OoD-Bench: Quantifying and Understanding Two Dimensions of Out-of-Distribution Generalization

Nanyang Ye[1*], Kaican Li[2*], Haoyue Bai[3†], Runpeng Yu[4†], Lanqing Hong[2], Fengwei Zhou[2], Zhenguo Li[2], Jun Zhu[4]

[1] Shanghai Jiao Tong University   [2] Huawei Noah's Ark Lab
[3] Hong Kong University of Science and Technology   [4] Tsinghua University

ynylincoln@sjtu.edu.cn, mjust.lkc@gmail.com, hbaiaa@cse.ust.hk, yrp19@mails.tsinghua.edu.cn,

{honglanqing, zhoufengwei, li.zhenguo}@huawei.com, dcszj@mail.tsinghua.edu.cn

Abstract

Deep learning has achieved tremendous success with independent and identically distributed (i.i.d.) data. However, the performance of neural networks often degenerates drastically when encountering out-of-distribution (OoD) data, i.e., when training and test data are sampled from different distributions. While a plethora of algorithms have been proposed for OoD generalization, our understanding of the data used to train and evaluate these algorithms remains stagnant. In this work, we first identify and measure two distinct kinds of distribution shifts that are ubiquitous in various datasets. Next, through extensive experiments, we compare OoD generalization algorithms across two groups of benchmarks, each dominated by one of the distribution shifts, revealing their strengths on one shift as well as limitations on the other shift. Overall, we position existing datasets and algorithms from different research areas seemingly unconnected into the same coherent picture. It may serve as a foothold that can be resorted to by future OoD generalization research. Our code is available at https://github.com/ynysjtu/ood_bench.

1. Introduction

Deep learning has been widely adopted in various applications of computer vision [32] and natural language processing [24] with great success, under the implicit assumption that the training and test data are drawn from the same distribution, which is known as the independent and identically distributed (i.i.d.) assumption. While neural networks often exhibit super-human generalization performance

on the training distribution, they can be susceptible to minute changes in the test distribution [74, 88]. This is problematic because sometimes true underlying data distributions are significantly underrepresented or misrepresented by the limited training data at hand. In the real world, such mismatches are commonly observed [28, 42], and have led to significant performance drops in many deep learning algorithms [11, 44, 55]. As a result, the reliability of current learning systems is substantially undermined in critical applications such as medical imaging [4, 20], autonomous driving [7, 22, 56, 80, 92], and security systems [37].

Out-of-Distribution (OoD) Generalization, the task of generalizing under such distribution shifts, has been fragmentarily researched in different areas, such as Domain Generalization (DG) [17, 59, 95, 106], Causal Inference [67, 69], and Stable Learning [104]. In the setting of OoD generalization, models usually have access to multiple training datasets of the same task collected in different environments. The goal of OoD generalization algorithms is to learn from these different but related training environments and then extrapolate to unseen test environments [8, 82]. Driven by this motivation, numerous algorithms have been proposed over the years [106], each claimed to have surpassed all its precedents on a particular genre of benchmarks. However, a recent work [31] suggests that the progress made by these algorithms might have been overestimated—most of the advanced learning algorithms tailor-made for OoD generalization are still on par with the classic Empirical Risk Minimization (ERM) [90].

In this work, we provide a quantification for the distribution shift exhibited in OoD datasets from different research areas and evaluate the effectiveness of OoD generalization algorithms on these datasets, revealing a possible reason as to why these algorithms appear to be no much better than ERM, which is left unexplained in previous work [31]. We find that incumbent datasets ex-

---

*[*]Nanyang Ye and Kaican Li are the joint first authors. Nanyang Ye is the corresponding author.

[†]Work is done during Haoyue and Runpeng's internships at Shanghai Jiao Tong University.

	PACS	Camelyon	ImageNet-V2	NICO	Colored MNIST
Training					
Test					

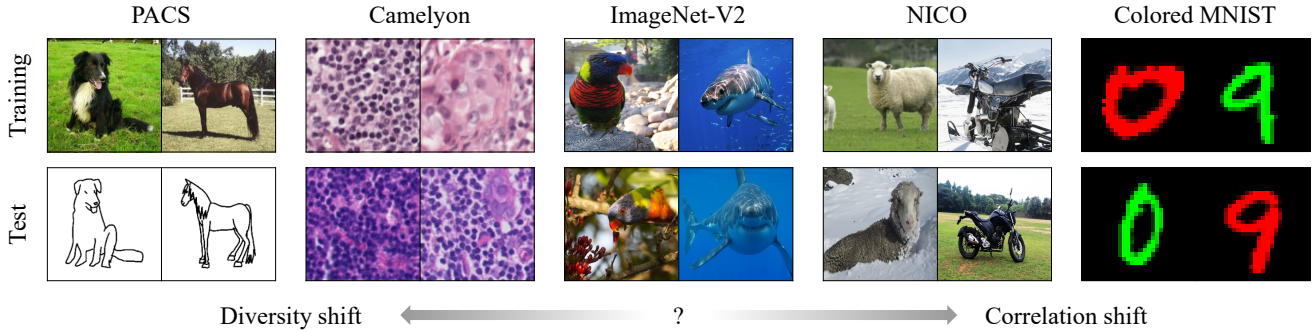Diversity shift ⟵——————— ? ———————⟶ Correlation shift

Figure 1. Examples of image classification datasets demonstrating different kinds of distribution shifts. While it is clear that the datasets at both ends exhibit apparent distribution shifts, in the middle, it is hard to distinguish the differences in distribution between the training dataset and the test dataset (*e.g.*, ImageNet [23] and ImageNet-V2 [74]), which represent a large body of realistic OoD datasets. This motivates us to quantify the distribution shifts in these OoD datasets.

hibiting distribution shifts can be generally divided into two categories of different characteristics, whereas the majority of the algorithms are only able to surpass ERM in at most one of the categories. We hypothesize that the phenomenon is due to the influence of two distinct kinds of distribution shift, namely *diversity shift* and *correlation shift*, while pre-existing literature often focuses on merely one of them. The delineation of diversity and correlation shift provides us with a unified picture for understanding distribution shifts. Based on the findings and analysis in this work, we make three recommendations for future OoD generalization research:

- Evaluate OoD generalization algorithms comprehensively on two types of datasets, one dominated by diversity shift and the other dominated by correlation shift. We provide a method to estimate the strength of these two distribution shifts on any labeled dataset.

- Investigate the nature of distribution shift in OoD problems before designing algorithms since the optimal treatment for different kinds of distribution shift may be different.

