Single-Domain Generalized Object Detection in Urban Scene via Cyclic-Disentangled Self-Distillation

Aming Wu, Cheng Deng*
School of Electronic Engineering, Xidian University, Xi'an, China
amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

Abstract

In this paper, we are concerned with enhancing the generalization capability of object detectors. And we consider a realistic yet challenging scenario, namely Single-Domain Generalized Object Detection (Single-DGOD), which aims to learn an object detector that performs well on many unseen target domains with only one source domain for training. Towards Single-DGOD, it is important to extract domain-invariant representations (DIR) containing intrinsical object characteristics, which is beneficial for improving the robustness for unseen domains. Thus, we present a method, i.e., cyclic-disentangled self-distillation, to disentangle DIR from domain-specific representations without the supervision of domain-related annotations (e.g., domain labels). Concretely, a cyclic-disentangled module is first proposed to cyclically extract DIR from the input visual features. Through the cyclic operation, the disentangled ability can be promoted without the reliance on domain-related annotations. Then, taking the DIR as the teacher, we design a self-distillation module to further enhance the generalization ability. In the experiments, our method is evaluated in urban-scene object detection. Experimental results of five weather conditions show that our method obtains a significant performance gain over baseline methods. Particularly, for the night-sunny scene, our method outperforms baselines by 3%, which indicates that our method is instrumental in enhancing generalization ability. Data and code are available at https://github.com/AmingWu/Single-DGOD.

1. Introduction

Recent years have witnessed the rapid development of deep learning based object detection [2,6,37,43,45], which assumes that the training and test data are from the same domain. However, in real applications, when object detectors trained on source domain data are applied to unseen target domains, these detectors usually suffer from poor generalization, due to the domain-shift impact [8,34].

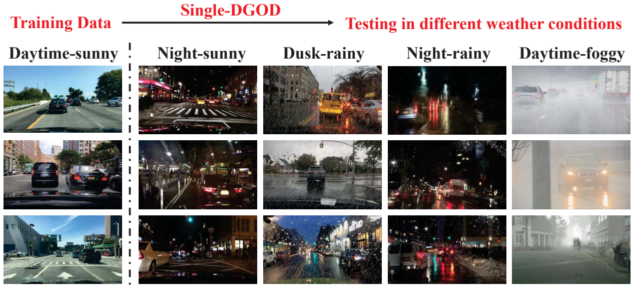

Figure 1. Illustration of Single-Domain Generalized Object Detection (Single-DGOD) with the urban-scene detection dataset. The dataset contains five domains with different weather conditions: daytime-sunny, night-sunny, dusk-rainy, night-rainy, and daytime-foggy. Single-DGOD aims to train a detector on one source domain dataset (e.g., the daytime-sunny scene) and generalize well to multiple target domains. Extracting domain-invariant representations is beneficial for generalizing a detector to unseen domains.

To alleviate the domain-shift impact, existing researches mainly focus on domain adaptation and domain generalization. In general, domain adaptation [4, 11, 36, 46] aims to align the data distribution from one annotated source domain to that from one target domain without annotations. During training, the aligned methods often need to access both source and target domain data, which results in these methods can not well adapt to other unseen target domains. Besides, when the target domain is a compound of multiple different data distributions [27], the generalization ability of the aligned methods tends to be weak, which attenuates the performance of object detection.

Domain generalization (DG) [24, 25, 39, 40, 52] aims to generalize a model to an unseen target domain by learning from multiple source domains, which is considered to be more challenging than domain adaptation. Generally, most DG methods [1, 12, 23, 25] try to learn a shared representation across multiple source domains. However, the performance of these methods highly depends on the number of source domains [9, 53]. And in the real world, collecting multiple source datasets is time-consuming and labor-intensive, which limits the application of DG methods.

[*] Corresponding author.

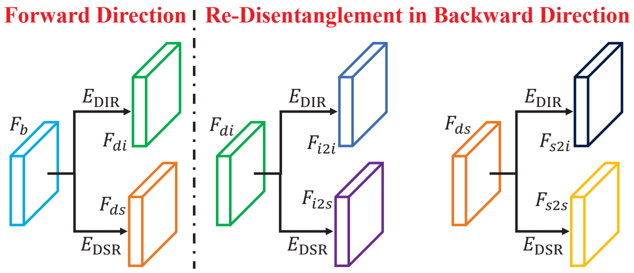

Figure 2. For Single-DGOD, to improve the disentangled ability without using domain-related annotations, we propose a cyclic-disentangled module. For the forward direction, given feature map F_b, two extractors E_{DIR} and E_{DSR} are designed to extract F_{di} and F_{ds}. For the backward direction, we separately take F_{di} and F_{ds} as the input of the module and perform re-disentanglement. It is worth noting that during the forward and backward direction, the parameters in E_{DIR} and E_{DSR} are shared.

To further explore improving the generalization ability of object detectors, we propose a realistic yet challenging task, namely Single-Domain Generalized Object Detection (Single-DGOD). As shown in Fig. 1, given one source domain dataset for training, e.g., the daytime-sunny scene, the goal of Single-DGOD is to well generalize an object detector to multiple unseen target domains, e.g., the night-sunny, dusk-rainy, night-rainy, and daytime-foggy scenes. Since multiple source domains and domain-related annotations (e.g., domain labels) are not accessible, existing DG methods can not be directly utilized to tackle this task.

Recent studies [7, 14, 18] have shown that extracting domain-invariant representations (DIR) containing intrinsical object characteristics is helpful for improving the generalization ability. To this end, many methods [31, 42, 44] directly utilize the domain-related annotations as the supervision to disentangle DIR from domain-specific representations (DSR). However, when domain-related annotations are not available, how to well extract DIR from the input visual features remains under-explored. Thus, in this paper, we mainly focus on disentangling DIR without using domain-related annotations. And we propose a method, i.e., cyclic-disentangled self-distillation, for Single-DGOD.

