Domain-Invariant Disentangled Network for Generalizable Object Detection

Chuang Lin[1*] Zehuan Yuan[2] Sicheng Zhao[3] Peize Sun[4] Changhu Wang[2] Jianfei Cai[1]
[1] Dept of Data Science and AI, Monash University [2] ByteDance AI Lab
[3] Columbia University [4] The University of Hong Kong

Abstract

We address the problem of domain generalizable object detection, which aims to learn a domain-invariant detector from multiple "seen" domains so that it can generalize well to other "unseen" domains. The generalization ability is crucial in practical scenarios especially when it is difficult to collect data. Compared to image classification, domain generalization in object detection has seldom been explored with more challenges brought by domain gaps on both image and instance levels. In this paper, we propose a novel generalizable object detection model, termed Domain-Invariant Disentangled Network (DIDN). In contrast to directly aligning multiple sources, we integrate a disentangled network into Faster R-CNN. By disentangling representations on both image and instance levels, DIDN is able to learn domain-invariant representations that are suitable for generalized object detection. Furthermore, we design a cross-level representation reconstruction to complement this two-level disentanglement so that informative object representations could be preserved. Extensive experiments are conducted on five benchmark datasets and the results demonstrate that our model achieves state-of-the-art performances on domain generalization for object detection.

1. Introduction

Object detection is a fundamental yet challenging problem in computer vision. It aims to identify and localize all object instances of certain categories in an image. In the past few years, we have witnessed significant breakthroughs of supervised object detection [9, 2, 30, 24, 10, 3] on various benchmark datasets [17, 33, 8, 42]. Nonetheless, performing object detection in practice remains challenging due to the complexity and diversity of natural scenes. Learning a general object detector requires collecting a large amount data, which is highly expensive in real word scenarios with various domains. An alternative is to transfer the learned knowledge from labeled source domains to another different but related target domain. However, because of the presence of *dataset bias* or *domain shift* [35], *i.e.* the joint probability distributions of observed data and labels are different in different domains, direct transfer may not perform well.

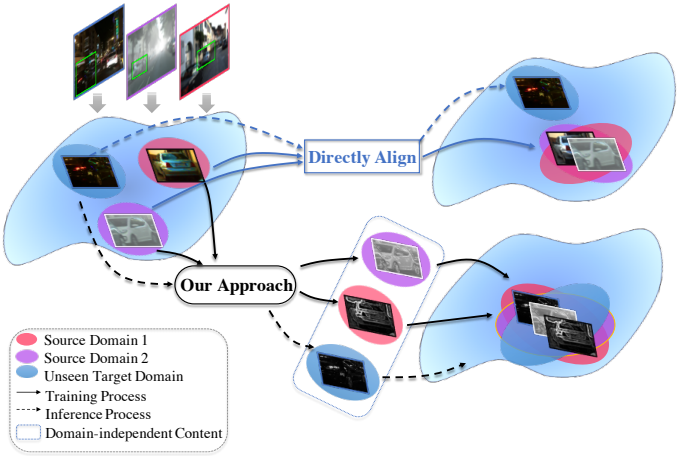


Figure 1. An illustration of our approach for domain generalizable object detection. If directly extending prior domain adaptation methods [12, 46, 40, 32, 47, 1] to the unseen domain, it will fail to generalize to the unseen domain since there is no data available in training. Our method first extracts the domain-independent object content to avoid fully matching all source domains. Further, we learn a shared feature space for domain generalization with preserved informative object representation.

Unsupervised domain adaptation (UDA) is one of the most popular attempts to remedy this problem and considerable efforts have been made [12, 46, 40, 32, 47, 1, 15]. Given the well labeled source data and the known target data without labels, the idea of UDA is to align the data distribution between the source and target domains so that the trained model on the source can well generalize to the target [45]. However, these methods still require pre-collecting target data and retraining the model for different target domains. Therefore, it is difficult to extend domain adaptation methods to the scenarios where target data is unavailable.

In this paper, we focus on domain generalizable object

*This work was performed while Chuang Lin worked as an intern at ByteDance.

detection, a more general problem which does not rely on target data and aims at learning a universal object detector directly from multiple source domains. Hopefully, the detector could perform well on any previously "unseen" target domains. The main challenge of generalizable object detection still lies in the notorious *domain shift* across multiple domains. On one hand, the domain shift is not only manifested on the image level (e.g, weather, time, scene layouts, *etc*.), but also on the instance level (e.g, object appearance, size, *etc*.). The resulting model needs to learn invariant representations on both levels. On the other hand, prior methods on domain adaptation align feature space by directly matching the distributions between the source domains and the known target domain, which is available during training. However, when target data is not observable, only matching multiple source domains would be insufficient for generalized object detection, because it might not learn a well aligned feature space for an unseen target domain, as illustrated in Figure 1.

