SimROD: A Simple Adaptation Method for Robust Object Detection

Rindra Ramamonjison[1], Amin Banitalebi-Dehkordi[1], Xinyu Kang[2], Xiaolong Bai[3], and Yong Zhang[1]

[1]Huawei Technologies Canada Co., Ltd
[2]University of British Columbia
[3]Huawei Cloud
rindranirina.ramamonjison@huawei.com, amin.banitalebi@huawei.com, xinyu.kang@alumni.ubc.ca, baixiaolong1@huawei.com,
yong.zhang3@huawei.com

Abstract

This paper presents a Simple and effective unsupervised adaptation method for Robust Object Detection (SimROD). To overcome the challenging issues of domain shift and pseudo-label noise, our method integrates a novel domain-centric data augmentation, a gradual self-labeling adaptation procedure, and a teacher-guided fine-tuning mechanism. Using our method, target domain samples can be leveraged to adapt object detection models without changing the model architecture or generating synthetic data. When applied to image corruptions and high-level cross-domain adaptation benchmarks, our method outperforms prior baselines on multiple domain adaptation benchmarks. SimROD achieves new state-of-the-art on standard real-to-synthetic and cross-camera setup benchmarks. On the image corruption benchmark, models adapted with our method achieved a relative robustness improvement of 15-25% AP50 on Pascal-C and 5-6% AP on COCO-C and Cityscapes-C. On the cross-domain benchmark, our method outperformed the best baseline performance by up to 8% and 4% AP50 on Comic and Watercolor respectively.[1]

1. Introduction

State-of-the-art object detection models are highly accurate when trained on images that have the same distribution as the test set [39]. However, they can fail when deployed to new environments due to domain shifts such as weather changes (e.g. rain or fog), light condition variations, or image corruptions (e.g. blur) [25]. Such failure is detrimental for mission-critical applications such as self-driving or automated retail checkout, in which domain shifts are inevitable. To make them reliable, it is important for detection models to be robust to domain shifts.

Different types of methods have been proposed to overcome domain shifts for object detection namely data augmentation [25, 14, 12], domain-alignment [6, 11, 38, 37, 27, 16, 23, 17], domain-mapping [3, 18, 23, 17], and self-labeling techniques [33, 30, 22, 18]. Augmentation methods can improve the performance on some fixed set of domain shifts but fail to generalize to the ones that are not similar to the augmented samples [1, 26, 32]. Domain-aligning methods use target domain samples to align intermediate features of networks. These methods require the addition of specialized modules such as gradient reversal layers, domain classifiers to the model. On the other hand, domain-mapping methods translate labeled source images to new images that look like target domain images using image-to-image translation networks. Similar to augmentation methods, they are suboptimal since the generated images do not always have a high similarity to real target domain images. Finally, self-labeling is a promising approach since it leverages unlabeled training samples form the target domain. However, generating accurate pseudo-labels under domain shift is hard; and when pseudo-labels are noisy, using target domain samples for adaptation is ineffective.

In this paper, we propose a Simple adaptation method for Robust Object Detection (SimROD), to mitigate the domain shifts using domain-mixed data augmentation and teacher-guided gradual adaptation. Our simple approach has three design benefits. First, it does not require ground-truth labels of target domain data and leverage unlabeled samples. Second, our approach requires neither complicated architecture changes nor generative models for creating synthetic data [18]. Third, our simple method is architecture-agnostic and is not limited to region-based detectors. The main contributions of this paper are summarized as follows:

1. We propose a simple method to improve the robustness of object detection models against domain shifts. Our

[1] Code and notebook are available at https://marketplace.huaweicloud.com/markets/aihub/notebook/detail/?id=d6d7162f-32b9-483d-97d7-b16b32b148e2

method first adapts a large teacher model using a gradual adaptation approach. The adapted teacher generates accurate pseudo-labels for adapting the student model.

2. We introduce a data augmentation called *DomainMix* for learning domain-invariant representations and for reducing the pseudo-label noise. It efficiently mixes the labeled source domain images with unlabeled samples from the target domain along with their (pseudo-)labels. The mixed training samples give strong supervision for adapting both the teacher and student models.
3. We conduct a comprehensive benchmark and ablation studies to demonstrate the effectiveness of SimROD in mitigating different domain shifts namely synthetic-to-real, cross-camera setup, real-to-artistic, and image corruptions. Our simple method are competitive with more complicated baselines and achieve new state-of-the-art results on some of these benchmarks.

2. Motivation and related works

In this section, we review the mainstream approaches relevant to our work and explain the motivation of our work.

Data augmentations for robustness to image corruption Data augmentation is an effective technique for improving the performance of deep learning models. Recent works have also explored the role of augmentation in enhancing the robustness to domain shifts. In particular, specialized augmentations have been proposed to combat the effect of image corruptions for image classification [13, 14, 12] and object detection [25, 8]. For example, AugMix [14] samples a set of geometric and color transformations which are applied sequentially to each image and mixes the original image with multiple augmented copies. DeepAugment [12] generates augmented samples using image-to-image translation networks whose weights are perturbed with random distortions. [25, 8] proposed style transfer [10] as augmentation for increasing the shape bias and improve robustness.

While these augmentation methods offer some improvement over the source baseline, they can overfit to few corruption types and fail to generalize to others. In fact, [1] provided empirical evidence that the perceptual similarity between the augmentation transformation and the corruption is a strong predictor of corruption error. [1] also observed that broader augmentation schemes perform better on dissimilar corruptions than more specialized ones. [32] showed that augmentation techniques that are tailored to synthetic corruptions have difficulty to generalize to natural distributions shifts. In their extensive study, training on more diverse data was the only intervention that effectively improved the robustness to natural distribution shifts.

