

# Unbiased Faster R-CNN for Single-source Domain Generalized Object Detection

Yajing Liu<sup>1,2,3</sup>, Shijun Zhou<sup>1,2,3</sup>, Xiyao Liu<sup>1,2</sup>, Chunhui Hao<sup>1,2</sup>, Baojie Fan<sup>4</sup>, Jiandong Tian<sup>1,2\*</sup>

<sup>1</sup>State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences

<sup>2</sup>Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences

<sup>3</sup>University of Chinese Academy of Sciences <sup>4</sup>Nanjing University of Posts and Telecommunications

{liuyajing, zhoushijun, liuxiyao, haochunhui, tianjd}@sia.cn, jobfbj@gmail.com

## Abstract

Single-source domain generalization (SDG) for object detection is a challenging yet essential task as the distribution bias of the unseen domain degrades the algorithm performance significantly. However, existing methods attempt to extract domain-invariant features, neglecting that the biased data leads the network to learn biased features that are non-causal and poorly generalizable. To this end, we propose an Unbiased Faster R-CNN (UFR) for generalizable feature learning. Specifically, we formulate SDG in object detection from a causal perspective and construct a Structural Causal Model (SCM) to analyze the data bias and feature bias in the task, which are caused by scene confounders and object attribute confounders. Based on the SCM, we design a Global-Local Transformation module for data augmentation, which effectively simulates domain diversity and mitigates the data bias. Additionally, we introduce a Causal Attention Learning module that incorporates a designed attention invariance loss to learn image-level features that are robust to scene confounders. Moreover, we develop a Causal Prototype Learning module with an explicit instance constraint and an implicit prototype constraint, which further alleviates the negative impact of object attribute confounders. Experimental results on five scenes demonstrate the prominent generalization ability of our method, with an improvement of 3.9% mAP on the Night-Clear scene.

## 1. Introduction

The problem of distribution shift in unseen domains often arises during the deployment of perception systems, leading to a notable decline in the model’s performance [3, 34]. Consequently, domain generalization [6, 19, 37, 50] has emerged as a branch of transfer learning, which aims at generalizing the knowledge from multiple source domains to the unseen target domain. Single-source domain gener-

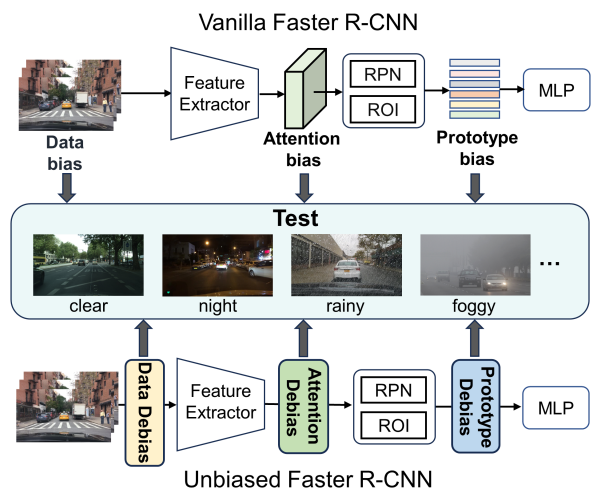


Figure 1. Comparison between vanilla Faster R-CNN (FR) [33] (top) and our proposed Unbiased Faster R-CNN (bottom). For vanilla FR [33], the biased distribution of the input data leads the network to learn biased features that favor the seen environment and are poorly generalizable to unseen test environments. The feature bias can be attributed to the image-level attention bias and object-level prototype bias. Our method mitigates the data bias in the input space and further learns unbiased attention and prototypes in the representation space.

alization (SDG) is a special case of domain generalization where there is only one source domain [31, 43, 51] and it focuses on exploring the robustness of the model under different image corruptions.

The existing two methods for SDG in object detection adopt different generalization strategies on Faster R-CNN [33]. The domain-invariant feature learning-based method [44] explicitly decomposes domain-invariant features and domain-specific features by imposing constraints on the network, without relying on data augmentation techniques. And the data augmentation-based method [38] perturbs the data distribution and increases the diversity of input data to enhance model generalization ability.

However, both of these strategies have certain limita-

\*Corresponding author

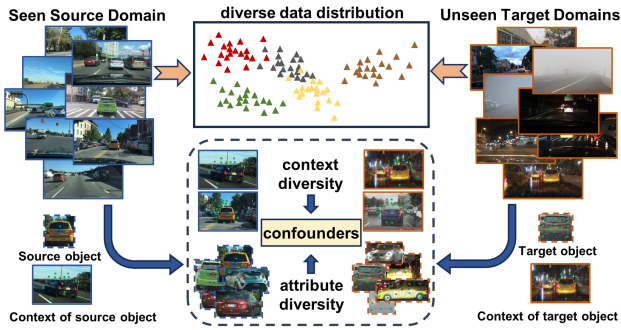


Figure 2. Illustration of highly changeable data distribution, diverse context and object attributes in unseen target domains.

tions. Firstly, it has been proven in [45] that domain-invariant features are inherently domain-dependent and biased, as features that are invariant to current domains may be variant to other domains. This problem can be attributed to the fact that domain-invariant features learned from a biased data distribution are not causal features and cannot adapt well to the unseen target environments. Secondly, the domain generalization methods that use only data augmentation without constraining network features fail to capture causal features from a rich augmented data distribution. This results in the network performing well in the augmented domains but remaining ineffective in the unseen domains, as confirmed in our Ablation Study in Section 5.3. We summarize the reasons for the above two limitations as the existence of data bias in the input space and feature bias in the representation space. The feature bias is further decomposed into image-level attention bias and object-level prototype bias, as illustrated in Fig. 1 and Fig. 2:

(1) **Data bias.** The data distribution of unseen target domains is highly changeable, as shown in Fig. 2. Therefore, learning invariant features solely from single-domain data will exacerbate the statistical dependence [28] between the input data and labels, leading to biased learned results that favor the seen environments.

