A Fine-Grained Analysis on Distribution Shift

Anonymous authors
Paper under double-blind review

Abstract

Robustness to distribution shifts is critical for deploying machine learning models in the real world. Despite this necessity, there has been little work in defining the underlying mechanisms that cause these shifts and evaluating the robustness of algorithms across multiple, different distribution shifts. To this end, we introduce a framework that enables fine-grained analysis of various distribution shifts. We provide a holistic analysis of current state-of-the-art methods by evaluating 19 distinct methods grouped into five categories across both synthetic and real-world datasets. Overall, we train more than 85K models. Our experimental framework can be easily extended to include new methods, shifts, and datasets. We find, unlike previous work (Gulrajani & Lopez-Paz, 2021), that progress has been made over a standard ERM baseline; in particular, pretraining and augmentations (learned or heuristic) offer large gains in many cases. However, the best methods are not consistent over different datasets and shifts. We will open source our experimental framework, allowing future work to evaluate new methods over multiple shifts to obtain a more complete picture of a method's effectiveness.

1 Introduction

If machine learning models are to be ubiquitous in critical applications such as driverless cars (Janai et al., 2020), medical imaging (Erickson et al., 2017), and science (Jumper et al., 2021), it is pivotal to build models that are robust to distribution shifts. Otherwise, models may fail surprisingly in ways that derail trust in the system. For example, Koh et al. (2020); Perone et al. (2019); AlBadawy et al. (2018); Heaven (2020); Castro et al. (2020) find that a model trained on images from one set of hospitals may not generalise to the imaging conditions of another; Alcorn et al. (2019); Dai & Van Gool (2018) find that a model for driverless cars may not generalise to new lighting conditions or object poses; and Buolamwini & Gebru (2018) find that a model may perform worse on subsets of the distribution, such as different ethnicities, if the training set has an imbalanced distribution. Thus, it is important to understand when we expect a model to generalise and when we do not. This would allow a practitioner to have confidence in the system (e.g. if a model is shown to be robust to the imaging conditions of different hospitals, then it can be deployed in new hospitals with confidence).

While domain generalization is a well studied area, Gulrajani & Lopez-Paz (2021); Schott et al. (2021) have cast doubt on the efficacy of existing methods, raising the question: *has any progress been made in domain generalization over a standard expectation risk minimization (ERM) algorithm?* Despite these discouraging results, there are many examples that machine learning models *do* generalise across datasets with different distributions. For example, CLIP (Radford et al., 2021), with well engineered prompts, generalizes to many standard image datasets. Taori et al. (2020) found that models trained on one image dataset generalise to another, albeit with some drop in performance; in particular, higher performing models generalise better. However, there is little understanding and experimentation on *when* and *why* models generalise, especially in realistic settings inspired by real-world applications. This begs the following question:

Can we define the important distribution shifts to be robust to and then systematically evaluate the robustness of different methods?

To answer the above question, we present a grounded understanding of robustness to distribution shifts. We draw inspiration from disentanglement literature (see section 6), which aims to separate images into an independent set of factors of variation (or attributes). In brief, we assume the data is composed of some (possibly extremely large) set of attributes. We expect models, having seen some distribution of values for an attribute, to be able to learn invariance to that attribute and so to generalise to unseen examples of the attribute and different distributions over that attribute. Using a simple example to clarify the setup, assume our data has two attributes (shape and color) among others. Given data with some distribution over the set of possible colors (e.g. red and blue) and the task of predicting shape (e.g. circle or square), we want our model to generalise to unseen colors (e.g. green) or a different distribution of colors (e.g. there are very few red circles in the training set, but the samples at evaluation are uniformly sampled from the set of possible colors and shapes).

Using this framework, we evaluate models across three distribution shifts: *spurious correlation*, *low-data drift*, and *unseen data shift* (illustrated in figure 1) and two additional conditions (label noise and dataset size). We choose these settings as they arise in the real world and harm generalization performance. Moreover, in our framework, these distribution shifts are the fundamental blocks of building more complex distribution shifts. We additionally evaluate models when there is varying amounts of label noise (as inspired by noise arising from human raters) and when the total size of the train set varies (to understand how models perform as the number of training examples changes). The unique ability of our framework to evaluate fine-grained performance of models across different distribution shifts and under different conditions is of critical importance to analyze methods under a variety of real-world settings. This work makes the following contributions:

- We propose a framework to unify different distribution shifts, defining how they arise and how different common approaches promote robustness to these shifts. We use this framework to define three, real world inspired distribution shifts. We then use this framework to create a systematic evaluation setup across real and synthetic datasets for different distribution shifts. Our evaluation framework is easily extendable to new distribution shifts, datasets, or methods to be evaluated. We will open source the code upon acceptance.
- We evaluate and compare 19 different methods (training more than 85K models) in these settings. These methods span 5 common techniques to improve robustness: architecture choice, data augmentation, domain generalization, adaptive algorithms, and representation learning. This allows for a direct comparison across different areas in machine learning.
- We find that simple techniques, such as data augmentation and pretraining are often effective and that domain generalization algorithms do work for certain datasets and distribution shifts. However, there is no easy way to select the best approach a-priori and results are inconsistent over different datasets and attributes, demonstrating there is still much work to be done to improve robustness in real-world settings.

