks such as PASCAL VOC [8] and MSCOCO [28]. Among all works, one of the most representative works is Faster RCNN [35], which extracts regions of interest (RoIs) by a Region Proposal Network (RPN), and then the prediction is made based on the feature sampled from the RoI. In this work,

we test our model using Faster RCNN [35] as our base detection network. But other detection networks should also be possible.

Cross Domain Object Detection: Recently, many works have been proposed to address the domain shift problem occurred in object detection, using different techniques. Many approaches utilize adversarial learning manner with a gradient reverse layer to obtain domain-invariant feature, such as DA-Faster [6], SCDA [56], SWDA [37], SPLAT [47], MAF [17], MDAL [50], CRDA [51], CDN [44], HTCN [4], *etc.* MTOR [1] performs Mean Teacher [46] to explore object relation in region-level consistency, inter-graph consistency and intra-graph consistency. Shan *et al.* [40] employs generative adversarial network and cycle consistency for image translation in the pixel space and minimizes domain discrepancy in features. DM [23] yields various distinctive shifted domains from source domain and employs multi-domain-invariant representation learning to encourage features to be indistinguishable among the domains. The label-level adaptation [21, 22, 36] has also been used for the task and produces improved detection performance. Our proposed UMT achieves improvements in large margins over the above methods.

3. The Unbiased Mean Teacher Model

In this section, we start from the Mean Teacher (MT) model [46] and discuss how to re-purpose it for cross-domain object detection. A simple MT model is firstly introduced by applying MT to object detection with necessary technical adjustment. Then, we give an analysis on the model bias problem of MT model in the cross-domain object detection task. Based on our analysis, we firstly address the model bias in teacher model through a cross-domain distillation for MT to maximally exploit the expertise of the teacher model and then propose to utilize the augmented training samples with pixel-level adaptation to further alleviate the bias in the student model. For the teaching process, we also employ an out-of-distribution detection strategy to select samples that most fit the current model to enhance the cross-domain distillation process.

3.1. The Mean Teacher Model

Mean Teacher (MT) [46] was initially proposed for semi-supervised learning. It consists of two models with identical architecture, a student model and a teacher model. The student model is trained using the labeled data as standard, and the teacher model uses the exponential moving average (EMA) weights of the student model. Each sample prediction of the teacher model can be seen as an ensemble of the student model's current and earlier versions, therefore it is more robust and stable. By enforcing the consistency of teacher and student models using a distillation loss based on unlabeled samples, the student model is then guided to

(a) Source sample   (b) Target-like sample

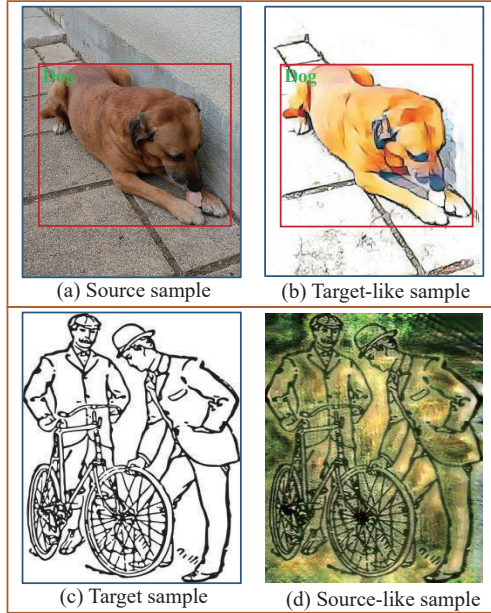(c) Target sample   (d) Source-like sample

Figure 1: Examples of translated images. (a) and (c) are respectively a source example image from the PASCAL VOC dataset [8], and a target example image from the Clipart1k dataset [19]. (b) is the target-like image by translating the source image (a) into the target style, and (d) is the source-like image by translating the target image (c) into the source style.

be more robust. The MT model has also been extended to unsupervised domain adaptation by using the target domain samples as the unlabeled data for distillation in [9].

During the distillation, a small perturbation is added to the unlabeled data. By selecting samples with high probabilities for prediction, the student model is encouraged to reduce more variance on the unlabeled target samples, thus further enhancing the model robustness. Therefore it is suitable to be applied to object detection, for addressing the issue that object detection model is sensitive to data variance due to simultaneously predicting the tangled bounding boxes and object classes.

However, since MT does not explicitly address the domain shift, when applying the MT model to the cross-domain object detection task, the considerable domain shift might cause the predictions from teacher model unreliable, making the distillation less effective (see our investigation in Section 3.3). The recent work [1] proposes to use the region graph to facilitate the distillation, however, it still does not directly address the intrinsic model bias of MT.

