# Multi-view Adversarial Discriminator: Mine the Non-causal Factors for Object Detection in Unseen Domains

Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, Lei Zhang*

School of Microelectronics and Communication Engineering, Chongqing University, China
Hikvision Research Institute, Hangzhou, China

{mingjunxu, lingyunQin}@cqu.edu.cn, {chenweijie5, pushiliang.hri}@hikvision.com,

leizhang@cqu.edu.cn

## Abstract

*Domain shift degrades the performance of object detection models in practical applications. To alleviate the influence of domain shift, plenty of previous work try to decouple and learn the domain-invariant (common) features from source domains via domain adversarial learning (DAL). However, inspired by causal mechanisms, we find that previous methods ignore the implicit insignificant non-causal factors hidden in the common features. This is mainly due to the single-view nature of DAL. In this work, we present an idea to remove non-causal factors from common features by multi-view adversarial training on source domains, because we observe that such insignificant non-causal factors may still be significant in other latent spaces (views) due to the multi-mode structure of data. To summarize, we propose a Multi-view Adversarial Discriminator (MAD) based domain generalization model, consisting of a Spurious Correlations Generator (SCG) that increases the diversity of source domain by random augmentation and a Multi-View Domain Classifier (MVDC) that maps features to multiple latent spaces, such that the non-causal factors are removed and the domain-invariant features are purified. Extensive experiments on six benchmarks show our MAD obtains state-of-the-art performance.*

## 1. Introduction

The problem of how to adapt object detectors to unknown target domains in real world has drawn increasing attention. Traditional object detection methods [11, 12, 25, 29, 30] are based on independent and identically distributed (i.i.d.) hypothesis, which assume that the training and testing datasets have the same distribution. However, the target distribution can hardly be estimated in real world and differs from the source domains, which is coined as domain
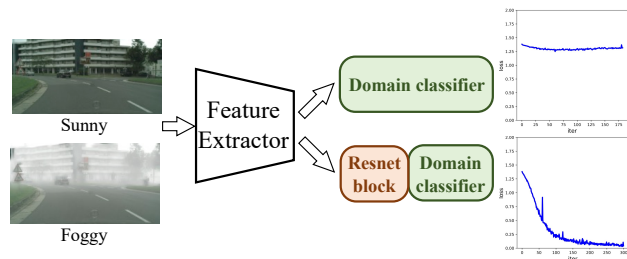


Figure 1. Illustration of the biased learning of conventional DAL. The domain classifier easily encounters early stop and fails.

shift [38]. And the performance of object detection models will sharply drop when facing the domain shift problem.

Domain adaptation (DA) [3, 6, 17, 34, 40, 44, 52] is proposed to deal with the domain shift problem, which enables the model to be adapted to the target distribution by aligning features extracted from the source and unlabeled target domains. However, the requirement of target domain datasets still limits the applicability of DA methods in reality. Domain generalization (DG) [49] goes one step further, aiming to train a model from single or multiple source domains that can generalize to unknown target domains.

Although lots of DG methods have been proposed in the image classification field, there are still some unresolved problems. In our opinion, the common features extracted by previous DG methods are still not pure enough. The main reason is that through a single-view domain discriminator in DAL, only the significant domain style information can be removed, while some implicit and insignificant non-causal factors in source domains may be absorbed by the feature extractor as a part of common features. This has never been noticed. This implies the multi-mode structure of data and single-view domain discriminator cannot fully interpret the data. There is a piece of evidence to support our claim.

To confirm our suspicions on the domain discriminator, we designed a validation experiment. As is shown in Fig. 1,

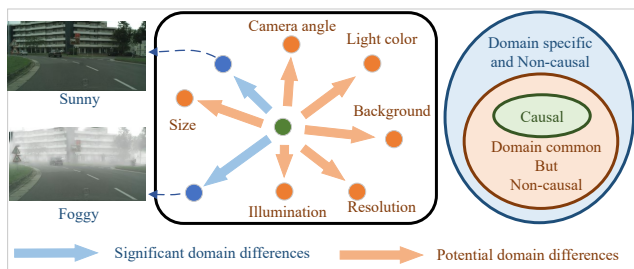---

*Corresponding author (Lei Zhang)

Figure 2. Relationships among causal factors, noncausal factors, domain specific feature and domain common feature.



Figure 3. An illustration of the multi-view idea and effect of MAD. Left: a toy example. Right: attention heatmaps of different views.

we use DANN model [9] with DAL strategy to train a common feature extractor. When domain classifier converges, we freeze feature extractor and re-train domain classifier with a newly added residual block [14]. We observe an interesting phenomenon: *when re-trained with the newly added residual block, the domain classifier loss continues to decline*. That is, some domain-specific information still exists. This phenomenon confirms our claim that in existing DG, DAL cannot explore and remove all domain specific features. This is because domain classifier only observes significant domain-specific feature in a single-view, while insignificant domain specific features in one view (space) can be significant in other views (latent spaces).

Based on the former experiment, we propose that mining common features through DAL in single-view on a limited number of domains is insufficient. By using traditional DAL, only the primary style information w.r.t. domain labels can be removed. Here we analyse this problem from the perspective of causality. As shown in Fig. 2, in a limited number of domains, the common features still contain non-causal factors such as light color, illumination, background, etc., which is expressed as the orange arrows in the figure. And such insignificant non-causal factors observed from one view may still be significant uninformative features in other latent spaces (views). So a natural idea is to explore and remove the implicit non-causal information from multiple views and purify the common features for generalizing to unseen domains.

