On the Impact of Spurious Correlation for Out-of-Distribution Detection

Yifei Ming, Hang Yin, Yixuan Li

Department of Computer Sciences, University of Wisconsin-Madison
{alvinming, hyin56, sharonli}@cs.wisc.edu

Abstract

Modern neural networks can assign high confidence to inputs drawn from outside the training distribution, posing threats to models in real-world deployments. While much research attention has been placed on designing new out-of-distribution (OOD) detection methods, the precise definition of OOD is often left in vagueness and falls short of the desired notion of OOD in reality. In this paper, we present a new formalization and model the data shifts by taking into account both the invariant and environmental (spurious) features. Under such formalization, we systematically investigate how spurious correlation in the training set impacts OOD detection. Our results suggest that the detection performance is severely worsened when the correlation between spurious features and labels is increased in the training set. We further show insights on detection methods that are more effective in reducing the impact of spurious correlation, and provide theoretical analysis on why reliance on environmental features leads to high OOD detection error. Our work aims to facilitate better understandings of OOD samples and their formalization, as well as the exploration of methods that enhance OOD detection. Code is available at https://github.com/deeplearning-wisc/Spurious_OOD.

Introduction

Modern deep neural networks have achieved unprecedented success in known contexts for which they are trained, yet they do not necessarily know what they don't know (Nguyen, Yosinski, and Clune 2015). In particular, neural networks have been shown to produce high posterior probability for test inputs from out-of-distribution (OOD), which should not be predicted by the model. This gives rise to the importance of OOD detection, which aims to identify and handle unknown OOD inputs so that the algorithm can take safety precautions.

Before we attempt any solution, an important yet often overlooked problem is: what do we mean by out-of-distribution data? While the research community lacks a consensus on the precise definition, a common evaluation protocol views data with non-overlapping semantics as OOD inputs (Hendrycks and Gimpel 2017). For example, an image of a *cow* can be viewed as an OOD *w.r.t* a model tasked to classify *cat* vs. *dog*. However, such an evaluation scheme is often oversimplified and may not capture the nuances and complexity of the problem in reality.

We begin with a motivating example where a neural network can rely on statistically informative yet *spurious* features in the data. Indeed, many prior works showed that modern neural networks can spuriously rely on the biased features (e.g., background or textures) instead of features of the object to achieve high accuracy (Beery, Van Horn, and Perona 2018; Geirhos et al. 2019; Sagawa et al. 2019). In Figure 1, we illustrate a model that exploits the spurious correlation between the *water background* and label *waterbird* for prediction. Consequently, a model that relies on spurious features can produce a high-confidence prediction for an OOD input with the same background (*i.e.*, water) but a different semantic label (*e.g.*, boat). This can manifest in downstream OOD detection, yet unexplored in prior works. In this paper, we systematically investigate how spurious correlation in the training set impacts OOD detection. We first provide a new formalization and explicitly model the data shifts by taking into account both **invariant** features and **environmental** features (Section). Invariant features can be viewed as essential cues directly related to semantic labels, whereas environmental features are non-invariant and can be spurious. Our formalization encapsulates two types of OOD data: (1) *spurious OOD*—test samples that contain environmental (non-invariant) features but no invariant features; (2) *non-spurious OOD*—inputs that contain neither the environmental nor invariant features, which is more in line with the conventional notion of OOD. We provide an illustration of both types of OOD in Figure 1.

Under the new formalization, we conduct extensive experiments and investigate the detection performance under both spurious and non-spurious OOD inputs (Section). Our results suggest that spurious correlation in the training data poses a significant challenge to OOD detection. For both spurious and non-spurious OOD samples, the detection performance is severely worsened when the correlation between spurious features and labels is increased in the training set. Further, we comprehensively evaluate common OOD detection approaches, and show that feature-based methods have a competitive edge in improving non-spurious OOD detection, while detecting spurious OOD remains challenging (Section). To further understand this, we provide theoretical insights on why reliance on non-invariant features leads to high OOD detection error (Section). We provably show the existence of spurious OOD inputs with arbitrarily high confidence,

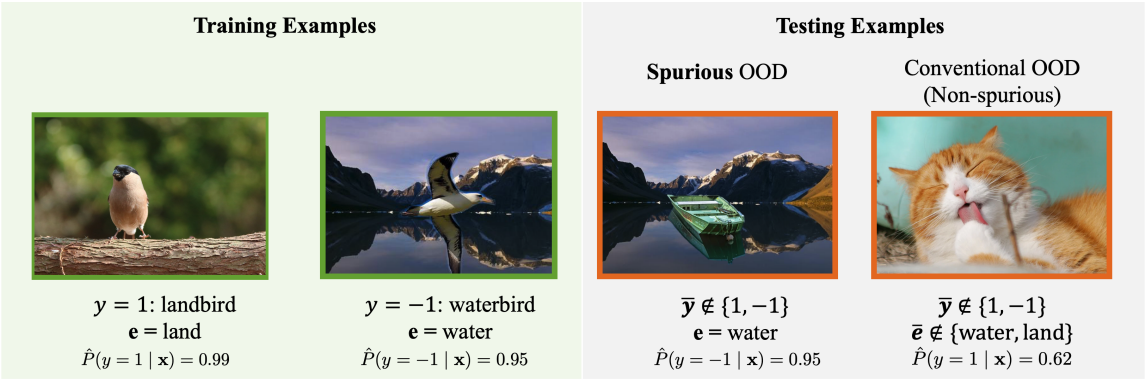


Figure 1: Left (train): The training examples $\mathbf{x}$ are generated by a combination of invariant features, dependent on the label y; and environmental features, dependent on the environment e. In Waterbirds dataset (Sagawa et al. 2019), $y \in \{\textit{waterbird}, \textit{landbird}\}$ is correlated with the environment $e \in \{\text{water}, \text{land}\}$. Right (test): During test time, we consider two types of OOD inputs. Spurious OOD inputs contain the environmental features, but no signals related to the in-distribution classes. Non-spurious OOD inputs have neither environmental features nor invariant features. Confidence scores are computed from a ResNet-18 model trained on Waterbirds (Sagawa et al. 2019).