Unsupervised domain adaptation for object detection Unsupervised domain adaptation (UDA) methods leverage unlabeled images from the target domain to explicitly mitigate the domain shift. In contrast to images obtained with augmentation, these unlabeled samples are more similar to the test samples. Moreover, they are cheap to collect and do not require a laborious annotation.

Several approaches have been proposed to solve the UDA problem for object detection. Adversarial training methods such as [6] learn domain-invariant representations of two-stage detector networks. Recent methods improved the performance, by mining important regions and aligning at the region-level [11], by using hierarchical alignment module [38], by coarse-to-fine feature adaptation [37], or by enforcing strong local alignment and weak global alignment [27]. [16] proposed a center-aware alignment method for anchor-free FCOS model. While alignment methods help reduce the domain shift, they require architecture changes since extra modules such as gradient reversal layers and domain classifiers must be added to the network.

Alternatively, domain-mapping methods tackle UDA by first translating source images to images that resemble the target domain samples using a conditional generative adversarial network (GAN) [3, 15]. The model is then fine-tuned with the domain-mapped images and the known source labels. For object detection, [23, 17] combined domain transfer with adversarial training. For instance, [23] generates a diverse set of intermediate domains between the source and target to discriminate and learn domain-invariant features.

Finally, recent works have shown that adapting batch normalization [19] layers can improve robustness to adversarial attacks [35] or image corruptions [28] and reduce domain shifts [24, 5].

Self-training for object detection adaptation Self-training enables a model to generate its own pseudo-labels on the unlabeled target samples. Recently, [30] applied pseudo-labeling in the STAC framework for semi-supervised object detection However, pseudo-labeling can degenerate the performance in the presence of domain shift since the pseudo-labels on target samples may become incorrect leading to poor supervision. Instead, our work tackles the domain shift between the original source training data and the unlabeled target training data. To reduce domain shift, [4] enforced region-level and graph-structures consistencies between a mean teacher model and the student model using additional regularization loss functions. Next, [22] proposed a method to directly mitigate the noisy pseudo-labels of Faster-RCNN detectors by modeling their proposal distribution. Unlike [22], our method is agnostic to the model architecture and can also work with single-stage object detectors too. Finally, [18] combined domain transfer with pseudo-labeling and is also architecture-agnostic.

Compared to prior works, our proposed method is simpler because it does not generate synthetic data using GANs, nor change the training loss function or model architecture. As will be shown in Section 4, our simple method

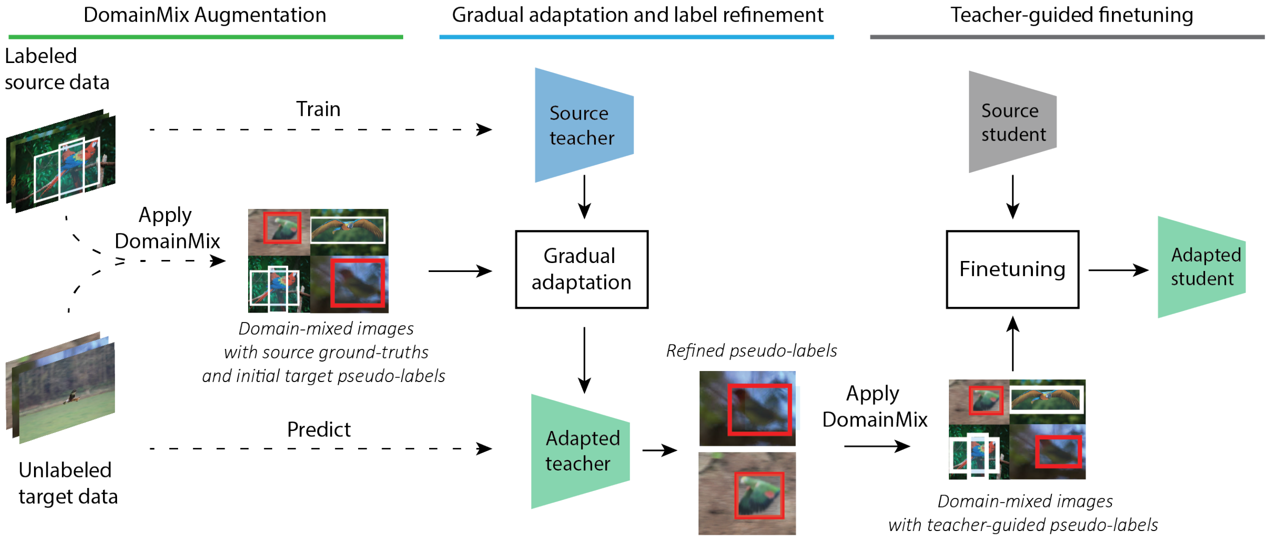


Figure 1. Our proposed adaptation method for robust object detection mitigates the domain shift and label noise using three simple steps. (1) The proposed DomainMix augmentation module randomly samples and mixes images from both the source and target domains along with their ground-truth and pseudo-labels. (2) These domain-mixed images are used to gradually adapt the batch norm and convolutional layers of a large source teacher model. During this step, the pseudo-labels of the target domain images are also refined. (3) New domain-mixed images with the refined pseudo-labels are used to finetune the source student model.

is very effective in reducing domain shifts and label noise.

3. Problem definition and proposed solution

In this section, we define the adaptation problem and describe our proposed solution.

3.1. Problem statement

We are given a source model $\mathtt{M}$ for an object detection task with parameters $\theta^s_{\mathtt{M}}$, which is trained with a source training dataset $\mathcal{D}{=}\{(\mathbf{x}_i, \mathbf{y}_i)\}$, where $\mathbf{x}_i$ is an image and each label y_i consists of object categories and bounding box coordinates. We consider scenarios in which there exists a covariate shift between the input distribution $p_S : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ of the original source data $\mathcal{D}$ and the target test distribution $p_T : \overline{\mathcal{X}} \times \mathcal{Y} \rightarrow \mathbb{R}^+$. More formally, we assume that $p_S\left(\mathbf{y} \mid \mathbf{x}\right) = p_T\left(\mathbf{y} \mid \mathbf{x}\right)$ but $p_S\left(\mathbf{x}\right) \neq p_T\left(\mathbf{x}\right)$ [31].