In order to remove the potential non-causal information, we rethink the domain discriminator in DAL and propose a multi-view adversarial domain discriminator (MAD) that can observe the implicit insignificant non-causal factors. In our life, in order to get the whole architecture of an object, we often need to observe it from multiple views/profiles. A toy example is shown in Fig. 3 (left part). When we observe the Penrose triangle from one specific view, we might misclassify it as a *triangle*, ignoring that it might also appear to be *L* from another perspective. Following this intuition, we construct a Multi-View Domain Classifier (MVDC) that can discriminate features in multiple views. Specifically,
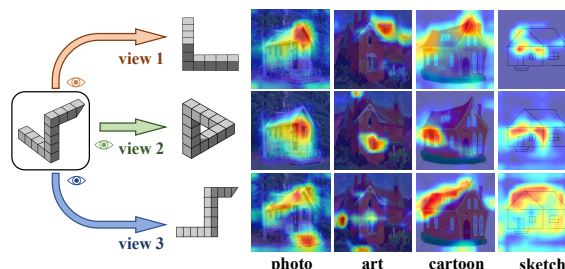
we simulate multi-view observations by mapping the features to different latent spaces with auto-encoders [16], and discriminate these transformed features via multi-view domain classifiers. By mining and removing as many non-causal factors as possible, MVDC encourages the feature extractor to learn more domain-invariant but causal factors. We conduct an experiment based on MVDC and show the heatmaps from different views in Fig. 3 (right part), which verifies our idea that different noncausal factors can be unveiled in different views.

Although the Multi-View Domain Classifier can remove the implicit non-causal features in principle, it still implies a sufficient diversity of source domains during training. So we further design a Spurious Correlation Generator (SCG) to increase the diversity of source domains. Our SCG generates non-causal spurious connections by randomly transforming the low-frequency and extremely high-frequency components, as [19] points out that in the spectrum of images, the extremely high and low frequency parts contain the majority of domain-specific components.

Combining MVDC and SCG, the Multi-view Adversarial Discriminator (MAD) is formalized. Cross-domain experiments on six standard datasets show our MAD achieves the SOTA performance compared to other mainstream DGOD methods. The contributions are three-fold:

1. We point out that existing DGOD work focuses on extracting common features but fails to mine and remove the potential spurious correlations from a causal perspective.

2. We propose a Multi-view Adversarial Discriminator (MAD) to eliminate implicit non-causal factors by discriminating non-causal factors from multiple views and extracting domain-invariant but causal features.

3. We test and analyze our method on standard datasets, verifying the effectiveness and superiority of our method.

## 2. Related Work

### 2.1. Domain Adaptive Object Detection

Object detection is a critical problem in computer vision, aiming to locate and classify the specified instances in spe-

cific images. Modern object detection methods can be divided into two categories: one-stage methods [25, 29] and two-stage methods [11, 12, 30]. However, traditional object detection methods suffer from domain shifts in practical applications. In order to alleviate the performance degradation caused by domain shift, lots of domain adaptive object detection (DAOD) methods are presented [3, 6, 17, 34, 40, 44, 52]. DAOD methods are trained with labeled source domains and unlabeled target domains, and alleviate the domain shift problem by DAL. The DAOD methods can be divided into two parts: adversarial-based methods and reconstruction-based methods. For the former, the domain adversarial learning structure is introduced to align feature maps by [6]. For the latter, [2] firstly uses CycleGAN [51] to generate pseudo samples that are similar to the target domain from the source domain samples. DAOD methods still have problems in real-world applications. On the one hand, they still require additional effort to collect unlabeled target domain datasets, which is expensive and even impossible. On the other hand, they cannot guarantee the causality of features. We hope to find domain-invariant but causal features that are more robust for unseen target domains.

## 2.2. Domain Generalization

Domain generalization has been studied for a long time in the image classification field. Existing domain generalization methods can be divided into the following three categories. First, domain augmentation methods aim to increase the diversity of source domains by transferring images to new domains. [35, 39, 50] augment source domains in image-space. [45] and [19] perform augmentation on the frequency spectrum. Second, representation learning methods aim to extract domain-invariant representation from source domains. [22] firstly adopts the idea of DAL in domain adaptation for domain generalization. Third, there are also learning strategies like [21] which firstly adopts meta-learning for domain generalization, following the idea of enabling the network learn how to learn domain-invariant components from different domains.

