Exploring Categorical Regularization for Domain Adaptive Object Detection

Chang-Dong Xu* Xing-Ran Zhao* Xin Jin Xiu-Shen Wei†
Megvii Research Nanjing, Megvii Technology
{xuchangdong, zhaoxingran, jinxin}@megvii.com, weixs.gm@gmail.com

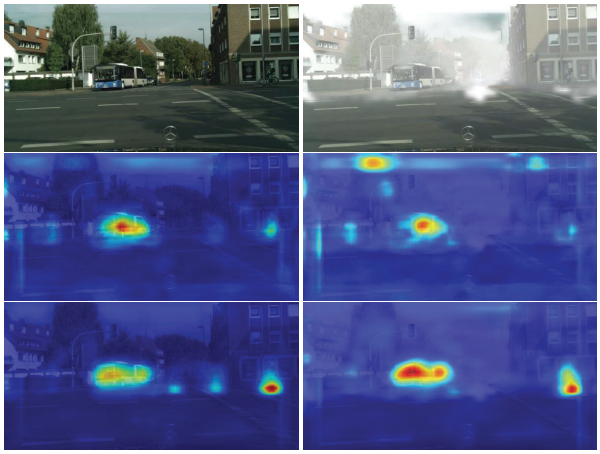
Abstract

In this paper, we tackle the domain adaptive object detection problem, where the main challenge lies in significant domain gaps between source and target domains. Previous work seeks to plainly align image-level and instance-level shifts to eventually minimize the domain discrepancy. However, they still overlook to match crucial image regions and important instances across domains, which will strongly affect domain shift mitigation. In this work, we propose a simple but effective categorical regularization framework for alleviating this issue. It can be applied as a plug-and-play component on a series of Domain Adaptive Faster R-CNN methods which are prominent for dealing with domain adaptive detection. Specifically, by integrating an image-level multi-label classifier upon the detection backbone, we can obtain the sparse but crucial image regions corresponding to categorical information, thanks to the weakly localization ability of the classification manner. Meanwhile, at the instance level, we leverage the categorical consistency between image-level predictions (by the classifier) and instance-level predictions (by the detection head) as a regularization factor to automatically hunt for the hard aligned instances of target domains. Extensive experiments of various domain shift scenarios show that our method obtains a significant performance gain over original Domain Adaptive Faster R-CNN detectors. Furthermore, qualitative visualization and analyses can demonstrate the ability of our method for attending on the key regions/instances targeting on domain adaptation. Our code is open-source and available at https://github.com/Megvii-Nanjing/CR-DA-DET.

1. Introduction

Object detection is a fundamental task in computer vision, which aims to identify and localize objects of interest in an image. In the past decade, remarkable progress has been witnessed for object detection, with the advances of large-scale benchmarks [19] and modern CNN-based detection frameworks, such as Fast/Faster R-CNN [8, 25]. However, state-of-the-art detectors require massive training images with bounding box annotations. This limits their generalization ability when facing new environments (*i.e.*, the target domain) where the object appearance, background, and even weather condition significantly differ from the training images (*i.e.*, the source domain). Meanwhile, due to the high cost of box annotations, it is not always feasible to acquire sufficient annotated training images from new environments.

In such situations, unsupervised domain adaptation offers an appealing solution by adapting object detectors from label-rich source domains to unlabeled target domains. Among a large number of methods, a promising manner for domain adaptation is to utilize the domain classifier to measure do-



(a) Source Domain (b) Target Domain

Figure 1. **First row**: Exampled images from Cityscapes [2] (source) and Foggy Cityscapes [29] (target). **Second row**: Heatmaps by the backbone network (VGG-16 [31]) of DA Faster R-CNN [1]. **Third row**: Heatmaps by the backbone network of DA Faster R-CNN trained *with* our categorical regularization framework. Our regularization framework enables more accurate alignment for crucial regions and important instances, and thus can assist the backbone network to activate the main objects of interest more accurately in *both domains*, and lead to better adaptive detection performance.

*Contributed equally.

†X.-S. Wei is the corresponding author (weixs.gm@gmail.com). This research was supported by National Key R&D Program of China (No. 2017YFA0700800).