2 Framework to Evaluate Generalization

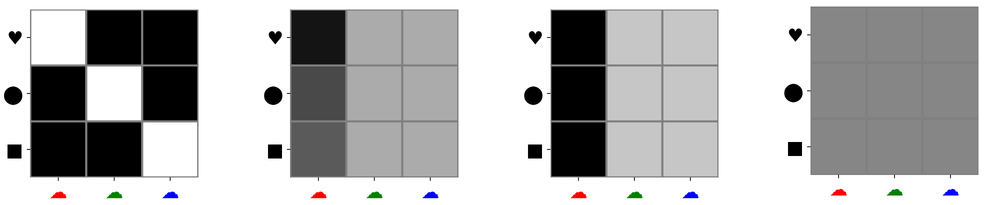
In this section we introduce our robustness framework for characterizing distribution shifts in a principled manner. We then relate three common, real world inspired distribution shifts.

2.1 Latent Factorisation

We assume a joint distribution p of inputs $\boldsymbol{x}$ and corresponding attributes $y^1, y^2, \ldots, y^K$ (denoted as $y^{1:K}$) with $y^k \in \mathbb{A}^k$ where $\mathbb{A}^k$ is a finite set. One of these K attributes is a label of interest, denoted as y^l (in a mammogram, the label could be cancer/benign and a nuisance attribute y^i with $i \neq l$ could be the identity of the hospital where the mammogram was taken). Our aim is to build a classifier f that minimizes the risk R. However, in real-world applications, we only have access to a finite set of inputs and attributes of size n. Hence, we minimize the empirical risk $\hat{R}$ instead:

$$R(f) = \mathbb{E}_{(\boldsymbol{x},y^l)\sim p}\left[\mathcal{L}(y^l, f(\boldsymbol{x}))\right] \qquad \hat{R}(f;p) = \frac{1}{n}\sum_{\{(y_i^l,\boldsymbol{x}_i)\sim p\}_{i=1}^n} \mathcal{L}(y_i^l, f(\boldsymbol{x}_i)).$$

where $\mathcal{L}$ is a suitable loss function. Here, all nuisance attributes y^k with $k \neq l$ are ignored and we work with samples obtained from the marginal $p(y^l, \boldsymbol{x})$. In practice, however, due to selection bias or other confounding factors in data collection, we are only able to train and test our models on data collected from two related but distinct distributions: $p_{\text{train}}, p_{\text{test}}$. For example, p_{train} and p_{test} may be concentrated on different subsets of hospitals and this discrepancy may result in a distribution shift; for example, hospitals may use different equipment, leading to different staining on their cell



(a) p_{train}**: SC.** (b) p_{train}**: LDD.** (c) p_{train}**: UDS.** (d) p_{test}**:** y^l, y^a **are IID.**

Figure 1: Visualization of the joint distribution for the different shifts we consider on the DSPRITES example. The lighter the color, the more likely the given sample. figure 1a-1c visualise different shifts over $p_{\text{train}}(y^l, y^a)$ discussed in 2.2: *spurious correlation* (SC), *low-data drift* (LDD), and *unseen data shift* (UDS). figure 1d visualises the test set, where the attributes are uniformly distributed.

images. While we train f on data from p_{train} by minimizing $\hat{R}(f; p_{\text{train}})$, we aim to learn a model that generalises well to data from p_{test}; that is, it should achieve a small $\hat{R}(f; p_{\text{test}})$.

While generalization in the above sense is desirable for machine learning models, it is not clear *why* a model f trained on data from p_{train} *should* generalise to p_{test}. It is worth noting that while p_{train} and p_{test} can be different, they are both related to the true distribution p. We take inspiration from disentanglement literature to express this relationship. In particular, that we can view data as being decomposed into an underlying set of factors of variations. We formalise various distribution shifts using a latent variable model for the true data generation process:

$$z \sim p(z) \qquad y^i \sim p(y^i|z) \quad i = 1 \dots K \qquad \boldsymbol{x} \sim p(\boldsymbol{x}|z) \tag{1}$$