## 2.3. Causal Mechanism

Methods based on causal mechanisms [26, 36] consider that the prediction based on statistical dependence is unreliable, because the statistical correlations contain both spurious non-causal correlations and causal correlations. For example, smoking, yellow teeth and lung cancer are closely related. Nevertheless, only smoking is the causal factor of lung cancer. To improve the generalization of methods, they try to mine these invariant causal correlations. In recent years, solving DG problems by finding causal factors is gaining more and more attention [27, 28]. Some methods attempt to obtain invariant causal mechanisms [15, 37, 42]. Meanwhile, other methods try to recover causality charac-
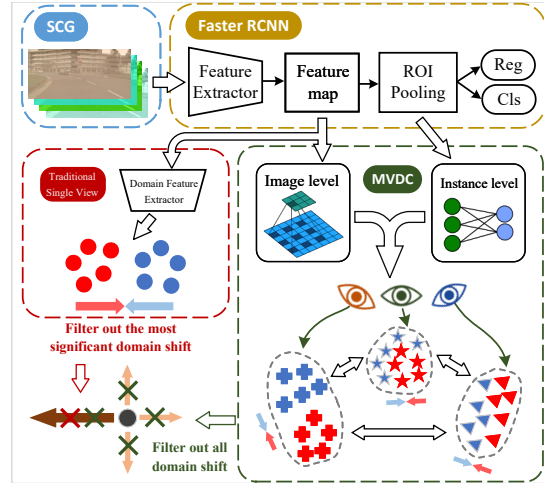


Figure 4. The overall structure of MAD can be divided into three parts. (1) Yellow part: **FasterRCNN** backbone network. (2) Blue part: **SCG** generates potential spurious correlations in the frequency domain. (3) Green part: **MVDC** maps features to different spaces for multi-view DAL and removes implicit non-causal features, such that the domain-invariant but causal features are obtained. Notably, the red part shows the conventional DAL.

teristics [4, 13, 24, 32]. Existing methods focus on looking for invariant causal factors. However, we argue that one should pay more attention to exploring the potential non-causal spurious correlations, because the domain-invariant representations learned by traditional DAL are often biased towards one view, as Fig. 3 shows. We propose to purify the domain-invariant features by removing implicit non-causal factors from multiple views in DAL.

## 3. Proposed MAD Approach

### 3.1. Overview

Existing DG methods learn common features with conventional DAL among finite domains [21, 22]. However, such common features extracted are often not pure due to the implicit non-causal factors. As discussed before, we propose a Multi-view Adversarial Discriminator (MAD) to explore and remove potential spurious correlations and encourage the model to extract purer domain-invariant but causal features. As is shown in Fig. 4, our MAD contains two new parts. First, a Spurious Correlation Generator (SCG) module is designed to increase the diversity of source domains and make the potential non-causal factors more significant. Second, a Multi-View Domain Classifiers (MVDC) module is designed to identify the non-causal factors for both image and instance levels, such that the domain adversarial learning is more sufficient and the non-causal factors are richer in different views, which instructs the feature extractor to ignore them. To summarize, SCG explores
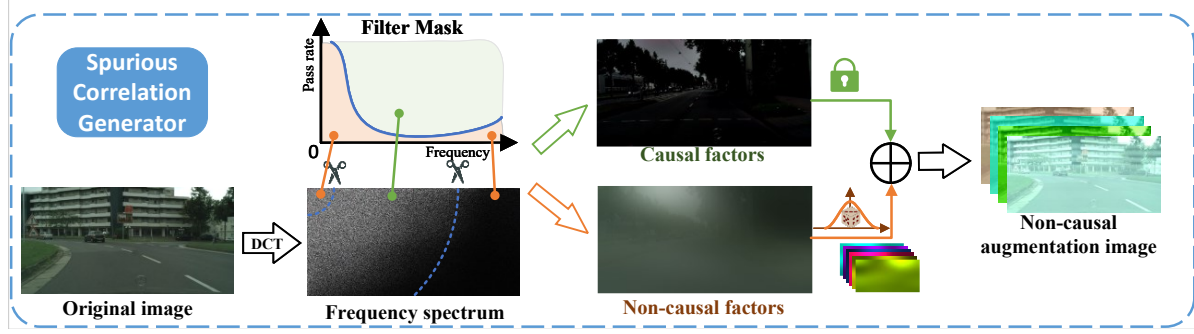
Figure 5. Structure of the Spurious Correlation Generator (SCG). The DCT spectrum of the original image is divided into causal and non-causal parts by the band-pass filter. The extremely high and low frequency components contain more non-causal factors and the remaining components is considered to contain more causal factors. The diversity of source domains is increased by randomising the non-causal part according to a Gaussian distribution. After IDCT, an image with potential non-causal factors is generated.

and exposes the potential non-causal factors, while MVDC discriminates and removes them.

We make the following definitions to formalize the domain generalization problem. The source domain is denoted as $D_s = \{X_s, Y_s\}$. The feature extractor $f(\cdot)$ can extract the features $S = f(X_s)$ from input images $X_s$. Feature $S$ contains causal and non-causal factors $\{s_{cau}, s_{non}\}$ in the finite source domains. Intuitively, not all common components $s_{com}$ are causal factors, but domain private components $s_{pri}$ are non-causal factors, and there is,

$$\begin{cases} s_{cau} \subset s_{com} \\ s_{non} \supset s_{pri} \end{cases} \quad (1)$$

The non-causal factors $s_{non}$ are supposed to obey Gaussian distribution, i.e., $s_{non} \sim \mathcal{N}(\mu, \sigma^2)$ [41].