3.2. A Simple MT Model for Object Detection

In the cross-domain object detection task, we have a set of source images annotated with object bounding boxes and their class labels, and a set of unlabeled target images. The

source and target domain samples are drawn from different data distribution while their label space is the same. Its goal is to learn a model that could achieve good performance for the target domain.

Formally, let us denote a source image as $\mathbf{I}^s$, which is annotated with multiple bounding boxes as well as their class labels. We denote by $\mathcal{B} = \{B_j|_{j=1}^M\}$ as the set of bounding box coordinates with each $B_j = (x, y, w, h)$ representing a bounding box. Accordingly, we denote by $\mathcal{C} = \{c_j|_{j=1}^M\}$ as the corresponding class labels, in which each $c_j \in \{0, 1, \ldots, C\}$ corresponds to B_j where 0 stands for background, the others for object classes, and C is the total number of classes. Then, the source domain can be represented as $\mathcal{D}_s = \{(\mathbf{I}_i^s, \mathcal{B}_i^s, \mathcal{C}_i^s)|_{i=1}^{N_s}\}$, where N_s is the number of source images. Similarly, the target domain can be defined as $\mathcal{D}_t = \{\mathbf{I}_i^t|_{i=1}^{N_t}\}$ where N_t is the number of target images. Note that labels for the target domain are not available.

Following the protocol of mean teacher, firstly, the labeled source samples are passed through the student model for training. In particular, we employ Faster RCNN [35] as our object detection backbone. Therefore the loss for training the student model with the labeled source samples can be written as:

$$\mathcal{L}_{\text{det}}(\mathcal{B}^s, \mathcal{C}^s, \mathbf{I}^s) = \mathcal{L}_{\text{rpn}}(\mathcal{B}^s; \mathbf{I}^s) + \mathcal{L}_{\text{roi}}(\mathcal{B}^s, \mathcal{C}^s; \mathbf{I}^s), \quad (1)$$

where $\mathcal{L}_{\text{rpn}}$ is the loss for the Region Proposal Network (RPN) module which is used for the candidate proposals generation, and $\mathcal{L}_{\text{roi}}$ is the loss for the prediction branch which performs bounding box regression and classification. More details can be found in [35].

Meanwhile, the unlabeled target samples are augmented with random cropping, padding and color jittering(i.e. brightness, contrast, hue and saturation augmentations). The augmented target samples are fed into teacher and student models, respectively, and then we take the instance predictions with high probabilities of teacher model to guide the student model via distillation. We denote the augmented target samples as $\hat{\mathbf{I}}^t$ for the student model, $\tilde{\mathbf{I}}^t$ for the teacher model. The loss for distillation can be defined as:

$$\mathcal{L}_{dist}(\tilde{\mathbf{I}}^t, \hat{\mathbf{I}}^t) = \mathcal{L}_{\text{det}}(\mathcal{T}_B(\tilde{\mathbf{I}}^t), \mathcal{T}_C(\tilde{\mathbf{I}}^t), \hat{\mathbf{I}}^t), \quad (2)$$

where $\mathcal{T}_B(\tilde{\mathbf{I}}_i^t)$ and $\mathcal{T}_C(\tilde{\mathbf{I}}_i^t)$ are the predicted bounding box coordinates and object classes with high maximum category score from the teacher model on the augmented image $\tilde{\mathbf{I}}_i^t$, and $\mathcal{L}_{\text{det}}$ is the Faster RCNN loss defined in Eq. (1).

For obtaining $\mathcal{T}_B(\tilde{\mathbf{I}}_i^t)$ and $\mathcal{T}_C(\tilde{\mathbf{I}}_i^t)$, we firstly input $\tilde{\mathbf{I}}_i^t$ into the teacher model to get a group of bounding boxes and their class scores. Then, for each foreground category, we sort all proposals by their scores on this category and take Non-Maximum Suppression (NMS) to eliminate redundant proposals. Finally, we select the bounding boxes with their
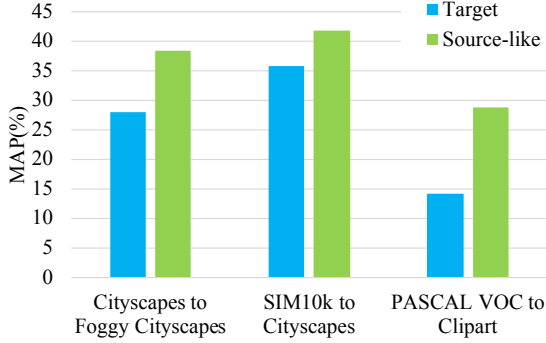
Figure 2: Evaluation of teacher models in mean teacher on target images and source-like images in different cross-domain object detection scenarios. Mean Average Precision (mAP) is used as the comparison metric. It can be observed that the teacher model produces significantly better detection results on source-like images, which clearly indicates the models are biased towards the source domain.