- Design large-scale datasets that more subtly capture real-world distribution shifts as imperceptible distribution shifts can also be conspicuous to neural networks.

2. Diversity Shift and Correlation Shift

Normally, datasets such as VLCS [89] and PACS [46] that consist of multiple domains are used to train and evaluate DG models. In these datasets, each domain represents a certain spectrum of diversity in data.[1] During experiments, these domains are further grouped into training and test domains, leading to the diversity shift. Although extensive research

---

[1] An implicit assumption of DG over the years is that each domain is *distinct* from one another, which is a major difference between DG and OoD generalization as the latter considers a more general setting.

efforts have been dedicated to the datasets dominated by diversity shift, it is not until recently that [9] draws attention to another challenging generalization problem stemmed from spurious correlations. Colored MNIST, a variant of MNIST [45], is constructed by coloring the digits with either red or green to highlight the problem. The colored digits are arranged into training and test environments such that the labels and colors are strongly correlated, but the correlation flips across the environments, creating correlation shift.

As shown in Figure 1, diversity and correlation shift are of clearly different nature. At the extremes, discrepancies between training and test environments become so apparent, causing great troubles for algorithms trying to generalize [9, 46]. Interestingly, in some real-world cases such as ImageNet versus ImageNet-V2 where the discrepancy is virtually imperceptible, neural networks are still unable to generalize satisfactorily, of which the reason is not fully understood [74]. In Figure 3, our estimation of diversity and correlation shift shed some light on the issue—there is a non-trivial degree of correlation shift between the original ImageNet and the variant. Besides, other OoD datasets have also shown varying degrees of diversity and correlation shift.

Formal Analysis. In the setting of supervised learning, every input $\mathbf{x} \in \mathcal{X}$ is assigned with a label $\mathbf{y} \in \mathcal{Y}$ by some fixed labeling rule $f : \mathcal{X} \to \mathcal{Y}$. The inner mechanism of f usually depends on a particular set of features $\mathcal{Z}_1$, whereas the rest of the features $\mathcal{Z}_2$ are not causal to prediction. For example, we assign the label "airplane" to the image of an airplane regardless of its color or whether it is landed or flying. The causal graph in Figure 2a depicts the interplay among the underlying random variables of our model: the input variable X is determined by the latent variables Z_1 and Z_2, whereas the target variable Y is determined by Z_1 alone. Similar graphs can be found in [2, 53, 57, 63].

Given a labeled dataset, consider its training environment
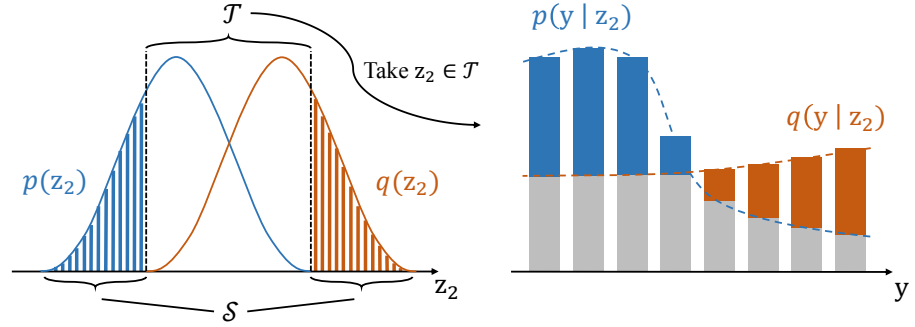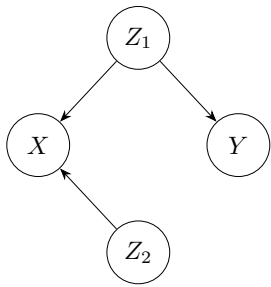
(a) Causal graph depicting the causal influence among the concerned variables.

(b) Illustration of diversity shift and correlation shift. Diversity shift amounts to half of the summed area of the colored regions in the left figure. Correlation shift is an integral over $\mathcal{T}$, where every integrand can be seen as the summed heights of the colored bars in the right figure then weighted by the square root of $p(\mathbf{z}_2) \cdot q(\mathbf{z}_2)$.

Figure 2. Explanatory illustrations for diversity and correlation shift. Diversity shift is defined by the support set's difference of the latent environment's distribution while correlation shift is defined by the probability density function's difference on the same support set.

$\mathcal{E}_{\text{tr}}$ and test environment $\mathcal{E}_{\text{te}}$ as distributions with probability functions p and q respectively. For ease of exposition, we assume no label shift [10] across the environments, i.e. $p(\mathbf{y}) = q(\mathbf{y})$ for every $\mathbf{y} \in \mathcal{Y}$.[2] Without loss of generality, we further assume that $\mathcal{E}_{\text{tr}}$ and $\mathcal{E}_{\text{te}}$ share the same labeling rule f, complementing the causal graph. To put it in the language of *causality* [66], it means that the *direct cause* of Y (which is Z_1) is observable in both environments and the causal mechanism that Z_1 exerts on Y is stable at all times. Formally, it dictates the following property for every $\mathbf{z} \in \mathcal{Z}_1$:

$$p(\mathbf{z}) \cdot q(\mathbf{z}) \neq 0 \ \wedge \ \forall \mathbf{y} \in \mathcal{Y} : p(\mathbf{y} \,|\, \mathbf{z}) = q(\mathbf{y} \,|\, \mathbf{z}). \quad (1)$$

The existence of such invariant features makes OoD generalization possible. On the other hand, the presence of $\mathbf{z} \in \mathcal{Z}_2$ possessing the opposite property,

$$p(\mathbf{z}) \cdot q(\mathbf{z}) = 0 \ \vee \ \exists \mathbf{y} \in \mathcal{Y} : p(\mathbf{y} \,|\, \mathbf{z}) \neq q(\mathbf{y} \,|\, \mathbf{z}), \quad (2)$$

makes OoD generalization challenging. From (2), we can see that $\mathcal{Z}_2$ consists of two kinds of features. *Intuitively, diversity shift stems from the first kind of features in $\mathcal{Z}_2$ since the diversity of data is embodied by novel features not shared by the environments; whereas correlation shift is caused by the second kind of features in $\mathcal{Z}_2$ which is spuriously correlated with some $\mathbf{y}$.* Based on this intuition, we partition $\mathcal{Z}_2$ into two subsets,

$$\begin{aligned} \mathcal{S} &:= \{\mathbf{z} \in \mathcal{Z}_2 \mid p(\mathbf{z}) \cdot q(\mathbf{z}) = 0\}, \\ \mathcal{T} &:= \{\mathbf{z} \in \mathcal{Z}_2 \mid p(\mathbf{z}) \cdot q(\mathbf{z}) \neq 0\}, \end{aligned} \quad (3)$$

that are respectively responsible for diversity shift and correlation shift between the environments. We then define the quantification formula of the two shifts as follows:

---

[2]As a side note, by this assumption we do *not* ignore the existence of label shift in datasets. In practice, datasets with label shift can be made to satisfy this assumption by techniques such as sample reweighting.

