Robustness to Spurious Correlations Improves Semantic Out-of-Distribution Detection

Lily H. Zhang, Rajesh Ranganath
New York University
60 5th Avenue
New York, NY 10011
lily.h.zhang@nyu.edu, rajeshr@cims.nyu.edu

Abstract

Methods which utilize the outputs or feature representations of predictive models have emerged as promising approaches for out-of-distribution (OOD) detection of image inputs. However, these methods struggle to detect OOD inputs that share nuisance values (e.g. background) with in-distribution inputs. The detection of shared-nuisance out-of-distribution (SN-OOD) inputs is particularly relevant in real-world applications, as anomalies and in-distribution inputs tend to be captured in the same settings during deployment. In this work, we provide a possible explanation for SN-OOD detection failures and propose *nuisance-aware* OOD detection to address them. Nuisance-aware OOD detection substitutes a classifier trained via Empirical Risk Minimization (ERM) and cross-entropy loss with one that 1. is trained under a distribution where the nuisance-label relationship is broken and 2. yields representations that are independent of the nuisance under this distribution, both marginally and conditioned on the label. We can train a classifier to achieve these objectives using Nuisance-Randomized Distillation (NURD), an algorithm developed for OOD generalization under spurious correlations. Output- and feature-based nuisance-aware OOD detection perform substantially better than their original counterparts, succeeding even when detection based on domain generalization algorithms fails to improve performance.

1 Introduction

Out-of-distribution (OOD) detection is the task of identifying inputs that fall outside the training distribution. A natural approach is to estimate the training distribution via a generative model and flag low-density inputs as OOD (Bishop 1994), but such an approach has been shown to perform worse than random chance on several image tasks (Nalisnick et al. 2019), likely due to model estimation error (Zhang, Goldstein, and Ranganath 2021). Instead, many detection methods utilize either the outputs or feature representations of a learned classifier, yielding results much better than those of deep generative models on many tasks (Salehi et al. 2021)

However, classifier-based OOD detection has been shown to struggle when OOD and in-distribution (ID) inputs share the same values of a *nuisance variable* that is of no inherent interest to the semantic task, e.g. the background of an image (Ming, Yin, and Li 2022). We call such OOD inputs *shared-nuisance* OOD (SN-OOD) examples. For example, in the Waterbirds dataset (Sagawa et al. 2020), the image background (water or land) is a nuisance in the task of classifying bird type (waterbird vs. landbird), and an image of a boat on water is an SN-OOD input, given the familiar background nuisance value but novel object label. Detection of SN-OOD images is worse than detection of OOD images with novel nuisance values. Moreover, the stronger the correlation between the nuisance and label in the training distribution, the worse the detection of SN-OOD inputs. Even when classifiers are trained via domain generalization algorithms intended for generalizing to new test domains, detection does not improve (Ming, Yin, and Li 2022).

This failure mode is far from a rare edge case, given the relevance of SN-OOD inputs in real-world applications. While an instance can be OOD with respect to labels (e.g. new object) or nuisances (e.g. new background), in most cases, the goal is *semantic* OOD detection, or detecting out-of-scope inputs (Yang et al. 2021). For instance, manufacturing plants are interested in product defects on the factory floor, not working products in new settings.

In this work, we introduce *nuisance-aware* OOD detection to address SN-OOD detection failures detection. Our contributions:

1. We present explanations for output- and feature-based SN-OOD detection failures based on the role of nuisances in the learned predictor and its representations (Section 3).
2. We illustrate why models that are robust to spurious correlations can yield better output-based SN-OOD detection and identify a predictor with such robustness guarantees to use for OOD detection (Section 4).
3. We explain why removing nuisance information from representations can improve feature-based SN-OOD detection and propose a joint independence constraint to achieve this end (Section 4).
4. We describe how domain generalization algorithms can fail to improve SN-OOD detection, providing insight into the empirical failures seen previously (Section 5).
5. We show empirically that nuisance-aware OOD detection improves detection on SN-OOD inputs while maintaining performance on non-SN-OOD ones (Section 7).

