Causal Transportability for Neural Representations

Chengzhi Mao[1] James Wang[1] Kevin Xia[1] Hao Wang[2] Junfeng Yang[1]
Elias Bareinboim[1] Carl Vondrick[1]

Abstract

Neural network representations often contain both robust and non-robust features that can be used in training. Existing vision classifiers fail on out-of-distribution samples because non-robust features are spurious correlations that can be changed in a new environment. We first analyze procedures for out-of-distribution generalization with a causal graph. We then introduce a causally motivated algorithm, which unlearns the non-robust representations from deep models and promotes the model's out-of-distribution generalization. We show that representations in deep models can be a feasible front-door mediating variable under empirical designs. We then use the established front-door criteria to estimate the causality for the prediction. Theoretical analysis, empirical results, and visualizations show that our approach estimates causal invariance and achieves improved generalization.

1. Introduction

The explosive growth of foundation models and representation learning has transformed computer vision [14, 15, 30, 41]. By learning from large image datasets, deep neural networks have been able to create excellent visual representations that improve many downstream vision tasks [14, 15]. However, central to this framework is the need to generalize to different visual distributions that are unforeseen at training time [2, 3, 6, 9, 19, 24, 39, 42, 57]. After learning from the Internet, for example, the representation needs to generalize to new environments in the real world.

The most popular technique to use representations is to fine-tune the backbone model or fit a linear model on the classification task [30]. Although this approach is effective on in-distribution benchmarks, the resulting classifier also inherits the biases from the original representation. This is because some undesirable features in the representation can correspond to spurious correlations in the training data, and hence the model acquires. The possibilities for spurious features are extensive, impacting the generalization and fairness of computer vision systems.

In this paper, we use the lens of causality to revisit how to make the right use of visual representations. We formulate a structured causal model for out-of-distribution image classification and show how different environments select a different set of robust and non-robust features in constructing the input. The training environment may tend to select specific nuisances with the given category, creating spurious correlations between the nuisances and the predicted class. Standard neural classifiers will use those spurious correlations, which analytically explains why they result in poor generalization performance to novel target distributions [23, 45, 50].

The causal graph allows us to use a rich toolbox from causal inference for identifying the robust features from observational data [7, 8, 11, 16, 32, 43]. The causal effect from the input to the output is invariant when the environment changes. Our main result is that image classification from deep representations is compatible with one of the most standard principles for causal effect identification based on the front-door adjustment [38, Sec. 3.3.2]. Natural images are challenging for causal identification because there are innumerable confounding factors with realistic data. However, the front door adjustment does not require observations of these factors, and instead only requires a mediating mechanism between the image and the labels. Under practical assumptions, we show that the learned neural representation can be a feasible front-door mediating variable, which allows us to disentangle and discard the spurious correlation from the representation using the front-door adjustment.

For both supervised and self-supervised representations, our experimental results show that incorporating the causal structure improves performance when generalizing to new domains. Our method is compatible with many existing representations without requiring re-training, making the approach particularly effective to deploy in practice. Compared to the standard techniques to use representations, which learns correlation, our causally motivated approach can obtain significant gain on simulated CMNIST (up to 40% gain), WaterBird (up to 25% gain), ImageNet-9 (up to 5%), ImageNet-Sketch (up to 8% gain), and ImageNet-Rendition (up to 7%) datasets. Our work is the first to con-

[1]Columbia University
[2]Rutgers University

nect deep representations to the front-door criteria in causal inference, and propose an empirical method that allows the learning of causality via the front-door criteria.

2. Related Work

Causal Inference. The connection between causality and generalization has been studied in past works [4,33,34]. Causal inference provides a principled framework for modeling the transportability of machine learning models across environments [7, 8, 11, 16, 32, 43]. For image classification, while a few works assume that labels cause images [22, 44], our work follows the set up that images causes the labels [4, 34, 52]. To estimate the causal effect, existing work on causality and generalization often assumes one can intervene on the data [28, 34] or exhaust all confounding factors [28, 53]. These assumptions are often overly optimistic for natural images, as most image data today is passive (preventing intervention) and there are too many confounding factors to enumerate. In addition, causal matching across different domains improves generalization [33] if multiple domains are available at training time.

Out of distribution Generalization. A large number of approaches have been proposed to learn classifiers that generalize to out-of-distribution and new environments [2, 6, 23, 39, 50, 57]. There are two major types of domain generalization(DG): the multi-source DG and the single-source DG. Multi-source domain generalization has been studied [1, 4, 10, 31, 48, 55], where the algorithm knows the domain index which the data points are sampled from. However, it is often challenging to collect images with accurate domain labels, such as from the Internet. Single domain generalization [22] does not require the domain index assumption, where all training data are assumed to be sampled from the same domain. However, domain generalization under this setup is more challenging due to lacking the domain information. Existing work achieves generalization via self-supervised learning [12], anticipating distribution shifting [40], creating pseudo domain split [36], adversarial self-challenging [26], and generative data augmentations [34]. Recently, the attention operation is also shown to be effective for improving robustness [18, 37]. However, a principled framework for modeling generalization to new environments is still missing.