which can fail to be distinguished from the ID data. Our **key contributions** are as follows:

- We provide a new formalization of OOD detection by explicitly taking into account the separation between invariant features and environmental features. Our formalization encapsulates both spurious and non-spurious OOD. Our work, therefore, provides a complementary perspective in the evaluation of OOD detection.
- We provide systematic investigations on how the extent of spurious correlation in the training set impacts OOD detection. We further show insights on OOD detection solutions that are more effective in mitigating the impact of spurious correlation, with up to 46.73% reduction of FPR95 in detecting non-spurious OOD data.
- We provide theoretical analysis, provably showing that detecting spurious OOD samples remains challenging due to the model's reliance on the environmental features.

Our study provides strong implications for future research on out-of-distribution detection. Our study signifies the importance for future works to evaluate OOD detection algorithms on spurious OOD examples besides standard benchmarks (most of which are non-spurious) to test the limits of the approaches. We hope that our work will inspire future research on the formalization of the OOD detection problem and algorithmic solutions.

A New Formalization of Out-of-Distribution

Data Model. We consider supervised multi-class classification, where $\mathcal{X} = \mathbb{R}^d$ denotes the input space and $\mathcal{Y} = \{1, 2, ..., K\}$ denotes the label space. We assume that the data is drawn from a set of E environments (domains) $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$. An input $\mathbf{x} := \tau(\mathbf{z}_{\text{inv}}, \mathbf{z}_e)$ is generated by a combination of invariant features $\mathbf{z}_{\text{inv}} \in \mathbb{R}^s$ which are dependent on the label y, and environmental features $\mathbf{z}_e \in \mathbb{R}^{d_e}$. τ is a function transformation from the latent features $[\mathbf{z}_{\text{inv}}, \mathbf{z}_e]^\top$ to the pixel-space $\mathcal{X}$. The signal $\mathbf{z}_{\text{inv}}$ are the cues essential for the recognition of $\mathbf{x}$ as y; examples include the color, the shape of beaks and claws, and fur patterns of birds for classifying *waterbird* vs. *landbird*. Environmental features $\mathbf{z}_e$, on the other hand, are cues not essential for the recognition but correlated with the target y. For example, many waterbird images are taken in water habitat, so water scenes can be considered as $\mathbf{z}_e$. Under the data model, we have a joint distribution $P(\mathbf{x}, y, e)$. Each $g = (y, e) \in \mathcal{Y} \times \mathcal{E}$ group has its own distribution over features $[\mathbf{z}_{\text{inv}}, \mathbf{z}_e] \in \mathbb{R}^{s+d_e}$. Let $\mathcal{D}_{\text{in}}^e$ denote the marginal distribution on $\mathcal{X}$ for environment e. The union of distributions $\mathcal{D}_{\text{in}}^e$ over all environments is the in-distribution $\mathcal{D}_{\text{in}}$.

Out-of-distribution Data. In practice, OOD refers to samples from an irrelevant distribution whose label set has no intersection with $\mathcal{Y}$, and therefore should not be predicted by the model. Under our data model, we define data distributional shifts by explicitly taking into account the separation between invariant features and environmental features. Concretely, our formalization encapsulates two types of OOD data defined below.

- **Spurious OOD** is a particularly challenging type of inputs, which contain the *environmental feature, but no invariant feature essential for the label*. Formally, we denote by $\mathbf{x} = \tau(\mathbf{z}_{\bar{Y}}, \mathbf{z}_e)$, where $\mathbf{z}_{\bar{Y}}$ is from an out-of-class label $\bar{Y} \notin \mathcal{Y}$. For example, this can be seen in Figure 1 (middle right), where the OOD example contains the semantic feature *boat* $\notin \{\textit{waterbird}, \textit{landbird}\}$, yet it has the environmental feature of water background.
- **Non-spurious (conventional) OOD** are inputs that contain *neither the environmental nor the invariant features*, i.e., $\mathbf{x} = \tau(\mathbf{z}_{\bar{Y}}, \mathbf{z}_{\bar{e}})$. In particular, $\mathbf{z}_{\bar{Y}}$ is sampled from an out-of-class label $\bar{Y} \notin \mathcal{Y}$, and $\mathbf{z}_{\bar{e}}$ is sampled from a different environment $\bar{e} \notin \mathcal{E}$. For example, an input of an *indoor cat* falls into this category, where both the semantic label *cat* and environment *indoor* are distinct from the in-distribution data of waterbirds and landbirds.

