Generative Interventions for Causal Learning

Chengzhi Mao[1] Augustine Cha[1*] Amogh Gupta[1*] Hao Wang[2] Junfeng Yang[1]
Carl Vondrick[1]
[1]Columbia University, [2]Rutgers University
{mcz, junfeng, vondrick}@cs.columbia.edu, {ac4612, ag4202}@columbia.edu, hoguewang@gmail.com

Abstract

We introduce a framework for learning robust visual representations that generalize to new viewpoints, backgrounds, and scene contexts. Discriminative models often learn naturally occurring spurious correlations, which cause them to fail on images outside of the training distribution. In this paper, we show that we can steer generative models to manufacture interventions on features caused by confounding factors. Experiments, visualizations, and theoretical results show this method learns robust representations more consistent with the underlying causal relationships. Our approach improves performance on multiple datasets demanding out-of-distribution generalization, and we demonstrate state-of-the-art performance generalizing from ImageNet to ObjectNet dataset.

1. Introduction

Visual recognition today is governed by empirical risk minimization (ERM), which bounds the generalization error when the training and testing distributions match [47]. When training sets cover all factors of variation, such as background context or camera viewpoints, discriminative models learn invariances and predict object category labels with the right cause [33]. However, the visual world is vast and naturally open. Collecting a representative, balanced dataset is difficult and, in some cases, impossible because the world can unpredictably change after learning.

Directly optimizing the empirical risk is prone to learning unstable spurious correlations that do not respect the underlying causal structure [11, 8, 24, 44, 4, 35]. Figure 1 illustrates the issue succinctly. In natural images, the object of interest and the scene context have confounding factors, creating spurious correlations. For example, ladle (the object of interest) often has a hand holding it (the scene context), but there is no causal relation between them. Several studies have exposed this challenge by demonstrating substantial performance degradation when the confounding bias no longer holds at testing time [41, 19]. For example, the ObjectNet [6] dataset removes several common spurious correlations from the test set, causing the performance of state-of-the-art models to deteriorate by 40% compared to the ImageNet validation set.

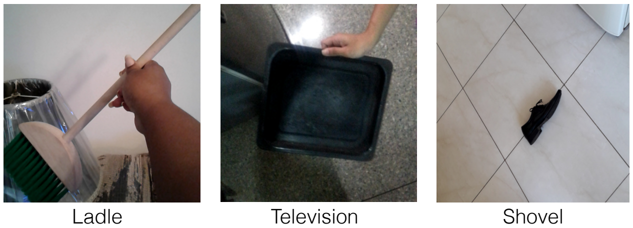

Figure 1. Top predictions from a state-of-the-art ImageNet classifier [21]. The model uses spurious correlations (scene contexts, viewpoints, and backgrounds), leading to incorrect predictions.[1] In this paper, we introduce a method to learn causal visual features that improve robustness of visual recognition models. The predictions of our model are in Figure 7.

A promising direction for fortifying visual recognition is to learn *causal* representations (see [43] for an excellent overview). If representations are able to identify the causal mechanism between the image features and the category labels, then robust generalization is possible. While the traditional approach to establish causality is through randomized control trials or interventions, natural images are passively collected, preventing the use of such procedures.

This paper introduces a framework for learning causal visual representations with natural images. Our approach is based on the observation that generative models quantify nuisance variables [23, 26], such as viewpoint or background. We present a causal graph that models both robust features and spurious features during image recognition. Crucially, our formulation shows how to learn causal features by steering generative models to perform interventions on realistic images, simulating manipulations to the camera and scene that remove spurious correlations. As our approach is model-agnostic, we are able to learn robust representations for any state-of-the-art computer vision model.

Our empirical and theoretical results show that our ap-

[*] Equal Contribution. Order by coin flip.

[1] The correct categories are clearly a broom, a tray, and a shoe.

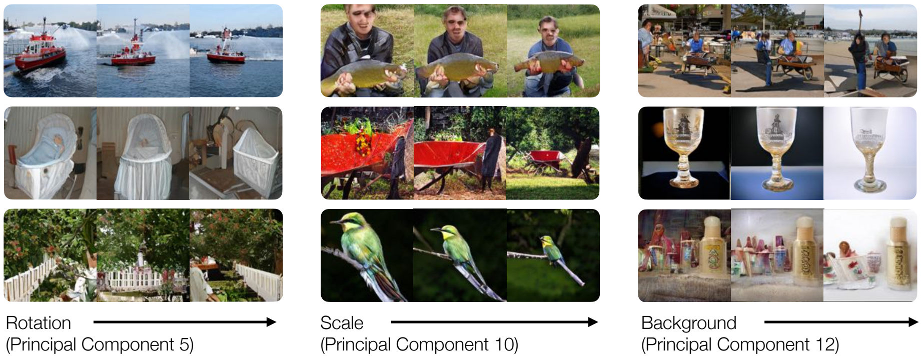

Figure 2. Generative adversarial networks are steerable [23, 26], allowing us to manipulate images and construct interventions on nuisances. The transformations transfer across categories. Each column in the figure presents images with one consistent intervention direction.

proach learns representations that regard causal structures. While just sampling from generative models will replicate the same training set biases, steering the generative models allows us to reduce the bias, which we show is critical for performance. On ImageNet-C [22] benchmark, we surpass established methods by up to 12%, which shows that our method helps discriminate based on the causal features. Our approach also demonstrates the state-of-the-art performance on the new ObjectNet dataset [6]. We obtain 39.3% top-1 accuracy with ResNet152 [21], which is over 9% gain over the published ObjectNet benchmark [6] while increasing accuracy on ImageNet and ImageNet-V2 [41]. We will release code, data, and models.

