

# Adapting Object Detectors with Conditional Domain Normalization

Peng Su<sup>1,2</sup>, Kun Wang<sup>2</sup>, Xingyu Zeng<sup>2</sup>, Shixiang Tang<sup>2</sup>  
Dapeng Chen<sup>2</sup>, Di Qiu<sup>2</sup>, and Xiaogang Wang<sup>1</sup>

<sup>1</sup> The Chinese University of Hong Kong

<sup>2</sup> SenseTime Research

{psu,xgwang}@ee.cuhk.edu.hk

**Abstract.** Real-world object detectors are often challenged by the domain gaps between different datasets. In this work, we present the Conditional Domain Normalization (CDN) to bridge the domain gap. CDN is designed to encode different domain inputs into a shared latent space, where the features from different domains carry the same domain attribute. To achieve this, we first disentangle the domain-specific attribute out of the semantic features from one domain via a domain embedding module, which learns a domain-vector to characterize the corresponding domain attribute information. Then this domain-vector is used to encode the features from another domain through a conditional normalization, resulting in different domains’ features carrying the same domain attribute. We incorporate CDN into various convolution stages of an object detector to adaptively address the domain shifts of different level’s representation. In contrast to existing adaptation works that conduct domain confusion learning on semantic features to remove domain-specific factors, CDN aligns different domain distributions by modulating the semantic features of one domain conditioned on the learned domain-vector of another domain. Extensive experiments show that CDN outperforms existing methods remarkably on both real-to-real and synthetic-to-real adaptation benchmarks, including 2D image detection and 3D point cloud detection.

## 1 Introduction

Deep neural networks have achieved remarkable success on visual recognition tasks [11, 13]. However, it is still very challenging for deep networks to generalize on a different domain, whose data distribution is not identical with original training data. Such a problem is known as dataset bias or domain shift [32]. For example, to guarantee safety in autonomous driving, the perception model is required to perform well under all conditions, like sunny, night, rainy, snowy, etc. However, even top-grade object detectors still face significant challenges when deployed in such varying real-world settings. Although collecting and annotating more data from unseen domains can help, it is prohibitively expensive, laborious and time-consuming. Another appealing application is to adapt from synthetic

data to real data, as it can save the amount of cost and time. However, current object detectors trained with synthetic data can rarely generalize on real data due to a significant domain distribution gap [28, 37, 39].

Adversarial domain adaptation emerges as a hopeful method to learn transferable representations across domains. It has achieved noticeable progress in various machine learning tasks, from image classification [23, 24, 27], semantic segmentation [37, 40, 48], object detection [34, 47] to reinforcement learning [20, 29, 39]. According to Ben-David’s theory [1], the empirical risk on the target domain is bounded by the source domain risk and the  $\mathcal{H}$  domain divergence. Adversarial adaptation dedicates to learn domain invariant representation to reduce the  $\mathcal{H}$  divergence, which eventually decreases the upper bound of the empirical error on the target domain.

However, existing adversarial adaptation methods still suffer from several problems. First, previous methods [4, 8, 39] directly feed semantic features into a domain discriminator to conduct domain confusion learning. But the semantic features contain both image contents and domain attribute information. It’s difficult to make the discriminator only focusing on removing domain-specific information without inducing undesirable influence on the images contents. Second, existing adversarial adaptation methods [4, 8, 39] use domain confusion learning at one or few convolution stages to handle the distribution mismatch, which ignores the differences of domain shifts at various representation levels. For example, the first few convolution layers’ features mainly convey low-level information of local patterns, while the higher convolution layers’ features include more abstract global patterns with semantics [44]. Such differences born within deep convolution neural networks naturally exhibit different types of domain shift at various convolution stages.

Motivated by this, we propose the Conditional Domain Normalization (CDN) to embed different domain inputs into a shared latent space, where the features of all different domains inputs carry the same domain attribute information. Specifically, CDN utilizes a domain embedding module to learn a domain-vector to characterize the domain attribute information, through disentangling the domain attribute out of the semantic features of domain inputs. We use this domain-vector to encode the semantic features of another domain’s inputs via a conditional normalization. Thus different domain features carry the same domain attributes information. We adopt CDN in various convolution stages to address different types of domain shift adaptively. The experiment on both real-to-real and synthetic-to-real adaptation benchmarks demonstrate that our method outperforms the-state-of-the-art adaptation methods. To summarize, our contributions are three folds: (1) We propose the Conditional Domain Normalization to bridge the domain distribution gap, through embedding different domain inputs into a shared latent space, where the features from different domains carry the same domain attribute. (2) CDN achieves state-of-the-art unsupervised domain adaptation performance on both real-to-real and synthetic-to-real benchmarks, including 2D image and 3D point cloud detection tasks. And we conduct both quantitative and qualitative comparisons to analyze the features learned

by CDN. (3) We construct a large-scale synthetic-to-real driving benchmark for 2D object detection, including a variety of public datasets.

## 2 Related work

**Object Detection** is the center topic in computer vision, which is crucial for many real-world applications, such as autonomous driving. In 2D detection, following the pioneering work of RCNN [11], a number of object detection frameworks based on convolutional networks have been developed like Fast R-CNN [10], Faster R-CNN [33], and Mask R-CNN [12], which significantly push forward the state of the art. In 3D detection, spanning from detecting 3d objects from 2d images [3], to directly generate 3D box from point cloud [30, 38], abundant works has been successfully explored. All these 2D and 3D objectors have achieved remarkable success on one or few specific public datasets. However, even top-grade object detectors still face significant challenges when deployed in real-world settings. The difficulties usually arise from the changes in environmental conditions.

**Domain Adaptation** generalizes a model across different domains, and it has been extensively explored in various tasks, spanning from image classification [2, 41, 24, 27, ?], semantic segmentation [15, 40, 37] to reinforcement learning [39, 29, 20]. For 2D detection, domain confusion learning via a domain discriminator has achieved noticeable progress in cross-domain detection. [4] incorporated a gradient reversal layer [8] into a Faster R-CNN model. [34, 47] adopt domain confusion learning on both global and local levels to align source and target distributions. In contrast to existing methods conducting domain confusion learning directly on semantic features, we explicitly disentangle the domain attribute out of semantic features. And this domain attribute is used to encode other domains’ features, thus different domain inputs share the same domain attribute in the feature space. For 3D detection, only a few works [46, 17] has been explored to adapt object detectors across different point cloud dataset. Different from existing works [46, 17] are specifically designed for point cloud data, our proposed CDN is a general adaptation framework that adapts both 2D image and 3D point cloud object detector through the conditional domain normalization.

