Causality Inspired Representation Learning for Domain Generalization

Fangrui Lv[1]   **Jian Liang**[2]   **Shuang Li**[1,*]   **Bin Zang**[1]   **Chi Harold Liu**[1]   **Ziteng Wang**[3]   **Di Liu**[2]

[1] Beijing Institute of Technology, China   [2]Alibaba Group, China   [3]Yizhun Medical AI Co., Ltd, China

[1]{fangruilv,shuangli,binzang}@bit.edu.cn, liuchi02@gmail.com
[2]{xuelang.lj,wendi.ld}@alibaba-inc.com   [3]ziteng.wang@yizhun-ai.com

Abstract

Domain generalization (DG) is essentially an out-of-distribution problem, aiming to generalize the knowledge learned from multiple source domains to an unseen target domain. The mainstream is to leverage statistical models to model the dependence between data and labels, intending to learn representations independent of domain. Nevertheless, the statistical models are superficial descriptions of reality since they are only required to model dependence instead of the intrinsic causal mechanism. When the dependence changes with the target distribution, the statistic models may fail to generalize. In this regard, we introduce a general structural causal model to formalize the DG problem. Specifically, we assume that each input is constructed from a mix of causal factors (whose relationship with the label is invariant across domains) and non-causal factors (category-independent), and only the former cause the classification judgments. Our goal is to extract the causal factors from inputs and then reconstruct the invariant causal mechanisms. However, the theoretical idea is far from practical of DG since the required causal/non-causal factors are unobserved. We highlight that ideal causal factors should meet three basic properties: separated from the non-causal ones, jointly independent, and causally sufficient for the classification. Based on that, we propose a Causality Inspired Representation Learning (CIRL) algorithm that enforces the representations to satisfy the above properties and then uses them to simulate the causal factors, which yields improved generalization ability. Extensive experimental results on several widely used datasets verify the effectiveness of our approach. [1]

1. Introduction

In recent years, with the increasing complexity of tasks in real world, out-of-distribution (OOD) problem has raised a severe challenge for deep neural networks based on the i.i.d. hypothesis [23, 24, 30]. Directly applying the model trained on source domain to an unseen target domain with different distribution typically suffers from a catastrophic performance degradation [13, 29, 31, 57]. In order to deal with the domain shift problem, Domain Generalization (DG) has attracted increasing attention, which aims to generalize the knowledge extracted from multiple source domains to an unseen target domain [2, 20, 22, 35].

In order to improve generalization capability, many DG methods have been proposed, which can be roughly categorized into invariant representation learning [9, 22, 25, 34], domain augmentation [53, 59, 63, 66], meta-learning [2, 6, 21], etc. Though promising results have been achieved, there exists one intrinsic problem with them. These efforts merely try to make up for the problems caused by OOD data and model the statistical dependence between data and labels without explaining the underlying causal mechanisms. It has been argued recently [43] that such practices may not be sufficient, and generalizing well outside the i.i.d. setting requires learning not mere statistical dependence between variables, but an underlying causal model [3, 39, 42, 43, 49, 54]. For instance, in an image classification task, it is very likely that all the giraffes are on the grass, showing high statistical dependence, which could easily mislead the model to make wrong predictions when the background varies in target domain. After all, the characteristics of giraffes such as head, neck, etc., instead of the background make a giraffe *giraffe*.

In this paper, we introduce a structural causal model (SCM) [48] to formalize the DG problem, aiming to excavate the intrinsic causal mechanisms between data and labels, and achieve better generalization ability. Specifically, we assume the category-related infor-
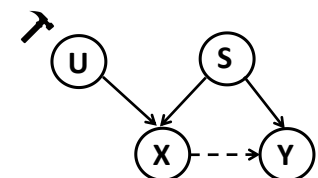

Figure 1. SCM of DG. The **solid arrow** indicates that the parent node causes the child one; while the **dash arrow** means there exists statistical dependence.

∗ Corresponding author.
[1]   Code is available at "https://github.com/BIT-DA/CIRL".

(a) noisy, dependent factorization   **(b) clean, independent factorization**

▲ Non-causal factors   ● Causal factors   ◌ Factorization of causal factors

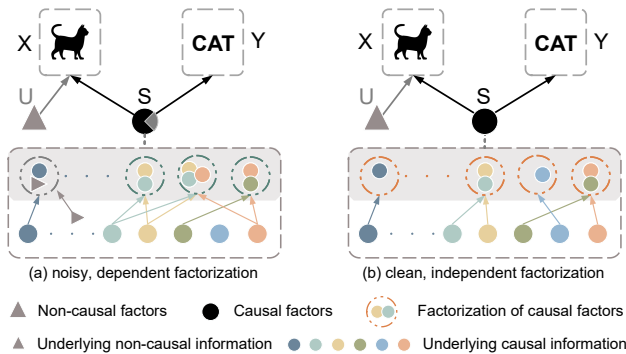▲ Underlying non-causal information   ● ● ● ● ● ● Underlying causal information

Figure 2. Illustration of the three properties of causal factors.

mation in data as causal factors, whose relationship with the label is independent of domain, e.g., *"shape"* in digit recognition. While the information independent of category is assumed as non-causal factors, which is generally domain-related information, e.g., *"handwriting style"* in digit recognition. Each raw data X is constructed from a mix of causal factors S and non-causal factors U and only the former causally effects the category label Y, as shown in Fig. 1. Our goal is to extract the causal factors S from raw input X and then reconstruct the invariant causal mechanisms, which can be done with the aid of causal intervention $P(Y|do(U), S)$. The do-operator $do(\cdot)$ [10] denotes intervention upon variables. Unfortunately, we cannot directly factorize raw input as $X = f(S, U)$ since the causal/non-causal factors are generally unobserved and cannot be formulated, which makes the causal inference particularly challenging [51, 55].

In order to make the theoretical idea into practice, we highlight that the causal factors S are expected to satisfy three properties based on the researches in [43, 46, 49]: 1) separated from the non-causal factors U; 2) the factorization of S should be jointly independent; 3) causally sufficient for the classification task $X \rightarrow Y$ in the sense of containing all the causal information. As shown in Fig. 2 (a), the mixture with U causes S to contain underlying non-causal information, while the jointly dependent factorization makes S redundant, further leading to the miss of some underlying causal information. In contrast, the causal factors S in Fig. 2 (b) are ideal ones that meet all the requirements. Inspired by this, we propose a Causality Inspired Representation Learning (CIRL) algorithm, enforcing the learned representations to possess the above properties and then exploiting each dimension of the representations to mimic the factorization of causal factors, which have stronger generalization ability.