2. Related Work

Data augmentation: Data augmentation often helps learn robust image classifiers. Most existing data augmentations use lower-level transformations [29, 46], such as rotate, contrast, brightness, and shear. Auto-data augmentation [13, 52] uses reinforcement learning to optimize the combination of those lower-level transformations. Other work, such as *cutout* [15] and *mixup* [51], develops new augmentation strategies towards improved generalization. [34, 54, 19] explored style transfer to augment the training data, however, the transformations for training are limited to texture and color change. Adversarial training, where images are augmented by adding adversarial noise, can also train robust models [50]. However, both adversarial training [50] and auto augmentation [13, 52] introduce orders of magnitude of computational overhead. In addition, none of the above methods can do high-level transformations such as changing the viewpoint or background [6], while our generative interventions can. Our method fundamentally differs from prior data augmentation methods because it learns a robust model by estimating the causal effects via generative interventions. Our method not only eliminates spurious correlations more than data augmentations, but also theoretically produces a tighter causal effect bound.

Causal Models: Causal image classifiers generalize well despite environmental changes because they are invariant to the nuisances caused by the confounding factors [8]. A large body of work studies how to acquire causal effects from a combination of association levels and intervention levels [31, 11, 32]. Ideally, we can learn an invariant representation across different environments and nuisances [8, 35] while maintaining the causal information [5]. While structural risk minimization, such as regularization [27], can also promote a model's causality, this paper focuses on training models under ERM [47].

Generative Models: Our work leverages recent advances in deep generative models [20, 28, 25, 10, 40]. Deep generative models capture the joint distribution of the data, which complements discriminative models [25, 37, 38]. Prior work has explored adding data sampled from a deep generator to the original training data to improve classification accuracy on ImageNet [39]. We denote it as GAN Augmentation in this paper. Other works improved classification accuracy under imbalanced and insufficient data by oversampling through a deep generator [16, 17, 49, 7]. However, sampling without intervention, the augmented data still follows the same training joint distribution, where unobserved confounding bias will continue to contaminate the generated data. Thus, the resulting models still fail to generalize once the spurious correlations changed. Ideally, we want to generate data independent of the spurious correlations while holding the object's causal features fixed.

Recent works analyzing deep generative models show that different variations, such as viewpoints and background, are automatically learned [26, 23]. We leverage deep generative models for constructing interventions in realistic visual data. Our work randomizes a large class of steerable variations, which shifts the observed data distribution to be independent of the confounding bias further. Our approach tends to manipulate high-level transformations orthogonal to traditional data augmentation strategies [12, 46, 29], and we obtain additional performance gains by combining them.

Domain Adaptation: Our goal is to train robust models that generalize to unforeseen data. Accessing the test data distribution, even unlabeled, could lead to overfitting

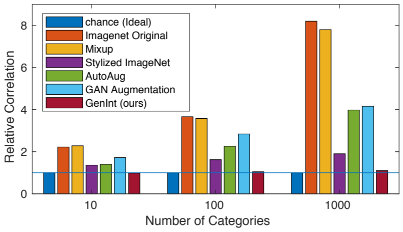

Figure 3. Do unwanted correlations exist between the nuisance factors (e.g. backgrounds, viewpoint) and labels on ImageNet? We measure correlation (y-axis) via how many times the classification accuracy is better than chance on the ImageNet validation set. The x-axis denotes the number of categories we select for prediction. To train causal models, nuisance factors should not be predictable for labels (chance). Our generative interventions (GenInt) reduce the unwanted correlations from the data better than existing data augmentation strategies [51, 13, 19, 39].

and fail to measure the true generalization. Our work thus is trained with no access to the test data. Our setting is consistent with ObjectNet's policy prohibiting any form of learning on its test set [6], and ImageNet-C's policy discouraging training on the tested corruptions. On the other hand, domain adaptation [3, 42, 48] needs access to the distributions of both the source domain and the target domain, which conflicts with our setting.

3. Causal Analysis

We quantify nuisances via generative models and propose the corresponding causal graph. We show how to train causal models via intervention on the nuisance factors. We theoretically show sufficient conditions for intervention strategy selection that promote causal learning.

3.1. Correlation Analysis

Nuisance factors do not cause the object label. If there is a correlation between the nuisance factors and the label in data, we cannot learn causal classifiers. While identifying such correlations is crucial, they are hard to quantify on large, real-world vision datasets, because nuisance factors such as viewpoint and backgrounds, are difficult and expensive to measure in natural images.

We propose to measure such nuisance factors via intervening on the conditional generative models. Prior work [23, 26] shows that nuisance transformations automatically emerge in generative models (Figure 2), which enables constructing desired nuisances via intervention. Given a category $\mathbf{y}$ and random noise vector $\mathbf{h}_0$, we first generate an exemplar image $\mathbf{x} = G(\mathbf{h}_0, \mathbf{y})$. We then conduct intervention $\mathbf{z}$ to get the intervened noise vector $\mathbf{h}_0^*$, and the intervened image $\mathbf{x}^* = G(\mathbf{h}_0^*, \mathbf{y})$, which corresponds to changing the viewpoints, backgrounds, and scene context of the exemplar. We thus get data with both image $\mathbf{x}^*$ and the corresponding nuisance manipulation $\mathbf{z}$. Implementation details are in the supplementary.

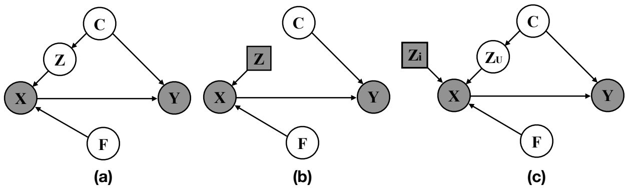

Figure 4. Causal graph for image classification. Gray variables are observed. (**a**) F is the variable that generates the object features. The unobserved confounder C causes both the background features Z and label Y, which creates a spurious correlation between the image X and label Y. (**b**) An ideal intervention blocks the backdoor path from Z to C, which produces causal models. (**c**) In practice, we cannot guarantee to intervene on all the Z variables. However, by properly intervening on even a small set of nuisance factors Z_i, the confounding bias of the observed distribution is mitigated, which is theoretically proven by Theorem 1.