Definition 1 (**Diversity Shift and Correlation Shift**). *Given the two sets of features $\mathcal{S}$ and $\mathcal{T}$ defined in (3), the proposed quantification formula of diversity shift and correlation shift between two data distributions p and q is given by*

$$D_{\text{div}}(p, q) := \frac{1}{2} \int_{\mathcal{S}} |p(\mathbf{z}) - q(\mathbf{z})| \, d\mathbf{z},$$

$$D_{\text{cor}}(p, q) := \frac{1}{2} \int_{\mathcal{T}} \sqrt{p(\mathbf{z}) \cdot q(\mathbf{z})} \sum_{\mathbf{y} \in \mathcal{Y}} |p(\mathbf{y} \,|\, \mathbf{z}) - q(\mathbf{y} \,|\, \mathbf{z})| \, d\mathbf{z},$$

where we assume $\mathcal{Y}$ to be discrete.

Figure 2b illustrates the above definition when $\mathbf{z}$ is unidimensional. It can be proved that D_{div} and D_{cor} are always bounded within $[0, 1]$ (see **Proposition 1** in Appendix B). In particular, the square root in the formulation of correlation shift serves as a coefficient regulating the integrand because features that hardly appear in either environment should have a small contribution to the correlation shift overall. Nevertheless, we are aware that these are not the only viable formulations, yet they produce intuitively reasonable and numerically stable results even when estimated by a simple method described next.

Practical estimation. Given a dataset sampled from $\mathcal{E}_{\text{tr}}$ and another dataset (of equal size) sampled from $\mathcal{E}_{\text{te}}$, a neural network is first trained to discriminate the environments. The network consists of a feature extractor $g : \mathcal{X} \to \mathcal{F}$ and a classifier $h : \mathcal{F} \times \mathcal{Y} \to [0, 1]$, where $\mathcal{F}$ is some learned representation of $\mathcal{X}$. The mapping induces two joint distributions over $\mathcal{X} \times \mathcal{Y} \times \mathcal{F}$, one for each environment, with probability functions denoted by $\hat{p}$ and $\hat{q}$. For every example from either $\mathcal{E}_{\text{tr}}$ or $\mathcal{E}_{\text{te}}$, the network tries to tell which environment the example is actually sampled from, in order to minimize the following objective:

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{E}_{\text{tr}}} \ell(\hat{e}_{\mathbf{x}, \mathbf{y}}, 0) + \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{E}_{\text{te}}} \ell(\hat{e}_{\mathbf{x}, \mathbf{y}}, 1), \quad (4)$$
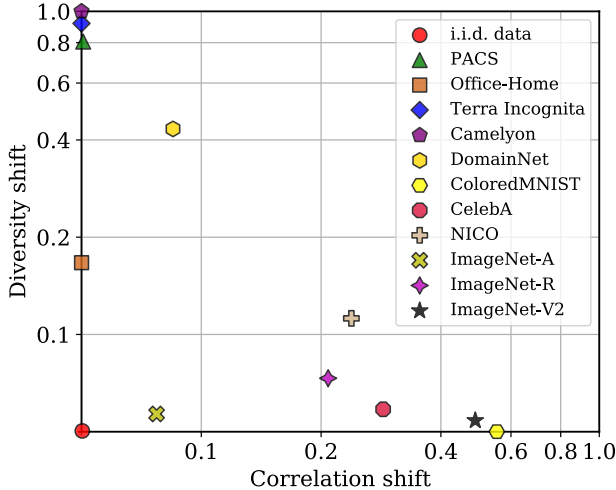
Figure 3. Estimation of diversity and correlation shift in various datasets. For ImageNet variants, the estimates are computed with respect to the original ImageNet. See Appendix F for the results in numeric form with error bars.

where $\hat{e}_{\mathbf{x},\mathbf{y}} = h(g(\mathbf{x}), \mathbf{y})$ is the predicted environment and ℓ is some loss function. The objective forces g to extract those features whose joint distribution with Y varies across the environments so that h could make reasonably accurate predictions. This is formalized by the theorem below.

Theorem 1. *The classification accuracy of a network trained to discriminate two different environments is bounded above by $\frac{1}{2} \int_{\mathcal{X}} \max\{p(\mathbf{x}), q(\mathbf{x})\}$, as the data size tends to infinity. This optimal performance is attained only when the following condition holds: for every $\mathbf{x} \in \mathcal{X}$ that is* not *i.i.d. in the two environments, i.e. $p(\mathbf{x}) \neq q(\mathbf{x})$, there exists some $\mathbf{y} \in \mathcal{Y}$ such that $\hat{p}(\mathbf{y}, \mathbf{z}) \neq \hat{q}(\mathbf{y}, \mathbf{z})$ where $\mathbf{z} = g(\mathbf{x})$.*

The proof is provided in Appendix B. After obtaining the features $\mathcal{F}$ extracted by g, we use Kernel Density Estimation (KDE) [65,76] to estimate $\hat{p}$ and $\hat{q}$ over $\mathcal{F}$. Subsequently, $\mathcal{F}$ is partitioned by whether $\hat{p}(\mathbf{z}) \cdot \hat{q}(\mathbf{z})$ is close to zero, in correspondence to $\mathcal{S}$ and $\mathcal{T}$, into two sets of features that are responsible for diversity and correlation shift respectively. The integrals in Definition 1 are then approximated by Monte Carlo Integration under importance sampling [61]. A caveat in evaluating the term $|p(\mathbf{y} \mid \mathbf{z}) - q(\mathbf{y} \mid \mathbf{z})|$ in $D_{\text{cor}}(p, q)$ is that the conditional probabilities are computationally intractable for $\mathbf{z}$ is continuous. Instead, the term is computed by the following equivalent formula as an application of Bayes' theorem:

$$\left| \frac{\hat{p}(\mathbf{y}) \cdot \hat{p}(\mathbf{z} \mid \mathbf{y})}{\hat{p}(\mathbf{z})} - \frac{\hat{q}(\mathbf{y}) \cdot \hat{q}(\mathbf{z} \mid \mathbf{y})}{\hat{q}(\mathbf{z})} \right|, \qquad (5)$$

where $\hat{p}(\mathbf{z} \mid \mathbf{y})$ and $\hat{q}(\mathbf{z} \mid \mathbf{y})$ can be approximated individually for every $\mathbf{y} \in \mathcal{Y}$ again by KDE. See Appendix C for more

details of our method including pseudo codes of the whole procedure.