3. Problem Formulation

Before we present our method, we ground the problem in a causal framework. We formulate a causal graph for image recognition and analyze its properties.

3.1. Causal Graph

We study how to train a classifier that can generalize to new image distributions. We denote the input image as X and its label as Y. We assume training data is sampled from environment π, and the testing data is sampled from a different environment π^*. When the training set and testing set are sampled from the same environment, existing classifiers generalize well. However, their performance drops significantly when tested on out-of-distribution samples.

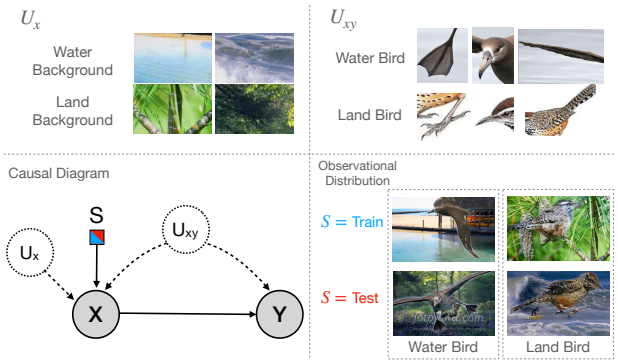


Figure 1. Causal graph for out-of-distribution image classification. We denote the set of nuisances features as U_x, and the core features for creating the category as U_{xy}. The out-of-distribution data is created due to the change of the environment S, where X combines 'waterbird' with 'water background' during the training and 'water bird' with 'land background' at testing.

To develop intuition on why generalization is challenging, we formulate a causal graph that makes explicit the causal mechanisms that generate the observed visual distributions. Causal graphs provide a powerful framework to integrate domain knowledge about images into visual classification. We draw our causal graph G in Figure 1, where the observational data distribution is $P(X, Y)$. We assume that there are underlying physical factors U_{xy}, which produce the core features in the image X and the corresponding label Y. For example, U_{xy} can contain the key shape of a waterbird and landbird. There are also nuisances factors U_x, such as the background or camera viewpoint, which are not the direct cause of the label Y, but affect the construction of input images X. For example, let the variable U_x denote the background, which contains 'water' and 'land'. Then each realized value u_x contains one realized water background (e.g., the ocean) and one realized land background (e.g., the forest) together in a set. Both U_x and U_{xy} produce the input image X. The ground truth label Y is caused by both the input image due to the annotation procedure of image recognition and the underlying features U_{xy}.

To understand out-of-distribution inference from a causal perspective, we consider different causal models for the training and test environments. The difference lies in the mechanism for f_X, which chooses features from U_x and U_{xy} to construct the image X, i.e. $X = f_X(U_x, U_{xy})$. In the causal diagram, we represent the difference with a switch node (S). S is a binary variable that decides which function to use (i.e. $S = 0$ refers to the f_X in the training domain, $S = 1$ for the testing domain). For example,

the training environment $S = 0$ often selects $water$ background from the realized nuisances set $u_x = \{water, land\}$ when it is given $u_{xy} = waterbird$, and the testing environment $S = 1$ often selects $land$ background from the nuisances set given by U_x when $U_{xy} = waterbird$.

3.2. Analysis of Existing Approach

The most common approach in visual recognition today is to formulate an optimization problem that minimizes the empirical risk, which uses any discriminative signal in the representation that correlates with the output label. For example, this approach will use both the water background and the bird shape to predict the waterbird category, because both features help reduce the loss.

When models are evaluated on a test set sampled from the same environment as the training set S, both the foreground and background create features that increase test accuracy, as the joint distribution $P(X, Y)$ is the same across training and testing. However, when the environment S changes, the joint distribution from the source domain $P(X, Y)$ is no longer the same as the joint distribution from the target domain $P^*(X, Y)$ due to the change in S. The original classifier is no longer effective because it has relied on both the foreground information and background information for prediction. Simply learning the correlation $P(Y|X)$ can fail to generalize on out-of-distribution data if the selection functions (f) of the domains π, π^* are different ($f_x \neq f_x^*$), even though everything else remains invariant.

However, generalization to novel environments is possible if we are able to learn the invariant signals. In our causal graph G in Figure 1, the only property that is invariant to the environment S is the causal path from X to Y, which is denoted as $P(Y|\text{do}(X = x))$. The $\text{do}(X = x)$ operation constructs an intervention to image X with x while keeping everything else the same. This is equivalent to removing all incoming edges to the variable X and setting X's value to x. After removing all incoming edges to X, the correlation from image X to the predicted label Y is the same across both environments.

Intuitively, given that we are talking about the interventional world, where the arrows towards X can be thought as removed, the effect of the selection mechanism is severed as well in both domains, which produces the same causal graph. Thus, our goal is to learn a model to approximate this causal effect $P(Y \mid \text{do}(X))$.

4. Approach

Based on this causal graph and analysis, we formulate how to identify the causal effect and learn classifiers than generalize across environments.

