Improving Cross-Domain Detection with Self-Supervised Learning

Anonymous CVPR submission

Paper ID 2

Abstract

Cross-Domain Detection (XDD) aims to train a domain-adaptive object detector using unlabeled images from a target domain and labeled images from a source domain. Existing approaches achieve this either by aligning the feature maps or the region proposals from the two domains, or by transferring the style of source images to that of target images. In this paper, rather than proposing another method following the existing lines, we introduce a new framework complementary to existing methods. Our framework unifies some popular Self-Supervised Learning (SSL) techniques (e.g., rotation angle prediction, strong/weak data augmentation, mean teacher modeling) and adapts them to the XDD task. Our basic idea is to leverage the unsupervised nature of these SSL techniques and apply them simultaneously across domains (source and target) and models (student and teacher). These SSL techniques can thus serve as shared bridges that facilitate knowledge transfer between domains. More importantly, as these techniques are independently applied in each domain, they are complementary to existing domain alignment techniques that relies on interactions between domains (e.g., adversarial alignment). We perform extensive analyses on these SSL techniques and show that they significantly improve the performance of existing methods. In addition, we reach comparable or even better performance than the state-of-the-art methods when integrating our framework with an old well-established method.

1. Introduction

Powered by deep learning, the task of recognizing and localizing objects of interest in a scene, i.e., object detection, has been tremendously advanced in recent years [15, 16, 18, 35, 39–42]. While a deep learning based object detector may have impressive performance on data within the same distribution as the data the detector was trained on, its performance often drops significantly when tested on data drawn from a different distribution. This is the so-called domain shift problem.

Cross-Domain Detection (XDD) addresses the domain shift problem by jointly training a detector with unlabeled data from the domain of interest (target domain) and labeled data from an auxiliary domain (source domain) [7]. By aligning the distributions of the two domains during training, the label supervision from the source domain becomes more shareable to the target domain and hence a detector of enhanced generalizability can be obtained.

Various approaches have been proposed to align domain distributions. The first category of approaches focus on feature alignment where images from both domains are fed to a detection network and are aligned with feature maps at different levels or extracted region proposals [7, 20, 23, 24, 34, 46, 52, 57, 60]. Adversarial learning is often used for the alignment. The second category of approaches are based on pseudo-labeling where the step of pseudo-label prediction and the step of model calibration are executed iteratively [25, 27, 28, 44]. The third category of approaches transforms the source images to resemble the target images with generative models [25, 29]. While similar to the first category in the philosophy of alignment, these methods operate on image pixels directly instead of the feature representations.

In this paper, rather than proposing another method that falls into the existing categories, we propose to address XDD in an orthogonal way by proposing a new framework complementary to existing methods. Our framework, dubbed ATMT[1], adapts and unifies some popular Self-Supervised Learning (SSL) techniques. ATMT takes advantage of the favorable property of these SSL techniques that requires no ground truth labels, and applies the SSL tasks *simultaneously* across domains and models. The shared SSL tasks thus push data from both domains towards common spaces, which mitigates domain shifts.

Specifically, ATMT learns two auxiliary tasks, the Rotation Prediction (RP) task and the Consistency Learning (CL) task, in parallel with the XDD learning task. The RP task trains the model to predict the rotation angle correctly for images, based on the extracted region proposals. It encourages the model to extract region proposals

[1] Short for Auxiliary Tasks and Mean Teacher modeling, which feature the key techniques.

from foreground regions because background regions usually lack semantics sufficient to predict the rotation angles. The CL task trains the model to be robust to changes in the image space by optimizing it to make consistent class predictions for region proposals regardless of various image perturbations. As the image perturbations simulate factors that account for domain shifts, training the detector to overcome them and make consistent predictions enhances cross-domain generalizability. As both auxiliary tasks do not require detection labels and can be applied on images from both domains indiscriminately, learning the two tasks in both domains hence helps push image from both domains towards shared spaces, and thus mitigates domain shifts. In some sense, these tasks serve as the shared bridges between domains, helping the detector overcome the domain gap.

With the enhanced detector learned with the auxiliary tasks, we further boost the performance with a novel mean teacher technique which includes a student and a teacher model with identical architecture [11, 50]. To train the student, we let the teacher and student take different augmented views of the same target image, but requiring their detection outputs to be consistent. The teacher is updated as the Exponential Moving Average (EMA) weights of the student, and therefore can be viewed as an ensemble of the student in different time steps, and ensemble models have shown have better generalization [26].

Our contributions can be summarized as follows:

- We propose the ATMT framework which addresses XDD from a new perspective orthogonal to existing methods. Though each part of ATMT is not *fundamentally* new, we are the first to introduce and unify them to address the XDD problem in a complementary way that results in a highly effective and flexible framework. So, our major technical contribution is rather than any individual part, but the whole framework.
- We conduct extensive analyses on ATMT, providing insights on each component and discuss the design choices to the XDD task. We believe this can inspire future researches. In addition, we reach comparable or even better results than the state-of-the-art methods by integrating ATMT with an old well-established method.
- We make unique modifications on existing techniques to adapt them for XDD. Most techniques incorporated in ATMT are originally applied on whole images; we apply them on region proposals, which fits the detection task. We further propose to address the heterogeneity of the different tasks by sharing the same set of region proposals across different tasks.