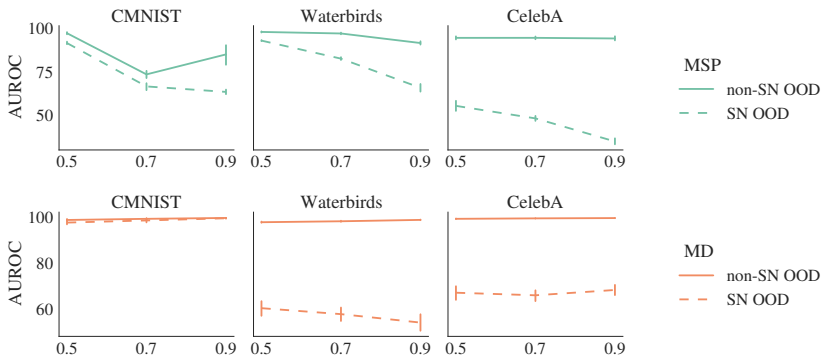


Figure 1: Output- and feature-based detection methods perform worse on shared-nuisance out-of-distribution inputs (see Section 7 for task details). The performance of output-based shared-nuisance OOD detection (MSP, top) degrades under increasing spurious correlation strength (x-axis). Figure 7 shows the same trend for other output-based methods. The performance of feature-based OOD detection (MD, bottom) is more stable but poor. See Appendix for plots of other common OOD detection methods, which follow a similar trend.

2 Background

Most methods which employ predictive models for out-of-distribution detection can be categorized as either output-based or feature-based. Output-based methods utilize some function of the logits as an anomaly, while feature-based methods utilize internal representations of the learned model.

Output-based Out-of-distribution Detection

Let $f : \mathbb{R}^D \to \mathbb{R}^K$ be the learned function mapping an D-dimensional input to K logits for the K ID classes. Output-based methods utilize some function of the logits $f(\mathbf{x}) \in \mathbb{R}^K$ as an anomaly score. Letting σ denote the softmax function, relevant methods include maximum softmax probability (MSP) or $\max_k \sigma(f(\mathbf{x}))_k$ (Hendrycks and Gimpel 2017), max logit or $\max_k f_k(\mathbf{x})$ (Hendrycks et al. 2022), the energy score or $-\log \sum_k \exp f_k(\mathbf{x})$ (Liu et al. 2020), and out-of-distribution detector for neural networks (ODIN) or $\max_k \sigma(f_k(\widetilde{\mathbf{x}})/T)$, where T is a learned temperature parameter and $\widetilde{\mathbf{x}}$ is an input perturbed in the direction of the gradient of the maximum softmax probability (Liang, Li, and Srikant 2018).

Feature-based Out-of-distribution Detection

Feature-based methods utilize internal representations of the learned classifier. Following prior work (Kamoi and Kobayashi 2020; Fort, Ren, and Lakshminarayanan 2021), we consider the penultimate feature activations $r(\mathbf{x})$. The most widely used feature-based method is the Mahalanobis distance (MD) (Lee et al. 2018), which models the feature representations of in-distribution data as class-conditional Gaussians with means μ_k and shared covariance Σ. At test time, the anomaly score is the minimum Mahalanobis distance from a new input's feature representations to each of these class distributions, $\min_k \sqrt{(r(\mathbf{x}) - \mu_k)\Sigma^{-1}(r(\mathbf{x}) - \mu_k)^\top} = \max_k p(r(\mathbf{x})|\mathbf{y} = k)$. Assuming minimal overlap in probability mass across each class-conditional Gaussian (e.g. tight and separated clusters), this method can approximate detection based on density estimation on the representations: $\max_k p(r(\mathbf{x})|\mathbf{y} = k) \propto \max_k p(r(\mathbf{x})|\mathbf{y} = k)p(\mathbf{y} = k) \approx \sum_k p(r(\mathbf{x})|\mathbf{y} = k)p(\mathbf{y} = k) = p(r(\mathbf{x}))$.

Failures in Shared-Nuisance OOD Detection

Ming, Yin, and Li (2022) find that output- and feature-based detection show worse performance on shared-nuisance OOD inputs than non-SN-OOD inputs, and that the performance of output-based methods degrades as the strength of the correlation between nuisance and label in the training data increases. Figure 1 corroborates and extends their findings, illustrating that across several datasets, performance is generally worse on shared nuisance inputs. Output-based detection of such inputs degrades under stronger spurious correlations and is sometimes comparable to or worse than random chance (AUROC < 50). Feature-based detection tends to be more stable across varying spurious correlations but can perform worse than output-based detection even under strong spurious correlations (e.g. Waterbirds). Our absolute numbers differ from those of Ming, Yin, and Li (2022) for the following reasons: First, our CelebA results are based on the blond/non-blond in-distribution task in (Sagawa et al. 2020) rather than gray/non-gray. Next, the Waterbirds results are sensitive data generation seed (see Figure 10 in Appendix for details). Finally, our feature-based results use the penultimate activations while Ming, Yin, and Li (2022) aggregate features over all layers, which requires additional validation OOD data. Even so, our results show the same trends.