category score larger than a threshold T. Specifically, we denote B_j, c_j and p_j are the bounding boxes coordinates, category and maximum probability of the j-th instance predictions from the teacher model for an input image. The selected instance predictions can be defined as follows:

$$(\mathcal{T}_B, \mathcal{T}_C) = \{(B_j, c_j) | p_j > T, \forall j\}. \tag{3}$$

Finally, the overall loss of the mean teacher can be obtained by putting together those two losses:

$$\mathcal{L} = \mathcal{L}_{\det}(\mathcal{B}^s, \mathcal{C}^s, \mathbf{I}^s) + \lambda \mathcal{L}_{dist}(\tilde{\mathbf{I}}^t, \hat{\mathbf{I}}^t), \tag{4}$$

where λ is a trade-off parameter.

3.3. Investigating the Model Bias in Mean Teacher

Although mean teacher can improve the robustness of predictions on the target domain, it is inevitable that learned models will be biased towards the source domain, as the supervision substantially comes from labeled source samples. As a result, in the distillation process, the prediction on the unlabeled target images produced by the teacher model could be deficient. It will degrade the object detection performance by a notable margin, especially for the unlabeled target domain. This issue might be relatively minor in the classification task. However, in object detection, minor biases in localization may cause a considerable difference in feature pooling and class prediction, and thus leads to inferior guidance to the student model.

To verify this, we experiment on the mean teacher models trained in the aforementioned way on several cross-domain object detection datasets. In particular, we aim to investigate whether the teacher model is biased to the source domain, by comparing its performance on source and target

samples. To ensure a fair comparison, we use the unpaired image-to-image translation by CycleGAN [55] to produce a source-like image for each target image. An example of the source-like image on PASCAL VOC→Clipart1k is shown in Fig. 1(d). We then feed the target samples and the translated source-like samples into the teacher model, and evaluate their detection performance, respectively.

The average precisions (APs) of the teacher models for two versions of samples on different datasets are plotted in Fig. 2. We observe that the APs of teacher models on source-like samples generally outperform their APs on target samples. This clearly confirms our analysis that the teacher model is biased towards the source domain.

3.4. Healing the Model Bias in Mean Teacher

3.4.1 Healing the Teacher Model Bias

Motivated by the above observation, we propose to remold the mean teacher model by pixel-level adaptation. Instead of using only target samples for distillation, we perform a *cross-domain distillation* by using paired images $(\mathbf{I}^t, \mathbf{P}^t)$, where $\mathbf{P}^t$ is the source-like version of the target image $\mathbf{I}^t$. As illustrated in Fig. 3, for each distillation iteration, we feed the source-like image $\mathbf{P}^t$ to the teacher network, and the target image $\mathbf{I}^t$ to the student network. In this way, the teacher network is expected to produce more precise predictions, thus being able to provide better guidance to the student network. Meanwhile, the student network is optimized over the original target samples for the distillation loss, which encourages its favor of target data. Thus, the original distillation loss in Eq. (2) is modified as:

$$\mathcal{L}_{dist}(\tilde{\mathbf{P}}^t, \hat{\mathbf{I}}^t) = \mathcal{L}_{\det}(\mathcal{T}_B(\tilde{\mathbf{P}}^t), \mathcal{T}_C(\tilde{\mathbf{P}}^t), \hat{\mathbf{I}}^t), \tag{5}$$

where $\tilde{\mathbf{P}}^t$ is augmented from $\mathbf{P}^t$ with small perturbations.