We have also shown that in theory the extracted features would converge to a unique solution as the network width grows to infinity using Neural Tangent Kernel [39]. It suggests that as long as the network has sufficient capacity, we can always obtain similar results within a small error bound. To empirically verify this, we have also experimented with different network architectures which demonstrates the stability of our estimation (see Appendix E).

The results in Figure 3 are obtained by the aforementioned method. Most of the existing OoD datasets lie over or near the axes, dominated by one kind of shift. For datasets under unknown distribution shift such as ImageNet-A [35], ImageNet-R [34], and ImageNet-V2, our method successfully decomposes the shift into the two dimensions of diversity and correlation, and therefore one may choose the appropriate algorithms based on the estimation. As shown by our benchmark results in the next section, such choices might be crucial as most OoD generalization algorithms do not perform equally well on two groups of datasets, one dominated by diversity shift and the other dominated by correlation shift.

3. Experiment

Previously, we have numerically positioned OoD datasets in the two dimensions of distribution shift. In this section, we run algorithms on these datasets to reveal the two-dimensional trend for existing datasets and algorithms. All experiments are conducted on Pytorch 1.4 with Tesla V100 GPUs. Our code for the following benchmark experiments is modified from the DomainBed [31] code suite.

3.1. Benchmark

Datasets. In our experiment, datasets are chosen to cover as much variety from different OoD research areas as possible. As mentioned earlier, the datasets demonstrated two-dimensional properties shown by their estimated diversity and correlation shift. The following datasets are dominated by diversity shift: **PACS** [46], **OfficeHome** [91], **Terra Incognita** [14], and **Camelyon17-WILDS** [42]. On the other hand, our benchmark also include three datasets dominated by correlation shift: **Colored MNIST** [9], **NICO** [33], and a modified version of **CelebA** [54]. See Appendix G for more detailed descriptions of the above datasets.

For PACS, OfficeHome, and Terra Incognita, we train multiple models in every run with each treating one of the domains as the test environment and the rest of the domains as the training environments since it is common practice for DG datasets. The final accuracy is the mean accuracy over all such splits. For other datasets, the training and test environments are fixed. A reason is that the leave-one-domain-out evaluation scheme would destroy the designated

Algorithm	PACS	OfficeHome	TerraInc	Camelyon17	Average	Ranking score
RSC [38]	$82.8 \pm 0.4^{\uparrow}$	$62.9 \pm 0.4^{\downarrow}$	$43.6 \pm 0.5^{\uparrow}$	$94.9 \pm 0.2^{\uparrow}$	71.1	+2
MMD [48]	$81.7 \pm 0.2^{\uparrow}$	$63.8 \pm 0.1^{\uparrow}$	$38.3 \pm 0.4^{\downarrow}$	$94.9 \pm 0.4^{\uparrow}$	69.7	+2
SagNet [60]	$81.6 \pm 0.4^{\uparrow}$	$62.7 \pm 0.4^{\downarrow}$	42.3 ± 0.7	$95.0 \pm 0.2^{\uparrow}$	70.4	+1
ERM [90]	81.5 ± 0.0	63.3 ± 0.2	42.6 ± 0.9	94.7 ± 0.1	70.5	0
IGA [43]	$80.9 \pm 0.4^{\downarrow}$	$63.6 \pm 0.2^{\uparrow}$	$41.3 \pm 0.8^{\downarrow}$	$95.1 \pm 0.1^{\uparrow}$	70.2	0
CORAL [85]	$81.6 \pm 0.6^{\uparrow}$	$63.8 \pm 0.3^{\uparrow}$	$38.3 \pm 0.7^{\downarrow}$	$94.2 \pm 0.3^{\downarrow}$	69.5	0
IRM [9]	$81.1 \pm 0.3^{\downarrow}$	$63.0 \pm 0.2^{\downarrow}$	42.0 ± 1.8	$95.0 \pm 0.4^{\uparrow}$	70.3	−1
VREx [44]	$81.8 \pm 0.1^{\uparrow}$	63.5 ± 0.1	$40.7 \pm 0.7^{\downarrow}$	$94.1 \pm 0.3^{\downarrow}$	70.0	−1
GroupDRO [79]	$80.4 \pm 0.3^{\downarrow}$	63.2 ± 0.2	$36.8 \pm 1.1^{\downarrow}$	$95.2 \pm 0.2^{\uparrow}$	68.9	−1
ERDG [105]	$80.5 \pm 0.5^{\downarrow}$	$63.0 \pm 0.4^{\downarrow}$	$41.3 \pm 1.2^{\downarrow}$	$95.5 \pm 0.2^{\uparrow}$	70.1	−2
DANN [27]	$81.1 \pm 0.4^{\downarrow}$	$62.9 \pm 0.6^{\downarrow}$	$39.5 \pm 0.2^{\downarrow}$	$94.9 \pm 0.0^{\uparrow}$	69.6	−2
MTL [16]	$81.2 \pm 0.4^{\downarrow}$	$62.9 \pm 0.2^{\downarrow}$	$38.9 \pm 0.6^{\downarrow}$	$95.0 \pm 0.1^{\uparrow}$	69.5	−2
Mixup [101]	$79.8 \pm 0.6^{\downarrow}$	63.3 ± 0.5	$39.8 \pm 0.3^{\downarrow}$	94.6 ± 0.3	69.4	−2
ANDMask [64]	$79.5 \pm 0.0^{\downarrow}$	$62.0 \pm 0.3^{\downarrow}$	$39.8 \pm 1.4^{\downarrow}$	$95.3 \pm 0.1^{\uparrow}$	69.2	−2
ARM [103]	$81.0 \pm 0.4^{\downarrow}$	63.2 ± 0.2	$39.4 \pm 0.7^{\downarrow}$	$93.5 \pm 0.6^{\downarrow}$	69.3	−3
MLDG [47]	$73.0 \pm 0.4^{\downarrow}$	$52.4 \pm 0.2^{\downarrow}$	$27.4 \pm 2.0^{\downarrow}$	$91.2 \pm 0.4^{\downarrow}$	61.0	−4
Average	80.7	62.5	39.8	94.6	69.4	−

Table 1. Performance of ERM and OoD generalization algorithms on datasets *dominated by diversity shift*. Every symbol ↓ denotes a score of -1, and every symbol ↑ denotes a score of $+1$; otherwise the score is 0. Adding up the scores across all datasets produces the ranking score for each algorithm.

training/test splits of these datasets. For more details about dataset statistics and environment splits, see Appendix G.