(2) **Attention bias.** When testing in unseen domains with complex context, as depicted in Fig. 2, the context features may confuse the object features and the network may exhibit attention bias, directing its focus towards the context of the object rather than the object itself. We refer to the scene context as **scene confounders**.

(3) **Prototype bias.** Each category possesses distinctive causal attributes, such as structural information, that are highly informative, as opposed to non-discriminative ones like view and color, which we refer to as **object attribute confounders**, as shown in Fig. 2. Therefore, if the learned category features are not constrained, the network may incorrectly take confounding attribute features that occur frequently in the source domain as category prototypes, which are biased and poorly generalizable.

To this end, we propose an Unbiased Faster R-CNN

(UFR) model, as illustrated in Fig. 1, with a comparison to vanilla Faster R-CNN [33]. The UFR model simulates diverse data distributions through data augmentation to mitigate the data bias and applies constraints on the features to learn scene-level causal attention and object-level causal prototype, thus eliminating the attention bias and prototype bias. Specifically, we define causality in the SDG task as the cause-and-effect relationship among scene, objects, causal features and non-causal features and construct a Structural Causal Model (SCM) to analyze the data bias and feature bias caused by scene confounders and object attribute confounders. Additionally, we propose a Global-Local Transformation (GLT) module, which augments the data as a whole in the frequency domain and leverages the segmentation capability of SAM [15] to augment local objects in the spatial domain. To address scene confounders, we propose a Causal Attention Learning module with an attention invariance loss, which transforms the learning of scene-level causal features into the learning of causal attention. To further learn causal features at the object level, we introduce a Causal Prototype Learning module, which encompasses an explicit constraint on the distribution of instance features and an implicit constraint on the relationship between prototypes. The effectiveness of our method is demonstrated through experiments conducted across five weather conditions. The contributions are summarized as follows:

(1) We are the first to investigate single-domain generalized object detection from a causal perspective and analyze three biases that limit the generalization ability of the detector, including data bias, attention bias and prototype bias.

(2) We construct a Structural Causal Model to analyze the biases caused by two types of confounders and further propose an Unbiased Faster R-CNN with a Global-Local Transformation module, a Causal Attention Learning module and a Causal Prototype Learning module to mitigate the data bias, attention bias and prototype bias respectively.

(3) We evaluate our method on five different weather conditions to demonstrate its effectiveness and superiority and our method achieves a remarkable 3.9% mAP improvement on the Night-Clear scene.

## 2. Related Works

### 2.1. Domain Generalization

Common strategies for addressing domain generalization problem include domain alignment [19, 37, 42, 45], meta learning [2, 6, 40, 50], data augmentation [14, 16, 38], ensemble learning [24, 29, 36], self-supervised learning [8, 10] and disentanglement learning [20, 39, 49]. As a special case of domain generalization, the solution for single-source domain generalization (SDG) can also be categorized into several of the aforementioned strategies. Many prior studies used data augmentation [17, 31, 38, 43] to gen-

erate out-of-distribution samples to extend the distribution of the source domain. For example, Wang et al. [43] proposed a style-complement module to generate diverse stylized images. Vedit et al. [38] proposed a semantic augmentation method for SDG in object detection with the help of a pre-trained CLIP [32]. In addition, some works employed the feature normalization strategy [4, 12, 26, 27] to learn domain-invariant features. For example, Fan et al. [7] proposed an ASR-Norm layer to learn standardization and rescaling statistics.

## 2.2. Causality in Computer Vision

Causal mechanism considers the fact that statistical dependence cannot reliably predict the labels of counterfactual inputs [28]. Thus, exploring causality [1] enables the acquisition of robust knowledge beyond what is supported by the observed data [35]. Recently, causal mechanism has gathered significant attention in computer vision [11, 25, 46, 48]. Many works use causal mechanism to tackle domain generalization [18, 22, 30, 41, 45] and domain adaptation [13, 21, 23, 47] problems. For example, Yue et al. [47] applied the causal mechanism to domain adaptation and used the domain-invariant disentanglement to identify confounders. Lv et al. [22] proposed a representation learning framework for causality-inspired domain generalization. Besides, Liu et al. [21] proposed a Decoupled Unbiased Teacher to solve the source-free domain adaptation problem. Moreover, Xu et al. [45] introduced a causality-inspired data-augmentation strategy and eliminated non-causal factors by a Multi-view Adversarial Discriminator. In this paper, we first apply the causal mechanism to the single-source domain generalized object detection task and propose an Unbiased Faster R-CNN to mine causal features at the image level and object level.

## 3. Structural Causal Model

Consider the data (e.g. images) from the observed environment (source domain) as  $X$  and its target (e.g. detection labels) as  $Y$ , the objective of single-domain generalized object detection is to generalize the model trained with  $(X, Y)$  in the observed environment to unseen environments (target domains). We represent images from a causal perspective and construct a Structural Causal Model (SCM) in Fig. 3 to describe the cause-and-effect relationship in object detection task and attempt to eliminate the data bias and feature bias caused by the scene confounders and object attribute confounders. The rationale behind the SCM is as follows:

(1)  $O \rightarrow X \leftarrow D$  denotes that an image is composed of scene  $D$  and a set of objects  $O$ .  $D$  is scene confounder, and the non-discriminative attributes of  $O$  are object attribute confounders. The GLT module is achieved by perturbing the scene confounders and object attribute confounders.

(2)  $O \rightarrow Z_V \leftarrow D$  denotes that the non-causal features

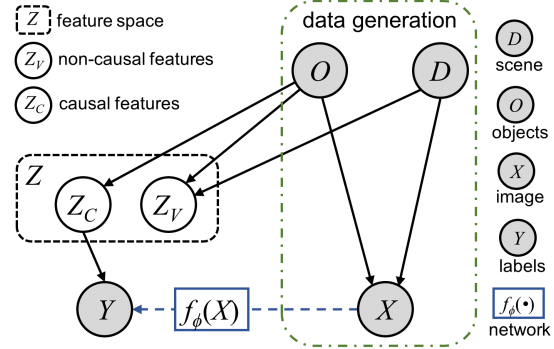


Figure 3. The constructed Structural Causal Model (SCM) for the object detection task. The nodes denote variables, the solid arrows denote the direct causal effect and the dashed arrow indicates that there exists data dependence.

