Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization

Jiaxin Qi[1(✉)], Kaihua Tang[1], Qianru Sun[2], Xian-Sheng Hua[3], and Hanwang Zhang[1]

[1] Nanyang Technological University, Singapore, Singapore
jiaxin003@e.ntu.edu.sg, {kaihua.tang,hanwangzhang}@ntu.edu.sg
[2] Singapore Management University, Singapore, Singapore
qianrusun@smu.edu.sg
[3] Damo Academy, Alibaba Group, Hangzhou, China

Abstract. Out-Of-Distribution generalization (OOD) is all about learning invariance against environmental changes. If the context (In this paper, the word "context" denotes any class-agnostic attributes such as color, texture and background. The formal definition can be found in Appendix, A.2.) in every class is evenly distributed, OOD would be trivial because the context can be easily removed due to an underlying principle: **class is invariant to context**. However, collecting such a balanced dataset is impractical. Learning on imbalanced data makes the model bias to context and thus hurts OOD. Therefore, the key to OOD is context balance. We argue that the widely adopted assumption in prior work—the context bias can be directly annotated or estimated from biased class prediction—renders the context incomplete or even incorrect. In contrast, we point out the ever-overlooked other side of the above principle: **context is also invariant to class**, which motivates us to consider the classes (which are already labeled) as the varying environments (The word "environments" [2] denotes the subsets of training data built by some criteria. In this paper, we take a class as an environment—our key idea.) to resolve context bias (without context labels). We implement this idea by minimizing the contrastive loss of intra-class sample similarity while assuring this similarity to be invariant across all classes. On benchmarks with various context biases and domain gaps, we show that a simple re-weighting based classifier equipped with our context estimation achieves state-of-the-art performance. *We provide the theoretical justifications in Appendix and codes on Github:* https://github.com/simpleshinobu/IRMCon.

1 Introduction

The gold standard for collecting a supervised training dataset of quality is to ensure the samples per class are as diverse as possible and the diversities across

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19806-9_6.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13685, pp. 92–109, 2022.
https://doi.org/10.1007/978-3-031-19806-9_6

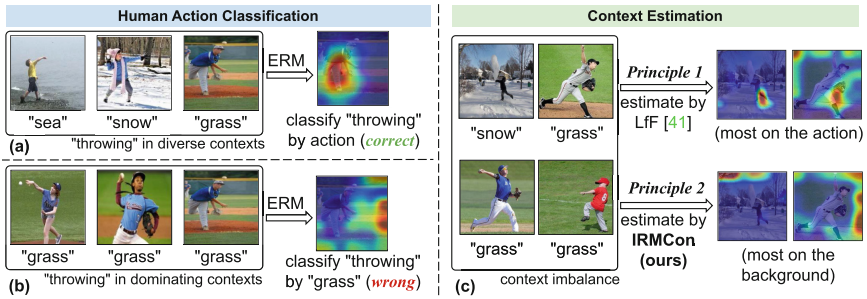

Fig. 1. GradCAM [51] visualizations of learned class and context. In (a) and (b): By using ERM, if the context is diverse and balanced within a class, the class feature is accurate—focused on the human's action; if the context dominates in the data, the class feature contains the context feature, e.g., the background "grass". In (c): The conventional context estimation [41] based on Principle 1 is biased to class (focusing on the class of human action "throwing"), while our IRMCon based on Principle 2 estimates better context (focusing on the background).

classes are as evenly distributed as possible [10,34]. For example, the "cat" class should contain cats of varying contexts, such as types, poses, and backgrounds, and the rule also applies in the "dog" class. As illustrated in Fig. 1 (a), on such a dataset, any Empirical Risk Minimization objective (ERM) [59], *e.g.*, the widely used softmax cross-entropy loss [16], can easily keep the class feature by penalizing inter-class similarities, while removing the context feature by favoring intra-class similarities. Thanks to the balanced context, the removal is clean. It can be summarized into the common principle:

Principle 1. *Class is invariant to context.*

For example, a "cat" sample is always a cat regardless of types, shapes, and backgrounds.

Given testing samples whose contexts are Out-Of-(training)Distribution (OOD), the above ERM model can still classify correctly thanks to its focus only on the context-invariant class feature[1]—model generalization emerges [17,19,33]. However in practice, due to the limited annotation budget, real-world datasets are far from the "golden" balance, and learning the class invariance on imbalanced datasets is challenging. As shown in Fig. 1 (b), if the context "grass" in class "throwing" dominates the training, the model will use the spurious correlation "most throwing actions happen in the grass" to predict "throwing". Therefore, the obstacle to OOD generalization is context imbalance.

Existing methods for context or context bias estimation fall into two categories (details in Sect. 2). First, they annotate the context directly [2,31], as shown in Fig. 2 (c). This annotation takes additional costs. Besides, it is elusive to annotate complex contexts. For example, it is easy to label the coarse scenes "water" and "grass" but hard to further tell their fine-grained differences. Thus, context supervision is usually incomplete.

[1] It is also known as causal or stable feature in literature [49,65,70].