To address the above challenges, in this paper we propose a novel framework for generalizable object detection. Inspired by the recent disentangled works [28] for image translation, we propose a Domain-Invariant Disentangled Network (DIDN) to learn a universal object detector. The network consists of three components: Image-level Disentanglement, Instance-level Disentanglement and Cross-level Reconstruction. Image-level and instance-level disentanglements aim to explicitly disentangle representation spaces to domain-independent and domain-exclusive parts. By integrating them into Faster R-CNN framework, the two-level disentanglement enables DIDN to extract generalized features suitable for object detection. We believe the consistent representation of objects in the two levels is more helpful to preserve the informative features for object detection. We further enforce a cross-level reconstruction to complement the detection model, since the two-level disentanglements are independent of each other.

In summary, the contributions of this paper are threefold:

1) We propose to generalize object detection from multiple sources to a previously unseen domain. To the best of our knowledge, this is the first work to explore domain generalization for object detection.

2) We develop a novel end-to-end learning framework termed DIDN, to learn domain-invariant representation on both image level and instance level for generalizable object detection.

3) We conduct extensive experiments on multiple benchmark datasets. DIDN outperforms the best baseline in terms of mAP by 2.2%, 2.1%, and 3.1% on Cityscapes, Foggy Cityscapes and BDD100k, respectively.

2. Related Work

2.1. Domain Adaptive Object Detection

To eliminate the domain shift, many methods have been proposed for unsupervised domain adaptive object detection [12, 46, 40, 37, 47, 1, 39, 15, 4, 20, 21, 18, 38, 19, 13]. They typically achieved feature space alignment by exploiting the target data distribution. For example, Hsu *et al*. fed source and target images to a shared feature extractor to generate center-aware features according to the centerness map module [12]. Zheng *et al*. aligned sources and target marginal distributions via multi-layer adversarial learning in the common feature space in a coarse-to-fine scheme [46]. A Graph-induced Prototype Alignment (GPA) framework was introduced for category-level domain alignment via elaborate prototype representations [40]. Saito *et al*. proposed a weak alignment model, which focused the adversarial alignment loss on sources and target that were globally similar and put less emphasis on globally dissimilar parts [32]. All these methods need to pre-collect target domain data, which may not fit the real-world situations. In our work, we aim to develop a generalizable object detection model trained on only multiple sources and tested on the unseen target.

2.2. Domain Generalization for Classification

As a practical task, domain generalization has been widely studied for image classification, which can be divided into two streams: learning domain invariance and augmenting source domains, where the former aims to align the feature space of multiple sources, while the latter broadens the learning feature space. In particular, the methods to learn domain invariance typically minimized the discrepancy among multiple source domains with adversarial loss or distance loss. Along this line, Muandet *et al*. learned an invariant transformation by minimizing Maximum Mean Discrepancy (MMD) across domains [26]. Li *et al*. extended adversarial autoencoders to align the distributions and match the aligned distribution to an arbitrary prior distribution [23]. In addition, meta-learning has also been applied to learning domain invariant features by dividing the training set into meta-train and meta-test setting. Qiao *et al*. relaxed the widely used worst-case constraint in a meta-learning scheme [29]. Li *et al*. developed a gradient-based meta-learning algorithm, which requires that the steps to improve the performance of the training domain should also improve the performance of the test domain [22]. To improve the possibility of covering the span of the target domain, the methods in the second stream source domains via diversifying samples. For image-level augmentation, Yue *et al*. transfered a source domain image to multiple styles, each dubbed an auxiliary domain [43]. For feature-level augmentation, Huang *et al*. iteratively challenged the dom-

inant features activated, and forced the network to activate remaining features that correlate with labels [14]. Different from these works, we focus on the object detection problem, which is more challenging as domain shift occurs on both image level and instance level. To the best of our knowledge, this is the first work from the domain generalizable object detection perspective.

2.3. Domain Generalization beyond Classification

Recently, the outstanding performance of domain generalization works [43, 16, 44, 36, 34] even surpassed domain adaptation methods, which stimulates its applications on other tasks, such as semantic segmentation [43, 44], person ReID [34], face presentation [36, 16] and so on. Particularly, Song *et al*. proposed a deep ReID model to learn a mapping between a person image and its identity classifier, using a single shot [34]. Jia *et al*. developed a feature generator to make the real faces from different domains indistinguishable, but excluding the fake ones, thus forming a single-side adversarial learning [16]. Zhang *et al*. proposed enhancing the generalization ability of the segmentation model by exploiting the model-agnostic learning in training and developing the target-specific normalization in testing [44]. However, object detection is a technically different problem from the above tasks. Compared with classification task, we would pay more attention to the domain-invariant of object of interest, denoted as region parts. In this work, we build a generalizable model for object detection with considering region parts in an end-to-end fashion.

3. Problem Setup

Let $\mathcal{X}$ denote a nonempty input space, and $\mathcal{Y}$ denote an arbitrary output space. We define $\mathbb{B}_{\mathcal{X}\times\mathcal{Y}}$ as a set of all probability distribution on $\mathcal{X}\times\mathcal{Y}$. Formally, a domain is a joint distribution $\mathbb{P}_{XY}$ sampled from $\mathbb{B}_{\mathcal{X}\times\mathcal{Y}}$. Domains are observed not directly but usually via datasets. We consider the domain generalization scenario with multiple labeled source domains $S_1, S_2, \cdots, S_M$, where M is the number of sources. In the ith source domain, $S_i = (x_i^j, y_i^j)_{j=1}^{N_i}$ is sampled from $\mathbb{P}_{XY}^i$, where N_i denotes the number of samples in S_i, x_i^j denote the observed images and $y_i^j = (b_i^j, c_i^j)$ denote the corresponding labels with the bounding-box coordinates b and their associated categories c. Although x_i^j $(i \in M)$ under different source domains are from the same input space, their distributions are different, likely with complex overlaps and interact relationships. Unless otherwise specified, we assume that the $y_i^j \in \mathcal{Y}$ share the same set of classes.