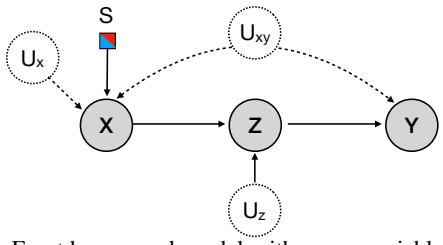


Figure 2. Frontdoor causal model with proxy variable Z. Gray nodes denote observed variables.

4.1. Causal Invariance

If we can estimate the invariant relationship $P(Y \mid \text{do}(X))$ between the input and the output task, then robust generalization is possible. However, from the given causal diagram in Figure 1, we cannot identify the causal effect $P(Y|\text{do}(X))$ from the observation distribution $P(X, Y)$ alone because there can be two different causal models that produce the same observation distribution $P(X, Y)$.

However, if we can add additional assumptions about the underlying data-generating model, we can obtain a refined causal diagram that is identifiable. One popular refinement is to assume we can observe the back-door variables, where the model assumes *all* the variations represented originally in the unobserved confounder U_{xy} is now observed [38, Sec. 3.3.1]. For example, prior work assumes that they can observe or estimate the unseen backdoor variables, such as the backgrounds, context, and style [34, 54]. By conditioning on the back-door variables when running correlation-based algorithms from the input image X to the output category Y, the spurious correlation from the back-door path can be disentangled from the correlation between X and Y. As a result, the correlation between X and Y will be the same even if the spurious correlations change from the environment. This assumption has been used in prior works such as IRM [4], MLLD [36], and DANN [1], where they assume the domain index is the back-door variable to be invariant to. However, the domain index is often not available on large scale datasets, such as ImageNet, which has only one training domain [17]. In addition, a major assumption for the success of the back-door based approach is to be able to exhaust all the backdoor variables. However, the variation and complexity of natural images makes this assumption unreasonable.

Our main observation is that we can create a mediating variable along the path from X to Y leveraging the representation in a deep network, which allows us to apply the front-door criterion without making assumptions about exhausting back-door variables. We denote the mediating variable as Z and introduce the front-door criteria.

4.2. The Front-door Criteria

To treat the representations in a neural network as a feasible front-door mediating variable, as shown in Figure. 2,

there are three properties we need to ascertain to satisfy the front-door condition:

1. There is no direct causal path from the latent confounding variable U_{xy} to the representation vector Z, i.e., $U_{xy} \perp\!\!\!\perp Z|X$. This is true because the representations are calculated from only the input image X without using U_{xy}.

2. There is no path that goes from X to Y directly, i.e., $Y \perp\!\!\!\perp X|Z, U_{xy}$. In words, all the information in X that causes Y is maintained in Z, such that a human annotator will label the same Y by looking at only X or only Z.

3. The representation Z cannot be a copy of the input X. This can be achieved by adding additional variance to the representation.

We describe how to construct a valid front-door causal graph in Section 4.3, we then can apply the front-door criteria to estimate the causal effect.

$$P(y|\text{do}(X = x)) = \sum_z P(z|x) \sum_{x'} P(y|z, x')P(x') \quad (1)$$

Intuitively, instead of estimating the causality from input X to output Y directly, the front-door criteria first estimates the causal effect from the input image X to the latent representation Z. This is simple as the causal effect from input to the representation is the same as the correlation between them, as there is no backdoor path from X to Z. It then estimates the causal effect from the representation Z to the output label Y. As there is a backdoor path $Z \rightarrow X \rightarrow U_{xy} \rightarrow Y$, the correlation between Z and Y contains spurious ones, which can be removed by conditioning on the observed input image X on the backdoor path (backdoor criteria).

To use the above formulation, we need to construct a valid front-door mediating variable Z, and also estimate $P(X)$, $P(Z|X)$ and $P(Y|X, Z)$. The term $P(X)$ is straightforward to calculate because we can assume it to be sampled from a uniform distribution. The other terms, however, need to be carefully constructed to satisfy the front-door criteria, which are discussed in the following sections.

4.3. Constructing $P(Z|X)$

We show that there are some classes of models that are valid ways to estimate $P(Z|X)$ for the front-door variable.

Variational Auto-Encoder (VAE) [29] is a major unsupervised representation learning approach, which aims to estimate latent distribution Z that can faithfully generate the input distribution. It maximizes the evidence lower bound for the distribution of X: $\mathcal{L} = -D_{KL}(q_E(z|x^{(i)})||p_\theta(z)) + \text{E}_{q_E(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)]$, where E is the encoder in the VAE. The latent representation Z of VAE models naturally satisfies the three properties for the front-door criteria. As VAEs are optimized to reconstruct input images via the term $\text{E}_{q_E(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)]$, the representation Z should contain all the causal information from the input images (which is validated by empirical results where images decoded from VAE can still be visually discriminated), and therefore Z can block all the information from X to Y. The representation Z is calculated via $q_E(z|x^{(i)})$ using only X, which is not directly caused by the confounders. VAE models are sampled from the latent distribution with additional randomness that is independent of the input image X, which means it is not a direct copy of the input X and enables us to estimate the term $P(z|x)$ in the front-door criteria.