Specifically, we first present a cyclic-disentangled module to obtain DIR. As shown in Fig. 2, for the forward direction of the cycle, we separately design a DIR and DSR extractor to disentangle DIR and DSR from the feature maps extracted by a backbone network, e.g., ResNet [17]. The backward direction takes the disentangled DIR and DSR as the input of the DIR and DSR extractors and performs re-disentanglement. We assume that when DIR and DSR extractors own well disentangled ability, inputting F_{di} (DIR) to the DIR extractor should output more domain-invariant information. And the output of inputting F_{ds} (DSR) to the DSR extractor should contain more domain-specific information. We design a contrastive loss to attain this assumption. Next, to further enhance the generalization ability, we explore employing self-distillation [21, 50] to distill the knowledge of the current detector. Concretely, the disentangled DIR is taken as teacher representations. By narrowing the distance between the DIR and the feature maps generated by middle layers of the backbone network, the feature maps could be promoted to contain more domain-invariant information, which is beneficial for improving generalization ability and detection performance. In the experiments, our method is evaluated on urban-scene object detection. Extensive experiments on five scenes with different weather conditions demonstrate the superiorities of our method.

The contributions are summarized as follows:

(1) To improve the generalization ability of object detectors, we propose a realistic yet challenging task, i.e., Single-Domain Generalized Object Detection (Single-DGOD), which aims to generalize a detector to multiple unseen target domains with only one source domain for training.

(2) To tackle Single-DGOD, we employ a method of cyclic-disentangled self-distillation to disentangle domain-invariant representations without the reliance on domain-related annotations (e.g., domain labels).

(3) We build a Diverse-Weather Dataset of urban-scene object detection to verify our method, which consists of five scenes with different weather conditions, i.e., daytime-sunny, night-sunny, dusk-rainy, night-rainy, and daytime-foggy. The significant performance gain over baseline methods shows the effectiveness of the proposed method.

2. Related Work

Domain Adaptive Object Detection. To alleviate the domain-shift impact, most existing methods [26, 28] try to align the feature-level distributions between the source and target domain. Particularly, Chen et al. [8] proposed to use the adversarial mechanism [14] to align both feature- and instance-level distributions. Based on this work, Saito et al. [34] proposed to align the local and global feature distributions to improve the generalization ability of detectors. Besides, some methods [3, 22] employed generative adversarial networks [54] to translate the style of the source domain to that of the target domain, which directly reduces the domain gap. Meanwhile, there exist some methods [4, 42] that explore to extract instance-invariant features to enhance the generalization ability. Though these methods have been shown to be effective, during training, they usually require to access both the source and target domain data. Therefore, these methods can not be used to solve Single-DGOD.

Single Domain Generalization. Recently, a new task of single domain generalization [38] is proposed, which aims to generalize a model trained on one source domain to many unseen target domains. Most existing methods employ data augmentation and feature normalization to solve this task. Particularly, Volpi et al. [38] and Qiao et al. [32] explored to utilize the adversarial mechanism to solve this task, which

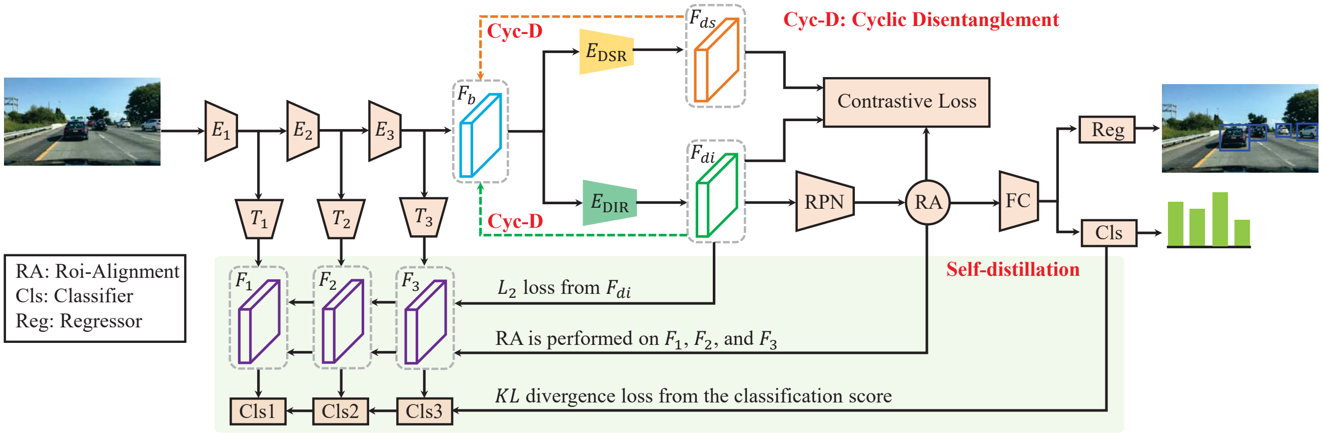

Figure 3. Illustration of Cyclic-Disentangled Self-Distillation. 'RPN' and 'FC' separately indicate region proposal network and fully-connected layers. L_2 is the L2-norm operation. This method mainly consists of cyclic disentanglement and self-distillation. Through a contrastive loss, cyclic disentanglement aims to disentangle DIR (F_{di}) from DSR (F_{ds}) without using domain-related annotations. Next, taking F_{di} as the teacher, self-distillation is employed to promote the generated representations (i.e., F_1, F_2, and F_3) to contain much domain-invariant information, which is beneficial for further improving the generalization ability.

is helpful for encouraging large domain transportation in the input space. Wang et al. [41] explored to improve the generalization by alternating diverse sample generation and discriminative style-invariant representation learning. Fan et al. [13] proposed a generic normalization approach, i.e., adaptive standardization and rescaling normalization, to improve the generalization. Though these methods are effective for image classification, as object detection contains localization and classification, these methods can not be directly applied to Single-DGOD.