In the unsupervised domain adaptation setting, we are also given a set of unlabeled images $\overline{\mathcal{D}} = \{(\overline{\mathbf{x}}_j)\}$ from the target domain, which we can use during training. Therefore, our objective is to update the model parameters $\theta^s_{\mathtt{M}}$ into $\theta^a_{\mathtt{M}}$ to achieve a good performance on both the source test set and a given target test set, i.e., improving its robustness to the domain shifts. To effectively exploit the additional information in $\overline{\mathcal{D}}$, we need to tackle two inter-related issues. First, the target training set $\overline{\mathcal{D}}$ does not come with ground-truth labels. Second, generating pseudo-labels for $\overline{\mathcal{D}}$ with the source model $\theta^s_{\mathtt{M}}$ leads to noisy supervision due to the domain shift and hinders the adaptation. In the following subsections, we present a simple approach for tackling these technical issues.

3.2. Simple adaptation for Robust Object Detection

We present our simple adaptation method SimROD for enabling robust object detection models. SimROD integrates a teacher-guided fine-tuning, a new DomainMix augmentation method and a gradual adaptation technique. Sec. 3.2.1 describes the overall method. Next, Sec. 3.2.2 presents the DomainMix augmentation, which is used for adapting both the teacher and student. Finally, Sec. 3.2.3 explains the gradual adaptation that overcomes the two interrelated issues of domain shift and pseudo-label noise.

3.2.1 Overall approach

Our simple approach is motivated by the fact that label noise is exacerbated by the domain shift. Therefore, our approach aims to generate accurate pseudo-labels on target domain images and use them together with mixed images from source and target domain so as to provide strong supervision for adapting the models.

Because the student target model may not have the capacity to generate accurate pseudo-labels and adapt itself, we propose to adapt an auxiliary teacher model first, which can later generate high-quality pseudo-labels for fine-tuning the student model. A flow diagram of SimROD is provided in Figure 1. Its steps are summarized as follows:

Step 1: We train a large source teacher model $\mathtt{T}$ with bigger capacity than the student model $\mathtt{M}$ to be adapted using the source data $\mathcal{D}$ and get parameters $\theta^s_{\mathtt{T}}$. The source teacher is used to generate initial pseudo-labels on target data.

Step 2: We adapt the large teacher model parameters from $\theta^s_{\mathtt{T}}$ to $\theta^a_{\mathtt{T}}$ using the gradual adaptation of Algorithm 2 (see Sec. 3.2.3). During this step, we use mixed images generated by the DomainMix augmentation (see Sec. 3.2.2)

Step 3: We refine the pseudo-labels on the target data $\overline{\mathcal{D}}$ using the adapted teacher model parameters $\theta^a_{\mathtt{T}}$. Then, we fine-tune the student model $\mathtt{M}$ using these pseudo-labels in line 2 and 8 of Algorithm 2.

One benefit of this approach is that it can adapt both small and large object detection models to domain shifts since it produces high quality pseudo-labels even when the student network is small. Another advantage of our method is that the teacher and student do not need to share the same architecture. Thus, it is possible to use a slow but accurate teacher for the purpose of adaptation while choosing a fast architecture for deployment.

3.2.2 DomainMix augmentation

Here, we present a new augmentation method named DomainMix. As illustrated in Figure 1, it uniformly samples images from both the source and target domains $\mathcal{D} \cup \overline{\mathcal{D}}$ and strongly mixes these images into a new image along with their (pseudo-)labels. Figure 2 shows an example of DomainMix images from natural and artistic domains.

DomainMix uses simple ideas with many benefits to mitigate domain shift and label noise:

- It produces a diverse set of images by randomly sampling and mixing crops from source and target sets with replacement. As a result, it uses a different sample of images at every epoch, thus increasing the effective number of training samples and preventing overfitting. In contrast, simple batching reuses same images at every epoch.

- It is data-efficient as it uses a weighted balanced sampling from both domains. This helps learning representations that are robust to data shifts even if the target dataset has limited samples or the source and target datasets are highly imbalanced. In [2], we provide ablation studies that demonstrate the data efficiency of DomainMix.

- It mixes ground-truth and pseudo-labels in the same image. This mitigates the effect of false labels during adaptation because the image always contains accurate labels from the source domain

- It enforces the model to detect small objects as the objects in original samples are scaled down.

Algorithm 1 DomainMix augmentation

Inputs: A batch β of B images, labels $\{\mathbf{y}_i\}$ from source data $\mathcal{D}$, unlabeled target data $\overline{\mathcal{D}}$, pseudo-labels $\{\overline{\mathbf{y}}_j\}$

Output: A batch of domain-mixed samples $\widehat{\beta}$

```
procedure DomainMix($\beta, \overline{\mathcal{D}}, \{\overline{\mathbf{y}}_j\}$)
    $\widehat{\beta} \leftarrow \emptyset$
    for $i \leftarrow 1, B$ do
        $\mathcal{S} \leftarrow \{(\mathbf{x}_i, \mathbf{y}_i)\}$
        for $j \leftarrow$ sample($\mathcal{D} \cup \overline{\mathcal{D}}, 3$) do
            if $j \in \overline{\mathcal{D}}$ then
                $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\overline{\mathbf{x}}_j, \overline{\mathbf{y}}_j)\}$
            else
                $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\mathbf{x}_j, \mathbf{y}_j)\}$
        Collate crops from 4 images in $\mathcal{S}$ into $\widehat{\mathbf{x}}_i$
        Recompute all box coordinates in $\mathcal{S}$ into $\widehat{\mathbf{y}}_i$
        $\widehat{\beta} \leftarrow \widehat{\beta} \cup \{(\widehat{\mathbf{x}}_i, \widehat{\mathbf{y}}_i)\}$
```