We train a model that predicts the nuisances $\mathbf{z}$ from input image $\mathbf{x}^*$. This model can then predict nuisances $\mathbf{z}$ from natural images $\mathbf{x}$. We read out the correlation between the nuisance $\mathbf{z}$ and label $\mathbf{y}$ by training a fully-connected classifier with input $\mathbf{z}$ and output $\mathbf{y}$. We measure the correlations via the times the classifier outperforms random. Generative models may capture only a subset of the nuisances, thus our estimated correlations are lower bounds. The true correlations maybe even more significant.

In Figure 3, the training data of five established methods [21, 51, 13, 19, 39] contains strong correlations that are undesirable. On the original ImageNet data, the undesirable correlation in the data is up to 8 times larger than chance. Our generative interventions reduce the unwanted correlations from the data significantly, naturally leading to robust classifiers that use the right cause.

3.2. Causal Graph

We build our causal graph based on the correlation analysis. We know that nuisances do not cause the label (context 'hand' does not cause the category 'ladle'), and there is no additional common outcome variable (collider) in our correlation prediction. If the correlation between the nuisances and the label is not chance, then there exists a confounder C that causes both Z and Y.

Figure 4(a) shows our causal graph for image recognition. We denote the unobserved confounder as C, which produces nuisance factors Z, and the corresponding annotated categorical label Y. Z produces the nuisance features X_Z in images. There is another variable F that generates the core object features X_F, which together with X_Z constructs the pixels of a natural image X. There is no direct arrow from F to Y since $Y \perp\!\!\!\perp F|X$, i.e., image X contains all the features for predicting Y. We can observe only X but not X_Z or Z_F separately. We draw a causal arrow from X to Y. Since nuisances Z are spuriously correlated to the label but not causing the label Y, classifiers are not causal if they predict Y from the nuisances Z better than chance. Note that "while a directed path is necessary for a total causal effect, it is not sufficient [36]." Thus, though there is a path $Z \to X \to Y$, Z does not cause Y.

3.3. Causal Discriminative Model

Generative interventions help in eliminating spurious correlations (Figure 3 and Section 3.1), leading to better generalization. We denote the causality from X to Y to be $P(\mathbf{y}|do(\mathbf{x}))$, which is the treatment effect of an input image X on label Y. To capture the right cause via correlations learned by empirical risk minimization, we need to construct data such that $P(Y|do(X)) = P(Y|X)$.

Natural images are often biased by unobserved confounding factors that are common causes to both the image X and the label Y. A passively collected vision dataset only enables us to observe the variables X and Y. Theoretically, we cannot identify the causal effect $P(Y|do(X))$ in Figure 4(a) with only the observed joint distribution $P(X, Y)$ because there is an unobserved common cause.

We thus want to collect faithful data independent of the confounding bias, so that we can identify the causal effect with only the observed data. We need to intervene on the data-generation process for the nuisances Z to be independent to the confounders, while keeping the core object features F unchanged. In the physical world, such interventions correspond to actively manipulating the camera or objects in the scene. In our paper, we perform such interventions via steering the generative models. The outcome of this intervention on Z is visualized in Figure 4(b), which manipulates the causal graph such that dependencies arriving at Z are removed. Removing the backdoor, the correlation is now equal to the causality, i.e., $P(Y|X) = P(Y|do(X))$. While this result is intuitive, performing perfect intervention in practice is challenging due to the complexity of the natural image distribution.

3.4. Causal Effect Bound

Imperfect interventions can eliminate only some spurious correlations. Though it is theoretically impossible to calculate the exact causal effect $P(\mathbf{y}|do(\mathbf{x}))$ when spurious correlations are not totally removed, we can still estimate the lower and upper bound for $P(\mathbf{y}|do(\mathbf{x}))$.

Given the observed joint distribution $P(\mathbf{x}, \mathbf{y})$, Pearl [33] identified that $P(\mathbf{y}|do(\mathbf{x}))$ can be bounded by $P(\mathbf{x}, \mathbf{y}) \leq P(\mathbf{y}|do(\mathbf{x})) \leq P(\mathbf{x}, \mathbf{y}) + 1 - P(\mathbf{x})$, which can be estimated by existing discriminative models without interventions.

Prior work augments the data by sampling from the GANs without explicit intervention [49, 7, 16, 17], which will yield the same causal bound as the original data. Since GANs capture the same distribution as the observational training set, the spurious correlations remain the same. The sampled transformations Z in Figure 4 (a) are still dependent on the confounders C. Thus, augmenting training data with GANs [39], without intervention is not an effective algorithm for causal identification.

In this paper, we aim to identify a tighter causal effect bound for $P(\mathbf{y}|do(\mathbf{x}))$ using generative interventions. This is desirable for robustness because it removes or reduces the overlap between the causal intervals, promoting causal predictions. Section 3.3 establishes that perfect interventions eliminate all spurious correlation and leads to better generalization. In practice, our generative interventions may only eliminate a subset of spurious correlations Z_i, while other nuisances Z_U remain unobserved and untouched. The next question is then: *what generative intervention strategy is optimal for tightening the causal effect bound?* We derive the following theory:

Theorem 1 (**Effective Intervention Strategy**). *We denote the images as x. The causal bound under intervention z_i is thus $P(\mathbf{y}, \mathbf{x}|\mathbf{z}_i) \leq P(\mathbf{y}|do(\mathbf{x})) \leq P(\mathbf{y}, \mathbf{x}|\mathbf{z}_i) + 1 - P(\mathbf{x}|\mathbf{z}_i)$. For two intervention strategies $\mathbf{z}_1$ and $\mathbf{z}_2$, $\mathbf{z}_1 \subset \mathbf{z}, \mathbf{z}_2 \subset \mathbf{z}$, if $P(\mathbf{x}|\mathbf{z}_1) > P(\mathbf{x}|\mathbf{z}_2)$, then $\mathbf{z}_1$ is more effective for causal identification.*

Proof. Figure 4(c) shows the causal graph after intervention Z_i, where $Z_i \perp\!\!\!\perp Y|X$. We add and remove the same term $\sum_c P(\mathbf{y}, \mathbf{x}, \mathbf{c}|\mathbf{z}_i)$:

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_c P(\mathbf{y}|\mathbf{x}, \mathbf{z}_i, \mathbf{c})P(\mathbf{c}) \qquad \text{(Backdoor Criteria)}$$

$$= \sum_{\mathbf{c}} P(\mathbf{y}, \mathbf{x}, \mathbf{c}|\mathbf{z}_i) + \sum_{\mathbf{c}} P(\mathbf{y}|\mathbf{x}, \mathbf{z}_i, \mathbf{c})(P(\mathbf{c}) - P(\mathbf{x}, \mathbf{c}|\mathbf{z}_i))$$

Since $0 \leq P(\mathbf{y}|\mathbf{x}, \mathbf{z}_i, \mathbf{c}) \leq 1$, we have the lower and upper bounds. We denote $\delta_1 = P(\mathbf{x}|\mathbf{z}_1) - P(\mathbf{x}|\mathbf{z}_2)$, thus $\delta_1 > 0$. In the causal graph (Figure 4(c)), since we intervene on $\mathbf{z}_i$, all incoming edges to $\mathbf{z}_i$ are removed; we then have $\mathbf{z}_i \perp\!\!\!\perp \mathbf{y}|\mathbf{x}$ and $P(\mathbf{x}, \mathbf{y}|\mathbf{z}_i) = P(\mathbf{y}|\mathbf{x}, \mathbf{z}_i)P(\mathbf{x}|\mathbf{z}_i) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x}|\mathbf{z}_i)$. Therefore $\delta_2 = P(\mathbf{x}, \mathbf{y}|\mathbf{z}_1) - P(\mathbf{x}, \mathbf{y}|\mathbf{z}_2) = \delta_1 \cdot P(\mathbf{y}|\mathbf{x})$. Since apparently $0 < P(\mathbf{y}|\mathbf{x}) < 1$, we have that $0 < \delta_2 < \delta_1$. Thus we obtain $[P(\mathbf{y}, \mathbf{x}|\mathbf{z}_1), P(\mathbf{y}, \mathbf{x}|\mathbf{z}_1) + 1 - P(\mathbf{x}|\mathbf{z}_1)] \subset [P(\mathbf{y}, \mathbf{x}|\mathbf{z}_2), P(\mathbf{y}, \mathbf{x}|\mathbf{z}_2) + 1 - P(\mathbf{x}|\mathbf{z}_2)]$, which means the intervention $\mathbf{z}_1$ results in a tighter causal effect bound. □

Our theorem shows that: the optimal intervention strategy should maximize $P(\mathbf{x}|\mathbf{z})$, which will tighten the causal effect bound $P(\mathbf{y}|do(\mathbf{x}))$. Also, the intervention strategy should be identically selected across all categories, so that they are independent of the confounding bias. While there are different choices of intervening on the generative model to create independence, we empirically select our generative intervention strategy that increases $P(\mathbf{x}|\mathbf{z})$, which we will discuss in Section 5.4.

4. Method

We show how deep generative models can be used to construct interventions on the spuriously correlated features in the causal graph. We combine these results to develop a practical framework for robust learning.

4.1. Learning Objective

We minimize the following training loss on our intervened data:

$$\begin{aligned}\mathcal{L} =&\mathcal{L}_e(\phi(\mathbf{X}), \mathbf{Y}) + \lambda_1 \mathcal{L}_e(\phi(\mathbf{X}_{int}), \mathbf{Y}') \\ &+ \lambda_2 \mathcal{L}_e(\phi(\mathbf{X}_{itr}), \mathbf{Y}'')\end{aligned} \tag{1}$$

where $\mathcal{L}_e$ denotes the standard cross entropy loss and $\lambda_i \in \mathbb{R}$ are hyper-parameters controlling training data choice. We denote the original data matrix as $\mathbf{X}$ with target labels $\mathbf{Y}$; the generated data matrix as $\mathbf{X}_{int}$ (Section 4.2) with target labels $\mathbf{Y}'$; the transfered data as $\mathbf{X}_{itr}$ (Section 4.3) with target labels $\mathbf{Y}''$; and the discriminative classifier as ϕ.

The last two terms of this objective are the interventions. In the remainder of this section, we present two different ways of constructing these interventions.

4.2. Generative Interventions

We construct interventions using conditional generative adversarial networks (CGAN). We denote the i-th layer's hidden representation as $\mathbf{h}_i$. CGAN learns the mapping $\mathbf{x} = G(\mathbf{h}_0, \mathbf{y})$, where $\mathbf{h}_0 \sim \mathcal{N}(0, I)$ is the input noise, $\mathbf{y}$ is the label, and $\mathbf{x}$ is a generated image of class $\mathbf{y}$ that lies in the natural image distribution. CGANs are trained on the joint data distribution $P(\mathbf{x}, \mathbf{y})$. While we can use any type of CGANs, we select BigGAN [10] in this paper since it produces highly realistic images. In addition, generative models learn a latent representation $\mathbf{h}_i$ equivariant to a large class of visual transformations and independent of the object category [23, 26], allowing for controlled visual manipulation. For example, GANSpace [23] showed that the principal components of $\mathbf{h}_i$ correspond to visual transformations over camera extrinsics and scene properties. The same perturbations in the latent space will produce the same visual transformations across different categories. Figure 2 visualizes this steerability for a few samples and different transformations. This property enables us to construct a new training distribution, where the nuisance features Z are not affected by the confounders.