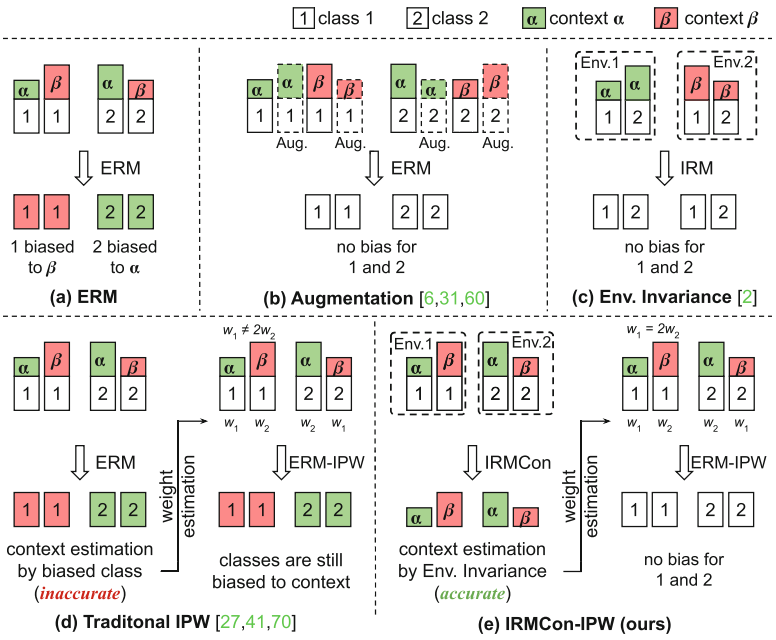

Fig. 2. Illustrations of the related approaches [2,6,27,31,41,60,70]. ERM is the baseline. Others and ours aim for mitigating context bias. The components are elaborated below. 1) The length of a context bar indicates the number of samples in that context—longer bar means the context is more prevailing. 2) A sole bar with the mixture of a color and a class number denotes the feature biased to the prevailing context. Our implementation method IRMCon-IPW is based on IRM and IPW, and our technical contribution (over the conventional methods of IRM or IPW) is the approach of disentangling context features not by using but by eliminating class features. We provide a theoretical justification in Sect. 4 and an empirical evaluation in Sect. 5.2.

Second, they estimate context bias by the biased class prediction [4,27,41], as shown in Fig. 2 (d). This relies on the contra-position of Principle 1 which is essentially an *indirect* context estimation.

Principle 1 *(Complement). If a feature is not invariant to context, it is not class but context.*

Here, the judgment of *"not invariant to context"* is implemented by using the biased prediction of a classifier, *i.e.*, if the classifier predicts wrongly, it is due to that the class invariance is not yet achieved in the classifier. Unfortunately, as the classifier is a combined effect of both class and context, it is ill-posed to disentangle if the bias is from biased context or immature class modeling. The reflection in the result is the incorrect context estimation mixed with class (see the upper part of Fig. 1 (c)). In fact, coinciding with recent findings [14,65], we show in Sect. 5 that existing methods with improper context estimation may

even under-perform the ERM baseline. In particular, if the data is less biased, such methods may catastrophically mistake context for class—this limits their applicability only in severely biased training data.

In this paper, we propose a more *direct* and accurate context estimation method without needing any context labels. Our inspiration comes from the other side of Principle 1:

Principle 2. *Context is also invariant to class.*

For example, the context "grass" is always grassy regardless of its foreground object class.

Principle 1 implies that the success of learning class invariance is due to the varying context. Similarly, Principle 2 tells us that we can learn context invariance with varying classes, and this is even easier for us to implement because the classes (taken as varying environments [2]) have been labeled and balanced—a common practice for any supervised training data with an equal sample size per class. In Sect. 4, as illustrated in Fig. 2 (e), we propose a context estimator trained by minimizing the contrastive loss of intra-class sample similarity which is invariant to classes (based on Principle 2). In particular, the invariance is achieved by Invariant Risk Minimization (IRM) [2] with our new loss term. We call our method **IRMCon** where **Con** stands for context. Figure 1 (c) illustrates that our IRMCon can capture better context feature. Based on IRMCon, we can simply deploy a re-weighting method, *e.g.*, [35], to generate the balancing weights for different contexts—context balance is achieved.

We follow DomainBed [14] for rigorous and reproducible evaluations, including 1) a strong Empirical Risk Minimization (ERM) baseline that is used to be mistakenly poor in OOD, and 2) a fair hyper-parameter tuning validation set. Experimental results in Sect. 5 demonstrate that our IRMCon can effectively learn context variance and eventually improve the context bias estimation, leading to a state-of-the-art OOD performance. Our another contribution in experiments is we propose a non-pretraining setting for OOD. It is known that many conventional experiment settings with pretraining, especially using the ImageNet [10], have data leakage issues as mentioned in related works [62,66]. We have an in-depth discussion on these issues in Sect. 5.2.

2 Related Work

OOD Tasks. Traditional machine learning heavily relies on the Independent and Identically Distributed (IID) assumption for training and testing data. Under this assumption, model generalization emerges easily [59]. However, this assumption is often violated by data distribution shift in practice—the Out-of-Distribution (OOD) problem causes the catastrophic performance drop [18,47]. In general, any test distribution unseen in training can be understood as OOD tasks, such as debiasing [8,11,24,32,63], long-tailed recognition [23,37,56], domain adaptation [5,12,58,69] and domain generalization [28,40,52]. In this

work, we focus on the most challenging one, where the distribution shift is unlabelled (*e.g.*, different from long-tailed recognition, where the shift of class distribution is known) and even unavailable (*e.g.*, different from domain adaptation, where the OOD data is available). We leave other related tasks as future work.

Invariant Feature Learning. The invariant class feature can help the model achieve robust classification when context distribution changes. The prevalent methods are: 1) *Data augmentation* [6,31,60,68]. They pre-define some augmentations for images to enlarge the available context distribution artificially. As the features are only invariant to the augmentation-related contexts, they cannot deal with other contexts out of the augmentation inventory. 2) *Context Annotation* [2,30,54]. They split data by different context annotation into environments, and penalize the model by the feature shifts among different environments. As the features are only invariant to the annotated context, the inaccurate and incomplete annotations will impact their feature invariance. 3) *Causal Learning* [39,44,46,65]. They learn the causal representations to capture the latent data generation process. Then, they can eliminate the context feature and pursue causal effect by intervention. These methods are essentially the re-weighting methods below in a causal perspective. 4) *Reweighting* [27,41,70]. They rebalance the context by re-weighting to help invariance feature learning. But, they improperly estimate the context weights by involving class learning into the context bias estimation. This inaccurate estimation problem severely influences the re-weighting and invariant feature learning. In contrast, IRMCon directly estimates the context without class prediction. The key difference is demonstrated in Fig. 2 (d) and (e): the output of our IRMCon does not contain class feature.