Objective Analysis. Using multiple seen domains $S_1, S_2, \cdots, S_M$, our goal is to produce a detection model that could perform well on target domain $S_T = \{x_t^j\}_{j=1}^{N_t}$ drawn from the unknown distribution $\mathbb{P}_{XY}^T$. Our motivation can be illustrated through the following example. Considering A is the domain containing family cars in foggy London, while B contains sports cars in German streets in rainy days. Our idea is to find a shared domain representing a car in the street regardless the domain specific information such as weather, city scene, and car style. .In other words, we want to disentangle the sources into a shared domain C: $g(S_i) \sim \mathbb{P}^C$ and specific domain D: $f(S_i) \sim \mathbb{P}^D$. In addition, to better regularize the disentanglement, we further introduce a function d to reconstruct the original distribution:

$$d(g(S_i), f(S_i)) \sim \mathbb{P}_{XY}^i. \tag{1}$$

The mapping g is expected to remove the domain-specific information from the sources, preserving the shared object information. In this way, the model trained on the shared domain can be expected to perform well on any previously "unseen" target domain.

4. Method

Figure 2 gives an overview of our proposed domain-invariant disentangled network for generalizable object detection. It consists of three major components: image-level disentanglement, instance-level disentanglement and cross-level reconstruction. The first two components aim to disentangle the representation space into the shared representation space C and the specific representation space D at image and instance levels, respectively, while the last component is to connect the two disentanglements via cross-level reconstruction. In the following, we describe each component in detail.

4.1. Image-level Disentanglement

Since the target domain data is unavailable, learning domain-independent representation from multiple sources is crucial for model generalization. To generalize to unseen target domain, it is necessary to learn domain-independent image content from multiple sources. Image-level disentanglement aims to explicitly disentangle image representation into domain-independent and domain-exclusive parts. Inspired by [28], the disentanglement is realized by the image reconstruction with domain adversarial training. Specifically, a set of encoders are learned to disentangle domain-independent image content from multiple sources. For each source domain S_i, we introduce an encoder E_i to extract the domain exclusive part, and another encoder E_c to extract domain-independent image content. E_c is shared among multiple sources and also serves as the detector backbone. Since the exclusive part and the shared image content ought to restore images, we employ a generator G_{img} to ensure information integrity. The corresponding reconstruction loss