3 Understanding Shared-Nuisance Out-of-Distribution Detection Failures

To understand SN-OOD failures, first note that the success of a classifier for detection depends on the classifier assigning different feature representations or outputs to OOD and ID inputs. Poor detection of SN-OOD inputs relative to non-SN-OOD inputs implies that the classifier fails to map SN-OOD inputs to outputs or representations sufficiently distinct from those expected of ID inputs. Given that both SN-OOD and

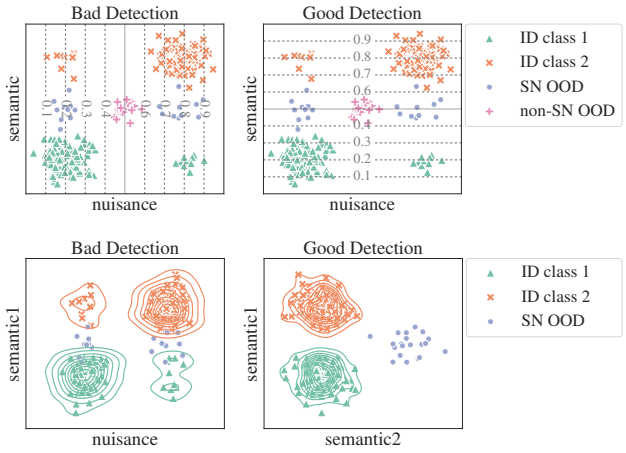


Figure 2: Top: Output-based OOD detection will perform poorly on shared-nuisance out-of-distribution inputs when the prediction output relies on nuisance (left) rather than semantics (right), even if non-SN-OOD inputs can be detected well in either case. Bottom: Feature-based OOD detection is easier when representations focus on semantics (right) rather than both semantics and nuisance (left). Nuisance-aware OOD detection encourages the top and bottom right scenarios to improve SN-OOD detection.

non-SN-OOD inputs have unseen semantics, we hypothesize that the difference in performance can be largely explained by how the learned classifier utilizes nuisances. We walk through an output-based and feature-based example below.

To establish notation, let $\mathbf{z}$ and $\mathbf{y}$ be the nuisance and label which together generate input $\mathbf{x}$. Let $\mathcal{Z}_{tr}, \mathcal{Y}_{tr}$ be the values of $\mathbf{z}$ and $\mathbf{y}$ which appear during training. In semantic OOD detection, an input is out-of-distribution if its semantic label was not seen in the training data, i.e. $\mathbf{y} \notin \mathcal{Y}_{tr}$. Then, the difference between shared-nuisance out-of-distribution inputs and non-shared-nuisance out-of-distribution inputs lies in $\mathbf{z}$: SN-OOD inputs have nuisances $\mathbf{z} \in \mathcal{Z}_{tr}$, while non-SN-OOD inputs do not. For instance, for the Waterbirds dataset, a non-bird image over a water background would be a shared-nuisance OOD input, while a non-bird image taken indoors would be a non-SN-OOD input.

Explaining Poor MSP Performance. Perfect MSP performance requires that all OOD inputs have less confident output probabilities than ID inputs. Worse detection on SN-OOD over non-SN-OOD inputs suggests that the former get more peaked confidences in their outputs, making them more similar to ID inputs. Since the difference between SN-OOD and non-SN-OOD inputs is whether $\mathbf{z} \in \mathcal{Z}_{tr}$, this worse performance can be attributed to the model's behavior on $\mathbf{z} \in \mathcal{Z}_{tr}$ vs. $\mathbf{z} \notin \mathcal{Z}_{tr}$. Poor SN-OOD results suggest that predictive models assign peaked output probabilities to inputs where $\mathbf{z} \in \mathcal{Z}_{tr}$ even if $\mathbf{y} \notin \mathcal{Y}_{tr}$. Such a phenomenon is possible if the learned function f is primarily a function of the nuisance, e.g. $f(\mathbf{x}) \approx g(\mathbf{z})$ (Figure 2, top left). Then, SN-OOD and ID inputs would yield similar outputs, whereas non-SN-OOD inputs could yield different outputs and still be detected.

Explaining Poor Mahalanobis Distance Performance. Mahalanobis distance performs well when OOD inputs have representations that are sufficiently different from ID ones, enough so to be assigned low density under a model estimating ID representations via class-conditional Gaussians. Detection is worse for SN-OOD inputs than non-SN-OOD ones, suggesting that OOD representations are assigned higher density when inputs have nuisance values $\mathbf{z} \in \mathcal{Z}_{tr}$. Such a scenario can only occur if nuisance information is present in the representations; otherwise, detection performance of SN-OOD and non-SN-OOD inputs should be similar, assuming their semantics are similarly different from ID semantics.