**Conditional Normalization** is a technique to modulate the neural activation using a transformation that depends on external data. It has been successfully used in the generative models and style transfer, like conditional batch normalization [6], adaptive instance normalization (AdaIN) [16] and spatial adaptive batch normalization [25]. [16] proposes AdaIN to control the global style of the synthesized image. [42] modulates the features conditioned on semantic masks for image super-resolution. [25] adopts a spatially-varying transformation, making it suitable for image synthesis from semantic masks. Inspired by these works, we propose Conditional Domain Normalization (CDN) to modulate one domain’s inputs condition on another domain’s attributes information. But our method exhibits significant difference with style transfer works: Style transfer works mod-

ify a content image conditioned on another style image, which is a conditional instance normalization by nature; but CDN modulates one domain’s features conditioned on the domain embedding learned from another domains’ inputs (a group of images), which is like a domain-to-domain translation. Hence we use different types of conditional normalization to achieve different goals.

### 3 Method

We first introduce the general unsupervised domain adaptation approach in section 3.1. Then we present the proposed Conditional Domain Normalization (CDN) in section 3.2. Last we adapt object detectors with the CDN in section 3.3.

#### 3.1 General Adversarial Adaptation Framework

Given source images and labels  $\{(x_i^S, y_i^S)\}_{i=1}^{N_S}$  drawn from  $P_s$ , and target images  $\{x_i^T\}_{i=1}^{N_T}$  from target domain  $P_t$ , the goal of unsupervised domain adaptation is to find a function  $f : x \rightarrow y$  that minimize the empirical error on target data. For object detection task, the  $f$  can be decomposed as  $f = G(\cdot; \theta_g) \circ H(\cdot; \theta_h)$ , where  $G(\cdot; \theta_g)$  represents a feature extractor network and  $H(\cdot; \theta_h)$  denotes a bounding box head network. The adversarial domain adaptation introduces a discriminator network  $D(\cdot; \theta_d)$  that tries to determine the domain labels of feature maps generated by  $G(\cdot; \theta_g)$ .

$$\begin{aligned} \min_{\theta_g, \theta_h} \quad & \mathcal{L}_{det} = \mathcal{L}_{cls}(G(x; \theta_g), H(x; \theta_h)) + \mathcal{L}_{reg}(G(x; \theta_g), H(x; \theta_h)) \\ \min_{\theta_d} \max_{\theta_g} \quad & \mathcal{L}_{adv} = \mathbb{E}_{x \sim P_s} [\log(D(G(x; \theta_g); \theta_d))] + \mathbb{E}_{x \sim P_t} [\log(1 - D(G(x; \theta_g); \theta_d))] \end{aligned} \quad (1)$$

As illustrated in Eq.1,  $G(\cdot; \theta_g)$  and  $H(\cdot; \theta_h)$  are jointly trained to minimize the detection loss  $\mathcal{L}_{det}$  by supervised training on the labeled source domain. At the same time, the backbone  $G(\cdot; \theta_g)$  is optimized to maximize the probability of  $D(\cdot; \theta_d)$  to make mistakes. Through this two-player min-max game, the final  $G(\cdot; \theta_g)$  will converge to extract features that are indistinguishable for  $D(\cdot; \theta_d)$ , thus domain invariant representations are learned.

#### 3.2 Conditional Domain Normalization

Conditional Domain Normalization is designed to embed source and target domain inputs into a shared latent space, where the semantic features from different domains carry the same domain attribute information. Formally, let  $v^s \in \mathbb{R}^{N \times C \times H \times W}$  and  $v^t \in \mathbb{R}^{N \times C \times H \times W}$  represent feature maps of source and target inputs, respectively.  $C$  is the channel dimension and  $N$  denotes the mini-batch size. We first learn a domain embedding vector  $e_{domain}^s \in \mathbb{R}^{1 \times C \times 1}$  to characterize the domain attribute of source inputs. It is accomplished by a domain embedding network  $\mathbf{F}_d(\cdot; W)$  parameterized by two fully-connect layers with ReLU non-linearity  $\delta$  as

$$e_{domain}^s = \mathbf{F}_d(v_{avg}^s; W) = \delta(W_2 \delta(W_1 v_{avg}^s)). \quad (2)$$

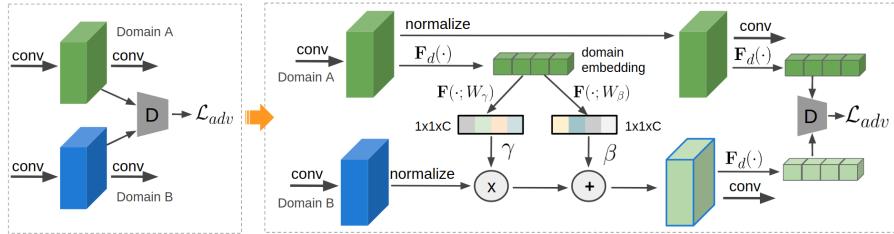


Fig. 1: (Left) Traditional domain adversarial approach. (Right) Conditional Domain Normalization (CDN). The green and blue cubes represent the feature maps of domain A and domain B respectively.