2. Related Work

Cross-domain detection. Previous work in Cross-Domain Detection (XDD) addresses the domain shift problem by aligning the features or region proposals from the source and target domains [5, 7, 19, 20, 23, 24, 33, 34, 46, 52, 57, 58, 60]. The alignment is often achieved by adversarial training where domain classifiers predict the domains of the pixels/images/proposals, while the detection model aims to deceive the classifiers . Another line of approaches trains the models iteratively by generating pseudo bounding box labels for target images and updating the models with the generated pseudo-labels [25, 27, 28, 44]. Different methods vary in how they generate the pseudo-labels or update the model. Some methods enhance the adaptation performance by improving the input images. They usually train a style-transfer model (e.g., CycleGAN [59]) using images from both domains and then apply the model to translate images from the source domain as the style of the target domain [25, 29]. As the image style difference narrows, adapting label supervision from the source domain to the target domain becomes easier. We address XDD in a perspective orthogonal to the existing methods by learning auxiliary tasks simultaneously in both domains and by the mean teacher model.

Self-supervised learning. Self-Supervised Learning (SSL) aims to use the data itself as supervision in a pretext task where the model can learn to extract informative representations from unlabeled data. Early efforts focus on designing various pretext tasks including image colorization [32, 55, 56], image rotation prediction [14], spatial context prediction [12], solving jigsaw puzzles [37], image inpainting [38], and contrastive learning [6, 17]. A comparison of some of these approaches can be found in [30]. It shows that the simple image rotation prediction task has shown promising results. SSL has also been introduced to address the domain adaptive classification problem [45, 49, 53] where SSL is used as an auxiliary task jointly trained along with the main alignment tasks. We follow this idea but focus on the detection problem instead. Thus, rather than performing SSL tasks with entire images, we apply it on region proposals. To our best knowledge, this is the first use of SSL to address the XDD problem.

Consistency learning Consistency learning regularizes model predictions to be invariant to moderate changes applied to input examples. It has been a popular technique in recent semi-supervised learning literature [1, 2, 47, 51]. Different consistency training methods vary in how data perturbations are generated and how the consistency loss is composed. Some methods perturb images by compositing various image transformation techniques, including translation, flipping, rotation, stretching, shearing, adding noise, etc. [1, 10, 13]. MixUp [54], a technique that performs linear interpolation between the samples to generate virtual samples, is used in [2]. Learning based augmentation approaches have also been proposed, such as AutoAugment [9] and population based augmentation [22] which employ reinforcement learning to search for the most effective com-

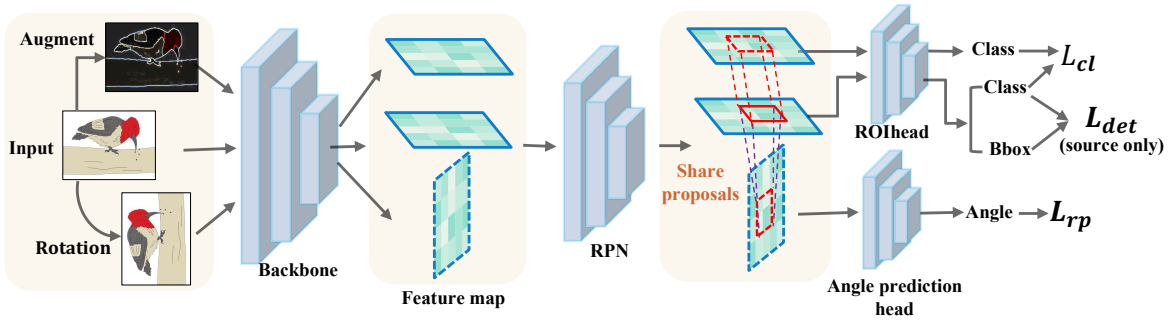

Figure 1. Illustration of the proposed framework. Our framework augments an existing XDD detector at training time with two additional tasks for rotation prediction (L_{rp}) and consistency learning (L_{cl}) which improve performance on the target domain. The domain alignment module (e.g., adversarial alignment) of the XDD detector is omitted for clarity.

binations of transformations. We adopt consistency learning from semi-supervised learning to address the domain shift problem for object detection. While existing methods apply the consistency constraint in image level, we enforce it on the region proposals.

3. Algorithm

Given a labeled dataset $\mathcal{S} = \{\mathcal{X}_s, \mathcal{Y}_s\}$ from the source domain and an unlabeled target dataset $\mathcal{T} = \{\mathcal{X}_t\}$, Cross Domain Detection (XDD) learns an object detector under the following framework:

$$L = L_{det}(\mathcal{X}_s, \mathcal{Y}_s) + \alpha L_{uda}(\mathcal{X}_s, \mathcal{X}_t), \quad (1)$$

where $\mathcal{X}_s$ and $\mathcal{X}_t$ are the images, $\mathcal{Y}_s$ denotes the labels which specify the locations and categories of the objects, and α is a hyper-parameter. The first term $L_{det}(\mathcal{X}_s, \mathcal{Y}_s)$ is the standard supervised learning objective for object detection. It includes the classification objective and bounding box regression objective using labeled images from the source domain. The second term $L_{uda}(\mathcal{X}_s, \mathcal{X}_t)$ is the unsupervised domain alignment objective that aims to align the distributions of the source and target domains. It is unsupervised in the sense that it works without the need of ground truth detection labels. The main effort of existing methods is to devise an effective $L_{uda}(\mathcal{X}_s, \mathcal{X}_t)$ (as well as the supporting model architectures).

Rather than replacing $L_{uda}(\mathcal{X}_s, \mathcal{X}_t)$ in Eq. (1) with another more effective one, we inherit it but boost it from two orthogonal perspectives. First, we append the learning objective with two more terms which correspond to two different auxiliary tasks (Sec. 3.1). Second, after learning the detection model with the enhanced learning objective, we further boost it with a mean teacher model (Sec. 3.3).