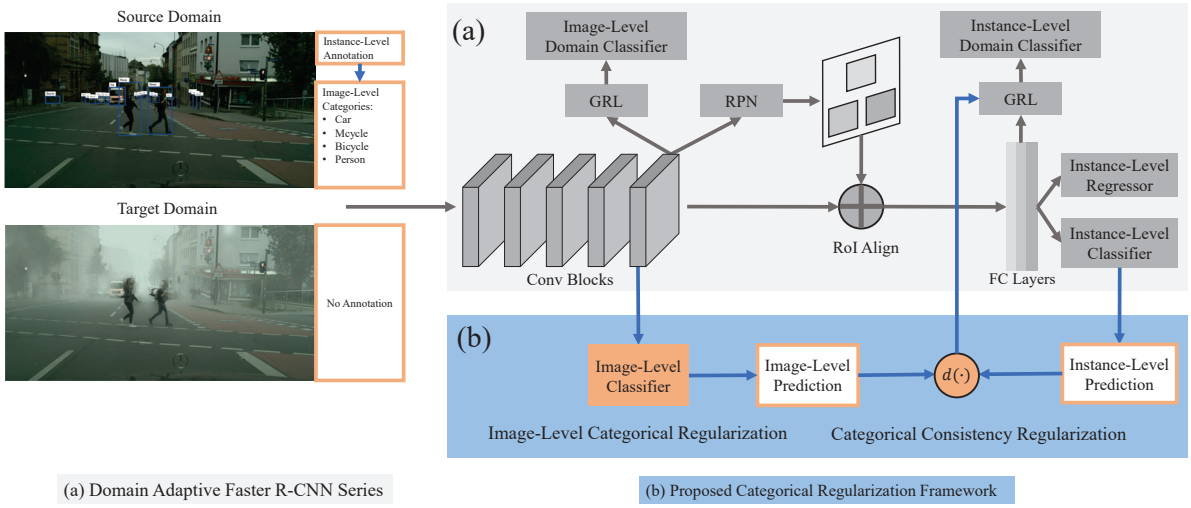


Figure 2. Overview of our categorical regularization framework: a plug-and-play component for the Domain Adaptive Faster R-CNN series [1, 28]. Our framework consists of two modules, *i.e.*, image-level categorical regularization (ICR) and categorical consistency regularization (CCR). The ICR module is an image-level multi-label classifier upon the detection backbone, which exploits the weakly localization ability of classification CNNs to obtain crucial regions corresponding to categorical information. The CCR module considers the consistency between the image-level and instance-level predictions as a novel regularization factor, which can be used to automatically hunt for hard aligned instances in the target domain during instance-level alignment.

main discrepancy, and train the domain classifier and feature extractor in an adversarial way [5, 33]. In the literature, adversarial training has been well-studied for domain adaptive image classification [5, 6, 22, 33], semantic segmentation [13, 30, 32] and object detection [1, 28, 41, 12].

Among many domain adaptive detection methods, Domain Adaptive (DA) Faster R-CNN [1] is the most representative work that integrates Faster R-CNN [25] with adversarial training. To address the domain shift problem, it aligns both the image and instance distributions across domains with adversarial training. Recently, DA Faster R-CNN has rapidly evolved into a successful series [28, 41, 12, 14]. Specifically, Saito *et al.* [28] and Zhu *et al.* [41] improved DA Faster R-CNN based on the observation that the plain image-level alignment forces to align non-transferable backgrounds, while the object detection task by nature focuses on local regions that may contain objects of interest. Furthermore, although instance-level alignment can match object proposals in both domains, current practices [1, 12] lack the ability of identifying the hard aligned instances from excessive low-value region proposals.

Aiming at these issues, we propose a novel categorical regularization framework, which can assist the Domain Adaptive Faster R-CNN series [1, 28] to focus on aligning the crucial regions and important instances cross domains. Thanks to the accurate alignment for such regions and instances, the detection backbone networks can activate objects of interest more accurately in *both domains* (cf. Figure 1), and thus lead to better adaptive object detection results.

Concretely, our framework consists of two regularization modules, *i.e.*, image-level categorical regularization (ICR), and categorical consistency regularization (CCR) (cf. Figure 2). For image-level categorical regularization, we attach the detection backbone network with an image-level multi-label classifier, and train it with categorical supervisions from the source domain. The classification manner enables the backbone to learn object-level concepts from the holistic images, without being affected by the distribution of non-transferable source backgrounds [39, 40]. It allows us to implicitly align the crucial regions on both domains at the image level. For categorical consistency regularization, we take into account the consistency between image-level predictions by the attached classifier and instance-level predictions by the detector. We adopt this categorical consistency as a novel regularization factor, and use it to increase the weights of the hard aligned instances in the target domain during instance-level alignment.

The main contributions of this work are three-fold:

- We present a novel categorical regularization framework for domain adaptive object detection, which can be applied as a plug-and-play component for the prominent Domain Adaptive Faster R-CNN series. Our framework is cost-free as requiring no further annotations, and also hyperparameter-free for performing on the vanilla detectors.

- We design two regularization modules, by exploiting the weakly localization ability of classification CNNs and the categorical consistency between image-level and instance-level predictions. They enable us to fo-

cus on aligning object-related regions and hard aligned instances that are directly pertinent to object detection.

- We conduct extensive experiments of various domain shift scenarios to validate the effectiveness of our categorical regularization framework. Our framework can significantly boost the performance of existing Domain Adaptive Faster R-CNN detectors [1, 28], and produce state-of-the-art results on benchmark datasets.

2. Preliminaries and Related Work

2.1. CNN-based Object Detection

In the past few years, the rise of deep convolutional neural networks led to a sharp paradigm shift of object detection [20]. Among a large number of approaches, the two-stage R-CNN series [9, 8, 25, 17] have become the mainstream detection framework. The pioneer work, *i.e.*, R-CNN [9], extracts region proposals from the image with low-level vision techniques [34], and applies a network to classify each region of interest (RoI) independently. Fast R-CNN [8] improves R-CNN by sharing convolutional features among RoIs, and thus enables fast training and inference. Faster R-CNN [25] advances the region proposal generation process with a Region Proposal Network (RPN). RPN shares the feature extraction backbone with the detection head, which in essence is a Fast R-CNN [8]. Faster R-CNN is a famous two-stage detection framework, and is the foundation for many follow-up works [7, 3, 17]. While recently single-stage detectors have emerged as a popular paradigm [24, 21, 18], many top-performing systems still adopt the proven two-stage pipeline [17, 10].