And  $v_{avg}^s \in \mathbb{R}^{N \times C \times 1}$  represents the channel-wise statistics of source feature  $v^s$  generated by global average pooling

$$v_{avg}^s = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W v^s(h, w). \quad (3)$$

To embed both source and target domain inputs into a shared latent space, where source and target features carry the same domain attributes while preserving individual image contents. We encode the target features  $v^t$  with the source domain embedding via an affine transformation as

$$\hat{v}^t = \mathbf{F}(e_{domain}^s; W_\gamma, b_\gamma) \cdot \left( \frac{v^t - \mu^t}{\sigma^t} \right) + \mathbf{F}(e_{domain}^s; W_\beta, b_\beta), \quad (4)$$

where  $\mu^t$  and  $\sigma^t$  denote the mean and variance of target feature  $v^t$ . The affine parameters are learned by function  $F(\cdot; W_\gamma, b_\gamma)$  and  $F(\cdot; W_\beta, b_\beta)$  conditioned on the source domain embedding vector  $e_{domain}^s$ ,

$$F(e_{domain}^s; W_\gamma, b_\gamma) = W_\gamma e_{domain}^s + b_\gamma, \quad F(e_{domain}^s; W_\beta, b_\beta) = W_\beta e_{domain}^s + b_\beta. \quad (5)$$

For the target feature mean  $\mu^t \in \mathbb{R}^{1 \times C \times 1}$  and variance  $\sigma^t \in \mathbb{R}^{1 \times C \times 1}$ , we calculate it with a standard batch normalization [19]

$$\mu_c^t = \frac{1}{NHW} \sum_{n=1}^N \sum_{h=1}^H \sum_{w=1}^W v_{nchw}^t, \quad \sigma_c^t = \sqrt{\frac{1}{NHW} \sum_{n=1}^N \sum_{h=1}^H \sum_{w=1}^W (v_{nchw}^t - \mu_c^t)^2 + \epsilon}, \quad (6)$$

where  $\mu_c^t$  and  $\sigma_c^t$  denotes  $c$ -th channel of  $\mu^t$  and  $\sigma^t$ . Finally, we have a discriminator to supervise the encoding process of domain attribute as

$$\min_{\theta_d} \max_{\theta_g} \mathcal{L}_{adv} = \mathbb{E}[\log(D(\mathbf{F}_d(v^s)); \theta_d)] + \mathbb{E}[\log(1 - D(\mathbf{F}_d(\hat{v}^t); \theta_d))], \quad (7)$$

where  $v^s$  and  $v^t$  are generated by  $G(\cdot; \theta_g)$ .

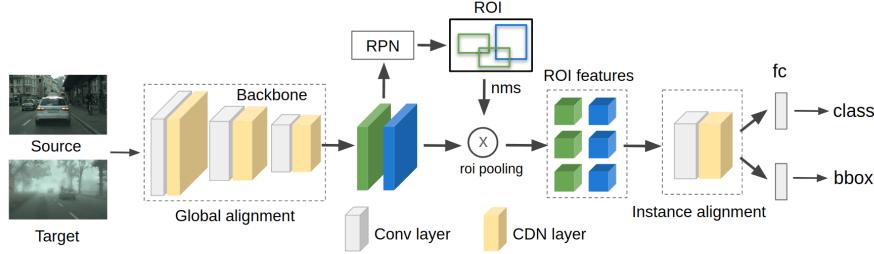


Fig. 2: Faster R-CNN network incorporates with CDN. The CDN is adopted in both backbone network and bounding box head network to adaptively address the domain shift at different representation levels.

**Discussion** CDN exhibits a significant difference compared with existing adversarial adaptation works. As shown in Fig. 1, previous methods conduct domain confusion learning directly on semantic features to remove domain-specific factors. However, the semantic features contain both domain attribute and image contents. It is not easy to enforce the domain discriminator only regularizing the domain-specific factors without inducing any undesirable influence on image contents. In contrast, we disentangle the domain attribute out of the semantic features via conditional domain normalization. And this domain attribute is used to encode other domains' features, thus different domain features carry the same domain attribute information.

### 3.3 Adapting Detector with Conditional Domain Normalization

Convolution neural network's (CNN) success in pattern recognition has been largely attributed to its great capability of learning hierarchical representations [44]. More specifically, the first few layers of CNN focus on low-level features of local pattern, while higher layers capture semantic representations. Given this observation, CNN based object detectors naturally exhibit different types of domain shift at various levels' representations. Hence we incorporate CDN into different convolution stages in object detectors to address the domain mismatch adaptively, as shown in Fig.2.

Coincident to our analysis, some recent works [34, 47] empirically demonstrate that global and local region alignments have different influences on detection performance. For easy comparison, we refer to the CDN located at the backbone network as global alignment, and CDN in the bounding box head networks as local or instance alignment.

As shown in Fig. 2, taking faster-RCNN model [33] with ResNet [13] backbone as an example, we incorporate CDN in the last residual block at each stage. Thus

the global alignment loss can be computed by

$$L_{adv}^{global} = \sum_{l=1}^L \mathbb{E}[\log(D_l(\mathbf{F}_d^l(v_l^s); \theta_d^l))] + \mathbb{E}[\log(1 - D_l(\mathbf{F}_d^l(\hat{v}_l^t); \theta_d^l))], \quad (8)$$

where  $v_l^s$  and  $v_l^t$  denote  $l$ -th layer's source feature and the encoded target feature, and  $D_l$  represents the corresponding domain discriminator parameterized by  $\theta_d^l$ .

As for bounding box head network, we adopt CDN on the fixed-size region of interest (ROI) features generated by ROI pooling [33]. Because the original ROIs are often noisy and the quantity of source and target ROIs are not equal, we randomly select  $\min(N_{roi}^S, N_{roi}^T)$  ROIs from each domain.  $N_{roi}^S$  and  $N_{roi}^T$  represent the quantity of source and target ROIs after non-maximum suppression (NMS). Hence we have instance alignment regularization for ROI features as

$$L_{adv}^{instance} = \mathbb{E}[\log(D_{roi}(\mathbf{F}_d^{roi}(v_{roi}^s); \theta_d^{roi}))] + \mathbb{E}[\log(1 - D_{roi}(\mathbf{F}_d^{roi}(\hat{v}_{roi}^t); \theta_d^{roi}))]. \quad (9)$$

The overall training objective is to minimize the detection loss  $\mathcal{L}_{det}$  (of the labeled source domain) that consists of a classification loss  $\mathcal{L}_{cls}$  and a regression loss  $\mathcal{L}_{reg}$ , and min-max a adversarial loss  $\mathcal{L}_{adv}$  of discriminator network

$$\begin{aligned} \min_{\theta_d} \max_{\theta_g} \mathcal{L}_{adv} &= \lambda L_{adv}^{global} + L_{adv}^{instance} \\ \min_{\theta_g, \theta_h} \mathcal{L}_{det} &= \mathcal{L}_{cls}(G(x; \theta_g), H(x; \theta_h)) + \mathcal{L}_{reg}(G(x; \theta_g), H(x; \theta_h)), \end{aligned} \quad (10)$$

where  $\lambda$  is a weight to balance the global and local alignment regularization.