3.1. Domain Alignment with Auxiliary Tasks

We propose to train two auxiliary tasks that are applicable to both the source and target domains to bridge the domain gap. The first one is the region proposal based image rotation prediction task which rotates an image and predicts the image rotation angle from the region proposals extracted from the unrotated image. The second task is the consistency learning task where the model is trained to make consistent classification predictions for the same set of region proposals within an image and its strongly augmented version. Figure 1 illustrates the framework.

3.1.1 Proposal-Based Rotation Angle Prediction

Training a model to predict the rotation angle of a given image was proposed in [14] for self-supervised learning. It is based on the intuition that a model can predict the rotation angle correctly if it has a deep understanding of the given image, including localization of salient objects, their orientation, the object type, etc. This inspires us to leverage this task to address the XDD problem because it does not require manually annotated labels, which suits the unsupervised domain adaptation setting well, and it helps localize salient objects and identify the object type, which is exactly the goal of object detection.

A straightforward way of exploiting this task is to learn the rotation prediction task jointly with the detection task by rotating the input image and training the model to predict the rotation angle from the feature representation of the given image. This is how this task is utilized for the classification problem [31, 48, 49, 53]. However, this practice is suboptimal for the detection problem because images used for detection are often much more complex, containing more salient objects in backgrounds with richer contexts. It may be too difficult for the model to learn a global representation for the whole image that encodes the essential information for all the salient objects.

Our insight is that classification and detection can be unified in the region proposal level: a region proposal, once extracted from a large scene, can be viewed as a single-object image typically used for classification. Based on this insight, we propose to predict the rotation angle from the region proposals. This practice has two merits. First, it encourages the detection model to extract region proposals from the foreground since the foreground contains semantic

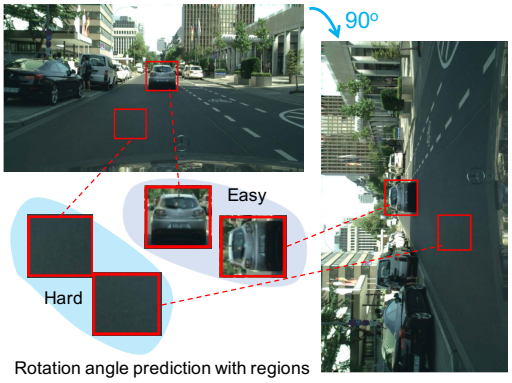

Figure 2. Predicting image rotation angle based on region proposals can help localize foreground regions.

information that is essential to predict the rotation angle. As shown in Figure 2, it is easy to tell the rotation angle from the car region, while hardly possible from the road region. Training the model to predict the rotation angle correctly encourages it to extract region proposals from foreground, which benefits the detection task. Second, this enhances the feature alignment of foreground regions as the model will activate more on the foreground regions and thus contribute more when aligning features from the two domains.

Formally, given a source image $\mathbf{s} \in \mathcal{X}_s$, we obtain $\mathbf{s}^r = Rot(x_s)$ by rotating $\mathbf{s}$ with an random angle θ_s from $[0°, 90°, 180°, 270°]$. From $\mathbf{s}^r$, we extract a set of region proposals $\mathcal{R}_s$ with the same rotation angle θ_s. Similarly, we can get a set of region proposals $\mathcal{R}_t$ with the same rotation angle θ_t for every target image $\mathbf{t} \in \mathcal{X}_t$. We align the domains by applying the rotation prediction task simultaneously on the two domains. Thus, our learning objective for this task is as follows:

$$L_{rp}(\mathcal{X}_s, \mathcal{X}_t) = \frac{1}{|\mathcal{X}_s||\mathcal{R}_s|}\sum_{\mathbf{s}\sim\mathcal{X}_s}\sum_{\mathbf{r}_s\sim\mathcal{R}_s} L(\mathbf{r}_s, \theta_s) + \frac{1}{|\mathcal{X}_t||\mathcal{R}_t|}\sum_{\mathbf{t}\sim\mathcal{X}_t}\sum_{\mathbf{r}_t\sim\mathcal{R}_t} L(\mathbf{r}_t, \theta_t), \tag{2}$$

where $L(\mathbf{r}_s, \theta_s)$ and $L(\mathbf{r}_t, \theta_t)$ are the cross-entropy losses for the source and target proposals, respectively.

3.1.2 Proposal-Based Consistency Learning

Consistency learning regularizes model predictions to be invariant to moderate changes applied to input examples. It has shown impressive performance for semi-supervised learning [8, 36, 47, 51] recently. Based on the insight that unsupervised domain adaptation is a special case of semi-supervised learning where the unlabeled data is drawn from a different data distribution due to the domain shift, we propose to use consistency learning to address the XDD problem. Same as the rotation prediction task, we apply consistency learning on region proposals.

For each source image $\mathbf{s} \in \mathcal{X}_s$, we apply data augmentation Φ and generate

$$\hat{\mathbf{s}} = \Phi(\mathbf{s}). \tag{3}$$

Following the previous methods [1, 47], we use RandAugment [10] as the data augmentation Φ, which produces highly perturbed images by uniformly sampling from the image processing transformations in Python Image Library, including polarization, solarization, brightness change, color change, etc. For ease of implementation, we exclude the transformations that change the positions of pixels (e.g., flipping, rotation, etc.). This ensures $\mathbf{s}$ and $\hat{\mathbf{s}}$ have pixel-to-pixel correspondence for every position. However, our framework could also work with transformations that change the position of pixels as long as the region proposals in the original image can be converted to the coordinates of the transformed image.

We map region proposals $\mathcal{R}_s$ (shared with the rotation prediction task) extracted from $\mathbf{s}$ to $\hat{\mathbf{s}}$, obtaining $\hat{\mathcal{R}}_s$. This ensures that every region proposal $\mathbf{r}_s \sim \mathcal{R}_s$ from $\mathbf{s}$ can find the corresponding $\hat{\mathbf{r}}_s \sim \hat{\mathcal{R}}_s$ from $\hat{\mathbf{s}}$ that localizes the same region in the scene. So, the pair of corresponding region proposals should be classified consistently by the classification branch of the detection model.

