Causal Transportability for Visual Recognition

Chengzhi Mao[1*] Kevin Xia[1*] James Wang[1] Hao Wang[2]
Junfeng Yang[1] Elias Bareinboim[1] Carl Vondrick[1]
[1]Columbia University [2]Rutgers University
{mcz,kevinmxia,jlw2247,junfeng,eb,vondrick}@cs.columbia.edu, hoguewang@gmail.com

Abstract

Visual representations underlie object recognition tasks, but they often contain both robust and non-robust features. Our main observation is that image classifiers may perform poorly on out-of-distribution samples because spurious correlations between non-robust features and labels can be changed in a new environment. By analyzing procedures for out-of-distribution generalization with a causal graph, we show that standard classifiers fail because the association between images and labels is not transportable across settings. However, we then show that the causal effect, which severs all sources of confounding, remains invariant across domains. This motivates us to develop an algorithm to estimate the causal effect for image classification, which is transportable (i.e., invariant) across source and target environments. Without observing additional variables, we show that we can derive an estimand for the causal effect under empirical assumptions using representations in deep models as proxies. Theoretical analysis, empirical results, and visualizations show that our approach captures causal invariances and improves overall generalization.

1. Introduction

Visual representations underlie most object recognition systems today [17,18,31,46]. By learning from large image datasets, convolutional networks have been able to create excellent visual representations that improve many downstream image classification tasks [17,18,33]. However, central to this framework is the need to generalize to new visual distributions at inference time [2, 3, 6, 12, 22, 26, 44, 47, 61].

The most popular technique to use representations is to fine-tune the backbone model or fit a linear model on the target classification task [31]. Although this approach is effective on in-distribution benchmarks, the resulting classifier also inherits the biases from the target dataset. Given the nature of how data is collected, essentially every realistic image dataset will have spurious features, which will impact the generalization of computer vision systems. Specifically, the learned representation will encode features that correspond to spurious correlations found in the training data.

In this paper, we investigate visual representations for object recognition through the lenses of causality [7,42,43]. Specifically, we will revisit the out-of-distribution image classification task through causal-transportability language [9, 11, 19], which will allow us to formally model both confounding and structural invariances shared across disparate environments. In our context, we will show how different environments select a distinct set of robust and non-robust features in constructing the input dataset. The training environment may tend to select specific nuisances with the given category, creating spurious correlations between the nuisances and the predicted class. In fact, standard classifiers will tend to use those spurious correlations, which analytically explains why they result in poor generalization performance to novel target distributions [25, 49, 55].

First, we will show that the association between image and label is not in generalizable (in causal language, transportable) across domains. We then note that the causal effect from the input to the output, which severs any spurious correlations, is invariant when the environment changes with respect to the features' distributions. This motivates us to pursue to an image classification strategy that will leverage causal effects, instead of merely the association, and will act as an anchor, providing stability across changing conditions and allowing extrapolation to more likely succeed. Getting the causal effect for natural images is challenging because there are innumerable unobserved confounding factors within realistic data. Under some relatively mild assumptions, we will be able to extract the robust features from observational data through both causal and deep representations [8, 14, 20, 34–36, 48], and then use the representations as proxies for identifying the causal effect without requiring observations of the confounding factors.

For both supervised and self-supervised representations, our experimental results show that incorporating the causal structure improves performance when generalizing to new

*Equal Contribution.

domains. Our method is compatible with many existing representations without requiring re-training, making the approach effective to deploy in practice. Compared to the standard techniques to use representations, our causally motivated approach can obtain significant gain on CMNIST (up to 40% gain), WaterBird (up to 25% gain), ImageNet-Sketch (up to 8% gain), and ImageNet-Rendition (up to 7%) datasets. Our work illustrates the importance of causal quantities in out-of-distribution image classification and proposes an effective empirical method that allows the learning of a classifier robust to domain change. Our code is available at `https://github.com/cvlab-columbia/CT4Recognition`.