Thanks to the flexibility of Faster R-CNN, recently, it is widely adapted for domain adaptive object detection [1, 28, 41, 12] with adversarial training [5]. Other approaches, such as self-training [16, 26], are also exploited for domain adaptive object detection in the literature.

2.2. Domain Adaptive Faster R-CNN Series

Domain Adaptive (DA) Faster R-CNN [1] is a prominent two-stage object detector for dealing with the challenging domain adaptive object detection problem. It is an intuitive extension of Faster R-CNN [25], which aligns both the image and instance distributions by learning domain classifiers in an adversarial manner. For the image-level alignment, the domain classifier is trained on each activation (channel-wise descriptor) from the feature map after the base convolutional layers, while for instance-level alignment, the domain classifier is trained with instance-level RoI features. Furthermore, the consistency between image-level and instance-level domain classifiers is enforced to learn the cross-domain robustness for RPN.

Formally, for a given image, let $D=0$ denote that it is from the source domain while $D=1$ denote that it is from the target domain. Let $\hat{D}^{(u,v)}$ denote the output of the image-level domain classifier for the activation located at (u,v) of the feature map, then the image-level alignment loss can be written as

$$\mathcal{L}_{img}=-\sum_{u,v}\Big[D\log\hat{D}^{(u,v)}+(1-D)\log(1-\hat{D}^{(u,v)})\Big]. \quad (1)$$

Let $\hat{D}_j$ denote the output of the instance-level domain classifier for the j-th region proposal, then the instance-level alignment loss is as follows

$$\mathcal{L}_{ins}=-\sum_{j}\Big[D\log\hat{D}_j+(1-D)\log(1-\hat{D}_j)\Big]. \quad (2)$$

Furthermore, let $\mathcal{L}_{cst}$ denote the consistency loss for image-level and instance-level domain classifiers, and let $\mathcal{L}_{det}$ be the original training loss for Faster R-CNN [25]. The overall objective $\mathcal{L}_{DAF}$ for DA Faster R-CNN can be written as

$$\mathcal{L}_{DAF}=\mathcal{L}_{det}+\lambda\cdot(\mathcal{L}_{img}+\mathcal{L}_{ins}+\mathcal{L}_{cst}), \quad (3)$$

where λ is a hyper-parameter to balance the detection loss and the domain adaptation components. The adversarial training for adaptation components is implemented by the gradient reverse layer (GRL) [5], where the sign of gradients is flipped when training the base convolutional layers.

As aforementioned, DA Faster R-CNN [1] may fail to align the crucial regions and important instances which are crucial for adaptive detection. Meanwhile, it tends to fit the distribution of non-transferable source backgrounds, as the training process involves a large amount of background proposals. Recent works attempted to improve DA Faster R-CNN by replacing the plain image-level alignment model with a weak alignment model [28] or a region-level alignment model [41], and found that the instance-level alignment model is not necessary in presence of other local alignment model [28]. We term to these methods collectively as Domain Adaptive Faster R-CNN series.

A high-level diagram of Domain Adaptive Faster R-CNN series is shown in Figure 2 (a), where we follow the paradigm of DA Faster R-CNN [1] but omit the part of $\mathcal{L}_{cst}$ which is not an essential ingredient in our regularization framework. Please note that Figure 2 (a) is a conceptual diagram, and not all components of the Domain Adaptive Faster R-CNN series strictly follow this structure.

2.3. Weakly Localization by Classification CNNs

It is widely acknowledged that CNNs trained for single-label image classification tend to produce high responses on the local regions containing the main objects [38, 40, 39]. Analogously, CNNs trained for multi-label classification also have the weakly localization ability for the objects associated with image-level categories [35, 36].

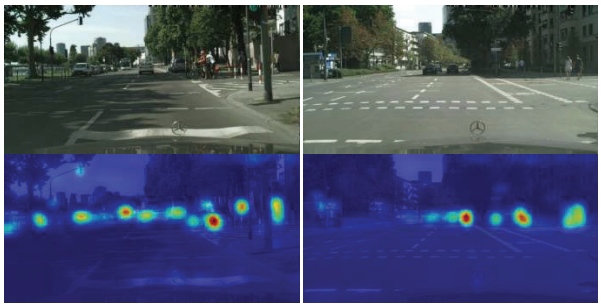


Figure 3. Visualization of the weakly localization ability of multi-label classification CNNs. The CNN model is VGG-16 trained on Cityscapes [2].

Taking the Cityscapes [2] dataset for an example, we collect all instance-level labels into an image-level label vector, and train VGG-16 [31] for multi-label image classification. Figure 3 shows the heatmaps for two exampled images from Cityscapes, where the main objects related to image-level categories such as "car", "person" and "rider" are weakly localized.

3. Approach

3.1. Framework Overview

The overview of our categorical regularization framework is illustrated in Figure 2. In general, our framework improves the DA Faster R-CNN series detectors [1, 28, 12] by exploring categorical regularization from two aspects: image-level categorical regularization (ICR) and categorical consistency regularization (CCR). Note that the ICR module does not depend on the CCR module, and thus it can be individually integrated with DA Faster R-CNN detectors which only perform image-level alignment [28].

Our framework enables better alignment of crucial regions and important instances across domains. Consequently, the detection backbone produces more accurate activations on objects of interest of both domains (cf. Figure 1), leading to better adaptive detection performance. Our framework is flexible and generalizable – it does not depend on specific algorithms for either image or instance alignment.