where z denotes latent factors. By a simple refactorization, we can write

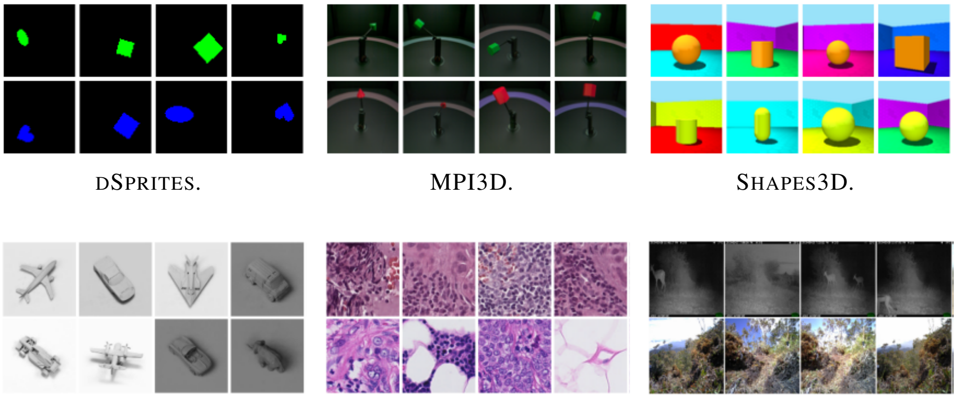
$$p(y^{1:K}, \boldsymbol{x}) = p(y^{1:K}) \int p(\boldsymbol{x}|z)p(z|y^{1:K})dz = p(y^{1:K})p(\boldsymbol{x}|y^{1:K}).$$

Thus, the true distribution can be expressed as the product of the marginal distribution of the attributes with a conditional generative model. We assume that distribution shifts arise when a new marginal distribution for the attributes is chosen, such as $p(y^{1:K}) \neq p_{\text{train}}(y^{1:K}) \neq p_{\text{test}}(y^{1:K})$, but otherwise the conditional generative model is shared across all distributions, i.e., we have $p_{\text{test}}(y^{1:K}, \boldsymbol{x}) = p_{\text{test}}(y^{1:K}) \int p(\boldsymbol{x}|z)p(z|y^{1:K})dz$, and similarly for p_{train}. This is a special case of covariate shift (Schölkopf et al., 2012) where the input $\boldsymbol{x}$ is the result of an unknown (but fixed) generating process conditioned on the attributes. This is realistic in practice, as the physical process (e.g. imaging in a hospital) is constrained but attributes such as age, sex, imaging equipment, hospital, etc. may vary.

To provide more context, as a running example, we use the color DSPRITES dataset (Matthey et al., 2017); where in our notation y^1 defines the color with $\mathbb{A}^1 = \{\text{red}, \text{green}, \text{blue}\}$, and y^2 defines the shape with $\mathbb{A}^2 = \{\text{ellipse}, \text{heart}, \text{square}\}$. We can imagine that a data collector (intentionally or implicitly) selects some marginal distribution over attributes $p_{\text{train}}(y^{1:K})$ when training; for example they select mostly blue ellipses and red hearts. This induces a new joint distribution over latent factors and attributes: $p_{\text{train}}(z, y^{1:K}) = p(z|y^{1:K})p_{\text{train}}(y^{1:K})$. Consequently, during training, we get images with a different joint distribution $p_{\text{train}}(\boldsymbol{x}, y^{1:K}) = \int p(\boldsymbol{x}|z)p_{\text{train}}(z, y^{1:K})$. This similarly applies when collecting data for the test distribution. We focus on common cases of distribution shifts visualized in figure 1; we discuss these in more detail in section 2.2.

The goal of enforcing robustness to distribution shifts is to maintain performance when the data generating distribution p_{train} changes. In other words, we would like to minimize risk on p, p_{test} given *only* access to p_{train}. We can achieve robustness in the following ways:

1. Weighted resampling. We can resample the training set using importance weights $W(y^{1:K}) = p(y^{1:K})/p_{\text{train}}(y^{1:K})$. Given the attributes, the i-th data point $(y_i^{1:K}, \boldsymbol{x}_i)$ in the training set is used with probability $W(y_i^{1:K})/\sum_{i'=1}^{n} W(y_{i'}^{1:K})$ rather than $1/n$. We refer to this empirical distribution as p_{reweight}. This requires access to the true distribution $p(y^{1:K})$, so to avoid bias and improve fairness, it is often assumed that all combinations of attributes happen uniformly at random.



dSprites. MPI3D. Shapes3D.

SmallNorb. Camelyon17. iWildCam.

Figure 2: **Dataset samples**. Each row fixes an attribute (e.g. color for dSprites, MPI3D, Shapes3D; azimuth for SmallNorb; hospital for Camelyon17; and location for iWildCam).

2. Heuristic Data Augmentation. Weighted resampling can lead to samples being reused many times, leading to overfitting. To mitigate overfitting, we can augment the training data with heuristic choices, such as color jitter. While not the typical interpretation, we can view this setup as encouraging invariance to nuisance attributes in the generative model, which leads us to the next approach.

3. Learned Data Augmentation: We can learn a generative model $\hat{p}(\boldsymbol{x}|y^{1:K})$ from the training data that aims to approximate $\int p(\boldsymbol{x}|z)p(z|y^{1:K})dz$, as the true conditional generator is by our assumption the same over all (e.g. train and test) distributions. If such a conditional generative model can be learned, we can sample new synthetic data at training time (e.g. according to the true distribution $p(y^{1:K})$) to correct for the distribution shift. More precisely, we can generate data from the augmented distribution $p_{\text{aug}} = (1-\alpha)p_{\text{train}} + \alpha\hat{p}(\boldsymbol{x}|y^{1:K})p(y^{1:K})$ and train a supervised classifier on this augmented dataset. Here, $\alpha \in [0, 1]$ is the percentage of synthetic data used for training.