2. Related Work

Causal Inference and Transportability Theory. Causal inference provides a principled framework for modeling structural invariances [42] and the problem of generalizing, or transporting, across environments and changing conditions [8–11, 14, 19, 20, 36, 48]. A few image generation works have modeled a causal connection between images and their labels, often assuming the labels are generating the images [24, 48], and some prior work studied the connection between causality and specific types of generalizations [4, 37, 38, 58]. Our work studies recognition and reverses this direction, on purpose, since we consider that the images generate the labels through a human-labeling process; this model is detailed in Sec. 3. To estimate arbitrary causal effects, one can construct a proxy causal-neural models [56], but in this paper we focus on directly computing and optimizing a specific causal estimand. Existing work on this often assumes one can intervene on the data [29, 38] or observe latent confounding factors [29, 59]. These assumptions are often overly optimistic for natural images, as image data is passive (preventing intervention) and does not allow us to observe additional confounding factors.

Out of distribution Generalization in Vision. 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, 13, 32, 53, 60], where the algorithm knows the domain index which the data points are sampled from. A large number of approaches have been proposed to learn classifiers that generalize to out-of-distribution and new environments [2, 6, 25, 44, 55, 61]. In practice, however, it is often challenging to collect images with accurate domain labels, such as from the Internet. Single domain generalization [24] does not require the domain index assumption, where all training data are assumed to be sampled from the same domain. Still, domain generalization under this setup is more challenging due to lacking the domain information. Existing work achieves generalization via self-supervised learning [15], anticipating distribution shifting [45], creating pseudo domain split

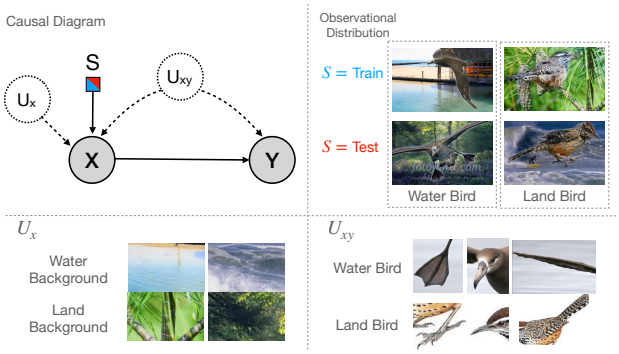


Figure 1. Causal graph for out-of-distribution image classification (top left). Image X is constructed from nuisance features U_X (bottom left) and concept features U_{XY} (bottom right). Label Y is created from X and U_{XY}. S, the transportability node, points to nodes with changes between domains, where X combines 'waterbird' with 'water background' during the training ($S = 0$) and 'water bird' with 'land background' at testing ($S = 1$) (top right).

[40], adversarial self-challenging [27], and generative data augmentations [38]. Recently, the attention operation is also shown to be effective for improving robustness [21, 39, 41]. However, a principled framework for modeling generalization to new environments is still missing.

3. Problem Formulation – Image Recognition Through Causal Lenses

We start by grounding the problem of image recognition in a causal framework to illustrate the key challenges of out-of-distribution generalization compared to its in-distribution counterpart.

3.1. Structural Modeling of the Classification Task

Let the pair X, Y represent the random variables related to images and their labels, and x, y the specific instantiations of the pixels and label. Given an input image $X = x$, the goal of the image classification task is to predict its label, $Y = y$. Taking a probabilistic interpretation, a standard strategy is to train a model to learn $P(Y \mid X)$ given data points of $X = x$ and $Y = y$, and then choose a class at inference time via $\text{argmax}_y P(Y = y \mid X = x)$.

We will take a causal approach here, and model the underlying generative process of X and Y using causal semantics. Specifically, we will use a class of generative processes known as a *structural causal model* (SCM, for short) [42, Ch. 7]. Each SCM M encodes a 4-tuple $\langle V = \{X, Y\}, U = \{U_X, U_{XY}\}, \mathcal{F} = \{f_X, f_Y\}, P(U)\rangle$, where V is the set of observed variables, in this case, the image (X) and its label (Y); U represents unobserved variables encoding external sources of variation not captured in the image and the label themselves (more details next); $\mathcal{F}$ is the set of mechanisms $\{f_X, f_Y\}$, which determine the generative processes of X and Y such that $X \leftarrow f_X(U_X, U_{XY})$

and $Y \leftarrow f_Y(X, U_{XY})$; $P(U)$ represents a probability distribution over the unobserved variables.