Self-Distillation. The purpose of self-distillation mechanism [21, 47, 49] is to train an effective student network by utilizing its own knowledge without a teacher network. Data augmentation based approach and auxiliary network based approach are two commonly used operations to perform self-distillation [20]. Specifically, the data augmentation based self-distillation [47] introduces a prediction consistency loss between the original data and its augmented data. The auxiliary network based method [21] employs additional branches in the middle layer of the model. And the additional branches are used to make similar outputs as the predictions of the network. Though these methods have been shown to be effective, they are rarely used to improve the generalization ability of object detectors. In this paper, based on the disentangled DIR, we utilize self-distillation to further enhance the generalization ability, which has been demonstrated to be effective in the experiments.

3. Cyclic-Disentangled Self-Distillation

As shown in Fig. 3, to tackle Single-DGOD, we propose a method of cyclic-disentangled self-distillation to disentangle DIR for object detection, which is conductive to improving the generalization ability for unseen domains.

3.1. Cyclic Disentanglement

Recently, many methods [31, 42, 44] try to use domain-related annotations, e.g., domain labels, to disentangle DIR, which could not be used for Single-DGOD as there is only one source domain for training. To disentangle DIR without the reliance on domain-related annotations, we propose a module of cyclic disentanglement.

Concretely, we adopt widely used Faster R-CNN [33] as the base detection model. Firstly, a backbone network, e.g., ResNet101 [17], is divided into three sections (i.e., E_1, E_2, and E_3) according to its depth and original structure, whose goal is to perform self-distillation. Given an input image, we use E_1, E_2, and E_3 to obtain feature map $F_b \in \mathbb{R}^{w \times h \times c}$, where w, h, and c respectively denote width, height, and the number of channels. Then, two extractors, i.e., E_{DIR} and E_{DSR}, are designed to separately extract domain-invariant features $F_{di} \in \mathbb{R}^{w \times h \times c}$ and domain-specific features $F_{ds} \in \mathbb{R}^{w \times h \times c}$. The processes are shown as follows:

$$F_{di} = E_{\text{DIR}}(F_b), \quad F_{ds} = E_{\text{DSR}}(F_b), \tag{1}$$

where E_{DIR} and E_{DSR} consist of multiple convolutional layers. The RPN module is performed on F_{di} to extract a set of object proposals O. After the Roi-Alignment operation, the output is $P \in \mathbb{R}^{n \times s \times s \times c}$, where n and s separately indicate the number of proposals and the size of proposals. Next, as shown in the right part of Fig. 2, in the backward direction, E_{DIR} and E_{DSR} separately take both F_{di} and F_{ds} as the input to perform re-disentanglement.

$$\begin{aligned} F_{i2i} &= E_{\text{DIR}}(F_{di}), \quad F_{i2s} = E_{\text{DSR}}(F_{di}), \\ F_{s2i} &= E_{\text{DIR}}(F_{ds}), \quad F_{s2s} = E_{\text{DSR}}(F_{ds}). \end{aligned} \tag{2}$$

We assume that when F_{di} contains abundant domain-invariant information, compared with F_{i2s}, F_{i2i} should involve more domain-invariant information that is related to F_{di}. Meanwhile, when F_{ds} contains sufficient domain-specific information, compared with the output F_{s2i}, F_{s2s} should involve more domain-specific information that is related to F_{ds}. We define a global- and instance-level contrastive loss [5, 15] to attain the assumption. Specifically, let $\mathrm{sim}(a, b)$ denote the average of the cosine similarity between all corresponding elements of feature map a and b. The global-level contrastive loss is calculated as follows:

$$\begin{aligned}\mathcal{L}_{gc} = -\,(\log \frac{\exp(\mathrm{sim}(F_{di}, F_{i2i})/\tau)}{\sum_{j=0}^{1} \exp(\mathrm{sim}(F_{di}, G[j])/\tau)} \\ + \log \frac{\exp(\mathrm{sim}(F_{ds}, F_{s2s})/\tau)}{\sum_{j=0}^{1} \exp(\mathrm{sim}(F_{ds}, D[j])/\tau)}),\end{aligned} \tag{3}$$

where $G = [F_{i2i}, F_{i2s}]$, $D = [F_{s2s}, F_{s2i}]$. τ is a hyper-parameter. In the experiments, τ is set to 1.0. By optimizing $\mathcal{L}_{gc}$, it is helpful for enlarging the gap between F_{i2i} and F_{i2s}, between F_{s2i} and F_{s2s}, and between F_{di} and F_{ds}. Meanwhile, this loss is beneficial for promoting F_{di} and F_{ds} to separately contain domain-invariant and domain-specific information, which improves the disentangled ability.

Next, to further facilitate E_{DIR} to own the ability of disentangling DIR, we define an instance-level contrastive loss. Based on the object proposals O from F_{di}, the Roi-Alignment operation is separately performed on F_{i2i} and F_{i2s} to obtain the output $P_{i2i} \in \mathbb{R}^{n\times s\times s\times c}$ and $P_{i2s} \in \mathbb{R}^{n\times s\times s\times c}$. This loss is computed as follows:

$$\mathcal{L}_{ic} = -\log \frac{\exp(\mathrm{sim}(P, P_{i2i})/\tau)}{\sum_{j=0}^{1} \exp(\mathrm{sim}(P, Q[j])/\tau)}, \tag{4}$$

where $Q = [P_{i2i}, P_{i2s}]$. By minimizing $\mathcal{L}_{ic}$, in addition to enlarging the gap between the instance-level features from F_{i2i} and F_{i2s}, it is beneficial for promoting the features P extracted from F_{di} to involve domain-invariant information, which further enhances the generalization ability and improves detection accuracy. Finally, the sum of the global-level and instance-level contrastive loss is taken as the training loss of this module, i.e., $\mathcal{L}_{cd} = \mathcal{L}_{gc} + \mathcal{L}_{ic}$.

3.2. DIR-based Self-Distillation

By the cyclic-disentangled module, the disentangled F_{di} could be facilitated to involve more domain-invariant information. Next, taking F_{di} as the teacher representations, we explore employing self-distillation mechanism to promote the feature maps extracted by the backbone network to own plentiful domain-invariant information, which further improves the generalization ability of object detectors.

