Cross-domain Object Detection through Coarse-to-Fine Feature Adaptation

Yangtao Zheng[1,2,3] Di Huang[1,2,3*] Songtao Liu[1,2,3] Yunhong Wang[1,3]
[1]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
[2]State Key Laboratory of Software Development Environment, Beihang University
[3]School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{ytzheng,dhuang,liusongtao,yhwang}@buaa.edu.cn

Abstract

Recent years have witnessed great progress in deep learning based object detection. However, due to the domain shift problem, applying off-the-shelf detectors to an unseen domain leads to significant performance drop. To address such an issue, this paper proposes a novel coarse-to-fine feature adaptation approach to cross-domain object detection. At the coarse-grained stage, different from the rough image-level or instance-level feature alignment used in the literature, foreground regions are extracted by adopting the attention mechanism, and aligned according to their marginal distributions via multi-layer adversarial learning in the common feature space. At the fine-grained stage, we conduct conditional distribution alignment of foregrounds by minimizing the distance of global prototypes with the same category but from different domains. Thanks to this coarse-to-fine feature adaptation, domain knowledge in foreground regions can be effectively transferred. Extensive experiments are carried out in various cross-domain detection scenarios. The results are state-of-the-art, which demonstrate the broad applicability and effectiveness of the proposed approach.

1. Introduction

In the past few years, Convolutional Neural Networks (CNN) based methods have significantly improved the accuracies of plenty of computer vision tasks [21, 38, 49]. These remarkable gains often rely on large-scale benchmarks, such as ImageNet [11] and MS COCO [35]. Due to a phenomenon known as domain shift or dataset bias [59], current CNN models suffer from performance degradation when they are directly applied to novel scenes. In practice, we are able to alleviate such an impact by building a task-specific dataset that covers sufficiently diverse samples. Unfortunately, it is rather expensive and time-consuming to annotate a large number of high-quality ground truths.

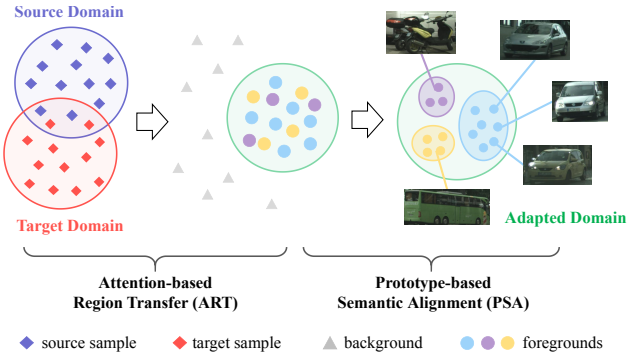

Figure 1. Illustration of the proposed coarse-to-fine feature adaptation approach. It consists of two components, *i.e.*, Attention-based Region Transfer (ART) and Prototype-based Semantic Alignment (PSA). The ART module figures out foregrounds from entire images in different domains and then aligns the marginal distributions on them. Further, the PSA module builds the prototype for each category to achieve semantic alignment. (Best viewed in color.)

To address this dilemma, one promising way is to introduce Unsupervised Domain Adaptation (UDA) to transfer essential knowledge from an off-the-shelf labeled domain (referred to as the *source domain*) to a related unseen but unlabeled one (the *target domain*) [44]. Recently, UDA methods have been greatly advanced by deep learning techniques, and they mostly focus on generating domain-invariant deep representation by reducing cross-domain discrepancy (*e.g.* Maximum Mean Discrepancy [20] or $\mathcal{H}$-divergence [1]), which have proved very competent at image classification [40, 15, 51, 12, 27] and semantic segmentation [23, 42, 64, 6]. Compared to them, object detection is more complex, which is required to locate and classify all instances of different objects within images; therefore, how to effectively adapt a detector is indeed a challenging issue.

In the literature, there are many solutions tackling this problem, including Semi-Supervised Learning (SSL) based [4], pixel-level adaptation based [31, 25, 50], and feature-

*corresponding author

level adaptation based [8, 22, 68, 52, 74]. The SSL based method reduces the domain gap through consistency regularization in a teacher-student scheme. However, the teacher does not always convey more meaningful knowledge than the student [28], and the detector thus tends to accumulate errors, leading to deteriorated detection performance. Pixel-level based methods first conduct style transfer [73] to synthesize a target-like intermediate domain, aiming to limit visual shift, and then train detectors in a supervised manner. Nevertheless, it still remains a difficulty to guarantee the quality of generated images, especially in some extreme cases, which may hurt the adapted results. Alternatively, feature-level adaptation based methods mitigate the domain shift by aligning the features across domains. Such methods work more conveniently with competitive scores, making them dominate the existing community.

In this category, domain adaptive Faster R-CNN [8] is a pioneer. It incorporates both image-level and instance-level feature adaptation into the detection model. In [52], Strong-Weak feature adaptation is launched on image-level. This method mainly makes use of the focal loss to transfer hard-to-classify examples, since the knowledge in them is supposed to be more intrinsic for both domains. Although they deliver promising performance, image-level or instance-level feature adaptation is not so accurate as objects of interest locally distribute with diverse shapes. [74] introduces K-means clustering to mine transferable regions to optimize the adaptation quality. While attractive, this method highly depends on the pre-defined cluster number and the size of the grouped regions, which is not flexible, particularly to real-world applications. Furthermore, in the object detection task, there are generally multiple types of objects, and each has its own sample distribution. But these methods do not take such information into account and regard the distributions of different objects as a whole for adaptation, thereby leaving space for improvement.