Our generative intervention strategy follows the GANSpace [23] method, which empirically steers the GAN with transformations independent of the categories. It contains three factors: the truncation value, the transformation type, and the transformation scale. The input noise $\mathbf{h}_0$ is sampled from Gaussian noise truncated by value t [10]. We define the transformations to be along the j-th principal directions $\mathbf{r}_j$ in the feature space [23], which are orthogonal and captures the major variations of the data. We select the top-k significant ones $\{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_k\}$ as the intervention directions. We then intervene along the selected directions with a uniformly sampled step size s' from a range $[-s, s]$. We intervene on the generator's intermediate layers with $\mathbf{h}_i^* = \mathbf{h}_i + \sigma s' \mathbf{r}_j - \mu$, where $\mathbf{h}_i^*$ are the features at layer i after interventions, σ is the standard deviation of noise on direction $\mathbf{r}$, and μ is the offset term. After the intervention, we follow the method in GANSpace [23] to recover $\mathbf{h}_0^*$ with regression and generate the new image $\mathbf{x}^* = G(\mathbf{h}_0^*, \mathbf{y})$. Using conditional generative models, we produce the causal features X_F by specifying the category. Our intervention removes the incoming edge from C to Z_i (Figure 4 (c)). We denote the intervention procedure as function I, and rewrite the generative interventions as:

$$\mathbf{X}_{int} = I(t, s, k, \mathbf{Y}')$$

Based on our Theorem 1, we choose the hyper-parameters t, k, s for intervention Z that maximizes $P(\mathbf{x}|\mathbf{z})$. We show ablation studies in Section 5.4.

4.3. Transfer to Natural Data

Maintaining the original training data $\mathbf{X}$ will add confounding bias to models. While our theory shows that our method still tightens the causal effect bound under the presence of spurious correlations, it is desirable to eliminate as many spurious correlations as possible. We will therefore also intervene on the original dataset.

One straightforward approach is to estimate the latent codes in the generator corresponding to the natural images, and apply our above intervention method. We originally tried projecting the images back to the latent space in the generative models [53, 2], but this did not obtain strong results, because the projected latent code cannot fully recover the query images [9].

Instead, we propose to transfer the desirable generative interventions from $\mathbf{X}_{int}$ to the original data $\mathbf{X}$ with neural style transfer [18]. The category information is maintained by the matching loss while the intervened nuisance factors are transferred via minimizing the maximum mean discrepancy [30]. Without projecting the images to the latent code, the transfer enables us to intervene on some of the nuisance factors z in the original data, such as the background. The transfer of the generative interventions $I(t, k, s, \mathbf{Y}')$ to natural data $\mathbf{X}$ is formulated as:

$$\mathbf{X}_{itr} = T(I(t, k, s, \mathbf{Y}'), \mathbf{X})$$

where T denote the style transfer mapping. The corresponding label $\mathbf{Y}''$ is the same label as for $\mathbf{X}$. Please see supplemental material for visualizations of these interventions.

5. Experiments

We present image classification experiments on four datasets — ImageNet, ImageNet-V2, Imagenet-C, and ObjectNet — to analyze the generalization capabilities of this

	ResNet 18				ResNet 152			
	Std. Augmentation		Add. Augmentation		Std. Augmentation		Add. Augmentation	
Training Distribution	top1	top5	top1	top5	top1	top5	top1	top5
ImageNet Only [21, 6]	20.48%	40.64%	24.42%	44.39%	30.00%	48.00%	37.43%	59.10%
Stylized ImageNet [19]	18.39%	37.29%	22.81%	42.27%	31.64%	52.56%	36.17%	57.95%
Mixup [51]	19.12%	37.78%	24.05%	44.17%	34.27%	55.68%	38.61%	60.36%
AutoAug [13]	21.20%	41.26%	21.20%	41.26%	33.96%	55.81%	33.96%	55.81%
GAN Augmentation [39]	20.63%	39.77%	23.72%	43.67%	33.17%	54.59%	36.37%	58.88%
GenInt (ours)	22.07%	**41.94%**	25.71%	46.39%	34.47%	55.63%	39.21%	61.06%
GenInt with Transfer (ours)	**22.34%**	41.65%	**27.03%**	**48.02%**	**34.69%**	**55.82%**	**39.38%**	**61.43%**

Table 1. Accuracy on the ObjectNet test set versus training distributions. By intervening on the training distribution with generative models, we obtain the state-of-the-art performance on the ObjectNet test set, even though the model was never trained on ObjectNet.