## 4 Experiments

We evaluate CDN on various real-to-real (KITTI to Cityscapes) and synthetic-to-real (Virtual KITTI/Synscapes/SIM10K to BDD100K, PreSIL to KITTI) adaptation benchmarks. We also report results on cross-weather adaptation, Cityscapes to Foggy Cityscapes. Mean average precision (mAP) with an intersection-over-union (IOU) threshold of 0.5 is reported for 2D detection experiments. We use Source and Target to represent the results of supervised training on source and target domain, respectively. For 3D point cloud object detection, PointR-CNN [38] with backbone of PointNet++ [30] is adopted as our baseline model. Following standard metric on KITTI benchmark [38], we use Average Precision(AP) with IOU threshold 0.7 for car and 0.5 for pedestrian/cyclist.

### 4.1 Dataset

**Cityscapes** [5] is a European traffic scene dataset, which contains 2,975 images for training and 500 images for testing.

**Foggy Cityscapes** [35] derives from Cityscapes with a fog simulation proposed by [35]. It also includes 2,975 images for training, 500 images for testing.

**KITTI** [9] contains 21,260 images collected from different urban scenes, which includes 2D RGB images and 3D point cloud data.

**Virtual KITTI** [7] is derived from KITTI with a real-to-virtual cloning technique proposed by [7]. It has the same number of images and categories as KITTI.

**Synscapes** [43] is a synthetic dataset of street scene, which consists of 25,000 images created with a photo-realistic rendering technique.

**SIM10K** [21] is a street view dataset generated from the realistic computer game Grand Theft Auto V (GTA-V). It has 10,000 training images and the same categories as in Cityscapes.

**PreSIL** [17] is synthetic point cloud dataset derived from GTA-V, which consists of 50,000 frames of high-definition images and point clouds.

**BDD100K** [45] is a large-scale dataset (contains 100k images) that covers diverse driving scenes. It is a good representative of real data in the wild.

## 4.2 Implementation Details

We train the Faster R-CNN [33] model for 12 epochs on all experiments. The model is optimized by SGD with multi-step learning rate decay. SGD uses the learning rate of 0.00625 multiplied by the batchsize, and momentum of 0.9. We adopt CDN layer in all convolution stages, including the backbone and bounding box network. All experiments use sync BN [26] with a batchsize of 32.  $\lambda$  is set as 0.4 by default in all experiments. On Cityscapes to Foggy Cityscapes adaptation, we follow [34, 47] to prepare the train/test split, and use an image shorter side of 512 pixels. On synthetic-to-real adaptation, for a fair comparison, we randomly select 7000 images for training and 3000 for testing, for all synthetic datasets and BDD100K dataset. For 3D point cloud detection, we use PointRCNN [38] model with same setting as [38]. We incorporated the CDN layer in the point-wise feature generation stage (global alignment) and 3D ROIs proposal stage (instance alignment).

## 5 Experimental Results and Analysis

### 5.1 Results on Cityscapes to Foggy Cityscapes

We compare CDN with the state-of-the-art methods in Table 1. Following [34, 47], we also report results using Faster R-CNN model with VGG16 backbone. As shown in Table 1, CDN outperforms previous state-of-the-art methods by a large margin of 1.8% mAP. The results demonstrate the effectiveness of CDN on reducing domain gaps. A detailed comparison of different CDN settings can be found at the ablation study 7. As shown in Fig. 3, our method exhibits good generalization capability under foggy weather conditions.

Method	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mAP
Source	29.3	31.9	43.5	15.8	27.4	9.0	20.3	29.9	26.1
DA-Faster [4]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.9
DT [18]	25.4	39.3	42.4	24.9	40.4	23.1	25.9	30.4	31.5
SCDA [47]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
DDMRL [22]	30.8	40.5	44.3	27.2	38.4	<b>34.5</b>	28.4	32.2	34.6
SWDA [34]	30.3	42.5	44.6	24.5	36.7	31.6	30.2	35.8	34.8
CDN (ours)	<b>35.8</b>	<b>45.7</b>	<b>50.9</b>	<b>30.1</b>	<b>42.5</b>	29.8	<b>30.8</b>	<b>36.5</b>	<b>36.6</b>

Table 1: Cityscapes to Foggy Cityscapes adaptation.

## 5.2 Results on KITTI to Cityscapes

Different camera settings may influence the detector performance in real-world applications. We conduct the cross-camera adaptation on KITTI to Cityscapes. Table 2 shows the adaptation results on *car* category produced by Faster R-CNN with VGG16. Global and Instance represent global and local alignment respectively. The results demonstrate that CDN achieves 1.7% mAP improvements over the state-of-the-art methods. We can also find that instance feature alignment contributes to a larger performance boost than global counterpart, which is consistent with previous discovery [34, 47].

## 5.3 Results on SIM10K to Cityscapes

Following the setting of [34], we evaluate the detection performance on *car* on SIM10K-to-Cityscapes benchmark. The results in Table 3 demonstrate CDN constantly performs better than the baseline methods. CDN with both global and instance alignment achieves 49.3% mAP on validation set of Cityscapes, which outperforms the previous state-of-the-art method by 1.6% mAP.