In this paper, we present a coarse-to-fine feature adaptation framework for cross-domain object detection. The main idea is shown in Figure 1. Firstly, considering that foregrounds between different domains share more common features compared to backgrounds [30], we propose an Attention-based Region Transfer (ART) module to highlight the importance of foregrounds, which works in a class-agnostic coarse way. We extract foreground objects of interest by leveraging the attention mechanism in high-level features, and underline them during feature distribution alignment. Through multi-layer adversarial learning, effective domain confusion can be performed with the complex detection model. Secondly, category information of objects tends to further refine preceding feature adaptation, and in this case it is necessary to distinguish different kinds of foreground objects. Meanwhile, there is no guarantee that foregrounds of source and target images in the same batch have consistent categories, probably incurring object mis-matches in some mini-batch, making semantic alignment in UDA rather tough. Consequently, we propose a Prototype-based Semantic Alignment (PSA) module to build the global prototype for each category across domains. The prototypes are adaptively updated at each iteration, thus suppressing the negative influence of false pseudo-labels and class mis-matches.

In summary, the contributions are three-fold as follows:

- A new coarse-to-fine adaptation approach is designed for cross-domain two-stage object detection, which progressively and accurately aligns deep features.
- Two adaptation modules, *i.e*., Attention-based Region Transfer (ART) and Prototype-based Semantic Alignment (PSA), are proposed to learn domain knowledge in foreground regions with category information.
- Extensive experiments are carried out in three major benchmarks in terms of some typical scenarios, and the results are state-of-the-art, demonstrating the effectiveness of the proposed approach.

2. Related Work

Object Detection. Object detection is a fundamental step in computer vision and has received increasing attention during decades. Most of traditional methods [63, 10, 13] depend on handcrafted features and sophisticated pipelines. In the era of deep learning, object detection can be mainly split into the one-stage detectors [48, 37, 34, 36] and the two-stage ones [18, 17, 49, 33]. However, those generic detectors do not address the domain shift problem that hurts detection performance in real-world scenes.

Domain Adaptation. Domain adaptation [2, 1] aims to boost performance in the target domain by leveraging common knowledge from the source domain, which has been widely studied in many visual tasks [67, 12, 72, 41, 7, 14]. With the advent of CNNs, many solutions reduce domain shift by learning domain-invariant features. Methods along this line can be divided into two streams: criterion-based [61, 39, 57] and adversarial learning-based [15, 60, 3, 46]. The former aligns the domain distributions by minimizing some statistical distances between deep features, and the latter introduces the domain classifier to construct minimax optimization with the feature extractor. Despite great success is achieved, the majority of them can only handle relatively simple tasks, such as image classification.

Cross-domain Object Detection. A number of traditional studies [66, 62, 43, 70] focus on adapting a specific model (*e.g*., for pedestrian or vehicle detection) across domains. Later, [47] proposes the adaptive R-CNN by subspace alignment [19]. More Recently, the methods can be mainly grouped into four categories, including (1) *Feature-level based*: [8] presents domain adaptive Faster R-CNN to