## 3.2. Spurious Correlations Generator

As [19] pointed out, the extremely high and low frequency parts of images contain more domain-private features. Our SCG aims to increase the diversity of source domains by keeping the causal features invariant and randomizing the non-causal frequency components according to a Gaussian distribution. Specifically, our SCG is implemented by the following steps as shown in Fig. 5. Firstly, by adopting Discrete Cosine Transform ($\mathscr{F}(\cdot)$) [1], we get the frequency spectrum of the input image $x \in \mathbb{R}^{H \times W}$ as $\mathscr{F}(x)$, of which the extremely high and low frequency parts contain more non-causal factors. Then, we separate the non-causal factors and causal factors in frequency domain by a band-pass filter:

$$\mathcal{M}(r) = e^{-\frac{u^2+v^2}{2R_H{}^2}} - e^{-\frac{u^2+v^2}{2R_L{}^2}} \quad (2)$$

where $u, v$ denotes the position of the spectrum and $r(R_L, R_H)$ denote the cut-off frequency of low and high frequency. We then randomize this non-causal factor $S$ according to a Gaussian distribution as $R_G(S) = S \cdot (1 +$

$\mathcal{N}(0, 1))$. Finally, we get the augmented image $\hat{x}$ with potential non-causal factors by adopting the Inverse Discrete Cosine Transform $\mathscr{F}'(\cdot)$ to the augmented spectrum. Our spurious correlations generator can be expressed as:

$$\hat{x} = \mathscr{F}'\left(R_G(\mathcal{M}(r) \cdot \mathscr{F}(x)) + (1 - \mathcal{M}(r)) \cdot \mathscr{F}(x)\right) \quad (3)$$

## 3.3. Multi-View Domain Classifier

### 3.3.1 Domain Adversarial Learning (DAL)

DAL [9] is a standard method to extract the common feature of different domains, which minimizes the $\mathcal{A} - Distance$ of the extracted features between different domains. DAL is a minimax optimization problem between feature extractor $\mathcal{F}$ and the ideal domain classifier:

$$\min_{\mathcal{F}} d_{\mathcal{A}}(D_{s1}, D_{s2}) = \underbrace{\max_{\mathcal{F}} \min_{h \in \mathcal{H}} err(h(s))}_{Standard\ DAL}$$

$$\Rightarrow \underbrace{\max_{\mathcal{F}} \sum_{i=1}^{M} \min_{h_i \in \mathcal{H}, e_i} err(h_i(e_i(s)))}_{Ours} \quad (4)$$

where $\mathcal{H}$ denotes a hypothesis set of all possible domain classifiers, $h(\cdot)$ is one of the domain classifiers in $\mathcal{H}$, and $e(\cdot)$ denotes encoders which map feature to divers latent spaces. A single $h$ depends on the most discriminative domain private features, so it ignores the insignificant domain specific components of the features and incorrectly takes such non-causal components as common features.

We, therefore, propose to improve the sensitivity of the domain classifier to potential non-causal factors by extending DAL to more views. Specifically, our MVDC can map features into multiple latent spaces with encoders $e_i$ and then discriminate features in each space with an independent domain classifier $h_i$. These domain classifiers encourage the feature extractor $\mathcal{F}$ to ignore the implicit non-causal factors and learn domain-invariant but causal features.
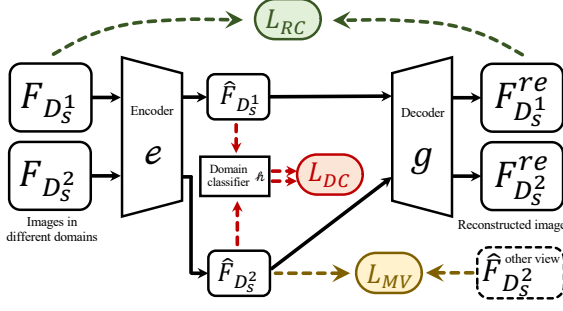
Figure 6. Structure of one branch of MVDC. It contains three constraints: $\mathcal{L}_{RC}$ ensures the mapped features contain complete semantic information. $\mathcal{L}_{DC}$ makes the domain discriminator of each view have domain classification ability. $\mathcal{L}_{MV}$ ensures that each auto-encoder maps features to different latent spaces.

### 3.3.2 Classifier Structure

Fig. 6 shows the structure of one branch of the MVDC, which represents one of the multiple views to observe the features. The complete structure of MVDC contains $M$ branches for image-level features and $M$ branches for instance-level features respectively. Each branch of MVDC contains an auto-encoder and a classifier in structure.

The encoder and decoder are the basic network structure of each branch, which map features into different latent spaces to show different profiles of the feature. The encoder part aims to compress the features and map them into different latent spaces. Then the latent features are fed into an independent domain classifier. Meanwhile, in order to ensure the semantic content invariance of features, the latent features are mapped back into the original space through a subsequent decoder.

To explore the non-causal factors hidden in the whole image and each instance, we make different designs on multi-view domain classifiers for image-level and instance-level respectively. For the image-level, we focus on the global non-causal factors of an image, such as the illumination, color and background texture. These global non-causal factors are similar across the image, so we use the convolutional layers to construct the encoder and decoder. In each branch, we use dilated convolution [47] with different dilation rates to extract different non-causal factors of domains. For the instance-level, we use fully connected layers to mind more semantic non-causal factors like the camera angle of each instance.

### 3.3.3 Loss Function

For training the object detector with MAD approach, several loss functions are introduced in the following text.