We enforce this consistency by optimizing the following objective function:

$$L^s_{cl} = \frac{1}{|\mathcal{R}_s|}\sum_{\mathbf{r}_s\sim\mathcal{R}_s, \hat{\mathbf{r}}_s\sim\hat{\mathcal{R}}_s} \big[\mathbb{1}(\max(\mathbf{p}_s) \geq \sigma) H(\mathbf{p}'_s, \hat{\mathbf{p}}_s)\big], \tag{4}$$

where $\mathbf{p}_s$ and $\hat{\mathbf{p}}_s$ are the classification probabilities of proposals $\mathbf{r}_s$ and $\hat{\mathbf{r}}_s$, respectively. $\mathbf{p}'_s = \arg\max(\mathbf{p}_s)$ returns a one-hot vector for the prediction; $H(.,.)$ is the cross-entropy of two possibility distributions; $\max(\mathbf{p}_s)$ returns the highest possibility score.

In essence, we enforce consistency of the class predictions for a pair of corresponding region proposals $(\mathbf{r}_s, \hat{\mathbf{r}}_s)$ by computing a pseudo label from $\mathbf{r}_s$ and apply the pseudo label on $\hat{\mathbf{s}}$ by computing the standard cross-entropy loss. To mitigate the impact of incorrect pseudo labels, only the samples with confident predictions (the highest probability scores are above a threshold) are used for loss computation.

We apply the same consistency learning task for every target image $\mathbf{t} \in \mathcal{X}_t$ as well. So, the learning objective for the consistency learning task is as follows:

$$L_{cl}(\mathcal{X}_s, \mathcal{X}_t) = \frac{1}{|\mathcal{X}_s|}\sum_{\mathbf{s}\sim\mathcal{X}_s} L^s_{cl} + \frac{1}{|\mathcal{X}_t|}\sum_{\mathbf{t}\sim\mathcal{X}_t} L^t_{cl}. \tag{5}$$

There are several merits of learning the above consistency learning task for the XDD problem. First, it introduces a form of consistency regularization, enforcing the model to be insensitive to the image perturbations and hence being stronger in detecting objects for unlabeled target images. Second, we generate pseudo labels for unlabeled target data and the pseudo labels share the same label space

Algorithm 1. Domain alignment with auxiliary tasks.

Input: Source set $\mathcal{S} = \{\mathcal{X}_s, \mathcal{Y}_s\}$ and target set $\mathcal{T} = \{\mathcal{X}_u\}$.
Output: Domain adaptive detector.

while not done **do**
1. Randomly sample $(\mathbf{s}, y_s) \sim \mathcal{S}$ and $\mathbf{t} \sim \mathcal{T}$.
2. Rotate $\mathbf{s}$ and get $(\mathbf{s}^r, \theta_s) = Rot(\mathbf{s})$; rotate $\mathbf{t}$ and get $(\mathbf{t}^r, \theta_t) = Rot(\mathbf{t})$; augment $\mathbf{s}$ and get $\hat{\mathbf{s}} = \Phi(\mathbf{s})$; augment $\mathbf{t}$ and get $\hat{\mathbf{t}} = \Phi(\mathbf{t})$.
3. Feed-forward $(\mathbf{s}, \mathbf{s}^r, \hat{\mathbf{s}}, \mathbf{t}, \mathbf{t}^r, \hat{\mathbf{t}})$ to the model.
3. Calculate the detection loss and unsupervised domain alignment loss in Eq. (1) using $(\mathbf{s}, y_s)$ and $\mathbf{t}$.
4. Calculate the rotation prediction loss in Eq. (2) using $(\mathbf{s}, \theta_s)$ and $(\mathbf{t}, \theta_t)$.
5. Calculate the consistency learning loss in Eq. (5) using $(\mathbf{s}, \hat{\mathbf{s}})$ and $(\mathbf{t}, \mathbf{t}')$.
6. Back-propagate the loss in Eq. (6).

end while

as the labeled source data. This facilitates label propagation from the labeled source domain to the unlabeled target domain. Third, we augment images with RandAugment [10], which applies various image processing transformations. These transformations and their combinations can model a wide range of factors that cause domain shifts. By training the detection model to be resistant to these factors, the generalizability of the model is thus enhanced.

Integrated learning objective Adding the learning objectives for the two tasks upon Eq. (1), we get our final learning objective as:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{det}(\mathcal{X}_s, \mathcal{Y}_s) + \alpha\mathcal{L}_{uda}(\mathcal{X}_s, \mathcal{X}_t) + \\ &\lambda_1\mathcal{L}_{rp}(\mathcal{X}_s, \mathcal{X}_t) + \lambda_2\mathcal{L}_{cl}(\mathcal{X}_s, \mathcal{X}_t)\end{aligned} \tag{6}$$

where λ_1 and λ_2 are the hyper-parameters.

Algorithm 1 outlines the main steps for this learning stage.

3.2. Discussions

Why do the auxiliary tasks help? The auxiliary tasks and detection task share the same image/proposal representations; by aligning the representations close in the auxiliary task spaces shared by both domains, we can get well-aligned representations and thus the decision boundaries learned from the source domain can generalize better to the target domain. As will be shown in the experiments (Table 2), applying the auxiliary tasks in both domains simultaneously reaches much better performance than that applying these tasks in the source domain alone.