3.4.2 Healing the Student Model Bias

As the teacher model is a moving average of the student model, the bias of teacher model essentially comes from the student model. Therefore we also aim to reduce the model bias from the side of student model. Towards this goal, we translate the source images into target-like images. Then the target-like images are used to train the student network, in addition to the supervision from the original source samples. Similarly as in generating source-like images, we utilize CycleGAN to obtain a target-like version image $\mathbf{P}^s$ for each source image $\mathbf{I}^s$ (an example of target-like images is shown in Fig. 1(b)). As the image translation process does not change the ground truth label (*i.e.*, bounding boxes), we use the same label information for target-like images. In this way, the target-like images can encourage the student model to be more favorable of the target domain data, and thus to reduce the bias towards source data. The loss for
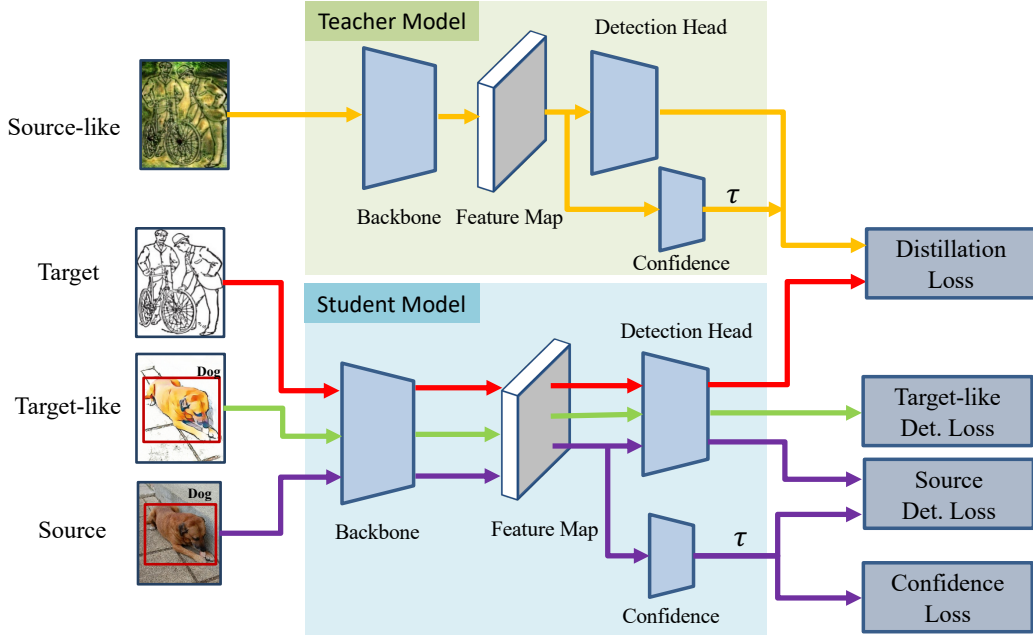
Figure 3: An overview of our proposed Unbiased Mean Teacher. In each mini-batch, four types of images are used: source-like images, target images, target-like images, and source images. The source images with annotations are used to optimize the object detection loss of the student model under the supervision of interpolated ground truth in Eq. (8) (*i.e.*, Source Det. Loss); the source-like images and target images are respectively fed into the teacher and student network to perform the cross-domain distillation (*i.e.*, Distillation Loss); the target-like images are used as additional training samples to train the student model (*i.e.*, Target-like Det. Loss); The confidence branch adopt proposal feature and predict a confidence score with the log penalty in Eq. (7) (*i.e.*, Confidence Loss). These training routes are performed jointly in an end-to-end manner.

target-like images can be written as:

$$\mathcal{L}_{\text{det}}(\mathcal{B}^s, \mathcal{C}^s, \mathbf{P}^s) = \mathcal{L}_{\text{rpn}}(\mathcal{B}^s; \mathbf{P}^s) + \mathcal{L}_{\text{roi}}(\mathcal{B}^s, \mathcal{C}^s; \mathbf{P}^s), \quad (6)$$

which has the same form with the loss in Eq. (1), only replacing $\mathbf{I}^s$ with $\mathbf{P}^s$.

3.4.3 Healing the Teaching Process

We have illustrated how to alleviate the model bias problem of the teacher and student models by using source-like and target-like images. However, one limitation is that these strategies are static, and cannot be automatically adapted during the training process. Specifically, with the source-like and target-like images, the MT model is expected to be gradually guided to fit the target domain, which means that the source-like images might not be ultimately suitable for the teacher model as the training process goes.

To this end, we further propose to dynamically adjust the teaching process of the MT model by selecting the most suitable samples for cross-domain distillation. In particular, as discussed in the out-of-distribution detection works [7, 18, 26, 25], the fitness of samples to a model can be measured with their confidences predicted by the

model*. Therefore, we could employ the target sample confidence scores to automatically select pairs that most fit the *current* teacher model to enhance the teaching process of the MT model during training.