First, a reconstruction loss is used to ensure that the semantics of features are not changed by the encoder. The mapped feature $e(s)$ should contain the semantic information required to reconstruct the original feature $s$. Only if the semantic information is guaranteed to be complete, the subsequent domain classifier is meaningful. So we use MSE loss to constrain the distance between the original feature $s$ and the reconstructed feature $g(e(s))$. The reconstruction loss can be described as:

$$\mathcal{L}_{RC} = \frac{1}{M} \sum_{m=1}^{M} MSE(s, g_m(e_m(s))) \quad (5)$$

Second, the adversarial domain classifier loss is used to ensure the mapped features are domain distinguishable (inner optimization) and domain confused (outer optimization). We use Cross-Entropy loss to adversarially train the K-domain classifiers in total $M$ branches. And the domain label in $k^{th}$ domain is denoted as $y_k$.

$$\mathcal{L}_{DC} = -\frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} y_k \cdot log(p(D_m(e_m(s_k)))) \quad (6)$$

The third constraint is the most critical view-different loss, which ensures the auto-encoders to map features into diverse latent spaces. Therefore, we propose to enlarge the feature difference between latent spaces (views), such that the insignificant non-causal factors become significant. So we construct the following MSE loss of each feature pair from $M$ different latent spaces.

$$\mathcal{L}_{MV} = -\frac{\sum_{i}^{M} \sum_{j,i \neq j}^{M} ||e_i(s) - e_j(s)||^2}{M^2 - M} \quad (7)$$

The fourth constraint is used to ensure the consistency of the results in these $2M$ branches of two levels. For each pair of image-level and instance-level branches, we adopt $l_2$ distance between the average value of image-level predictions $p_i^{(u,v)}$ and each instance-level prediction $p_{j,n}$ as the consistency constraint. We suppose the feature map of each image contains $|I|$ pixels and $N$ instances in total. Then, the consistency loss of the whole model can be described as:

$$\mathcal{L}_{cst} = \sum_{i,j}^{M} \sum_{n}^{N} \|\frac{1}{|I|} \sum_{u,v} p_i^{(u,v)} - p_{j,n}\|_2 \quad (8)$$

The MVDC loss in both image level and instance level can be presented as:

$$\mathcal{L}_{MVDC}^{(img,ins)} = \mathcal{L}_{RC} + \mathcal{L}_{DC} + \mathcal{L}_{MV} \quad (9)$$

Then, we obtain the overall loss of MAD by trade-off the object detector loss and MVDC loss with $\lambda$ as:

$$\mathcal{L}_{MAD} = \mathcal{L}_{det} + \lambda(\mathcal{L}_{MVDC}^{img} + \mathcal{L}_{MVDC}^{ins} + \mathcal{L}_{cst}) \quad (10)$$

| Source \ Target | Method | Cityscapes | Foggy Cityscapes | Rain Cityscapes | BDD100k |
|---|---|---|---|---|---|
| Cityscapes | Source-only | — | 27.2 | 36.3 | 24.0 |
| | MLDG | — | 29.2 | 42.1 | 21.0 |
| | FACT | — | 25.3 | 39.9 | 26.0 |
| | FSDR | — | 31.0 | **42.8** | 26.2 |
| | DANN+**SCG** | — | 37.5 | 39.1 | 26.1 |
| | **MAD(Ours)** | — | **38.6** | 42.3 | **28.0** |
| Foggy Cityscapes | Source-only | 29.9 | — | 38.4 | 17.5 |
| | MLDG | 30.4 | — | 38.6 | 18.0 |
| | FACT | 30.0 | — | 38.7 | 20.2 |
| | FSDR | 31.3 | — | 40.8 | 20.4 |
| | DANN+**SCG** | 38.4 | — | 40.4 | 22.4 |
| | **MAD(Ours)** | **41.3** | — | **43.3** | **24.4** |
| BDD100k | Source-only | 33.6 | 27.2 | 34.3 | — |
| | MLDG | 24.7 | 17.1 | 20.0 | — |
| | FACT | 32.4 | 24.3 | 33.9 | — |
| | FSDR | 32.4 | 27.8 | 34.7 | — |
| | DANN+**SCG** | 35.8 | 29.3 | 33.9 | — |
| | **MAD(Ours)** | **36.4** | **30.3** | **36.1** | — |

Table 1. Results on four domains (**C**, **F**, **R**, **B**) trained on single source domain. Note that during training, the target domain is unseen according to DG setting. The best mAP are highlighted in bold.

## 4. Experiments

### 4.1. Datasets

We adopt six cross-domain object detection benchmark datasets, which will be introduced below. Cityscapes [7] dataset mainly contains daytime scenery in the streets and Foggy Cityscapes [7] and Rain Cityscapes [18] are datasets of synthesized images with different weather conditions based on the depth information from the Cityscapes. SIM10k [20] dataset contains rendered images of rendered 3D models.KITTI [10] is an autonomous driving dataset.PASCAL VOC [8] dataset is collected from the real world.The BDD100k [46] dataset is a large-scale dataset for autonomous driving. We abbreviate {SIM 10k, Cityscapes, Foggy Cityscapes, Rain Cityscapes, BDD100k, KITTI, PASCAL VOC} as {S, C, F, R, B, K, V} respectively in the following text.