In particular, we call U_{XY} the "concept vector", as it represents all underlying factors that produce both the core features of the object in image x and its label, y. For example, one instantiation of $U_{XY} = u_{XY}$ may encode the concepts of "flippers" and "wing," which are translated into an image of a "waterbird" when passed into f_X. U_X represents nuisance factors, such as the background, that affect the generation process of the image. Likewise, f_Y may represent someone who is labeling image x and will have a conceptual understanding of waterbird through u_{XY}. One natural, albeit critical observation, is that if f_X selects the color "flippers" and the background "water" more likely together, there would be a strong association between these two concepts, given the image. Together, the underlying distribution over $P(U_{XY}, U_X)$ combined with functions f_X and f_Y induce a distribution over $P(X, Y)$, which is how the data is generated. The SCM M is almost never observable, and it is in general, in a formal sense, impossible to recover the structural functions ($\mathcal{F}$) and probability over the exogenous variables ($P(U)$) from observational data alone ($P(V)$) [7, Thm. 1].

3.2. Modeling In vs. Out-of-Distribution Generalization through Transportability

When training a classifier for in-distribution problems, both training and test data come from the same domain. In the out-of-distribution case, also known as the *transportability* problem in the causal inference literature [9, 11, 19], training data may come from a domain π that differs from the test domain, π^*. We assume that the labeling process and underlying concepts are consistent across domains (i.e. f_Y and $P(U_{XY})$ remain the same in both settings), but the generative process of the image X may change (i.e. f_X^* and $P^*(U_X)$ may differ from f_X and $P(U_X)$, respectively).

In general, we do not know the true underlying mechanisms f_X, f_X^*, and f_Y, nor can we observe the immeasurably large space of $P(U_X, U_{XY})$. However, we can represent the structural invariances across domains by leveraging a graphical representation shown in Fig. 1. The disparities across domains π and π^* are usually modeled by a transportability node called S [11], which can be interpreted as a switch across domains; i.e., f_X will be active if $S = 0$, and f_X^* otherwise. For concreteness, consider two different categories of birds, the waterbird and the landbird, between which we want to discriminate. Both bird categories have their own underlying features U_{XY} that cause an annotator to label them as a waterbird or landbird. However, while waterbirds are typically paired with water backgrounds in images generated in the source domain ($S = 0$), this factor may change in the target domain ($S = 1$), where waterbirds are now commonly shown in land backgrounds.

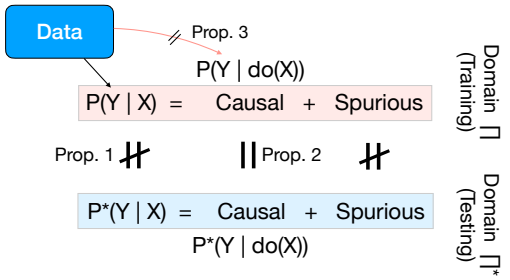


Figure 2. Visualization comparing quantities between domains π and π^*. Prop. 1 shows that $P(Y \mid X)$, which contains both causal and spurious information, does not match $P^*(Y \mid X)$. Prop. 2 shows that the causal effect is invariant across settings, i.e., $P(Y \mid \text{do}(X)) = P^*(Y \mid \text{do}(X))$. However, Prop. 3 shows that unlike $P(Y \mid X)$, $P(Y \mid \text{do}(X))$ is not identifiable from π-data .

In the in-distribution case, the more traditional strategy of learning $P(Y|X)$ is logical, in the sense that it leverages all possible information to maximize the chance of predicting the correct label. However, given the way the data generation process is modeled, it is easy to see why this same strategy fails in the out-of-distribution case. Since only data from domain π is given, we can only train a model on $P(Y \mid X)$, which does not adequately model $P^*(Y \mid X)$.

Proposition 1. *Let M and M^* be the two underlying SCMs representing the source and target domains, π and π^*, and compatible with the assumptions represented in the causal graph in Fig. 1. Then, $P^*(Y \mid X) \neq P(Y \mid X)$.*

In words, the classifier represented by the quantity $P(Y \mid X)$, in π, is not *transportable* across settings and cannot be used to make statements about $P^*(Y \mid X)$, even when everything aside from the mechanism of X (f_X) remains invariant (including the labeler f_Y). Intuitively, this is due to the unobserved confounding, or spurious effects, between X and Y through U_{XY}. By conditioning on X, the variables Y and S become d-connected via the path through U_{XY}, i.e. $P(Y \mid X, S = 0) \neq P(Y \mid X, S = 1)$. This result is also shown pictorially in Fig. 2.