Given an input image, we respectively extract the feature map F_{e1} from E_1, F_{e2} from E_2, and F_{e3} from E_3, where the size and channel number of F_{e1} and F_{e2} are different from F_{di}. Instead, the size and channel number of F_{e3} are the same as F_{di}. Then, we define three networks consisting of multiple convolutional layers, i.e., T_1, T_2, and T_3, to perform a transformation on F_{e1}, F_{e2}, and F_{e3}. The output is $F_1 \in \mathbb{R}^{w\times h\times u}$, $F_2 \in \mathbb{R}^{w\times h\times v}$, and $F_3 \in \mathbb{R}^{w\times h\times c}$, where u and v are the number of channels. Next, we respectively define a feature- and classification-level constraint to promote the extracted feature maps to distill the knowledge from F_{di}. For the feature-level constraint, we separately employ a convolutional layer $\Phi_1 \in \mathbb{R}^{1\times 1\times u\times c}$, $\Phi_2 \in \mathbb{R}^{1\times 1\times v\times c}$, and $\Phi_3 \in \mathbb{R}^{1\times 1\times c\times c}$ to project F_1, F_2, and F_3 into the teacher space. The constraint is calculated as follows:

$$\begin{aligned}\mathcal{L}_{fc} = dist(\Phi_1(F_1), F_{di}) + dist(\Phi_2(F_2), F_{di}) \\ + dist(\Phi_3(F_3), F_{di}),\end{aligned} \tag{5}$$

where $dist(\cdot,\cdot)$ denotes a distance function, e.g., L2-norm. By narrowing the distance, the teacher representation F_{di} could guide the representations extracted by the backbone network to learn domain-invariant information, which enhances the generalization ability of object detectors.

For the classification-level constraint, as shown in Fig. 3, based on the proposals O, the Roi-Alignment operation is separately performed on F_1, F_2, and F_3 to obtain the output $P_1 \in \mathbb{R}^{n\times s\times s\times u}$, $P_2 \in \mathbb{R}^{n\times s\times s\times v}$, and $P_3 \in \mathbb{R}^{n\times s\times s\times c}$. Then, we define three classifiers, which take P_1, P_2, and P_3 as the input and output the predicted probability y_1, y_2, and y_3. Next, the Kullback-Leibler (KL) divergence is used to make the predicted probability approximate the classification probability y that is computed based on P from F_{di}.

$$\mathcal{L}_{cc} = KL(y, y_1) + KL(y, y_2) + KL(y, y_3). \tag{6}$$

Minimizing the loss $\mathcal{L}_{cc}$ could further promote F_1, F_2, and F_3 to distill the category-related knowledge from F_{di}, which is helpful for improving the detection accuracy. Finally, the sum of the feature-level and classification-level constraint is taken as the training loss of the self-distillation module, i.e., $\mathcal{L}_{sd} = \mathcal{L}_{fc} + \mathcal{L}_{cc}$.

During training, our method is trained end-to-end. The joint training loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{loc} + \lambda(\mathcal{L}_{cd} + \mathcal{L}_{sd}), \tag{7}$$

where $\mathcal{L}_{rpn}$ is the RPN loss to distinguish foreground from background and refine bounding-box anchors. $\mathcal{L}_{cls}$ and $\mathcal{L}_{loc}$ separately denote the classification loss and bounding-box regression loss. λ is a hyper-parameter, which is set to 0.01 in the experiments. During inference, we take the predictions calculated based on F_{di} as the detection results.

3.3. Further Discussion

In this section, we will further discuss the fundamental issue that our method can improve the generalization ability.

For Single-DGOD, disentangling DIR is a feasible solution to generalize an object detector trained on one source domain to multiple unseen target domains. However, most existing methods [31, 42, 44] try to utilize domain-related annotations (e.g., domain labels) to attain disentanglement. When there is no domain-related annotations available, how to extract DIR remains under-explored.

To this end, we present a cyclic-disentangled module to extract DIR. By the cyclic operation, the gap between the domain-invariant features (e.g., F_{i2i} and F_{s2i}) and domain-specific features (e.g., F_{i2s} and F_{s2s}) could be enlarged, which promotes E_{DIR} and E_{DSR} to own the disentangled ability. Meanwhile, during the cyclic processes, since the parameters in E_{DIR} and E_{DSR} are shared, the two extractors could facilitate the disentangled F_{di} and F_{ds} to keep separable. Next, by minimizing the two contrastive losses (Eq. (3) and (4)), the relevance between F_{di} and F_{i2i} and between F_{ds} and F_{s2s} could be strengthened, which is helpful for guiding F_{di} and F_{ds} to respectively involve domain-invariant and domain-specific information. Finally, taking F_{di} as the teacher representation, using self-distillation further improves the generalization ability.

4. Experiments

To evaluate the generalization capability of our method, we conduct experiments on the urban-scene object detection with five weather conditions. Data and code are available at https://github.com/AmingWu/Single-DGOD.

4.1. Experimental Setup

In the real scenario, e.g., autonomous driving, the data in the daytime-sunny scene is easily collected and labeled. Thus, we train our model on the daytime-sunny dataset and show its performance on other datasets (e.g., night-sunny, dusk-rainy, night-rainy, and daytime-foggy) to measure the generalization ability on unseen domains. For all the quantitative experiments, mean average precisions (mAP) is used as the evaluation metric.