It is worth noting that, if all the semantic information needed to distinguish ID and OOD is present in the representations, then a detection method based on perfect density estimation of the ID representation distribution would successfully detect OOD inputs, regardless of whether the representations additionally contain nuisance information or not. However, in the absence of perfect estimation, representations with more dimensions related to nuisance can be more sensitive to estimation error since they have fewer dimensions dedicated to semantics where ID and SN-OOD are non-overlapping; for instance, when representations only differ over one dimension, accurate detection requires very accurate modeling of the single relevant dimension, unknown *a priori* (Figure 2, bottom left). In contrast, representations where ID and SN-OOD inputs differ over more dimensions are more robust to misestimation over any one dimension (Figure 2, bottom right).

4 Nuisance-Aware OOD Detection

We summarize above observations and explanations below:

1. **Observation:** Output-based OOD detection is worse on SN-OOD inputs and degrades with increasing correlation between the nuisance and label in the training distribution.

   Explanation: The learned predictor adjusts its output based on the nuisance, particularly when there is a strong spurious correlation between nuisance and label in the training data. The result is that SN-OOD outputs look like ID outputs.

2. **Observation:** Feature-based OOD detection is worse on SN-OOD inputs even though its performance is fairly stable across different correlations.

   Explanation: Regardless of the nuisance-label correlation, the learned representations contain information about the nuisance in addition to semantics, making SN-OOD representations look more similar to ID ones.

To address both issues, we propose *nuisance-aware* out-of-distribution detection, utilizing knowledge of nuisances to improve detection. To improve output-based detection, we substitute a classifier trained via empirical risk minimization with one that is robust to spurious correlations, defined by good classification performance on all distributions that differ from the training distribution in nuisance-label relationship. To improve feature-based detection, we train a classifier such that its penultimate representation cannot predict the

nuisance by itself or conditioned on the label. We motivate and describe our approach below.

Addressing Spurious Correlations via Reweighting

To improve output-based SN-OOD detection, we recall that poor output-based SN-OOD detection can occur when the learned function f can be approximated by a function of only the nuisance, i.e. $f(\mathbf{x}) \approx g(\mathbf{z})$. Is there a way to avoid learning such functions given only in-distribution data?

First, if a predictor behaves like a function of the nuisance in order to predict the label well on a given data distribution, then it can perform arbitrarily poorly on a new distribution where the relationship between the nuisance and label has changed. Given a data distribution p_D, let $\mathcal{F}$ be a family of distributions that differ from p_D only in the nuisance-label relationship: $\mathcal{F} = \{p_{D'} = p_D(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) p_{D'}(\mathbf{z} \mid \mathbf{y}) p_D(\mathbf{y})\}$, $\text{supp}(p_{D'}(\mathbf{z} \mid \mathbf{y})) = \text{supp}(p_D(\mathbf{z} \mid \mathbf{y}))$ for all $\mathbf{y} \in \text{supp}(p_D(\mathbf{y}))$. We call the nuisance-label relationship a spurious correlation. A predictor that performs well across all distributions in $\mathcal{F}$, i.e. is robust to the spurious correlation between nuisance and label, cannot rely on a function of only nuisance to make its prediction and thus is more likely to succeed at SN-OOD detection. In other words, *models that are robust to spurious correlations are also likely to be better for output-based* SN-OOD *detection.*

Theoretical Motivation. We propose to improve output-based detection by training models that are robust to spurious correlations. Let $p_{\perp\!\!\!\perp}$ be a distribution in $\mathcal{F}$ where the label is independent of the nuisance: $\mathbf{z} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{y}$. Puli et al. (2022a) prove that for all representation functions $r \in \mathcal{R}$ such that $\mathbf{z} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{y} | r(\mathbf{x})$, the predictor $p_{\perp\!\!\!\perp}(\mathbf{y} \mid r(\mathbf{x}))$ is guaranteed to perform as well as marginal label prediction on any distribution in $\mathcal{F}$, a guarantee that does not always hold for other predictors. Moreover, when the identity function is in $\mathcal{R}$, i.e. $\mathbf{z} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{y} | \mathbf{x}$, then the predictor $p_{\perp\!\!\!\perp}(\mathbf{y} \mid \mathbf{x})$ yields simultaneously optimal performance across all distributions in $\mathcal{F}$ relative to any representation $r \in \mathcal{R}$ and is minimax optimal for a sufficiently diverse $\mathcal{F}$ among all predictors $p_D(\mathbf{y} \mid \mathbf{x})$. In other words, $p_{\perp\!\!\!\perp}(\mathbf{y} \mid \mathbf{x})$ enjoys substantial performance guarantees when $\mathbf{z} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{y} | \mathbf{x}$.