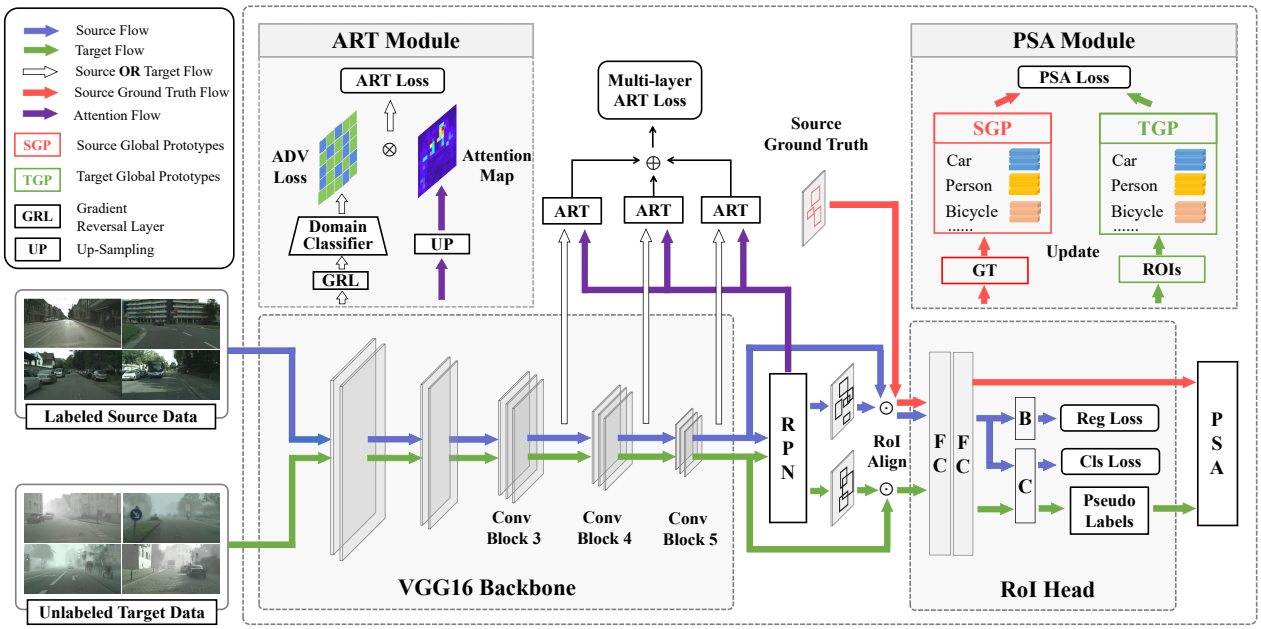

Figure 2. Overview of the proposed feature adaptation framework. We address the problem of domain shift on foreground regions by coarse-to-fine scheme with the ART and PSA modules. First, we utilize the attention map learned from the RPN module to localize foregrounds. Combined with multiple domain classifiers, the ART module puts more emphasis on aligning feature distributions of foreground regions, which achieves a coarse-grained adaptation in a category-agnostic way. Second, the PSA module makes use of ground truth labels (for source) and pseudo labels (for target) to maintain global prototypes for each category, and delivers fine-grained adaptation on foreground regions in a category-ware mode.

alleviate image-level and instance-level shifts, and [22, 68] extend this idea to multi-layer feature adaptation. [52] exploits strong-weak alignment components to attend strong matching in local features and weak matching in global features. [74] mines discriminative regions that contain objects of interest and aligns their features across domains. (2) *SSL based*: [4] integrates object relations into the measure of consistency cost with the mean teacher [58] model. (3) *Pixel-level based*: [25, 50] employ CycleGAN to translate the source domain to the target-like style. [31] uses domain diversification and multi-domain invariant representation learning to address the imperfect translation and source-biased problem. (4) *Others*: [29] establishes a robust learning framework that formulates the cross-domain detection problem as training with noisy labels. [30] introduces weak self-training and adversarial background score regularization for domain adaptive one-stage object detection. [71] minimizes the wasserstein distance to improve the stability of adaptation. [54] explores a gradient detach based stacked complementary loss to adapt detectors.

As mentioned, feature-level adaptation is the main branch in cross-domain object detection, and its performance is currently limited by inaccurate feature alignment. The proposed method concentrates on two-stage detectors and substantially improves the quality of feature alignment by a coarse-to-fine scheme, where the ART module learns the adapted importance of foreground areas and the PSA module encodes the distribution property of each class.

3. Method

3.1. Problem Formulation

In the task of cross-domain object detection, we are given a labeled source domain $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where x_i^s and $y_i^s = (b_i^s, c_i^s)$ denote the i-th image and its corresponding labels, *i.e.*, the coordinates of the bounding box b and its associated category c respectively. In addition, we have access to an unlabeled target domain $\mathcal{D}_T = \{x_i^t\}_{i=1}^{N_t}$. We assume that the source and target samples are drawn from different distributions (*i.e.*, $\mathcal{D}_S \neq \mathcal{D}_T$) but the categories are exactly the same. The goal is to improve the detection performance in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$.

3.2. Framework Overview

As shown in Figure 2, we introduce a feature adaptation framework for cross-domain object detection, which contains a detection network and two adaptation modules.

Detection Network. We select the reputed and powerful Faster R-CNN [49] model as our base detector. Faster R-CNN is a two-stage detector that consists of three major components: 1) a backbone network G that extracts image features, 2) a Region Proposal Network (RPN) that simultaneously predicts object bounds and objectness scores, and 3) a Region-of-Interest (RoI) head, including a bounding box regressor B and a classifier C for further refinement. The overall loss function of Faster R-CNN is defined as:

$$\mathcal{L}_{det}(x) = \mathcal{L}_{rpn} + \mathcal{L}_{reg} + \mathcal{L}_{cls} \quad (1)$$

where $\mathcal{L}_{rpn}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{cls}$ are the loss functions for the RPN, RoI based regressor and classifier, respectively.

Adaptation Modules. Different from most of the existing studies which typically reduce domain shift in the entire feature space, we propose to conduct feature alignment on foregrounds that are supposed to share more common properties across domains. Meanwhile, in contrast to current methods that regard the samples of all objects as a whole, we argue that the category information contributes to this task and thus highlight the distribution of each category to further refine feature alignment. To this end, we design two adaptation modules, *i.e.*, *Attention-based Region Transfer* (ART) and *Prototype-based Semantic Alignment* (PSA), to fulfill a coarse-to-fine knowledge transfer in foregrounds.

3.3. Attention-based Region Transfer

The ART module is designed to raise more attention to align the distributions across two domains within the regions of foregrounds. It is composed of two parts: the domain classifiers and the attention mechanism.

To align the feature distributions across domains, we integrate multiple domain classifiers D into the last three convolution blocks in the backbone network G, where a two-player minimax game is constructed. Specifically, the domain classifiers D try to distinguish which domain the features come from, while the backbone network G aims to confuse the classifiers. In practice, G and D are connected by the Gradient Reverse Layer (GRL) [15], which reverses the gradients that flow through G. When the training process converges, G tends to extract domain-invariant feature representation. Formally, the objective of adversarial learning in the l-th convolution block can be written as:

$$\begin{aligned}\mathcal{L}_{ADV}^{l} = \min_{\theta_{G_l}} \max_{\theta_{D_l}} & \mathbb{E}_{x_s \sim \mathcal{D}_S} \log D_l(G_l(x_s)) \\ & + \mathbb{E}_{x_t \sim \mathcal{D}_T} \log(1 - D_l(G_l(x_t)))\end{aligned} \tag{2}$$

where θ_{G_l} and θ_{D_l} are the parameters of G_l and D_l respectively. $D_l(\cdot)^{(h,w)}$ stands for the probability of the feature in location (h, w) from the source domain.