Algorithms. We have selected Empirical Risk Minimization (**ERM**) [90] and several representative algorithms from different OoD research areas for our benchmark: Group Distributionally Robust Optimization (**GroupDRO**) [79], Inter-domain Mixup (**Mixup**) [100, 101], Meta-Learning for Domain Generalization (**MLDG**) [47], Domain-Adversarial Neural Networks (**DANN**) [27], Deep Correlation Alignment (**CORAL**) [85], Maximum Mean Discrepancy (**MMD**) [48], Invariant Risk Minimization (**IRM**) [9], Variance Risk Extrapolation (**VREx**) [44], Adaptive Risk Minimization (**ARM**) [103], Marginal Transfer Learning (**MTL**) [16], Style-Agnostic Networks (**SagNet**) [60], Representation Self Challenging (**RSC**) [38], Learning Explanations that are Hard to Vary (**ANDMask**) [64], Out-of-Distribution Generalization with Maximal Invariant Predictor (**IGA**) [43], and Entropy Regularization for Domain Generalization (**ERDG**) [105].

Model selection methods. As there is still no consensus on what model selection methods should be used in OoD generalization research [31], appropriate selection methods are chosen for each dataset in our study. To be consistent with existing lines of work [19, 38, 44, 46, 60], models trained on PACS, OfficeHome, and Terra Incognita are selected by *training-domain validation*. As for Camelyon17-WILDS and NICO, *OoD validation* is adopted in respect of [42]

and [12]. The two remaining datasets, Colored MNIST and CelebA, use *test-domain validation* which has been seen in [1, 9, 44, 70]. Another reason for using test-domain validation is that it may be improper to apply training-domain validation to datasets dominated by correlation shift since under the influence of spurious correlations, achieving excessively high accuracy in the training environments often leads to low accuracy in novel test environments. More detailed explanations of these model selection methods are provided in Appendix H.

Implementation details. Unlike DomainBed, we use a simpler model, ResNet-18 [32], for all algorithms and datasets excluding Colored MNIST, as it is the common practice in previous works [19, 25, 38, 60, 105]. Moreover, we believe smaller models could enlarge the gaps in OoD generalization performance among the algorithms, as larger models are generally more robust to OoD data [34] and thus the performance is easier to saturate on small datasets. The ResNet-18 is pretrained on ImageNet and then finetuned on each dataset with only one exception—NICO, which contains photos of animals and vehicles largely overlapped with ImageNet classes. For simplicity, we continue to use a two-layer perceptron following [9, 44, 70] for Colored MNIST. Our experiments further differs from DomainBed in several minor aspects. First, we do not freeze any batch normalization layer in ResNet-18, nor do we use any dropout, to be consistent with most of prior works in DG. Second, we use a larger portion (90%) of data from training environments for

Algorithm	Colored MNIST	CelebA	NICO	Average	Prev score	Ranking score
VREx [44]	$56.3 \pm 1.9^{\uparrow}$	87.3 ± 0.2	71.5 ± 2.3	71.7	−1	+1
GroupDRO [79]	$32.5 \pm 0.2^{\uparrow}$	87.5 ± 1.1	71.0 ± 0.4	63.7	−1	+1
ERM [90]	29.9 ± 0.9	87.2 ± 0.6	72.1 ± 1.6	63.1	0	0
IRM [9]	$60.2 \pm 2.4^{\uparrow}$	$85.4 \pm 1.2^{\downarrow}$	73.3 ± 2.1	73.0	−1	0
MTL [16]	29.3 ± 0.1	87.0 ± 0.7	70.6 ± 0.8	62.3	−2	0
ERDG [105]	$31.6 \pm 1.3^{\uparrow}$	$84.5 \pm 0.2^{\downarrow}$	72.7 ± 1.9	62.9	−2	0
ARM [103]	$34.6 \pm 1.8^{\uparrow}$	86.6 ± 0.7	$67.3 \pm 0.2^{\downarrow}$	62.8	−3	0
MMD [48]	$50.7 \pm 0.1^{\uparrow}$	$86.0 \pm 0.5^{\downarrow}$	$68.9 \pm 1.2^{\downarrow}$	68.5	+2	−1
RSC [38]	$28.6 \pm 1.5^{\downarrow}$	$85.9 \pm 0.2^{\downarrow}$	$74.3 \pm 1.9^{\uparrow}$	61.4	+2	−1
IGA [43]	29.7 ± 0.5	$86.2 \pm 0.7^{\downarrow}$	71.0 ± 0.1	62.3	0	−1
CORAL [85]	30.0 ± 0.5	$86.3 \pm 0.5^{\downarrow}$	70.8 ± 1.0	61.5	−1	−1
Mixup [101]	$27.6 \pm 1.8^{\downarrow}$	87.5 ± 0.5	72.5 ± 1.5	60.6	−2	−1
MLDG [47]	$32.7 \pm 1.1^{\uparrow}$	$85.4 \pm 1.3^{\downarrow}$	$66.6 \pm 2.4^{\downarrow}$	56.6	−4	−1
SagNet [60]	30.5 ± 0.7	$85.8 \pm 1.4^{\downarrow}$	$69.8 \pm 0.7^{\downarrow}$	62.0	+1	−2
ANDMask [64]	$27.2 \pm 1.4^{\downarrow}$	$86.2 \pm 0.2^{\downarrow}$	71.2 ± 0.8	61.5	−2	−2
DANN [27]	$24.5 \pm 0.8^{\downarrow}$	$86.0 \pm 0.4^{\downarrow}$	$69.4 \pm 1.7^{\downarrow}$	59.7	−2	−3
Average	34.5	86.4	70.8	63.7	–	–

Table 2. Performance of ERM and OoD generalization algorithms on datasets *dominated by correlation shift*. Every symbol ↓ denotes a score of −1, and every symbol ↑ denotes a score of +1; otherwise the score is 0. Adding up the scores across all datasets produces the ranking score for each algorithm. Prev scores are the scores of corresponding algorithms in Table 1.

training and the rest for validation. Third, we use a slightly different data augmentation scheme following [19].