OOD Type	Test Set	r=0.5		r=0.7		r=0.9	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC ↑
Spurious OOD		59.89 ± 12.40	88.54 ± 4.81	74.22 ± 13.12	80.98 ± 4.45	74.39 ± 12.50	79.81 ± 8.43
Non-spurious OOD	iSUN	19.69 ± 10.66	91.88 ± 4.52	43.22 ± 12.50	91.81 ± 3.32	57.40 ± 15.54	82.45 ± 7.98
	LSUN	22.60 ± 12.08	90.80 ± 3.33	43.30 ± 16.66	90.09 ± 4.51	52.68 ± 13.70	84.56 ± 8.56
	SVHN	15.32 ± 5.05	95.71 ± 2.20	25.53 ± 8.11	95.60 ± 2.45	43.89 ± 23.80	93.27 ± 6.90

Table 1: OOD detection performance of models trained on Waterbirds (Sagawa et al. 2019). Increased spurious correlation in the training set results in worsen performance for both non-spurious and spurious OOD samples. In particular, spurious OOD is more challenging than non-spurious OOD samples. Results (mean and std) are estimated over 4 runs for each setting.

Out-of-distribution Detection. OOD detection can be viewed as a binary classification problem. Let $f : \mathcal{X} \to \mathbb{R}^K$ be a neural network trained on samples drawn from the data distribution defined above. During inference time, OOD detection can be performed by exercising a thresholding mechanism $G_\lambda(\mathbf{x}; f) = \mathbf{1}\{S(\mathbf{x}; f) \geq \lambda\}$, where samples with higher scores $S(\mathbf{x}; f)$ are classified as ID and vice versa. The threshold λ is typically chosen so that a high fraction of ID data (*e.g.,* 95%) is correctly classified.

How Does Spurious Correlation Impact OOD Detection?

During training, a classifier may learn to rely on the association between environmental features and labels to make its predictions. Moreover, we hypothesize that such a reliance on environmental features can cause failures in the downstream OOD detection. To verify this, we begin with the most common training objective empirical risk minimization (ERM). Given a loss function ℓ, ERM finds the model w that minimizes the average training loss:

$$\hat{\mathcal{R}}(w) = \mathbb{E}_{(\mathbf{x},y,e)\sim\hat{P}}[\ell(w; (\mathbf{x}, y, e))]. \tag{1}$$

We now describe the datasets we use for model training and OOD detection tasks. We consider three tasks that are commonly used in the literature. We start with a natural image dataset Waterbirds, and then move onto the CelebA dataset (Liu et al. 2015). Due to space constraints, a third evaluation task on ColorMNIST is in the Appendix[1].

Evaluation Task 1: Waterbirds. Introduced in (Sagawa et al. 2019), this dataset is used to explore the spurious correlation between the image background and bird types, specifically $\mathcal{E} \in \{\textit{water}, \textit{land}\}$ and $\mathcal{Y} \in \{\textit{waterbirds}, \textit{landbirds}\}$. We also control the correlation between y and e during training as $r \in \{0.5, 0.7, 0.9\}$. The correlation r is defined as $r = P(e = \textit{water} \mid y = \textit{waterbirds}) = P(e = \textit{land} \mid y = \textit{landbirds})$. For spurious OOD, we adopt a subset of images of land and water from the Places dataset (Zhou et al. 2017). For non-spurious OOD, we follow the common practice and use the *SVHN* (Netzer et al. 2011), *LSUN* (Yu et al. 2015), and *iSUN* (Xu et al. 2015) datasets.

Evaluation Task 2: CelebA. In order to further validate our findings beyond background spurious (environmental) features, we also evaluate on the CelebA (Liu et al. 2015) dataset. The classifier is trained to differentiate the hair color (grey vs. non-grey) with $\mathcal{Y} = \{\textit{grey hair}, \textit{nongrey hair}\}$. The environments $\mathcal{E} = \{\textit{male}, \textit{female}\}$ denote the gender of the person. In the training set, "Grey hair" is highly correlated with "Male", where 82.9% ($r \approx 0.8$) images with grey hair are male. Spurious OOD inputs consist of *bald male*, which contain environmental features (gender) without invariant features (hair). The non-spurious OOD test suite is the same as above (*SVHN*, *LSUN*, and *iSUN*). Figure 2 illustates ID samples, spurious and non-spurious OOD test sets. We also subsample the dataset to ablate the effect of r; see results are in the Appendix.

Results and Insights. We train on ResNet-18 (He et al. 2016) for both tasks. See Appendix for details on hyperparameters and in-distribution performance. We summarize the OOD detection performance in Table 1 (Waterbirds), Table 2 (CelebA) and Table 4 (ColorMNIST).

There are several salient observations. **First**, for both spurious and non-spurious OOD samples, the detection performance is severely worsened when the correlation between spurious features and labels is increased in the training set. Take the Waterbirds task as an example, under correlation $r = 0.5$, the average false positive rate (FPR95) for spurious OOD samples is 59.89%, and increases to 74.39% when $r = 0.9$. Similar trends also hold for other datasets. **Second**, spurious OOD is much more challenging to be detected compared to non-spurious OOD. From Table1, under correlation $r = 0.7$, the average FPR95 is 37.35% for non-spurious OOD, and increases to 74.22% for spurious OOD. Similar observations hold under different correlation and different training datasets. **Third**, for non-spurious OOD, samples that are more semantically dissimilar to ID are easier to detect. Take Waterbirds as an example, images containing scenes (e.g. LSUN and iSUN) are more similar to the training samples compared to images of numbers (e.g. SVHN), resulting in higher FPR95 (e.g. 43.22% for iSUN compared to 25.53% for SVHN under $r = 0.7$).