$Z_V$  are derived from two components, including the scene confounders and non-causal object attribute confounders.

(3)  $O \rightarrow Z_C \rightarrow Y$  represents that the causal features  $Z_C$  are determined by the discriminative attributes of objects, such as shape. And the prediction labels  $Y$  are derived from the causal features  $Z_C$ . The complete feature space  $Z$  consists of non-causal features  $Z_V$  and causal features  $Z_C$ .

(4)  $X \rightarrow f_\phi(X) \rightarrow Y$  models the data stream of the object detection network parameterized by  $\phi$ .

Based on the constructed SCM, the ideal feature mapping of the network is  $f_\phi(X) = Z_C$ , which is also formulated as:

$$f_\phi(X^{(a_l, d_l)}) = f_\phi(X^{(a_k, d_k)}), \quad (1)$$

where  $a_l \neq a_k, d_l \neq d_k$  and the superscript  $(a_l, d_l)$  denotes assigning scene and object attributes the distribution  $d_l$  and  $a_l$ , respectively. Eq. (1) indicates that the ideal representation of an image learned by the model is invariant under different data augmentations.

To achieve the above goal, we decompose the objective into image-level attention invariance learning and object-level prototype invariance learning and further propose an Unbiased Faster R-CNN model, as depicted in Fig. 4, and the details of which are illustrated in Section 4.

## 4. Unbiased Faster R-CNN

### 4.1. Global-Local Transformation

The complete process of Global-Local Transformation is depicted in Fig. 5, which consists of Global Transformation (GT) in the frequency domain and Local Transformation (LT) in the spatial domain.

Inspired by the frequency-domain augmentation in [45], the global transformation is formulated as:

$$GT(x) = \mathcal{F}'(G(H(r)) \cdot \mathcal{F}(x)) + (1 - H(r)) \cdot \mathcal{F}(x), \quad (2)$$

where  $\mathcal{F}$  denotes the Fourier Transformation and  $\mathcal{F}'$  is the

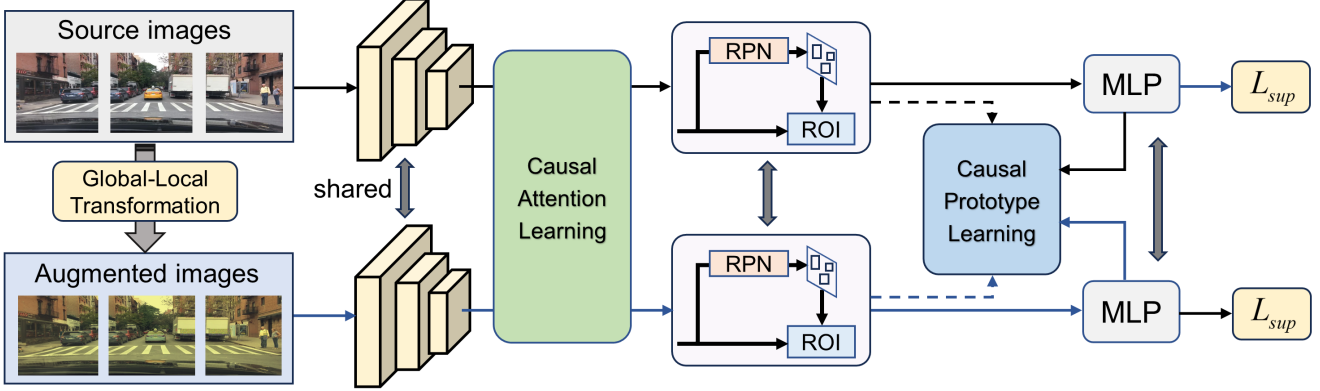


Figure 4. The overall structure of the proposed Unbiased Faster R-CNN. The input source images are augmented through the Global-Local Transformation (GLT) module. Both the original images and augmented images are fed into the network for training. During training, the role of the Causal Attention Learning module is to constrain the network to learn scene-level causal attention and select causal features to feed into the RPN. The purpose of the Causal Prototype Learning module is to constrain the network to learn object-level causal features with the help of an explicit instance constraint (solid arrows) and an implicit prototype constraint (dashed arrows).

inverse one.  $H(r)$  is the band-pass filter with a filter radius of  $r$ .  $G(\cdot)$  is a randomization function according to a Gaussian distribution and  $G(X) = X \cdot (1 + \mathcal{N}(0, 1))$ .

For local transformation, as shown in Fig. 5, we first obtain the object masks with the help of Segment Anything Model (SAM) [15]:

$$\text{SAM}(x, B) = m^O, \quad (3)$$

where  $x$  is the input image,  $B$  is the bounding box set of the objects and  $m^O$  is the set of the obtained masks, where  $O = \{o_1, o_2, \dots, o_{n_O}\}$  and  $n_O$  is the number of objects in the input image. Then we extract object images by :

$$x^{o_k} = x \odot m^{o_k}. \quad (4)$$

Then the local transformation is formulated as:

$$\text{LT}(x) = \mathcal{T}_0(x^{bg}) + \sum_{k=1}^{n_O} \mathcal{T}_k(x^{o_k}), \quad (5)$$

where  $x^{bg}$  is the background image and  $\mathcal{T}_0(\cdot)$  and  $\mathcal{T}_k(\cdot)$  denote the randomly selected augmentation strategies for the background  $x^{bg}$  and the object  $x^{o_k}$ .

The augmented image output by the GLT module is denoted as:

$$\text{GLT}(x) = \alpha \cdot \text{GT}(x) + (1 - \alpha) \cdot \text{LT}(x), \quad (6)$$

where  $\alpha$  is the fusion weight to balance the GT image and the LT image and the fusion strategy also increases the diversity of the augmented images.