In addition to the spurious effects, X and Y still co-vary due to the direct link $X \rightarrow Y$. In other words, the labeling process can be seen as moving unobserved co-variation that goes through U_{xy} to the observed link $X \rightarrow Y$. These variations are known as the causal effect of X on Y. Intuitively, one can think of the causal effect $P(Y \mid \text{do}(X))$ as describing the interventional world where arrows towards X can be thought of as removed. This includes the S-node, which no longer has an influence on X when X is forced to take a certain value, say x. This is promising since if a quantity is not affected by S, that implies that it is invariant across domains. As shown next, this is indeed the case with $P(Y \mid \text{do}(X))$.

Proposition 2. *Let M and M^* be the two underlying SCMs representing the source and target domains, π and π^*, and*

compatible with the causal graph in Fig. 1. Then, $P^*(Y \mid do(X)) = P(Y \mid do(X))$.

Regardless of the change in the mechanism of f_X^* and $P^*(U_X)$, it is guaranteed that the causal effect of X on Y will remain invariant across π and π^*. In causal language, $P^*(Y \mid \text{do}(X))$ is transportable across settings.

3.3. Identifiability

Given that the causal effect is invariant across domains, we consider using $P(Y \mid \text{do}(X))$ as a surrogate for $P^*(Y \mid X)$ for classification purposes (out-of-distribution), instead of the classifier trained in the source, $P(Y \mid X)$. That leaves the question of how to identify (and then estimate) this quantity given observational data, $P(X, Y)$. Unfortunately, this is still not possible in the general case.

Proposition 3. *Let M be the SCM representing domain π and described through the causal diagram G in Fig. 1. The interventional distribution $P(Y \mid do(X))$ is not identifiable from G and the observational distribution $P(X, Y)$.*

In words, non-identifiability suggests that there are multiple SCMs that are consistent with $P(X, Y)$ and that induce different distributions $P(Y \mid \text{do}(X))$. This means that $P(X, Y)$ is too weak, in some sense, and it is too underspecified to allow one to deduce $P(Y \mid \text{do}(X))$. Additional assumptions are needed to identify (and then estimate) this causal effect.

In fact, some prior work has assumed that *all* back-door variables can be observed [42, Sec. 3.3.1], which means that all the variations represented originally in the unobserved confounder U_{xy} are, in some sense, captured by the model. When additional domain index information is available (e.g. styles of the images), prior works such as IRM [4], MLLD [40], and DANN [1] have performed adjustment-like operations with the domain index. In most image datasets that contain only images and their labels, the assumption that all back-door variables (and sources of co-variation) are observable is overly stringent. Even when additional data is available, it is unlikely that such data contains all possible variations encapsulated by the concept vector. Our goal now is to identify the effect of X on Y without having knowledge of the back-door variables.

4. Neural Representation Approach to Deriving a Causal Estimand

Following the previous understanding that $P(Y \mid \text{do}(X))$ is a suitable proxy for the classifier in the target domain, $P^*(Y|X)$, we discuss in this section sufficient assumptions that would allow us to estimate such a quantity. Further, we discuss methods that could allow the practical realizability of these assumptions in the context of image recognition.

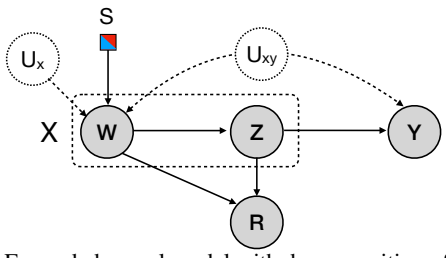


Figure 3. Expanded causal model with decomposition of image X and representation R. Gray nodes denote observed variables.

To realize the goal of estimating the target causal effect, we build two neural network models: $\hat{P}(R \mid X)$, which generates visual representations R from images X, and $\hat{P}(Y \mid R, X)$, which uses both R and X to classify Y. We make the following assumptions about the structure of image X and the properties of these networks:

Assumption 1 (Decomposition). *Each image X can be decomposed into causal factors Z and spurious factors W (i.e. $X = (Z, W)$), and the generative process follows the causal graph in Fig. 3.*

One may be tempted to surmise that this is an innocent assumption, but it does make strong claims about the generative process. The interpretation is that W contains all of the lower level signals or patches of the image, which may contain concepts confounding with Y. On the other hand, Z refines these patches into interpretable factors, which is what is visually used by the labeler. Since Z is a direct function of W, these factors are not confounded. For example, while W might include various pieces of information such as patches of blue in the water or texture of feathers, Z refines all of these signals into factors such as "waterbird shape," which is then used by the labeler to choose "waterbird" for Y. While this assumption may not be true in all settings, we believe that many practical, image settings can be approximated by this assumption.