### 4.2. Experimental Setup

**Implementation Details.** First, to verify the effectiveness of our MAD method, we conduct cross-test experiments on {**C**, **F**, **R**, **B**}, which means we train a model on one of these datasets and test it on the rest datasets. For each source and target pair, we only calculate the result on the intersection of their label space. To uniform the annotation styles of these datasets, we regard labels {*motor*, *motorcycle* and *motorbike*} as {*motor*} and labels {*bike* and *bicycle*} as {*bike*}.

Second, to verify the superiority of our MAD method, we compare with the existing DGOD methods like MLDG [21], DIDN [23], FACT [45] and FSDR [19]. Several DAOD methods such as DAF [6], SW-DA [34], SC-DA [52], MTOR [3], GPA [44] are also compared under the task from cityscapes to foggy cityscapes. We train a total of 10 epochs. In training process, we set the initial learning rate to 0.002, and start to attenuate the learning rate to 0.0002 at the 7th epoch to make the model converge better. In our experiments, we train the models with Mind-Spore [5] and PyTorch frameworks. Our code is available at github.com/K2OKOH/MAD. Mean average precisions (mAP) with a IoU threshold of 0.5 is reported.

**Baseline.** We build our method on the basis of Faster-RCNN [30] framework with vgg16 pre-trained on ImageNet [33] as the backbone and adopt the Stochastic Gradient Descent (SGD) [31] as the optimization method.

### 4.3. Results and Discussion

The results in Tab. 1 show that our method can achieve better results in most cross-domain scenarios. Trained with limited number of source domains, our SCG method can add non-causal factors in more directions to the existing images, which better simulate the potential target domain distribution. This makes our method superior to MLDG, FACT and FSDR that extract features over finite known source domains or fixed augmented domains. Comparing the single-view DANN with our MAD, we can also find that our MAD

| Methods | | Dataset used | person | rider | car | truck | bus | train | motor | bike | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Source-only | | Single Source | 27.1 | 39.3 | 36.0 | 14.2 | 31.4 | 9.4 | 26.9 | 33.4 | 27.2 |
| DA | DAF [6] | Single Source & Target images (without labels) | 31.6 | 43.6 | 42.8 | 23.6 | 41.3 | 21.2 | 28.9 | 32.6 | 33.2 |
| | SW-DA [34] | | 31.8 | 44.3 | 48.9 | 21.0 | 43.8 | 28.0 | 28.9 | 35.8 | 35.3 |
| | SC-DA [52] | | 33.8 | 42.1 | 52.1 | 26.8 | 42.5 | 26.5 | 29.2 | 34.5 | 35.9 |
| | MTOR [3] | | 30.6 | 41.4 | 44.0 | 21.9 | 38.6 | 40.6 | 28.3 | 35.6 | 35.1 |
| | ICR-CCR [43] | | 32.9 | 43.8 | 49.2 | 27.2 | 45.1 | 36.4 | 30.3 | 34.6 | 37.4 |
| | Coarse-to-Fine [48] | | 34.0 | **46.9** | 52.1 | **30.8** | 43.2 | 29.9 | 34.7 | 37.4 | 38.6 |
| | GPA [44] | | 32.9 | 46.7 | 54.1 | 24.7 | **45.7** | **41.1** | **32.4** | **38.7** | 39.5 |
| | Center-Aware [17] | | **41.5** | 43.6 | **57.1** | 29.4 | 44.9 | 39.7 | 29.0 | 36.1 | **40.2** |
| DG | DIDN [23] | Multiple Source | 31.8 | 38.4 | **49.3** | **27.7** | 35.7 | 26.5 | 24.8 | 33.1 | 33.4 |
| | LMDG [21] | Single Source | 32.2 | 41.7 | 38.9 | 19.2 | 33.0 | 9.1 | 23.5 | 36.3 | 29.2 |
| | FACT [45] | | 26.2 | 41.2 | 35.9 | 13.6 | 27.7 | 3.0 | 23.3 | 31.3 | 25.3 |
| | FSDR [19] | | 31.2 | 44.4 | 43.3 | 19.3 | 36.6 | 11.9 | 27.1 | 34.1 | 31.0 |
| | **MAD** | | **34.2** | **47.4** | 45.0 | 25.6 | **44.0** | **42.4** | **30.28** | **40.12** | **38.6** |
| Oracle - Train on target | | Target | 37.8 | 47.4 | 53.0 | 31.6 | 52.9 | 34.3 | 37.0 | 40.6 | 41.8 |

Table 2. Results of DG and DA experiments tested on Foggy Cityscapes (**F**). We compare our MAD method with typical DAOD methods and DGOD methods. DAOD methods are trained on **C** and unlabeled **F**, multi-domain DGOD methods are trained on **C** and **B**, and single-domain DGOD methods are trained on **C**. The best AP in each class and mAP are highlighted in bold.

| Method | F | R | B | V | S | K |
|---|---|---|---|---|---|---|
| SourceOnly | 36.0 | 39.0 | 41.3 | 62.0 | 39.2 | 73.4 |
| DAF | 42.8 | 52.9 | 41.4 | 59.2 | 39.0 | 72.1 |
| MLDG | 38.9 | 52.7 | 39.4 | 61.4 | 37.2 | 63.9 |
| FACT | 35.9 | 48.8 | 42.0 | 65.3 | 41.2 | 73.2 |
| FSDR | 43.3 | 52.7 | **45.4** | 63.4 | 42.2 | 73.8 |
| **MAD** | **45.0** | **54.0** | 42.4 | **67.6** | **43.2** | **74.1** |