What's unique of applying the auxiliary tasks? Both tasks are originally applied for entire images, here we apply them on region proposals, which suits better for detection. As will be shown in the experiments (Table 3), the proposal-based strategy leads to better performance than the naive image-based strategy. Besides, we address the heterogeneity of different tasks (the main detection task, and the two auxiliary tasks) by sharing the same set of proposals extracted in the original images. Moreover, we apply on images from both domains with RandAugment which includes various image transformations. The combinations of these transformations model a wide range of factors that cause domain shifts. Training the model to be resistant with these factors thus encourages it to extract domain-invariant features across domains. As will be shown in the experiments (Table 4), the strong augmentation technique makes the CL task more effective.

Why not other auxiliary tasks? Rotation prediction and contrastive learning[2] are the two most popular SSL tasks. Other SSL tasks might help as long as they are executed simultaneously in both domains. However, some auxiliary tasks, e.g., Jigsaw Puzzles [37], that change the structure of images shall not fit because objects might be fragmented and unable to be detected.

3.3. Domain Alignment with Mean Teacher

Let h be the cross-domain model (the rotation prediction head was dropped off) learned with Eq. (6). We further propose to enhance h with a mean teacher model. We first use h to initialize the teacher model h_t and the student model h_s that have identical architecture with h. For each unlabeled target image $\mathbf{t} \in \mathcal{X}_t$, we generate a strongly augmented view $\hat{\mathbf{t}} = \Phi(t)$. We feed $\mathbf{t}$ to the teacher model and generate a set of region proposals $\mathcal{R}_t$, and get the corresponding classification probabilities $\mathcal{P}_t$ and bounding box regression offsets $\mathcal{O}_t$ after feeding $\mathcal{R}_t$ to the ROIhead. Similarly, we feed $\hat{\mathbf{t}}$ to the student model, but instead of generating region proposals again, we reuse $\mathcal{R}_t$ and produce the classification probabilities $\hat{\mathcal{P}}_t$ and bounding box offsets $\hat{\mathcal{O}}_t$ in the context of $\hat{\mathbf{t}}$. Then, we back-propagate the following loss to train the student model,

$$L_{mtm} = \mathcal{L}_{det}(\mathcal{X}_s, \mathcal{Y}_s) + \lambda_3\mathcal{L}_{mt}(\mathcal{X}_t), \tag{7}$$

where the first term is the standard object detection loss using label source images. The second term is defined as

$$L_{mt}(\mathcal{X}_t) = \frac{1}{|\mathcal{X}_t||\mathcal{R}_t|}\sum_{\mathbf{t}\sim\mathcal{X}_t}\sum_{\mathbf{r}_t\sim\mathcal{R}_t} D_{KL}(\hat{\mathbf{p}}\|\mathbf{p}) + \|\hat{\mathbf{o}} - \mathbf{o}\|_2, \tag{8}$$

where $\mathbf{p} \in \mathcal{P}_t$ and $\hat{\mathbf{p}} \in \hat{\mathcal{P}}_t$ are the classification probabilities produced by the teacher and student, respectively, with the proposal $\mathbf{r}_t \in \mathcal{R}_t$; Similarly, $\mathbf{o} \in \mathcal{O}_t$ and $\hat{\mathbf{o}} \in \hat{\mathcal{O}}_t$ are the regression offsets produced by the teacher and the student. D_{KL} calculates the KL divergence.

Following the standard practice of mean teacher modeling, the teacher h_t is updated by the student h_s with exponential moving average (EMA) [50].

$$h_t = \eta h_t + (1 - \eta)h_s, \tag{9}$$

[2]Our consistent learning based technique can be viewed as a special case of contrastive learning that without using negative pairs [43].

Algorithm 2. Domain alignment with mean teacher.

Input: Pretrained object detector h, source set $\mathcal{S} = \{\mathcal{X}_s, \mathcal{Y}_s\}$ and target set $\mathcal{T} = \{\mathcal{X}_u\}$.
Output: Teacher model h_t and student model h_s.

1. Initialize student model $h_s = h$ and teacher model $h_t = h$.
while not done **do**
  while not done **do**
    2. Randomly sample $(\mathbf{s}, y_s) \sim \mathcal{S}$ and $\mathbf{t} \sim \mathcal{T}$.
    3. Apply strong augmentation Φ and get $\hat{\mathbf{s}} = \Phi(\mathbf{s})$ and $\hat{\mathbf{t}} = \Phi(\mathbf{t})$.
    4. Produce region proposals $\mathcal{R}_t$ on $\mathbf{t}$ using h_t, and get the corresponding classification probability $\mathcal{P}_t$ and the bounding box regression offset $\mathcal{O}_t$.
    5. Get the classification probability $\hat{\mathcal{P}}_t$ and the bounding box regression offset $\hat{\mathcal{O}}_t$ of proposals $\mathcal{R}_t$ using h_s on $\hat{\mathbf{t}}$.
    6. Train h_s by back-propagating the loss in Eq. (7).
  end while
  7. Update h_t using Eq. (9).
end while

where η is the coefficient we set as $\eta = 0.999$.

It is worth noting that mean teacher has been introduced to address XDD before. Qi et al proposed to enforce the consistency of the relation graphs of region proposals constructed with the student and the teacher [3]. Deng et al proposed to use CycleGAN to generate source-like target images and target-like source images as input to the student and the teacher, respectively, to mitigate model bias [11]. Our method differs from the two methods in that we use strong augmentation to generate a different view for an image and take it as an input to the mean teacher model, which can be achieved easily in an online fashion without relying on external algorithms. Besides, we enforce consistency of the outputs of the student model and the teacher model with KL divergence loss and mean square loss, neither requires hard pseudo labels (one-hot vectors obtained by thresholding the classification probability) that might contain noises and therefore lead to harmful impacts.