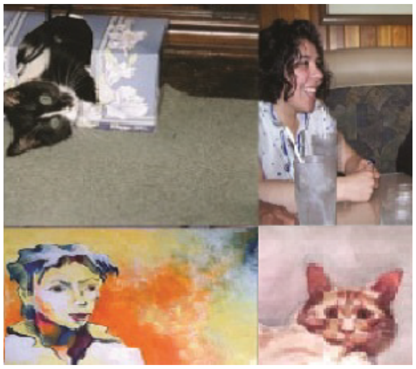


Figure 2. An example image generated by DomainMix mixing real images from Pascal VOC and artistic images from Watercolor2K.

The steps of DomainMix augmentation are listed in Algorithm 1. For each image in a batch, we randomly sample three additional images from source and target data $\mathcal{D} \cup \overline{\mathcal{D}}$ and mix random crops of these images to create a new domain-mixed image in a 2 × 2 collage. In addition, we collate the pseudo-labels $\overline{\mathbf{y}}_j$ for the unlabeled examples $\overline{\mathbf{x}}_j$ in $\overline{\mathcal{D}}$ with the ground-truth labels of source images. The bounding box coordinates of the objects are computed based on the relative position of each crop in the new mixed image. Furthermore, we employ a weighted balanced sampler to sample uniformly from the two domains.

3.2.3 Gradual self-labeling adaptation

Next, we present a gradual adaptation for optimizing the parameters of the detection model. This algorithm mitigates the effects of label noise, which is exacerbated by the domain shift. In fact, the pseudo-labels generated by the source models can be noisy on target domain images (e.g. it cannot detect objects or detects them inaccurately). If these initial pseudo-labels are used to adapt all the layers of the model at the same time, it results in poor supervision and hinders the model adaptation.

Instead, we propose a phased approach. First, we freeze all convolutional layers and adapts only the BN layers in the first w epochs. After this first phase, BN layers' trainable coefficients are updated. The partially adapted model is then used to generate more accurate pseudo-labels, which

Algorithm 2 Gradual self-labeling adaptation

```
Inputs: Source model $\theta^s_{\mathtt{M}}$, labeled source data $\mathcal{D}$, unlabeled
    target data $\overline{\mathcal{D}}$, warmup epochs $w$, total epochs T, steps
    per epoch $N$, and batch size $B$
Output: Adapted model $\theta^a_{\mathtt{M}}$
procedure ADAPT($\theta^s_{\mathtt{M}}, \mathcal{D}, \overline{\mathcal{D}}$)
    for $\overline{\mathbf{x}}_j \leftarrow \overline{\mathcal{D}}$ do $\overline{\mathbf{y}}_j \leftarrow$ GenPseudo($\overline{\mathbf{x}}_j, \theta^s_{\mathtt{M}}$)
    Initialize $\theta \leftarrow \theta^s_{\mathtt{M}}$
    for $layer \leftarrow \theta$.layers do
        if layer is not BatchNorm then Freeze layer
    for $epoch \leftarrow 1, \dots, T$ do
        if epoch == w then                    ▷ switch to Phase 2
            for $\overline{x} \leftarrow \overline{\mathcal{D}}$ do $\overline{\mathbf{y}}_j \leftarrow$ GenPseudo($\overline{\mathbf{x}}_j, \theta$)
            Unfreeze all layers
        for step $\leftarrow 1, \dots, N$ do
            Sample a batch $\beta = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{B}$ from $\mathcal{D}$
            $\widehat{\beta} \leftarrow$ DomainMix($\beta, \overline{\mathcal{D}}, \{\overline{\mathbf{y}}_j\}$) as in Algo 1.
            Update $\theta$ to minimize the loss with $\widehat{\beta}$
    $\theta^a_{\mathtt{M}} \leftarrow \theta$
```

is done offline for simplicity. In the second phase, all layers are unfrozen and then fine-tuned using the refined pseudo-labels. Note that during these two phases, we use the mixed image samples generated by the DomainMix augmentation. The detailed steps of this gradual adaptation are listed in Algorithm 2.

In contrast to prior works [24, 28], which used BN Adaption on its own, we integrate it within a self-training framework to effectively overcome the inevitable label noise caused by the domain shift [18]. As will be shown in Section 4, when used with the DomainMix augmentation, the resulting method is effective in adapting object detection models to different kinds of domain shifts.

Note that [18] also used a two-phase progressive adaptation method but they used synthetic domain-mapped images, which are generated by a conditional GAN, to fine-tune the model in the first phase. In contrast, our method leverages actual target domain images, which are mixed with source domain images using DomainMix augmentation, during the entire adaptation process.

4. Experiments results

In this section, we evaluate the effectiveness of SimROD to combat different kinds of domain shifts, compare the performance with prior works on standard benchmarks, and conduct ablation studies. For our experiments, we adopted the single-stage detection architecture Yolov5 [20] and used different model sizes by scaling the input size, width and depth. We study synthetic-to-real and camera-setup shifts [6] in Section 4.1, cross-domain artistic shifts [18] in Section 4.2, and robustness against image corruptions [25] in Section 4.3. Training details and additional results are provided in the supplementary materials [2].

4.1. Synthetic-to-real and cross-camera benchmark

Datasets. We used Sim10k [21] to Cityscapes [7] and KITTI [9] to Cityscapes benchmarks to study the ability to adapt in synthetic-to-real and cross-camera shifts, respectively. Following prior works, only the *car* class was used.