Concisely, for each input, we first exploit a causal intervention module to separate the causal factors S from non-causal factors U via generating new data with perturbed domain-related information. The generated data have different non-causal factors U but the same causal-factors S compared with the original ones, so the representations are enforced to remain invariant. Besides, we propose a factorization module that makes each dimension of the rep-

resentations jointly independent and then can be used to approximate the causal factors. Furthermore, to be causally sufficient towards classification, we design an adversarial mask module which iteratively detects dimensions that contain relatively less causal information and forces them to contain more and novel causal information via adversarial learning between a masker and the representation generator. The contributions of our work are as follows:

- We point out the insufficiency of only modeling statistical dependence and introduce a causality-based view into DG to excavate the intrinsic causal mechanisms.

- We highlight three properties that the ideal causal factors should possess, and propose a CIRL algorithm to learn causal representations that can mimic the causal factors, which have better generalization ability.

- Extensive experiments on several widely used datasets and analytical results demonstrate the effectiveness and superiority of our method.

2. Related Work

Domain Generalization (DG) aims to extract knowledge from multiple source domains that are well-generalizable to unseen target domains. A promising and prevalent solution is to align the distribution of domains by learning domain-invariant representation via kernel-based optimization [8, 35], adversarial learning [22, 25, 34], second-order correlation [41] or using Variational Bayes [62]. Data augmentation is also an important technique to empower the model with generalization ability by enriching source diversity. Several researches have been explored in previous works: [53] perturbs images according to adversarial gradients induced by domain discriminator. [63, 66] mix the styles of training instances across domains by mixing feature statistics [66] or amplitude spectrums [63]. [65] generates more training synthetic data by maximizing a divergence measure. Another popular way that has been investigated is meta-learning, which simulates domain shift by dividing meta-train and meta-test domains from the original source domains [2, 6, 21, 26]. Other DG works also explore low-rank decomposition [45], secondary task as solving jigsaw puzzles [4] and gradient-guided dropout [14]. Different from all the methods above, we tackle DG problem from a causal viewpoint. Our method focuses on excavating intrinsic causal mechanisms by learning causal representations, which has shown better generalization ability.

Causal Mechanism [15, 40, 42] focuses on the fact that statistical dependence ("seeing people take medicine suggests that he is sick") cannot reliably predict the outcome of a counterfactual input ("stopping taking medicine does not make him healthy"). Generally, it can be viewed as components of reasoning chains [19] that provide predictions for

situations that are very far from the observed distribution. In that sense, excavating causal mechanisms means acquiring robust knowledge that holds beyond the support of observed data distributions [50]. The connection between causality and generalization has gained increasing interest in the past few years [33, 43]. Many causality-based methods have been proposed to gain invariant causal mechanisms [12, 56, 61] or recover causal features [5, 10, 27, 47] and hence improve OOD generalization. It is worth noting that they generally rely on restrictive assumptions on the causal diagram or structural equations. Very recently, MatchDG [32] introduces causality into DG literature by enforcing the inputs across domains have the same representation via contrastive learning if they are derived from the same object. Our CIRL is related to MatchDG in its efforts to learn causal representations. However, CIRL differs in the fact that it is done explicitly with exploiting dimension-wise representations to mimic causal factors based on a much theoretical formulation and only relies on a more general causal structural model without restrictive assumptions. Essentially, CIRL can be seen as causal factorization with intervention, which is clearly different from the object conditional MatchDG.

3. Method

In this section, we consider DG from the causal view with a general structural causal model as Fig. 1 shows. We demonstrate that the intrinsic causal mechanisms (formalized as conditional distributions) can be feasible to construct if the causal factors are given. However, as discussed in [1], it is hard to recover the causal factors exactly since they are unobservable. Therefore, we propose to learn causal representations based on the properties of causal factors as a mimic, while inheriting the superior generalization ability.

3.1. DG from the Causal View

The mainstream of DG focuses on modeling the statistical dependence between observed inputs and corresponding labels, i.e., $P(X, Y)$, which is assumed variant across domains. To obtain an invariant dependence, they generally enforce the distribution to be domain-invariant marginally or conditionally, i.e., minimizing the gap across domains in $P(X)$ or $P(X \mid Y)$. However, since the statistical dependence cannot explain the intrinsic causal mechanism between inputs and labels, it tends to vary with domain. Therefore, the learned invariant dependence among source domains may still fail on an unseen target domain. Meanwhile, causal mechanisms usually keep stable across domains [43]. We first articulated the connection between causality and statistical dependence as Reichenbach [46] claimed in Principle 1.

Principle 1 ([46]). ***Common Cause Principle****: if two observables X and Y are statistically dependent, then there exists a variable S that causally influences both and explains*

all the dependence in the sense of making them independent when conditioned on S.

Based on Principle 1, we formalize the following structural causal model (SCM) to describe the DG problem:

$$X := f(S, U, V_1), S \perp\!\!\!\perp U \perp\!\!\!\perp V_1,$$
$$Y := h(S, V_2) = h(g(X), V_2), V_1 \perp\!\!\!\perp V_2. \quad (1)$$

where X, Y represent the input images and corresponding labels respectively. S denotes the causal factors that causally influences both X and Y, i.e., the category-related information such as "*shape*" in digit recognition, while U denotes the non-causal factors that only causally influence X, which is generally domain-related information such as *"style"*. V_1, V_2 are unexplained noise variables that are jointly independent. As for f, h, g, they can be regarded as unknown structural functions. Thus, for any distributions $P(X, Y) \in \mathcal{P}$, if the causal factors S is given, there exists a general conditional distribution $P(Y \mid S)$, i.e., an invariant causal mechanism. Based on the discussion above, if we can access the causal factors, then it is easy to obtain the causal mechanism that generalizes well outside i.i.d. assumption by optimizing h :

$$h^* = \arg\min_h \mathbb{E}_P[\ell(h(g(X)), Y)] = \arg\min_h \mathbb{E}_P[\ell(h(S), Y)], \quad (2)$$

where $\ell(\cdot, \cdot)$ is the cross entropy loss.