	Model	**mCE** ↓	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
	AlexNet	100.00	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
ResNet 18 [21]	ImgNet Only [21]	87.16	89.5	90.4	93.0	86.01	93.3	87.7	90.0	87.5	86.4	80.0	73.7	80.5	91.5	85.5	92.4
	Stylized ImgNet [19]	80.83	79.1	80.9	81.7	81.7	87.6	80.0	90.0	78.3	80.2	76.2	72.5	77.2	**84.1**	76.2	86.7
	Mixup [51]	86.06	86.8	88.1	90.8	88.7	95.6	89.1	89.3	82.5	**72.8**	71.9	75.9	76.5	96.2	89.5	97.2
	AutoAug [13]	84.00	84.3	83.7	84.5	87.9	93.6	87.7	93.5	85.7	83.4	**71.0**	**67.4**	**63.5**	97.8	85.3	90.5
	GAN Augmentation [39]	86.48	86.4	87.5	90.5	87.0	92.4	87.2	90.3	88.0	86.3	82.8	73.3	82.8	90.7	84.5	87.5
	GenInt with Transfer (ours)	**74.68**	**67.0**	**68.4**	**67.3**	**75.0**	**80.5**	**76.0**	**84.2**	**77.4**	75.9	77.5	68.8	76.6	87.5	**59.8**	**77.5**
ResNet 152 [21]	ImgNet Only [21]	69.27	72.5	73.4	76.3	66.9	81.4	65.7	74.5	70.7	67.8	62.1	51.0	67.1	75.6	68.9	65.1
	Stylized ImgNet [19]	64.19	63.3	63.1	64.6	66.1	77.0	63.5	71.6	**62.4**	65.4	59.4	52.0	62.0	73.2	55.3	62.9
	Mixup [51]	66.43	69.0	71.1	73.8	67.3	83.4	65.5	74.6	63.5	**56.9**	**55.2**	49.4	62.4	75.4	65.0	63.7
	AutoAug [13]	69.20	71.7	72.8	75.6	67.2	82.1	67.7	76.7	70.3	67.7	61.8	50.5	65.0	76.0	68.3	64.6
	GAN Augmentation [39]	69.01	71.8	73.1	75.9	67.3	82.3	67.5	76.2	69.9	68.1	59.2	51.3	62.5	76.6	67.7	65.7
	GenInt with Transfer (ours)	**61.70**	**59.2**	**60.2**	**62.4**	**60.7**	**70.8**	**59.5**	**69.9**	64.4	63.8	58.3	**48.7**	**61.5**	**70.9**	**55.2**	**60.0**

Table 2. The mCE ↓ rate (the smaller the better) on ImageNet-C validation [22] set with 15 different corruptions. Our GenInt model, without training on any of the corruptions, reduces the mCE by up to **12.48%**. From column 'Gauss.' to column 'JPEG,' we show individual Error Rate on each corruption method. Without adding similar corruptions in the training set, our generative causal learning approach learns models that naturally generalize to unseen corruptions.

method and validate our theoretical results. We call our approach **GenInt** for generative interventions, and compare the different intervention strategies.

5.1. Datasets

In our experiments, all the models are first trained on **ImageNet** [14] (in addition to various intervention strategies). We train only on ImageNet without any additional data from other target domains. We directly evaluate the models on the following out-of-distribution testing sets:

ObjectNet [6] is a test set of natural images that removes background, context, and camera viewpoints confounding bias. Improving performance on ObjectNet—without fine-tuning on it—indicates that a model is learning causal features. ObjectNet's policy prohibits any form of training on the ObjectNet data. We measure performance on the 113 overlapping categories between ImageNet and ObjectNet.

ImageNet-C [22] is a benchmark for model generalization under 15 common corruptions, such as 'motion,' 'snow,' and 'defocus.' Each corruption has 5 different intensities. We use mean Corruption Error (mCE) normalized by AlexNet as the evaluation metric [22]. Note that we do not train our model with any of these corruptions, thus the performance gain measures our model's generalization to unseen corruptions.

ImageNet-V2 [41] is a new test set for ImageNet, aiming to quantify the generalization ability of ImageNet models. It contains three sampling strategies: MatchedFrequency, Threshold0.7, and TopImages. While current models are overfitting to the ImageNet test set, this dataset measures the ability to generalize to a new test set.

5.2. Baselines

We compare against several established data augmentation baselines:

Stylized ImageNet refers to training the model using style transferred dataset [19], which trains classifiers that are not biased towards texture.

Mixup [51] does linear interpolation to augment the dataset. We use their best hyperparameters setup ($\alpha = 0.4$).

AutoAug [13] systematically optimizes the strategy for data augmentation using reinforcement learning.

GAN Augmentation refers to the method that augments the ImageNet data by directly sampling from the BigGAN [39]. They provide an extensive study for hyper-parameter selection. We use their best setup as our baseline: 50% of synthetic data sampled from BigGAN with truncation 0.2.

ImageNet only refers to training the standard model on ImageNet dataset only [21].

5.3. Empirical Results

Our GenInt method demonstrates significant gains on four datasets over five established baselines. We report

		ImageNet-V2 Grouped by Sampling Strategy [41]						Original ImageNet Val	
		"TopImages"		"Threshold0.7"		"MatchedFrequency"			
	Training Distribution	top1	top5	top1	top5	top1	top5	top1	top5
ResNet 18 [21]	ImageNet Only [21]	71.77%	91.11%	65.41%	87.39%	56.18%	79.35%	68.82%	88.96%
	Stylized ImageNet [19]	69.55%	89.97%	62.92%	85.38%	54.13%	77.30%	66.95%	87.42%
	Mixup [51]	69.90%	90.16%	63.42%	86.40%	54.42%	77.94%	66.00%	86.93%
	AutoAug [13]	72.05%	91.49%	65.32%	87.32%	56.25%	79.16%	69.24%	88.91%
	GAN Augmentation [39]	72.01%	91.24%	65.72%	87.58%	56.43%	79.42%	69.19%	88.85%
	GenInt (ours)	72.80%	**91.89%**	66.26%	**88.30%**	**57.86%**	**80.11%**	**70.41%**	**89.59%**
	GenInt with Transfer (ours)	**72.84%**	91.85%	**66.49%**	88.11%	57.35%	79.61%	70.25%	89.33%
ResNet 152 [21]	ImageNet Only [21]	81.01%	96.21%	76.17%	94.12%	67.76%	87.57%	78.57%	94.29%
	Stylized ImageNet [19]	79.40%	95.72%	74.02%	92.88%	65.12%	86.22%	77.27%	93.76%
	Mixup [51]	80.68%	96.28%	75.91%	94.00%	67.11%	87.66%	78.78%	94.45%
	AutoAug [13]	80.61%	96.30%	75.90%	94.06%	67.35%	87.61%	78.95%	94.56%
	GAN Augmentation [39]	80.10%	96.00%	75.60%	93.74%	66.89%	87.04%	78.53%	94.21%
	GenInt (ours)	80.77%	**96.38%**	76.20%	**94.24%**	67.74%	**87.83%**	79.46%	94.71%
	GenInt with Transfer (ours)	**81.24%**	96.28%	**76.60%**	93.95%	**68.08%**	87.70%	**79.59%**	**94.79%**