Table 3. Results of SourceOnly, DAF, MLDG, FACT, FSDR and our MAD from **C** to **F**, **B**, **V**, **S** and **K** on shared category {*car*}. We train DGOD methods with single source domain **C**, while using both **C** and unlabeled **F** in the training process of DAF. The best mAP is highlighted in bold.
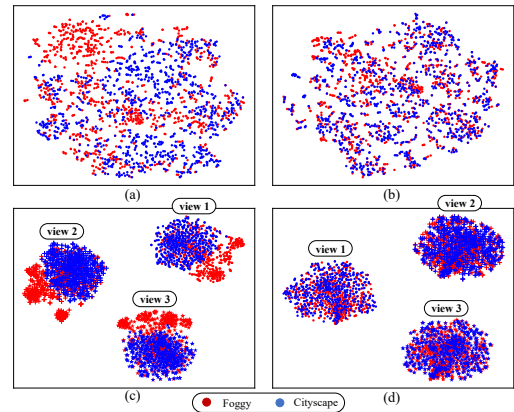


Figure 7. Visualization of the feature distribution via t-SNE. (a) Result of Faster RCNN. (b) Result of DANN under single-view. (c) Result of DANN under multi-view. (d) Result of MAD.

performs better, which shows that our Multi-View domain classifier can further help mine and remove non-causal factors from the simulated target distribution.

In Tab. 2, we compare our MAD with mainstream DG and DA methods. Under both single-source and multi-source DG settings, our method has the best generalization ability among DG methods and exceeds Multi-Source methods in most categories. Furthermore, our target-free MAD can even surpass some of the domain adaptive methods, which are trained with unlabeled target images.

We further conduct experiments on the common categories *car* in six datasets (**C**, **F**, **R**, **S**, **K**, **V** and **B**) to verify the domain generalization ability of our MAD. As shown in Tab. 3, our method also performs the best in most unseen target domains.

As is shown in Fig. 7, we also perform visualization of feature distribution via t-SNE under the task from **C** to **F**.

(a) shows the feature distribution of cars in datasets **C** and **F** extracted with the original Faster RCNN model, in which the difference of distribution between domains is clear. (b) shows the feature distribution of the same datasets and category extracted by DANN, from which we can see that DANN can align the distribution of different domains in a single view. However, as shown in (c), the aligned feature distributions by DANN are still separated with multi-view discriminators by our MVDC, which means that DANN can only remove the significant non-causal factors and the remained insignificant non-causal factors are still domain discriminative. Compared to (c), (d) shows the feature distribution in multi-view extracted by our MAD, and we can see

| Source | Target | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ERM | | | | ERM+SCG | | | | DANN+SCG | | | | MVDC+SCG (MAD) | | | |
| PACS | P | A | C | S | P | A | C | S | P | A | C | S | P | A | C | S |
| P | - | 61.9 | 26.2 | 31.9 | - | 62.8 | 29.3 | 40.1 | - | 63.1 | 35.3 | 43.1 | - | 66.6 | 40.9 | 44.2 |
| A | 90.6 | - | 67.3 | 57.2 | 90.8 | - | 68.7 | 61.7 | 91.4 | - | 70.7 | 64.3 | 92.6 | - | 71.2 | 68.9 |
| C | 79.5 | 64.1 | - | 65.6 | 78.6 | 64.3 | - | 69.0 | 79.2 | 63.6 | - | 69.3 | 79.9 | 64.6 | - | 70.9 |
| S | 48.0 | 42.8 | 60.5 | - | 49.4 | 51.5 | 62.2 | - | 48.7 | 53.8 | 63.4 | - | 53.2 | 57.4 | 63.8 | - |
| VLCS | V | L | C | S | V | L | C | S | V | L | C | S | V | L | C | S |
| V | - | 39.6 | 96.1 | 68.9 | - | 40.1 | 97.6 | 69.2 | - | 43.4 | 98.3 | 69.5 | - | 47.2 | 98.5 | 71.4 |
| L | 61.3 | - | 82.6 | 43.8 | 61.7 | - | 83.7 | 46.9 | 61.7 | - | 83.7 | 46.9 | 62.2 | - | 86.7 | 51.8 |
| C | 50.6 | 20.7 | - | 42.7 | 51.2 | 21.9 | - | 43.5 | 51.7 | 27.2 | - | 44.9 | 51.8 | 29.6 | - | 46.0 |
| S | 60.2 | 45.5 | 72.7 | - | 60.9 | 47.4 | 72.9 | - | 62.4 | 50.0 | 74.9 | - | 64.0 | 51.3 | 75.4 | - |

Table 4. Classification results on PACS and VLCS datasets.

| Methods | | | | mAP | | |
|---|---|---|---|---|---|---|
| SCG | INS | IMG | CST | C to F | C to R | C to B |
| | | | | 27.2 | 36.3 | 24.0 |
| ✓ | | | | 34.1 | 37.9 | 25.4 |
| ✓ | ✓ | | | 38.2 | 40.7 | 26.7 |
| ✓ | ✓ | ✓ | | 38.3 | 41.0 | 26.2 |
| ✓ | ✓ | ✓ | ✓ | **38.6** | **42.3** | **28.0** |

Table 5. Ablation results of each component in MAD trained on **C** and tested on **F**, **R**, **B**. SCG: Spurious Correlations Generator. IMG: Image-level Multi-View Discriminator. INS: Instance-level Multi-View Discriminator. CST: Consistent loss of img and ins.

that our MAD can indeed map features into different spaces and well-align different domains under each view.