Our results suggest that spurious correlation poses a significant threat to the model. In particular, a model can produce high-confidence predictions on the spurious OOD, due to the reliance on the environmental feature (*e.g.*, background information) rather than the invariant feature (*e.g.*, bird species). To verify that the spurious feature causes poor detection performance, we show that the classifier frequently predicts the spurious OOD as the ID class with the same environmental feature. For Waterbirds, on average 93.9% of OOD samples

[1]Appendix is available at https://arxiv.org/abs/2109.05642

Figure 2: For CelebA, the classifier is trained to differentiate the hair color (grey vs. non-grey). Left: Training environments. 82.9% images with grey hair are male, whereas 82.9% images with non-grey hair are female. Middle: Spurious OOD inputs contain the environmental feature (male) without invariant features (hair). Right: Non-spurious OOD samples consist of images with diverse semantics without human faces.

with water background is classified as waterbirds, and 80.7% of OOD samples with land background is classified as land birds. For the CelebA dataset, on average 86.5% of spurious OOD samples (bold male) are classified as grey hair. Note that our results here are based on the energy score (Liu et al. 2020), which is one competitive detection method derived from the model output (logits) and has shown superior OOD detection performance over directly using the predictive confidence score. Next, we provide an expansive evaluation using a broader suite of OOD scoring functions in Section .

OOD Type	Test Set	FPR95 ↓	AUROC ↑
Spurious OOD		71.28 ± 4.12	82.04 ± 2.64
Non-spurious OOD	iSUN	17.35 ± 2.97	97.03 ± 0.30
	LSUN	18.85 ± 2.44	96.90 ± 0.17
	SVHN	5.63 ± 2.60	98.64 ± 0.21

Table 2: OOD detection performance of models trained on CelebA (Liu et al. 2015) with $r \approx 0.8$. Spurious OOD test data incurs much higher FPR than non-spurious OOD data. Results (mean and std) are estimated over 4 runs for each setting.

How to Reduce the Impact of Spurious Correlation for OOD Detection?

The results in the previous section naturally prompt the question: how can we better detect spurious and non-spurious OOD inputs when the training dataset contains spurious correlation? In this section, we comprehensively evaluate common OOD detection approaches, and show that feature-based methods have a competitive edge in improving non-spurious OOD detection, while detecting spurious OOD remains challenging (which we further explain theoretically in Section).

Feature-based vs. Output-based OOD Detection. Section suggests that OOD detection becomes challenging for output-based methods especially when the training set contains high spurious correlation. However, the efficacy of using representation space for OOD detection remains unknown. In this section, we consider a suite of common scoring functions including maximum softmax probability (MSP) (Hendrycks and Gimpel 2017), ODIN score (Liang, Li, and Srikant 2018; Hsu et al. 2020), Mahalanobis distance-based score (Lee et al. 2018), energy score (Liu et al. 2020), and Gram matrix-based score (Sastry and Oore 2020)—all of which can be derived *post hoc*[2] from a trained model. Among those, Mahalanobis and Gram Matrices can be viewed as feature-based methods. For example, Lee et al. (2018) estimates class-conditional Gaussian distributions in the representation space and then uses the maximum Mahalanobis distance as the OOD scoring function. Data points that are sufficiently far away from all the class centroids are more likely to be OOD.

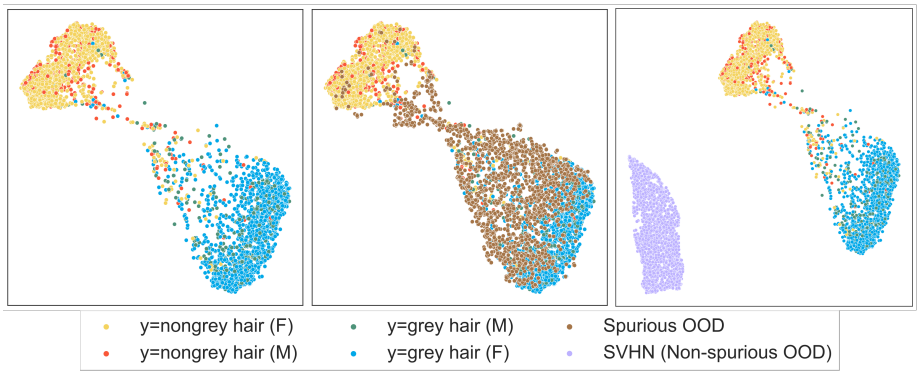
Results. The performance comparison is shown in Table 3 (full table in the Arxiv version). Several interesting observations can be drawn. **First**, we can observe a significant performance gap between *spurious OOD* (SP) and *non-spurious OOD* (NSP), irrespective of the OOD scoring function in use. This observation is in line with our findings in Section . **Second**, the OOD detection performance is generally improved with the feature-based scoring functions such as Mahalanobis distance score (Lee et al. 2018) and Gram Matrix score (Sastry and Oore 2020), compared to scoring functions based on the output space (*e.g.*, MSP, ODIN, and energy). The improvement is substantial for non-spurious OOD data. For example, on Waterbirds, FPR95 is reduced by 46.73% with Mahalanobis score compared to using MSP score. For spurious OOD data, the performance improvement is most pronounced using the Mahalanobis score. Noticeably, using the Mahalanobis score, the FPR95 is reduced by 28.02% on ColorMNIST, compared to using the MSP score. Our results suggest that feature space preserves useful information that can more effectively distinguish between ID and OOD data.