Unfortunately, the causal factors S are not provided to us a priori, we are given raw images X instead which are generally unstructured. It is somehow impractical to directly reconstruct the causal factors as well as mechanisms since they are unobservable and ill-defined. What's more, as discussed by [28, 50, 51], which factors can be extracted and their granularity depends on which distribution shifts, supervision signals and explicit interventions are available. Nevertheless, what is clear is that the causal factors still need to obey certain requirements. Previous works [43, 49] declare that causal factors should be jointly independent, as expressed in Principle 2.

Principle 2 ([43, 49]). ***Independent Causal Mechanisms (ICM) Principle****: The conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

Since the S in Eq. (1) represents the set of all causal factors $\{s_1, s_2, \ldots, s_N\}$, the principle tells us that 1) changing (or performing intervention upon) one mechanism $P(s_i \mid PA_i)$ does not change any of the other mechanisms $P(s_j \mid PA_j), i \neq j$ [49] (PA_i denotes s_i's parents in causal graph, which can be regarded as causal information that s_i contains since S is already the root node), 2) knowing some other mechanisms $P(s_i \mid PA_i)$ does not give us information about a mechanism $P(s_j \mid PA_j)$ [17]. Thus, we can factorize the joint distribution of causal factors into conditional as
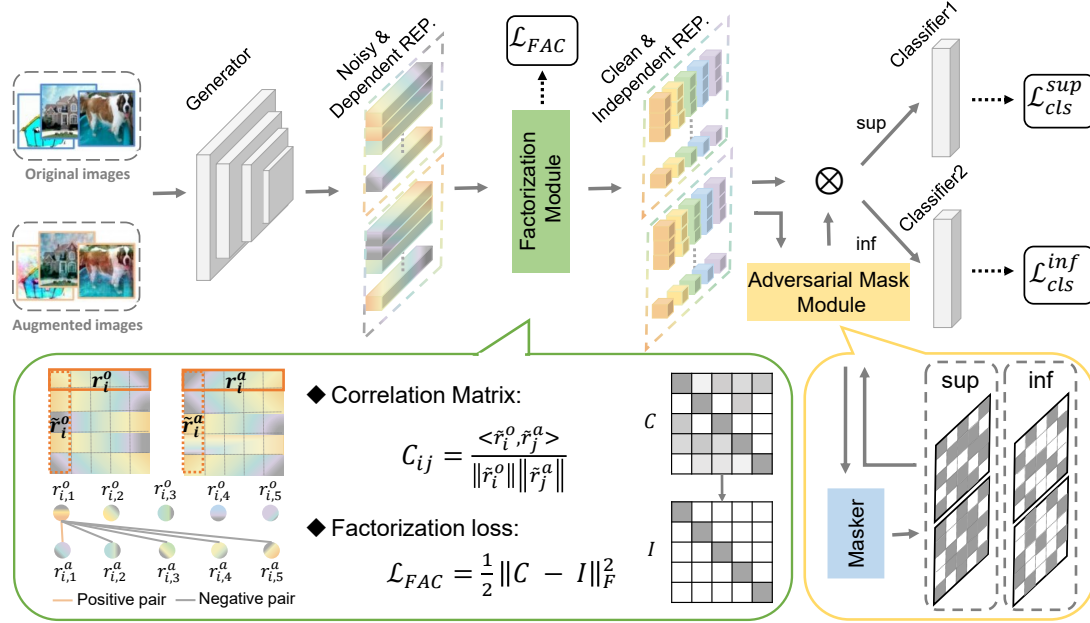
Figure 3. The framework of CIRL. We first generate augmented images by a causal intervention module with intervention upon non-causal factors. Both the representations of original and augmented images are sent to the factorization module, which imposes a factorization loss to force the representations to be separated from non-causal factors and jointly independent. At last, the adversarial mask module conducts an adversary between the generator and a masker, rendering the representations to be causally sufficient for classification.

follows, which refers to causal factorization:

$$P(s_1, s_2, \ldots, s_N) = \prod_{i=1}^{N} P(s_i \mid PA_i), \qquad (3)$$

Therefore, we highlight that the causal factors S should satisfy three basic properties based on the definition of causal variables in Common Cause Principle (Principle 1) and the nature of causal mechanism in ICM Principle (Principle 2):

- The causal factors S should be separated from the non-causal factors U, i.e., $S \perp\!\!\!\perp U$. Thus, performing an intervention upon U does not make changes to S.

- The factorization $s_1, s_2, \ldots, s_N$ should be jointly independent, none of which entails information of others.

- The causal factors S should be causally sufficient to the classification task $X \rightarrow Y$, i.e., contain information that can explain all the statistical dependencies.

Thus, instead of directly reconstructing the causal factors, we propose to learn causal representations as an alternative by forcing them to have the same properties as causal factors. We will explain the details in Sec. 3.2

3.2. Causality Inspired Representation Learning

In this section, we illustrate our proposed representation learning algorithm inspired by the causality discussed above, which consists of three modules: causal intervention module, causal factorization module and adversarial mask module. The whole framework is shown in Fig. 3.

3.2.1 Causal Intervention Module

We first aim to separate the causal factors S from the mixture of non-causal ones U by causal intervention. Specifically, although the explicit form of causal factor extractor $g(\cdot)$ in Eq. (2) is unknown in general, we have prior knowledge that the causal factors S should remain invariant to the intervention upon U, i.e., $P(S \mid do(U))$. While in DG literature, we do know that some domain-related information cannot determine the category of inputs, which can be regarded as non-causal factors and captured by some techniques [63,65,66]. For instance, Fourier transformation has a well-known property: the phase component of Fourier spectrum preserves high-level semantics of the original signal, while the amplitude component contains low-level statistics [38,44]. Thus, we conduct the intervention upon U by disturbing the amplitude information while keeping the phase information unchanged as [63] does. Formally, given an original input image $\boldsymbol{x}^o$, its Fourier transformation can be formulated as :

$$\mathcal{F}(\boldsymbol{x}^o) = \mathcal{A}(\boldsymbol{x}^o) \times e^{-j \times \mathcal{P}(\boldsymbol{x}^o)}, \qquad (4)$$