Algorithm 1 Front-door Model Training

Input: Training set D over $\{(X, Y)\}$.
Phase 1: Compute $P(Z|X)$ from representation of VAE or pretrained model.
Phase 2:
for $i = 1, ..., K$ **do**
  Sample x_i, z_i, y_i from the joint distribution $D' = (X, Z, Y)$
  Random sample x'_i from the same category as x_i
  Train $P(Y|X', Z)$ via minimizing the classification loss $\mathcal{L}$ through gradient descent.
end for
Output: Model $P(Z|X)$ and $P(Y|X, Z)$

Constrastive Learning is trained to produce representations that can discriminate pairs of images augmentations from the same instance. It has been shown that contrastive learning inverts the generative process [58] and reduces the (labeled) sample complexity on classification tasks [5] than on raw image input. This means that contrastive representations have larger mutual information with the predicted label Y. As constrastive learning are trained to be invairant under data augmentaion, it still maintains all causal information from the input images (property 2). Since representations in contrastive learning are calculated via a deep network using only X, they are not directly caused by the confounders. In addition, by random sampling the representation, we can make sure Z is not a direct copy of the input image X.

4.4. Constructing $P(Y|Z, X)$

To perform the front-door adjustment, we also need to estimate $P(Y|Z, X)$, which we do with a neural network. However, estimating the probability of $P(Y|Z, X)$ is challenging in our setting because $P(Y|Z, X) = \frac{P(Y,Z,X)}{P(Z,X)}$, where the probability density estimation can be incorrect if $P(Z, X) \approx 0$. This happens when the data pair (z, x) is not covered in the training distribution and the learned model

Algorithm 2 Front-door Causal Inference

Input: Query x, training distribution D over $\{(X, Y)\}$, model $P(Z|X)$ and $P(Y|X', Z)$, the sampling time N_i for the mediating variable Z, and the sampling time N_j for X'.
for $i = 1, ..., N_i$ **do**
  $\mathbf{z}_i \leftarrow P(z|x)$
  for $j = 1, ..., N_j$ **do**
    Random sample $\mathbf{x}'_{ij}$ from Training Distribution D.
    Compute $P(Y|x'_{ij}, z_i)$
  end for
end for
Calculate the causal effect $P(y|\text{do}(X = x)) = \sum_i P(z_i|x) \sum_j P(y|z_i, x'_{ij}) P(x'_{ij})$
Output: Class $\hat{y} = \text{argmax}_y P(y|\text{do}(X = x))$.

does not extrapolate well.

There are two major reasons for this to occur in our application: First, as we train on the domain π and generalize to the out of distribution domain π^*, the data pair (Z^*, X) at inference time is unlikely to have a similar data pair in the training distribution, especially for image data of high dimensionality. Second, the front-door inference algorithm requires sampling random images X to pair with the query Z at the inference time, while the training algorithm by default samples the data pairs (z, x) from the same image instance. To avoid the above issues, we need to carefully formulate the data choice and the model design.

Data Choice for Training $P(Y|Z, X)$. We can mitigate the problem of $P(Z, X) \approx 0$ by training on a wider range of Z and X. To improve the coverage of Z, we add significant amount of noise to the representation vector when creating the mediating variable Z, such that the resulting variable Z has a larger coverage of the space due to the increased variance. To improve the coverage of X given Z, we sample x' from the same category as X and construct the data (z, x') instead of sampling x from the same image instance as z. This requires the assumption that data from the same category often share the same confounding bias in the same environment.

Model Design for $P(Y|Z, X)$. While adding noise with a large variance to data Z can help estimate the probability density $P(Y|Z, X)$, it may cause $P(Y|Z, X) = P(Y|X)$ due to the increased variance in Z. Ideally, we need $Y \not\perp\!\!\!\perp Z|X$ so that there is some additional information from Z about Y than X. We can explicitly apply this constraint by optimizing the mutual information, $I(Y; Z|X) > 0$, in our learning objective. For simplicity, we can empirically limit the information that is used from X through the inductive bias from the architecture of the neural network. Specifically, we discard some pixels from the input image X with convolution operation that has larger stride size than the kernel size. This enables $Y \not\perp\!\!\!\perp Z|X$, as X is subsampled and does not contain all the information of Z; the trained neural network will then be encouraged to use both the information from X and latent representation Z. In addition, prior work on self-supervision shows that one can predict Y with higher accuracy using self-supervised representations Z than just learning from X [14, 25, 35]. Thus, the representation Z from self-supervision can provide more information to predict Y than just using X, which promotes $Y \not\perp\!\!\!\perp Z|X$. Furthermore, by restricting the capacity of the neural network with shallow architecture and fewer parameters, we can also smooth the prediction $P(Y|Z, X)$.

Intuitively, by limiting the capacity of $P(Y|Z, X)$, the model tends to attend to low level features from the input images X while using high level deep features from the latent representation Z. Traditional correlation based approach only uses $P(Y|Z)$, which can also include spurious features such as the texture and backgrounds. With our approach, the low level spurious features tend to be learned by the model that condition on the input pixels X directly, and the model will discard those features after marginalizing over the variable X in the front-door formulation.