4. Representation Learning: An alternative factorization of a data generating distribution (e.g. train) is $p_{\text{train}}(y^{1:K}, \boldsymbol{x}) = \int p(z|\boldsymbol{x})p_{\text{train}}(y^{1:K}|z)dz$. We can learn an unsupervised representation that approximates $p(z|\boldsymbol{x})$ using the training data only, and attach a classifier to learn a task specific head that approximates $p_{\text{train}}(y^l|z)$. Again, by our assumption $p(z|\boldsymbol{x}) \propto p(\boldsymbol{x}|z)p(z)$. Given a good guess of the true prior, the learned representation would not be impacted by the specific attribute distribution and so generalise to p_{test}, p.

2.2 Distribution Shifts

While distribution shifts can happen in a continuum, we consider three types of shifts, inspired by real-world challenges. We discuss these shifts and two additional, real-world inspired conditions. Our experimental framework can be used to compare more complex shifts, discussed in section E.

Test distribution p_{test}. We assume that the attributes are distributed uniformly: $p_{\text{test}}(y^{1:K}) = 1/\prod_i |\mathbb{A}^i|$. This is desirable, as all attributes are represented and a-priori independent.

Shift 1: Spurious correlation – Attributes are correlated under p_{train} but not p_{test}. Spurious correlation arises in the wild for many reasons including capture bias, environmental factors, and geographical bias (Beery et al., 2018; Torralba & Efros, 2011). These spurious correlations lead to surprising results and poor generalization. Therefore, it is important to build models that are robust to such challenges. In our framework, spurious correlation arises when two attributes y^a, y^b are correlated at training time, but this is not true of p_{test}, for which attributes are independent: $p_{\text{train}}(y^a|y^1 \ldots y^b \ldots y^K) > p_{\text{train}}(y^a|y^1 \ldots y^{b-1}, y^{b+1} \ldots y^K)$. This is especially problematic if one attribute y^b is the label. Using the running dSprites example, shape and color may be correlated and the model may find it easier to predict color. If color is the label, the model will generalise. However, if shape is the label, the model's reliance on color will lead to poor generalization.

Shift 2: Low-data drift – Attribute values are unevenly distributed under p_{train} but not under p_{test}. Low-data drift arises in the wild (e.g. in (Buolamwini & Gebru, 2018) for different ethnicities) when data has not been collected uniformly across different attributes. When deploying models in the wild, it is important to be able to reason and have confidence that the final predictions will be consistent and fair across different attributes. In the framework above, low-data shifts arise when certain values in the set $\mathbb{A}^a$ of an attribute y^a are sampled with a much smaller probability than in p_{test}: $p_{\text{train}}(y^a = v) \ll p_{\text{test}}(y^a = v)$. Using the DSPRITES example, only a handful of red shapes may be seen at training time, yet in p_{test} all colors are sampled with equal probability.

Shift 3: Unseen data shift – Some attribute values are unseen under p_{train} but are present under p_{test}. This is a special case of *shift 2: low-data drift*, which we make explicit due to its importance in real world applications. Unseen data shift arises when a model, trained in one setting is expected to work in another, disjoint setting. For example: a model trained to classify animals in images at certain times of day should generalise to other times of day. In our framework, *unseen data shift* arises when some values in the set $\mathbb{A}^a$ of an attribute y^a are not in p_{train} but are in p_{test}:

$$p_{\text{train}}(y^a = v) = 0 \qquad p_{\text{test}}(y^a = v) > 0 \qquad |\{v | p_{\text{train}}(y^a = v)\}| > 1 \tag{2}$$

This is a stronger constraint than in standard out-of-distribution generalization (see section 6), as multiple values for $\mathbb{A}^a$ must be seen under p_{train}, which allows the model to learn invariance to y^a. In the DSPRITES example, we can assume there are two colors (green, blue) in train but the color red is unseen at train time. All colors are in p_{test}.

Discussion. We choose these shifts as they are building blocks of more complex distribution shifts. Consider two attributes: the label and a nuisance attribute. The marginal distribution of the label, it decomposes into two terms: the conditional probability and the probability of a given attribute value: $p(y^l) = \sum_{y^a} p(y^l|y^a)p(y^a)$. The three shifts we consider control these terms independently: *unseen data shift* and *low-data drift* control $p(y^a)$ whereas *spurious correlation* controls $p(y^l|y^a)$. The composition of these terms describes any distribution shift for these two variables.

2.3 Conditions

Label noise. We investigate the change in performance due to noisy information. This can arise when there are disagreements and errors among the labellers (e.g. in medical imaging (Castro et al., 2020)). We model this as an observed attribute (e.g. the label) being corrupted by noise. $\hat{y}^i \sim c(y^i)$, where $y^i \in \mathbb{A}^i$ is the true label, $\hat{y}^i \in \mathbb{A}^i$ the corrupted, observed one, and c the corrupting function.