where $\mathcal{A}(\boldsymbol{x}^o), \mathcal{P}(\boldsymbol{x}^o)$ denote the amplitude and phase components respectively. The Fourier transformation $\mathcal{F}(\cdot)$ and its inverse $\mathcal{F}^{-1}(\cdot)$ can be calculated with the FFT algorithm [37] effectively. We then perturb the amplitude information via linearly interpolating between the amplitude spectrums of the original image $\boldsymbol{x}^o$ and an image $(\boldsymbol{x}')^o$ sampled randomly from arbitrary source domains:

$$\hat{\mathcal{A}}(\boldsymbol{x}^o) = (1 - \lambda)\mathcal{A}(\boldsymbol{x}^o) + \lambda \mathcal{A}((\boldsymbol{x}')^o), \qquad (5)$$

where $\lambda \sim U(0, \eta)$ and η controls the strength of perturbation. Then we combine the perturbed amplitude spectrums with the original phase component to generate the augmented image $\boldsymbol{x}^a$ by inverse Fourier transformation :

$$\mathcal{F}(\boldsymbol{x}^a) = \hat{\mathcal{A}}(\boldsymbol{x}^o) \times e^{-j \times \mathcal{P}(\boldsymbol{x}^o)}, \boldsymbol{x}^a = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x}^a)). \quad (6)$$

Denote the representation generator implemented by a CNN model as $\hat{g}(\cdot)$ and the representations as $\boldsymbol{r} = \hat{g}(\boldsymbol{x}) \in \mathbb{R}^{1 \times N}$, where N is the number of dimensions. To simulate the causal factors that remain invariant to the intervention upon U, we optimize $\hat{g}$ to enforce the representations to keep unchanged dimension-wisely to the above intervention:

$$\max_{\hat{g}} \frac{1}{N} \sum_{i=1}^{N} COR(\tilde{\boldsymbol{r}}_i^o, \tilde{\boldsymbol{r}}_i^a), \quad (7)$$

where $\tilde{\boldsymbol{r}}_i^o$ and $\tilde{\boldsymbol{r}}_i^a$ denote the Z-score normalized i-th column of $\mathbf{R}^o = [(\boldsymbol{r}_1^o)^T, \dots, (\boldsymbol{r}_B^o)^T]^T \in \mathbb{R}^{B \times N}$ and $\mathbf{R}^a = [(\boldsymbol{r}_1^a)^T, \dots, (\boldsymbol{r}_B^a)^T]^T$, respectively, $B \in \mathbb{Z}_+$ is the batch size, $\boldsymbol{r}_i^o = \hat{g}(\boldsymbol{x}_i^o)$ and $\boldsymbol{r}_i^a = \hat{g}(\boldsymbol{x}_i^a)$ for $i \in \{1, \dots, B\}$. We leverage a COR function to measure the correlation of representations before and after the intervention. Thus, we can realize the first step of simulating causal factors S with representations R by making them independent of U.

3.2.2 Causal Factorization Module

As we proposed in Sec. 3.1 that the factorization of causal factors $s_1, s_2 \dots, s_N$ should be jointly independent in the sense of none of them entails information of others. Therefore, we intend to make any two dimensions of the representations independent of each other:

$$\min_{\hat{g}} \frac{1}{N(N-1)} \sum_{i \neq j} COR(\tilde{\boldsymbol{r}}_i^o, \tilde{\boldsymbol{r}}_j^a), i \neq j, \quad (8)$$

Note that for saving computation cost, we omit the constraints within $\mathbf{R}^o$ or $\mathbf{R}^a$. To unify the optimization goals of Eq. (7) and Eq. (8), we build a correlation matrix $\boldsymbol{C}$:

$$\boldsymbol{C}_{ij} = \frac{< \tilde{\boldsymbol{r}}_i^o, \tilde{\boldsymbol{r}}_j^a >}{\|\tilde{\boldsymbol{r}}_i^o\| \|\tilde{\boldsymbol{r}}_j^a\|}, i, j \in 1, 2, \dots, N, \quad (9)$$

where $< \cdot >$ denotes the inner product operation. Thus, the same dimension of $\mathbf{R}^o$ and $\mathbf{R}^a$ can be taken as positive pairs which need to maximize the correlation, while the different dimensions can be taken as negative pairs which need to minimize the correlation. Based on that, we design a Factorization loss $\mathcal{L}_{Fac}$ which can be formulated as follows:

$$\mathcal{L}_{Fac} = \frac{1}{2} \|\boldsymbol{C} - \boldsymbol{I}\|_F^2. \quad (10)$$

Remark 1. The objective in Eq. (10) can make the diagonal elements of correlation matrix C approximate to 1, which means that the representations before and after the

intervention to non-causal factors are invariant. It indicates that we can effectively separate the causal factors from the mixture of the non-causal ones. Moreover, it also makes the non-diagonal elements of C close to 0, i.e., enforcing the dimensions of representations to be jointly independent. Thus, by minimizing $\mathcal{L}_{Fac}$, we can make the noisy and dependent representations into clean and independent ones, satisfying the first two properties of ideal causal factors.

3.2.3 Adversarial Mask Module

To succeed on the classification task $X \to Y$, the representations should be causally sufficient that entails all the support information. The most straightforward way is to utilize the supervision labels y in the multiple source domains:

$$\mathcal{L}_{cls} = \ell(\hat{h}(\hat{g}(\boldsymbol{x}^o)), y) + \ell(\hat{h}(\hat{g}(\boldsymbol{x}^a)), y) \quad (11)$$

where $\hat{h}$ is the classifier. However, this straightforward way cannot guarantee that every dimension of our learned representations are important, i.e., contain sufficient underlying causal information for classification. Specifically, there may be inferior dimensions that carry relatively less causal information and then make a small contribution for classification. Therefore, we propose to detect these dimensions and enforce them to contribute more. Since the dimensions are also required to be jointly independent with the aid of our factorization module, the detected inferior dimensions are rendered to contain more and novel causal information that is not included by other dimensions, which makes the whole representations to be more causality sufficient.