## 4.2. Causal Attention Learning

We transform the learning of scene-level causal features into the learning of causal attention, which eliminates the

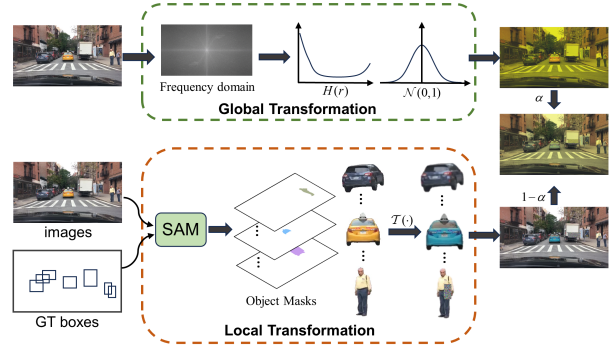


Figure 5. Overview of the Global-Local Transformation (GLT) module. The Global Transformation (GT) performs overall augmentation of an image in the frequency domain. And The Local-Transformation (LT) performs augmentation of local objects obtained by SAM [15] in the spatial domain. The final augmented image is obtained by fusing the GT image with the LT image.

need for explicit decoupling of causal features from non-causal features, instead focusing on selecting causal features based on attention.

Specifically, it is desirable for the network to capture object features precisely in different scene context. Thus, we enforce the feature attention maps of different data distribution images output by the backbone to converge and the attention invariance loss is defined as:

$$\mathcal{L}_{att} = \text{Dice}(g(F_{att}^{(a_0, d_0)}), g(F_{att}^{(a_k, d_k)})), \quad (7)$$

$$F_{att}^{(a, d)} = \sigma(E(x^{(a, d)})), \quad (8)$$

where  $g(\cdot)$  is a pixel-level binarization function,  $F_{att}^{(a, d)}$  is the calculated attention map,  $\sigma(\cdot)$  is the Sigmoid function,  $E$  denotes the backbone,  $x^{(a_0, d_0)}$  is the original input image and  $x^{(a_k, d_k)}$  denotes the randomly augmented image

obtained from GLT. Besides,  $\text{Dice}(\cdot, \cdot)$  denotes the dice loss which is commonly used for measuring the regional similarity between two samples, which is defined as:

$$\text{Dice}(X_1, X_2) = 1 - \frac{2|X_1 \cap X_2| + 1}{|X_1| + |X_2| + 1}, \quad (9)$$

where  $X_1$  and  $X_2$  are two binary maps. The regions with a value of 1 for the maps are salient attention regions, while the regions with a value of 0 are non-salient attention regions.

Then we select causal features according to the attention map for the subsequent Region Proposal Network (RPN) to generate object proposals  $P$ :

$$P = \text{RPN}(E(x^{(a,d)}) \odot F_{att}^{(a,d)}). \quad (10)$$

### 4.3. Causal Prototype Learning

To facilitate the learning of object-level causal features, we introduce the Causal Prototype Learning module which encompasses an explicit constraint and an implicit constraint.

The explicit constraint is imposed on ROI features extracted from different data distributions. Specifically, given the proposal set  $P$  generated from the source image, we select a group of proposals with a confidence higher than the threshold  $t$ , denoted as  $P(t)$ . We represent the ROI features generated from  $P(t)$  as  $f(x, P(t))$  for simplicity. Then the explicit constraint is defined as:

$$\mathcal{L}_{exp} = KL(p^{do(a_0, d_0)}, p^{do(a_k, d_k)}), \quad (11)$$

where  $KL$  denotes the Kullback-Leibler divergence,  $p^{do(a_0, d_0)}$  is the short for  $p(y|f(x^{(a_0, d_0)}, P(t)))$ ,  $p^{do(a_k, d_k)}$  is the short for  $p(y|f(x^{(a_k, d_k)}, P(t)))$  and  $p^{do(a, d)} = \text{MLP}(f(x^{(a, d)}, P(t)))$ .

The explicit constraint  $\mathcal{L}_{exp}$  encourages the within-class distance of the learned class representations from different data distributions to be concentrated and also gives supervisory information to the data-augmented image at the object feature level, which enhances the saliency of the object region, thereby improving the object localization performance of the data-augmented images.

On the other hand, the implicit invariance constraint constrains the relationship between category prototypes across different data distributions. We hypothesize that the distance between causal prototypes of the same category for different data distributions is smaller compared to the distance between prototypes of different categories:

$$\text{dist}(v_{c_i}^{(a_0, d_0)}, v_{c_j}^{(a_k, d_k)}) > \text{dist}(v_{c_i}^{(a_0, d_0)}, v_{c_i}^{(a_k, d_k)}) \approx 0, \quad (12)$$

where  $i \neq j$ ,  $\text{dist}(\cdot)$  is a distance metric function,  $v_{c_i}^{(a_0, d_0)}$  is the prototype of category  $c_i$  of the source data distribution and  $v_{c_j}^{(a_k, d_k)}$  is the prototype of category  $c_j$  of the augmented data distribution. The prototype  $v_c$  is computed by

dynamically averaging the ROI features belonging to the category  $c$ .

We transform the implicit constraint that satisfy the above hypothesis into a prototype contrastive loss:

$$\mathcal{L}_{imp} = -\log \frac{\exp(s(i, i))}{\exp(s(i, i)) + \sum_j \exp(s(i, j))}, \quad (13)$$

$$s(i, j) = \text{sim}(v_{c_i}^{(a_0, d_0)}, v_{c_j}^{(a_k, d_k)}) / \tau, \quad (14)$$

where  $i \neq j$ ,  $\text{sim}(\cdot, \cdot)$  is the cosine similarity and  $\tau$  is a temperature parameter.

The loss function of CPL module is presented as:

$$\mathcal{L}_{prot} = \mathcal{L}_{exp} + \mathcal{L}_{imp}. \quad (15)$$

The constraints of this module will further refine the image-level attention, thereby allowing the features fed into the RPN module to contain more discriminative information about the objects.

### 4.4. Model Optimization

The data streams of the original images and the augmented images share the network parameters and the total training loss is:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{att} + \lambda_2 \mathcal{L}_{prot}, \quad (16)$$

where  $\mathcal{L}_{sup}$  is the supervised object detection loss for the original images and the augmented images and  $\lambda_{1/2}$  are hyperparameters.