## 5.4 Results on Synthetic to Real Data

To thoroughly evaluate the performance of the state-of-the-art methods on synthetic to real adaptation, we construct a large-scale synthetic-to-real adaptation benchmark on various public synthetic datasets, including Virtual KITTI, Synscapes and SIM10K. “All” represents using the combination of 3 synthetic datasets. Compared with SIM10K-to-Cityscapes, the proposed benchmark is more challenging in terms of much larger image diversity in both real and synthetic domains. We compare CDN with the state-of-the-art method SWDA[34] in Table 4. CDN consistently outperforms SWDA under different backbones, which achieves average 2.2% mAP and 2.1% mAP improvements on Faster-R18 and Faster-R50 respectively. Using the same adaptation method, the detection performance strongly depends on the quality of synthetic data. For instance, the adaptation performance of SIM10K is much better than Virtual KITTI. Some example predictions produced by our method are visualized in Fig. 3.

Method	Global	Instance	mAP(%)
Source only			37.1
DA-Faster [4]	✓	✓	38.3
SWDA [34]	✓	✓	43.2
SCDA [47]	✓	✓	42.9
CDN	✓		40.2
		✓	43.1
	✓	✓	<b>44.9</b>

Table 2: KITTI to Cityscapes.

Method	Global	Instance	mAP(%)
Source only			34.3
DA-Faster[4]	✓	✓	38.3
SWDA [34]	✓	✓	47.7
SCDA [47]	✓	✓	44.1
CDN	✓		41.2
		✓	45.8
	✓	✓	<b>49.3</b>

Table 3: SIM10K to Cityscapes.

Model	Method	Virtual KITTI	Synscapes	SIM10K	All
Faster-R18	Source	9.8	24.5	37.7	38.2
	SWDA[34]	15.6	27.0	40.2	41.3
	CDN	<b>17.5</b>	<b>29.1</b>	<b>42.7</b>	<b>43.6</b>
	Target		70.5		
Faster-R50	Source	13.9	29.1	41.6	42.8
	SWDA[34]	19.7	31.5	42.9	44.3
	CDN	<b>21.8</b>	<b>33.4</b>	<b>45.3</b>	<b>47.2</b>
	Target		75.6		

Table 4: Adaptation from different synthetic data to real data. mAP on car is reported on BDD100K validation. The results of supervised training on BDD100K are highlighted in gray.



Fig. 3: Example results on Foggy Cityscapes/Synscapes/SIM10K/BDD100K (from top to bottom). The results are produced by a Faster R-CNN model incorporated with CDN. The class and score predictions are at the top left corner of the bounding box. Zoom in to visualize the details.

### 5.5 Adaptation on 3D Point Cloud Detection

We evaluate CDN on adapting 3D object detector from synthetic point cloud (PreSIL) to real point cloud data (KITTI). PointRCNN [38] with backbone of PointNet++ [30] is adopted as our baseline model. Following standard metric on KITTI benchmark [38], we use Average Precision(AP) with IOU threshold 0.7 for car and 0.5 for pedestrian / cyclist. Table 5 shows that CDN constantly outperforms the state-of-the-art method PointDAN [31] across all categories, with an average improvement of 1.9% AP. We notice that instance alignment contributes to a larger performance boost than global alignment. It can be attributed by the fact that point cloud data spread over a huge 3D space but most information is stored in the local foreground points (see Fig. 4).

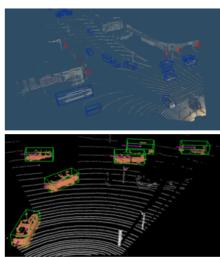


Fig. 4: Top:PreSIL;  
Bottom:KITTI.

Model	Global Instance	Car	Pedestrian	Cyclist
Source		15.7	9.6	5.6
CycleGAN[36]	✓	16.5	10.3	5.9
PointDAN[31]	✓	17.1	10.9	7.5
CDN	✓	17.3	10.5	6.0
		18.5	12.8	8.7
	✓	<b>19.0</b>	<b>13.2</b>	<b>9.1</b>
Target		75.7	41.7	59.6

Table 5: Adapting from synthetic (PreSIL) to real (KITTI) pint cloud. AP of moderate level on KITTI test is reported.

## 6 Analysis

### 6.1 Visualize and Analyze the Feature Maps

Despite the general efficiency on various benchmarks, we are also interested in the underlying principle of CDN. We interpret the learned domain embedding via appending a decoder network after the backbone to reconstruct the RGB images from the feature maps. As shown in Fig. 5, the top row shows the original inputs from Foggy Cityscapes, SIM10K and Synscapes (left to right), and the bottom row shows the reconstructed images from the corresponding features encoded with the domain embedding of another domain. The reconstructed images carry the same domain style of another domain, suggesting the learned domain embedding captures the domain attribute information and CDN can effectively transform the domain style of different domains.

Furthermore, we compute Frchet Inception Distance (FID)[14] to quantitatively investigate the difference between source and target features. FID has been a popular metric to evaluate the style similarity between two groups of images in GANs. Lower FID score indicates a smaller style difference. For easy comparison, we normalize the FID score to  $[0, 1]$  by dividing the maximum score. As shown in Table 6, the feature learned with CDN achieves significantly smaller FID score



Fig. 5: Top row: Original inputs from Foggy Cityscapes, SIM10K and Synscapes (left to right); Bottom row: Reconstructed images from features encoded with the learned domain embedding of another domain.

compared with feature learned on source domain only, suggesting CDN effectively reduces the domain gap in the feature space. Obviously, supervised joint training on source and target data gets the smallest FID score, which is verified by the best detection performance achieved by joint training. As shown in Fig. 6, synthetic-to-real has larger FID score than real-to-real dataset, since the former owns larger domain gaps.

Method	SIM to BDD		City to Foggy	
	FID	mAP	FID	mAP
Source	0.94	37.7	0.83	26.1
Joint training	0.67	79.3	0.41	49.5
SWDA [34]	0.83	40.2	0.76	34.8
CDN	0.71	42.7	0.60	36.6

Table 6: FID score and mAP.



Fig. 6: FID scores on all datasets.

## 6.2 Analysis on Domain Discrepancy

We adopt symmetric KullbackLeibler divergence to investigate the discrepancy between source and target domain in feature space. To simplify the analysis, we assume source and target features are drawn from the multivariate normal distribution. The divergence is calculated with the Res5-3 features and plotted in log scale. Fig. 7 (a) and (c) show that the domain divergence continues decreasing during training, indicating the Conditional Domain Normalization keeps reducing domain shift in feature space. Benefiting from the reduction of domain divergence, the adaptation performance on the target domain keeps increasing. Comparing with SWDA, CDN achieves lower domain discrepancy and higher adaptation performance.