Recall that the detection task is required to localize and classify objects, and RoIs are usually more important than backgrounds. However, the domain classifiers align all the spatial locations of the whole image without focus, which probably degrades adaptation performance. To deal with this problem, we propose an attention mechanism to achieve foreground-aware distribution alignment. As mentioned in [49], the RPN in Faster R-CNN serves as the attention to tell the detection model where to look, and we naturally utilize the high-level feature in RPN to generate the attention map, as shown in Figure 3. To be specific, given an image x from an arbitrary domain, we denote $F_{rpn}(x) \in \mathbb{R}^{H \times W \times C}$ as the output feature map of the convolutional layer in the RPN module, where $H \times W$ and C are the spatial dimensions and the number of channels of the feature map, respectively. Then, we construct a spatial attention map by averaging activation values across the channel dimension. Further, we filter out (set to zero) those values that are less than the given threshold, which are more likely to belong to the background regions. The attention map $A(x) \in \mathbb{R}^{H \times W}$ is formulated as:

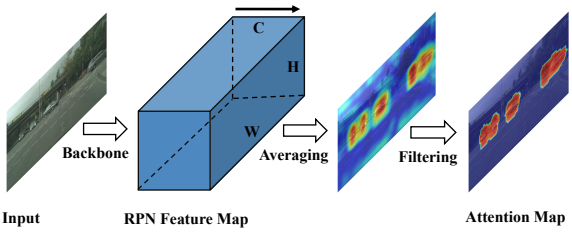

Figure 3. Illustration of the attention mechanism. We first extract the feature map from the RPN module. Then, we construct a spatial attention map by averaging values across the channel dimension. At last, filtering is applied to suppress the noise.

$$M(x) = S(\frac{1}{C} \sum_{c} |F_{rpn}^{c}(x)|) \tag{3}$$

$$T(x) = \frac{1}{HW} \sum_{h,w} M(x)^{(h,w)} \tag{4}$$

$$A(x) = I(M(x) > T(x)) \otimes M(x) \tag{5}$$

where $M(x)$ stands for the attention map before filtering. $S(\cdot)$ is the sigmoid function and $I(\cdot)$ is the indication function. $F_{rpn}^{c}(x)$ represents the c-th channel of the feature map. $\otimes$ denotes the element-wise multiplication. Threshold $T(x)$ is set to the mean value of $M(x)$.

As the size of the attention map is not compatible with the features in different convolution blocks, we adopt bilinear interpolation to perform up-sampling, thus producing the corresponding attention maps. Due to the fact that the attention map may not always be so accurate, if a foreground region is mistaken for background, its attention weight is set to zero and cannot contribute to adaptation. Inspired by the residual attention network in [65], we add a skip connection to the attention map to enhance its performance.

The total objective of the ART module is defined as:

$$\mathcal{L}_{ART} = \sum_{l,h,w} (1 + U_l(A(x)^{(h,w)})) \cdot \mathcal{L}_{ADV}^{l,h,w} \tag{6}$$

where $U_l(\cdot)$ is the up-sampling operation and $\mathcal{L}_{ADV}^{l,h,w}$ stands for the adversarial loss on pixel (h, w) in the l-th convolution block. Combining adversarial learning with the attention mechanism, the ART module aligns the feature distributions of foreground regions that are more transferable for the detection task.

3.4. Prototype-based Semantic Alignment

Since the attention map from RPN carries no information about classification, the ART module aligns the feature distributions of foregrounds in a category-agnostic way. To achieve class-aware semantic alignment, a straightforward method is to train domain classifiers for each category. Nevertheless, there are two main disadvantages: (1) training multiple class-specific classifiers is inefficient; (2) false pseudo-labels (*e.g.*, backgrounds or misclassified foregrounds) occurred in the target domain may hurt the performance of semantic alignment.

Inspired by the prototype-based methods in few-shot learning [56] and cross-domain image classification [69, 5, 45], we propose the PSA module to handle the above problems. Instead of directly training classifiers, PSA tries to minimize the distance between the pair of prototypes (P_k^S, P_k^T) with the same category across domains, thus maintaining the semantic consistency in the feature space. Formally, the prototypes can be defined as:

$$P_k^S = \frac{1}{|GT_k|} \sum_{r \in GT_k} F(r) \tag{7}$$

$$P_k^T = \frac{1}{|RoI_k|} \sum_{r \in RoI_k} F(r) \tag{8}$$

where P_k^S and P_k^T represent the prototypes of the k-th category in the source and target domain respectively. $F(r)$ denotes the feature of foreground region r after the second fully-connected (FC) layer in the RoI head. We use the ground truth GT_k to extract the foreground regions in the source domain. Due to the absence of target annotations, we employ the RoI_k provided by the RoI head module as the pseudo labels in the target domain. $|\cdot|$ indicates the number of regions.

The benefits of prototypes are two-fold: (1) the prototypes have no extra trainable parameters and can be calculated in linear time; (2) the negative influence of false pseudo-labels can be suppressed by the correct ones whose number is much larger when generating the prototypes. It should be noted that the prototypes above are built over all samples. In the training process, the size of each mini-batch is usually small (*e.g.*, 1 or 2) for the detection task, and the foreground objects of source and target images in the same batch may have inconsistent categories, making categorical alignment not practical for all classes at this batch. For example, two images (one for each domain) are randomly selected for training, but *Car* only appears in the source image. As a consequence, we cannot align the prototypes of *Car* across domains in this batch.