Thus, to detect the inferior dimensions, we design an adversarial mask module. We build a neural-network-based masker, denoted by $\hat{w}$ to learn the contribution of each dimension, and the dimensions correspond to the largest $\kappa \in (0, 1)$ ratio are regarded as superior dimensions while the rest are regarded as the inferior ones:

$$\boldsymbol{m} = \text{Gumbel-Softmax}(\hat{w}(\boldsymbol{r}), \kappa N) \in \mathbb{R}^N, \quad (12)$$

where we employ the commonly-used derivable Gumbel-Softmax trick [16] to sample a mask with κN values approaching 1. The details of the trick are deferred to supplementary materials. By multiplying the learned representations by the obtained masks $\boldsymbol{m}$ and $1 - \boldsymbol{m}$, we can acquire the superior and inferior dimensions of the representations, respectively. Then, we feed them into two different classifiers $\hat{h}_1, \hat{h}_2$. Eq. (11) can be rewritten as follows:

$$\mathcal{L}_{cls}^{sup} = \ell(\hat{h}_1(\boldsymbol{r}^o \odot \boldsymbol{m}^o), y) + \ell(\hat{h}_1(\boldsymbol{r}^a \odot \boldsymbol{m}^a), y),$$
$$\mathcal{L}_{cls}^{inf} = \ell(\hat{h}_2(\boldsymbol{r}^o \odot (1 - \boldsymbol{m}^o)), y) + \ell(\hat{h}_2(\boldsymbol{r}^a \odot (1 - \boldsymbol{m}^a)), y),$$
$$(13)$$

We optimize the masker by minimizing $\mathcal{L}_{cls}^{sup}$ and maximizing $\mathcal{L}_{cls}^{inf}$, while optimize the generator $\hat{g}$ and classifiers $\hat{h}_1, \hat{h}_2$ by minimizing the two supervision loss.

Remark 2. The proposed adversarial mask module can precisely detect the inferior dimensions because 1) for an optimized $\hat{h}_2$ to minimize $\mathcal{L}_{cls}^{inf}$ based on existing masked dimensions, learning m to select dimensions for maximizing $\mathcal{L}_{cls}^{inf}$ can find inferior dimensions with less contribution, and 2) superior and inferior dimension-sets complement each other such that if one dimension is not treated as superior then it will be treated as inferior, thus the selection of the superior ones will help the selection of the inferior ones. Furthermore, compared with optimizing Eq. (11) only, the adversarial mask module combined with our causal factorization module can help to generate more causally sufficient representations, because by optimizing $\hat{g}$ to minimize both $\mathcal{L}_{cls}^{inf}$ and $\mathcal{L}_{Fac}$, the inferior dimensions are forced to carry more causal information and be independent with existing superior dimensions. Finally, the learned representations will approach causal sufficient by iteratively "replacing" the inferior representations to be novel superior ones.

To be clear, the overall optimization objective of our proposed CIRL is summarized as follows:

$$\min_{\hat{g},\hat{h}_1,\hat{h}_2} \mathcal{L}_{cls}^{sup} + \mathcal{L}_{cls}^{inf} + \tau\mathcal{L}_{Fac}, \quad \min_{\hat{w}} \mathcal{L}_{cls}^{sup} - \mathcal{L}_{cls}^{inf}, \quad (14)$$

where τ is the trade-off parameter. Note that the whole representation r and classifier $\hat{h}_1$ are utilized during inference.

Remark 3. Note that the effect of the number of feature dimension is negligible. Through the cooperative optimization of the three modules, the total amount of causal information contained in the whole representations will increase until the learned representations can explain all the statistical dependence between inputs and labels, regardless of the feature dimension. Experimental analyses are provided in supplementary materials, verifying our demonstration.

4. Experiment

4.1. Datasets

Digits-DG [64] encompasses four digit domains including *MNIST* [18], *MNIST-M* [7], *SVHN* [36] and *SYN* [7], which present dramatic differences in the font style, background and stroke color. Following [64], we randomly select 600 images per class for each domain, and then split 80% of the data for training and 20% of the data for validation.

PACS [20] is specially proposed for DG, which contains 9,991 images from four domains (*Art-Painting*, *Cartoon*, *Photo* and *Sketch*) with large style discrepancy. In each domain, there are 7 categories: dog, elephant, giraffe, guitar, house, horse, and person. For a fair comparison, the original training-validation split provided by [20] is used.

Office-Home [58] is an object recognition dataset in the office and home environments, which collects 15,500 images of 65 categories. The 65 categories are shared by four domains (*Art*, *Clipart*, *Product* and *Real-World*), which dif-

Table 1. Leave-one-domain-out results on Digits-DG. The best and second-best results are bold and underlined, respectively.

Methods	MNIST	MNIST-M	SVHN	SYN	Avg.
DeepAll [64]	95.8	58.8	61.7	78.6	73.7
Jigen [4]	96.5	61.4	63.7	74.0	73.9
CCSA [34]	95.2	58.2	65.5	79.1	74.5
MMD-AAE [22]	96.5	58.4	65.0	78.4	74.6
CrossGrad [53]	$\underline{96.7}$	61.1	65.3	80.2	75.8
DDAIG [64]	96.6	64.1	68.6	81.0	77.6
L2A-OT [65]	$\underline{96.7}$	63.9	68.6	83.2	78.1
FACT [63]	**97.9**	$\underline{65.6}$	$\underline{72.4}$	**90.3**	$\underline{81.5}$
CIRL (*ours*)	96.08	**69.87**	**76.17**	$\underline{87.68}$	**82.5**

fers in the viewpoint and image style. Following [63], each domain is split into 90% for training and 10% for validation.

4.2. Implementation Details

Following the commonly used leave-one-domain-out protocol [20], we specify one domain as the unseen target domain for evaluation and train with the remaining domains. For Digits-DG, all images are resized to 32×32, we train the network from scratch using the mini-batch SGD optimizer with batch size 128, momentum 0.9, and weight decay $5e\text{-}4$ for 50 epochs. And the learning rate is decayed by 0.1 every 20 epochs. As for PACS and Office-Home, all images are resized to 224×224. The network is trained from scratch using the mini-batch SGD with batch size 16, momentum 0.9 and weight decay $5e\text{-}4$ for 50 epochs, and the learning rate is decayed by 0.1 at 80% of the total epochs. For the hyperparameters κ and τ, their values are selected according to the results on the source validation set, since the target domain is unseen during the training. Specifically, we set $\kappa = 60\%$ for Digits-DG and PACS while $\kappa = 80\%$ for Office-Home. τ is set as 2 for Digits-DG and 5 for the others. All results are reported based on the average accuracy over three repetitive runs. More details are given in the supplementary details.