3.2. Image-Level Categorical Regularization

Image-level categorical regularization (ICR) is exploited to obtain the sparse but crucial image regions corresponding to categorical information. We achieve this with a weakly supervised solution, which can learn discriminative features for objects of interest, without being affected by the distribution of non-transferable source backgrounds. While the standard training for Faster R-CNN can learn discriminative features for objects of interest, it tends to fit the source backgrounds due to the large amount of background RoIs sampled for training. Since the patterns of source backgrounds are non-transferable, plain image-level alignment may lead to noisy activations in target domains (cf. Figure 1).

In our proposal, as illustrated in Figure 2 (b), we attach the detection backbone with an image-level multi-label classifier, and train it with supervisions from the source domain. Such categorical supervisions are cost-free for detection datasets, and can be easily acquired by collecting all instance-level categories in an image into an image-level categorical vector.

Given the detection backbone network, we perform global average pooling on the output of the last convolutional layer, and feed the pooled features into a plain multi-label classifier implemented by a 1×1 convolution. We train this image-level classifier with the standard cross-entropy multi-label loss by

$$\mathcal{L}_{ICR} = \sum_{c=1}^{C} y^c \log(\hat{y}^c) + (1-y^c)\log(1-\hat{y}^c), \quad (4)$$

where C is the total number of categories of a detection dataset, y^c is the ground truth label, and $\hat{y}^c$ is the predicted one. $y^c = 1$ denotes that there is at least one object of category c appearing in this image, while $y^c = 0$ means there is no object of category c in the image.

The image-level categorical supervisions encourage the detection backbone to learn category-specific features that can activate object-related regions. This allows us to align the crucial regions of both domains with an image-level alignment model (*e.g.*, Equation (2)). Meanwhile, because there is no background supervision involved in the training process of our image-level multi-label classifier, the risk of fitting (even over-fitting) non-transferable source backgrounds is greatly reduced.

3.3. Categorical Consistency Regularization

We design a categorical consistency regularization (CCR) module to automatically hunt for the hard aligned instances in target domains. Our motivation lies in two aspects. First, current instance alignment models [1, 12] may be dominated by the excessive low-value background proposals, as they can not identify the hard foreground instances in the target domain. Second, the attached image-level classifier and the instance-level detection head are complementary, because the former exploits the whole image-level context while the latter enjoys more accurate RoI features.

Building upon those above considerations, we adopt the categorical consistency between the image-level and instance-level predictions as a measure for the hardness of classifying a certain target instance. Intuitively, if the image-level classifier predicts that there is no "person" in a target image while the detection head classifies a certain instance as "person", this instance should be a hard but informative sample for current detection model. Therefore, we utilize this consistency as a regularization factor to increase the weight of hard aligned samples in target domains during instance-level alignment.

Specifically, assume that the detection head classifies the j-th instance in a target image as category c, we let $\hat{p}_j^c$ denote the estimated probability. Using the notation in Equation (4), we let $\hat{y}^c$ denote the image-level estimation of the probability that this image contains objects of category c. We define the following distance function to measure the categorical consistency between the instance-level and image-level predictions as

$$d_j = e^{|\hat{p}_j^c - \hat{y}^c|}. \quad (5)$$

Here the exponent form characterizes the intuition that while a small disagreement may come from the model's variance, a large disagreement should be attributed to the hardness in classifying this instance.

We use Equation (5) to weight the instance-level adversarial loss, which in implementation is equivalent to weight the gradients passed through the gradient reversal layer (GRL) during training. Take the instance alignment model (*i.e.*, Equation (2)) in DA Faster R-CNN [1] for an example, the instance-level alignment loss with CCR can be written as

$$\mathcal{L}_{ins}^{CCR} = -\sum_j d_j \left[D \log \hat{D}_j + (1-D) \log(1-\hat{D}_j) \right]. \quad (6)$$

It is worth noting that, we only apply Equation (5) to weight *foreground* instances from the target domain, according to the predictions of detection head. We keep the weights for source instances and the background instances from the target domain unchanged (*i.e.*, $d_j = 1$), as the former have supervision signals from the source domain, while the latter are not as important as foreground proposals.

3.4. Integration with DA Faster R-CNN Series

In this work, we take the DA Faster R-CNN [1] and the state-of-the-art strong-weak aligned Faster R-CNN [28] as our baseline detectors. In the following, we term them as "DA-Faster" and "SW-Faster" for simplicity. In fact, other Domain Adaptive Faster R-CNN detectors [12, 41] may also be compatible with our framework with minor modifications.

Integration with DA-Faster. Integrating our framework with DA-Faster [1] is straightforward. We attach an image-level multi-label classifier to the backbone, by adding a global averaging pooling layer and a 1×1 convolution layer. Furthermore, we use our CCR to weight the gradients passed through the reverse gradient layer (GRL) for instance-level alignment. The modified overall objective of DA-Faster with our regularization framework can be written as

$$\mathcal{L}_{DAF}^* = \mathcal{L}_{det} + \mathcal{L}_{ICR} + \lambda \cdot (\mathcal{L}_{img} + \mathcal{L}_{ins}^{CCR} + \mathcal{L}_{cst}), \quad (7)$$

where λ is set to 0.1 in [1], and our method does not introduce additional hyper parameters.