Analysis and Visualizations. To provide further insights on why the feature-based method is more desirable, we show the visualization of embeddings in Figure 3a. The visualization is based on the CelebA task. From Figure 3a (left), we observe a clear separation between the two class labels. Within each class label, data points from both environments are well mixed (e.g., see the green and blue dots). In Figure 3a (middle), we visualize the embedding of ID data together with spurious OOD inputs, which contain the environmental feature (*male*). Spurious OOD (bold male) lies between the two ID clusters, with some portion overlapping with the ID samples, signifying the hardness of this type of OOD. This

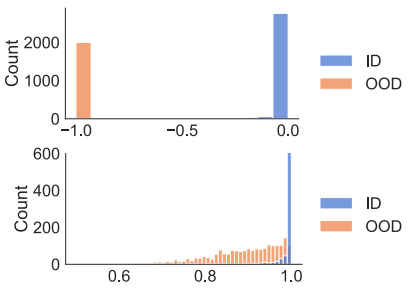
[2]Note that Generalized-ODIN requires modifying the training objective and model retraining. For fairness, we primarily consider strict post-hoc methods based on the standard cross-entropy loss.

Method	MSP				ODIN				Mahalanobis				Energy			
Metric	FPR95↓		AUROC↑		FPR95↓		AUROC↑		FPR95↓		AUROC↑		FPR95↓		AUROC↑	
ID Data	SP	NSP	SP	NSP	SP	NSP	SP	NSP	SP	NSP	SP	NSP	SP	NSP	SP	NSP
CMNIST	42.99	3.15	77.75	99.13	38.06	1.88	78.78	99.01	14.97	0.04	88.65	99.54	30.45	7.65	86.74	97.54
Waterbirds	74.68	47.53	79.22	92.34	77.25	34.06	81.04	93.48	69.35	0.80	82.73	99.51	74.22	37.35	80.98	92.50
CelebA	83.70	22.60	68.22	90.21	81.07	11.49	75.22	89.11	78.75	2.33	83.12	98.93	71.28	13.94	82.04	97.51

Table 3: Performance for different post-hoc OOD detection methods when the spurious correlation is high in the training set. We choose $r = 0.45$ for ColorMNIST, $r = 0.7$ for Waterbirds, and $r = 0.8$ for CelebA. SP stands for Spurious OOD test set. NSP denotes non-spurious OOD, where the results are averaged over 3 OOD test sets.



(a) UMAP (McInnes et al. 2018) Visualization of feature embeddings



(b) Histograms of different scores

Figure 3: (a) Left: Feature for in-distribution data only. (a) Middle: Feature for both ID and spurious OOD data. (a) Right: Feature for ID and non-spurious OOD data (SVHN). M and F in parentheses stand for male and female respectively. (b) Histogram of Mahalanobis score (top) and MSP score (bottom) for ID and the non-spurious OOD dataset SVHN. Full results for other non-spurious OOD datasets (iSUN and LSUN) are in the Appendix.

is in stark contrast with non-spurious OOD inputs shown in Figure 3a (right), where a clear separation between ID and OOD (purple) can be observed. This shows that feature space contains useful information that can be leveraged for OOD detection, especially for conventional non-spurious OOD inputs. Moreover, by comparing the histogram of Mahalanobis distance (top) and MSP score (bottom) in Figure 3b, we can further verify that ID and OOD data is much more separable with the Mahalanobis distance. Therefore, our results suggest that feature-based methods show promise for improving non-spurious OOD detection when the training set contains spurious correlation, while there still exists large room for improvement on spurious OOD detection.

Why Is It Hard to Detect Spurious OOD?

Given the results above, a natural question arises: why is it hard to detect spurious OOD inputs? To better understand this issue, we now provide theoretical insights. In what follows, we first model the ID and OOD data distributions and then derive mathematically the model output of invariant classifier, where the model aims not to rely on the environmental features for prediction.

Setup. We consider a binary classification task where $y \in \{-1, 1\}$, and is drawn according to $\eta := P(y = 1)$. We assume both the invariant features $\mathbf{z}_{\text{inv}}$ and environmental features $\mathbf{z}_e$ are drawn from Gaussian distributions:

$$\mathbf{z}_{\text{inv}} \sim \mathcal{N}\left(y \cdot \boldsymbol{\mu}_{\text{inv}}, \sigma_{\text{inv}}^2 I\right), \quad \mathbf{z}_e \sim \mathcal{N}\left(y \cdot \boldsymbol{\mu}_e, \sigma_e^2 I\right)$$

where $\boldsymbol{\mu}_e \in \mathbb{R}^{d_e}$, $\boldsymbol{\mu}_{\text{inv}} \in \mathbb{R}^s$, and I is the identity matrix. Note that the parameters $\boldsymbol{\mu}_{\text{inv}}$ and σ_{inv}^2 are the same for all environments. In contrast, the environmental parameters $\boldsymbol{\mu}_{\text{e}}$ and σ_e^2 are different across e, where the subscript is used to indicate the dependence on the environment and the index of the environment. In what follows, we present the results, with detailed proof deferred in the Appendix.

Lemma 1 *(Bayes optimal classifier) For any feature vector which is a linear combination of the invariant and environmental features $\Phi_e(\mathbf{x}) = M_{inv}\mathbf{z}_{inv} + M_e\mathbf{z}_e$, the optimal linear classifier for an environment e has the corresponding coefficient $2\Sigma_\Phi^{-1}\boldsymbol{\mu}_\Phi$, where:*

$$\boldsymbol{\mu}_\Phi = M_{inv}\boldsymbol{\mu}_{inv} + M_e\boldsymbol{\mu}_e$$
$$\Sigma_\Phi = M_{inv}M_{inv}^T\sigma_{inv}^2 + M_eM_e^T\sigma_e^2$$

Note that the Bayes optimal classifier uses environmental features which are informative of the label but non-invariant. Rather, we hope to rely *only* on invariant features while ignoring environmental features. Such a predictor is also referred to as *optimal invariant predictor* (Rosenfeld, Ravikumar, and Risteski 2021), which is specified in the following. Note that this is a special case of Lemma 1 with $M_{\text{inv}} = I$ and $M_e = 0$.