In particular, as shown in Fig. 3, we utilize a side-branch to predict the model confidence for each proposal as inspired by [7]. During the training process, on one hand, we encourage the confidence score to be maximized for all labeled training samples. Let us denote by τ_j as the confidence score for the j-th proposal B_j, a log penalty is used to train the confidence prediction branch:

$$\mathcal{L}_\tau = \sum_j -log(\tau_j). \quad (7)$$

On the other hand, the confidence score is used to adjust classification loss through interpolating between the model prediction and the ground-truth label. Formally, we denote $\mathbf{p}_j = [p_j^1, \ldots, p_j^C]^\top$ as the predicted class probability for B_j and $\mathbf{y}_j$ as its one-hot representation of the category label. We give the hints through interpolating between the

---

*Note that the *confidence* here refers to the confidence of a sample being an in-distribution sample, which is different from the prediction confidence (*i.e.*, the maximum category probability) used in other works.

model prediction $\mathbf{p}_j$ and the category distribution $\mathbf{y}_j$. The interpolation degree is indicated by the confidence score τ_j:

$$\mathbf{p}'_j = \tau_j \cdot \mathbf{p}_j + (1 - \tau_j) \cdot \mathbf{y}_j, \qquad (8)$$

in which τ_j is clipped in the range of $[0, 1]$ if $\tau_j > 1$.

During training, we replace the original one-hot target category probability with the above soft label $\mathbf{p}'$. Therefore, Eq. (1) can be updated as follows:

$$\mathcal{L}_{\text{det}}(\mathcal{B}^s, \mathcal{P}, \mathbf{I}^s) = \mathcal{L}_{\text{rpn}}(\mathcal{B}^s; \mathbf{I}^s) + \mathcal{L}_{\text{roi}}(\mathcal{B}^s, \mathcal{P}; \mathbf{I}^s), \quad (9)$$

where $\mathcal{P} = \{\mathbf{p}'_j|_{j=1}^M\}$.

To enhance the teaching process, we utilize the confidence branch to predict the confidence score for source-like image $\tilde{\mathbf{P}}^t$ in the teacher model. A high confidence score indicates the source-like sample is more likely an in-distribution sample for the current model, and is beneficial to be used for the teaching process in the MT model. Therefore, we updated Eq. (3) as follows:

$$(\hat{\mathcal{T}}_B, \hat{\mathcal{T}}_C) = \left\{ (B_j, c_j) | \sqrt{\tau_j \cdot p_j} > T, \forall j \right\}, \qquad (10)$$

Accordingly, the cross domain distillation loss in Eq. (5) is updated with the above defined $(\hat{\mathcal{T}}_B, \hat{\mathcal{T}}_C)$ as follows:

$$\hat{\mathcal{L}}_{dist}(\tilde{\mathbf{P}}^t, \hat{\mathbf{I}}^t) = \mathcal{L}_{\text{det}}(\hat{\mathcal{T}}_B(\tilde{\mathbf{P}}^t), \hat{\mathcal{T}}_C(\tilde{\mathbf{P}}^t), \hat{\mathbf{I}}^t). \qquad (11)$$

3.4.4 Overall Model

We illustrate the overall architecture of our Unbiased Mean Teacher in Fig. 3. The source-like and target-like data are generated offline. Then the model is trained jointly by optimizing all losses in an end-to-end manner. The overall training objective can be written as:

$$\begin{aligned} \mathcal{L} = &\mathcal{L}_{\text{det}}(\mathcal{B}^s, \mathcal{P}, \mathbf{I}^s) + \mathcal{L}_{\text{det}}(\mathcal{B}^s, \mathcal{C}^s, \mathbf{P}^s) \\ &+ \lambda \cdot \hat{\mathcal{L}}_{dist}(\tilde{\mathbf{P}}^t, \hat{\mathbf{I}}^t) + \gamma \cdot \mathcal{L}_\tau, \end{aligned} \qquad (12)$$

where the loss terms are respectively the detection loss on source samples defined in Eq. (9), the detection loss on target-like samples defined in Eq. (6), the cross-domain distillation loss defined in Eq. (11), and the confidence loss defined in Eq. (7), and λ and γ are the trade-off parameters.

4. Experiments

To validate the effectiveness of our approach, we compare with state-of-the-art methods for cross-domain object detection on benchmark datasets with three different types of domain shifts, including 1) real images to artistic images, 2) normal weather to adverse weather, 3) synthetic images to real images.

As a common practice, we adopt the protocol of unsupervised cross-domain object detection in [6]. Full annotations

including the bounding boxes and the corresponding category labels of objects are available for the source domain training data, while the target domain only contains unlabeled images. Moreover, we can access only the unlabeled train set in the target domain, while the target domain test set is strictly held out during the training phase.