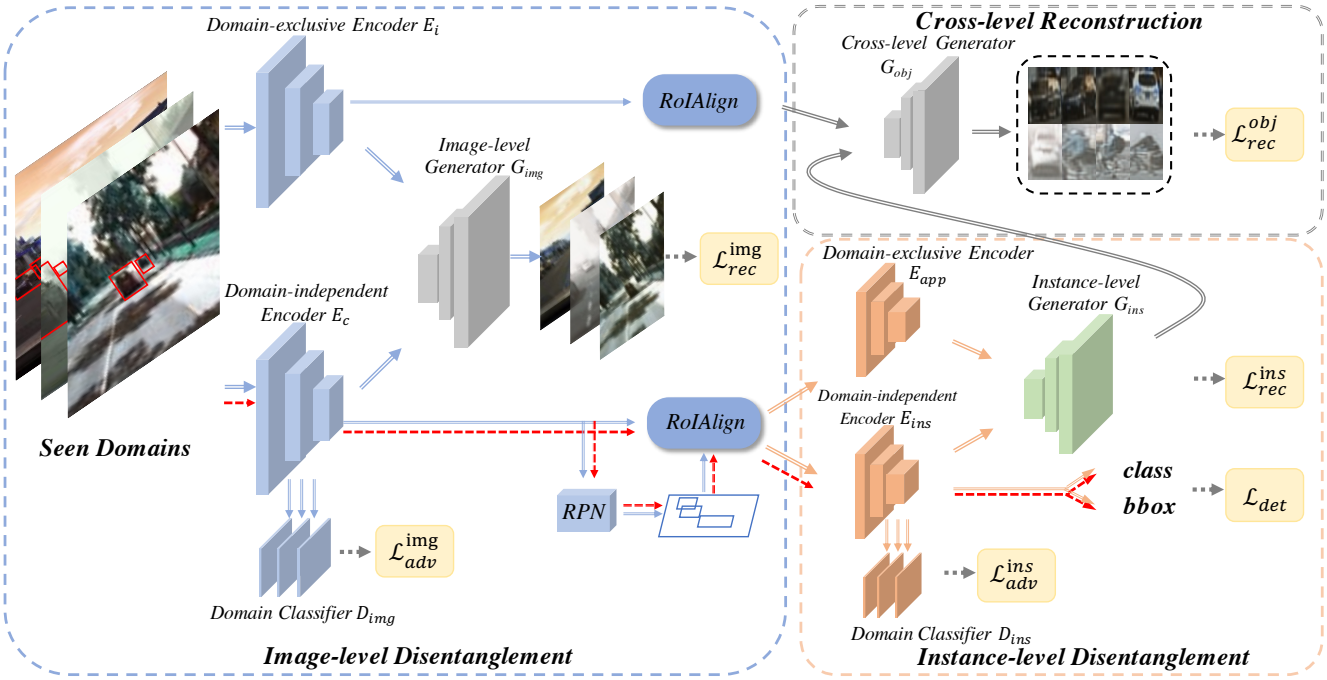


Figure 2. Overview of our proposed novel framework termed DIDN. The color solid lines with arrows indicate the operations in the training stage, while the red arrows represent the inference pipeline.

is defined as:

$$\mathcal{L}_{rec}^{img} = \mathbb{E}_{x_i \sim S_i}(G_{img}(E_i(x_i), E_c(x_i)) - x_i)^2. \quad (2)$$

To encourage the image content from different domains being similar, a set of pair-wise classifiers $D^{i,j} = \{D_{l_{th}}^{i,j}\}_{i \neq j}$ are trained to distinguish source i and source j in the l-th convolution block, while the encoder E_c aims to fool them. The hierarchical domain classifiers are constructed to catch rich intermediate information from the domain-independent encoder. We alternatively optimize the encoders and the hierarchical domain classifiers to form a min-max game for the domain-independent feature space. The corresponding adversarial loss function can be written as:

$$\begin{aligned}\mathcal{L}_{adv}^{img} = \mathbb{E}_{x_i \sim S_i} \sum_{l=1}^{L} \log[1 - D_l^{i,j}(E_c(x_i))] \\ + \mathbb{E}_{x_j \sim S_j} \sum_{l=1}^{L} \log D_l^{i,j}(E_c(x_j)).\end{aligned} \quad (3)$$

Eq. (3) essentially encourages D_l^{ij} to classify the shared content features of x_i and x_j into 0 and 1, while encouraging E_c to fool D_l^{ij} to generate the opposite results. By doing this over all pair-wise domain classifiers, we ensure the domain-specific feature is largely removed from the content representation $E_c(x_i)$, i.e. unable to recognize domain i from $E_c(x_i)$. After that, the content features are fed to the object detector.

4.2. Instance-level Disentanglement

In generalizable object detection, the domain shift is not only manifested on the image level, but also on the instance level, which includes object appearance, size, viewpoint, *etc*.

For each possible region proposals $r_{i,k}$, an RoIAlign layer (RoI) [10] is utilized to extract a fixed-size feature map $p_{i,k} = RoI(E_c(r_{i,k}))$. Similar to the image level, instance-level disentanglement is realized by the instance reconstruction with domain adversarial training. Additionally, as the instance is related to the category information, we employ conditional domain classifiers. Specifically, we apply a pair of encoders E_{app} and E_{ins} to extract the appearance and the instance content information, respectively. E_{ins} also serves as the object encoder in RoI Head of Faster R-CNN. A generator G_{ins} is further used to reconstruct the instance feature map $p'_{i,k}$. The instance-level reconstruction loss function is defined as:

$$\mathcal{L}_{rec}^{ins} = \mathbb{E}_{x_i \sim S_i}(p'_{i,k} - p_{i,k})^2. \quad (4)$$

$$p'_{i,k} = G_{ins}(E_{app}(p_{i,k}), E_{ins}(p_{i,k})). \quad (5)$$

Specifically, we extend the origin domain classifiers to conditional domain classifiers $\{D_{ins}^{i,j}\}_{i \neq j}$ with the predicted category information $c_{i,k}$ of proposal $r_{i,k}$. Similar to

Dataset	Size	Scene	Weather	Time	*#C*
Cityscapes **(C)**	3475	City street	Good/med. weather conditions	Daytime	8
Fog Cityscapes **(F)**	3475	City street	Fog	Daytime	8
SIM 10k **(S)**	10000	Synthetic street	Sun, Fog, Rain, Haze	Night, Morning, Dusk	1
KITTI **(K)**	7481	City street, Highway, Rural	Good/med. weather conditions	Daytime	9
BDD100k **(B)**	100000	City street, Highway	Rain, Snow, Cloud, Overcast	Daytime, Night	11

Table 1. Comparison of datasets. "#C" is the number of categories in the datasets. It is clear that the existing datasets suffer from the domain shift problem, due to *size, scene, weather, time* and *categories* interacted and disjoint.

Eq. (3), the instance-level adversarial loss can be written as:

$$\begin{aligned}\mathcal{L}_{adv}^{ins} = \log[1 - D_{ins}^{ij}(E_{ins}(p_{i,k})|c_{i,k})] \\ + \log D_{ins}^{ij}(E_{ins}(p_{j,k})|c_{j,k}).\end{aligned} \tag{6}$$

In this way, the disentanglements at the image-level and the instance-level lead to the alignments of multiple sources from image content to instance representation in a domain-invariant way.