Table 3. Accuracy on ImageNet V2 validation set [41] and original ImageNet validation set. Our method improves the performance upon the baselines, which suggests our causal learning approach does not hurt the performance on original test set while becoming robust.

results for two different network architectures (ResNet18, ResNet152). All ResNet18 models are trained with SGD for 90 epochs, we follow the standard learning rate schedule where we start from 0.1, and reduce it by 10 times every 30 epochs. For ResNet152 models, we train "ImageNet only" models using the above mentioned method, and finetune all the other methods from the baseline for 40 epochs given that it is computationally expensive to train ResNet-152 models from scratch. For GenInt, we all use $\lambda_1 = 0.05$ and $\lambda_2 = 0$ for ResNet18 and $\lambda_1 = 0.2$ and $\lambda_2 = 0$ for ResNet152. For **GenInt with Transfer**, we use $\lambda_1 = 0.02$ and $\lambda_2 = 1$ for our experiments on Resnet18, and $\lambda_1 = 0.2$ and $\lambda_2 = 0.2$ for our finetuning on ResNet152. We select hyperparameters of our intervention strategy in Section 5.4. Implementation details are in the supplementary.

ObjectNet: Table 1 shows that our model can learn more robust features, and consequently generalizes better to ObjectNet without any additional training. Our results consistently outperform the naive sampling from generative models [39] and other data augmentation strategies [51, 19, 13] for multiple metrics and network architectures, highlighting the difference between traditional data augmentation and our generative intervention. Our approach enjoys benefits by combining with additional data augmentations, demonstrated by the differences between the "Std. Augmentation" columns and the "Add. Augmentation" columns.[2] This improvement suggests that our generative intervention can manipulate additional nuisances (viewpoints, backgrounds, and scene contexts) orthogonal to traditional augmentation, which complements existing data augmentation methods. Moreover, our results suggest that intervening on the generative model is more important than just sampling from it.

ImageNet-C: To further validate that our approach learns causality, and not just overfits, we measure the same models' generalization to unseen corruptions on ImageNet-C. We evaluate performance with mean corruption error (mCE) [22] normalized by CE of AlexNet. Table 2 shows that directly sampling from GAN as augmentation (GAN Augmentation) slightly improves performance (less than 1%). Stylized ImageNet achieves the best performance among all the baselines, but it is still worse than our approach in mCE. In addition, Stylized ImageNet hurts the performance on ObjectNet, which suggests its high performance on corruptions is overfitting to the correlations instead of learning the causality. Our approach outperforms baseline by up to 12.48% and 7.57% on ResNet18 and ResNet152 respectively, which validates that our generative interventions promote causal learning.

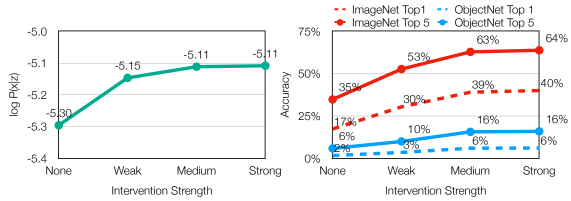

Figure 5. As the strength of the intervention increases, the value of log $P(x|z)$ increases (value calculated on sampled set), which improves the performance of ResNet-18 model.

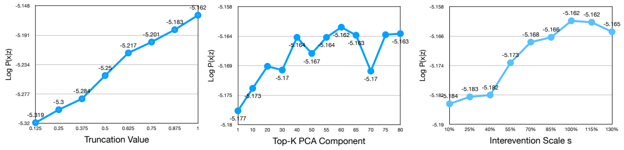

Figure 6. log $P(x|z)$ for causal effect bound under different intervention strategies. The x-axis of each subfigure changes one hyper-parameter for intervention strategy: truncation value t (left), PCA number k (middle), and the intervention scale s (right). Based on theorem 1, we choose the hyper-parameters t, k, s that produces the highest value for log $P(x|z)$ from individual figure.

ImageNet and ImageNet-V2: Table 3 shows the accuracy on both validation sets. Some baselines, such as Stylized ImageNet, hurt the performances on the ImageNet validation set, while our approach improves the performance.

Overall, without trading-off the performance between

[2] Standard augmentation only uses random crop and horizontal flips [1]. Additional augmentation method uses rotation and color jittering [46].

Training Dist.	Truncation	ImageNet top1	ImageNet top5
Obervational GAN [39]	1.0	39.07%	62.97%
Obervational GAN [39]	1.5	42.65%	65.92%
Obervational GAN [39]	2.0	40.98%	64.37%
Interventional GAN (ours)	1.0	**45.06%**	**68.48%**

Table 4. We show performance for ResNet50 trained only on BigGAN. Our intervention model surpasses performance of the best established benchmark [39]

different datasets, our approach achieves improved performance for all test sets, which highlights the advantage of our causal learning approach.

5.4. Analysis

Causal Bound and Performance: Does tighter causal bound lead to a better classifier? Following Theorem 1, we measure the tightness of causal bound after intervention, where we use the log likelihood $\log P(x|z) = \sum_i \sum_{x'_j} \log(P(x_i|x'_j)P(x'_j|z))$, where x_i is the query image from the held out ImageNet validation set, and x'_j is the data generated by intervention z. We train ResNet18 on our generated data.[3] By varying the intervention strength, we increase the value of $P(x|z)$, which corresponds to a tighter causal bound. Figure 5 shows that, as the causal bound getting tighter (left), performance steadily increases (right).