During inference, our UFR model maintains the same parameter size as vanilla Faster R-CNN [33], ensuring that it does not introduce additional space complexity. However, a key difference lies in the calculation of features that are fed into the RPN network, which is determined by Eq. (10).

## 5. Experiments

### 5.1. Experimental Setup

**Datasets.** We conduct experiments on the dataset built in [44]. The dataset consists of five different weather conditions, including Daytime-Clear, Daytime-Foggy, Dusk-Rainy, Night-Clear and Night-Rainy. The Daytime-Clear scene is used as the source domain, comprising 19,395 images for training and 8,313 images for testing. The other four scenes are used as unseen target domains, consisting of 3,775 images in Daytime-Foggy condition, 3,501 images in Dusk-Rainy condition, 26,158 images in Night-Clear condition and 2,494 images in Night-Rainy condition. The dataset contains annotations for seven categories of objects, including person, car, bike, rider, motor, bus and truck.

**Implementation Details.** We adopt Faster R-CNN [33] with ResNet-101 [9] as the object detector. The backbone is initialized with weights pre-trained on ImageNet [5]. We train the model with Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 for 80k iterations. During

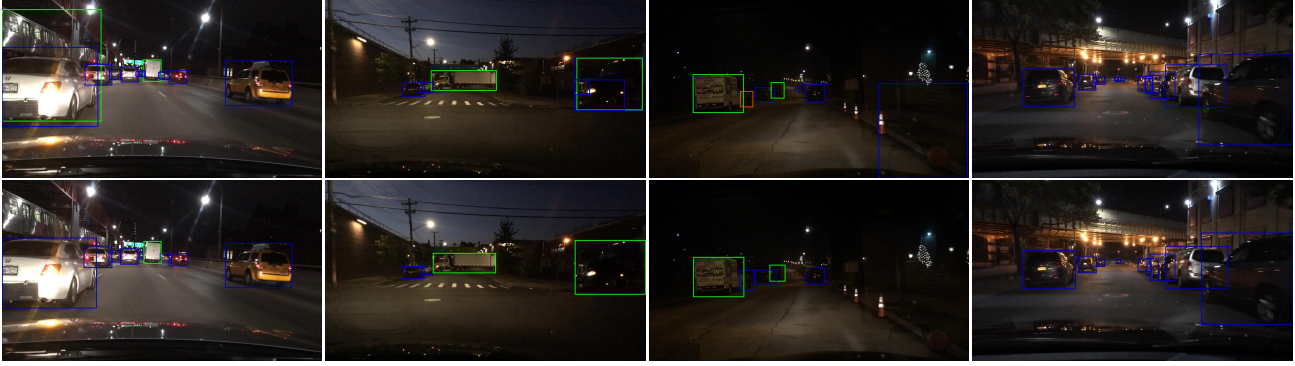


Figure 6. Qualitative evaluation results of the model’s generalization ability on the Night-Clear scene. The **top-row** images are the results of **vanilla Faster R-CNN [33]**. The **bottom-row** images are the results of **our method**.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [33]	66.9	45.9	69.8	46.5	<b>50.6</b>	49.4	64.0	56.2
SW [27]	62.3	42.9	53.3	49.9	39.2	46.2	60.6	50.6
IBN-Net [26]	63.6	40.7	53.2	45.9	38.6	45.3	60.7	49.7
IterNorm [12]	58.4	34.2	42.4	44.1	31.6	40.8	55.5	43.9
ISW [4]	62.9	44.6	53.5	49.2	39.9	48.3	60.9	51.3
SDGOD [44]	<b>68.8</b>	50.9	53.9	<b>56.2</b>	41.8	<b>52.4</b>	<b>68.7</b>	56.1
CLIP-Gap [38]	55.0	47.8	67.5	46.7	49.4	46.7	54.7	52.5
UFR (Ours)	66.8	<b>51.0</b>	<b>70.6</b>	55.8	49.8	48.5	67.4	<b>58.6</b>

Table 1. Quantitative results (%) on the Daytime-Clear scene.

training, the learning rate is set to 0.001, the batch size is set to 4. Besides, the threshold  $t$  is set to 0.7, the  $\lambda_1$  and  $\lambda_2$  are both set to 0.1.

**Data Augmentation Setting.** For local transformation in spatial domain, we randomly apply the augmentation strategies consisting of gaussian blurring, color jittering, random erasing and grayscale. The fusion weight  $\alpha$  in Eq. (6) is a random scalar in  $[0, 1]$ .

## 5.2. Comparison with State-of-the-arts

Following the setting in [44], we use the Mean Average Precision (mAP) as our metric and report the mAP@0.5 results. We compare our method with feature normalization-based SDG methods, including SW [27], IBN-Net [26], IterNorm [12], and ISW [4], and the existing two single domain generalized object detection methods, including SDGOD [44] and CLIP-Gap [38]. Besides, we also compare the performance of our model with vanilla Faster R-CNN (FR) [33].

**Results on Daytime-Clear Scene.** We evaluate the model performance on the source domain. As shown in Table 1, our method achieves an optimal result of 58.6% mAP and has a 2.4% mAP gain against vanilla FR [33]. It shows that our method is able to maintain or even improve the performance of the FR [33] model in the source domain while improving its generalization ability in unseen domains.