To tackle the problem, we dynamically maintain global prototypes, which are adaptively updated by local prototypes at each mini-batch as follows:

$$\alpha = sim(P_k^{(i)}, GP_k^{(i-1)}) \tag{9}$$

Algorithm 1: The coarse-to-fine feature adaptation framework for cross-domain object detection.

Input: Labeled source domain $\mathcal{D}_S$.
Unlabeled target domain $\mathcal{D}_T$.
Batch size B. Category number C.
Output: An adaptive detector $F(\cdot;\theta)$.

Calculate the initial global prototypes $GP_k^{S(0)}$ and $GP_k^{T(0)}$ using the pretrained detector based on $\mathcal{D}_S$
for $i = 1$ **to** max_iter **do**
  $X_S, Y_S \leftarrow$ `Sample(`$\mathcal{D}_S$, $B/2$`)`
  $X_T \leftarrow$ `Sample(`$\mathcal{D}_T$, $B/2$`)`
  Supervised Learning:
  Calculate $\mathcal{L}_{det}$ according to Eq. (1)
  Coarse-grained Adaptation:
  Calculate $A(X_S)$ and $A(X_T)$ by Eq. (5)
  Calculate $\mathcal{L}_{ART}$ by Eq. (6)
  Fine-grained Adaptation:
  $\hat{Y_T} \leftarrow F(X_T;\theta)$
  for $k = 1$ **to** C **do**
    Calculate $P_k^{S(i)}$ and $P_k^{T(i)}$ by Eq. (7) and (8)
    Update $GP_k^{S(i)}$ and $GP_k^{T(i)}$ by Eq. (10)
  Calculate $\mathcal{L}_{PSA}$ according to Eq. (11)
  Optimize the detection model by Eq. (12)

$$GP_k^{(i)} = \alpha P_k^{(i)} + (1-\alpha) GP_k^{(i-1)} \tag{10}$$

where $sim(x_1, x_2) = (\frac{x_1^\top \cdot x_2}{\|x_1\|\|x_2\|} + 1)/2$ denotes the cosine similarity. $P_k^{(i)}$ represents the local prototypes of the k-th category at i-th iteration. It is worth noting that we calculate the initial global prototypes $GP_k^{(0)}$ by Eq. (7) (for source) and Eq. (8) (for target) based on the pretrained model from the labeled source domain.

We do not directly align the local prototypes, but minimize the L_2 distance between the source global prototypes GP_k^S and the target global prototypes GP_k^T to achieve semantic alignment. The objective of the PSA module at i-th iteration can be formulated as following:

$$\mathcal{L}_{PSA} = \sum_k \|GP_k^{S(i)} - GP_k^{T(i)}\|^2 \tag{11}$$

3.5. Network Optimization

The training procedure of our proposed framework integrates three major components, as shown in Algorithm 1.

1. **Supervised Learning.** The supervised detection loss $\mathcal{L}_{det}$ is only applied to the labeled source domain $\mathcal{D}_S$.
2. **Coarse-grained Adaptation.** We utilize the attention mechanism to extract the foregrounds in images. Then, we focus on aligning the feature distributions of those regions by optimizing $\mathcal{L}_{ART}$.

Cityscapes → FoggyCityscapes												
Method	Arch.	Bus	Bicycle	Car	Motor	Person	Rider	Train	Truck	**mAP**	**mAP***	**Gain**
MTOR [4]	R	38.6	35.6	44.0	28.3	30.6	41.4	40.6	21.9	35.1	26.9	8.2
RLDA [29]	I	45.3	36.0	49.2	26.9	35.1	42.2	27.0	30.0	36.5	31.9	4.6
DAF [8]	V	35.3	27.1	40.5	20.0	25.0	31.0	20.2	22.1	27.6	18.8	8.8
SCDA [74]	V	39.0	33.6	48.5	28.0	33.5	38.0	23.3	26.5	33.8	26.2	7.6
MAF [22]	V	39.9	33.9	43.9	29.2	28.2	39.5	33.3	23.8	34.0	18.8	15.2
SWDA [52]	V	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3	20.3	14.0
DD-MRL [31]	V	38.4	32.2	44.3	28.4	30.8	40.5	**34.5**	27.2	34.6	17.9	16.7
MDA [68]	V	41.8	36.5	44.8	30.5	33.2	44.2	28.7	28.2	36.0	22.8	13.2
PDA [25]	V	44.1	35.9	**54.4**	29.1	**36.0**	45.5	25.8	24.3	36.9	19.6	17.3
Source Only	V	25.0	26.8	30.6	15.5	24.1	29.4	4.6	10.6	20.8	-	-
3DC (Baseline)	V	37.9	37.1	51.6	33.1	32.9	45.6	27.9	28.6	36.8	20.8	16.0
Ours w/o ART	V	41.6	35.4	51.5	**36.9**	33.5	45.2	26.6	28.2	37.4	20.8	16.6
Ours w/o PSA	V	**45.2**	37.3	51.8	33.3	33.9	46.7	25.5	29.6	37.9	20.8	17.1
Ours	V	43.2	**37.4**	52.1	34.7	34.0	**46.9**	29.9	**30.8**	**38.6**	20.8	**17.8**
Oracle	V	49.5	37.0	52.7	36.0	36.1	47.1	56.0	32.1	43.3	-	-

Table 1. Results (%) of different methods in the Normal-to-Foggy adaptation scenario. "V", "R" and "I" represent the VGG16, ResNet50 and Inception-v2 backbones respectively. "Source Only" denotes the Faster R-CNN model trained on the source domain only. "3DC" stands for the Faster R-CNN model integrated with three domain classifiers, which is our baseline method. "Oracle" indicates the model trained on the labeled target domain. **mAP*** shows the result of "Source Only" for each method, and **Gain** displays its the improvement after adaptation. The best results are **bolded** and the second best results are underlined among the methods with the VGG16 backbone.