Datasets. Based on multiple existing datasets, we build an urban-scene detection dataset (as shown in Fig. 1) that consists of five different weather conditions. Particularly, for the daytime-sunny scene, we select 27,708 daytime-sunny images from the Barkeley Deep Drive 100k (BDD-100k) dataset [48] that contains 100,000 driving videos. And 19,395 images are chosen for training. 8,313 images are utilized for testing. The model trained on the daytime-sunny scene is used to perform evaluation on other unseen domains. For the night-sunny scene, we also select 26,158 images from the BDD-100k dataset. For the dusk-rainy and night-rainy scenes, we utilize the recently proposed dataset [44], which renders the rainy images that are from the BDD-100k dataset to enlarge the gap between the source and target domain. The dusk-rainy and night-rainy scenes separately include 3,501 and 2,494 images. Finally, for the daytime-foggy scene, we collect foggy images from FoggyCityscapes [35] and Adverse-Weather [16] datasets. This scene contains 3,775 images. We can see the built dataset includes multiple challenging scenes, which is helpful for evaluating the generalization ability of object detectors. Besides, the BDD-100k and FoggyCityscapes datasets separately contain ten and eight categories. Here, we choose seven commonly used categories, which do not include the category of light, sign, and train.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster R-CNN	63.4	42.9	53.4	49.4	39.8	48.1	60.8	51.1
SW [30]	62.3	42.9	53.3	49.9	39.2	46.2	60.6	50.6
IBN-Net [29]	63.6	40.7	53.2	45.9	38.6	45.3	60.7	49.7
IterNorm [19]	58.4	34.2	42.4	44.1	31.6	40.8	55.5	43.9
ISW [9]	62.9	44.6	53.5	49.2	39.9	48.3	60.9	51.3
Ours	**68.8**	**50.9**	**53.9**	**56.2**	**41.8**	**52.4**	**68.7**	**56.1**

Table 1. Results (%) on the daytime-sunny scene. Here, 'motor' indicates the motorcycle category.

Implementation Details. We use Faster R-CNN [33] as the base detector. RseNet101 [17] is taken as the backbone. We use the weights pre-trained on ImageNet [10] in initialization. We separately design a network consisting of three convolutional layers as the DIR extractor E_{DIR} and DSR extractor E_{DSR}. Meanwhile, T_1, T_2, and T_3 all consist of three convolutional layers with BatchNorm operation. All parameters in these networks are randomly initialized. During training, our model is trained using SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001. The learning rate is set to 0.001. And the batch size is set to 4. More details can be seen in the supplementary material.

4.2. Performance Analysis of Single-DGOD

We compare our method with four baseline methods, i.e., SW [30], IBN-Net [29], IterNorm [19], and ISW [9]. These approaches all use the idea of feature normalization to improve the generalization ability of models. In order to compare with these normalization methods fairly, we directly plug these methods into Faster R-CNN [33].

Results on Daytime-Sunny Scene. Table 1 shows the results of the daytime-sunny scene. We can see that our method outperforms the compared methods significantly. This shows that when the training and test sets are from the same domain, the proposed method could improve the performance of the current domain. Besides, we can also see that the feature normalization methods [9, 19, 29, 30] do not significantly improve the performance of Faster R-CNN [33]. The reason may be that the discrimination ability is affected by feature normalization, which weakens the detection performance. This further shows that compared with feature normalization methods, our method is beneficial for enhancing the discrimination ability.

Results on Night-Sunny Scene. Table 2 shows the detection results on the night-sunny scene. Here, we directly use the model trained on the daytime-sunny scene to per-

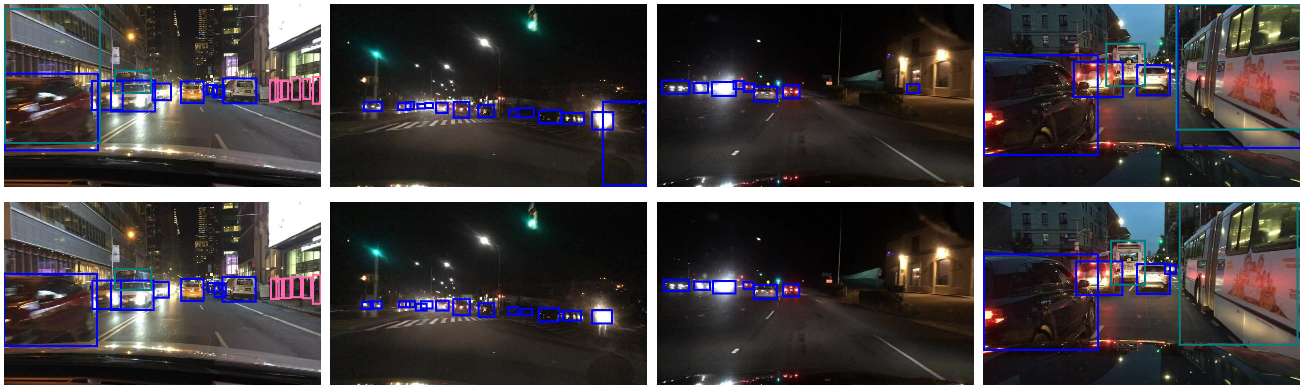

Figure 4. Detection results on the night-sunny scene. The first and second rows separately show the results from Faster R-CNN [33] and our method. We can see that compared with Faster R-CNN [33], our method could detect the objects accurately, e.g., the car, person, bus.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster R-CNN	37.7	30.6	49.5	15.4	31.5	28.6	40.8	33.5
SW [30]	38.7	29.2	49.8	16.6	31.5	28.0	40.2	33.4
IBN-Net [29]	37.8	27.3	49.6	15.1	29.2	27.1	38.9	32.1
IterNorm [19]	38.5	23.5	38.9	15.8	26.6	25.9	38.1	29.6
ISW [9]	38.5	28.5	49.6	15.4	31.9	27.5	41.3	33.2
Ours	**40.6**	**35.1**	**50.7**	**19.7**	**34.7**	**32.1**	**43.4**	**36.6**

Table 2. Results (%) on the night-sunny scene.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster R-CNN	36.8	15.8	50.1	12.8	18.9	12.4	39.5	26.6
SW [30]	35.2	16.7	50.1	10.4	**20.1**	13.0	38.8	26.3
IBN-Net [29]	37.0	14.8	50.3	11.4	17.3	13.3	38.4	26.1
IterNorm [19]	32.9	14.1	38.9	11.0	15.5	11.6	35.7	22.8
ISW [9]	34.7	16.0	50.0	11.1	17.8	12.6	38.8	25.9
Ours	**37.1**	**19.6**	**50.9**	**13.4**	19.7	**16.3**	**40.7**	**28.2**