4.5. Algorithm

We describe our training procedure in Algorithm 1. In the first phase, we estimate $P(Z|X)$, where we either train representation with our proposed VAE or contrastive learning approach, or we use representations from a pretrained deep model. In the second phase, we train $P(Y|X, Z)$ where we sample random images X from the same category as the representation Z. We describe our inference procedure in Algorithm 2, where we infer the $P(y|\text{do}(X = x))$ via the front-door criteria. We first randomly sample Z. Then, for each Z, we sample images X from random categories. We then make the prediction through the front-door criteria.

4.6. Imperfect Representations

Even when a supervised learning algorithm cannot guarantee to contain all the causal information from the input X to predict the output Y, our approach can still discard the incorrect spurious features from the imperfect representation. We consider such causal graph in Figure 3. We assume X can be separated into two sub-variables, X_1 and X_2, where X_1 denotes the information that learned by our representation, and X_2 denotes

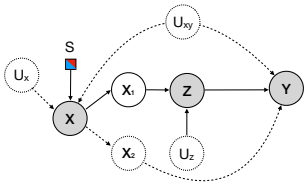


Figure 3. Causal diagram under imperfect front-door variable.

	Test Accuracy	
	In-Distribution	Out-of-Distribution
Chance	10%	10%
ERM [49]	99.45%	8.26%
IRM* [4]	87.32%	18.49%
JiGEN [12]	99.11%	10.29%
M-ADA [40]	**99.93%**	12.73%
DG-MMLD [36]	99.09%	12.53%
RSC [27]	96.58%	20.62%
GenInt [34]	58.48%	29.62%
Ablation	97.41%	38.82%
Ours	82.85%	**51.40%**

Table 1. Accuracy on the CMNIST dataset. Our method advances the state-of-the-art GenInt [34] method by over 20% on the out-of-distribution test set.

the information that is missed. The learned representation blocks all the information that flows from X_1 to the output Y. There is another causal path from X_2 to Y that is unblocked. Thus, the traditional way of using representation estimates the result via $P(Y|X_1)$, and our algorithm estimates via $P(Y|\text{do}(X = x_1))$. While both of approaches cannot capture the causal effect $P(Y|\text{do}(X = x_2))$, our approach can capture a subset of the causal features.

5. Experiment

5.1. Datasets

CMNIST. We use the more challenging setup of colored MNIST dataset with 10 categories [34]. The function $F_X(U_x, U_{xy})$ will combine digits with different background colors from the training domain, creating an out-of-distribution (OOD) dataset. **WaterBird** [45] The Waterbird dataset [46] contains two classes of foreground birds, the waterbird and the landbird, and two types of backgrounds: water and land. The testing is OOD to the training because of the different mechanism in combining the foreground and background. **PACS** [6] contains seven classes over four domains (Art, Cartoon, Photo, Sketch). Models are evaluated by training on three domains and test on the fourth domain. **VLCS.** [2] contains five classes over four domains. Models are evaluated by training on three domains and test on the fourth domain. **ImageNet-Rendition** [23] has renditions of 200 ImageNet classes, including art, cartoons, etc, which is an OOD test set for ImageNet. **ImageNet-Sketch** [50] contains sketch of 1000 ImageNet classes, which evaluate classifiers' robustness without texture and color clue. **ImageNet-9** [51] has 8 variations of ImageNet validation set to measure the impact of backgrounds. It contains Mixed-Same, Mixed-Rand, Mixed-Next, etc. **ImageNet Backgrounds Challenge** [51] studies the classifier's vulnerability to adversarially chosen backgrounds.

5.2. Baselines

Our paper studies generalization on the out-of-distribution test set without domain index for training samples. We compare with the following 6 baselines below:

ERM [20, 49] is the standard way to train deep network classifiers. *JiGEN* [12] uses self-supervision to improve generalization. *M-ADA [40]* uses adversarial training to anticipate OOD populations for generalization. *DG-MMLD* [36] uses the EM algorithm to discover latent domains, and then trains domain-invariant model using the generated pseudo domain labels. *GenInt* [34] learns a causal classifier by steering the generative models to simulate interventions. *RSC* [27] uses representation self-challenging to improve generation to the OOD data, where features that are significant in ERM will be punished. We also compare with the popular IRM [4] which uses domain index information.

Method	Domain ID	Train	In-Distribution	Out-of-Distribution
GDRO* [46]	Yes	100.0%	**97.4%**	76.9%
ERM	No	100.0%	97.3%	52.0%
RSC	No	92.2%	95.6%	49.7%
Ablation	No	99.4%	96.8%	71.6%
Ours	No	99.4%	96.8%	**77.9%**

Table 2. Accuracy on the WaterBird dataset. Our causal method improves ERM model's worst group OOD generalization significantly. Our approach achieves performance on par with group invariant training (GDRO) without needing the domain index.