3. **Fine-grained Adaptation.** At first, pseudo labels are predicted in the target domain. We further update the global prototypes for each category adaptively. Finally, semantic alignment on foreground objects is achieved by optimizing $\mathcal{L}_{PSA}$.

With the terms aforementioned, the overall objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda_1 \mathcal{L}_{ART} + \lambda_2 \mathcal{L}_{PSA} \tag{12}$$

where λ_1 and λ_2 denote the trade-off factors for the ART module and the PSA module, respectively.

4. Experiments

4.1. Datasets and Scenarios

Datasets. Four datasets are used in evaluation. (1) ***Cityscapes*** [9] is a benchmark for semantic urban scene understanding. It contains 2,975 training images and 500 validation images with pixel-level annotations. Since it is not designed for the detection task, we follow [8] to use the tightest rectangle of an instance segmentation mask as the ground truth bounding box. (2) ***FoggyCityscapes*** [53] derives from Cityscapes by adding synthetic fog to the original images. Thus, the train/val split and annotations are the same as those in Cityscapes. (3) ***SIM10k*** [26] is a synthetic dataset containing 10,000 images, which is rendered from the video game Grand Theft Auto V (GTA5). (4) ***KITTI*** [16] is another popular dataset for autonomous driving. It consists of 7,481 labeled images for training.

Scenarios. Following [8], we evaluate the framework under three adaptation scenarios as follows:

(1) ***Normal-to-Foggy*** (Cityscapes → FoggyCityscapes). It aims to perform adaptation across different weather conditions. During the training phase, we use the training set of *Cityscapes* and *FoggyCityscapes* as the source and target domain respectively. Results are reported in the validation set of *FoggyCityscapes*.

(2) ***Synthetic-to-Real*** (SIM10k → Cityscapes). Synthetic images offer an alternative to alleviate the data annotation problem. To adapt the synthetic scenes to the real one, we utilize the entire *SIM10k* dataset as the source domain and the training set of *Cityscapes* as the target domain. Since only *Car* is annotated in both domains, we report the performance of *Car* in the validation set of *Cityscapes*.

(3) ***Cross-Camera*** (Cityscapes → KITTI). Images captured by different devices or setups also incur the domain shift problem. To simulate this adaptation, we use the training set of *Cityscapes* as the source domain and the training set of *KITTI* as the target domain. Note that the classification standards of categories in the two domains are different, we follow [68] to classify {*Car*, *Van*} as *Car*, {*Pedestrian*, *Person sitting*} as *Person*, *Tram* as *Train*, *Cyclist* as *Rider* in *KITTI*. The results are reported in the training set of *KITTI*, which is the same as in [8, 68].

4.2. Implementation Details

In all experiments, we adopt the Faster R-CNN with the VGG16 [55] backbone pre-trained on ImageNet [11]. We resize the shorter sides of all images to 600 pixels. The batch size is set to 2, *i.e.*, one image per domain. The detector is trained with SGD for 50k iterations with the learning rate of 10^{-3}, and it is then dropped to 10^{-4} for another 20k iterations. Domain classifiers are trained by the Adam optimizer [32] with the learning rate of 10^{-5}. The factor λ_1 is set at 1.0. Since prototypes in the target domain are unreliable at the beginning, the PSA module is employed after 50k iterations with λ_2 set at 0.01. We report mAP with an IoU threshold of 0.5 for evaluation.

SIM10k → Cityscapes				
Method	Arch.	**AP on Car**	**AP***	**Gain**
RLDA [29]	I	42.6	31.1	11.5
MTOR [4]	R	46.6	39.4	7.2
DAF [8]	V	39.0	30.1	8.9
MAF [22]	V	41.1	30.1	**11.0**
SWDA [52]	V	42.3	34.6	7.7
MDA [68]	V	42.8	34.3	8.5
SCDA [74]	V	43.0	34.0	9.0
Source Only	V	35.0	-	-
3DC (Baseline)	V	42.3	35.0	7.3
Ours w/o ART	V	42.7	35.0	7.7
Ours w/o PSA	V	43.4	35.0	8.4
Ours	V	**43.8**	35.0	8.8
Oracle	V	59.9	-	-

Table 2. Results (%) of the Synthetic-to-Real adaptation scenario.

4.3. Results

We conduct extensive experiments and make comparison to the state-of-the-art cross-domain object detection methods, including (1) **Semi-Supervised Learning**: MTOR [4], (2) **Robust Learning**: RLDA [29], (3) **Feature-level adaptation**: DAF [8], SCDA [74], MAF [22], SWDA [52] and MDA [68], and (4) **Pixel-level adaptation + Feature-level adaptation**: DD-MRL [31] and PDA [25]. Moreover, we also provide ablation studies to validate the effectiveness of each module. Our baseline method is referred as 3DC, which is the Faster R-CNN model integrated with three domain classifiers. We alternately remove the ART and PSA module from the entire framework and report the performance. Note that removing the ART means we only remove the attention map while domain classifiers are still kept.