3 Common Pipeline: Invariance as Class

Model generalization in supervised learning is based on the fundamental assumption [20,64]: any sample x is generated from the two disentangled features (or independent causal mechanisms [55]), $x = g(\mathbf{x}_c, \mathbf{x}_t)$, where $\mathbf{x}_c$ is the class feature, $\mathbf{x}_t$ is the context feature, $g(\cdot)$ is a generative function that transforms the two features in vector space to sample space (*e.g.*, pixels). In particular, the disentanglement naturally encodes the two principles. To see this for Principle 1, if we only change the context of x and obtain a new image x', we have $\mathbf{x}_c = \mathbf{x}'_c$ but $\mathbf{x}_t \neq \mathbf{x}'_t$—class is invariant to context; Principle 2 can be interpreted in a similar way. Therefore, we'd like to learn a feature extractor $\phi_c(x) = \mathbf{x}_c$ that helps the subsequent classifier to predict robustly across varying contexts.

3.1 Empirical Risk Minimization (ERM)

If the training data per class is balanced and diverse, *i.e.*, containing sufficient samples in different contexts, ERM has been theoretically justified that it can learn the class feature extractor $\phi_c(x)$ by minimizing a contrastive based loss such as softmax cross-entropy (CE) loss [64]:

$$\mathcal{L}_{\text{ERM}}(\phi_c, f) = \frac{1}{N}\sum_{i=1}^{N} \text{CE}(y_i, \hat{y}_i = f(\phi_c(x_i))), \tag{1}$$

where y_i is the ground-truth label of x_i and $\hat{y}_i$ is the predicted label by the softmax classifier $f(\cdot)$.

However, when the data is imbalanced and less diverse, ERM cannot learn $\phi_c(x) = \mathbf{x}_c$. We illustrate this in Fig. 2 (a): if more class `1` samples contain context β than α, the resultant $\phi_c(x)$ will be biased to the prevailing context, *e.g.*, features for classifying class `1` will be entangled with context β. To this end, augmentation-based methods [6,61] aim to compensate for the imbalance (Fig. 2 (b)). However, as contexts are complex, augmentation will be far from enough to compensate for all of them.

3.2 Invariant Risk Minimization (IRM)

If context annotation is available, we can use IRM [2] to learn ϕ_c by applying Principle 1 that ϕ_c should be invariant to different contexts. Compared to ERM on balanced data that achieves invariance in a passive way via random trials [3], IRM on imbalanced data adopts the active intervention, taking contexts as the environments:

$$\mathcal{L}_{\text{IRM}}(\phi_c, \theta) = \sum_e \frac{1}{|e|} \sum_{(x_i, y_i) \in e} \left[\text{CE}(y_i, \hat{y}_i) + \lambda \|\nabla_\theta \, \text{CE}(y_i, \hat{y}_i^\theta)\|^2 \right], \tag{2}$$

where $\hat{y}_i^\theta = f(\phi_c(x_i) \cdot \theta)$, e is one of the environments of the training data according to context labels, and $\lambda > 0$ is a trade-off hyper-parameter for the invariance regularization term. θ is a dummy classifier, whose gradient is not applied to update itself but to calculate the regularization term in Eq. (2). The regularization term encourages ϕ_c to be equally optimal in different environments, *i.e.*, become invariant to environments (contexts). We follow IRM [2] to set θ as 1.

As illustrated in Fig. 2 (c), if we want to learn a common classifier that discriminates `1` and `2` in both environments, the only way is to remove the context α and β. However, it has been demonstrated by [36,65] that the context annotation is usually incomplete and using it may even under-perform ERM.

3.3 Inverse Probability Weighting (IPW)

When context annotation is unavailable, we can estimate the context and then re-balance data according to context. We begin with the following ERM-IPW loss [22,50]:

$$\mathcal{L}_{\text{ERM-IPW}}(\phi_c, \phi_t, f) = \frac{1}{N} \sum_{i=1}^{N} \text{CE}(y_i, \hat{y}_i = f(\phi_c(x_i))) \cdot \frac{1}{P(x_i \mid \phi_t(x_i))}. \tag{3}$$

We can see that the key difference between ERM-IPW and ERM is the sample-level IPW term $1/P(x_i|\phi_t(x_i))$, where $\phi_t(x) = \mathbf{x}_t$ is the context feature extractor. This IPW implies that if x is more likely associated with its context $\mathbf{x}_t$, *i.e.*, the class feature counterpart $\mathbf{x}_c$ is also more likely associated with $\mathbf{x}_t$, we should under-weight the loss because we need to discourage such a context bias.

However, the context estimation of ϕ_t is almost challenging as learning ϕ_c. Instead, a prevailing strategy is to estimate it by a biased classifier [27,41], *e.g.*,

$$P(x|\phi_t(x)) \propto \frac{\mathrm{CE}(y, \hat{y} = f(\phi_c(x))) + \mathrm{CE}(y, \hat{y} = f_b(\phi_b(x)))}{\mathrm{CE}(y, \hat{y} = f_b(\phi_b(x)))}, \tag{4}$$

where ϕ_b is the bias feature extractor and f_b is the bias classifier. ϕ_b and f_b are minimized by ERM equipped with generalized cross entropy (GCE) loss [71]:

$$\mathcal{L}_{\mathrm{ERM}}(\phi_b, f_b) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{GCE}(y_i, \hat{y}_i = f_b(\phi_b(x_i))), \tag{5}$$

where $\mathrm{GCE}(y, \hat{y}) = \sum_{k=1}^{n} y_k \cdot \frac{1-\hat{y}_k^q}{q}$ is used to amplify the bias, where q is a constant, k is the index of class and n is the class number. However, the loss in Eq. (5) inevitably includes the effect from the class feature $\mathbf{x}_c$, due to the aforementioned assumption $x = g(\mathbf{x}_c, \mathbf{x}_t)$. In other words, such a combined effect cannot distinguish whether the bias is from class or context, resulting in inaccurate context estimation. We show the illustration in Fig. 2 (d). Specifically, the weights are estimated from class and context, and thus inaccurate to balance the context. In addition, the experimental results in Fig. 6 (Bottom) testify that: inaccurate context estimation will severely hurt the performance, *i.e.*, fail to derive unbiased classifiers.