Algorithm 2 outlines the steps for training the mean teacher model.

4. Experiments

Datasets. Following the previous methods [7, 46], we conduct experiments on the following three common benchmarks: (1) adaptation from *PASCAL VOC* to *Clipart*, (2) adaptation from *PASCAL VOC* to *Watercolor*, and (3) *Cityscape* to *Foggy Cityscape*. We use ResNet-101 for the first two benchmarks and VGG-16 for the last benchmark as the backbones and pretrain the backbones on ImageNet.

Implementation details. Our ATMT framework can serve as a plug-and-play component to existing XDD methods. To make fair comparison, we keep the architecture and experimental settings unchanged when integrating ATMT with existing methods. Here we only introduce the designs specific to ATMT. When using VGG-16 as the backbone, the rotation prediction branch is structurally identical to the last three FC layers in the standard VGG-16 network, except the output dimension of the last FC layer is 4. When ResNet-101 is used as the backbone, we use a lighter architecture for the rotation prediction branch to save GPU memory. The structure is "Conv3 $\rightarrow$ ReLU $\rightarrow$ Conv1 $\rightarrow$ ReLU". We use mean pooling over the output feature map to get a vector representation for each proposal, which is then used for rotation prediction. To train the mean teacher model, we adopt the same setting as that of the stage of training with auxiliary tasks. We set the hyper-parameters $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$ in Eq. (6)[3], and $\lambda_3 = 10$ in Eq. (7) for all our experiments. For the threshold σ in Eq. (4), we set it as $\sigma = 0.8$ for all our experiments.

4.1. Integrating ATMT with existing XDD methods

ATMT is orthogonal to various existing domain alignment techniques and is expected to further enhance the performance when it is integrated with existing XDD methods. To verify this, we implement ATMT on top of two most well-established XDD models, DAF [7] and SWDA [46]. **Note other more recent XDD methods should also benefit from our ATMT framework; we leave this as a future work and only verify the effectiveness on the well established methods. In addition, as will be shown later, ATMT is able to achieve state-of-the-art performance even with these "old and less-advanced" methods.** As a baseline, we also report the result of integrating ATMT on the source-only model which does not include any domain alignment technique, i.e., without using the second term in Eq. (6).

Table 1 shows the experimental results for the adaptation from *Cityscape* to *Foggy Cityscape*. We can see that ATMT significantly improves performance of existing XDD methods: It raises the Source-only model from 18.8 to 32.7, DAF from 31.9 to 36.6, and SWDA from 34.3 to 38.8, for the mAP, respectively. The ablation study also substantiates the effectiveness of both auxiliary tasks and the mean teacher technique.

It is noted that classes response differently to the proposed techniques, i.e, the AP scores of some classes increase while others decrease after adding the proposed techniques. We analyze the reason could be that due to the application of the Non-Maximum Suppression (NMS) operator, which reduces overlapped proposals, object detection results are often a trade-off among all classes. The proposed technique recalibrate the feature space and decision boundaries for all classes as a whole. Classes that benefit more from them may warp the feature space to the detri-

[3]The hyper-parameter α is not introduced by our framework. It varies in different XDD methods. We keep it unchanged.

	RP	CL	MT	person	rider	car	truck	bus	train	mbike	bicycle	mAP
Source-only				17.8	23.6	27.1	11.9	23.8	9.1	14.4	22.8	18.8
	✓			35.1	30.2	41.9	21.6	28.9	35.8	16.1	20.2	28.7
		✓		32.2	34.7	43.0	28.3	30.9	39.6	10.4	17.3	29.6
	✓	✓		35.8	33.9	44.1	24.2	31.6	40.1	19.8	20.0	31.2
	✓	✓	✓	34.7	36.9	44.2	24.4	29.6	40.0	30.7	20.7	32.7
DAF [7]				31.5	40.9	43.9	21.4	34.2	20.2	27.8	35.4	31.9
	✓			32.7	41.3	44.5	20.6	39.5	28.0	27.8	35.3	33.7
		✓		33.8	43.0	44.7	24.3	38.3	10.9	30.5	39.4	33.1
	✓	✓		34.2	47.1	49.0	25.1	37.7	13.4	33.9	38.9	34.9
	✓	✓	✓	36.1	46.6	51.9	27.5	41.9	15.6	32.7	40.2	36.6
SWDA [46]				29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
	✓			39.8	37.8	48.1	32.0	32.9	41.6	31.8	25.3	36.2
		✓		41.8	34.3	47.7	30.8	33.2	43.1	34.5	28.3	36.7
	✓	✓		47.6	35.0	49.4	33.8	33.6	44.5	31.8	28.3	38.0
	✓	✓	✓	44.7	37.6	51.1	34.0	34.0	46.7	35.1	27.1	38.8

Table 1. Integrating ATMT with existing XDD methods. "RP", "CL", and "MT" stand for the proposed rotation prediction task, the consistency learning task and the mean teacher technique, respectively.

	Cross entropy	RP	CL	RP + CL
Source domain	18.8	20.3	25.0	25.8
Both domains	-	28.7	29.6	31.2

Table 2. Effect of applying the auxiliary tasks on both domains.

	SWDA	SWDA + ImgRot	SWDA + PropRot
mAP	34.3	34.6	36.2

Table 3. Rotation prediction based on entire images (ImgRot) versus that based on region proposals (PropRot).

	mAP
Without CL	34.3
CL with flipped images	34.9
CL with strongly-augmented images	36.7

Table 4. Effect of strong augmentation for CL.