Normal-to-Foggy. As shown in Table 1, we achieve an mAP of 38.6% on the weather transfer task, which is the best result among all the counterparts. Since detection performance before adaptation is different for each method, we point out that "Gain" is also a key criterion for fair comparison, which is ignored by previous work. In particular, we achieve a remarkable increase of +17.8% over the source only model. Among all the feature-level adaptation methods, we improve the mAP by +2.6% compared to MDA [68]. Although we do not leverage extra pixel-level adaptation, our method still outperforms previous state-of-the-art PDA [25] by +1.7%. Besides, with the help of coarse-to-fine feature adaptation on foregrounds, the proposed method brings improvements on all the categories than the 3DC model does, which shows that feature alignment on foregrounds can boost performance. Additionally, we find that the proposed method is comparable to or even better than the oracle model in several categories. It suggests that the performance which is similar to that of supervised learning methods can be achieved, if we effectively transfer knowledge across domains.

Synthetic-to-Real. Table 2 displays the results on the Synthetic-to-Real task. We obtain an average precision of 43.8% on *Car* and find that there is a slight gain of +0.8% compared to SCDA [74]. The reason is that knowledge transfer is much easier for single category, and many other methods can also adapt well. Further, one may wonder why the PSA module is still effective for single category adaptation, and we argue that it serves as another attention mechanism that focuses on foreground regions, which conveys some complementary clues to the ART module in this case.

Cityscapes → KITTI									
Method	Arch.	Person	Rider	Car	Truck	Train	**mAP**	**mAP***	**Gain**
DAF [8]	V	40.9	16.1	70.3	23.6	21.2	34.4	34.0	0.4
MDA [68]	V	**53.0**	24.5	72.2	28.7	**25.3**	40.7	34.0	6.7
Source Only	V	48.1	23.2	**74.3**	12.2	9.2	33.4	-	-
3DC (Baseline)	V	45.8	27.0	73.9	26.4	18.4	38.3	33.4	4.9
Ours w/o ART	V	50.2	27.3	73.2	29.5	17.1	39.5	33.4	6.1
Ours w/o PSA	V	50.5	27.8	73.3	26.8	20.5	39.8	33.4	6.4
Ours	V	50.4	**29.7**	73.6	**29.7**	21.6	**41.0**	33.4	**7.6**
Oracle	V	71.1	86.6	88.4	90.7	90.1	85.4	-	-

Table 3. Results (%) of the Cross-Camera adaptation scenario.

Cross-Camera. In Table 3, we illustrate the performance comparison on the cross-camera task. The proposed method reaches an mAP of 41.0% with a gain of +7.6% over the non-adaptive model. Due to the fact that scenes are similar across domains and the *Car* sample dominate the two datasets, we can observe that the score on *Car* is already good for the source only model. Compared with DAF [8] and MDA [68], our method reduces the influence of negative transfer in *Car* detection. Meanwhile, our method also outperforms the baseline model (3DC) in the rest categories.

4.4. Further Analysis

Feature distribution discrepancy of foregrounds. The theoretical result in [2] suggests that $\mathcal{A}$-distance can be used as a metric of domain discrepancy. In practice, we calculate the Proxy $\mathcal{A}$-distance to approximate it, which is defined as $d_{\mathcal{A}} = 2(1 - \epsilon)$. ϵ is the generalization error of a binary classifier (linear SVM in our experiments) that tries to distinguish which domain the input features come from. Figure 5 displays the distances for each category on the *Normal-to-Foggy* task with the features of ground truth foregrounds extracted from the models of *Source Only*, *SWDA* and *Ours*. Compared with the non-adaptive model, *SWDA* and *Ours* reduce the distances in all the categories by large margins, which demonstrates the necessity of domain adaptation. Besides, since we explicitly optimize the prototypes of each category by PSA, we achieve a smaller feature distribution discrepancy of foregrounds than the others do.

Error analysis of highest confident detections. To further validate the effect of the proposed framework for cross-domain object detection, we analyze the errors of the models of *Source Only*, *SWDA* and *Ours* caused by highest confident detections on the *Normal-to-Foggy* task. We follow [24, 8, 4] to categorize the detections into three error

(a) Source Only (b) SWDA (c) Ours (d) Attention Map

Figure 4. Qualitative results on the *Normal-to-Foggy* adaptation scenario. (a)-(c): The detection results of the Source Only model, SWDA and the proposed method. (d): Visualization of the corresponding attention maps (best viewed by zooming in).

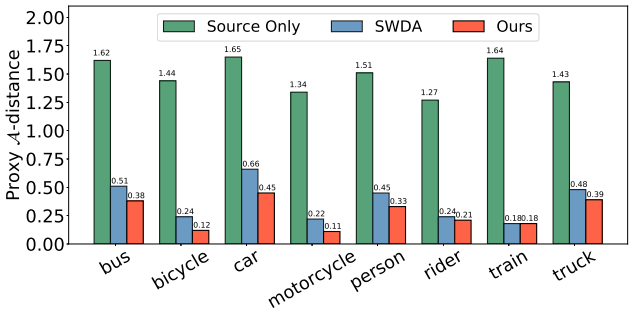

Figure 5. Feature distribution discrepancy of foregrounds.