Implementation Details: Following [6], we take the Faster RCNN [35] model as the base object detection model for our Unbiased Mean Teacher approach. The ResNet-101 [16] or VGG16 [43] model pre-trained on ImageNet [24] is used as the backbone for the Faster RCNN model. Following the implementation of Faster RCNN with ROI-alignment [53, 14], we rescale all images by setting the shorter side of the image to 600 while keeping the image aspect ratios.

For the mean teacher and our model, unless otherwise stated, we set the trade-off parameter $\lambda = 0.01$ and $\gamma = 0.1$ for all the experiments. We set the confidence threshold $T = 0.8$ in all our experiments. We train the student network with a learning rate of 0.001 for the first 50k iterations and schedule linear decay for the learning rate of 0.0001 for the next 30k iterations. Each batch consists of four image samples: source, target, source-like, and target-like. The weight smooth coefficient parameter α of the exponential moving average for the teacher model is set to 0.99. Other experimental hyper-parameters settings in our model follow the setup in [35].

To understand the individual impact of the proposed components, we include several special versions of our UMT model for ablation study as follows: 1) UMT$_S$ is the simple mean teacher model by optimizing the loss in Eq. (4); 2) UMT$_{SC}$ is the mean teacher model with our cross-domain distillation strategy as described in Section 3.4.1; and 3) UMT$_{SCA}$ is the mean teacher model with both our cross-domain distillation strategy in Section 3.4.1 and using the target-like images to augment the training set for the student model as described in Section 3.4.2.

4.1. Real to Artistic Adaptation

Datasets: In this experiment, we test our model with domain shift between the real image domain and the artistic image domain. Following [37, 42], we combine the PASCAL VOC 2007 and PASCAL VOC 2012 datasets as the source domain, and use the Clipart1k and Watercolor2k datasets as target domains, respectively. The Clipart1k dataset contains $1,000$ images from the same 20 classes as the PASCAL VOC dataset, which is split equally into a training set and a test set, containing 500 images each. We use the training set as the unlabeled target domain samples for domain adaptation in the training phase, and the test set is held out for evaluation. The Watercolor2k consists of $2,000$ images from 6 classes in common with the PASCAL VOC dataset. Similarly, we use $1,000$ images as the target

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.

Method	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	MAP
Source Only	26.0	45.9	23.2	22.1	20.1	51.7	29.8	9.4	34.6	13.6	30.1	0.9	33.7	50.0	37.2	46.2	18.9	6.7	34.1	20.5	27.7
SCL [42]	**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 [37]	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 [23]	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
CRDA [51]	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
HTCN [4]	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$_S$	30.9	51.8	27.2	28.0	31.4	59.0	34.2	10.0	35.1	19.6	15.8	9.3	41.6	54.4	52.6	40.3	22.7	28.8	37.8	41.4	33.6
UMT$_{SC}$	40.1	**69.3**	26.8	29.0	24.9	39.4	42.7	8.6	39.8	63.0	14.9	18.8	43.6	66.1	63.0	40.7	**31.7**	8.7	27.5	53.0	37.6
UMT$_{SCA}$	39.5	60.0	30.5	39.7	37.5	56.0	42.7	11.1	**49.6**	59.5	21.0	29.2	**49.5**	71.9	66.4	48.0	21.2	13.5	38.8	50.4	41.8
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**
Oracle	33.3	47.6	43.1	38.0	24.5	82.0	57.4	22.9	48.4	49.2	37.9	46.4	41.1	54.0	73.7	39.5	36.7	19.1	53.2	52.9	45.0

Table 2: The 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.

Method	bike	bird	car	cat	dog	person	MAP
Source Only	74.3	49.0	35.2	33.9	25.3	60.6	46.4
SCL [42]	82.2	55.1	51.8	39.6	38.4	64.0	55.2
DM [23]	-	-	-	-	-	-	52.0
SWDA [37]	82.3	**55.9**	46.5	32.7	35.5	66.7	53.3
UMT$_S$	76.2	53.4	46.2	39.3	34.9	**71.5**	53.6
UMT$_{SC}$	79.7	49.5	50.1	**45.5**	30.6	69.8	54.2
UMT$_{SCA}$	86.6	51.3	**52.6**	42.1	33.5	67.5	55.6
UMT	**88.2**	55.3	51.7	39.8	**43.6**	69.9	**58.1**
oracle	49.8	50.6	40.2	38.9	53.3	69.4	50.4

unlabeled training data for training models, and the remaining $1,000$ images are held out for testing.