	VOC→Clipart	VOC→Watercolor	Cityscape→Foggy Cityscape
Without MT	43.7	59.0	38.0
MT (normal)	44.0	59.5	38.1
MT (augmented)	45.2	60.2	38.8

Table 5. Effect of strong augmentation for MT.

ment of other classes. Additionally, pseudo-labels may benefit classes unequally if the per-class pseudo-label accuracy varies greatly.

4.2. Analysis

We conduct experiments to analyze ATMT with the adaptation from *Cityscape* to *Foggy Cityscape*, unless otherwise specified.

Auxiliary tasks in one domain *vs.* both domains. Table 2 shows that applying the auxiliary tasks only in the source domain indeed helps improve the generalization performance in the target domain. But the improvement is much less significant than applying the auxiliary tasks in both domains simultaneously. This is because the auxiliary tasks push images from the two domains along the same direction, which alleviates domain shifts.

RP with images *vs.* with proposals. One of the unique aspects of ATMT for the RP task is that it predicts rotation angles based on region proposals, rather than entire images. The merit is that this can encourage the model to extract region proposals from foreground regions and thus enhance detection performance. To validate this, we implement the image-based rotation prediction task and train SWDA jointly with this task. Table 3 shows that the image-based rotation prediction task (ImgRot) produces only a marginal improvement, which is far lower than our proposal-based rotation prediction task.

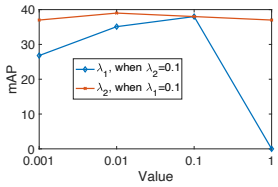

Figure 3. Parameter analysis for λ_1 and λ_2.

λ_3	mAP
0.1	39.4
1.0	39.3
10	38.8
20	38.4
50	37.5

Table 6. Parameter analysis for λ_3.

Mean Teacher (MT) without Strong Augmentation (SA). In MT, we feed the student with images applied with SA. Table 5 shows that if we feed the student with normal (unperturbed) images instead, the mAP scores drop, but are still better than the baseline that without using MT. This substantiates the efficacy of MT as well as SA.

Consistency Learning (CL) without SA. The core idea of CL is to enforce the consistency of the different views of the same image. We use the original image as one view and the strong augmentation to generate the other view. Table 4 shows that if we use the standard augmentation technique, i.e., image flipping, to generate the other view instead, the mAP drops and is only slightly better than the baseline result that without CL. This shows that CL is more effective when the two views of an image are more different.

Parameter analysis. In the phase of alignment with the auxiliary tasks, we have two hyper-parameters, λ_1 and λ_2 that balance the two task losses. Fig. 3 shows the results ATMT-SWDA (ATMT on top of SWDA) is quite robust with λ_2, but is more sensitive to λ_1 and the performance drops to 0 when λ_1 equals 1. This is because the model fails to converge when the rotation prediction loss is weighted too much. In the phase of alignment with the mean teacher model, we have λ_3 that balances the detection loss and the mean teacher loss. Table 6 shows that ATMT-SWDA is not sensitive to this hyper-parameter.

	aero	bike	bird	boat	bot	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
Source-only	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
DAF [7]	26.0	58.3	24.0	23.0	28.1	44.5	29.4	10.4	32.0	39.0	17.5	15.9	31.1	58.2	49.3	44.0	19.1	19.0	30.6	43.0	32.1
SWDA [46]	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
HTCN [5]	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
DDMRL [29]	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
ATF [21]	41.9	67.0	27.4	36.4	41.0	48.5	42.0	13.1	39.2	75.1	33.4	7.9	41.2	56.2	61.4	50.6	42.0	25.0	53.1	39.1	42.1
DBGL [4]	28.5	52.3	34.3	32.8	38.6	66.4	38.2	25.3	39.9	47.4	23.9	17.9	38.9	78.3	61.2	51.7	26.2	28.9	56.8	44.5	41.6
UMT [11]	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
ATMT-SWDA	37.5	63.4	37.9	29.8	45.1	62.7	41.2	19.5	43.7	57.4	22.9	25.3	39.6	87.1	70.9	50.6	29.1	32.2	58.4	50.5	**45.2**

Table 7. Results on adaptation from *PASCAL VOC* to *Clipart*. The best results are in **bold**.

	bike	bird	car	cat	dog	person	mAP
Source-only	68.8	46.8	37.2	32.7	21.3	60.7	44.6
DAF [7]	89.6	45.3	37.5	25.5	24.4	47.9	45.0
SWDA [46]	82.3	55.9	46.5	32.7	35.5	66.7	53.3
WST-BSR [28]	75.6	45.8	49.3	34.1	30.3	64.1	49.9
MAF [20]	73.4	55.7	46.4	36.8	28.9	60.8	50.3
ATF [21]	78.8	59.9	47.9	41.0	34.8	66.9	54.9
DBGL [4]	83.1	49.3	50.6	39.8	38.7	61.3	53.8
UMT [11]	88.2	55.3	51.7	39.8	43.6	69.9	58.1
ATMT-SWDA	88.8	57.7	49.5	44.2	48.4	72.2	**60.2**

Table 8. Results on adpatation from *PASCAL VOC* to *Watercolor*.

	person	rider	car	truck	bus	train	mbike	bicycle	mAP
Source-only	17.8	23.6	27.1	11.9	23.8	9.1	14.4	22.8	18.8
DAF [7]	31.5	40.9	43.9	21.4	34.2	20.2	27.8	35.4	31.9
DAF* [7]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SWDA [46]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
SC-DA [60]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
MAF [20]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
DAM [29]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
GA-CA [23]	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
ECR-DAF [52]	29.7	37.3	43.6	20.8	37.3	12.8	25.7	31.7	29.9
ECR-SWDA [52]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
PDA [24]	36.0	45.5	54.4	24.3	44.1	25.8	29.1	35.9	36.9
RPN-PA [57]	43.6	36.8	50.5	29.7	33.3	45.6	42.0	30.4	39.0
HTCN [5]	33.2	47.5	47.9	31.6	47.4	40.9	32.3	37.1	39.8
UMT [11]	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3	**41.7**
ATMT-SWDA	44.7	37.6	51.1	34.0	34.0	46.7	35.1	27.1	38.8

Table 9. Results of adapting *Cityscapes* to *Foggy Cityscapes*. "DAF*" indicates the results reported in the paper, while "DAF" represents the reimplemented results.