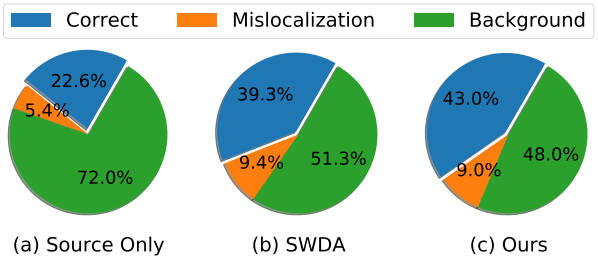

Figure 6. Error analysis of highest confident detections.

types: **1) Correct** (IoU with GT ≥ 0.5), **2) Mislocalization** ($0.3 \leq$ IoU with GT < 0.5), and **3) Background** (IoU with GT < 0.3). For each category, we select top-K predictions to analyze the error type, where K is the number of ground truths in this category. We report the mean percentage of each type across all categories in Figure 6. We can see that the *Source Only* model seems to take most of backgrounds as false positives (green color). Compared with *SWDA*, we improve the percentage of correct detections (blue color) from 39.3% to 43.0% and reduce other error types simultaneously. The results indicate that the proposed framework can effectively increase true positives and reduce false positives, resulting in better detection performance.

Qualitative results. Figure 4 shows some qualitative results. Due to the domain shift problem, the *Source Only* model simply detects some salient objects as shown in (a). From (b) to (c), we can observe that the proposed method not only increases true positives (detects more cars in the first and second row), but also reduces false positives (discards persons in the third row), which is consistent with previous analysis. Further, we visualize the attention maps generated from the ART module. Despite some noise, the attention maps well locate the foreground regions, which is beneficial to knowledge transfer across domains.

5. Conclusion

In this paper, we present a novel coarse-to-fine feature adaptation approach to address the issue of cross-domain object detection. The proposed framework achieves the goal with the incorporation of two delicately designed modules, *i.e.*, ART and PSA. The former highlights the importance of the foreground regions figured out by the attention mechanism in a category-agnostic way, and aligns their feature distributions across domains. The latter takes the advantage of prototypes to perform fine-grained adaptation of foregrounds at the semantic level. Comprehensive experiments are conducted on various adaptation scenarios and state-of-the-art results are reached, demonstrating the effectiveness of the proposed approach.

Acknowledgment. This work is funded by the National Key Research and Development Plan of China under Grant 2018AAA0102301, the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2019ZX-03), and the Fundamental Research Funds for the Central Universities.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, 2007.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.

[4] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019.

[5] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, 2019.

[6] M. Chen, H. Xue, and D. Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, 2019.

[7] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, and J. Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*, 2019.

[8] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[12] Z. Ding, S. Li, M. Shao, and Y. Fu. Graph adaptive knowledge transfer for unsupervised domain adaptation. In *ECCV*, 2018.

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[14] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, 2019.

[15] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[16] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[17] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[19] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

[20] A. Gretton, K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola. A kernel method for the two-sample problem. In *NeurIPS*, 2008.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[22] Z. He and L. Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019.

[23] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[24] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012.

[25] H.-K. Hsu, W.-C. Hung, H.-Y. Tseng, C.-H. Yao, Y.-H. Tsai, M. Singh, and M.-H. Yang. Progressive domain adaptation for object detection. In *CVPR Workshops*, 2019.

[26] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*, 2017.

[27] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, 2019.

[28] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *ICCV*, 2019.

[29] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, 2019.

[30] S. Kim, J. Choi, T. Kim, and C. Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, 2019.

[31] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, 2019.

[32] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[36] S. Liu, D. Huang, and Y. Wang. Receptive field block net for accurate and fast object detection. In *ECCV*, 2018.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[39] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[40] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2014.

[41] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *ICCV*, 2019.

[42] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019.

[43] F. Mirrashed, V. I. Morariu, B. Siddiquie, R. S. Feris, and L. S. Davis. Domain adaptive object detection. In *WACV*, 2013.

[44] S. J. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010.

[45] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, 2019.

[46] Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018.

[47] A. Raj, V. Namboodiri, and T. Tuytelaars. Subspace alignment based domain adaptation for rcnn detector. In *BMVC*, 2015.

[48] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[49] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *T-PAMI*, 39(6):1137–1149, 2017.

[50] A. L. Rodriguez and K. Mikolajczyk. Domain adaptation for object detection via style consistency. In *BMVC*, 2019.

[51] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, 2017.

[52] K. Saito, Y. Ushiku, T. Harada, and K. Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019.

[53] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018.

[54] Z. Shen, H. Maheshwari, W. Yao, and M. Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019.

[55] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv:1409.1556, 2014.

[56] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[57] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016.

[58] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[59] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[60] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[61] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[62] D. Vázquez, A. M. López, and D. Ponsa. Unsupervised domain adaptation of virtual and real worlds for pedestrian detection. In *ICPR*, 2012.

[63] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[64] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.

[65] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *CVPR*, 2017.

[66] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *CVPR*, 2011.

[67] Y. Xia, D. Huang, and Y. Wang. Detecting smiles of young children via deep transfer learning. In *ICCV Workshops*, 2017.

[68] R. Xie, F. Yu, J. Wang, Y. Wang, and L. Zhang. Multi-level domain adaptive learning for cross-domain detection. In *ICCV Workshops*, 2019.

[69] S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, 2018.

[70] J. Xu, S. Ramos, D. Vázquez, and A. M. López. Domain adaptation of deformable part-based models. *T-PAMI*, 36(12):2367–2380, 2014.

[71] P. Xu, P. Gurram, G. Whipps, and R. Chellappa. Wasserstein distance based domain adaptation for object detection. *arXiv preprint arXiv:1909.08675*, 2019.

[72] S. Zhao, H. Fu, M. Gong, and D. Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *CVPR*, 2019.

[73] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[74] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, 2019.