4 Our Approach: Invariance as Context

To tackle the inaccurate context estimation of $\phi_t(x)$, we propose to apply Principle 2 as a way out. As illustrated in Fig. 2 (e), if we consider each class as the environment, we can clearly see that the *unique* environmental change is the class which has been already labeled. This motivates us to apply IRM to learn invariance as context by removing the environment-equivariant class. The crux is how to design the contrastive based loss—more specifically, how to modify θ and $\mathrm{CE}(\cdot)$ in Eq. (2). The following is our novel solution.

We design a new contrastive loss based on the intra-class (environment) sample similarity, as follows,

$$\mathcal{L}_{ct}(\phi_t, e, \theta) = \sum_{x_i \in e} -log \frac{exp(\phi_t(x_i)^T \phi_t(\texttt{Aug}(x_i)) \cdot \theta)}{\sum_{x_i' \in e} exp(\phi_t(x_i)^T \phi_t(x_i') \cdot \theta)}, \tag{6}$$

where $\texttt{Aug}(\cdot)$ is the common augmentations, such as flip and Gaussian noise (used in standard contrastive losses [7,13,15]), e is the environment split by class, *e.g.*, under the environment e_1, any $x_i \in e_1$ has the class label $\texttt{1}$, θ is the dummy classifier, we add θ here for the convenience to introduce Eq. (7). The reason for using contrastive loss is that it preserves all the intrinsic features of each sample [43,64]. Yet, without the invariance to class, $\phi_t(x) \neq \mathbf{x}_t$. Then, based on Eq. (2), our proposed IRMCon for learning "invariance as context" is:

$$\mathcal{L}_{\mathrm{IRMCon}}(\phi_t, \theta) = \sum_{e} \frac{1}{|e|} [\mathcal{L}_{ct}(\phi_t, e, \theta) + \lambda |\nabla_\theta \mathcal{L}_{ct}(\phi_t, e, \theta)|], \tag{7}$$

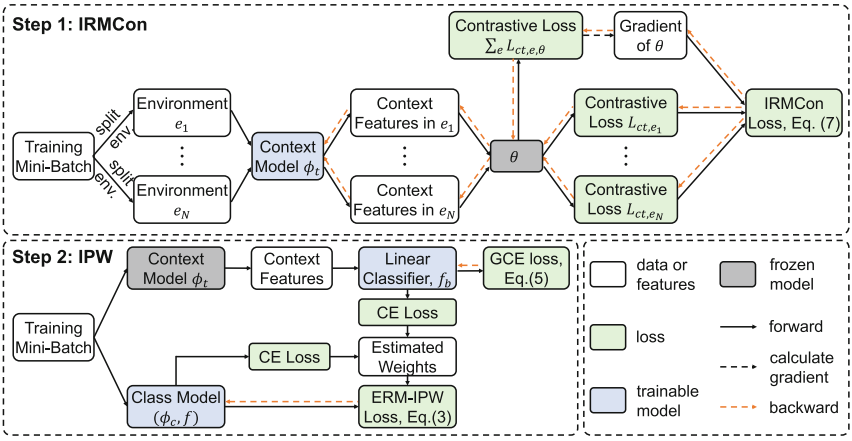

Fig. 3. The training pipeline of our IRMCon-IPW. 1) "split env." denotes we split the training samples in mini-batch into subsets based on class labels, i.e., samples of each class in one subset, forming N environments $\{e_i\}_1^N$; 2) θ is a dummy classifier, whose gradient is for regularizing ϕ_t become invariant to classes. See the detailed algorithm in Appendix

where θ plays the same role in Eq. (2), to regularize ϕ_t be invariant to environments (classes). We can prove that solving Eq. (7) achieves $\phi_t(x) = \mathbf{x}_t$, *i.e.*, the context feature is disentangled (see Appendix). As demonstrated in Fig. 4, ϕ_t can extract accurate context features. Thanks to ϕ_t, we can further improve IPW:

$$P(x|\phi_t(x)) \propto \frac{\mathrm{CE}(y, \hat{y} = f(\phi_c(x))) + \mathrm{CE}(y, \hat{y} = f_b(\mathbf{x}_t))}{\mathrm{CE}(y, \hat{y} = f_b(\mathbf{x}_t))}, \tag{8}$$

where $\mathbf{x}_t = \phi_t(x)$. We train f_b by using GCE loss, just replacing $\phi_b(x)$ with $\mathbf{x}_t$ in Eq. (5). ϕ_t is trained by IRMCon and then fixed when estimating the context.

As shown in Fig. 5, our biased classifier can estimate more accurate weights to perform better reweighting than the traditional one We streamline the proposed IRMCon-IPW in Fig. 3 and summarize our algorithm in Appendix.

5 Experiments

We introduce the benchmarks of two OOD generalization tasks, removing context bias (also called debias) and mitigating domain gaps (also called domain generalization and termed DG), and our implementation details in Sect. 5.1. Then, we evaluate the effectiveness of our approach based on the experimental results in Sect. 5.2.

5.1 Datasets and Settings

Context Biased Datasets. We follow LfF [41] to use two synthetic datasets,

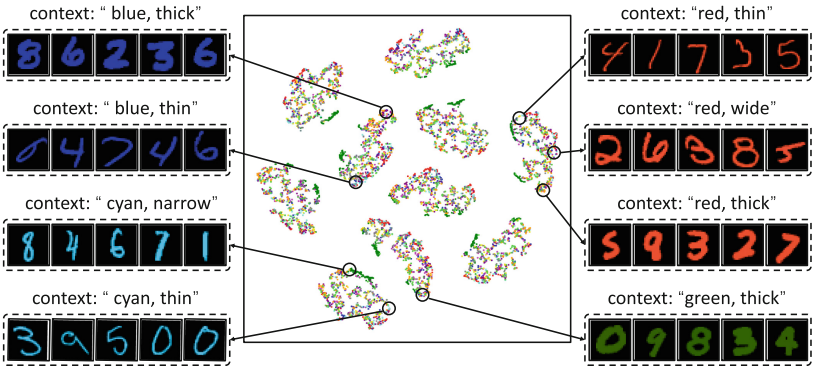

Fig. 4. t-SNE [38] visualizations of our context features of the *Colored MNIST* test samples. The color of points denotes their class labels. IRMCon is trained on the 99% biased training set. Features are naturally clustered by context. As there is no context ground-truth, the context labels are interpreted by us.