Finally, we adopt the following hyperparameter search protocol, the same as in DomainBed: a 20-times random search is conducted for every pair of dataset and algorithm, and then the search process is repeated for another two random series of hyperparameter combinations, weight initialization, and dataset splits. Altogether, the three series yield the three best accuracies over which a mean and standard error bar is computed for every dataset-algorithm pair. See Appendix J for the hyperparameter search space for every individual algorithm.

Results. The benchmark results are shown in Tab. 1 and Tab. 2. In addition to mean accuracy and standard error bar, we report a ranking score for each algorithm with respect to ERM. For every dataset-algorithm pair, depending on whether the attained accuracy is lower than, within, or higher than the standard error bar of ERM accuracy on the same dataset, we assign score −1, 0, +1 to the pair. Adding up the scores across all datasets produces the ranking score for each algorithm. We underline that the ranking score does not indicate whether an algorithm is definitely better or worse than the other algorithms. It only reflects a relative degree of robustness against diversity and correlation shift.

From Tab. 1 and Tab. 2, we observe that none of the OoD generalization algorithms achieves consistently better performance than ERM on both OoD directions. For example, on the datasets dominated by diversity shift, the

ranking scores of RSC, MMD, and SagNet are higher than ERM, whereas on the datasets dominated by correlation shift, their scores are lower. Conversely, the algorithms (VREx and GroupDRO) that outperform ERM in Tab. 2 are worse than ERM on datasets of the other kind. This supports our view that *OoD generalization algorithms should be evaluated on datasets embodying both diversity and correlation shift*. Such a comprehensive evaluation is of great importance because real-world data could be tainted by both kinds of distribution shift, *e.g.*, the ImageNet variants in Figure 3.

In the toy case of Colored MNIST, several algorithms are superior to ERM, however, in the more realistic and complicated cases of CelebA and NICO, none of the algorithms surpasses ERM by a large margin. Hence, we argue that *contemporary OoD generalization algorithms are by large still vulnerable to spurious correlations*. In particular, IRM that achieves the best accuracy on Colored MNIST among all algorithms, fails to surpass ERM on the other two datasets. It is in line with the theoretical results discovered by [77]: IRM does not improve over ERM unless the test data are sufficiently similar to the training distribution. Besides, we have also done some experiments on ImageNet-V2, and the result again supports our argument (see Appendix I).

Due to inevitable noises and other changing factors in the chosen datasets and training process, whether there is any compelling pattern in the results across the datasets dominated by the same kind of distribution shift is unclear. So, it is important to point out that the magnitude of diversity and correlation shift does not indicate the absolute level of diffi-
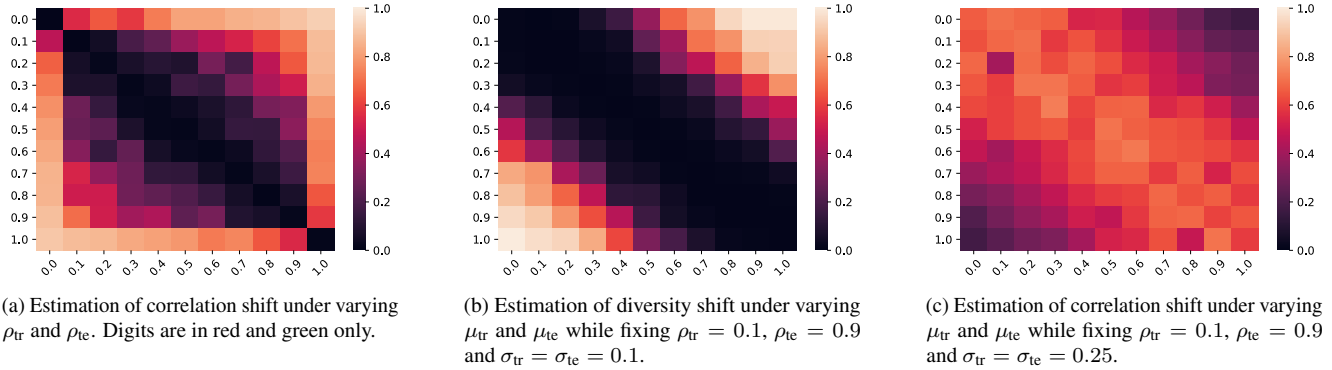
(a) Estimation of correlation shift under varying ρ_{tr} and ρ_{te}. Digits are in red and green only.

(b) Estimation of diversity shift under varying μ_{tr} and μ_{te} while fixing $\rho_{tr} = 0.1$, $\rho_{te} = 0.9$ and $\sigma_{tr} = \sigma_{te} = 0.1$.

(c) Estimation of correlation shift under varying μ_{tr} and μ_{te} while fixing $\rho_{tr} = 0.1$, $\rho_{te} = 0.9$ and $\sigma_{tr} = \sigma_{te} = 0.25$.

Figure 4. Estimation of diversity and correlation shift under varying color distribution in Colored MNIST. Another color, blue, uncorrelated with the labeled classes, is added onto the digits to create diversity shift. The intensity of blue is sampled from a truncated Gaussian distribution for every image. Assuming only one training and one test environment, ρ_{tr} and ρ_{te} stand for the correlation between red/green and the digits; μ_{tr} and μ_{te} stand for the mean intensities of blue; σ_{tr} and σ_{te} stand for the standard deviations.

culty for generalization. Instead, it represents a likelihood that certain algorithms will perform better than some other algorithms under the same kind of distribution shift.

3.2. Further Study

In this section, we conduct further experiments to check the reliability of our estimation method for diversity and correlation shift and compare our method against other existing metrics for measuring the non-i.i.d. property of datasets, demonstrating the robustness of our estimation method and the significance of diversity and correlation shift.

Sanity check and numerical stability. To validate the robustness of our estimation method, we check whether it can produce stable results that faithfully reflect the expected trend as we manipulate the color distribution of Colored MNIST. For simplicity, only one training environment is assumed. To start with, we manipulate the correlation coefficients ρ_{tr} and ρ_{te} between digits and colors in constructing the dataset. From Figure 4a, we can observe that when ρ_{tr} and ρ_{te} have similar values, the estimated correlation shift is negligible. It aligns well with our definition of correlation shift that measures the distribution difference of features present in both environments. As for examining on the estimation of diversity shift, another color, blue, is introduced in the dataset. The intensity (between 0 and 1) of blue added onto each digit is sampled from truncated Gaussian distributions with means μ_{tr}, μ_{te} and standard deviations σ_{tr}, σ_{te} for training and test environment respectively. Meanwhile, the intensity of red and green is subtracted by the same amount. From Figure 4b, we observe that as the difference in color varies between red/green and blue, the estimate of diversity shift varies accordingly (at the corners). Lastly, we investigate the behavior in the estimation of correlation shift while keeping the correlation coefficients fixed and manipulating μ_{tr} and μ_{te} that controls diversity shift. Figure 4c shows a

trade-off between diversity and correlation shift, as implied by their definitions. Experiments in every grid cell are conducted only once, so the heatmaps also reflect the variance in our estimation, which can be compensated by averaging over multiple runs.