4.3. Cross-level Reconstruction

To further complement the two-level disentanglement, we design a cross-level reconstruction to preserve the information. Since Faster R-CNN has to account for varying sizes of proposals $r_{i,k}$, some object information is lost in RoIAlign. To circumvent this issue, a cross-level generator is trained to reconstruct objects at the pixel level for the information integrity of the two-level disentanglement, for which the domain-exclusive information at the image level is also fed into the generator.

We reconstruct the pixel-level objects with the domain clues in image level and the reconstructed feature map $p'_{i,k}$ in instance level with the following cross-level reconstruction loss:

$$\mathcal{L}_{obj}^{rec} = \sum_{k \in \mathcal{O}} (G_{obj}(p'_{i,k}, RoI(E_i(r_{i,k}))) - r_{i,k})^2, \tag{7}$$

where $\mathcal{O}$ denotes all possible region proposals.

4.4. DIDN Learning

The overall objective loss function of DIDN can be written as:

$$\begin{aligned}\mathcal{L}_{DIDN} = \mathcal{L}_{det} + \lambda_a(\mathcal{L}_{adv}^{img} + \mathcal{L}_{adv}^{ins}) \\ + \lambda_r(\mathcal{L}_{rec}^{img} + \mathcal{L}_{rec}^{ins}) + \lambda_c \mathcal{L}_{obj}^{rec},\end{aligned} \tag{8}$$

where λ_a, λ_r and λ_c denote the weights to balance the adversarial losses, the reconstruction losses and the cross-level reconstruction loss at the image and instance levels, and L_{det} includes all the standard detection losses in Faster R-CNN. The training process is essentially to solve the following optimization problem:

$$R^* = \arg \min_{E,G} \max_{D} \mathcal{L}_{DIDN}, \tag{9}$$

where E, G, and D represent all the encoders, generators, domain classifiers in our model. The inference is of light weight (see red arrows in Fig. 2), since many building blocks are only needed for training. The inference speed is the same as Faster R-CNN.

5. Experiments

5.1. Experimental Settings

Datasets. Consider that we aim to simulate a real-world scenario where a detection model is likely to be trained with many public datasets, in the hope that it can generalize well to an unseen domain. To this end, we adopt many existing large-scale object detection datasets including Cityscapes [5], Foggy Cityscapes [5], SIM 10k [17], KITTI [8] and BDD100k [41]. We summarize these five datasets in Table 1.

Cityscapes [5] dataset is an urban scene dataset for driving scenarios. The images are captured by a car-mounted video camera. *Foggy Cityscapes* [5] is a synthetic foggy dataset in that it simulates fog on real scenes. The images are rendered using the images and depth maps from *Cityscapes*. They both have 2,975 images in the training set, and 500 images in the validation set. *BDD100k* [41] is collected by a real driving platform and captured on the streets. It is a satisfying large-scale, diversified dataset with time information. We only use the validation set including 10,000 images in our experiments. *KITTI* [8] is constructed by a standard station wagon with two high-resolution video cameras, which thus has cross camera difference. *KITTI* [8] is a autonomous driving database which contains 7,481 images. *SIM 10k* [17] includes 10,000 images which are rendered by the gaming engine Grand Theft Auto.

Implementation Details. There are eight shared categories with instance labels in Cityscapes, Foggy Cityscape and BDD100k. However, only *Car* is annotated in SIM 10k and KITTI. Therefore, we conduct experiments in two settings. The first one is using Cityscapes, Foggy Cityscapes and BDD100k with eight shared categories, where two of them serve as the source domains and the left one serves as the target domain. Note that for BDD100k, since there are few "train" objects, we only evaluate seven common categories of BDD100k. The second setting is using all