Metrics. For a fair comparison, we grouped different model/method pairs whose "Source" models (trained only on the labeled source data) have a similar average precision $\text{AP}^{50}(\theta^s)$ on the target test set (i.e. Cityscapes val). We compared each group based on three metrics: (1) $\text{AP}^{50}(\theta^a)$ of their "Adapted" models, (2) absolute adaptation gains τ, and (3) their effective adaptation gains ρ defined as:

$$\tau = \text{AP}^{50}(\theta^a) - \text{AP}^{50}(\theta^s), \tag{1}$$

$$\rho = 100 \times \frac{\text{AP}^{50}(\theta^a) - \text{AP}^{50}(\theta^s)}{\text{AP}^{50}(\text{Oracle}) - \text{AP}^{50}(\theta^s)}, \tag{2}$$

where "Oracle" is the model that is trained with the labeled target domain data. The gain metric τ was proposed by [37] to compare methods that may share same base architecture but have different performance before adaptation. For a better comparison, we also analyze the effectiveness of the adaptation method using the metric ρ. This metric helps understand if an adaptation method offers higher performance on the target test set beyond what is expected from having high performance on the source test set. A method that fails to adapt a model will have an effective gain of $\rho = 0\%$ for that model whereas a method that gives a target performance close to the Oracle will have $\rho = 100\%$.

Sim10K to Cityscapes. Table 1 shows that SimROD achieved new SOTA results on both the target AP50 performance and on the effective adaptation gain. We use two student models S320 and S416, which have the same Yolov5s architecture but different input sizes of 320 and 416 pixels to compare with prior methods that have comparable Source AP50 performance. For example, our S320 models achieves $AP50 = 44.70\%$ and $\rho = 72.93\%$ compared to $AP50 = 43.8\%$ and $\rho = 35.34\%$ for Coarse-to-Fine [37]. Similar results were observed when comparing the performance of our adapted S416 model with that of the FCOS model adapted with EPM [16]. Fig. 3 demonstrates the effectiveness of SimROD to adapt models from Sim10K to Cityscapes compared to prior baselines. Models adapted with SimROD enjoyed up to 70-75% of the target AP performance (that is obtained if the model was trained with a fully labeled target dataset). In contrast, the baseline methods achieved only about 30% of their Oracle performance.

KITTI to Cityscapes benchmark. Table 2 shows the results of this experiment, where SimROD outperformed the baselines. With the S416 model, it achieves slightly higher AP50 performance than the best baseline PDA [17].

Method	Arch.	Backbone	Source	AP50	Oracle	τ	ρ	Reference
DAF [6]	F-RCNN	V	30.10	39.00	-	8.90	-	CVPR 2018
MAF [11]	F-RCNN	V	30.10	41.10	-	11.00	-	ICCV 2019
RLDA [22]	F-RCNN	I	**31.08**	**42.56**	68.10	**11.48**	**31.01**	ICCV 2019
SCDA [38]	F-RCNN	V	34.00	43.00	-	9.00	-	CVPR 2019
MDA [36]	F-RCNN	V	34.30	42.80	-	8.50	-	ICCV 2019
SWDA [27]	F-RCNN	V	34.60	42.30	-	7.70	-	CVPR 2019
Coarse-to-Fine [37]	F-RCNN	V	**35.00**	43.80	59.90	8.80	35.34	CVPR 2020
SimROD (self-adapt)	YOLOv5	S320	33.62	38.73	48.81	5.11	33.66	Ours
SimROD (w. teacher X640)	YOLOv5	S320	33.62	**44.70**	48.81	**11.08**	**72.93**	Ours
MTOR [4]	F-RCNN	R	39.40	46.60	-	7.20	-	CVPR 2019
EveryPixelMatters [16]	FCOS	V	**39.80**	49.00	69.70	9.20	30.77	ECCV 2020
SimROD (self adapt)	YOLOv5	S416	39.57	44.21	56.49	4.63	27.37	Ours
SimROD (w. teacher X1280)	YOLOv5	S416	39.57	**52.05**	56.49	**12.47**	**73.73**	Ours

Table 1. Results of different method/model pairs for the Sim10K-to-Cityscapes adaptation scenario. “V”, “I” and “R” represent the VGG16, ResNet50, Inception-v2 backbones respectively. ”S320”, “M416”, “X640”, “X1280” represent different scales of Yolov5 model with increasing depth, width and input size. “Source” refers to the model trained only using source images without domain adaptation. For a fair comparison, we group together method/model pairs whose “Source” performance are similar. We report the AP50 (%) performance of the adapted model and the “Oracle” model which is trained with labeled target data, as well each method’s absolute and effective gains (%) when available. τ and ρ are the absolute gain and the effective gain respectively as defined in (1) and (2).

Method	Arch.	Backbone	Source	AP50	Oracle	τ	ρ	Reference
DAF [6]	F-RCNN	V	30.20	38.50	-	8.30	-	CVPR 2018
MAF [11]	F-RCNN	V	30.20	41.00	-	10.80	-	ICCV 2019
RLDA [22]	F-RCNN	I	31.10	42.98	68.10	11.88	32.11	ICCV 2019
PDA [17]	F-RCNN	V	30.20	43.90	55.80	13.70	53.52	WACV 2020
SimROD (self-adapt)	YOLOv5	S416	31.61	35.94	56.15	4.33	17.65	Ours
SimROD (w. teacher X1280)	YOLOv5	S416	**31.61**	**45.66**	56.15	**14.05**	**57.27**	Ours
SCDA [38]	F-RCNN	V	**37.40**	42.60	-	5.20	-	CVPR 2019
EveryPixelMatters [16]	FCOS	R	35.30	45.00	70.40	9.70	27.64	ECCV 2020
SimROD (self adapt)	YOLOv5	M416	36.09	42.94	59.29	6.85	29.51	Ours
SimROD (w. teacher X1280)	YOLOv5	M416	36.09	**47.52**	59.29	**11.43**	**49.26**	Ours

Table 2. Results of different method/model pairs on the KITTI-to-Cityscapes adaptation scenario. τ and ρ are the absolute gain and the effective gain respectively as defined in (1) and (2).