Visualized results. Figure 4 shows some detection samples from the Clipart dataset using the ATMT-SWDA. As a comparison, we also show the detection results of SWDA on the same images. We can see from the figure that ATMT-SWDA produces fewer false negatives (real objects but not detected) and false positives (objects detected but not real).

4.3. Comparison with the State-of-the-Art

To compare with state-of-the-art performance, we implement ATMT on top SWDA [46] and get a method we call ATMT-SWDA. It is worth noting that SWDA is an old but well-established method in this field; its performance is far behind the current state-of-the-art. While we integrate ATMT with SWDA for ease of implementation, ATMT is not limited it (as verified in Table 1). ATMT has the potential of integrating with more recent XDD methods and gets performance better than integrating with SWDA.

Table 7, 8 and 9 show the results for the adaptation from *PASCAL VOC* to *Clipart*, from *PASCAL VOC* to *Watercolor* and from *Cityscape* to *Foggy Cityscape*, respectively. We can see that ATMT-SWDA reaches comparable or even better performance than most recent state-of-the-art methods, even though SWDA is far behind the state-of-the-art. Remarkably, the results show ATMT-SWDA is worse than the state-of-the-art methods for the adaptation from *Cityscape* to *Foggy Cityscape*, but better for the adaptation from *PASCAL VOC* to *Clipart*, and from *PASCAL VOC* to *Watercolor*. We speculate the reason is that the adaptation from *Cityscape* to *Foggy Cityscape* is easier than the other two adaptation experiments, since the two domains are similar. The existing state-of-the-art methods seem more competitive to handle light domain shift, while the proposed ATMT-SWDA is more capable of handling severe domain shift.

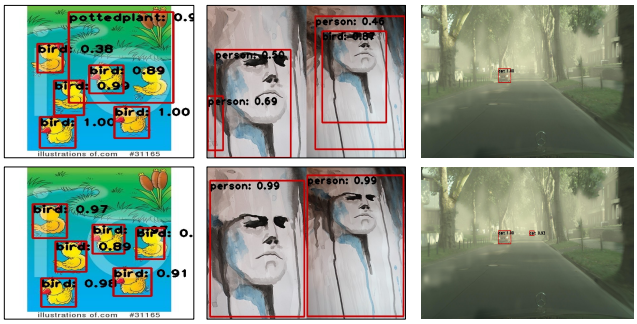

Figure 4. Detection results on *Clipart* (Left), *Watercolor* (Middle), and *Foggy Cityscape* (Right) of SWDA [46] (Top) and ATMT-SWDA (Bottom).

5. Conclusions

We introduce in this paper the ATMT framework which augments existing XDD methods with self-supervised learning techniques. The two auxiliary tasks, proposal based rotation prediction and proposal based consistency learning, are learned simultaneously with images from both domains and thus push the domains towards shared spaces. The enhanced model learned with the auxiliary tasks is further boosted by the proposed mean teacher model, which enhances generalizability by enforcing the consistency of the outputs by the teacher model and the student models for different views of the same unlabeled target images. Experiments show that ATMT significantly improves the performance of existing XDD methods and is able to boost performance of an old well-established method to the level comparable or even better than the state-of-the-art.

References

[1] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2019. 2, 4

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2

[3] 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*, 2019. 6

[4] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, 2021. 8

[5] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, 2020. 2, 8

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[7] 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*, 2018. 1, 2, 6, 7, 8

[8] Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*, 2018. 4

[9] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 2

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019. 2, 4, 5

[11] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2021. 2, 6, 8

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2

[13] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NeurIPS*, 2014. 2

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2, 3

[15] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 1

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[19] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *CVPR*, 2022. 2

[20] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019. 1, 2, 8

[21] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. *arXiv preprint arXiv:2007.01571*, 2020. 8

[22] Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, and Pieter Abbeel. Population based augmentation: Efficient learning of augmentation policy schedules. In *ICML*, 2019. 2

[23] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, 2020. 1, 2, 8

[24] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *WACV*, 2020. 1, 2, 8

[25] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 1, 2

[26] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 2

[27] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, 2019. 1, 2

[28] 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*, 2019. 1, 2, 8

[29] 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*, 2019. 1, 2, 8

[30] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, 2019. 2

[31] D. Zhao etal L. Xiao, J. Xu. Self-supervised domain adaptation with consistency training. In *ICPR*, 2020. 3

[32] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017. 2

[33] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, 2022. 2

[34] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *ICCV*, 2021. 1, 2

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 1

[36] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018. 4

[37] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2, 5

[38] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1

[40] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 1

[41] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NuerIPS*, 2015. 1

[43] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020. 5

[44] 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*, 2019. 1, 2

[45] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. In *NeurIPS*, 2020. 2

[46] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019. 1, 2, 6, 7, 8

[47] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2, 4

[48] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *ECCV*, 2020. 3

[49] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 2, 3

[50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2, 5

[51] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 2, 4

[52] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, 2020. 1, 2, 8

[53] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019. 2, 3

[54] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[55] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[56] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 2

[57] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, 2021. 1, 2, 8

[58] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *CVPR*, 2022. 2

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

[60] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, 2019. 1, 2, 8