Proposition 1 *(Optimal invariant classifier using invariant features) Assume the featurizer recovers the invariant feature* $\Phi_e(\mathbf{x}) = [\mathbf{z}_{inv}]\ \forall e \in \mathcal{E}$*, the optimal invariant classifier has the corresponding coefficient* $2\boldsymbol{\mu}_{inv}/\sigma^2_{inv}$.[3]

The optimal invariant classifier explicitly ignores the environmental features. However, an invariant classifier learned does not necessarily depend only on the invariant features. Next Lemma shows that *it can be possible to learn an invariant classifier that relies on the environmental features while achieving lower risk than the optimal invariant classifier.*

Lemma 2 *(Invariant classifier using non-invariant features) Suppose* $E \leq d_e$*, given a set of environments* $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$ *such that all environmental means are linearly independent. Then there always exists a unit-norm vector* $\mathbf{p}$ *and positive fixed scalar* β *such that* $\beta = \mathbf{p}^T\boldsymbol{\mu}_e/\sigma^2_e\ \forall e \in \mathcal{E}$*. The resulting optimal classifier weights are*

$$\hat{w} = \begin{bmatrix} \beta_{inv} \\ 2\beta \end{bmatrix} = \begin{bmatrix} 2\boldsymbol{\mu}_{inv}/\sigma^2_{inv} \\ 2\mathbf{p}^\top\boldsymbol{\mu}_e/\sigma^2_e \end{bmatrix}.$$

Note that the optimal classifier weight 2β is a constant, which does not depend on the environment (and neither does the optimal coefficient for $\mathbf{z}_{\text{inv}}$). The projection vector $\mathbf{p}$ acts as a "short-cut" that the learner can use to yield an insidious surrogate signal $\mathbf{p}^\top\mathbf{z}_e$. Similar to $\mathbf{z}_{\text{inv}}$, this insidious signal can also lead to an invariant predictor (across environments) admissible by invariant learning methods. In other words, despite the varying data distribution across environments, the optimal classifier (using non-invariant features) is the same for each environment. We now show our main results, where OOD detection can fail under such an invariant classifier.

Theorem 1 *(Failure of OOD detection under invariant classifier) Consider an out-of-distribution input which contains the environmental feature:* $\Phi_{out}(\mathbf{x}) = M_{inv}\mathbf{z}_{out} + M_e\mathbf{z}_e$*, where* $\mathbf{z}_{out} \perp \boldsymbol{\mu}_{inv}$*. Given the invariant classifier (cf. Lemma 2), the posterior probability for the OOD input is* $p(y = 1 \mid \Phi_{out}) = \sigma\left(2\mathbf{p}^\top\mathbf{z}_e\beta + \log\eta/(1-\eta)\right)$*, where* σ *is the logistic function. Thus for arbitrary confidence* $0 < c := P(y = 1 \mid \Phi_{out}) < 1$*, there exists* $\Phi_{out}(\mathbf{x})$ *with* $\mathbf{z}_e$ *such that* $\mathbf{p}^\top\mathbf{z}_e = \frac{1}{2\beta}\log\frac{c(1-\eta)}{\eta(1-c)}$.

Our theorem above signifies the existence of OOD inputs that can trigger high-confidence predictions on in-distribution classes yet contain no meaningful feature related to the labels in $\mathcal{Y} = \{1, -1\}$ at all. An OOD detector can fail to detect these inputs with predictions that are indistinguishable from ID data. We provide a simple toy example to explain this phenomenon further.

An Intuitive Example. An illustrative example with two environments is in Figure 4 (Left). The feature representations for examples in environments 1 and 2 are shown as circle and diamond, respectively. In-distribution samples with different colors correspond to different labels: yellow indicates $y = 1$ and green indicates $y = -1$. The decision boundary of classification is denoted by the dashed line, which

[3]The constant term in the classifier weights is $\log\eta/(1-\eta)$, which we omit here and in the sequel.

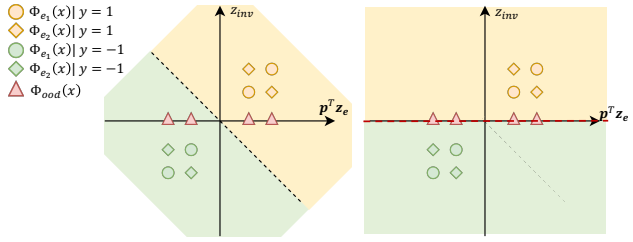


Figure 4: Left: The invariant decision boundary (dashed line) is based on both the invariant feature z_{inv} and environmental features z_e. OOD inputs (red triangles) can be predicted as in-distribution with high confidence, therefore can fail to be detected by OOD methods (e.g., using predictive confidence threshold). Right: An ideal case when the invariant decision boundary is purely based on z_{inv} (red dashed line). The OOD inputs lie on the decision boundary and will be predicted as $y = 1$ or $y = -1$ with probability 0.5.

relies on both the invariant features $\mathbf{z}_{\text{inv}}$ and environmental features $\mathbf{z}_e$. It can be seen that if the feature representation relies on environmental features $\mathbf{p}^\top\mathbf{z}_e$, spurious OOD samples (red triangles) can trick the classifier into recognizing OOD samples as one of the in-distribution classes with high confidence, posing severe threats to OOD detection.

In contrast, under an ideal case when the invariant classifier only uses invariant features $\mathbf{z}_{\text{inv}}$, the optimal decision boundary is a horizontal dashed line, as shown in Figure 4 (Right). OOD inputs (red triangles) will be predicted with a probability of 0.5 since they lie on the decision boundary.