Dataset size. We investigate how performance changes with the size of the training dataset. This setting arises when it is unrealistic or expensive to collect additional data (e.g. in medical imaging or in camera trap imagery). Therefore, it is important to understand how performance degrades given fewer total samples. We do this by limiting the total number of samples from p_{train}.

3 Models Evaluated

We evaluate 19 algorithms to cover a broad range of approaches that can be used to improve model robustness to distribution shifts and demonstrate how they relate to the three ways to achieve robustness, outlined in section 2. We believe this is the first paper to comprehensively evaluate a large set of different approaches in a variety of settings. These algorithms cover the following areas: architecture choice, data augmentation, domain adaptation, adaptive approaches and representation learning. Further discussion on how these models relate to our robustness framework is in appendix F.

Architecture choice. We use the following standard models: ResNet18, ResNet50, ResNet101 (He et al., 2016), ViT (Dosovitskiy et al., 2021), and an MLP (Vapnik, 1992). We use weighted resampling to oversample from the parts of p_{train} that have a lower probability of being sampled from. Performance depends on how robust the learned representation is to distribution shift.

Heuristic data augmentation. These approaches encourage invariance to nuisance attributes in the generative model to improve robustness. We analyze the following augmentation methods: standard ImageNet augmentation (He et al., 2016), AugMix without JSD (Hendrycks et al., 2020), RandAugment (Cubuk et al., 2020), and AutoAugment (Cubuk et al., 2019). Performance depends on how well the heuristic augmentations enforce invariance to useful attributes.

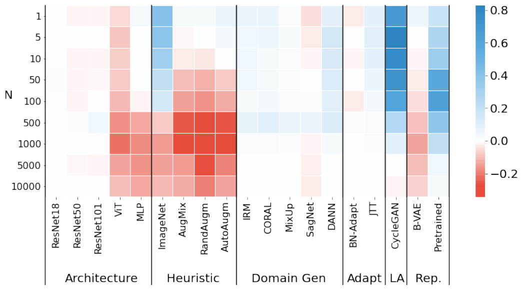


Figure 3: **Spurious Correlation.** We use all correlated samples and vary the number of samples N from the true, uncorrelated distribution. We plot the percentage change over the baseline ResNet, averaged over all seeds and datasets. Blue is better, red worse. CycleGAN performs consistently best while ImageNet augmentation and pretraining on ImageNet also consistently boosts performance.

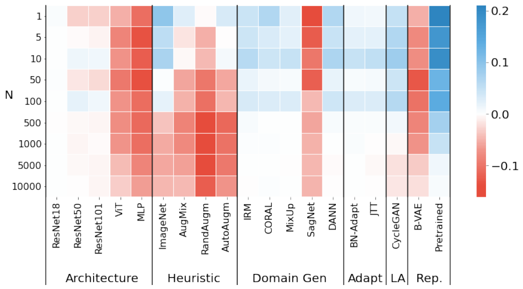


Figure 4: **Low-data drift.** We use all samples from the high data regions and vary the number of samples N from the low-data region. We plot the percentage change over the baseline ResNet, averaged over all seeds and datasets. Blue is better, red worse. Pretraining on ImageNet performs consistently best, while CycleGAN, most domain generalization methods and ImageNet augmentation also provide some boost in performance.

Learned data augmentation. These approaches approximate the true underlying generative model $p(\boldsymbol{x}|y^{1:K})$ by learning augmentations conditioned on the nuisance attribute. The learned augmentations can be used to transform any image $\boldsymbol{x}$ to have a new attribute, while keeping the other attributes fixed. We follow Goel et al. (2020), who use CycleGAN (Zhu et al., 2017), but we do not use their SGDRO objective in order to evaluate the performance of learned data augmentation alone. Performance depends on how well the learned augmentations approximate the true generative model.

Domain generalization. These approaches aim to recover a representation z that is independent of the attribute: $p(y^a, z) = p(y^a)p(z)$ to allow generalization over that attribute. We evaluate IRM (Arjovsky et al., 2019), DeepCORAL (Sun & Saenko, 2016), domain MixUp (Gulrajani & Lopez-Paz, 2021), DANN (Ganin et al., 2016), and SagNet (Nam et al., 2021). We train these models by treating values of the nuisance attribute y^a as different domains. Performance depends on the invariance of the learned representation z.

Adaptive approaches. These works modify p_{reweight} dynamically. We evaluate JTT (Liu et al., 2021) and BN-Adapt (Schneider et al., 2020). These methods do not give performance guarantees.

Representation learning. These works aim to learn a robust representation of z that describes the true prior. We evaluate using a β-VAE (Higgins et al., 2017a) and pretraining on ImageNet (Deng et al., 2009). Performance depends on the quality of the learned representation for the specific task.

4 Experiments

We evaluate the 19 different methods across these six datasets, three distribution shifts, varying label noise, and dataset size. We plot aggregate results in figures 3-7 by averaging results over the different datasets and complete results in the appendix in figures 10-12, in which the results are broken down by dataset. All results reported are on the test set. We distill the results into seven concrete takeaways in section 4.1 and four practical tips in section 4.2.