Comparison with other measures of distribution shift. We also compare OoD-Bench with other measures of distribution shift. The results on the variants of Colored MNIST are shown in Tab. 3. We empirically show that general metrics for measuring the discrepancy between distributions, such as EMD [78] and MMD [30], are not very informative. Specifically, EMD and MMD are insensitive to the correlation shift in the datasets, while EMD is also insensitive to the diversity shift. Although NI [33] can produce comparative results on correlation shift, it is still unidimensional like EMD and MMD, not discerning the two kinds of distribution shift present in the datasets. In comparison, our method provides more stable and interpretable results. As ρ_{tr} an ρ_{te} gradually become close, the estimated correlation shift reduces to zero. On the other hand, the estimated diversity shift remains constant zero until the last scenario where our method again produces the expected answer.

4. Related Work

Quantification on distribution shifts. Non-i.i.d. Index (NI) [33] quantifies the degree of distribution shift between training and test set with a single formula. There are also a great number of general distance measures for distributions: Kullback-Leibler (KL) divergence, EMD [78], MMD [30], and $\mathcal{A}$-distance [15], etc. However, they all suffer from the same limitation as NI, not being able to discern different kinds of distribution shifts. To the best of our knowledge, we are among the first to formally identify the two-dimensional distribution shift and provide quantitative results on various

ρ_{te}	Dominant shift	EMD	MMD	NI	Div. shift (ours)	Cor. shift (ours)
0.9	Cor. shift	0.08 ± 0.01 ✗	0.01 ± 0.00 ✗	1.40 ± 0.06 ✓	0.00 ± 0.00	0.67 ± 0.04
0.7	Cor. shift	0.07 ± 0.00 ✗	0.01 ± 0.00 ✗	1.05 ± 0.03 ✓	0.00 ± 0.00	0.48 ± 0.06
0.5	Cor. shift	0.07 ± 0.00 ✗	0.00 ± 0.00 ✗	0.72 ± 0.04 ✓	0.00 ± 0.00	0.34 ± 0.06
0.3	Cor. shift	0.06 ± 0.00 ✗	0.00 ± 0.00 ✗	0.57 ± 0.04 ✓	0.00 ± 0.00	0.18 ± 0.05
0.1	None	0.06 ± 0.00 ✗	0.00 ± 0.00 ✓	0.39 ± 0.02 ✗	0.00 ± 0.00	0.00 ± 0.00
$0.1^{\dagger}$	Div. shift	0.29 ± 0.01 ✗	1.00 ± 0.00 ✓	10.76 ± 0.43 ✗	0.93 ± 0.01	0.00 ± 0.00

Table 3. Existing metrics on measuring the distribution shift in Colored MNIST with only one training environment where $\rho_{tr} = 0.1$. All environments contain only red and green digits except the last. †Blue is added with $\mu_{tr} = 0$, $\mu_{te} = 1$ and $\sigma_{tr} = \sigma_{te} = 0.1$. Results are averaged over 5 runs.

OoD datasets. Notably, a concurrent work [99] studies three kinds of distribution shift, namely *spurious correlation*, *low-data drift*, and *unseen data shift*, which are very similar to correlation and diversity shift. Their findings are mostly in line with ours, but they do not provide any quantification formula or estimation method for the shifts.

OoD generalization. Without access to test distribution examples, OoD generalization always requires additional assumptions or domain information. In the setting of DG [17, 59, 89], it is often assumed that multiple training datasets sampled from similar but distinct domains are available. Hence, most DG algorithms aim at learning a domain-invariant data representation across training domains. These algorithms take various approaches include domain adversarial learning [3, 5, 6, 27, 48, 100, 101, 105], meta-learning [13, 25, 47, 52, 103], image-level and feature-level domain mixup [55, 100], adversarial data augmentation [81], domain translation/randomization [62, 75, 108], feature alignment [68, 85], gradient alignment [43, 73, 83], gradient orthogonalization [12], invariant risk minimization [1, 9, 44], self-supervised learning [96, 107], prototypical learning [26], and kernel methods [16, 29, 51, 59]. There are also DG algorithms that do not assume multiple training domains. Many of them instead assume that variations in the style/texture of images is the main cause of distribution shift. These algorithms mostly utilize AdaIN [36] or similar operations to perform style perturbations so that the learned classifier would be invariant to various styles across domains [40, 49, 60, 84, 97, 109]. Other approaches include [19] which designs a self-supervision objective enforcing models to focus on global image structures such as shapes of objects, and [94] which introduces an explicit adversarial learning objective so that the learned model would be invariant to local patterns. More general single-source DG algorithms (that do not assume the style/texture bias) and other OoD generalization algorithms include distributionally robust optimization [79], self-challenging [38], spectral decoupling [70], feature augmentation [50], adversarial data augmentation [71, 93], gradient alignment [64], sample reweighting [33, 104], test-time training [87], removing bias with bias [11], contrastive learning [41], causal discovery [58], and variational bayes that leverages causal structures of data [53, 86]. For a more comprehensive summary of existing OoD generalization and DG algorithms, we refer readers to these survey papers [82, 95, 106].

DomainBed. The living benchmark is created by [31] to facilitate disciplined and reproducible DG research. After conducting a large-scale hyperparameter search, the performances of fourteen algorithms on seven datasets are reported. The authors then arrive at the conclusion that ERM beats most of DG algorithms under the same fair setting. Our work differs from DomainBed mainly in three aspects. First, we not only provide a benchmark for algorithms but also for datasets, helping us gain a deeper understanding of the distribution shift in the data. Second, we compare different algorithms in a more informative manner in light of diversity and correlation shift, recovering the fact that some algorithms are indeed better than ERM in appropriate scenarios. Third, we experiment with several new algorithms and new datasets, especially those dominated by correlation shift.