We include the results from state-of-the-art methods DM [23] , SWDA [37], HTCN [4], CRDA [51] and SCL [42] for comparison. Besides, we report the oracle result 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. All methods are built on the Faster RCNN model, where ImageNet pre-trained ResNet-101 [16] is used as the backbone network.

Results: We report the average precision (AP) of each class as well as the mean AP over all classes in Table 1 and Table 2 for object detection on the Clipart1k and Watercolor2k datasets, respectively.

We take the Clipart1k dataset as an example to explain the experimental results. In particular, the simple MT model UMT$_S$ obtains a mean AP of 33.6%, which outperforms the result of 27.7% from the source only baseline. This proves that the mean teacher model could help to improve the robustness of object detection model against data variance

considerably. However, the improvement is not as significant as other state-of-the-art methods like CRDA, HTCN, and SCL, possibly due to the model bias problem as analyzed in Section 3.3. By using the cross-domain distillation the result is boosted to 37.6% (*i.e.*, UMT$_{SC}$), which is further improved to 41.8% (*i.e.*, UMT$_{SCA}$) by additionally using the target-like augmentation strategy. Note that the result of UMT$_{SC}$ is already on par with the state-of-the-art result on this dataset (*i.e.*, DM), which clearly demonstrates the effectiveness of our strategies for handling the model biases in mean teacher. By dynamically adjusting the teaching process of the MT model by selecting the most suitable samples for cross-domain distillation, our final UMT model reaches 44.1%, which gives the new state-of-the-art performance for cross-domain object detection on the Clipart1k dataset. We have similar observations for the Watercolor2k dataset.

4.2. Adaptation in Inverse Weather

Datasets: In this experiment, we follow the setting in [6]. The training set of the Cityscapes dataset is used as the source domain, and the Foggy Cityscapes dataset [38] is used as the target domain. The Cityscapes dataset is collected from the urban street scene captured in 50 cities. The dataset contains $2,975$ images in the train set and 500 images in the validation set. The Foggy Cityscapes is a synthetic foggy scene dataset rendered using the images and depth maps from Cityscapes, which therefore has the same data split as the Cityscapes dataset, *i.e.* a training set of $2,975$ images a validation set of 500 images. 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 unsupervised domain adaptation.

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

Method	bus	bicycle	car	mcycle	person	rider	train	truck	MAP
Source Only	24.7	29.0	27.2	16.4	24.3	31.5	9.1	12.1	21.8
PF [40]	-	-	-	-	-	-	-	-	28.9
WD [52]	39.9	34.4	44.2	25.4	30.2	42.0	26.5	22.2	33.1
SCL [42]	41.8	36.2	44.8	33.6	31.6	44.0	40.7	30.4	37.9
DA-Faster [6]	35.3	27.1	40.5	20.0	25.0	31.0	20.2	22.1	27.6
SCDA [56]	39.0	33.6	48.5	28.0	33.5	38.0	23.3	26.5	33.8
DM [23]	38.4	32.2	44.3	28.4	30.8	40.5	34.5	27.2	34.6
MAF [17]	39.9	33.9	43.9	29.2	28.2	39.5	33.3	23.8	34.0
MTOR [1]	38.6	35.6	44.0	28.3	30.6	41.4	40.6	21.9	35.1
SWDA [37]	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
CRDA [51]	45.1	34.6	49.2	30.3	32.9	43.8	36.4	27.2	37.4
HTCN [4]	47.4	37.1	47.9	**32.3**	33.2	**47.5**	40.9	31.6	39.8
iFAN [57]	45.5	33.0	48.5	22.8	32.6	40.0	31.7	27.9	35.3
CDN [44]	42.5	36.5	**50.9**	30.8	**35.8**	45.7	29.8	30.1	36.6
UMT$_S$	30.1	31.3	36.1	22.4	27.9	38.2	20.2	21.5	28.5
UMT$_{SC}$	43.4	38.0	50.6	33.7	33.4	45.9	36.4	31.9	39.2
UMT$_{SCA}$	48.2	**38.9**	49.8	33.0	33.8	47.3	42.1	30.0	40.4
UMT	**56.5**	37.3	48.6	30.4	33.0	46.7	**46.8**	**34.1**	**41.7**
Oracle	50.0	36.2	49.7	34.7	33.2	45.9	37.4	35.6	40.3

Besides the baselines compared previously, we further include DA-Faster [6], PF [40], SCDA [56], MAF [17], WD [52], MTOR [1], iFAN [57] and CDN [44] for comparison. The setup for special cases of our UMT approach and the oracle method is the same as those in the previous experiments. All methods are built on Faster RCNN, where the VGG16 [43] pre-trained on ImageNet [24] is used as the backbone.