5.3. Experimental Settings

We construct $P(Y|X', Z)$ by first using a 3-layer convolution network to process the input X', concatenating the obtained feature with Z, and then using 2-layer fully connected network to predict Y. We set $N_j = 256$ and $N_i = 10$ for all experiments and denotes it as **Ours**. We also conduct a variant with $N_j = 1$ and $N_i = 1$ and denote it as **Ablation**, where everything is the same as 'Ours' but the inference procedure is a traditional single forward pass. For CMNIST, WaterBird, PACS, VLCS dataset, we select the model with the highest validation accuracy. For ImageNet-Rendition, ImageNet-Sketch, ImageNet-9, and the Background Challenge, we report the validation accuracy as there is no validation/test split available.

5.4. Results on Simulated Datasets

CMNIST. Our approach uses the latent representation from VAE to construct the front-door variable. We report the accuracy in Table 1. Our front-door method outperforms existing methods including the causal GenInt method by over 20%.

WaterBird. Following prior work, we use the representation from a pre-trained ResNet50. We train the model for 20 epochs and randomly drop 90% of the features to construct Z. We show the result in Table 2. Without using domain index information, our causal approach improves the worst group test performance by over 25% compared with ERM, and even 1% higher than the state-of-the-art GDRO [46] method which uses domain index information.

BackBone	Algorithm	MixNext	MixRand	FG	MixSame	NoFG ↓	only BG B ↓	only BG T ↓	Original	Adversarial
Moco-v2	ERM	75.43%	76.25%	82.47%	85.09%	39.88%	13.09%	13.21%	92.40%	14.59%
Moco-v2	Ablation	74.39%	75.67%	77.48%	84.37%	36.64%	8.10%	11.08%	94.02%	17.04%
Moco-v2	Ours	**76.27%**	**77.43%**	**83.13%**	**85.45%**	**37.55%**	**7.21%**	**10.10%**	**94.44%**	**18.02%**
SWAV	ERM	69.65%	72.47%	80.88%	83.06%	37.21%	10.76%	14.44%	94.64%	20.00%
SWAV	Ablation	76.79%	78.14%	81.87%	85.25%	41.48%	**8.41%**	10.96%	94.79%	20.25%
SWAV	Ours	**78.00%**	**79.70%**	**85.88%**	**85.68%**	**41.18%**	9.16%	**10.17%**	**94.98%**	**20.42%**
SimCLR	ERM	80.46%	82.64%	88.22%	88.27%	53.11%	20.13%	16.74%	95.35%	27.73%
SimCLR	Ablation	82.17%	83.87%	85.67%	89.26%	48.27%	**12.67%**	11.31%	95.43%	28.44%
SimCLR	Ours	**83.18%**	**84.86%**	**89.33%**	**89.65%**	**47.85%**	13.72%	**10.69%**	**95.77%**	**29.41%**

Table 3. Model accuracy on variants of ImageNet-9 classification. ↓ indicates lower value is better. Our causal models tend to make the prediction with the right foreground objects, and getting lower accuracy if foreground objects are absent.

Algorithm	PACS Out-of-distribution accuracy					VLCS Out-of-distribution accuracy				
	Art	Cartoon	Sketch	Photo	**Average**	Caltech	LABELME	PAS	SUN	**Average**
IRM* [4]	71.32%	57.30%	74.30%	93.11%	74.01%	94.79%	61.78%	65.19%	72.57%	73.58%
M-ADA [40]	64.29%	72.91%	67.21%	88.23%	73.16%	74.33%	48.38%	45.31%	33.82%	50.46%
JiGen [12]	79.42%	75.25%	71.35%	96.03%	80.51%	96.17%	62.06%	70.93%	71.40%	75.14%
DG-MMLD [36]	81.28%	77.16%	72.29%	96.09%	81.83%	97.01%	62.20%	73.01%	72.49%	76.18%
RSC [27]	83.43%	80.31%	80.85%	95.99%	85.15%	96.21%	62.51%	73.81%	72.10%	76.16%
ERM [20]	76.61%	73.60%	76.08%	93.31%	79.90%	91.86%	61.81%	67.48%	68.77%	72.48%
Ablation	83.39%	81.18%	82.13%	95.33%	85.51%	98.11%	66.87%	76.30%	72.48%	78.59%
Ours	**83.54%**	**81.31%**	**82.71%**	**96.70%**	**86.06%**	**98.58%**	**67.12%**	**78.58%**	**73.40%**	**79.12%**

Table 4. Classification accuracy on the domain generalization dataset PACS and VLCS. Models are trained on three domains and tested on the unseen fourth domains. We use *ResNet18* for all the models. All approaches do not use the domain index information except for the IRM. Our causal approach outperforms the existing methods.