**Results on Night-Clear Scene.** The qualitative and quantitative evaluation results of the model’s generalization per-

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [33]	43.5	31.2	49.8	17.5	36.3	29.2	43.1	35.8
SW [27]	38.7	29.2	49.8	16.6	31.5	28	40.2	33.4
IBN-Net [26]	37.8	27.3	49.6	15.1	29.2	27.1	38.9	32.1
IterNorm [12]	38.5	23.5	38.9	15.8	26.6	25.9	38.1	29.6
ISW [4]	38.5	28.5	49.6	15.4	31.9	27.5	41.3	33.2
SDGOD [44]	40.6	35.1	50.7	<b>19.7</b>	34.7	<b>32.1</b>	43.4	36.6
CLIP-Gap [38]	37.7	34.3	58.0	19.2	37.6	28.5	42.9	36.9
Ours	<b>43.6</b>	<b>38.1</b>	<b>66.1</b>	14.7	<b>49.1</b>	26.4	<b>47.5</b>	<b>40.8</b>

Table 2. Quantitative results (%) on the Night-Clear scene.

formance in the Night-Clear scene are shown in Table 2 and Fig. 6, respectively. As shown in Table 2, our method achieves the best result of 40.8% mAP, outperforming Clip-Gap [38] by 3.9% mAP. Besides, compared with vanilla FR [33], our method achieves a 5.0% mAP gain, which indicates that the learned features is beneficial for improving model generalization ability. Besides, compared with SDGOD [44] that learns domain-invariant features, our method has an improvement of 4.2% mAP, which suggests that causal features are more discriminative features and can generalize better. For qualitative results demonstrated in Fig. 6, it is obviously observed that our method detects objects more accurately and has fewer false positives compared with vanilla FR [33].

**Results on Dusk-Rainy and Night-Rainy Scenes.** Table 3 and Table 4 demonstrate the model’s ability to generalize to Dusk-Rainy and Night-Rainy scenes. Our method achieves the best mAP in both scenes, with gains of 0.9% mAP and 0.5% mAP respectively compared with CLIP-Gap [38]. Besides, compared with vanilla FR [33], our method achieves improvements of 5.2% mAP and 5.0% mAP respectively. Furthermore, the feature normalization-based methods [4, 12, 26, 27] have a poor performance in both scenes, especially in motor category in Night-Rainy scene, with mAP of less than 1.0%, which reinforces the challenge of these two scenes and the superiority of our method.

**Results on Daytime-Foggy Scene.** Table 5 shows the gen-



Figure 7. Qualitative evaluation results of the model’s generalization ability on the Daytime-Foggy scene. The top-row and bottom-row images demonstrate the results of vanilla Faster R-CNN [33] and our method, respectively. Besides, the first two columns are the results of real foggy images and the last two columns are the results of synthetic foggy images.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [33]	34.2	21.8	47.9	16.0	22.9	18.5	34.9	28.0
SW [27]	35.2	16.7	50.1	10.4	20.1	13.0	38.8	26.3
IBN-Net [26]	37.0	14.8	50.3	11.4	17.3	13.3	38.4	26.1
IterNorm [12]	32.9	14.1	38.9	11.0	15.5	11.6	35.7	22.8
ISW [4]	34.7	16.0	50.0	11.1	17.8	12.6	38.8	25.9
SDGOD [44]	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
CLIP-Gap [38]	<b>37.8</b>	<b>22.8</b>	60.7	<b>16.8</b>	26.8	<b>18.7</b>	42.4	32.3
Ours	37.1	21.8	<b>67.9</b>	16.4	<b>27.4</b>	17.9	<b>43.9</b>	<b>33.2</b>

Table 3. Quantitative results (%) on the Dusk-Rainy scene.

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [33]	21.3	7.7	28.8	6.1	8.9	10.3	16.0	14.2
SW [27]	22.3	7.8	27.6	0.2	10.3	10.0	17.7	13.7
IBN-Net [26]	24.6	10.0	28.4	0.9	8.3	9.8	18.1	14.3
IterNorm [12]	21.4	6.7	22.0	0.9	9.1	10.6	17.6	12.6
ISW [4]	22.5	11.4	26.9	0.4	9.9	9.8	17.5	14.1
SDGOD [44]	24.4	11.6	29.5	<b>9.8</b>	10.5	<b>11.4</b>	19.2	16.6
CLIP-Gap [38]	28.6	<b>12.1</b>	<b>36.1</b>	9.2	12.3	9.6	22.9	18.7
Ours	<b>29.9</b>	11.8	<b>36.1</b>	9.4	<b>13.1</b>	10.5	<b>23.3</b>	<b>19.2</b>

Table 4. Quantitative results (%) on the Night-Rainy scene.

eralization results of our model in Daytime-Foggy scene. We can see that our method outperforms all the methods in the table and achieves 39.6% mAP. Specifically, compared with SDGOD [44] and CLIP-Gap [38], our method achieves improvements of 6.1% mAP and 1.1% mAP, respectively. Besides, some visualization results are demonstrated in Fig. 7. Compared with vanilla FR [33], our method achieves more accurate object localization and classification in both real and synthetic foggy environments.

### 5.3. Ablation Study

In this section, we conduct several experiments on Daytime-Clear, Night-Clear and Daytime-Foggy scenes to analyze the role of each component of the Unbiased Faster R-CNN model. Specifically, the experimental results in Table 6 are obtained by training on the Daytime-Clear scene and testing

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
FR [33]	34.5	29.6	49.3	26.2	33.0	35.1	26.7	33.5
SW [27]	30.6	36.2	44.6	25.1	30.7	34.6	23.6	30.8
IBN-Net [26]	29.9	26.1	44.5	24.4	26.2	33.5	22.4	29.6
IterNorm [12]	29.7	21.8	42.4	24.4	26.0	33.3	21.6	28.4
ISW [4]	29.5	26.4	49.2	27.9	30.7	34.8	24.0	31.8
SDGOD [44]	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5
CLIP-Gap [38]	36.2	34.2	57.9	<b>34.0</b>	38.7	<b>43.8</b>	25.1	38.5
Ours	<b>36.9</b>	<b>35.8</b>	<b>61.7</b>	33.7	<b>39.5</b>	42.2	<b>27.5</b>	<b>39.6</b>

Table 5. Quantitative results (%) on the Daytime-Foggy scene.

on the three weather conditions.

**Analysis of the GLT module.** As shown in Table 6, the results of vanilla FR [33] are improved when combined with the GLT module, with gains of 4.2% mAP in Daytime-Clear scene, 2.8% mAP in Night-Clear scene, and 2.6% mAP in Daytime-Foggy scene. Besides, the 60.4% mAP in Daytime-Clear scene is optimal, which indicates that the GLT module is an effective data augmentation method that increases data diversity and performs exceptionally well in the source domain with supervised learning.