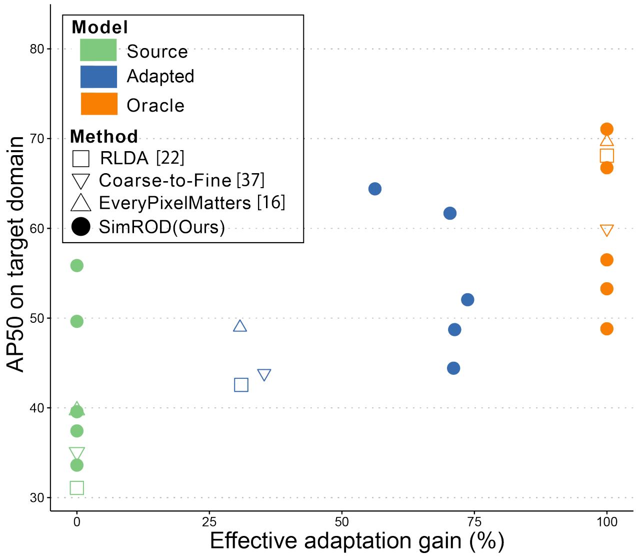


Figure 3. AP50 on test vs effective gain for Sim10K to Cityscapes. We use five different backbones S320, M320, S416, S640 and M640 for the student and the same backbone X1280 for teacher.

When using the medium size M416 model, SimROD also outperformed prior baselines with comparable Source AP50 performance namely SCDA [38] and EPM [16].

4.2. Cross-domain artistic benchmark

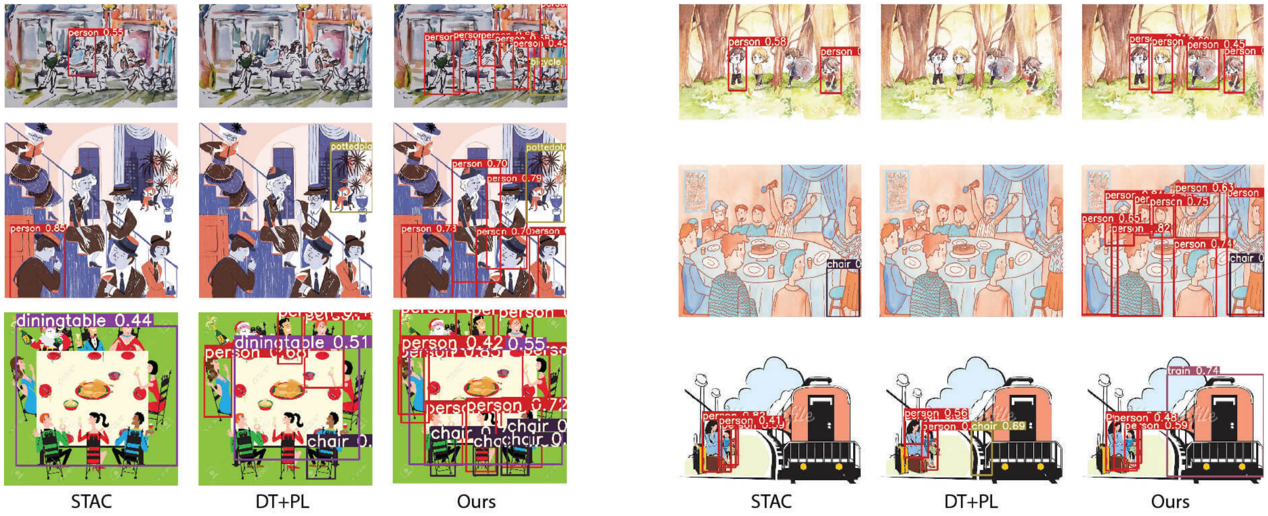
Datasets and metrics. The cross-domain artistic benchmark consists of three domain shifts where the source data is VOC07 trainval and the target domains are Clipart1k, Watercolor2k and Comic2k datasets [18]. We use the same benchmark metrics as in Sec. 4.1.

Results. Our method outperformed the baselines by significant margins. Compared to DT+PL [18], our method further improved the AP50 of the yolov5s model by absolute gains of +8.45, +12 and +10.69 % points on Clipart, Comic, and Watercolor respectively. While DT+PL outperformed the augmentation-based baselines on Clipart, it did slightly worse than STAC on Comic and Watercolor. Finally, SimROD was effective in adapting models of different sizes. Without generating synthetic data or using domain adversarial training, SimROD’s effective gain ρ was consistently above 70% and could reach up to 97% when a large adapted teacher was used to refine the pseudo-labels.

In Table 3, we give a detailed benchmark for the VOC to Watercolor benchmark, from which we used 1000 unlabeled images as target data. In [2], we present detailed results on Clipart and Comic dataset as well as more ablation results when using extra unlabeled data for adaptation.