Integration with SW-Faster. SW-Faster [28] improves the strong image-level alignment model of DA-Faster with a weak global alignment model, and replaces the instance-level alignment model with a strong local alignment model. Since our categorical regularization framework is independent of the specific algorithms for alignment, our ICR module can be straightly integrated into SW-Faster. Furthermore, we add an instance-level alignment model, which is the same to that of DA-Faster, into the pipeline of SW-Faster during training. This allows us to apply our CCR module to further improve SW-Faster. The modified overall objective for SW-Faster with our regularization framework can be written as

$$\mathcal{L}_{SWF}^* = \mathcal{L}_{det} + \mathcal{L}_{ICR} + \lambda' \cdot (\mathcal{L}_{ins}^{CCR} + \mathcal{L}_{global} + \mathcal{L}_{local}), \quad (8)$$

where λ' is set to 1.0, and $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ denote the global alignment loss and local alignment loss in [28].

4. Experiments

4.1. Empirical Setup

Datasets. Five public datasets are utilized in our experiments, including Cityscapes [2], Foggy Cityscapes [29], BDD100k [37], PASCAL VOC [4], and Clipart1k [15].

- **Cityscapes** [2] focuses on capturing high variability of outdoor street scenes in common weather conditions from different cities. It contains 2,975 training images and 500 validation images with dense pixel-level labels. We transform the instance segmentation annotations into bounding boxes for our experiments.
- **Foggy Cityscapes** [29] is built upon the images in the Cityscapes dataset [2]. This dataset simulates the foggy weather using depth maps provided in Cityscapes with three levels of foggy weather, and thus is suitable to conduct weather adaptation experiments.
- **BDD100k** [37] consists of 100k images, with 70k training images and 10k validation images annotated with bounding boxes. We extract a subset of BDD100k with images labeled as *daytime*, including 36,728 training and 5,258 validation images. We use this subset for scene adaptation experiments.
- **PASCAL VOC** [4] is a real-world dataset containing 20 categories of common objects with bounding box annotations. Following [28], we employ PASCAL VOC 2007 and 2012 training and validation images (16,551 images in total) for experiments.
- **Clipart1k** [15] contains 1k clipart images, which shares the same instance categories with PASCAL VOC but exhibits a large domain shift. We follow the practice in [28], and use all images of Clipart1k for both training (without labels) and test.

Baselines and Comparison Methods. We consider DA-Faster [1] and the state-of-the-art SW-Faster [28] as our

Table 1. **Weather Adaptation**: Results on Foggy Cityscapes, using models trained on Cityscapes.

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Faster R-CNN (Source)	24.4	30.5	32.6	10.8	25.4	9.1	15.2	28.3	22.0
MA-Faster [12]	28.4	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
Selective-Faster [41]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
DA-Faster [1]	28.7	36.5	43.5	19.5	33.1	12.6	24.8	29.1	28.5
DA-Faster-ICR (Ours)	28.7	37.3	43.0	21.9	36.9	9.2	25.9	31.9	29.4
DA-Faster-ICR-CCR (Ours)	29.7	37.3	43.6	20.8	37.3	12.8	25.7	31.7	29.9
SW-Faster [28]	32.3	42.2	47.3	23.7	41.3	27.8	28.3	35.4	34.8
SW-Faster-ICR (Ours)	**33.1**	**44.2**	48.8	**27.7**	44.9	27.9	29.4	**36.2**	36.5
SW-Faster-ICR-CCR (Ours)	32.9	43.8	**49.2**	27.2	**45.1**	**36.4**	**30.3**	34.6	**37.4**
Faster R-CNN (Oracle)	36.2	47.7	53.0	34.7	51.9	41.0	36.8	37.8	42.4

Table 2. **Scene Adaptation**: Results of **7 common categories** on the *daytime* subset of BDD100k, using models trained on Cityscapes.

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Faster R-CNN (Source)	26.9	22.1	44.7	17.4	16.7	-	17.1	18.8	23.4
DA-Faster [1]	29.4	26.5	44.6	14.3	16.8	-	15.8	20.6	24.0
DA-Faster-ICR (Ours)	29.1	28.6	44.8	14.9	15.8	-	17.1	22.4	24.7
DA-Faster-ICR-CCR (Ours)	29.3	28.4	45.3	17.5	17.1	-	16.8	22.7	25.3
SW-Faster [28]	30.2	29.5	45.7	15.2	18.4	-	17.1	21.2	25.3
SW-Faster-ICR (Ours)	30.9	31.2	45.6	15.9	18.4	-	**19.3**	23.7	26.4
SW-Faster-ICR-CCR (Ours)	**31.4**	**31.3**	**46.3**	**19.5**	**18.9**	-	17.3	**23.8**	**26.9**
Faster R-CNN (Oracle)	35.3	33.2	53.9	46.3	46.7	-	25.6	29.3	38.6

baseline methods, and re-implement them for fair comparisons. Our re-implementations achieve comparable or even better accuracies compared to the original papers. When comparing with other state-of-the-art methods, we report the results from original papers. Furthermore, we also train Faster R-CNN [25] only using source images, as well as directly using annotated target images. We refer to models of these two settings as "Faster R-CNN (Source)" and "Faster R-CNN (Oracle)", respectively.