Assumption 2 (Sufficient representation). *The neural representations $R \sim \hat{P}(R \mid Z, W)$ are learned such that they do not lose information w.r.t. Z. In words, for two samples r_1 and r_2 from $\hat{P}(R \mid z_1, w_1)$ and $\hat{P}(R \mid z_2, w_2)$, respectively, $r_1 \neq r_2$ if $z_1 \neq z_2$.*

This is a somewhat more technical assumption, which says that the neural representation has enough capacity to represent unambiguously the causal factors. This assumption should hold in general given a proper choice of model for $\hat{P}(R \mid X)$, which we further elaborate in Sec. 4.1.

Assumption 3 (Selective prediction). *Consider two images of X, $x = (z, w)$ and $x' = (z', w')$, with neural output $\hat{P}$, and the true labeling probability P. Let $R = r$ be a representation of x, sampled from $\hat{P}(R \mid x)$. Then, $\hat{P}(Y = y \mid R = r, X = x') = P(y \mid z, w')$.*

The details on how to select the specific architectural design for constructing $\hat{P}(Y \mid R, X)$ that satisfies this assumption is discussed in more detail in Sec. 4.2. Still, in words, the assumption says that once inputted with two images x and x' (x in its representation form, r), the network will make the same prediction y as if it were the true labeler when inputted with the causal feature z, from the first image, and the spurious feature w', from the second image.

Putting all these observations together, we now state one of the main results of the paper:

Theorem 1 (**Causal Identification**). *Given the assumptions about the generative process encoded in the causal graph in Fig. 3 together with assumptions 1, 2, 3, the causal effect can be computed using neural representation R via $P(Y = y|do(X = x)) = \sum_r \hat{P}(r|x) \sum_{x'} \hat{P}(y|r, x')P(x')$.*

Proof. We first derive the following steps.

$$
\begin{aligned}
& P(y \mid \text{do}(x)) \\
&= P(y \mid \text{do}(z, w)) && \text{Assumption 1} \\
&= P(y \mid \text{do}(z)) && \text{Do-Calculus Rule 3 [42]} \\
&= \sum_{w'} P(y \mid z, w')P(w') && \text{Backdoor Criterion} \\
&= \sum_{z', w'} P(y \mid z, w')P(z', w') && \text{Marginalization}
\end{aligned}
$$

By Assumptions 2 and 3, the last expression can be rewritten as

$$= \sum_{x'} \hat{P}(y \mid r, x' = (z', w'))P(x')$$

where r is sampled from $\hat{P}(R \mid x)$. Since Assumption 3 applies for any sampled value of R, we can average across samples of R,

$$= \sum_{r} \hat{P}(r \mid x) \sum_{x'} \hat{P}(y \mid r, x')P(x'),$$

concluding the proof. □

The intuition behind this derivation is that if the image x can be decomposed into causal factors (z) and spurious factors (w), as shown in Fig. 3, then the causal effect is isolated in z, and w can be ignored. By conditioning on $W = w'$, using another image, all the backdoor paths from Z to Y are blocked, which leads to an identifiable result (i.e., without do-terms). That leaves the question of how to obtain the z component from image x, and w' from x'. The general idea behind assumptions 2 and 3, and the last two lines of the derivation, is that $\hat{P}(Y \mid R, X)$ is able to extract all of the causal information z from the representation r, and extract the spurious information w' from the second image x', which will happen through the design of the neural net.

Algorithm 1 Causal-Transportability Model Training

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

Algorithm 2 Causal-Transportability Effect Evaluation

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

Altogether, Theorem 1 allows us to estimate the causal effect through [1]:

$$P(y|\text{do}(X = x)) = \sum_r \hat{P}(r|x) \sum_{x'} \hat{P}(y|r, x')P(x') \quad (1)$$

To use this formula, we need to construct the neural models to satisfy the three assumptions and properly estimate $P(X)$, $\hat{P}(R|X)$, and $\hat{P}(Y|X, R)$. The term $P(X)$ is straightforward to calculate because we can assume it is sampled from a uniform distribution [52]. The other terms, however, require a more careful construction so as to satisfy the aforementioned assumptions, which are discussed in the following sections.