Colored MNIST and *Corrupted CIFAR-10*, and one real-world dataset, *Biased Action Recognition* (*BAR*) [41] for evaluation.

On each dataset, we manually control the context bias ratio by generating (in synthetic datasets) or sampling (in the real-world dataset) training images.

In specific, on *Colored MNIST*, we follow LfF to generate 10 colors as 10 contexts. We connect each digit (class) with a specific color and dye them with the ratio from {99.9%, 99.8%, 99.5%, 99.0%, 98.0%, 95.0%} to construct each biased training set. In the test set, 10 colors are uniformly distributed on the samples of each class. For *Corrupted CIFAR-10*, we follow LfF to use {Saturate, Elastic, Impulse, Brightness, Contrast, Gaussian, Defocus Blur, Pixelate, Gaussian Blur, Frost} as 10 contexts. Similar to *Colored MNIST*, we generate context biased training set by pairing a context and a class with a ratio chosen from {99.5%, 99.0%, 98.0%, 95.0%}. In the test set, 10 corruptions are uniformly distributed.

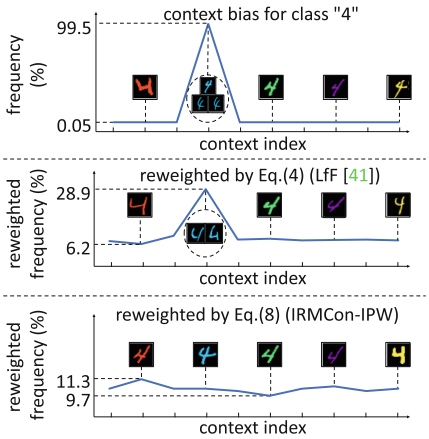

Fig. 5. Illustrations of the reweighted sample frequencies for 10 color contexts. All models are trained on the 99.5% biased *Colored MNIST*. The reweighted frequency of a context indicates the normalized sum over the inverse probabilities of the samples in this context. **Top**: Biased context distribution in the training set. **Middle**: Biased context distribution derived by using LfF [41]. **Bottom**: Relatively balanced context distribution by using our method.

The real-world dataset *BAR* contains six kinds of action-place bias, and each one is between human action and background, e.g., "throwing" always happens with the "grass" background; We choose a bias ratio in {99.0%, 95.0%}.

Domain Gap Dataset. We use *PACS* [28] to testify our method. It consists of seven object categories spanning four image domains: *Photo*, *Art-painting*, *Cartoon*, and *Sketch*. We follow DomainBed [14] to each time select three domains for training and the left one for testing. More details about datasets, *e.g.*, the number and size of the training images, are given in Appendix.

Table 1. Accuracy (%) on context biased datasets compared with SOTA methods. We reproduced the methods and averaged the results over three independent trials (mean±std). "*": For reproducing mismatch issues, performance is quoted from the original paper. Our reproduced results are reported in Appendix. "-": no report in that setting.

Dataset	Bias ratio (%)	Methods					
		ERM	Rebias [4]	EnD* [57]	LfF [41]	Feat-Aug* [27]	IRMCon-IPW (ours)
Colored MNIST	99.9	20.4 ± 1.1	20.8 ± 0.6	-	56.8 ± 1.6	-	**66.7** ± 2.3
	99.8	26.4 ± 0.4	28.3 ± 0.9	-	68.3 ± 1.5	-	**75.5** ± 1.5
	99.5	42.9 ± 1.1	44.4 ± 0.5	34.3 ± 1.2	77.0 ± 1.5	65.2 ± 4.4	**81.0** ± 0.9
	99.0	59.2 ± 0.5	58.6 ± 0.4	49.5 ± 2.5	82.5 ± 1.7	81.7 ± 2.3	**85.3** ± 0.3
	98.0	72.5 ± 0.2	73.5 ± 1.0	68.5 ± 2.2	84.1 ± 1.5	84.8 ± 1.0	**88.3** ± 0.2
	95.0	85.7 ± 0.5	85.5 ± 0.5	81.2 ± 1.4	86.8 ± 0.5	89.7 ± 1.1	**92.2** ± 0.5
Corrupted Cifar-10	99.5	22.7 ± 0.5	22.7 ± 0.7	22.9 ± 0.3	26.1 ± 0.7	30.0 ± 0.7	**31.0** ± 0.6
	99.0	25.8 ± 0.6	24.9 ± 0.7	25.5 ± 0.4	31.8 ± 0.7	36.5 ± 1.8	**37.1** ± 0.4
	98.0	28.7 ± 0.1	29.1 ± 0.7	31.3 ± 0.4	38.9 ± 1.0	41.8 ± 2.3	**42.5** ± 1.0
	95.0	39.9 ± 1.6	38.9 ± 1.7	40.3 ± 0.9	51.3 ± 0.9	51.1 ± 1.3	**53.8** ± 1.3
BAR	99.0	52.9 ± 0.7	52.1 ± 0.5	-	48.1 ± 2.7	52.3 ± 1.0	**55.3** ± 0.6
	95.0	65.2 ± 1.9	65.0 ± 1.8	-	60.6 ± 2.6	63.5 ± 1.5	**67.9** ± 0.8

Comparing Methods. As the two types of datasets have their own state-of-the-art (SOTA) methods, we compare with different SOTA methods in context biased benchmark and domain gap benchmark, respectively.

For context biased datasets, we compare with Rebias [4], End [57], LfF [41], and Feat-Aug [27]. For domain gap dataset (DG task), we compare with domain-label based methods, such as DANN [1], fish [53], and TRM [67], as well as domain-label free methods, such as RSC [21] and StableNet [70]. As we claimed at the end of Sect. 3.1, we train all models from scratch. This makes some DG methods (*e.g.*, MMD [30] and CDANN [42]) hard to converge.