Table 3. Results (%) on the dusk-rainy scene.

form the evaluation. We can see that when the training and test data are from different domains, the model's performance degrades significantly. This indicates that improving the generalization capability of object detectors is meaningful. Compared with Faster R-CNN [33], our method is 3.1% higher than its performance, which indicates the proposed cyclic disentanglement is capable of disentangling DIR. By taking the DIR as the teacher, self-distillation further improves the generalization ability. Besides, we can see that feature normalization methods [9, 19, 29, 30] do not lead to the performance improvement. In addition to the factor of the weak discrimination, the large gap between the daytime-sunny and night-sunny scenes may be another reason that weakens the performance. This further shows that Single-DGOD is filled with challenges. Enhancing the generalization ability is an effective solution.

In Fig. 4, we show some detection examples in the night-sunny scene. Due to the impact of low light, the detections in the night scene are very challenging. Compared with the results from Faster R-CNN [33], our method could detect the objects in the night images accurately, which further indicates the effectiveness of our method.

Results on Dusk-Rainy and Night-Rainy Scenes. Table 3 and 4 separately show the detection performance on the dusk-rainy and night-rainy scenes. We can see that when the scenario is full of challenges, e.g., the rainy weather and low light, the detection performance degrades significantly. This shows that the adverse weather enlarges the gap between the training and test set, which weakens the detection performance. Compared with Faster R-CNN [33], our method is 1.6% and 2.1% higher than its performance. This indicates that extracting domain-invariant features is indeed helpful for alleviating the domain-shift impact on object detection. Besides, the performance of our method still outperforms the feature normalization methods [9, 19, 29, 30], which further demonstrates that our method is beneficial for enhancing the generalization capability of object detectors.

Results on Daytime-Foggy Scene. Table 5 shows the results of the daytime-foggy scenario. Due to the foggy impact, the images become very obscure, which affects the accuracy of object localization and classification severely. We can see that our method outperforms the compared methods [9, 19, 29, 30, 33] significantly. This demonstrates that our method could well extract domain-invariant features that lead to the performance improvement. Fig. 5 shows the detection results. We can see that our method accurately detects objects in the foggy images, which demonstrates that our method owns the generalization ability.

4.3. Ablation Analysis

In this section, we use the model trained on the daytime-sunny scene to make an ablation analysis of our method. Here, the daytime-sunny and night-sunny scenes are used to make evaluations. Table 6 shows the results. We can see that the performance of only using the forward step to perform disentanglement (as shown in the left part of Fig. 2) is very weak. This indicates that when there is no domain-related annotations available, only using the forward step

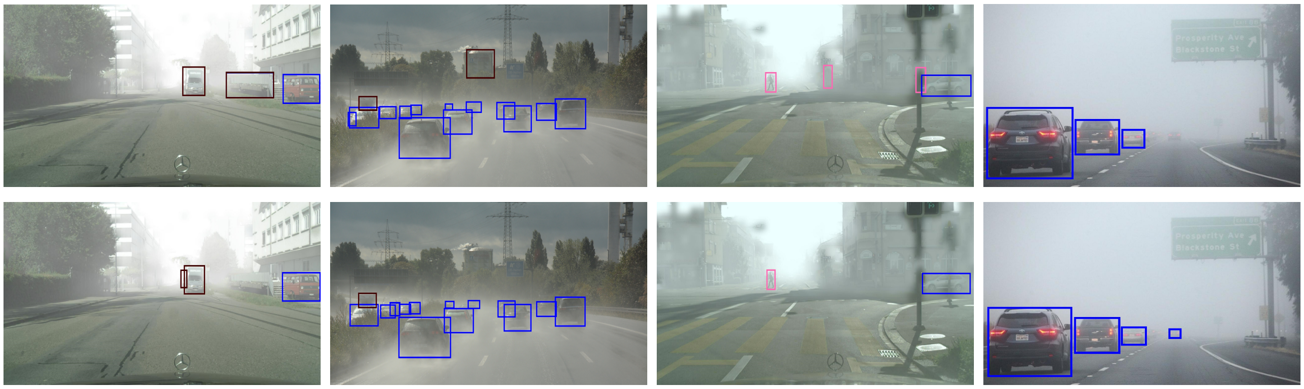

Figure 5. Detection results on the daytime-foggy scene. Compared with the results from Faster R-CNN [33] (as shown in the first row), our method (as shown in the second row) accurately detects objects in the foggy images, e.g., the car, person, truck.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster R-CNN	22.6	11.5	27.7	0.4	10.0	10.5	19.0	14.5
SW [30]	22.3	7.8	27.6	0.2	10.3	10.0	17.7	13.7
IBN-Net [29]	**24.6**	10.0	28.4	0.9	8.3	9.8	18.1	14.3
IterNorm [19]	21.4	6.7	22.0	0.9	9.1	10.6	17.6	12.6
ISW [9]	22.5	11.4	26.9	0.4	9.9	9.8	17.5	14.1
Ours	24.4	**11.6**	**29.5**	**9.8**	**10.5**	**11.4**	**19.2**	**16.6**

Table 4. Results (%) on the night-rainy scene.