Datasets. We evaluate these approaches on six vision, classification datasets (see figure 2 – dSprites (Matthey et al., 2017), MPI3D (Gondal et al., 2019), SmallNorb (LeCun et al., 2004), Shapes3D (Burgess & Kim, 2018), Camelyon17 (Koh et al., 2020; Bandi et al., 2018), and iWildCam (Koh et al., 2020; Beery et al., 2018). These datasets consist of multiple (potentially an arbitrarily large number) attributes. We select two attributes y^l, y^a for each dataset and make one y^l the label. We then use these two attributes to build the three shifts. These datasets differ in their properties: dSprites, MPI3D, Shapes3D are all synthetic; SmallNorb consists of black and white images of toys; Camelyon17 consists of medical imagery and iWildCam of camera trap imagery. While we aggregate results in the main text, results differ across datasets so please refer to

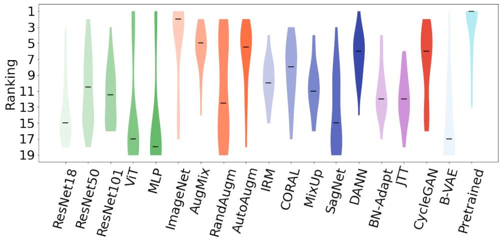


Figure 5: **Unseen data shift.** We rank the methods (where best is 1, worst 19) for each dataset and seed and plot the rankings, with the overall median rank as the black bar. Pretraining on ImageNet and ImageNet augmentation perform consistently best. DANN, CycleGAN and other heuristic augmentations perform consistently well.

figures 10-12 for performance on each dataset. Additional samples, description, and discussion on the shifts are given in appendix D.1, D.2.

Model selection. When investigating heuristic data augmentation, domain generalization, learned augmentation, adaptive approaches, and representation learning, we use a ResNet18 for the simpler, synthetic datasets (DSPRITES, MPI3D, SHAPES3D, and SMALLNORB) but a ResNet50 for the more complex, real world ones (CAMELYON17 and IWILDCAM). To perform model selection, we choose the best model according to the in-distribution validation set ($\mathcal{D}_{\text{valid}} \subset \mathcal{D}_{\text{train}}$). In the *unseen data shift* setting for the CAMELYON17 and IWILDCAM, we use the given out-of-distribution validation set, which is a different, distinct set in $\mathcal{D}$ that is independent of $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$. (We consider using the in-distribution validation set in appendix B.4.)

Hyperparameter choices. We perform a sweep over the hyperparameters (appendix G.8). We run each set of hyperparameters for five seeds for each setting. To choose the best model for each seed, we perform model selection over *all* hyperparameters using the top-1 accuracy on the validation set. In the *low-data* and *spurious correlation* settings, we choose a different set of samples from the low-data region with each seed. We report the mean and standard deviation over the five seeds.

Additional data. Only the models pretrained on ImageNet use additional data to $\mathcal{D}_{\text{train}}$.

4.1 TAKEAWAYS

Takeaway 1: While we can improve over ERM, no one method always performs best. The relative performance between methods varies across datasets and shifts. Under *spurious correlation* (figure 3), CYCLEGAN consistently performs best but in figure 4, under *low-data drift*, pretraining consistently performs best. Under *unseen data shift* (figure 5), pretraining is again one of the best performing models. However, if we drill down on the results in figure 10 (appendix B.1), we can see pretraining performs best on the synthetic datasets, but not on CAMELYON17 (where using augmentation or DANN is best) or IWILDCAM (where using ViT or an MLP is best).

Takeaway 2: Pretraining is a powerful tool across different shifts and datasets. While pretraining is not always helpful (e.g. in appendix B.1 on CAMELYON17, IWILDCAM in figures 10-11), it often provides a strong boost in performance. This is presumably because the representation z learned during pretraining is helpful for the downstream task. For example, the representation may have been trained to be invariant to certain useful properties (e.g. scale, shift, and color).

Takeaway 3: Heuristic augmentation does not always improve generalization. In all settings (figures 3-5), ImageNet augmentation generally improves performance. However, RandAugment, AugMix, and AutoAugment have more variable performance (as further shown in figures 10-12). These methods are compositions of different augmentations. We investigate the impact of each augmentation in RandAugment in appendix B.2 and find variable performance. Mintun et al. (2021) investigated a similar problem: the impact of different corruptions on robustness to other corruptions. Augmentations that promote invariance to nuisance attributes in the true underlying generative model $p(\boldsymbol{x}|y^{1:K})$ lead to the best results; otherwise, the model may waste capacity. For example, on MPI3D (which consists of a robotic arm with different colored shapes), color augmentations (such as color jitter, contrast and brightness improve performance) but spatial transformations (such as translation and sheering) which do not exist in the dataset, as the robotic arm is always in the same position, harm performance.