DG Setting	Methods	person	rider	car	truck	bus	train	motor	bike	mAP
F & B to C	Single-best	36.0	39.8	53.6	15.8	31.4	11.5	26.9	35.0	31.2
	Source-combined	43.0	48.9	62.7	**42.7**	55.9	39.4	34.8	37.9	45.3
	Directly Align	41.6	**49.2**	61.5	40.3	57.7	42.2	35.0	38.5	45.7
	DIDN (Ours)	**43.6**$_{\uparrow 0.6}$	46.2$_{\downarrow 3.0}$	**63.2**$_{\uparrow 0.5}$	41.9$_{\downarrow 0.8}$	**60.9**$_{\uparrow 3.2}$	**51.1**$_{\uparrow 8.9}$	**36.0**$_{\uparrow 1.0}$	**41.3**$_{\uparrow 2.8}$	**47.9**$_{\uparrow 2.2}$
	Oracle - Train on Target	44.7	51.6	63.5	42.0	58.6	45.8	42.0	44.4	49.1
C & B to F	Single-best	25.0	30.0	30.0	14.2	18.5	5.0	15.0	26.6	20.5
	Source-combined	31.7	**39.5**	48.9	**28.2**	34.3	12.9	21.8	32.8	31.3
	Directly Align	25.6	39.3	42.7	22.1	34.0	19.5	22.1	30.1	27.4
	DIDN (Ours)	**31.8**$_{\uparrow 0.1}$	38.4$_{\downarrow 1.1}$	**49.3**$_{\uparrow 0.4}$	27.7$_{\downarrow 0.5}$	**35.7**$_{\uparrow 1.4}$	**26.5**$_{\uparrow 7.0}$	**24.8**$_{\uparrow 2.7}$	**33.1**$_{\uparrow 0.3}$	**33.4**$_{\uparrow 2.1}$
	Oracle - Train on Target	36.1	47.1	52.7	32.1	49.5	56.0	36.0	37.0	43.3
F & C to B	Single-best	27.9	27.5	43.1	16.6	15.1	-	5.6	21.0	19.6
	Source-combined	30.0	22.6	44.6	16.5	11.6	-	6.2	20.1	18.9
	Directly Align	31.3	21.4	**44.8**	18.6	13.3	-	5.8	20.9	19.1
	DIDN (Ours)	**34.5**$_{\uparrow 3.2}$	**30.4**$_{\uparrow 7.8}$	44.2$_{\downarrow 0.6}$	**21.2**$_{\uparrow 2.6}$	**19.0**$_{\uparrow 3.9}$	-	**9.2**$_{\uparrow 3.0}$	**22.8**$_{\uparrow 1.8}$	**22.7**$_{\uparrow 3.1}$
	Oracle - Train on Target	35.5	32.1	50.9	33.7	28.9	-	13.5	27.5	30.8

Table 2. Results (%) of the domain generalization on Cityscapes (C) [5], Foggy Cityscapes (F) [5], and BDD100k (B) [41]. The best category AP and mAP are highlighted in bold. *Single-best* indicates choosing the best performance from Faster R-CNN trained on each source, *Source-combined* indicates combining all source domain as a traditional single domain, and *Directly Align* indicates extending the domain adaptation method which directly matches all source domains in feature space.

Methods	***C&B&S&F to K***	***C&F&K&S to B***
Single-best	74.3	38.6
Source-combined	75.2	48.2
Directly Align	75.6	45.1
DIDN (Ours)	**76.8**$_{\uparrow 1.2}$	**52.3**$_{\uparrow 4.1}$

Table 3. Results (%) of the domain generalization on Cityscapes (C) [5], Foggy Cityscapes (F) [5], BDD100k (B) [41], Sim 10k (S) [17], and KITTI (K) [8].

the five datasets but considering only the *Car* category for multi-source domain generalization. In the experiments, we choose four datasets as source domains, and the left one as unseen target domain. Note that SIM 10k is only considered as a source domain dataset but not a target dataset. This is because it is a simulation dataset and the generalization from simulation to real is meaningful, but not the reverse.

We adopt Faster R-CNN [31] with RoIAlign [10] and implement our model with maskrcnn-benchmark [25] in Pytorch [27]. Although single-stage detectors have emerged as a popular paradigm, Faster R-CNN is considered as the most representative of two-stage detectors and is still a top-performing detector. We will consider other backbones in future. ResNet-50 [11] pre-trained on ImageNet [6] is used as the backbone of the detector, which is also the domain-independent encoder in our model. In all experiments, unless specified, all training and testing images are resized such that their shorter side has 600 pixels. We use SGD optimizer, first trained with a learning rate of $lr = 0.002$ for $120K$ iterations, and then $lr = 0.0002$ for another $60K$ iterations. The learning rate warm up strategy [31] is used in the first 200 iterations of training. We follow [25] to set

DG Setting	*Img*	*Ins*	*Comp*	mAP
F & B to C	✓			47.1
		✓		46.1
	✓	✓		47.3
	✓	✓	✓	**47.9**
C & B to F	✓			32.1
		✓		31.9
	✓	✓		33.2
	✓	✓	✓	**33.4**
F & C to B	✓			21.1
		✓		20.3
	✓	✓		22.2
	✓	✓	✓	**22.7**

Table 4. Effectiveness of each component in DIDN for domain generalization on Cityscapes (C), Foggy Cityscapes (F), and BDD100k (B). *Img*: Faster R-CNN with image-level disentanglement; *Ins*: Faster R-CNN with instance-level disentanglement; *Comp*: connecting the two-level disentanglements with cross-level reconstruction.

the hyper-parameters.

In our experiments, two NVIDIA V100 GPUs are used for training. Each batch is composed of two images from each source domain, e.g. eight images per batch from four sources to fit the two GPUs. We employ mean average precisions (mAP) with a threshold of 0.5 to evaluate the results of all the classes.