Implementation Details. Following the default settings in [1, 28], all training and test images are resized such that the shorter side has a length of 600 pixels. By default, the backbone models are initialized using pre-trained weights of VGG-16 [31] on ImageNet, but for the dissimilar domain adaptation experiments from PASCAL VOC [4] to Clipart1k [15], we follow the practices in [28] and use ResNet-101 [11] as the detection backbone. We fine-tune the network with a learning rate of 1×10^{-3} for 50k iterations and then reduce the learning rate to 1×10^{-4} for another 20k iterations. Each batch is composed of two images, one from source and another from target. The momentum of 0.9 and the weight decay of 5×10^{-4} is used for VGG-16 based detectors, while for ResNet-101 based detectors, we set the weight decay as 1×10^{-4}. In all experiments, we employ RoIAlign [10] for RoI feature extraction.

4.2. Comparison Results

Weather Adaptation. In real-world scenarios, object detectors may be applied under different weather conditions. We study the weather adaptation from clear weather to a foggy environment, using Cityscapes' training set and Foggy Cityscapes' validation set as the source domain and the target domain, respectively.

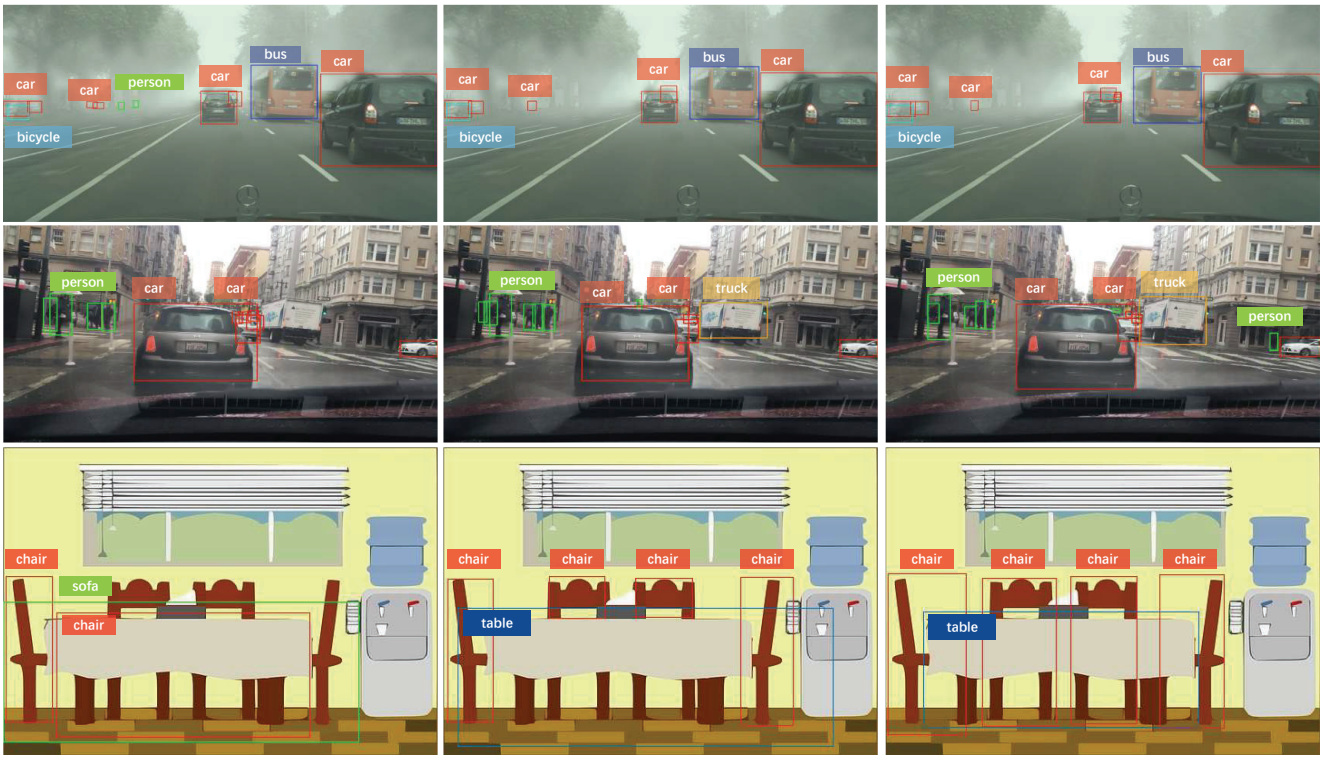
Table 1 shows the comparison results. Our categorical regularization framework can consistently boost the performance of DA-Faster and SW-Faster detectors, with 1.4% and 2.6% mAP improvements, respectively. In particular, our CCR module can greatly improve the detection results for some difficult categories such as "train". It clearly verifies the importance of increasing the weight of hard foreground instances in target domains for instance-level alignment. It is worth noting that our categorical regularization framework helps to reduce the performance gap between the domain adaptive detector and oracle detector trained with annotated target images to about 5% mAP.

Scene Adaptation. Scene layout changes frequently occur in real-life applications of object detection, *e.g.*, automatic driving from one city to another. To study the effectiveness of our regularization framework for scene adaptation, we choose the Cityscapes [2] training set as the source domain and a subset of BDD100k [37] as the target domain. In particular, we choose a subset of the BDD100k dataset annotated as *daytime* to be our target domain and consider the city scene as the adaptation factor, since there only exists daytime data in the Cityscapes dataset. We report the detection results on seven common categories on both datasets.