4.3. Experimental Results

Results on Digits-DG are presented in Table 1, where CIRL beats all the compared baselines in terms of the average accuracy. Note that CIRL surpasses the domain-invariant representation based methods CCSA [34] and MMD-AAE [22] by a large margin of 8.0% and 7.9%, respectively, which indicates the importance of excavating the intrinsic causal mechanisms between data and labels, instead of the superficial statistical dependence. Besides, we also compare CIRL with FACT [63], since our causal intervention module adopts the same augmentation technique. It is worth mentioning that FACT is a quite state-of-the-art method in DG community and 1.0% performance improvement is challenging. While CIRL achieves 1.0% improvement over FACT, which further validates the effectiveness of our method.

Table 2. Leave-one-domain-out results on PACS with ResNet-18.

Methods	Art	Cartoon	Photo	Sketch	Avg.
DeepAll [64]	77.63	76.77	95.85	69.50	79.94
MetaReg [2]	83.70	77.20	95.50	70.30	81.70
JiGen [4]	79.42	75.25	96.03	71.35	80.51
DDAIG [64]	84.20	78.10	95.30	74.70	83.10
CSD [45]	78.90	75.80	94.10	76.70	81.40
MASF [6]	80.29	77.17	94.99	71.69	81.04
L2A-OT [65]	83.30	78.20	96.20	73.60	82.80
EISNet [60]	81.89	76.44	95.93	74.33	82.15
MatchDG [32]	81.32	**80.70**	<u>96.53</u>	79.72	84.56
RSC [14]	83.43	80.31	95.99	80.85	85.15
FACT [63]	<u>85.90</u>	79.35	**96.61**	<u>80.88</u>	<u>85.69</u>
CIRL (*ours*)	**86.08**	<u>80.59</u>	95.93	**82.67**	**86.32**

Table 3. Leave-one-domain-out results on PACS with ResNet-50.

Methods	Art	Cartoon	Photo	Sketch	Avg.
DeepAll [64]	84.94	76.98	97.64	76.75	84.08
MetaReg [2]	87.20	79.20	97.60	70.30	83.60
MASF [6]	82.89	80.49	95.01	72.29	82.67
EISNet [60]	86.64	81.53	97.11	78.07	85.84
MatchDG [32]	85.61	82.12	**97.94**	78.76	86.11
FACT [63]	**90.89**	<u>83.65</u>	97.78	<u>86.17</u>	<u>89.62</u>
CIRL (*ours*)	<u>90.67</u>	**84.30**	97.84	**87.68**	**90.12**

Table 4. Leave-one-domain-out results on Office-Home

Methods	Art	Clipart	Product	Real	Avg.
DeepAll [64]	57.88	52.72	73.50	74.80	64.72
CCSA [34]	59.90	49.90	74.10	75.70	64.90
MMD-AAE [22]	56.50	47.30	72.10	74.80	62.70
CrossGrad [53]	58.40	49.40	73.90	75.80	64.40
DDAIG [64]	59.20	52.30	74.60	76.00	65.50
L2A-OT [65]	<u>60.60</u>	50.10	<u>74.80</u>	**77.00**	65.60
Jigen [4]	53.04	47.51	71.47	72.79	61.20
RSC [14]	58.42	47.90	71.63	74.54	63.12
FACT [63]	60.34	<u>54.85</u>	74.48	76.55	<u>66.56</u>
CIRL (*ours*)	**61.48**	**55.28**	**75.06**	<u>76.64</u>	**67.12**

Table 5. Ablation study of CIRL on PACS dataset with ResNet-18.

Method	CInt.	CFac.	AdvM.	Art	Cartoon	Photo	Sketch	Avg.
Variant 1	✓	-	-	85.99	78.67	95.81	79.49	84.99
Variant 2	-	✓	-	80.52	78.75	94.37	78.57	83.05
Variant 3	✓	✓	-	84.23	79.44	95.51	82.54	85.43
Variant 4	-	✓	✓	82.18	79.69	95.99	80.68	84.64
CIRL	✓	✓	✓	86.08	80.59	95.93	82.67	**86.32**

Results on Office-Home based on ResNet-18 are summarized in Table 4. The larger number of categories and samples makes Office-Home a more challenging benchmark than PACS for domain generalization. On this challenging benchmark, CIRL still achieves the best average performance of 67.12%, surpassing FACT [63] by a margin of 0.56%. The improvements further justify the efficacy of CIRL.

4.4. Analytical Experiments

Ablation Study. We discuss the influences of the Causal Intervention (*CInt.*) module, Causal Factorization (*CFac.*) module and Adversarial Mask (*AdvM.*) module in CIRL. Table 5 presents the results of different variants of CIRL on PACS dataset with ResNet-18 as the backbone. Comparing variants 1, 2 with variant 3, we can observe that the performance of combining both CInt. and CFac. modules is much better, which indicates that only separating the representation from non-causal factors or making the dimensions independent is not sufficient for well modeling the causal factors. Besides, the improved performance of variant 4 over variant 2 implies that the AdvM. module contributes to integrating more informative information into representations for classification. Finally, CIRL performs best, showing that the three modules complement and promote mutually, none of which is indispensable for the superior generalization ability.

Visual Explanation. To visually verify the claim that the representations learned by CIRL can simulate the causal factors, we provide the attention maps of the last convolutional layer for baseline (i.e., DeepAll) and CIRL methods, utilizing the visualization technique in [52]. The results are shown in Fig. 4. We can see that the representations learned by CIRL are more category-related, compared with the baseline method. Taking the giraffe for example, its long neck can be regarded as one of the causal factors to classify, which is precisely captured by CIRL. While the baseline focuses on non-causal factors (e.g., the texture resembling the mane), which leads to the misclassification. More visualization results are given in the supplementary materials.