5. Conclusion

In this paper, we have identified diversity shift and correlation shift as two of the main forms of distribution shift in OoD datasets. The two-dimensional characterization positions disconnected datasets into a unified picture and have shed light on the nature of unknown distribution shift in some real-world data. In addition, we have demonstrated some of the strengths and weaknesses of existing OoD generalization algorithms. The results suggest that future algorithms should be more comprehensively evaluated on two types of datasets, one dominated by diversity shift and the other dominated by correlation shift. Lastly, we leave an open problem regarding whether there exists an algorithm that can perform well under both diversity and correlation shift. If not then our method can be used for choosing the appropriate algorithms.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *ICML*, 2020. 5, 8

[2] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? a sample complexity perspective. In *ICLR*, 2021. 2

[3] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *ECML-PKDD*, 2019. 8

[4] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 2018. 1

[5] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv:1911.00804*, 2019. 8

[6] Isabela Albuquerque, João Monteiro, Tiago H Falk, and Ioannis Mitliagkas. Adversarial target-invariant representation learning for domain generalization. *arXiv:1911.00804*, 2019. 8

[7] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *CVPR*, 2019. 1

[8] Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020. 1

[9] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019. 2, 4, 5, 6, 8, 19

[10] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *ICLR*, 2019. 3

[11] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020. 1, 8

[12] Haoyue Bai, Rui Sun, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S-H Gary Chan, and Zhenguo Li. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. *arXiv:2012.09382*, 2020. 5, 8

[13] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *NeurIPS*, 2018. 8

[14] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. 4, 19

[15] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *NeurIPS*, 2007. 7

[16] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv:1711.07910*, 2017. 5, 6, 8

[17] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NeurIPS*, 2011. 1, 8

[18] Péter et al. Bándi. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *TMI*, 2019. 19

[19] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 5, 6, 8

[20] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 2020. 1

[21] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *ICLR*, 2021. 19

[22] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *ITSC*, 2018. 1

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018. 1

[25] Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019. 5, 8

[26] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *CVPR*, 2021. 8

[27] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 5, 6, 8

[28] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 1

[29] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016. 8

[30] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *NeurIPS*, 2006. 7

[31] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 1, 4, 5, 8, 19, 20, 21

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5

[33] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 2020. 4, 7, 8, 19

[34] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler

Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv:2006.16241*, 2020. 4, 5

[35] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv:1907.07174*, 2019. 4

[36] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 8

[37] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 2020. 1

[38] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *arXiv:2007.02454*, 2020. 5, 6, 8

[39] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv:1806.07572*, 2018. 4, 16, 17

[40] Seogkyu Jeon, Kibeom Hong, Pilhyeon Lee, Jewook Lee, and Hyeran Byun. Feature stylization and domain-aware contrastive learning for domain generalization. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 8

[41] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *ICCV*, 2021. 8

[42] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv:2012.07421*, 2020. 1, 4, 5, 19, 20

[43] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv:2008.01883*, 2020. 5, 6, 8

[44] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv:2003.00688*, 2020. 1, 5, 6, 8

[45] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 2, 19

[46] D. Li, Y. Yang, Y. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 2, 4, 5, 19

[47] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 5, 6, 8

[48] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 5, 6, 8

[49] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *CVPR*, 2021. 8

[50] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *ICCV*, 2021. 8

[51] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI*, 2018. 8

[52] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, 2019. 8

[53] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In *NeurIPS*, 2021. 2, 8

[54] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 4, 19

[55] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *ECCV*, 2020. 1, 8

[56] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv:1907.07484*, 2019. 1

[57] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *ICLR*, 2021. 2

[58] S Chandra Mouli and Bruno Ribeiro. Asymmetry learning for counterfactually-invariant classification in ood tasks. In *ICLR*, 2021. 8

[59] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 1, 8

[60] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap via style-agnostic networks. *arXiv:1910.11645*, 2019. 5, 6, 8

[61] M Newman and G Barkema. *Monte carlo methods in statistical physics chapter 1-4*. Oxford University Press, 1999. 4

[62] A. Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. In *NeurIPS*, 2021. 8

[63] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *arXiv:2111.12525*, 2021. 2

[64] Giambattista Parascandolo, Alexander Neitz, ANTONIO ORVIETO, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *ICLR*, 2021. 5, 6, 8

[65] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 1962. 4

[66] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000. 3

[67] Judea Pearl. Causal inference. *Causality: Objectives and Assessment*, pages 39–58, 2010. 1

[68] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 8

[69] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 1

[70] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *arXiv:2011.09468*, 2020. 5, 8, 19

[71] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, 2020. 8

[72] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *CVPR*, 2019. 19

[73] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv:2109.02934*, 2021. 8

[74] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 1, 2

[75] Alexander Robey, George Pappas, and Hamed Hassani. Model-based domain generalization. In *NeurIPS*, 2021. 8

[76] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 1956. 4

[77] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *ICLR*, 2021. 6

[78] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *ICCV*, 1998. 7

[79] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020. 5, 6, 8, 19

[80] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jack Jia, Xue Lin, and Qi Alfred Chen. Security of deep learning based lane keeping system under physical-world adversarial attack. *arXiv:2003.01782*, 2020. 1

[81] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv:1804.10745*, 2018. 8

[82] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv:2108.13624*, 2021. 1, 8

[83] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv:2104.09937*, 2021. 8

[84] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv:2006.11207*, 2020. 8

[85] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 5, 6, 8

[86] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. In *NeurIPS*, 2021. 8

[87] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 8

[88] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1

[89] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 2, 8

[90] Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998. 1, 5, 6

[91] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 4, 19

[92] Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust CNN-based object detection through augmentation with synthetic rain variations. In *ITSC*, 2019. 1

[93] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018. 8

[94] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 8

[95] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv:2103.03097*, 2021. 1, 8

[96] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, 2020. 8

[97] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *ICCV*, 2021. 8

[98] Zeyu Wang, Klint Qinami, Ioannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020. 19

[99] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *ICLR*, 2022. 8

[100] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020. 5, 8

[101] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv:2001.00677*, 2020. 5, 6, 8

[102] Amir Zandieh, Insu Han, Haim Avron, Neta Shoham, Chaewon Kim, and Jinwoo Shin. Scaling neural tangent kernels via sketching and random features. *arXiv:2106.07880*, 2021. 16

[103] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group distribution shift. *arXiv:2007.02931*, 2020. 5, 6, 8

[104] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization, 2021. 1, 8

[105] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. In *NeurIPS*, 2020. 5, 6, 8

[106] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv:2103.02503*, 2021. 1, 8

[107] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, 2020. 8

[108] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020. 8

[109] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2020. 8