Implementation Details. We first introduce two implementation details to deal with the implementation issues we met, and then provide training details. 1) *Weighted sample strategy.* This strategy is for the biased dataset. For example, under the 99.9% biased training set, in a mini-batch, all the images may have

Table 2. Accuracy (%) on the domain generalization dataset *PACS* [28]. We reproduced all the methods by the DomainBed [14] code base without pretraining. Results are averaged over 3 independent trials (mean±std). "-" denotes that methods fail to converge when training from scratch.

Methods		PACS				
		Art.	*Cartoon*	*Photo*	*Sketch*	Avg.
w/ domain supervision	IRM [2]	31.1 ± 1.4	38.7 ± 2.5	-	44.4 ± 2.2	-
	DRO [48]	39.0 ± 1.9	53.8 ± 1.2	63.6 ± 2.9	**62.4** ± 0.6	54.7
	InterMix [68]	42.2 ± 0.5	52.8 ± 1.9	61.0 ± 2.4	58.4 ± 1.0	53.6
	MLDG [29]	38.8 ± 0.7	53.5 ± 0.7	63.3 ± 0.1	60.2 ± 1.2	54.0
	DANN [1]	31.5 ± 1.1	48.2 ± 1.6	58.1 ± 1.5	44.9 ± 0.7	45.7
	V-REx [26]	33.9 ± 1.2	40.9 ± 1.2	-	55.1 ± 2.9	-
	Fish [53]	**43.1** ± 2.1	**57.4** ± 0.4	**64.8** ± 2.7	61.1 ± 0.8	**56.6**
	TRM [67]	41.8 ± 1.8	54.9 ± 0.8	-	61.3 ± 2.3	-
w/o domain supervision	ERM	40.4 ± 0.7	54.3 ± 0.3	63.7 ± 0.4	58.9 ± 2.6	54.3
	SD [45]	39.1 ± 0.8	54.4 ± 1.4	61.7 ± 3.8	51.3 ± 3.2	51.6
	RSC [21]	40.7 ± 1.1	49.8 ± 6.0	58.0 ± 1.9	53.3 ± 4.3	50.5
	LfF [41]	38.2 ± 1.4	50.4 ± 0.9	58.0 ± 0.6	60.4 ± 1.2	51.8
	IRMCon-IPW	**40.9** ± 1.7	**56.0** ± 2.9	**64.9** ± 0.7	**61.1** ± 2.5	**55.7**

the same context in a class, unless we can sample over 1,000 images per class to get 1 sample with non-biased context. To solve this issue, we use the bias model from LfF [41] to learn an inaccurate context estimator, and based on its inverse probability we sample a relative context-balanced mini-batch. This strategy frees us from sampling a very large batch to learn Eq. (6).

2) *Strategy for learning augmentation-related context.* It is hard to learn augmentation related context, when using contrastive loss. To minimize contrastive loss, the model needs to learn invariance on augmentations, *i.e.*, augmentation related features will be removed. On *Corrupted Cifar-10*, we add the classification loss in Eq. (5) to our IRMCon loss to train the context extractor. Please note that we use this strategy only for *Corrupted Cifar-10* as context on this dataset is dominated by augmentation-related context, such as 95% "car" has augmentation-related context 'Gaussian noise". Due to space limits, we put other details in Appendix.

3) *Training details.* On the *Colored MNIST*, we use 3-layers MLPs to model ϕ_c, ϕ_b and ϕ_t. On the *Corrupted Cifar-10*, we use ResNet-18 for ϕ_c and 3-layers CNNs for ϕ_b and ϕ_t. On the *BAR* and *PACS*, we use ResNet-18 for ϕ_c, ϕ_b and ϕ_t. For optimization in context biased datasets, we follow LfF [41] to use Adam [25] optimizer with the learning rate as 0.001. Other detailed settings, *e.g.*batch size, epochs, and λ in each setting, can be found in Appendix.

On all datasets, we follow DomainBed [14] to randomly split the original unbiased test set into 20% and 80% as the validation set and test set, respectively, and select the best model based on validation results. We average the results of three independent runs, and report them in the format of "mean accuracy ± standard deviation".

5.2 Results and Analyses

IRMCon-IPW Achieves SOTA. We show our results of context biased datasets in Table 1 and domain gap dataset in Table 2.

1) Table 1 presents that our IRMCon-IPW achieves very clear margins over the related methods.

In particular, the improvements are more obvious in the settings of higher bias ratios. The possible reason is when the bias ratio is higher, the "rare" context samples become less. Reweighting methods are more sensitive to the accuracy of context weights estimation. Therefore, accurate context estimation plays a more essential role. Compared to related methods, our IRMCon can estimate more accurate context, *i.e.*, extract high-quality context features like the illustration in Fig. 4, whose gain over others is more obvious when increasing the context bias ratio.

2) Table 2 presents that on the domain gap dataset, our method outperforms ERM and also achieves the best average performance over all the domain label-free methods. In addition, it achieves comparable results to the other DG methods (in the upper block) which need domain labels.

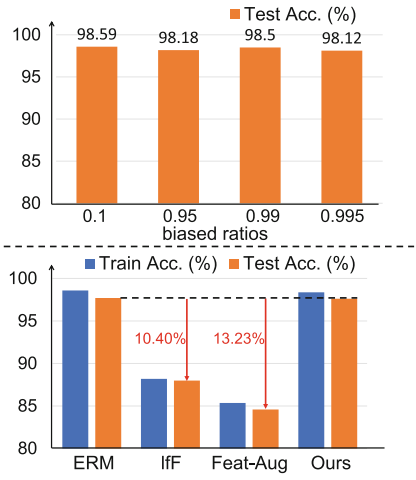

Fig. 6. Accuracy (%) of models when training on *Colored MNIST* context-balance set. **Top**: ERM is stable in test sets with varying context biases; **Bottom**: due to the incorrect context estimation, traditional reweighting methods degenerate significantly compared to ERM when training on context-balance set. Thanks to the correct context estimation, our IRMCon-IPW achieves comparable performance to ERM.

Why does ERM perform so well in most cases? On *PACS*, we follow the DomainBed [14] to implement a strong ERM baseline. On *BAR*, we use the strong augmentation strategy, Random Augmentation [9], which can be considered as an OOD method as shown in Fig. 2 (b). If we do not apply such strong augmentations, ERM performance drops significantly. We show the corresponding results in Appendix.