Algorithm	ImageNet Rendition				ImageNet Sketch			
	ERM	RSC	Ablation	Ours	ERM	RSC	Ablation	Ours
Moco-v2	26.92%	26.14%	25.96%	**28.70%**	17.29%	16.43%	14.11%	**19.09%**
SWAV	31.77%	30.47%	30.32%	**33.32%**	21.51%	21.03%	17.26%	**22.48%**
SimCLR	37.82%	34.06%	35.74%	**38.25%**	27.43%	19.26%	24.90%	**29.51%**
ResNet50	25.02%	**33.34%**	30.96%	32.22%	14.45%	22.54%	19.19%	**22.57%**
ResNet152	30.53%	**37.86%**	34.94%	36.07%	18.53%	26.60%	24.61%	**27.07%**
ResNet101-2x	31.44%	35.50%	35.82%	**36.70%**	19.92%	26.38%	25.07%	**27.41%**

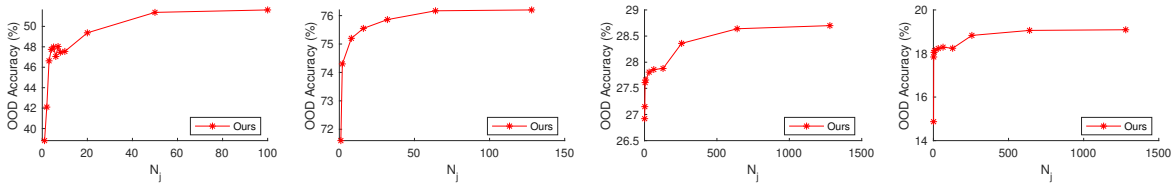
Table 5. Robust accuracy on ImageNet-Rendition and ImageNet-Sketch. For contrastive learning based representations, our model achieves improved robustness than standard ERM and the state-of-the-art RSC approach. On superivsed learning representations, the representation may fail to capture all the causal information, where RSC method out-performs ours on two variants on ImageNet Rendition. Overall, our method improves robustness by estimating the causal effect from the representation.

ImageNet-9 and Adversarial Backgrounds. We assess our model's robustness on testing distributions where the foreground and the background are manipulated to be different from the training distribution. In Table 3, after training on the ImageNet, we report models' accuracy on 8 foreground-background variants, including applying background from different categories to the foreground object ('MixNext', 'MixRand'), foreground object only ('FG'), and no foreground object ('NoFG', 'only BG B', 'only BG T', where lower values indicate better performance). We also test on images with adversarially chosen backgrounds. We experiment on three variants of contrastive loss based self-supervised learning approaches, including Moco-v2 [15], SWAV [13], and SimCLR [14]. Overall, our approach performs better when foreground object is present even if background is changed.

5.5. Real-world Out of Distribution Generalization

PACS and **VLCS** are two major OOD domain generalization datasets. Following prior work [12, 40, 56], we use the penultimate representation layer from a pre-trained ResNet-18 [21] model to create the front-door variable Z. We train our model on a mixture of any three of the domains (we do not use the domain index information) and test on the fourth domain. We show results in Table 4. Overall, our front-door-based causal approach outperforms the baselines on both datasets. Our results show that our causal approach is more robust in generalizing to new distributions.

ImageNet-Rendition and **ImageNet-Sketch** are two OOD test sets for ImageNet. We study the representation from constrastive loss based self-supervision learning approaches including SimCLR, MoCo-v2, and SWAV. In addition, we also study the representations from supervised



(a) CMNIST, $K = 10$ (b) WaterBird, $K = 2$ (c) Rendition, $K = 1000$ (d) Sketch, $K = 1000$

Figure 4. OOD generalization accuracy under different number of N_j. At inference time, by increasing N_j that samples more images X', OOD generalization improve because the spurious correlation is better removed through the front-door criteria.