As shown in Table 2, we observe a significant performance gap between the domain adaptive detectors and the oracle detector, which suggests that scene layout shift is a challenging factor that hinders the performance of domain adaptive detection. Even under this difficult setting, our categorical regularization framework can also improve DA-

Table 3. **Dissimilar Domain Adaptation**: Results on the Clipart1k dataset, using models trained on the PASCAL VOC training set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Faster R-CNN (Source)	21.9	42.2	22.9	19.0	30.8	43.1	28.9	10.7	27.4	18.1	13.5	10.3	25.0	50.7	39.0	37.4	6.9	18.1	39.2	34.9	27.0
Kim *et al.* [16]	28.0	**64.5**	23.9	19.0	21.9	64.3	**43.5**	**16.4**	**42.2**	25.9	**30.5**	7.9	25.5	67.6	54.5	36.4	10.3	**31.2**	**57.4**	43.5	35.7
DA-Faster [1]	**38.0**	47.5	27.7	24.8	41.3	41.2	38.2	11.4	36.8	39.7	19.6	12.7	31.9	47.8	55.6	46.3	12.1	25.6	51.1	45.5	34.7
DA-Faster-ICR (Ours)	31.0	53.9	29.2	**28.2**	**41.5**	56.6	38.3	8.1	37.4	43.1	22.0	12.4	27.8	49.8	55.0	**48.2**	11.0	22.7	54.2	46.9	35.9
DA-Faster-ICR-CCR (Ours)	30.2	57.0	30.6	26.2	38.0	57.1	36.1	12.7	36.4	44.8	18.2	14.6	30.0	56.7	56.6	45.9	17.8	25.3	50.5	**48.5**	36.7
SW-Faster [28]	29.2	53.1	30.2	24.4	41.4	52.5	34.6	14.0	36.3	43.5	17.6	**16.6**	**33.4**	**78.1**	59.1	42.1	15.8	24.9	45.5	43.7	36.8
SW-Faster-ICR (Ours)	25.2	54.0	31.7	23.4	40.3	**65.8**	35.4	12.1	37.6	48.1	18.6	14.2	31.3	73.6	59.9	46.5	19.5	25.9	46.0	45.6	37.7
SW-Faster-ICR-CCR (Ours)	28.7	55.3	**31.8**	26.0	40.1	63.6	36.6	9.4	38.7	**49.3**	17.6	14.1	33.3	74.3	**61.3**	46.3	**22.3**	24.3	49.1	44.3	**38.3**



(a) SW-Faster (b) SW-Faster-ICR-CCR (**Ours**) (c) Ground Truth

Figure 4. Detection examples from three target datasets, from top to bottom: Foggy Cityscapes [29], BDD100k [37], and Clipart1k [15]. Our categorical regularization framework enables SW-Faster [28] to produce more accurate detection results with large domain shifts.

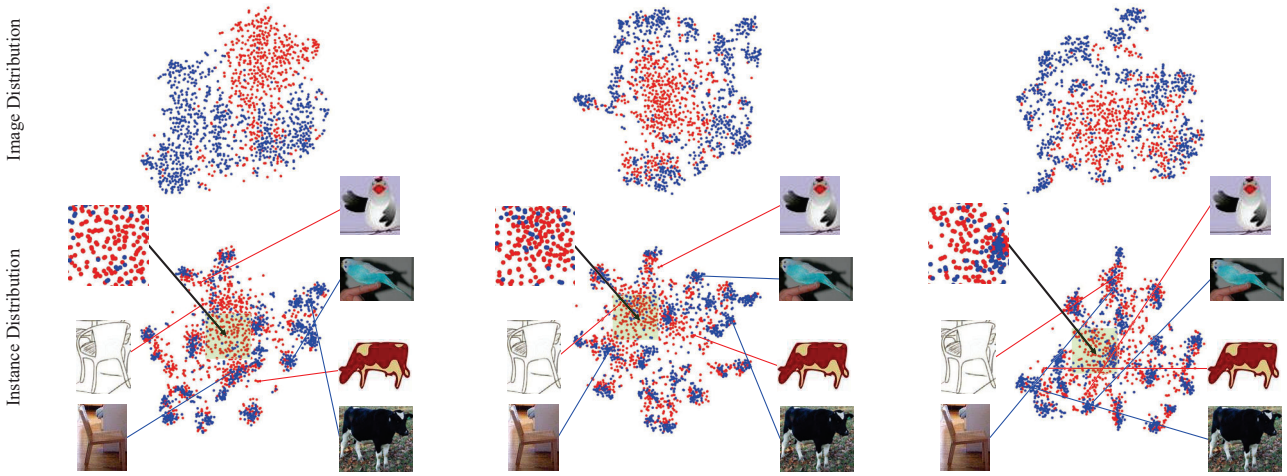
Faster and SW-Faster by 1.3% and 1.6%, respectively. Similar to the observations on weather adaptation experiments, our CCR module can significantly improve the detection results of some difficult objects such as "truck".

Dissimilar Domain Adaptation. Both weather adaptation and scene adaptation can be considered as adaptation between similar domains. We further show experiments on the dissimilar domain adaptation from real images to artistic images. We utilize Pascal VOC [4] as the real source domain and the Clipart1k [15] as the target domain. Clipart1k contains 1k comical images in total, which have the same 20 categories as PASCAL VOC. Following [28], all images in Clipart1k are used for both training (without labels) and testing, and thus there is no oracle detector for this dataset.

As shown in Table 3, for dissimilar domain adaptation, our regularization framework also achieves considerable improvements over the baseline DA-Faster and SW-Faster by 2.0% and 1.5% mAP, respectively. Furthermore, our methods also outperform recent state-of-the-art one-stage object adaptive detector [16] that employs self training for domain adaptation.