Why do we train models from scratch for OOD problems? We challenge the traditional pretraining settings in some OOD tasks, such as Domain Generalization, because we are concerned that the data or knowledge of the test set has been leaked to the model when pretrained on large-scale image datasets. Data leakage is a usual problem in pretraining settings, such as ImageNet [10] leaks to CUB [62]. Such problem will severely destroy the validity of the OOD task [66]. Empirically,

we provide an observation in Domain Generalization to justify our challenge. In pretraining settings, ERM achieves the "impressive" 98% test accuracy [14] when *Photo* domain is used for testing. This number is significantly higher (around 20% higher) than using *Cartoon* and *Sketch* in testing. However, this is not the case if there is no pretraining on ImageNet, see Table 2, bottom block first line, ERM method. The reason is that ImageNet, collected from the real world, leaks more real images in *Photo*, compare to artificial images in *Cartoon* and *Sketch*. Therefore, we propose the non-pretraining setting for all OOD benchmarks to prevent the leakage problem.

How to evaluate the context feature learned in IRMCon-IPW? We visualize the comparisons between the context features learned by IRMCon-IPW and LfF in Fig. 7. We show the training and test accuracies of the linear classifiers (we call bias classification heads) that are trained with context features and class labels, *i.e.*, to learn the bias intentionally. We can see from the figures that ours shows the almost same learning behavior as the upper bound case: context is invariant to class and should predict class by random chance. It means that IRMCon-IPW is able to recover the oracle distribution of contexts in the image. This can be taken as a support to the bottom illustration in Fig. 5 where using our weights can achieve a balanced context distribution—the ground truth distribution.

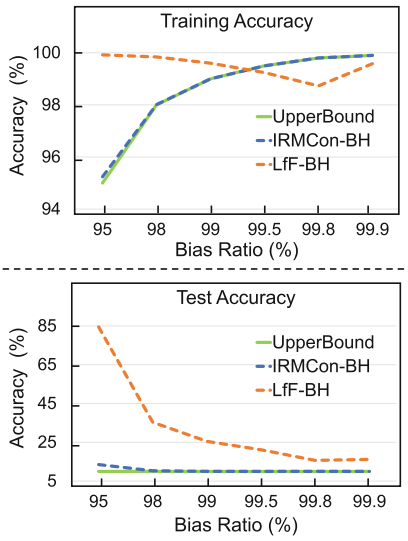

Fig. 7. Comparing the bias classification heads in LfF [41] (LfF-BH) and in ours (IRMCon-BH) on *Colored MNIST* with different bias ratios. The bias classification heads (BH) intentionally use context to predict class. Our bias head is almost the same as the upper bound case in test set—random class prediction (10%).

How does IRMCon-IPW tackle domain gap issues? Compared to the datasets with pre-defined context distribution in training (*e.g.*, set color distribution in each class in *Colored MNIST* dataset [41]), the domain gap dataset such as *PACS* does not have such explicit context settings. While it has implicit context distribution related to the domain. This distribution is often imbalanced which leads to context bias problems (similar to context biased datasets such as *BAR*). Therefore, our method

can help *PACS* to "debias". We notice that, compared to ERM, our improvement for *PACS* is not as significant as that on the context biased datasets. This might be because the context bias in *PACS* is not as severe as that in context biased datasets.

Failure Cases. We show some failure cases of our IRMCon in Fig. 8. The failure cases are selected if their IRMCon-IPW classification results are wrong. As expected, we see that the key reasons for failure are the incorrect context estimation, *e.g.*, the contexts are mixed with the foreground or wrongly attended to the foreground. By inspecting the *BAR* dataset, we find that some contexts, *e.g.*, "pool" for the class "diving", are relatively unique for certain classes. This implies that the context is NOT invariant to class. To resolve this, we conjecture that this is a dataset failure and the only way out is to bring external knowledge.

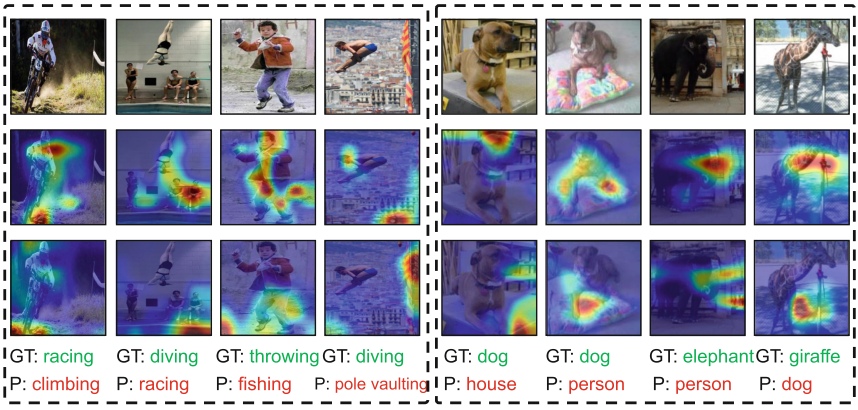

Fig. 8. GradCAM [51] visualizations of IRMCon-IPW failure cases. **Top**: input test images; **Middle**: context visualization by bias classifier of IRMCon; **Bottom**: class visualization. Left four columns are selected from *BAR* test set, the model is trained on the 99% biased training set; right four are selected from the *Photo* domain of *PACS*, model is trained on the other three domains. GT: ground-truth label; P: predicted label.

6 Conclusions

Context imbalance is the main challenge in learning class invariance for OOD generalization. Prior work tackles this challenge in two ways: 1) relying on context supervision and 2) estimating context bias by classifier failures. We showed how they fail and hence proposed a novel approach called IRM for Context (IRMCon) that directly learns the context feature without context supervision. The success of IRMCon is based on: *context is invariant to class*, which is the overlooked other side of the common principle—class is invariant to context.