Method	Arch.	Backbone	Source	AP50	Oracle	τ	ρ	Reference
DAF [6]	F-RCNN	V	39.80	34.30	NA	-5.50	NA	CVPR 2018
DAM [23]	F-RCNN	V	**39.80**	52.00	NA	12.20	NA	CVPR 2019
DeepAugment [12]	YOLOv5	S416	37.46	45.19	56.07	7.73	41.54	arXiv 2020
BN-Adapt [19]	YOLOv5	S416	37.46	45.72	56.07	8.26	44.39	NeurIPS 2020
Stylize [10]	YOLOv5	S416	37.46	46.26	56.07	8.80	47.29	arXiv 2019
STAC [30]	YOLOv5	S416	37.46	49.83	56.07	12.37	66.47	arXiv 2020
DT+PL [18]	YOLOv5	S416	37.46	44.86	56.07	7.40	39.77	CVPR 2018
SimROD (self-adapt)	YOLOv5	S416	37.46	52.58	56.07	15.12	81.26	Ours
SimROD (teacher X416)	YOLOv5	S416	37.46	**55.55**	56.07	**18.09**	**97.21**	Ours
ADDA [34]	SSD	V	49.60	49.80	58.40	0.20	2.27	CVPR 2017
DT+PL [18]	SSD	V	**49.60**	54.30	58.40	4.70	53.41	CVPR 2018
SWDA [27]	F-RCNN	V	44.60	56.70	58.60	12.10	**86.43**	CVPR 2019
DeepAugment [12]	YOLOv5	M416	46.95	54.02	66.34	7.07	36.47	arXiv 2020
BN-Adapt [19]	YOLOv5	M416	46.95	55.75	66.34	8.80	45.39	NeurIPS 2020
Stylize [10]	YOLOv5	M416	46.95	55.24	66.34	8.29	42.76	arXiv 2019
STAC [30]	YOLOv5	M416	46.95	57.82	66.34	10.87	56.07	arXiv 2020
DT+PL [18]	YOLOv5	M416	46.95	49.14	66.34	2.19	11.30	CVPR 2018
SimROD (self-adapt)	YOLOv5	M416	46.95	60.08	66.34	13.13	67.72	Ours
SimROD (teacher X416)	YOLOv5	M416	46.95	**63.47**	66.34	**16.52**	85.22	Ours

Table 3. Benchmark results on Real (VOC) to Watercolor2K domain shift.



(a) Pseudo-labels on unlabeled target samples

(b) Predictions on test target samples

Figure 4. Qualitative comparison: (a) pseudo-labels generated on unlabeled target examples and (b) test predictions with adapted Yolov5s.

4.3. Image corruptions benchmark

Datasets. We evaluate our method's robustness to image corruption using the standard benchmarks Pascal-C, COCO-C, and Cityscapes-C [25]. For Pascal-C, we used VOC07 trainval split as the source training data. For COCO-C and Cityscapes-C, we divided the train split and used the first half as source training data. There are $N_c = 15$ different corruption types for each dataset. Thus, we applied each corruption type on the VOC12 trainval or on the second half of COCO-C and Cityscapes-C train as unlabeled target data. Precisely, we applied each corruption type with middle severity onto each image using the *imagecorruptions* library [25]. More details are given in [2].

Metrics. For image corruption benchmark, we followed the evaluation protocol from [13, 25, 32] and measured the mean performance under corruption (mPC), relative performance under corruption (rPC), and the relative robustness τ_c of the adapted model averaged over N_c corruption types:

$$\text{mPC}^x = \frac{1}{N_c}\sum_{c=1}^{N_c}\frac{1}{N_s}\sum_{s=1}^{5}\text{AP}^x_{c,s}. \tag{3}$$

$$\text{rPC}^x = \frac{\text{mPC}^x}{\text{AP}^x_{\text{clean}}}. \tag{4}$$

$$\tau_c = \text{mPC}(\theta^a) - \text{mPC}(\theta^s). \tag{5}$$

where $\text{AP}^x_{\text{clean}}$ and $\text{AP}^x_{c,s}$ denote the average precision of the test data with corruption type c and severity level s. The relative robustness τ_c quantifies the effect of adaptation on the performance under distribution shift (mPC).

Baselines. We use the following baselines which were proposed to improve the robustness to image corruptions: Stylize [10], BN-Adapt [19], DeepAugment [12], STAC [30], and DT+PL [18]. Unless specified, we employed weak data augmentations such as RandomHorizon-

Method	AP$^{50}_{\text{clean}}$	mPC^{50}	rPC	τ_c	ρ
Source	83.13	53.78	64.69	0.00	0
Stylize	84.79	62.92	74.21	9.14	36.62
BN-Adapt	83.01	64.60	77.82	10.82	43.35
DeepAugment	85.05	64.88	76.28	11.10	44.47
STAC	**87.00**	66.88	76.87	13.10	52.48
SimROD (ours)	86.97	**75.40**	**86.70**	**21.62**	**86.62**
Oracle	86.75	78.74	90.77	24.96	100

Table 4. Performance comparison on Pascal-C benchmark.

Method	AP$^{50}_{\text{clean}}$	mPC^{50}	rPC	τ_c	ρ
Source	**36.85**	22.03	59.78	0.00	0
Stylize	35.75	23.82	66.63	1.79	22.02
BN-Adapt	36.24	24.79	68.41	2.76	33.95
DeepAugment	35.51	24.33	68.52	2.30	28.29
STAC	36.76	24.80	67.46	2.77	34.07
SimROD (ours)	36.79	**28.46**	**77.36**	**6.43**	**79.09**
Oracle	36.23	30.16	83.25	8.13	100

Table 5. Performance benchmark on COCO-C dataset.

Method	AP$^{50}_{\text{clean}}$	mPC^{50}	rPC	τ_c	ρ
Source	19.48	11.53	59.19	0.00	0
Stylize	21.77	14.62	67.16	3.09	25.81
DeepAugment	20.28	14.79	72.93	3.26	27.23
STAC	**24.54**	15.39	62.71	3.86	32.25
SimROD (ours)	24.06	**18.01**	**74.85**	**6.48**	**54.14**
Oracle	26.58	23.50	88.41	11.97	100

Table 6. Performance benchmark on Cityscapes-C dataset.

talFlip and RandomCrop for all baselines.

Main results. Table 4, 5 and 6 show the results of Yolov5m model for Pascal-C, COCO-C, and Cityscapes-C, respectively. We report the results with different model sizes in [2]. We used the large model Yolov5x model as a teacher. An ablation study on Pascal-C is provided in Table 7 and will be discussed later.