[1] Interestingly, the derivation of this expression is somewhat similar to the well-known identification strategy named the front-door criterion [42, Sec. 3.3.2]. One of the key assumptions made by the front-door is that there exists a variable M that acts as an (unconfounded) mediator between X and Y. In spirit, R, our deep representation, resembles M. Despite the syntactical appearances, the variable R in the case here is not exactly a mediator, in the original sense, since it acts as a proxy for both X and Z.

	Test Accuracy	
	In-distribution	Out-of-distribution
Chance	10.0%	10.0%
ERM [54]	99.5%	8.3%
IRM* [4]	87.3%	18.5%
RSC [28]	96.6%	20.6%
GenInt [38]	58.5%	29.6%
Ablation	97.4%	38.8%
Ours	82.9%	**51.4%**

Table 1. Accuracy on the CMNIST dataset. Our method advances the state-of-the-art GenInt [38] method by over 20% on the out-of-distribution test set.

4.1. Constructing $P(R|X)$

We discuss some classes of models that are valid ways of estimating $\hat{P}(R|X)$ while satisfying Assumption 2.

Variational Auto-Encoder (VAE) [30] is an unsupervised representation learning approach, which aims to estimate a latent distribution R that can faithfully generate the input distribution. It maximizes the evidence lower bound for the distribution of X: $\mathcal{L} = -D_{KL}(q_E(r|x^{(i)})||p_\theta(r)) + \mathrm{E}_{q_E(r|x^{(i)})}[\log p_\theta(x^{(i)}|r)]$, where E is the encoder in the VAE. As VAEs are optimized to reconstruct input images via the term $\mathrm{E}_{q_E(r|x^{(i)})}[\log p_\theta(x^{(i)}|r)]$, the representation R should contain all the causal information from the input images, satisfying Assumption 2.

Constrastive Learning is another unsupervised learning approach that produces representations that can align views of the same image while separating views of different images. Given enough negative examples, contrastive learning will produce representations that are invariant under data augmentation, which still maintains all causal information from the input images, also satisfying Assumption 2.

Pretrained models from larger dataset. Empirically, deep neural networks show better generalization when pretrained from large datasets. This suggests that their representation R does not drop robust features for classification and keeps the information about Z, satisfying Assumption 2.

4.2. Constructing $P(Y|R, X)$

To properly evaluate Eq. 1, we also need to estimate a $\hat{P}(Y|R, X)$ such that Assumption 3 is satisfied. We discuss some neural network designs to achieve this.

Model Design for $\hat{P}(Y|R, X)$. In addition to the representation R, we use as input a bag of patches, which are subsampled from input image X into the branch that takes the input X. A bag of image patches corrupts the global shape information and often contains local features that are spurious, such as color, texture and background [39]. During training, the causal features Z in the image tend to be ignored by the read-out model. Specifically, we have $\hat{P}(Y|R \sim \hat{P}(R \mid Z, W), X = (Z, W)) = \hat{P}(Y|R \sim \hat{P}(R \mid Z, W), W)$. During training, the image X and the representation R are sampled from the same instance. During testing, the image X can be sampled from an arbitrary instance.

Method	Domain ID	Train	I.I.D	OOD
GDRO* [50]	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.

	OOD Test Accuracy		
	Moco-v2	SWAV	SimCLR
ERM [54]	14.59%	20.00%	27.73%
Ablation	17.04%	20.25%	28.44%
Ours	**18.02%**	**20.42%**	**29.41%**

Table 3. Accuracy on the Imagenet-9 adversarial backgrounds.

The model $\hat{P}(Y|R, X)$ has limited capacity. Given that the model has learned the information about W, learning W from R again will not further decrease the empirical loss. Thus, the model will learn Z from the representation R and ignore the W from the representation. In addition, The pretrained representations R, such as the ones from contrastive learning, can reduce the (labeled) sample complexity on classification tasks [5] than on raw image input, which allows the model to learn Z from R efficiently. This satisfies Assumption 3.