**Analysis of  $\mathcal{L}_{att}$ ,  $\mathcal{L}_{exp}$  and  $\mathcal{L}_{imp}$ .** As depicted in Table 6, incorporating  $\mathcal{L}_{att}$ ,  $\mathcal{L}_{exp}$  and  $\mathcal{L}_{imp}$  in addition to the GLT module decreases the model performance in the Daytime-Clear scene while improving the performance in the Night-Clear and Daytime-Foggy scenes, which suggests that these constraints limit the model’s performance in the seen environment with supervised learning, but improve the model’s performance in unseen environments. The phenomenon is due to that these constraints encourage the model to extract causal knowledge, thereby preventing the model from acquiring domain-specific knowledge. The domain-specific knowledge, although not generalizable, carries some supplementary information for supervised learning. Furthermore, we can observe that the attention invariance loss  $\mathcal{L}_{att}$  improves the model generalization performance more significantly, which indicates that the scene confounders have a greater impact on the model generalization ability. Be-

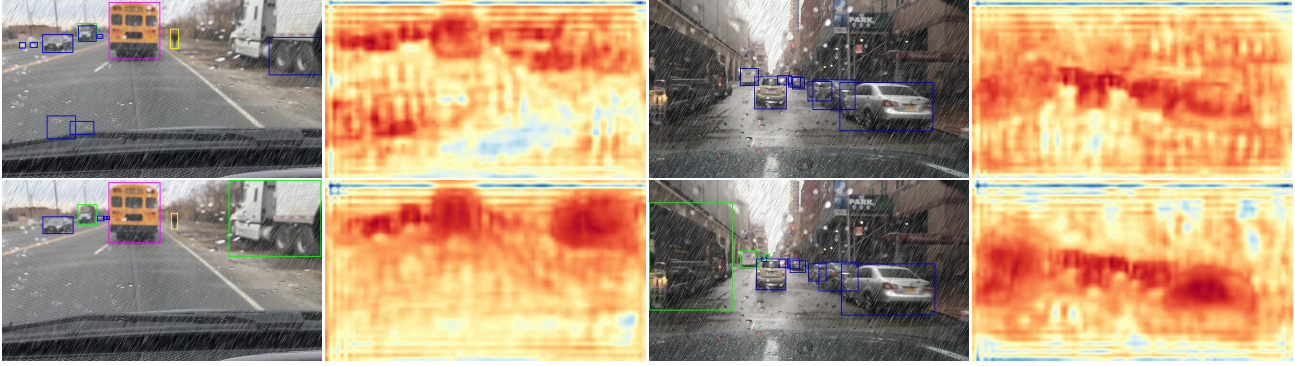


Figure 8. Visualization of the detection results and attention maps on the Dusk-Rainy scene. The top-row and bottom-row images are the results of vanilla Faster R-CNN [33] and our method respectively. The dark red region represents the area where the attention is salient.

Methods	$\mathcal{L}_{att}$	$\mathcal{L}_{exp}$	$\mathcal{L}_{imp}$	D-Clear	N-Clear	D-Foggy
FR [33]				56.2	35.8	33.5
+GLT				<b>60.4</b>	38.6	36.1
		✓		59.5	39.7	38.1
	✓	✓		59.1	40.4	39.0
	✓	✓	✓	59.3	40.0	38.5
	✓	✓	✓	58.6	<b>40.8</b>	<b>39.6</b>

Table 6. Ablation study results (%) of our proposed UFR method. ‘+GLT’ denotes incorporating the GLT to vanilla FR [33]. The loss check mark denotes the addition of the loss to FR [33] with GLT.

sides, compared to the implicit constraint  $\mathcal{L}_{imp}$ , the explicit constraint  $\mathcal{L}_{exp}$  is more binding and contributes more to generalization. Therefore, purely implicit constraint cannot guarantee the model to learn causal prototypes, and explicit constraint on the predicted distribution is necessary.

#### 5.4. Further Analysis

**Attention visualization.** We conduct experiments on Dusk-Rainy scene and compare the attention maps produced by vanilla FR [33] and our UFR model, as shown in Fig. 8. We can observe that the attention maps generated by FR [33] are unfocused and have more attention on irrelevant background areas. In contrast, our method produces more effective category-related attention and has less attention on background, which suggests that our method can capture discriminative object features in unseen environments with a superior generalization performance.

**Hyperparameter analysis.** As demonstrated in Fig. 9, we train our model with different settings of  $t$ ,  $\lambda_1$  and  $\lambda_2$  and test the generalization performance on the Night-Clear and Daytime-Foggy scenes. The model performance with different settings of threshold  $t$  in Eq. (11) is reported in Fig. 9(a). The best threshold  $t$  is 0.7. Besides, we also report the results of our model in Fig. 9(b) with different settings of  $\lambda_1$  and  $\lambda_2$  in Eq. (16). It is obvious that the model achieves the best performance with  $\lambda_1 = 0.1$  and  $\lambda_2 = 0.1$ .

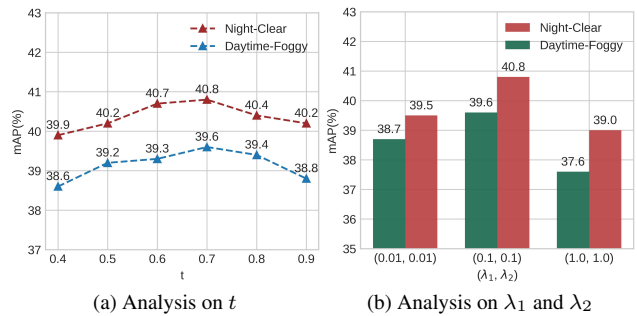


Figure 9. Results of hyperparameter analysis.

## 6. Conclusions

In this paper, we analyze the factors that contribute to the limited single-domain generalization ability of Faster R-CNN from a causal perspective and summarize them as data bias, attention bias and prototype bias. To tackle these challenges, we propose a novel Unbiased Faster R-CNN. Our model leverages a Structural Causal Model to analyze the biases in the task that arise from both scene confounders and object attribute confounders. We design a Global-Local Transformation module for data augmentation to mitigate the data bias. To learn features that are robust to scene confounders, we introduce a Causal Attention Learning module for image-level causal feature selection. To further mitigate the influence of object attribute confounders, we develop a Causal Prototype Learning module that learns object-level causal features. Experimental results and further analysis demonstrate the effectiveness of our method.