Takeaway 4: Learned data augmentation is effective across different conditions and distribution shifts. This approach is highly effective in the *spurious correlation* setting (figure 3). It can also help in the *low-data* and *unseen data shift* settings (figure 4,5) (though the gains for these two

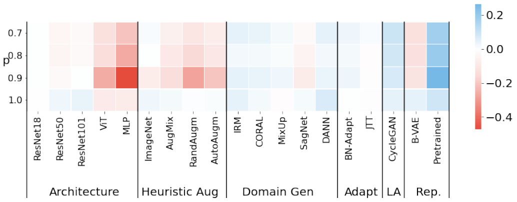


Figure 6: **Condition 1: Noisy labels.** We vary the amount of noise p in the labels. We plot the percentage change over the baseline ResNet, averaged over all seeds and datasets.

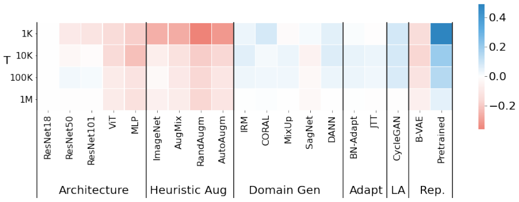


Figure 7: **Condition 2: Fixed data.** We vary the total size of the dataset T. We plot the percentage change over the baseline ResNet, averaged over all seeds and datasets.

shifts are not as large as for pretraining). The effectiveness of this approach can be explained by the fact that if the augmentations are learned perfectly, then augmented samples by design are from the true underlying generative model and can cover missing parts of the distribution.

Takeaway 5: Domain generalization algorithms offer limited performance improvement. In some cases these methods (in particular DANN) do improve performance, most notably in the *low-data drift* and *unseen data shift* settings (figures 4-5). However, this depends on the dataset (see figures 10-12) and performance is rarely much better than using heuristic augmentation.

Takeaway 6: The best algorithms may differ under the precise conditions. When labels have varying noise in figure 6, relative performance is reasonably consistent. When the dataset size decreases in figure 7, heuristic augmentation methods perform poorly. However, using pretraining and learned augmentation is consistently robust.

Takeaway 7: The precise attributes we consider directly impacts the results. For example, on DSPRITES, if we make color y^l and shape y^a, we find that *all* methods generalise perfectly in the *unseen data shift* setting (as demonstrated in appendix B.3) unlike when shape is y^l (figure 10).

4.2 PRACTICAL TIPS

While there is no free lunch in terms of the method to choose, we recommend the following tips.

Tip 1: If heuristic augmentations promote invariance to nuisance attributes, use them. Following on from *takeaway 3*, heuristic augmentations can significantly improve performance. How to select these augmentations without trying all combinations is an open question.

Tip 2: If heuristic augmentations do not help, learn the augmentation. If some subset of the generative model can be learned by conditioning on known attributes, this is a promising way to further improve performance. However, the utility of such an approach depends heavily on the quality of the learned generative model, which may be more challenging for more complex datasets.

Tip 3: Use pretraining. In general, pretraining was found to be useful to learn robust representations, corroborating other papers such as Kornblith et al. (2019). However, this was not true for all datasets (e.g. CAMELYON17, IWILDCAM where the task is more specific and the datasets are large). An area to be investigated is the utility of self-supervised pre-training (Hendrycks et al., 2019).

Tip 4: More complex approaches lead to limited improvements. Domain generalization, adaptive approaches and disentangling lead to limited improvements across the different datasets and shifts. How to make these approaches generically useful for robustness is still an open question.

5 DISCUSSION

Our experiments demonstrate that no one method performs best over all shifts and that performance is dependent on the precise attribute being considered. This leads to the following considerations.

There is no way to decide a-priori on the best method given only the dataset. It would be helpful for practitioners to be able to select the best approaches without requiring comprehensive evaluations and comparisons. Moreover, it is unclear how to pinpoint the precise distribution shift (and thereby methods to explore) in a given application. This is an important future area of investigation.

We should focus on the cases where we have knowledge about the distribution shift. We found that the ability of a given algorithm to generalize depends heavily on the attribute and dataset being considered. Instead of trying to make one algorithm for any possible shift, it makes sense to have adaptable algorithms which can use auxiliary information if given. Moreover, algorithms should be evaluated in the context for which we will use them.

It is pivotal to evaluate methods in a variety of conditions. Performance varies due to the number of examples, amount of noise, and size of the dataset. Thus it is important to perform comprehensive evaluations when comparing different methods, as in our framework. This gives others a more realistic view of different models' relative performance in practice.

6 Related Work

We briefly summarize benchmarks on distribution shift, leaving a complete review to appendix C.

Benchmarking robustness to out of distribution (OOD) generalization. While a multitude of methods exist that report improved OOD generalization, Gulrajani & Lopez-Paz (2021) found that in actuality no evaluated method performed significantly better than a strong ERM baseline on a variety of datasets. However, Hendrycks et al. (2021) found that, when we focus on better augmentation, larger models and pretraining, we can get a sizeable boost in performance. This can be seen on the Koh et al. (2020) benchmark (the largest boosts come from larger models and better augmentation). Our work is complementary to these methods, as we look at a range of approaches (pretraining, heuristic augmentation, learned augmentation, domain generalisation, adaptive, disentangled representations) on a range of both synthetic and real-world datasets. Moreover, we allow for a fine-grained analysis of methods over different distribution shifts.