Optimal Intervention Strategy: Since tighter causal bound produces better models, we investigate the optimal intervention strategy for tightening causal bounds. We study the effect of changing t, k, s for our intervention on the causal bound (Section 4.2). We conduct ablation studies and show the trend in Figure 6. We choose $t = 1$, $k = 60$, and $s = 100\%$ as our intervention strategy for tightest causal bound, which produces $\log P(x|z) = -5.162$ and yields the optimal accuracy of 45.06% (Table 4) in practice.

Importance of Intervention: Our results show that creating interventions with a GAN is different from simply augmenting datasets with samples from a GAN. To examine this, Table 4 shows performance on ImageNet when the training sets only consist of images from the GAN. We use the best practices from [39], which comprehensively studies GAN image generation as training data. Our results show that creating interventions, not just augmentations, improves classification performance by 2.4%-6.0%.

5.5. Model Visualization

By removing the confounding factors in the dataset, we expect the model to learn to attend tightly to the spatial regions corresponding to the object, and not spuriously correlated contextual regions. To analyze this, Figure 7 uses GradCAM [45] to visualize what regions the models use for making prediction. While the baseline often attends to the background or other nuisances for prediction, our method focuses on the spatial features of the object. For example, for the first 'Broom' image, the baseline uses spurious context 'hand,' leading to a misprediction 'Ladle,' while our model predicts the right 'Broom' by looking at its shape. This suggests that, in addition to performance gains, our model predicts correctly for the right reasons.

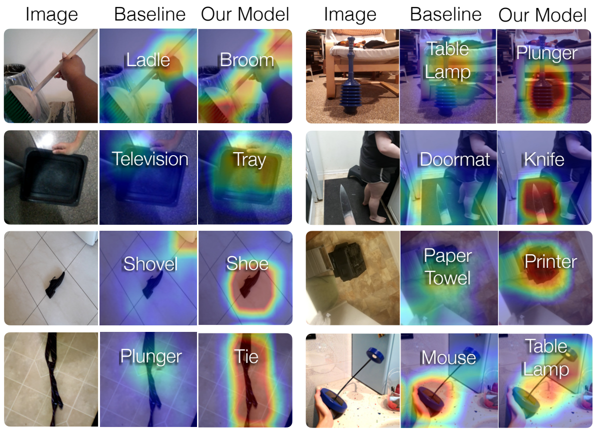

Figure 7. We visualize the input regions that the model uses to make predictions. Blue implies the model ignores the region for discrimination, while red implies the region is very discriminative. The white text shows the model's top prediction. The baseline frequently latches onto spurious background context (e.g., hand spuriously correlated with ladle, chair spuriously correlated with tablelamp), and consequently makes the wrong prediction. Meanwhile, our model often predicts correctly for the right reasons.

6. Conclusion

Fortifying visual recognition for an unconstrained environment remains an open challenge in the field. We introduce a method for learning discriminative visual models that are consistent with causal structures, which enables robust generalization. By steering generative models to construct interventions, we are able to randomize many features without being affected by confounding factors. We show a theoretical guarantee for learning causal classifiers under imperfect interventions, and demonstrate improved performance on ImageNet, ImageNet-C, ImageNet-V2, and the systematically controlled ObjectNet.

Acknowledgments: This research is based on work partially supported by NSF NRI Award #1925157, the DARPA SAIL-ON program under PTE Federal Award No. W911NF2020009, an Amazon Research Gift, NSF grant CNS-15-64055, ONR grants N00014-16-1- 2263 and N00014-17-1-2788, a JP Morgan Faculty Research Award, and a DiDi Faculty Research Award. We thank NVidia for GPU donations. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government.

[3] We sample an observational and intervention data from BigGAN with truncation 0.5 [10]. Please see supplementary material for full details.

References

[1] Pytorch imagenet tutorial.

[2] Transforming and projecting images to class-conditional generative networks.

[3] Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

[4] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *Advances in Neural Information Processing Systems*, pages 9447–9460, 2018.

[5] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[6] 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.

[7] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks, 2017.

[8] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

[9] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4502–4511, 2019.

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[11] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 181–190, Arlington, Virginia, USA, 2015. AUAI Press.

[12] Shorten Connor and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 2019.

[13] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2018.

[14] 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.

[15] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

[16] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans, 2019.

[17] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 289–293, 2018.

[18] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[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.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

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

[23] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020.

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

[25] Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *In Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1998.

[26] Ali Jahanian*, Lucy Chai*, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

[27] Dominik Janzing. Causal regularization. In *Advances in Neural Information Processing Systems*, pages 12704–12714, 2019.

[28] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[30] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the*

26th International Joint Conference on Artificial Intelligence, pages 2230–2236, 2017.

[31] Raha Moraffah, Kai Shu, Adrienne Raglin, and Huan Liu. Deep causal representation learning for unsupervised domain adaptation, 2019.

[32] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks, 2019.

[33] Judea Pearl. Causality: Models, reasoning, and inference, 2003.

[34] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[35] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, Oct 2016.

[36] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[38] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing, 2020.

[39] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 12268–12279. Curran Associates, Inc., 2019.

[40] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14866–14876, 2019.

[41] 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.

[42] Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[43] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[44] Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, pages 10220–10230, 2019.

[45] 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. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019.

[46] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need?, 2020.

[47] V. Vapnik. Principles of risk minimization for learning theory. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 831–838. Morgan-Kaufmann, 1992.

[48] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *ICML*, 2020.

[49] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[50] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

[51] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[52] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. In *International Conference on Learning Representations*, 2020.

[53] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016.

[54] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 349–360, Cham, 2018. Springer International Publishing.