Except for the quantitative measure of domain divergence, we also visualize the t-SNE plot of instance features extracted by a Faster R-CNN incorporated

with CDN. Fig. 7 (b)(d) shows the t-SNE plot of instance features extracted by a Faster R-CNN model incorporated with CDN. The same category features from two domains group in tight clusters, suggesting source and target domain distributions are well aligned in feature space. Besides, features of different categories own clear decision boundaries, indicating discriminative features are learned by our method. These two factors contribute to the detection performance on target domain.

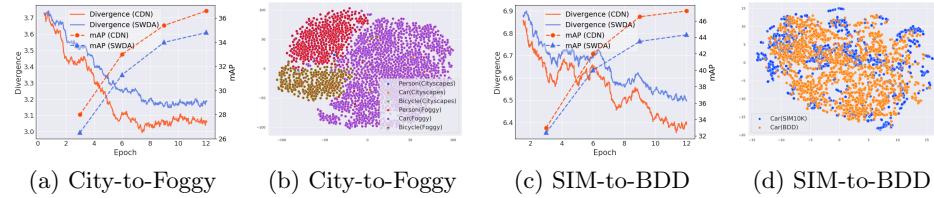


Fig. 7: (a)(c): Divergence and adaptation performance. (b)(d): t-SNE plot of instance features.

## 7 Ablation Study

For the ablation study, we use a Faster R-CNN model with ResNet-18 on SIM10K to BDD100K adaptation benchmark, and a Faster R-CNN model with VGG16 on Cityscapes-to-Foggy Cityscapes adaptation benchmark. G and I denote adopting CDN in the backbone and bounding box head network, respectively.

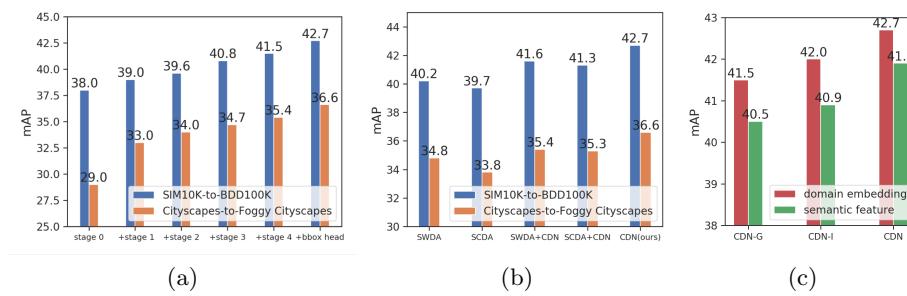


Fig. 8: (a) Adopt CDN at different convolution stages of ResNet; (b) Adopt CDN in existing adaptation frameworks; (c) Domain embedding vs. semantic features.

**Adopting CDN at different convolution stages.** Fig. 8(a) compares the results of Faster R-CNN models adopting CDN at different convolution stages.

We follow [13] to divide ResNet into 5 stages. Bbox head denotes the bounding box head network. From left to right, adding more CDN layers keeps boosting the adaptation performance on both benchmarks, benefiting from adaptive distribution alignments across different levels' representation. It suggests that adopting CDN in each convolution stage is a better choice than only aligning domain distributions at one or two specific convolution stages.

**Comparing with existing domain adaptation frameworks adopting CDN.** Fig. 8(b) shows the results of adopting CDN layer in existing adaptation methods like SWDA [34] and SCDA [47]. Directly adopting CDN in SWDA and SCDA can bring average 1.3% mAP improvements on two adaptation benchmarks, suggesting CDN is more effective to address domain shifts than traditional domain confusion learning. It can be attributed to that CDN disentangle the domain-specific factors out of the semantic features via learning a domain-vector. Leveraging the domain-vector to align the different domain distributions can be more efficient.

**Compare domain embedding with semantic features.** In Eq. 7, we can either use semantic features ( $v^s, \hat{v}^t$ ) or domain embedding ( $\mathbf{F}_d(v^s), \mathbf{F}_d(\hat{v}^t)$ ) as inputs of discriminator. Fig. 8(c) compares the adaptation performance of using semantic features with using domain embedding. Although semantic features can improve the performance over baseline, domain embedding consistently achieves better results than directly using semantic features. Suggesting the learned domain embedding well captures the domain attribute information, and it is free from some undesirable regularization on specific image contents.

**Value of  $\lambda$**  In Eq. 10, we use  $\lambda$  controls the balance between global and local regularization. Fig. 9 (left) shows the influence on adaptation performance by different  $\lambda$ . Because object detectors naturally focus more on local regions, we can see stronger instance regularization largely contributes to detection performance. In our experiments,  $\lambda$  between 0.4 and 0.5 gives the best performance.

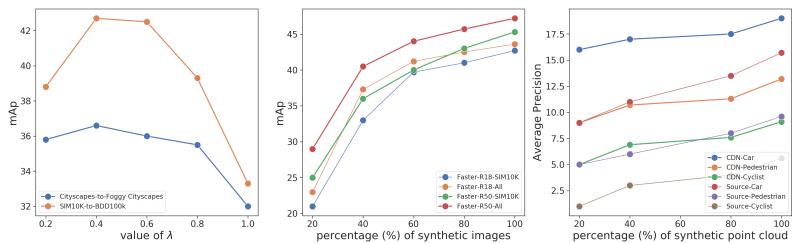


Fig. 9: Left: mAP vs. Value of  $\lambda$ ; Middle: mAP vs. Percentage (%) of synthetic image data; Right: AP vs. Percentage (%) of synthetic point cloud.

**Scale of target domain dataset** Fig. 9 middle/right quantitatively investigate the relation between real data detection performance and percentage of synthetic data used for training. “All” means to use the combination of 3 differ-

ent synthetic datasets. The larger synthetic dataset provides better adaptation performance, on both 2D image and 3D point cloud detection.