Baselines. We include the original Faster R-CNN model as a baseline, which is trained on the source domains, without considering the domain gap. We consider two variants: (1) single-best, i.e. trained on each single source, and we choose the single best performanceof all the trained models

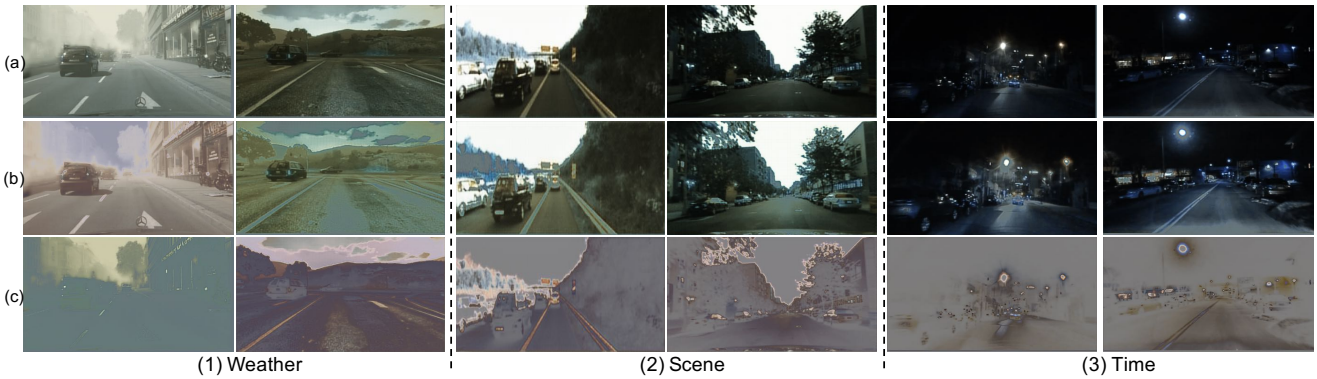


Figure 3. Visualization of the image-level disentanglement results. (a),(b),(c) are respectively the source images, and disentangled content and style images.

on the unseen target; (2) source-combined, i.e. all source domains are combined into a traditional single source. In addition, we extend the domain adaptation method [7] as another baseline named Directly Align, where a domain classifier was trained to directly align all sources in a shared feature space. This baseline is constructed by adding a domain classifier connected to the Faster R-CNN backbone, to ensure the feature distribution from different domains as indistinguishable as possible. We also report the results of an oracle setting in Table 2, where the model is both trained and tested on the target domain.

5.2. Domain Generalization Results

Table 2 and Table 3 give the quantitative results of different methods under the two different settings, respectively.. From the results, we have the following observations.

(1) The source-only methods including single-best and source-combined obtain the worst performance. The source-only method i.e. directly transferring the Faster R-CNN trained on the sources to the target drops greatly in all settings. This is due to domain shift or dataset bias. The specific domain clues cause the low transferability of Faster R-CNN to unseen target domain.

(2) Training with simply combing multiple sources does not guarantee a better performance than the corresponding single-best method. For example, in the setting of **F**&**C** to **B** in Table 2, the performance of single-best is better than the source-combined method. This suggests that although combining multiple sources results in more training data, they may interfere with each other.

(3) Directly aligning the distributions of all source domains is insufficient for generalizable object detection. In both settings of **C**&**B** to **F** and **C**&**F**&**K**&**S** to **B**, directly aligning sources failed to increase generalization ability. This indicates domain adaption methods would not work in domain generalization, where there is no target data available during training.

(4) Table 2 shows that as compared to the baselines, our method achieves 2.2%, 2.1% and 3.1% improvement in Cityscapes, Foggy Cityscapes and BDD100k, respectively. We can see that the proposed method is able to alleviate domain gap over most of the categories. Specifically, even there are only a few "train" annotations in BDD100k, the performance of "train" has been significantly improved. Moreover, our method has a significant improvement in "bike" and "motor" categories, which are highly similar in object appearance. These results further reveal that the instance-level disentanglement in our model is able to learn the domain-invariant instance representation. We find that the "bus" and "car" often appear concurrently in an image. The improvement indicates the proposed method is able to eliminate spatial interaction. Additionally, we find that the proposed method is comparable to or even better than the oracle model in several categories.

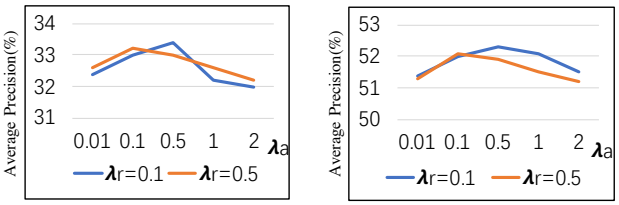
(5) As shown in Table 3, for the domain generalization on Cityscapes, Foggy Cityscapes, BDD100k, Sim10k, and KITTI, the proposed method achieves 1.2% and 4.1% improvement, respectively. Compared with the results of BDD100k, the improvement in KITTI is relatively small. The reason is that the performance is already good for the single source generalizing to target, due to the fact that scenes are similar between Cityscape and KITTI. From another perspective, all four sources have a clear gap to BDD100k, which proves our model has the ability to generalize well on unseen target.