Method	bus	bike	car	motor	person	rider	truck	mAP
Faster R-CNN	30.7	26.7	**49.7**	26.2	30.9	35.5	23.2	31.9
SW [30]	30.6	26.2	44.6	25.1	30.7	34.6	23.6	30.8
IBN-Net [29]	29.9	26.1	44.5	24.4	26.2	33.5	22.4	29.6
IterNorm [19]	29.7	21.8	42.4	24.4	26.0	33.3	21.6	28.4
ISW [9]	29.5	26.4	49.2	27.9	30.7	34.8	24.0	31.8
Ours	**32.9**	**28.0**	48.8	**29.8**	**32.5**	**38.2**	**24.1**	**33.5**

Table 5. Results (%) on the daytime-foggy scene.

does not disentangle domain-invariant features. Meanwhile, we can also see that only using the self-distillation mechanism without domain-invariant features does not obtain better detection results than Faster R-CNN [33]. This indicates that for self-distillation, taking the domain-invariant features as the teacher representations is beneficial for improving the generalization ability of object detectors. Finally, we can see that utilizing the cyclic disentangled operation improves the detection performance, which shows that employing the cyclic operation and contrastive losses (as shown in Eq. (3) and (4)) is indeed helpful for extracting domain-invariant features. Next, with the help of the self-distillation mechanism, the model's generalization ability could be further enhanced, which leads to the performance improvement for Single-DGOD.

Analysis of contrastive losses in Eq. (3) and Eq. (4). To disentangle domain-invariant features without the reliance on domain-related annotations, we propose a cyclic-disentangled module and design two contrastive losses to enhance the disentangled ability. Particularly, $\mathcal{L}_{gc}$ in Eq. (3) and $\mathcal{L}_{ic}$ in Eq. (4) separately aim to enlarge the gap between the domain-invariant and domain-specific features from the global-level and instance-level perspectives. Here, we make an ablation analysis. Based on the night-sunny scene, for the case of only using $\mathcal{L}_{gc}$ and self-distillation, the performance is 34.8%. Meanwhile, the performance of only using $\mathcal{L}_{ic}$ and self-distillation is 35.2%. This shows that utilizing the two contrastive losses is helpful for improving disentanglement. And combining with self-distillation could further enhance the generalization ability.

Analysis of the losses in Eq. (5) and Eq. (6). To further enhance the generalization ability, we take the disentangled domain-invariant features as the teacher and design a self-distillation module with two losses. Specifically, the loss in Eq. (5) aims to narrow the feature-level distance between the teacher and the extracted middle-layer representations, which is beneficial for promoting the middle-layer representations to involve domain-invariant information. And the goal of the loss in Eq. (6) is to distill the category-related knowledge from the teacher, which is helpful for improving the detection accuracy. Here, based on the night-sunny scene and the output of the cyclic-disentangled module, for the case of only using $\mathcal{L}_{fc}$ in Eq. (5), the performance is 35.0%. Meanwhile, the performance of only using $\mathcal{L}_{cc}$ in Eq. (6) is 35.4%. This indicates that using the two losses is helpful for attaining self-distillation, which leads to the generalization enhancement of the object detector.

4.4. Comparison with Domain Adaptation Methods

To further evaluate the generalization ability, based on the night-sunny scene, we compare our method with some domain adaptation methods [3,44,46] that require to access the target domain during training. Table 7 shows the results. Compared with the results of only using the source domain data (i.e., Source Only), these domain adaptation methods do not obtain superior performance. The reason may be that the gap between the source and target domain is very large, which makes the adaptation more difficult. We can see that

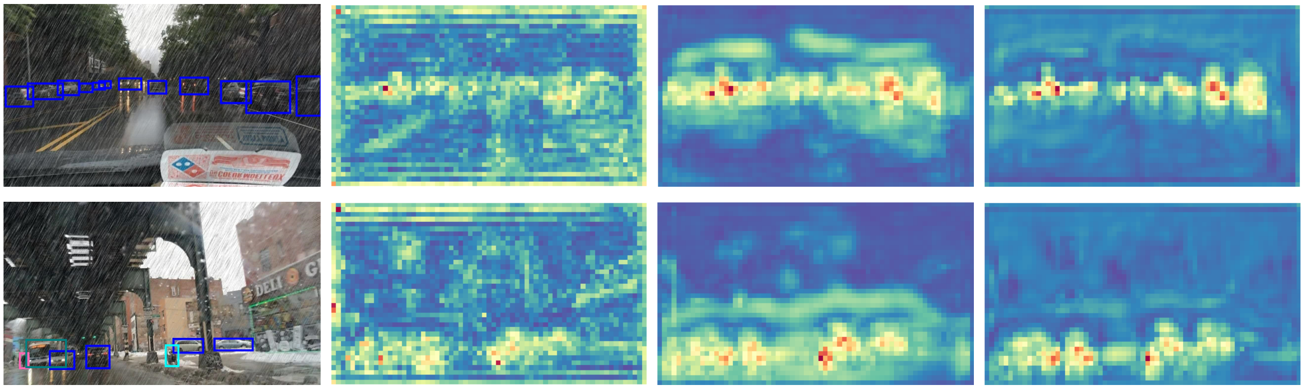

Figure 6. Visualization analysis of our method based on the dusk-rainy scene. The first column is the detection results of our method. The second, third, and fourth columns separately indicate the features F_b, F_{di}, and F_{i2i} (as shown in Eq. (1) and (2)). For each feature map, the channels corresponding to the maximum value are selected for visualization.

Method	DT	Cyc-D	SD	Daytime-Sunny	Night-Sunny
CDSD	✓			48.7%	30.4%
CDSD	✓		✓	50.4%	31.3%
CDSD			✓	51.0%	33.3%
CDSD		✓		52.3%	34.2%
CDSD		✓	✓	**56.1%**	**36.6%**

Table 6. Ablation analysis of our proposed cyclic-disentangled self-distillation (CDSD). Here, we use mAP as the metric. 'DT' indicates that we only use a forward step to perform disentanglement (as shown in the left part of Fig. 2). 'Cyc-D' denotes the cyclic disentanglement. 'SD' is the self-distillation. The model trained on daytime-sunny scene is used to make the analysis.

though there is no target domain data available, our method outperforms these adaptation methods significantly. Particularly, for each category, the performance of our method is the best. This not only demonstrates that disentangling DIR is helpful for enhancing the generalization capability but also indicates that the proposed cyclic-disentangled self-distillation could extract DIR effectively.