Unlabeled target samples improved robustness to image corruption. The source models suffered from performance drop due to image corruptions. By adapting the models with SimROD, the mean performance under corruption mPC^{50} was significantly improved by +21.62, +6.43, and +6.48 absolute percentage points on Pascal-C, COCO-C, and Cityscapes-C, respectively. Our method outperformed the Stylize, DeepAugment, BNAdapt baselines on all metrics. In fact, STAC, which also used unlabeled target samples, achieved the second best performance. This shows that augmentation or batch norm adaptation is not sufficient to fix the domain shift on all possible corruptions. Instead, using unlabeled samples from target domain is more effective to combat image corruptions.

Pseudo-label refinement ensured performance close to Oracle. Moreover, Tables 4, 5 and 6 show that the performance of our unsupervised method was close to that of the Oracle, which uses ground-truth labels for *target* domain data. This was possible because the adapted teacher produces highly accurate pseudo-labels, which could be used along with DomainMix augmentation to effectively adapt the student model.

Ablation Study. Next, we present an ablation study using the Yolov5m model on Pascal-C in Table 7 to gain some insights about the contributions of the three parts of our method. First, BN-Adapt improved the mean performance under corruption by 10.82% AP50. Applying DomainMix augmentation on top of BN-Adapt improved the performance by 2.18%. Next, the teacher-guided (TG) pseudo-label refinement was particularly useful in adapting small models. When using our full method, the performance increased by 10.8% compared to BN-Adapt. Compared to self adaptation, TG improved the Yolov5 model's performance mPC by +3.7 %. Finally, the gradual adaptation (GA) also played an important role in refining pseudo-labels and in improving the model's robustness. For example, if we did not use GA and skipped the BN adaptation in the first phase, the performance dropped by 1.95% compared to the full method. Our method organically integrates these parts to tackle UDA for object detection. While the parts may appear simple, their synergy helped mitigate the challenging issues of domain shift and pseudo-label noise.

Method	TG	DMX	GA	FT	mPC^{50}	τ_c
Source					53.78	0.0
BN-Adapt			✓		64.60	10.8
BN-A + DMX		✓	✓		66.78	13.0
SimROD w/o TG		✓	✓	✓	71.81	18.0
SimROD w/o GA	✓	✓		✓	73.45	19.7
SimROD	✓	✓	✓	✓	75.40	21.7

Table 7. Ablation study on Pascal-C with yolov5m. See [2] for ablations with other models. TG, GA, DMX, and FT denote Teacher Guidance, Gradual Adaption, DomainMix, and Fine-Tuning.

Qualitative analysis Finally, we illustrate the effectiveness of our method by showing the pseudo-labels generated with our method on the unlabeled target training images on Comic dataset. As seen in Figure 4(a), our method generated highly accurate pseudo-labels despite the domain shift. In contrast, STAC and DT+PL generated sparse labels since they missed to detect many objects. The performance difference transferred to the quality of predictions on the test set as shown in Figure 4(b).

5. Conclusion

We proposed a simple and effective unsupervised method for adapting detection models under domain shift. Our simple method gradually adapts the model with the help of a new domain-centric data augmentation and a teacher-guided pseudo-label refinement procedure. Our method achieved significant gains in terms of model robustness compared to baselines both for small and large models. Our method could mitigate different kinds of domain shifts from low-level image corruptions to high-level cross-domain or stylistic differences. Through ablation study, we got some insights on why gradual adaptation works and how the teacher-guided pseudo-label refinement can help adapt the models. We hope that this simple method will serve as a strong baseline and will guide future research progress.

References

[1] Anonymous. Is robustness robust? on the interaction between augmentations and corruptions. In *Submitted to International Conference on Learning Representations*, 2021. under review. 1, 2

[2] Authors. SimROD: A Simple Adaptation Method for Robust Object Detection, 2021. Supplementary materials 8083_supplementary.zip. 4, 5, 6, 7, 8

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 1, 2

[4] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 2, 6

[5] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019. 2

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1, 2, 5, 6, 7

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, M. Enzweiler, Rodrigo Benenson, Uwe Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 5

[8] Sebastian Cygert and Andrzej Czyżewski. Toward robust pedestrian detection with data augmentation. *IEEE Access*, 8:136674–136683, 2020. 2

[9] Andreas Geiger, Philip Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 5

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 2, 7

[11] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019. 1, 2, 6

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020. 1, 2, 7

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2, 7

[14] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 1, 2

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, pages 1994–2003, 2018. 2

[16] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 1, 2, 5, 6

[17] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 749–757, 2020. 1, 2, 5, 6

[18] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 1, 2, 5, 6, 7

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456, 2015. 2, 7

[20] Glenn Jocher et al. ultralytics/yolov5: v1.0 - initial release. Zenodo, June 2020. 5

[21] M. Johnson-Roberson, Charles Barto, R. Mehta, S. N. Sridhar, Karl Rosaen, and R. Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 746–753, 2017. 5

[22] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019. 1, 2, 6

[23] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 1, 2, 7

[24] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017. 2, 5

[25] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 1, 2, 5, 7

[26] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *arXiv preprint arXiv:2102.11273*, 2021. 1

[27] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 1, 2, 6, 7

[28] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551, 2020. 2, 5

[29] Garrett Smith et al. Guildai. *Github. Note: https://github.com/guildai/guildai*, 2017.

[30] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 2, 7

[31] Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments - Introduction to Covariate Shift Adaptation*. MIT Press, 2012. 3

[32] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification, 2020. 1, 2, 7

[33] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. 42(2):245–284, Feb. 2015. 1

[34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017. 7

[35] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020. 2

[36] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[37] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13766–13775, 2020. 1, 2, 5, 6

[38] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019. 1, 2, 6

[39] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey, 2019. 1