4.3. Visualization and Analyses

Detection Examples. In Figure 4, we show some detection examples from three target datasets, *i.e.*, Foggy Cityscapes [29], BDD100k [37] and Clipart1k [15]. Compared to the baseline SW-Faster [28] method, our SW-Faster-ICR-CCR method produces more accurate detection results under complex environments and large domain shifts.



(a) SW-Faster (mAP: 36.8) (b) SW-Faster-ICR (**Ours**) (mAP: 37.7) (c) SW-Faster-ICR-CCR (**Ours**) (mAP: 38.3)

Figure 5. Visualization of image features and instance features with t-SNE [23], where the blue points represent source samples from PASCAL VOC [4] and the red ones represent target samples from Clipart1k [15]. **Top Row:** Holistic image features obtained by applying global average pooling to the output of the detection backbone network. **Second Row:** Instance features obtained by applying RoIAlign on the ground truth instances, where we also show three pairs of instances from different domains, and zoom in to the local regions of the most poorly matched instances. Compared to original SW-Faster [28], our method better aligns both the image-level and instance-level features on both domains, and enables two dissimilar instances of the same category from different domains to stay close in the feature space.

Feature Visualization. We visualize the image and instance features learned for dissimilar domain adaptation (from PASCAL VOC [4] to Clipart1k [15]) using t-SNE [23]. For this experiment, we randomly sample 100 ground truth instances for each category, 50 from the source domain and 50 from the target domain. For some categories that have less than 50 instances in a certain domain, we sample all instances in that domain and the same number of instances from the other domain. The images containing these instances are sampled for image-level visualization. The image features are extracted by applying global average pooling on the output of the detection backbone network, while the instance features are extracted by RoIAlign.

As shown in Figure 5, the blue points represent source samples and the red ones represent target samples. We also show three pairs of instances from different domains, and zoom in to the local regions of the most poorly matched instances. The dissimilar instance pairs of the same category from different domains stay closer in the feature space of our methods. Even for the most poorly matched region, our method still have better alignment performance than the baseline SW-Faster method [28]. Furthermore, thanks to the accurate instance-level alignment, our image-level alignment performance is also better than the baseline method.

Domain Distance. Besides visualization understanding, we also calculate a quantitative metric for domain distance, where both domains are represented by object instances. For this experiment, we use the same instance samples as the feature visualization experiment. Specifically, we adopt Earth Mover's Distance (EMD) [27] as the metric for measuring domain distance. With this metric, domain distance computed for SW-Faster [28], SW-Faster-ICR and SW-Faster-ICR-CCR are 8.84, 8.59, 8.15, respectively.

The consistency between domain distance and model accuracy verifies the motivation of our work. That is, domain adaptive object detection relies heavily on aligning the crucial local regions and important instances on both domains. Our regularization framework assists the DA Faster R-CNN series to achieve this goal.

5. Conclusions

In this work, we presented a categorical regularization framework upon Domain Adaptive Faster R-CNN series for improving the adaptive detection performance. Specifically, we exploited the weakly localization ability of multi-label classification CNNs and the categorical consistency between image-level and instance-level predictions, which allows us to focus on aligning object-related local regions and hard aligned instances. In experiments, our framework significantly boosted the performance of existing Domain Adaptive Faster R-CNN detectors and produced state-of-the-art results on public benchmark datasets. Visualization and analyses can validate the effectiveness of our method. In the future, we will investigate how to apply our regularization framework to improve adaptive detectors beyond the Domain Adaptive Faster R-CNN series.

References

[1] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2, 3, 4, 5, 6, 7

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1, 4, 5, 6

[3] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016. 3

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, (88):303–338, 2010. 5, 6, 7, 8

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 2, 3

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 2

[7] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware CNN model. In *CVPR*, pages 1134–1142, 2015. 3

[8] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 1, 3

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 3

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 3, 6

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

[12] Zhenwei He and Lei Zhang. Multi-adversarial Faster-RCNN for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019. 2, 3, 4, 5, 6

[13] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[14] Han-Kai Hsu, Wei-Chih Hung, Hung-Yu Tseng, Chun-Han Yao, Yi-Hsuan Tsai, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *CVPR Workshops*, pages 1–5, 2019. 2

[15] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 5, 6, 7, 8

[16] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6092–6101, 2019. 3, 7

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 1

[20] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *IJCV*, (128):261–318, 2020. 3

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016. 3

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

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 9(11):2579–2605, 2008. 8

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 3

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1, 2, 3, 6

[26] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *CVPR*, pages 780–790, 2019. 3

[27] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 8

[28] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 2, 3, 4, 5, 6, 7, 8

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 1, 5, 7

[30] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, pages 3752–3761, 2018. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–8, 2015. 1, 4, 6

[32] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018. 2

[33] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017. 2

[34] Jasper Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 3

[35] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. CNN-RNN: A unified framework for multi-label image classification. In *CVPR*, pages 2285–2294, 2016. 3

[36] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. HCP: A flexible CNN framework for multi-label image classification. *IEEE TPAMI*, 38(9):1901–1907, 2015. 3

[37] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 5, 6, 7

[38] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014. 3

[39] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, pages 1–10, 2015. 2, 3

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2, 3

[41] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, and Jianping Shi. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 2, 3, 5, 6