## Acknowledgements

This work was supported by National Natural Science Foundation of China under Grants U2013210 and 62303447; Liaoning Science Fund for Distinguished Young Scholars under Grant 2023JH6/100500005; Postdoctoral Innovation Talents Support Program, No. BX20230399.



## References

- [1] Peter Bühlmann. Invariance, Causality and Robustness. *Statistical Science*, 35:404–426, 2020. [3](#)
- [2] Jin Chen, Zhi Gao, Xinxiao Wu, and Jiebo Luo. Meta-causal learning for single domain generalization. In *CVPR*, pages 7683–7692, 2023. [2](#)
- [3] 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](#)
- [4] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, pages 11580–11590, 2021. [3](#), [6](#), [7](#)
- [5] 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, 2009. [5](#)
- [6] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *ECCV*, pages 200–216, 2020. [1](#), [2](#)
- [7] 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](#)
- [8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. [2](#)
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [5](#)
- [10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. [2](#)
- [11] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling Causal Effect of Data in Class-Incremental Learning. In *CVPR*, pages 3956–3965, 2021. [3](#)
- [12] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *CVPR*, pages 4874–4883, 2019. [3](#), [6](#), [7](#)
- [13] Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. In *NeurIPS*, pages 20782–20794, 2022. [3](#)
- [14] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style Neophile: Constantly Seeking Novel Styles for Domain Generalization. In *CVPR*, pages 7120–7130, 2022. [2](#)
- [15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. [2](#), [4](#)
- [16] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. WildNet: Learning Domain Generalized Semantic Segmentation from the Wild. In *CVPR*, pages 9926–9936, 2022. [2](#)
- [17] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xi-aoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *CVPR*, pages 224–233, 2021. [2](#)
- [18] Shanshan Li, Qingjie Zhao, Changchun Zhang, and Yuanbing Zou. Deep discriminative causal domain generalization. *Information Sciences*, 645:119335, 2023. [3](#)
- [19] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018. [1](#), [2](#)
- [20] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-Invariant Disentangled Network for Generalizable Object Detection. In *ICCV*, pages 8751–8760, 2021. [2](#)
- [21] Xinyu Liu, Wuyang Li, and Yixuan Yuan. Decoupled unbiased teacher for source-free domain adaptive medical object detection. *IEEE TNNLS*, 2023. [3](#)
- [22] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality Inspired Representation Learning for Domain Generalization. In *CVPR*, pages 8036–8046, 2022. [3](#)
- [23] Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *NeurIPS*, 2018. [3](#)
- [24] Massimiliano Mancini, Samuel Rota Buló, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *ICIP*, pages 1353–1357, 2018. [2](#)
- [25] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022. [3](#)
- [26] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, pages 464–479, 2018. [3](#), [6](#), [7](#)
- [27] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *ICCV*, pages 1863–1871, 2019. [3](#), [6](#), [7](#)
- [28] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 19(2):3, 2000. [2](#), [3](#)
- [29] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-Aware Domain Generalized Segmentation. In *CVPR*, pages 2584–2595, 2022. [2](#)
- [30] Jiaxin Qi, Kaihua Tang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Class is invariant to context and vice versa: on learning invariance for out-of-distribution generalization. In *ECCV*, pages 92–109, 2022. [3](#)
- [31] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, pages 12556–12565, 2020. [1](#), [2](#)

- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. [3](#)
- [33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. [1](#), [2](#), [5](#), [6](#), [7](#), [8](#)
- [34] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-Weak Distribution Alignment for Adaptive Object Detection. In *CVPR*, pages 6949–6958, 2019. [1](#)
- [35] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint arXiv:2102.11107*, 2021. [3](#)
- [36] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, pages 68–83, 2020. [2](#)
- [37] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031, 2019. [1](#), [2](#)
- [38] Vedit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *CVPR*, pages 3219–3229, 2023. [1](#), [2](#), [3](#), [6](#), [7](#)
- [39] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *CVPR*, pages 6678–6687, 2020. [2](#)
- [40] Lianzhe Wang, Shiji Zhou, Shanghang Zhang, Xu Chu, Heng Chang, and Wenwu Zhu. Improving generalization of meta-learning with inverted regularization at inner-level. In *CVPR*, pages 7826–7835, 2023. [2](#)
- [41] Yunqi Wang, Furui Liu, Zhitang Chen, Yik-Chung Wu, Jianye Hao, Guangyong Chen, and Pheng-Ann Heng. Contrastive-ACE: Domain Generalization Through Alignment of Causal Mechanisms. *IEEE TIP*, 32:235–250, 2023. [3](#)
- [42] Ziqi Wang, Marco Loog, and Jan Van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *ICPR*, pages 9756–9763, 2021. [2](#)
- [43] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *ICCV*, pages 834–843, 2021. [1](#), [2](#), [3](#)
- [44] Aming Wu and Cheng Deng. Single-Domain Generalized Object Detection in Urban Scene via Cyclic-Disentangled Self-Distillation. In *CVPR*, pages 837–846, 2022. [1](#), [5](#), [6](#), [7](#)
- [45] Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *CVPR*, pages 8103–8112, 2023. [2](#), [3](#)
- [46] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *CVPR*, pages 9588–9597, 2021. [3](#)
- [47] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting Causal Mechanisms for Unsupervised Domain Adaptation. In *ICCV*, pages 8579–8588, 2021. [3](#)
- [48] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal Intervention for Weakly-Supervised Semantic Segmentation. In *NeurIPS*, pages 655–666, 2020. [3](#)
- [49] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *CVPR*, pages 8024–8034, 2022. [2](#)
- [50] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, pages 6277–6286, 2021. [1](#), [2](#)
- [51] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization: A Survey. *IEEE TPAMI*, 45:4396–4415, 2023. [1](#)