Thanks to the class supervision which has been already provided as environments in training data, IRMCon can achieve context invariance by using IRM on the intra-class sample similarity contrastive loss. We used the context feature for Inverse Probability Weighting (IPW): a method for context balancing, to learn the final classifier that generalizes to OOD. IRMCon-IPW achieves state-of-the-art results on several OOD benchmarks.

Acknowledgements. This research was supported by the Alibaba-NTU Singapore Joint Research Institute (JRI), and Artificial Intelligence Singapore (AISG), Alibaba Innovative Research (AIR) programme, A*STAR under its AME YIRG Grant (Project No. A20E6c0101).

References

1. Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M.: Domain-adversarial neural networks. In: NIPS (2014)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
3. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. In: Multivariate Behavioral Research (2011)
4. Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J.: Learning de-biased representations with biased representations. In: ICML (2020)
5. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al.: Analysis of representations for domain adaptation. In: NIPS (2007)
6. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving Jigsaw puzzles. In: CVPR (2019)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
8. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: ensemble based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683 (2019)
9. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
11. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
12. Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., Schölkopf, B.: Domain adaptation with conditional transferable components. In: ICML (2016)
13. Grill, J.B., et al.: Bootstrap your own latent - a new approach to self-supervised learning. Adv. Neural. Inf. Process. Syst. **33**, 21271–21284 (2020)
14. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: ICLR (2021)

15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. He, Y., Shen, Z., Cui, P.: Towards non-IID image classification: a dataset and baselines. Pattern Recogn. **110**, 107383 (2021)
18. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
19. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
20. Higgins, I., et al.: Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230 (2018)
21. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 124–140. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_8
22. Jung, Y., Tian, J., Bareinboim, E.: Learning causal effects via weighted empirical risk minimization. In: NIPS (2020)
23. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans. Neural Netw. Learn. Syst. (2017)
24. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: training deep neural networks with biased data. In: CVPR, pp. 9012–9020 (2019)
25. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Krueger, D., et al.: Out-of-distribution generalization via risk extrapolation (rex). In: International Conference on Machine Learning (2021)
27. Lee, J., Kim, E., Lee, J., Lee, J., Choo, J.: Learning debiased representation via disentangled feature augmentation. In: NIPS (2021)
28. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5542–5550 (2017)
29. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: meta-learning for domain generalization. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
30. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR (2018)
31. Li, Y., et al.: Deep domain generalization via conditional invariant adversarial networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 647–663. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_38
32. Li, Y., Vasconcelos, N.: Repair: removing representation bias by dataset resampling. In: CVPR, pp. 9572–9581 (2019)
33. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2018)
34. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
35. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, vol. 793. Wiley, Hoboken (2019)

36. Liu, J., Hu, Z., Cui, P., Li, B., Shen, Z.: Heterogeneous risk minimization. In: ICML (2021)
37. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: CVPR (2019)
38. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
39. Mahajan, D., Tople, S., Sharma, A.: Domain generalization using causal matching. In: International Conference on Machine Learning, pp. 7313–7324. PMLR (2021)
40. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: ICML (2013)
41. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: training debiased classifier from biased classifier. In: NIPS (2020)
42. Okumura, R., Okada, M., Taniguchi, T.: Domain-adversarial and-conditional state space model for imitation learning. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2020)
43. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
44. Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) 947–1012 (2016)
45. Pezeshki, M., Kaba, S.O., Bengio, Y., Courville, A., Precup, D., Lajoie, G.: Gradient starvation: a learning proclivity in neural networks. In: NIPS (2021)
46. Pfister, N., Bühlmann, P., Peters, J.: Invariant causal prediction for sequential data. J. Am. Stat. Assoc. **114**(527), 1264–1276 (2019)
47. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: ICML (2019)
48. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. In: ICLR (2020)
49. Schölkopf, B., et al.: Toward causal representation learning. Proc. IEEE **109**(5), 612–634 (2021)
50. Seaman, S.R., Vansteelandt, S.: Introduction to double robust methods for incomplete data. Stat. Sci. Rev. J. Inst. Math. Stat. (2018)
51. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
52. Shen, Z., et al.: Towards out-of-distribution generalization: a survey. arXiv preprint arXiv:2108.13624 (2021)
53. Shi, Y., et al.: Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937 (2021)
54. Sun, B., Saenko, K.: Deep CORAL: correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 443–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_35
55. Suter, R., Miladinovic, D., Schölkopf, B., Bauer, S.: Robustly disentangled causal mechanisms: validating deep representations for interventional robustness. In: ICML (2019)
56. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: NIPS (2020)
57. Tartaglione, E., Barbano, C.A., Grangetto, M.: End: entangling and disentangling deep representations for bias correction. In: CVPR (2021)

58. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017)
59. Vapnik, V.: Principles of risk minimization for learning theory. In: Advances in Neural Information Processing Systems (1992)
60. Volpi, R., Murino, V.: Addressing model vulnerability to distributional shifts over image transformation sets. In: ICCV (2019)
61. Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: NIPS (2018)
62. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset. California Institute of Technology (2011)
63. Wang, H., He, Z., Lipton, Z.C., Xing, E.P.: Learning robust representations by projecting superficial statistics out. arXiv preprint arXiv:1903.06256 (2019)
64. Wang, T., Yue, Z., Huang, J., Sun, Q., Zhang, H.: Self-supervised learning disentangled group representation as feature. In: NIPS (2021)
65. Wang, T., Zhou, C., Sun, Q., Zhang, H.: Causal attention for unbiased visual recognition. In: ICCV (2021)
66. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. Pattern Anal. Mach. Intell. (2018)
67. Xu, Y., Jaakkola, T.: Learning representations that support robust transfer of predictors. arXiv preprint arXiv:2110.09940 (2021)
68. Yan, S., Song, H., Li, N., Zou, L., Ren, L.: Improve unsupervised domain adaptation with mixup training. arXiv preprint arXiv:2001.00677 (2020)
69. Yue, Z., Sun, Q., Hua, X.S., Zhang, H.: Transporting causal mechanisms for unsupervised domain adaptation. In: ICCV (2021)
70. Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z.: Deep stable learning for out-of-distribution generalization. In: CVPR (2021)
71. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: NIPS (2018)