## 8 Conclusion

We present the Conditional Domain Normalization (CDN) to adapt object detectors across different domains. CDN aims to embed different domain inputs into a shared latent space, where the features from different domains carry the same domain attribute. Extensive experiments demonstrate the effectiveness of CDN on adapting object detectors, including 2D image and 3D point cloud detection tasks. And both quantitative and qualitative comparisons are conducted to analyze the features learned by our method.

## Appendix

### A1. Interpret the Domain Embedding



Fig. 10: Top row: Original inputs of Foggy Cityscapes; Middle row: Reconstructed results from features of original inputs; Bottom row: Reconstructed results from features encoded with the domain embedding of Cityscapes.

Conditional domain normalization disentangles the domain-specific attribute out of the semantic features from one domain via a learning a domain embedding to characterize the domain attribute information. In this section, we interpret

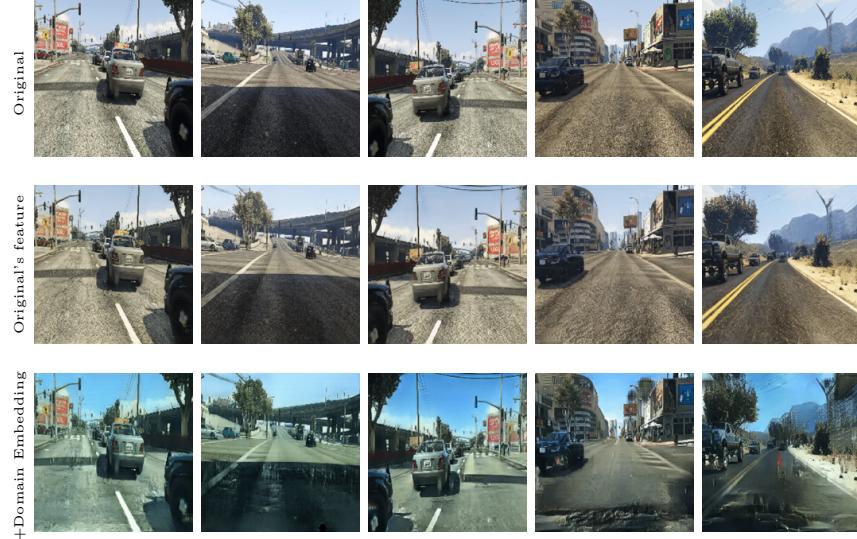


Fig. 11: Top row: Original inputs of SIM10K; Middle row: Reconstructed results from features of original inputs; Bottom row: Reconstructed results from features encoded with the domain embedding of BDD100K.

the learned domain embedding via reconstructing the RGB images from the features. As shown in Fig. 12, we first built a decoder network  $Decoder(\cdot; \theta_{dec})$  upon the backbone network  $G(\cdot; \theta_g^*)$  of fixed weights. The parameters of the backbone network are obtained in the adaptation training (see Eq. 1). The decoder network mostly mirrors the backbone network, with all pooling layers replaced by nearest up-sampling and all normalization layers removed. The decoder network is trained to reconstruct the RGB images from the features extracted by the backbone,

$$\arg \min_{\theta_{dec}} \mathcal{L} = ||Decoder(G(x; \theta_g^*); \theta_{dec}) - x||_2. \quad (11)$$

For contrast analysis, only single domain images are used to train the decoder network, i.e. the decoder for Cityscapes experiment is trained on Foggy Cityscapes images, the decoder for SIM10K experiment is trained on SIM10K images. After we got a trained decoder network, we use it to reconstruct the RGB image from features encoded with the domain embedding.

Fig. 10 shows the effect of domain embedding learned in Cityscapes to Foggy Cityscapes adaptation experiments (Section 5.1). The top row shows the inputs of Foggy Cityscapes; the middle row shows the reconstructed results from features of Foggy Cityscapes inputs; the bottom row is reconstructed results from Foggy Cityscapes features encoded with the domain embedding learned on Cityscapes. With the help of the domain embedding learned on Cityscapes, the reconstructed results from Foggy Cityscapes features no longer exhibit foggy

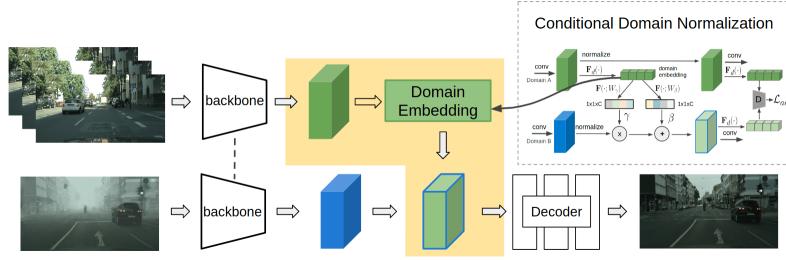


Fig. 12: Interpreting the learned domain embedding with a decoder network.

characteristics, suggesting that both Cityscapes and Foggy Cityscapes inputs are embedded into a shared latent space, where their features carry the same domain attribute. Given the domain gap bridged, the object detector supervised trained on Cityscapes also works on Foggy Cityscapes.

Fig. 11 and 13 show the reconstructed results from synthetic data’s features encoded with domain embedding of real data (BDD100K), which are learned in SIM10K-to-BDD100K and Synscapes-to-BDD10K adaptation experiments, respectively (see Section 5.4). Without the domain embedding of real data, the reconstructed images (middle row of Fig 11 and 13) still exhibit characteristic of CG (computer graphic), that look identical to the original images. When the same features of original inputs are encoded with the domain embedding of real data, the reconstructed images (bottom row of Fig 11 and 13) obviously becomes more realistic. For example, the color of the sky, the texture of the road and objects in the reconstructed images look similar to the real images. It proves that the learned domain embedding well captures the domain attribute information of real data, and it can be used to effectively translate the synthetic images towards real images.

**A.2 Visualize the Feature Maps** Despite the general efficiency on various benchmarks, we are also interested in the underlying principle of CDN. We first visualize the features of different domain images. As shown in Fig. 14, we can not easily distinguish the domain label from feature maps alone, suggesting the features from synthetic and real domain carry the same domain attribute. Besides, the same category objects across synthetic and real domain share similar activation patterns and contours, indicating the our method well preserves the feature semantics.