4.5. Visualization Analysis

In Fig. 6, we make a visualization analysis. The second column is F_b that is used to perform disentanglement. The third column indicates the disentangled feature map F_{di} (as shown in Eq. (1)). And the fourth column denotes the cyclical output F_{i2i} that is based on F_{di} (as shown in Eq. (2)). We can see that compared with F_b, F_{di} contains much more object-related information and less background that mainly reflects the domain-related information. Moreover, through the cyclic operation (as shown in Eq. (2)), compared with F_{di}, the background information in the output F_{i2i} is further reduced. Finally, we can see that our method accurately detects objects in the rainy images, e.g., the car, person, motorcycle. This indicates that without the reliance on domain-related annotations, our method could extract domain-invariant features containing intrinsical object characteristics, which is beneficial for Single-DGOD.

Method	target	bus	bike	car	motor	person	rider	truck	mAP
Source Only	–	37.7	30.6	49.5	15.4	31.5	28.6	40.8	33.5
DAF [8]	✓	36.2	29.1	49.3	16.0	33.1	29.3	40.2	33.3
CT [51]	✓	34.1	22.1	46.4	12.8	26.5	19.8	31.5	27.6
SCL [36]	✓	27.1	18.6	45.9	11.6	24.0	19.0	32.3	25.5
SW [34]	✓	34.2	23.6	48.0	13.4	26.4	23.7	37.5	29.5
ICCR [46]	✓	36.1	23.2	48.9	15.5	29.1	23.8	39.4	30.9
HTCN [3]	✓	30.5	17.6	44.7	11.0	22.9	20.6	31.3	25.5
VDD [44]	✓	35.4	29.6	49.8	14.5	31.3	28.0	39.9	32.6
Ours	–	**40.6**	**35.1**	**50.7**	**19.7**	**34.7**	**32.1**	**43.4**	**36.6**

Table 7. Results (%) on the adaptation from the daytime-sunny scene to the night-sunny scene. 'target' indicates that during training, the model needs to access the target domain data (i.e., the night-sunny data). '–' denotes that during training, we do not use the target domain data. Here, the released codes of the compared methods are directly run to obtain the results.

5. Conclusion

In this paper, we frame a new paradigm of object detection, Single-DGOD, which aims to generalize object detectors to multiple unseen target domains with only one source domain for training. Towards Single-DGOD, we focus on extracting DIR without the reliance on domain-related annotations, and a method of cyclic-disentangled self-distillation is proposed accordingly. Firstly, we design a cyclic-disentangled module to disentangle DIR. Then, taking the DIR as the teacher representation, a self-distillation module is used to further enhance the generalization capability of the object detector. Experimental results and visualization analyses show the superiorities of our method.

Acknowledgement. Our work was supported in part by the National Key R&D Program of China under Grant 2017YFE0104100, in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, Grant 62071361, and Grant 62102293, in part by Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03, and in part by the Fundamental Research Funds for the Central Universities ZDRC2102.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *NeurIPS*, 31:998–1008, 2018. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8869–8878, 2020. 2, 7, 8

[4] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *CVPR*, pages 12576–12585, 2021. 1, 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 4

[6] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 1

[7] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, pages 1081–1090, 2019. 2

[8] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2, 8

[9] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne Taery Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, pages 11580–11590, 2021. 1, 5, 6, 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5

[11] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 1

[12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *NeurIPS*, 32:6450–6461, 2019. 1

[13] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *CVPR*, pages 8208–8217, 2021. 3

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

[15] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 4

[16] M. Hassaballah, Mourad A. Kenk, Khan Muhammad, and Shervin Minaee. Vehicle detection and tracking in adverse weather using a deep learning framework. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4230–4242, 2021. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 3, 5

[18] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. *ICCV*, 2019. 2

[19] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *CVPR*, pages 4874–4883, 2019. 5, 6, 7

[20] Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, and Il-Chul Moon. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *CVPR*, pages 10664–10673, 2021. 3

[21] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation: A simple way for better generalization. *arXiv preprint arXiv:2006.12000*, 2020. 2, 3

[22] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, pages 12456–12465, 2019. 2

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[24] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. In *ICCV*, October 2019. 1

[25] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018. 1

[26] Wanyi Li, Fuyu Li, Yongkang Luo, Peng Wang, et al. Deep domain adaptive object detection: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1808–1813. IEEE, 2020. 2

[27] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *CVPR*, pages 12406–12415, 2020. 1

[28] Poojan Oza, Vishwanath A Sindagi, Vibashan VS, and Vishal M Patel. Unsupervised domain adaption of object detectors: A survey. *arXiv preprint arXiv:2105.13502*, 2021. 2

[29] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, pages 484–500, 2018. 5, 6, 7

[30] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *ICCV*, pages 1863–1871, 2019. 5, 6, 7

[31] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *ICML*, 2019. 2, 3, 5

[32] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, pages 12556–12565, 2020. 2

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 3, 5, 6, 7

[34] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 2, 8

[35] 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. 5

[36] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 1, 8

[37] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 1

[38] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, volume 31, pages 5334–5344, 2018. 2

[39] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021. 1

[40] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *CVPR*, pages 7289–7298, 2019. 1

[41] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. *arXiv preprint arXiv:2108.11726*, 2021. 3

[42] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 5

[43] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9567–9576, 2021. 1

[44] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. *ICCV*, 2021. 2, 3, 5, 7, 8

[45] Aming Wu, Suqi Zhao, Cheng Deng, and Wei Liu. Generalized and discriminative few-shot object detection via svd-dictionary enhancement. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[46] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11724–11733, 2020. 1, 7, 8

[47] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5565–5572, 2019. 3

[48] 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 *CVPR*, pages 2636–2645, 2020. 5

[49] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *CVPR*, pages 13876–13885, 2020. 3

[50] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, pages 3712–3721, 2019. 2

[51] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, pages 86–102. Springer, 2020. 8

[52] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *NeurIPS*, 33, 2020. 1

[53] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021. 1

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, Oct 2017. 2