Remark. As a special case, if the representation consists purely of environmental features, i.e., $\Phi_e(\mathbf{x}) = [\mathbf{z}_e]$, the resulting optimal classifier weights are $2\mathbf{p}^\top\mu_e/\sigma^2_e = 2\beta$, a fixed scalar that is still invariant across environments. Lemma 3 below shows that such a predictor can yield low risks under certain conditions. Our main theorem above still holds under such a predictor.

Lemma 3 *(Existence of purely environmental predictors with low risks (Rosenfeld, Ravikumar, and Risteski 2021)) There exists a representation constructed purely relying on environmental features based on the short-cut direction* $\mathbf{p}$ *that achieves lower risk than the optimal invariant predictor on every environment e such that* $\sigma_e\beta > \sigma^{-1}_{inv}\left\|\mu_{inv}\right\|_2$ *and* $2\sigma_e\beta\sigma^{-1}_{inv}\left\|\mu_{inv}\right\|_2 \geq |\log\eta/(1-\eta)|$.

Summary. To summarize, the theoretical analysis demonstrates the difficulty of recovering the invariant classifier without using environmental features. In particular, there exists an invariant classifier that uses non-invariant features, and achieves lower risks than the classifiers only based on invariant features. As a result, spurious OOD samples can utilize environmental clues to deteriorate the OOD detection performance. Our main theorem provably shows the existence of OOD inputs with arbitrarily high confidence, and can fail to be distinguished from the ID data.

Extension: Empirical Validation of Theoretical Analysis. To further validate our analysis above, we evaluate the OOD

detection performance of models that are trained with recent prominent domain invariance learning objectives (Arjovsky et al. 2019; Bahng et al. 2020; Krueger et al. 2020; Ganin et al. 2016; Li et al. 2018b; Sagawa et al. 2019) (Section E in Appendix). The results align with our theoretical analysis.

Discussion and Related Works

Out-of-Distribution Uncertainty Estimation. The phenomenon of neural networks' overconfidence to out-of-distribution data is revealed by Nguyen *et al.* (Nguyen, Yosinski, and Clune 2015). Early works attempt to improve the OOD uncertainty estimation by proposing the ODIN score (Liang, Li, and Srikant 2018) and Mahalanobis distance-based confidence score (Lee et al. 2018). Recent work by Liu *et al.* (Liu et al. 2020) proposed using an energy score for OOD detection, which demonstrated advantages over the softmax confidence score both empirically and theoretically. Huang and Li (Huang and Li 2021) proposed a group-based OOD detection method for large-scale datasets. Recent work by Lin *et al.* (Lin, Roy, and Li 2021) proposed a dynamic OOD inference framework to improve the computational efficiency. However, previous methods primarily focused on convention non-spurious OOD. We introduce a new formalization of OOD detection that encapsulates both spurious and non-spurious OOD data.

A parallel line of approaches resorts to generative models (Goodfellow et al. 2014; Kingma and Dhariwal 2018) that directly estimate in-distribution density (Nalisnick et al. 2019; Ren et al. 2019; Serrà et al. 2020; Xiao, Yan, and Amit 2020; Kirichenko, Izmailov, and Wilson 2020). In particular, Ren et al. (2019) addressed distinguishing between background and semantic content under unsupervised generative models. Generative approaches yield limiting performance compared with supervised discriminative models and typically suffer from high computational complexity. Notably, none of the previous works systematically investigate the influence of spurious correlation for OOD detection. Our work presents a novel perspective for defining OOD data and investigates the impact of spurious correlation in the training set. Moreover, our general formulation extends beyond the background spurious correlation (*e.g.* gender bias).

Hard OOD Evaluations. Our proposed spurious OOD can be viewed as a form of hard OOD evaluation. Orthogonal to our work, previous works (Winkens et al. 2020; Roy et al. 2021) considered the hard cases where the *semantics* of OOD inputs are similar to that of ID data (*e.g.*, CIFAR-10 vs. CIFAR-100). In our setting, spurious OOD inputs may have very different semantic labels but are statistically close to the ID data due to shared environmental features (*e.g.*, boat vs. waterbird in Figure 1). While other works have considered domain shift (Hsu et al. 2020) or covariate shift (Ovadia et al. 2019), they are more relevant for evaluating model generalization—in which case the goal is to make the model classify accurately into the ID classes and should not be confused with OOD detection. We emphasize that semantic label shift (i.e., change of invariant feature) is more akin to OOD detection, where the inputs have disjoint labels from ID data and therefore should not be predicted by the model.

Out-of-Distribution Generalization. Recently, various works have been proposed to tackle the issue of domain generalization, which aims to achieve high classification accuracy on new test environments consisting of inputs *with invariant features*, and does not consider the change of invariant features at test time (i.e., label space $\mathcal{Y}$ remains the same)—a key difference from our focus. Literature in OOD detection is commonly concerned about model reliability and detection of shifts where the OOD inputs have disjoint labels and therefore should not be predicted by the model. In other words, we consider samples *without invariant features*, regardless of the presence of environmental features or not.