**A.3 More Qualitative Results** Fig. 15 shows 3D point cloud detection results on the KITTI dataset.



Fig. 13: Top row: Original inputs of Syncscapes; Middle row: Reconstructed results from features of original inputs; Bottom row: Reconstructed results from features encoded with the domain embedding of BDD100K.

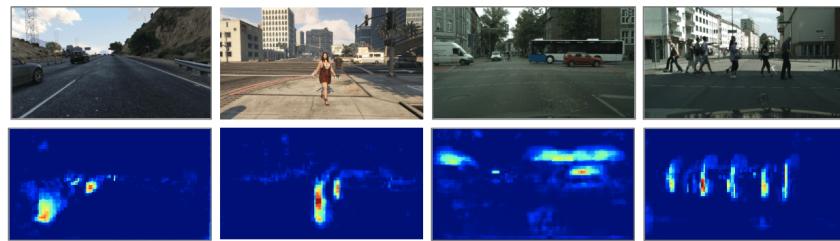


Fig. 14: Res5-3 features learned by Faster R-CNN with CDN. Left two images are from synthetic data (SIM10K) and the right two images are from real data (Cityscapes).

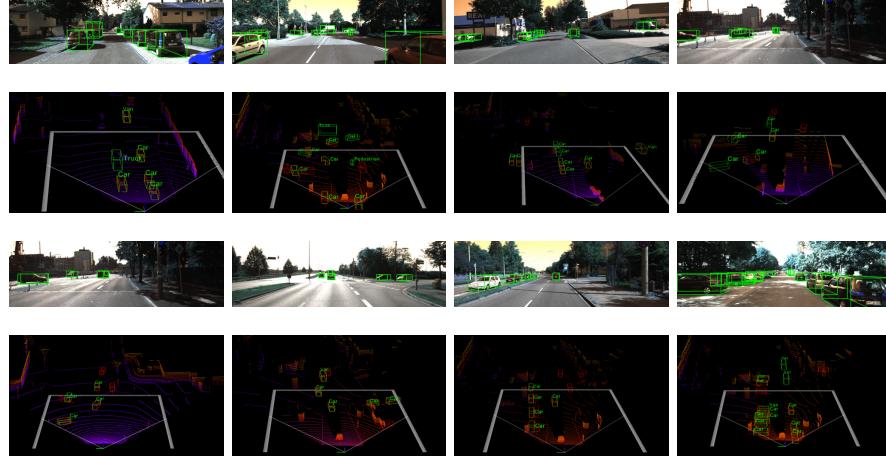


Fig. 15: Example results of PointRCNN model with Conditional Domain Normalization (with AP of moderate level of 19.0). For each example, the upper part is the image and the lower part is the corresponding point cloud. The detected objects are shown with red 3D bounding boxes. The green bounding boxes represent the ground truth.

## References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. *Machine learning* **79**(1-2), 151–175 (2010)
2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3722–3731 (2017)
3. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: Advances in Neural Information Processing Systems. pp. 424–432 (2015)
4. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3339–3348 (2018)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
6. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016)
7. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016)

8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. pp. 1180–1189 (2015)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
10. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017)
15. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
16. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
17. Hurl, B., Czarnecki, K., Waslander, S.: Precise synthetic image and lidar (presil) dataset for autonomous vehicle perception. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 2522–2529. IEEE (2019)
18. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
20. James, S., Wohlhart, P., Kalakrishnan, M., Kalashnikov, D., Irpan, A., Ibarz, J., Levine, S., Hadsell, R., Bousmalis, K.: Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12627–12637 (2019)
21. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983 (2016)
22. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12456–12465 (2019)
23. Liu, Z., Miao, Z., Pan, X., Zhan, X., Yu, S.X., Lin, D., Gong, B.: Compound domain adaptation in an open world (2019)
24. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Advances in Neural Information Processing Systems. pp. 1640–1650 (2018)
25. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. arXiv preprint arXiv:1903.07291 (2019)

26. Peng, C., Xiao, T., Li, Z., Jiang, Y., Zhang, X., Jia, K., Yu, G., Sun, J.: Megdet: A large mini-batch object detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6181–6189 (2018)
27. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
28. Peng, X., Usman, B., Saito, K., Kaushik, N., Hoffman, J., Saenko, K.: Syn2real: A new benchmark for synthetic-to-real visual domain adaptation. CoRR **abs/1806.09755** (2018)
29. Peng, X.B., Andrychowicz, M., Zaremba, W., Abbeel, P.: Sim-to-real transfer of robotic control with dynamics randomization. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 1–8. IEEE (2018)
30. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. pp. 5099–5108 (2017)
31. Qin, C., You, H., Wang, L., Kuo, C.C.J., Fu, Y.: Pointdan: A multi-scale 3d domain adaption network for point cloud representation. In: Advances in Neural Information Processing Systems. pp. 7190–7201 (2019)
32. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. The MIT Press (2009)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
34. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
35. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (Sep 2018), <https://doi.org/10.1007/s11263-018-1072-8>
36. Saleh, K., Abobakr, A., Attia, M., Iskander, J., Nahavandi, D., Hossny, M., Nahavandi, S.: Domain adaptation for vehicle detection from bird’s eye view lidar point cloud data. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
37. Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3752–3761 (2018)
38. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–779 (2019)
39. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 23–30. IEEE (2017)
40. Tsai, Y.H., Sohn, K., Schulter, S., Chandraker, M.: Domain adaptation for structured output via discriminative representations. arXiv preprint arXiv:1901.05427 (2019)
41. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176 (2017)

42. Wang, X., Yu, K., Dong, C., Change Loy, C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR (2018)
43. Wrenninge, M., Unger, J.: Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv preprint arXiv:1810.08705 (2018)
44. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
45. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 (2018)
46. Yue, X., Wu, B., Seshia, S.A., Keutzer, K., Sangiovanni-Vincentelli, A.L.: A lidar point cloud generator: from a virtual world to autonomous driving. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 458–464 (2018)
47. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
48. Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 289–305 (2018)