Independence of Causal Representation. Fig. 5a and 5b show the independence degree between dimensions of representations on task Sketch. Specifically, we use $\|C\|_F^2 - \|diag(C)\|_2^2$ as metric and smaller value denotes better independence degree, where C is the correlation matrix in Section 3.2.2. It is clear that on both ResNet-18 and ResNet-50, the baseline methods have poor independence
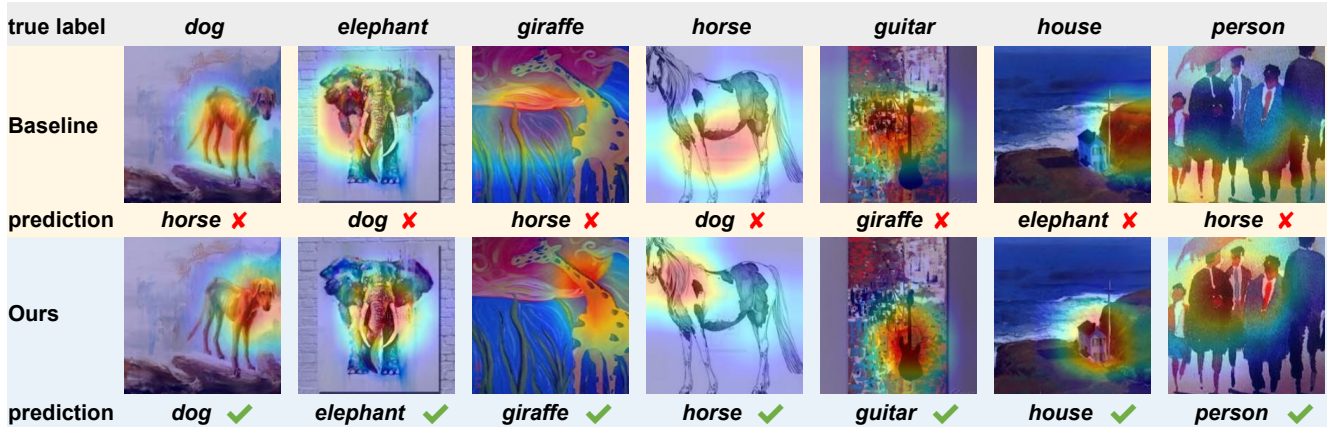
Results on PACS based on ReNet-18 and ResNet-50 are reported in Table 2 and 3, respectively. It can be observed that CIRL obtains the highest average accuracy among all the compared methods on both backbones. Specifically, compared with MatchDG [32], which also introduces causality into DG problem, CIRL outperforms MatchDG by a large margin of 1.76% on ResNet-18 and 4.01% on ResNet-50. Because CIRL explicitly learns causal representations based on a more theoretical formulation, instead of in an implicit manner. There also exist cases where CIRL performs relatively poorly, this may be due to the tasks being fairly saturated in performance such as the *photo* task, or due to bad image qualities like noise samples which contain damaged causal information. However, we still achieve the second-best on these tasks and our overall performance outperforms others. In general, the encouraging results demonstrate the superiority of our causal reconstruction technique in CIRL.

Figure 4. Visualization of attention maps of the last convolutional layer on PACS dataset, with Art-Painting as the unseen target domain.
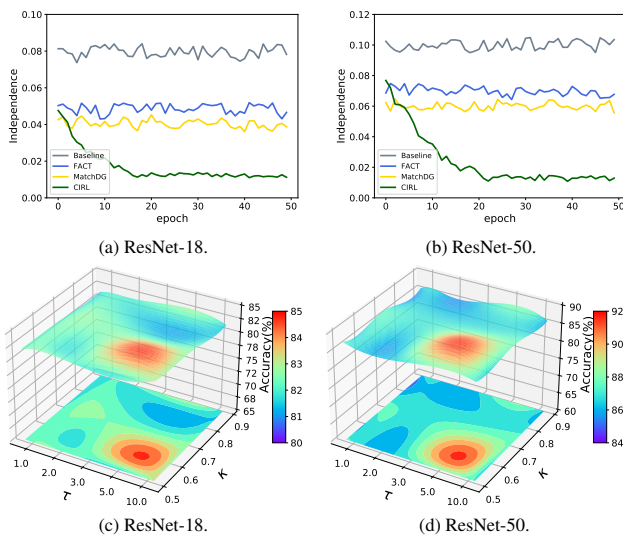

(a) ResNet-18.

(b) ResNet-50.

(c) ResNet-18.

(d) ResNet-50.

Figure 5. (a) and (b) are the Independence Degree, (c) and (d) are the sensitivity of CIRL to hyper-parameters τ and κ. All results are conducted on PACS dataset with Sketch as unseen target domain.

Table 6. Comparisons of the representation importance for different methods on PACS dataset. [Std. denotes the standard deviation.]

Methods	Mean	Std.
DeepAll [64]	3.10	0.25
FACT [63]	3.12	0.28
MatchDG [32]	3.17	0.19
CIRL (*ours*)	**3.22**	**0.09**

while the independence of CIRL increases with the training epochs and finally achieves a high degree with stabilization. The results demonstrate the efficacy of our designed causal factorization module, which allows our learned representations can be served as the mimic of causal factors.

Representation Importance. Ideally, we hope each dimension of the representations can be important that involves as much and as conducive causal information as possible, making the whole representations causally sufficient for classification. Inspired by [11], we exploit the weights

for the first layer of the classifier to estimate such importance of each representation dimension. Note that we conduct normalization on the weights for each dimension, i.e., $(x - min)/(max - min)$ to deal with the scale issue. We show the mean and standard deviation of the importance for all dimensions in Table 6 since the number of dimensions is large. It can be observed that MatchDG [32] and CIRL present higher representation importance since they embed the causal information that truly affects the classification into representations. Moreover, the highest mean value and the lowest standard deviation that CIRL achieves indicate that each dimension of our learned representations is important, which further shows the superiority of CIRL.

Parameter Sensitivity. Fig. 5c and 5d show the sensitivity of CIRL to hyper-parameters τ and κ. Specifically, the value of τ varies from $\{1.0, 2.0, 3.0, 5.0, 10.0\}$, while κ changes from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. It can be observed that CIRL achieves competitive performances robustly under a wide range of hyper-parameter values, i.e., $5.0 \le \tau \le 10.0$ and $0.5 \le \kappa \le 0.6$, with either ResNet-18 or ResNet-50 as backbone, which further verify the stability of our method.