5.3. Ablation Study

Effectiveness of each component. To prove the effectiveness of the different components in the proposed model, we provide the results of each component in Table 4. Both image-level and instance-level disentanglements achieve better results than the baselines, which demonstrates the domain gap exists in both image style and instance appearance. Image-level and instance-level disentan-

Methods	person	rider	car	truck	bus	train	motor	bike	mAP
Source-only	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
SW-DA [32] *CVPR'19*	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
SC-DA [47] *CVPR'19*	33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.5	35.9
MTOR [1] *CVPR'19*	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
ICR-CCR [39] *CVPR'20*	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
Coarse-to-Fine [46] *CVPR'20*	34.0	46.9	52.1	**30.8**	43.2	29.9	34.7	37.4	38.6
GPA [40] *CVPR'20*	32.9	**46.7**	54.1	24.7	45.7	**41.1**	**32.4**	38.7	39.5
Center-Aware [12] *ECCV'20*	**41.5**	43.6	**57.1**	29.4	44.9	39.7	29.0	36.1	40.2
DIDN(Ours)	38.3	44.4	51.8	28.7	**53.3**	34.7	**32.4**	**40.4**	**40.5**
Oracle - Train on Target	36.1	47.1	52.7	32.1	49.5	56.0	36.0	37.0	43.3

Table 5. Results (%) of doamin adaptation from Cityscape ***(C)*** to Foggy Cityscape ***(F)***. The best category AP and mAP are emphasized in bold.



(a) Foggy Cityscapes ***(F)*** as Target (b) BDD100k ***(B)*** as Target

Figure 4. Sensitivity of λ in Eq. (8) of the proposed ***DIDN***.

glements are complement to each other, so adding them together leads the performance improvement. Adding cross-level object reconstruction could further improve the precision.

Parameter Sensitivity. We analyze the impact of the hyper-parameter values in Eq. (8) to the object detection results by using λ_r and λ_a as examples. For different weight values λ_r on the reconstruction loss, we compute the achieved average precision over different weight values λ_a on the adversarial loss for both **C**&**B** to **F** and **C**&**F**&**K**&**S** to **B**. The results are reported in Figure 4. From the results, we find it is a good choice to set $\lambda_a = 0.5, \lambda_r = 0.1$.

Visualization. We visualize the results of image-level disentanglement to demonstrate the interpretability of the proposed method in Figure 3. We can see that our model is effective in disentangling the source image to the domain-independent content and domain-specific style. For bad weather cases in columns (1), the result of the content image is a good elimination of fog or overcast information, leaving the structure of objects preserved. For different scenes (city and highway) in columns (2), the corresponding content images are uniformly changed to the shared style. For the night cases in columns (3), compared with the sources, the detected objects are clearer, which indicates that the structures are preserved, e.g. the black car in the dark can be easily observed after the generalization.

5.4. Extension to Domain Adaptation

Now we conduct more experiments using the domain adaptation setting and compare our results with previous state-of-the-art works. In this section, we use the labeled source data and unlabeled target data as domain adaptation setting. We view source-only, i.e. train on the source domains and directly test on the target domain, as a lower bound of DA. The compared previous state-of-the-art methods include SW-DA [32], SC-DA [47], MTOR [1], GPA [40] and so on. Due to the limitation of our method that the cross-level reconstruction component need ground truth bounding box, we just simplify our model and remove it. Since most of the previous works conducted adaptation on weather transfer task, we present the adaptation mAP comparison on Cityscapes to Foggy Cityscapes. As seen in Table 5, DIDN has also achieved the best results under unsupervised domain adaptation setting. It is worth noting that in some categories, our method even surpasses the results of supervised methods that use labeled target domain.

6. Conclusion

In this paper, we have proposed a novel framework, termed Domain-Invariant Disentangled Network (DIDN), for generalizable object detection. To handle data from unseen domain, we integrate a two-level disentanglement into Faster R-CNN. We have conducted extensive experiments on five benchmark datasets which demonstrates the superior performance of our proposed method. For further study, we will explore the scenario where there are unseen object categories in target domain and investigate multi-modal DG, e.g. consider both images and LiDAR data.

Acknowledgments

This research is partially supported by Monash FIT Start-up Grant.

References

[1] 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 Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[2] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 354–370. Springer, 2016.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[4] 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.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] 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. *arXiv preprint arXiv:2008.08574*, 2020.

[13] 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 Winter Conference on Applications of Computer Vision*, pages 749–757, 2020.

[14] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *arXiv preprint arXiv:2007.02454*, 2020.

[15] 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.

[16] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020.

[17] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016.

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

[19] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6092–6101, 2019.

[20] 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 Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.

[21] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020.

[22] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. *arXiv preprint arXiv:1710.03463*, 2017.

[23] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[25] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of instance segmentation and object detection algorithms in pytorch. *Accessed: Apr*, 29:2019, 2018.

[26] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[28] Ori Press, Tomer Galanti, Sagie Benaim, and Lior Wolf. Emerging disentanglement in auto-encoder based unsupervised image content transfer. *arXiv preprint arXiv:2001.05017*, 2020.

[29] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

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

[33] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[34] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 719–728, 2019.

[35] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[36] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6678–6687, 2020.

[37] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019.

[38] 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 International Conference on Computer Vision Workshops*, 2019.

[39] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020.

[40] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020.

[41] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020.

[42] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[43] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2100–2110, 2019.

[44] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Generalizable semantic segmentation via model-agnostic learning and target-specific normalization. *arXiv preprint arXiv:2003.12296*, 2020.

[45] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

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

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