Results: Object detection in foggy scene images is extremely challenging due to low visibility. The current best state-of-the-art result on this dataset is 39.8% from the recent work HTCN [4]. However, the special case of our UMT$_{SC}$ model which using mean teacher with cross-domain distillation already approaches HTCN with a mean AP of 39.2%. This again validates the effectiveness of the cross-domain distillation strategy for improving the mean teacher model for cross-domain object detection. By healing the student model bias and teaching process, our final UMT model shows improved performance compared to HTCN and reaches a mean AP of 41.7%. Interestingly, this result exceeds the oracle result on this dataset, 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.

4.3. Synthetic-to-Real Adaptation

Datasets: Following [6], the SIM10K dataset [20] is used as the source domain, and the Cityscapes dataset is used as the target domain. The SIM10K dataset contains $10,000$ images of the computer-rendered driving scene

Table 4: The average precision (AP, in %) of different methods for cross-domain object detection on the validation set of Cityscapes for **SIM10K→Cityscapes** adaptation.

Method	AP on *car*
Source Only	34.3
WD [52]	40.6
SCL [42]	42.6
DA-Faster [6]	38.97
SCDA [56]	43.0
SWDA [37]	40.1
MAF [17]	41.1
HTCN [4]	42.5
UMT$_S$	40.8
UMT$_{SC}$	42.0
UMT$_{SCA}$	42.6
UMT	**43.1**
Oracle	53.0

from the Grand Theft Auto (GTAV) game. The training set of Cityscapes is used as target training samples, and the validation set is used for evaluation. All methods are built on Faster RCNN, where the VGG16 [43] pre-trained on ImageNet [24] is used as the backbone.

Results: The AP on detecting cars for different approaches are reported in Table 4. Similarly, as in the previous experiments, we observe that our UMT approach gradually improves the MT model by addressing its model bias with different strategies. Our final model also achieves the new state-of-the-art AP of 43.1% on this dataset using VGG16 [43] as the backbone, which again demonstrates the effectiveness of our proposed approach.

5. Conclusion

In this work, we provide a novel perspective to study the cross-domain object detection problem by exploiting the observation that a detection model can be easily biased towards source images. For that, we present a new method, named Unbiased Mean Teacher (UMT), by designing three highly effective strategies to remedy the model bias. In particular, we firstly introduce cross-domain distillation to maximally exploit the expertise of the teacher model. Then, we further augment the training samples for the student model through pixel-level adaptation to reduce its model bias. Lastly, we employ an out-of-distribution estimation strategy to select samples that most fit the current model to enhance the cross-domain distillation process. Extensive experiments are conducted on multiple benchmark datasets, and the results clearly show that our UMT surpasses the existing state-of-the-art models by relatively large margins.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[3] Fabio Maria Cariucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. AutoDIAL: Automatic domain alignment layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5067–5075, 2017.

[4] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020.

[5] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019.

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.

[7] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

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

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

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

[12] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1440–1448, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[17] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019.

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[19] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5009, 2018.

[20] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016.

[21] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019.

[22] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6092–6101, 2019.

[23] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[25] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[26] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016.

[30] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[32] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, pages 2208–2217, 2017.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[34] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019.

[37] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.

[38] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[39] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[40] Yuhu Shan, Wen Feng Lu, and Chee Meng Chew. Pixel and feature level based domain adaptation for object detection in autonomous driving. *Neurocomputing*, 367:31–38, 2019.

[41] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. DSOD: Learning deeply supervised object detectors from scratch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1927, 2017.

[42] 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.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. *arXiv preprint arXiv:2003.07071*, 2020.

[45] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450, 2016.

[46] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017.

[47] Eric Tzeng, Kaylee Burns, Kate Saenko, and Trevor Darrell. Splat: Semantic pixel-level adaptation transforms for detection. *arXiv preprint arXiv:1812.00929*, 2018.

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[49] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[50] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[51] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020.

[52] Pengcheng Xu, Prudhvi Gurram, Gene Whipps, and Rama Chellappa. Wasserstein distance based domain adaptation for object detection. *arXiv preprint arXiv:1909.08675*, 2019.

[53] Jianwei Yang, Jiasen Lu, Dhruv Batra, and Devi Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017.

[54] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018.

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2223–2232, 2017.

[56] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019.

[57] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew Scott. ifan: Image-instance full alignment networks for adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13122–13129, 2020.