5. Conclusion

This paper points out the insufficiency of statistical knowledge and presents a causal view of DG. The main idea is to reconstruct causal factors and excavate intrinsic causal mechanisms. We then propose a framework CIRL to learn causal representations that can mimic the causal factors based on the ideal properties we highlight. Comprehensive experiments demonstrate the effectiveness and superiority of CIRL. Considering the mainstream of related work is generally based on statistical dependence between variables, we hope our work can shed some lights into the community.

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. 3

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, pages 1006–1016, 2018. 1, 2, 7

[3] Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.*, 14(1):3207–3260, 2013. 1

[4] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2229–2238, 2019. 2, 6, 7

[5] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. Invariant rationalization. In *ICML*, pages 1448–1458, 2020. 3

[6] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, pages 6447–6458, 2019. 1, 2, 7

[7] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 6

[8] Muhammad Ghifary, David Balduzzi, W. Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE TPAMI*, 39(7):1414–1430, 2017. 2

[9] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559, 2015. 1

[10] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 3

[11] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *ICLR*, 2020. 8

[12] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Mach. Learn.*, (2):303–348, 2021. 3

[13] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 1

[14] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pages 124–140, 2020. 2, 7

[15] G. W. Imbens and D. B. Rubin. Causal inference for statistics, social, and biomedical sciences: Contents. 2015. 2

[16] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 5

[17] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Trans. Inf. Theory*, 56(10):5168–5194, 2010. 3

[18] Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278 – 2324, 1998. 6

[19] David Lewis. Causation. *The journal of philosophy, 70(17):*, pages 556–567, 1974. 2

[20] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5543–5551, 2017. 1, 6

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, pages 3490–3497, 2018. 1, 2

[22] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018. 1, 2, 6, 7

[23] Shuang Li, Fangrui Lv, Binhui Xie, Chi Harold Liu, Jian Liang, and Chen Qin. Bi-classifier determinacy maximization for unsupervised domain adaptation. In *AAAI*, pages 8455–8464, 2021. 1

[24] Shuang Li, Binhui Xie, Qiuxia Lin, Chi Harold Liu, Gao Huang, and Guoren Wang. Generalized domain conditioned adaptation network. *CoRR*, abs/2103.12339, 2021. 1

[25] 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 647–663, 2018. 1, 2

[26] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy M. Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, pages 3915–3924, 2019. 2

[27] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *ICML*, pages 6804–6814, 2021. 3

[28] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *ICML*, pages 6348–6359, 2020. 3

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, pages 136–144, 2016. 1

[30] Fangrui Lv, Jian Liang, Kaixiong Gong, Shuang Li, Chi Harold Liu, Han Li, Di Liu, and Guoren Wang. Pareto domain adaptation. In *NeurIPS*, 2021. 1

[31] Xinhong Ma, Tianzhu Zhang, and Changsheng Xu. Deep multi-modality adversarial networks for unsupervised domain adaptation. *IEEE Trans. Multim.*, 21(9):2419–2431, 2019. 1

[32] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, pages 7313–7324, 2021. 3, 7, 8

[33] Meinshausen, Nicolai, Buhlmann, Peter, Peters, and Jonas. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B. Statistical Methodology*, 2016. 3

[34] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5716–5726, 2017. 1, 2, 6, 7

[35] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18, 2013. 1, 2

[36] Yuval Netzer, Tiejie Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS*, 2011. 6

[37] H. J. Nussbaumer. *Fast Fourier Transform and Convolution Algorithms*. Springer-Verlag, 1981. 4

[38] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. *Proc IEEE*, 69(5):529–541, 1981. 4

[39] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *ICML*, pages 4033–4041, 2018. 1

[40] J. Pearl. Causality: Models, reasoning, and inference, second edition. *Cambridge University Press*, 2000. 2

[41] Xingchao Peng and Kate Saenko. Synthetic to real adaptation with generative correlation alignment networks. In *WACV*, pages 1982–1991, 2018. 2

[42] S. Peter, G. Clark, and S. Richard. Causation, prediction, and search. *The British Journal for the Philosophy of Science*, (4):4, 2000. 1, 2

[43] J. Peters, D. Janzing, and B Schölkopf. Elements of causal inference - foundations and learning algorithms. *MIT Press, Cambridge, MA, USA*, 2017. 1, 2, 3

[44] L. N. Piotrowski and F. W. Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–46, 1982. 4

[45] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, pages 7728–7738, 2020. 2, 7

[46] H. Reichenbach. *The Direction of Time.* University of California Press, Berkeley, CA,, 1956. 2, 3

[47] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard E. Turner, and Jonas Peters. Invariant models for causal transfer learning. *J. Mach. Learn. Res.*, 19:36:1–36:34, 2018. 3

[48] Bernhard Schölkopf. Causality for machine learning. *CoRR*, abs/1911.10500, 2019. 1

[49] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *ICML*, 2012. 1, 2, 3

[50] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *CoRR*, abs/2102.11107, 2021. 3

[51] B Schölkopf, D. Janzing, and D. Lopezpaz. Causal and statistical learning. 2016. 2, 3

[52] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 7

[53] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018. 1, 2, 6, 7

[54] Ram Shanmugam. Causality: Models, reasoning, and inference : Judea on causal and anticausal learning; cambridge university press, cambridge, uk, 2000, pp 384, ISBN 0-521-77362-8. *Neurocomputing*, 41(1-4):189–190, 2001. 1

[55] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *ICLR*, 2020. 2

[56] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *AISTATS*, pages 3118–3127, 2019. 3

[57] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020. 1

[58] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 6

[59] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, pages 5339–5349, 2018. 1

[60] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, volume 12354, pages 159–176, 2020. 7

[61] 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. *CoRR*, abs/2106.00925, 2021. 3

[62] Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, and Cees Snoek. A bit more bayesian: Domain-invariant learning with uncertainty. In *ICML*, pages 11351–11361, 2021. 2

[63] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021. 1, 2, 4, 6, 7, 8

[64] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 130 25–13032, 2020. 6, 7, 8

[65] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pages 561–578, 2020. 2, 4, 6, 7

[66] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 1, 2, 4