The multi-view adversarial discriminator is substantially orthogonal to the computer vision tasks and thus can also be applicable to DG-based image classification tasks. Therefore, we conduct single-source DG experiments on the widely used PACS and VLCS datasets, and compared with ERM and DANN frameworks, as shown in Tab. 4. The results show the effectiveness of our MAD.

### 4.4. Ablation Study

We conducted an ablation study on our MAD methods to verify the validity of each part. Our method can be divided into four parts in total, namely spurious correlations generator (SCG), image-level and instance-level multi-view domain classifier (IMG, INS) and the consistency constraints (CST). We study the contribution of each part by adding them sequentially and observing the change in mAP performance. We train MAD on domain **C** and test it on other domains **F**, **R**, **B** to conduct ablation experiments.

Tab. 5 reflects the effectiveness of each part of our MAD. By introducing SCG, potential spurious correlations are injected into the network. The MVDC consisting of three submodules (IMG, INS, CST) further mines and removes insignificant spurious correlations in the domains. Specifically, the image-level adversarial submodule (IMG) eliminates overall non-causal factors, and the instance-level submodule (INS) eliminates the semantic non-causal factors in each instance. The consistency loss (CST) ensures the consistency of the domain discriminators in two stages.
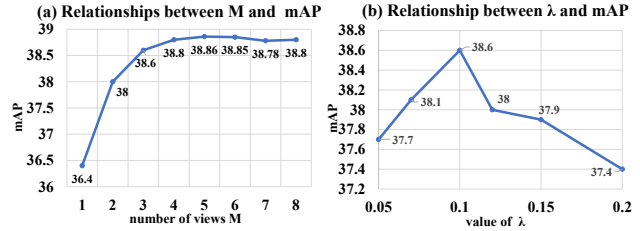


Figure 8. Ablation results of hyperparameters. (a) reflects the performance w.r.t. the number of views (branches). (b) reflects the performance w.r.t. the trade-off coefficient $\lambda$.

### 4.5. Hyper-parameters Analysis

We tested two hyper-parameters of our MAD method.

First, the number $M$ of views is the key hyper-parameter in MAD. More views lead to better performance, but too many auto-encoders increase model complexity with diminishing marginal effect. As we can see in Fig. 8 (a), we found that performance improved until $M = 5$ and then converged with further views until $M = 8$. Thus, we set $M = 3$ as a balance between performance and cost.

Second, the trade-off coefficient $\lambda$ of the domain adversarial loss in Eq. (10) is used to balance the main task of object detection and the MVDC part. We take several values from 0.05 to 0.2 for testing. As can be seen from Fig. 8 (b), we set $\lambda = 0.1$ in MAD for all the experiments.

### 5. Conclusion

This paper analyzes the problem of domain adversarial learning (DAL) from the perspective of causal mechanisms. We point out that existing DG methods fail to remove potential non-causal factors implied in common features, because DAL is biased by the single-view nature of the domain discriminator. To overcome this problem, we propose a Multi-view Adversarial Discriminator (MAD) to learn domain-invariant but causal features. Our MAD includes an SCG that generates potential spurious correlations to diversify the source domains and an MVDC that constructs multi-view domain classifiers to remove implicit non-causal factors in latent spaces. Finally, MAD purifies the domain-invariant features and the causality is augmented. Extensive experiments on benchmarks for cross-domain object detection verify the generalization ability to unseen domains.

# References

[1] Nasir Ahmed, T̲ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 4

[2] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 3

[3] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 1, 3, 6, 7

[4] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020. 3

[5] Lei Chen. *Deep Learning and Practice with MindSpore*. Springer Nature, 2021. 6

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1, 3, 6, 7

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 6

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 4

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1, 3

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1, 3

[13] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[15] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017. 3

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2

[17] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 1, 3, 7

[18] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[19] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2, 3, 4, 6, 7

[20] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 6

[21] 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. 3, 6, 7

[22] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 3

[23] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021. 6, 7

[24] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pages 6804–6814. PMLR, 2021. 3

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1, 3

[26] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000. 3

[27] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. 3

[28] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 3

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1, 3, 6

[31] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 6

[32] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018. 3

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[34] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 1, 3, 6, 7

[35] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 3

[36] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. 3

[37] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019. 3

[38] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1521–1528, 2011. 1

[39] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018. 3

[40] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019. 1, 3

[41] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[42] Yunqi Wang, Furui Liu, Zhitang Chen, Qing Lian, Shoubo Hu, Jianye Hao, and Yik-Chung Wu. Contrastive ace: domain generalization through alignment of causal mechanisms. *arXiv preprint arXiv:2106.00925*, 2021. 3

[43] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 7

[44] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020. 1, 3, 6, 7

[45] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 3, 6, 7

[46] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6

[47] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 5

[48] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13766–13775, 2020. 7

[49] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[50] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. 3

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[52] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019. 1, 3, 6, 7