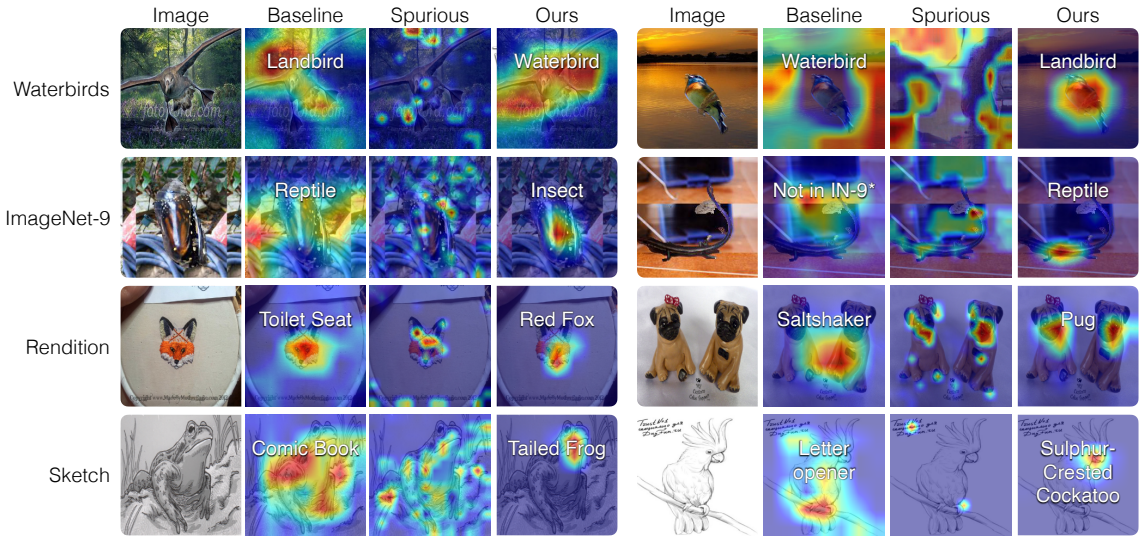


Figure 5. We visualize the input regions that the models use for prediction. We use GradCAM [47] and highlight the the discriminative regions that the model relies on with red. The white text shows the model's prediction. The correlation based ERM method often attends to spurious background context. By marginalizing over the spurious features (visualized in the Spurious column), our front-door model captures the right, causal features, which predict the right thing for the right reason.

learning, though they may be imperfect representations. We show results in Table 5. Our front-door algorithm estimates the causal invariance, which improves OOD generalization. The exception is that the supervised trained models, ResNet50 and ResNet152, are not trained with contrastive learning and therefore may lose causal information. Such loss can be explained by our discussion on imperfect representations in Section 4.6.

5.6. Analysis

Importance of Front-door Sampling. Our front-door adjustment requires to marginalize over the input images X at inference time. Sampling less X can speed up the inference, however, at a cost of not estimate the accurate causal effect. In Figure 4, we vary the number for front-door sampling N_j and test the peroformance on four datasets. In general, We find for datasets with K categories, using $N_j > K$ can significantly improves generalization.

GradCam Visualization. Using the front-door criteria that estimate the causal effect, we expect our model to attend to the spatial regions corresponding to the object, instead of the spurious context. In Figure 5, we validate this by visualizing the regions that the models use for classification with the GradCAM [47]. We examine on four datasets, including the WaterBird, ImageNet-9, ImageNet-Rendition, and ImageNet-Sketch. We visualize the ERM model in the 'Baseline' column, the branch that condition on the variable X of model $P(Y|Z, X)$ in the 'Spurious' Column, and our front-door based causal method in 'Ours'. By discarding the information in the 'Spurious' model through marginalizing over X', our model focus on the right object for prediction.

6. Conclusion

We present a causal inference framework that learns causality from deep representations for out-of-distribution generalization. We first formulate the problem of domain generalization into a causal model, and then propose effective algorithm that estimate the invariant causal effect through the front-door criteria. Our results demonstrate improved out-of-distribution robustness on both simulated and real-world datasets. Our findings suggest integrating causal structure into deep representation is a promising direction to improve generalization.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, Franccois Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014. 2, 3

[2] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H. Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching, 2020. 1, 2, 6

[3] Julian Alverio William Luo Christopher Wang Dan Gutfreund Josh Tenenbaum Andrei Barbu, David Mayo and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *In Advances in Neural Information Processing Systems 32*, page 9448–9458, 2019. 1

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 2, 3, 6, 7

[5] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 4

[6] Nader Asadi, Amir M. Sarfi, Mehrdad Hosseinzadeh, Zahra Karimpour, and Mahdi Eftekhari. Towards shape biased unsupervised representation learning for domain generalization, 2020. 1, 2, 6

[7] Elias Bareinboim and Judea Pearl. Causal transportability with limited experiments. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, page 95–101. AAAI Press, 2013. 1, 2

[8] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. 1, 2

[9] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 05 2010. 1

[10] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017. 2

[11] Peter Bühlmann. Invariance, causality and robustness, 2018. 1, 2

[12] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles, 2019. 2, 6, 7

[13] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 7

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 5, 7

[15] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 7

[16] Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10902–10912. Curran Associates, Inc., 2020. 1, 2

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3

[18] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2

[19] 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 *International Conference on Learning Representations*, 2019. 1

[20] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020. 6, 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 7

[22] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017. 2

[23] 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. *ICCV*, 2021. 1, 2, 6

[24] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 1

[25] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019. 5

[26] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *arXiv preprint arXiv:2007.02454*, 2, 2020. 2

[27] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 6, 7

[28] Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions, 2020. 2

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 4

[30] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 1

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[32] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1, 2

[33] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 2

[34] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning, 2021. 2, 3, 6

[35] Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 661–671, October 2021. 5

[36] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020. 2, 3, 6, 7

[37] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021. 2

[38] Judea Pearl. Causality: Models, reasoning, and inference, 2000. 1, 3

[39] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 1, 2

[40] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12556–12565, 2020. 2, 6, 7

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1

[42] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1

[43] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. 1, 2

[44] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018. 2

[45] 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. *arXiv preprint arXiv:1911.08731*, 2019. 1, 6

[46] 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, 2020. 6

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[48] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation, 2016. 2

[49] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992. 6

[50] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 1, 2, 6

[51] 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. 6

[52] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8599–8608, 2021. 2

[53] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2734–2746. Curran Associates, Inc., 2020. 2

[54] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000*, 2020. 3

[55] 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 preprint arXiv:2007.02931*, 2020. 2

[56] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization, 2021. 7

[57] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation, 2020. 1, 2

[58] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process, 2021. 4