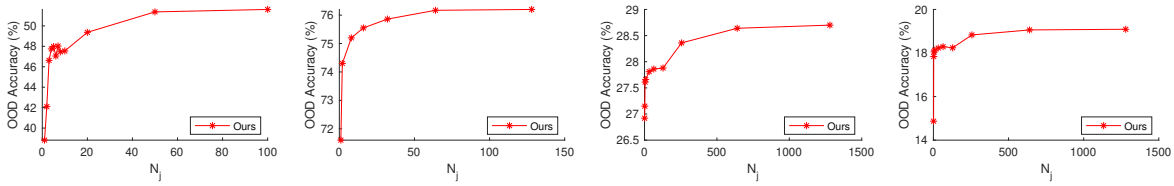
By limiting the capacity of $\hat{P}(Y|R, X)$, the model tends to use low-level features from the input images X while using high-level deep features from the latent representation R. Traditional correlation-based approaches only use $\hat{P}(Y|R)$, which can also include spurious features such as the texture and backgrounds from the representation R. With our approach, the low-level spurious features tend to be learned by the model that conditions on the input X, and the model will discard those features after marginalizing over the variable X.

4.3. Algorithm

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

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 4. 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.



(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 our approach.

5. Experiment

5.1. Datasets

CMNIST. We use the more challenging setup of colored MNIST dataset with 10 categories [38]. 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** dataset [50] 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 mechanisms in combining the foreground and background. **ImageNet-Rendition** [25] has renditions of 200 ImageNet classes, including art, cartoons, etc, which is an OOD test set for ImageNet. **ImageNet-Sketch** [55] contains sketch of 1000 ImageNet classes, which evaluate classifiers' robustness without texture and color clue. **ImageNet-9 Backgrounds Challenge** [57] studies the classifier's vulnerability to adversarially chosen backgrounds on ImageNet.

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 baselines:

ERM [23, 54] is the standard way to train deep network classifiers. *GenInt* [38] learns a causal classifier by steering the generative models to simulate interventions. *RSC* [28] 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.

5.3. Experimental Settings

We construct the low capacity network $\hat{P}(Y|X', R)$ with 3 random convolution layers applied to a bag of patches from X', concatenating the obtained feature with R, and then using 2-layer fully connected network to predict Y. Except for CMNIST where the input is low dimension and we do not use convolution layer. 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 and WaterBird datasets, we select the model with the highest validation accuracy. For ImageNet-Rendition and ImageNet-Sketch, we report the best 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 representation variable. We report the accuracy in Table 1. Our 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 10 epochs. In Table 2, without using domain index information, our causal approach improves the out-of-distribution test performance by over 25% compared with ERM, and even 1% higher than the state-of-the-art GDRO [50] method which uses domain index information.

ImageNet-9 Adversarial Backgrounds. We assess our model's robustness on testing distributions where the foreground and the background are manipulated to be different

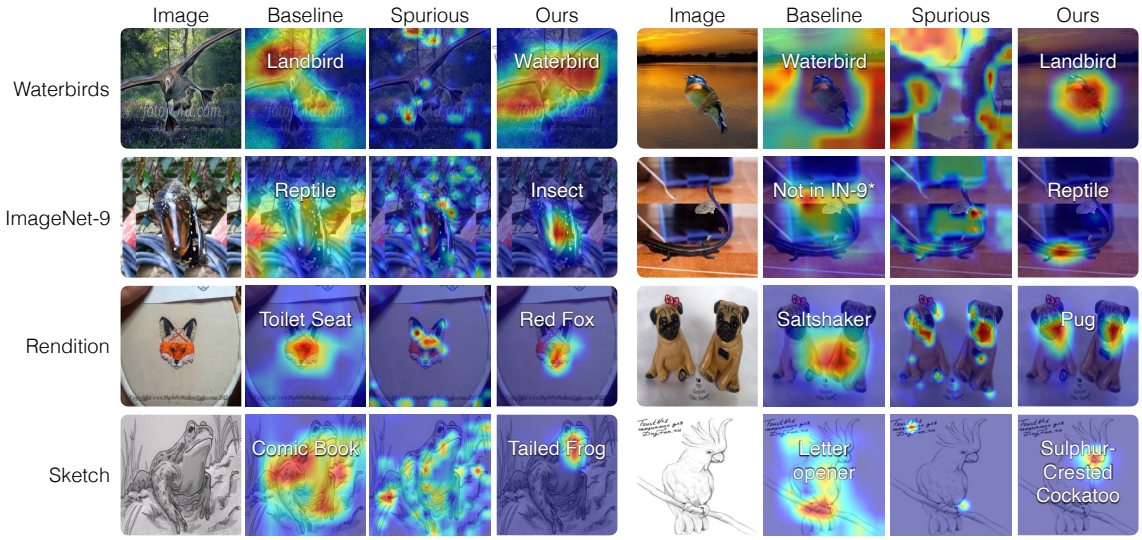


Figure 5. We visualize the input regions that the models use for prediction. We use GradCAM [51] and highlight the the discriminative regions that the model relies on with red. The white text shows the model's prediction. The correlation based ERM method often attends to spurious background context. By marginalizing over the spurious features (visualized in the Spurious column), our model captures the right, causal features, which predict the right thing for the right reason.