A plethora of algorithms are proposed: learning invariant representation across domains (Ganin et al. 2016; Li et al. 2018b; Sun and Saenko 2016; Li et al. 2018a), minimizing the weighted combination of risks from training domains (Sagawa et al. 2019), using different risk penalty terms to facilitate invariance prediction (Arjovsky et al. 2019; Krueger et al. 2020), causal inference approaches (Peters, Bühlmann, and Meinshausen 2016), and forcing the learned representation different from a set of pre-defined biased representations (Bahng et al. 2020), mixup-based approaches (Zhang et al. 2018; Wang, Li, and Kot 2020; Luo, Song, and Zhang 2020), etc. A recent study (Gulrajani and Lopez-Paz 2021) shows that no domain generalization methods achieve superior performance than ERM across a broad range of datasets.

Contextual Bias in Recognition. There has been a rich literature studying the classification performance in the presence of contextual bias (Torralba 2003; Beery, Van Horn, and Perona 2018; Barbu et al. 2019). The reliance on contextual bias such as image backgrounds, texture, and color for object detection are investigated in (Zhu, Xie, and Yuille 2017; Baker et al. 2018; Geirhos et al. 2019; Zech et al. 2018; Xiao et al. 2021; Sagawa et al. 2019). However, the contextual bias for OOD detection is underexplored. In contrast, our study systematically investigates the impact of spurious correlation on OOD detection and how to mitigate it.

Conclusion

Out-of-distribution detection is an essential task in open-world machine learning. However, the precise definition is often left in vagueness, and common evaluation schemes can be too primitive to capture the nuances of the problem in reality. In this paper, we present a new formalization where we model the data distributional shifts by considering the invariant and non-invariant features. Under such formalization, we systematically investigate the impact of spurious correlation in the training set on OOD detection and further show insights on detection methods that are more effective in mitigating the impact of spurious correlation. Moreover, we provide theoretical analysis on why reliance on environmental features leads to high OOD detection error. We hope that our work will inspire future research on the understanding and formalization of OOD samples, new evaluation schemes of OOD detection methods, and algorithmic solutions in the presence of spurious correlation.

Acknowledgements

This work is supported by the Office of the Vice Chancellor for Research and Graduate Education (OVCRGE) with funding from the Wisconsin Alumni Research Foundation (WARF).

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bahng, H.; Chun, S.; Yun, S.; Choo, J.; and Oh, S. J. 2020. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, 528–539. PMLR.

Baker, N.; Lu, H.; Erlikhman, G.; and Kellman, P. J. 2018. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12): 1–43.

Barbu, A.; Mayo, D.; Alverio, J.; Luo, W.; Wang, C.; Gutfreund, D.; Tenenbaum, J.; and Katz, B. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32: 9453–9463.

Beery, S.; Van Horn, G.; and Perona, P. 2018. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 456–473.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *Proceedings of International Conference on Learning Representations*.

Hsu, Y.-C.; Shen, Y.; Jin, H.; and Kira, Z. 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10951–10960.

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-distribution Detection for Large Semantic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 10215–10224.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why Normalizing Flows Fail to Detect Out-of-Distribution Data. *Advances in Neural Information Processing Systems*, 33.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. 2020. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 7167–7177.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018a. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5400–5409.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018b. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 624–639.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations, ICLR*.

Lin, Z.; Roy, S. D.; and Li, Y. 2021. MOOD: Multi-level Out-of-distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. *Advances in Neural Information Processing Systems*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Luo, C.; Song, C.; and Zhang, Z. 2020. Generalizing Person Re-Identification by Camera-Aware Invariance Learning and Cross-Domain Mixup. In *European Conference on Computer Vision*.

McInnes, L.; Healy, J.; Saul, N.; and Grossberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, 3(29): 861.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2019. Do Deep Generative Models Know What They Don't Know? In *International Conference on Learning Representations*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for

unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32: 13991–14002.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 947–1012.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 14680–14691.

Rosenfeld, E.; Ravikumar, P. K.; and Risteski, A. 2021. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*.

Roy, A. G.; Ren, J.; Azizi, S.; Loh, A.; Natarajan, V.; Mustafa, B.; Pawlowski, N.; Freyberg, J.; Liu, Y.; Beaver, Z.; et al. 2021. Does Your Dermatology Classifier Know What It Doesn't Know? Detecting the Long-Tail of Unseen Conditions. *arXiv preprint arXiv:2104.03829*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations, ICLR*.

Sastry, C. S.; and Oore, S. 2020. Detecting Out-of-Distribution Examples with In-distribution Examples and Gram Matrices. In *Proceedings of the 37th International Conference on Machine Learning*.

Serrà, J.; Álvarez, D.; Gómez, V.; Slizovskaia, O.; Núñez, J. F.; and Luque, J. 2020. Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models. In *International Conference on Learning Representations*.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, 443–450. Springer.

Torralba, A. 2003. Contextual priming for object detection. *International journal of computer vision*, 53(2): 169–191.

Wang, Y.; Li, H.; and Kot, A. C. 2020. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3622–3626. IEEE.

Winkens, J.; Bunel, R.; Roy, A. G.; Stanforth, R.; Natarajan, V.; Ledsam, J. R.; MacWilliams, P.; Kohli, P.; Karthikesalingam, A.; Kohl, S.; et al. 2020. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*.

Xiao, K. Y.; Engstrom, L.; Ilyas, A.; and Madry, A. 2021. Noise or Signal: The Role of Image Backgrounds in Object Recognition. In *International Conference on Learning Representations*.

Xiao, Z.; Yan, Q.; and Amit, Y. 2020. Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder. *Advances in Neural Information Processing Systems*, 33.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zech, J. R.; Badgeley, M. A.; Liu, M.; Costa, A. B.; Titano, J. J.; and Oermann, E. K. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11).

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.

Zhu, Z.; Xie, L.; and Yuille, A. 2017. Object Recognition with and without Objects. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 3609–3615.