Benchmarking spurious correlation and low-data drift. Studies on fairness and bias (surveyed by Mehrabi et al. (2021)) have demonstrated the pernicious impact of low-data in face recognition (Buolamwini & Gebru, 2018), medical imaging (Castro et al., 2020), and conservation (Beery et al., 2018) and spurious correlation in classification (Geirhos et al., 2019) and conservation (Beery et al., 2020). Arjovsky et al. (2019) hypothesized that spurious correlation may be the underlying reason for the poor generalization of models. To our knowledge, there has been no systematic large scale work on understanding the benefits of different methods across these distribution shifts using multiple datasets and with fine-grained control on the amount of shift. We introduce a framework for creating these shifts in a controllable way to allow such challenges to be investigated robustly.

Benchmarking disentangled representations. A related area, disentangled representation learning, aims to learn a representation where the factors of variation in the data are separated. If this could be achieved, then models should be able to generalise effortlessly to unseen data as investigated in multiple settings such as reinforcement learning (Higgins et al., 2017b). Despite many years of work in disentangled representations (Higgins et al., 2017a; Burgess et al., 2017; Kim & Mnih, 2018; Chen et al., 2018), a benchmark study by Locatello et al. (2019) found that, without supervision or implicit model or data assumptions, one cannot reliably perform disentanglement; however, weak supervision appears sufficient to do so (Locatello et al., 2020). Dittadi et al. (2021); Schott et al. (2021); Montero et al. (2020) further investigated whether representations (disentangled or not) can interpolate, extrapolate, or compose properties; they found that when considering complex combinations of properties and multiple datasets, representations do not do so reliably.

7 Conclusion

This work has put forward a general, comprehensive framework to reason about distribution shifts. We analyzed 19 different methods, spanning a range of techniques, over three distribution shifts – *spurious correlation*, *low-data drift*, and *unseen data shift*, and two additional conditions – *label noise* and *dataset size*. We found that while results are not consistent, a number of methods do better than an ERM baseline in some settings. We hope that our framework and comprehensive benchmark spurs research on in this area and provides a useful tool for practitioners to evaluate which methods work best under which conditions and shifts.

REFERENCES

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.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, 2018.

Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 2020.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 2018.

Chris Burgess and Hyunjik Kim. 3D shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. In *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems*, 2017.

Fabio M Carlucci, Paolo Russo, Tatiana Tommasi, and Barbara Caputo. Hallucinating agnostic images to generalize across domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

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

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *International Conference on Intelligent Transportation Systems*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. In *Proceedings of the International Conference on Learning Representations*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *Radiographics*, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations*, 2019.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

Muhammad Waleed Gondal, Manuel Wüthrich, Dorde Miladinović, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *arXiv preprint arXiv:1906.03292*, 2019.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Sven Gowal, Chongli Qin, Po-Sen Huang, Taylan Cemgil, Krishnamurthy Dvijotham, Timothy Mann, and Pushmeet Kohli. Achieving robustness in the wild via adversarial mixing with disentangled representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Keren Gu, Xander Masotto, Vandana Bachani, Balaji Lakshminarayanan, Jack Nikodem, and Dong Yin. An instance-dependent simulation framework for learning with label noise. *arXiv preprint arXiv:2107.11413*, 2021.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *Proceedings of the International Conference on Learning Representations*, 2021.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Will Douglas Heaven. Google's medical ai was super accurate in a lab. real life was a different story., 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*, 2018.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the International Conference on Machine Learning*, 2019.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *Advances in Neural Information Processing Systems*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *Proceedings of the International Conference on Computer Vision*, 2021.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations*, 2017a.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017b.

Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 2020.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *Proceedings of the International Conference on Learning Representations*, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the International Conference on Machine Learning*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Yann LeCun, Fu Jie Huang, and Léon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the International Conference on Computer Vision*, 2017.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision*, pp. 624–639, 2018.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving group robustness without training group information. In *Proceedings of the International Conference on Machine Learning*, 2021.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the International Conference on Machine Learning*, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 2021.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *arXiv preprint arXiv:2102.11273*, 2021.

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *Proceedings of the International Conference on Learning Representations*, 2020.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, 2010.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the International Conference on Computer Vision*, 2019.

Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International Conference on Machine Learning*, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *Proceedings of the International Conference on Learning Representations*, 2020.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Proceedings of the International Conference on Learning Representations*, 2020.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *Proceedings of the International Conference on Machine Learning*, 2012.

Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *Proceedings of the International Conference on Learning Representations*, 2021.

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? *arXiv preprint arXiv:1906.02168*, 2019.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision*, 2016.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, 1992.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or Signal: The role of image backgrounds in object recognition. *ArXiv preprint arXiv:2006.09994*, 2020.

Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *Proceedings of the International Conference on Learning Representations*, 2020.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI Conference on Artificial Intelligence*, 2020.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. MixUp: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*, 2018.

Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 2019.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI Conference on Artificial Intelligence*, 2020.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision*, 2017.