from the training distribution. In Table 3, we experiment on three variants of contrastive loss based self-supervised learning approaches, including Moco-v2 [18], SWAV [16], and SimCLR [17]. Overall, our approach performs better when the foreground object is present even if the background is changed.

5.5. Real-world Out of Distribution Generalization

ImageNet-Rendition and **ImageNet-Sketch** are two OOD test sets for ImageNet. We study the representation from contrastive-loss-based self-supervision learning approaches including SimCLR, MoCo-v2, and SWAV. In addition, we also study the representations from supervised learning, though they may be imperfect representations. We show results in Table 4. Our 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.

5.6. Analysis

Importance of Image Sampling. Our approach requires marginalizing over random input images x' at inference time. Sampling fewer x' can speed up the inference, however, at a cost of not estimating the accurate causal effect. In Figure 4, we vary the number of samples N_j and test the performance on four datasets. In general, We find for datasets with K categories, using $N_j > K$ can significantly improve generalization.

GradCam Visualization. Using the criterion derived in the previous section, 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 [51]. We examine four datasets, including the WaterBird, ImageNet-9, ImageNet-Rendition, and ImageNet-Sketch. We visualize the ERM model in the 'Baseline' column, the branch that conditions on the variable X of model $P(Y|R, X)$ in the 'Spurious' Column, and our 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. Conclusions

Generalization is a fundamental problem in visual recognition. This paper uses causal transportability theory to revisit and formulate the problem of out-of-distribution classification, since associational relations are not generalizable across domains. Our results demonstrate improved out-of-distribution robustness on both simulated and real-world datasets. Our findings suggest integrating causal knowledge and tools into visual representations is a promising direction to improve generalization.

Acknowledgments: CM, JW, and CV are partially supported by DARPA SAIL-ON and DARPA GAIL. CM and JF are partially supported by DiDi Faculty Research Award, J.P. Morgan Faculty Research Award, Accenture Research Award, ONR N00014-17-1-2788, and NSF CNS-1564055. EB and KX are partially supported by NSF, ONR, Amazon, JP Morgan, and The Alfred P. Sloan Foundation. HW is partially supported by NSF Grant IIS-2127918 and an Amazon Faculty Research Award.

References

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

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

[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.

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

[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.

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

[7] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*. Association for Computing Machinery, NY, USA, 1st edition, 2022.

[8] E. Bareinboim, S. Lee, V. Honavar, and J. Pearl. Transportability from multiple environments with limited experiments. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 136–144. Curran Associates, Inc., 2013.

[9] E. Bareinboim and J. Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

[10] E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 280–288. Curran Associates, Inc., 2014.

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

[12] 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.

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

[14] Peter Bühlmann. Invariance, causality and robustness, 2018.

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

[16] 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.

[17] 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.

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

[19] J. Correa and E. Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In S. Kraus, editor, *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1661–1667, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization.

[20] 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.

[21] 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.

[22] 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.

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

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

[25] 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.

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

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

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

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

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

[31] 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.

[32] 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.

[33] Wanyu Lin, Zhaolin Gao, and Baochun Li. Shoestring: Graph-based semi-supervised classification with severely limited labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4174–4182, 2020.

[34] Wanyu Lin, Hao Lan, and Baochun Li. Generative Causal Explanations for Graph Neural Networks. In *Proc. International Conference on Machine Learning*, 2021.

[35] Wanyu Lin, Hao Lan, Hao Wang, and Baochun Li. Orphicx: A causality-inspired latent variable model for interpreting graph neural networks. In *CVPR*, 2022.

[36] 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.

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

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

[39] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. *arXiv preprint arXiv:2111.10493*, 2021.

[40] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020.

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

[42] Judea Pearl. Causality: Models, reasoning, and inference, 2000.

[43] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Cosma Shalizi. Advanced data analysis from an elementary point of view. page 456, 2013.

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

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

[55] 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.

[56] Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34, 2021.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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