When the input $\mathbf{x}$ determines the nuisance $\mathbf{z}$ (e.g., looking at an image tells you its background), then $\mathbf{z} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{y} \mid \mathbf{x}$ holds trivially. Consequently, $p_{\perp\!\!\!\perp}(\mathbf{y} \mid \mathbf{x})$ has the robustness guarantees summarized above.

Method: Reweighting. We can estimate $p_{\perp\!\!\!\perp}(\mathbf{y} \mid \mathbf{x})$ by reweighting: given $p_D(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_D(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) p_D(\mathbf{y} \mid \mathbf{z}) p_D(\mathbf{z})$, we can construct $p_{\perp\!\!\!\perp}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_D(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) p_D(\mathbf{y}) p_D(\mathbf{z})$ as follows (Puli et al. 2022a):

$$p_{\perp\!\!\!\perp}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_D(\mathbf{x}, \mathbf{y}, \mathbf{z}) \frac{p_D(\mathbf{y})}{p_D(\mathbf{y} \mid \mathbf{z})}. \tag{1}$$

We reweight both the training and validation loss using Equation (1) and perform model selection based on the best reweighted validation loss.

Reweighting when Group Labels are Unavailable. When nuisance values are present as metadata, reweighting based on Equation (1) is straightforward. When we do not have access to exact group labels, following Puli et al. (2022b), we can use functions of the input as nuisance values, e.g. via masking. For instance, for images with centered objects, the outer border of the image can be used as a proxy for background. Then, a well-calibrated classifier predicting the label from the masked input can approximate $p_D(\mathbf{y} \mid \mathbf{z})$.

Addressing Nuisance Features via Independence Constraints

Can we improve feature-based methods on SN-OOD inputs, even if they are stable in performance regardless of spurious correlation strength? We hypothesize that removing nuisance information from the learned representations makes ID and SN-OOD inputs easier to distinguish. First, shared nuisance information is not helpful for distinguishing ID and SN-OOD inputs by definition, so removing it should not hurt detection performance. Furthermore, when representations contain this information, SN-OOD inputs can more easily go undetected, e.g. by looking like ID inputs over more of the principal components of variation in the representation. More generally, nuisance information can be additional modeling burden for a downstream feature-based method by introducing additional entropy; for instance, given a discrete representation $r(\mathbf{x})$ that is independent of a discrete nuisance, a representation $r'(\mathbf{x})$ which additionally includes nuisance information (i.e. $p(r'(\mathbf{x})) = p(r(\mathbf{x}), g(\mathbf{z}))$) has strictly higher entropy H: $\text{H}(r'(\mathbf{x})) = \text{H}(r(\mathbf{x})) + \text{H}(g(\mathbf{z}) \mid r(\mathbf{x})) > \text{H}(r(\mathbf{x}))$.

Theoretical Motivation. To remove nuisance information, we propose enforcing $r(\mathbf{x}) \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}$ and $r(\mathbf{x}) \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z} \mid \mathbf{y}$. The former ensures that the representations cannot predict nuisance on their own, while the latter ensures that within each label class, the representations do not provide information about the nuisance. Without the latter condition, the representations can be a function of nuisance within the ID classes such that marginal independence is enforced but SN-OOD representations overlap with ID ones (see the Appendix A for an example). To avoid this situation and encourage disjoint representations, we enforce marginal and conditional independence, equivalent to joint independence $\mathbf{z} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{y}, r(\mathbf{x})$.

Method: Joint Independence. To enforce joint independence, we penalize the estimated mutual information between the nuisance $\mathbf{z}$ and the combined representations $r(\mathbf{x})$ and label $\mathbf{y}$. When $\mathbf{z}$ is high-dimensional, e.g. a masked image, we estimate the mutual information via the density-ratio estimation trick (Sugiyama, Suzuki, and Kanamori 2012), following Puli et al. (2022a). Concretely, we use a binary classifier distinguishing between samples from $p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}, \mathbf{z})$ and $p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}) p_{\perp\!\!\!\perp}(\mathbf{z})$ to estimate the ratio $\frac{p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}, \mathbf{z})}{p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}) p_{\perp\!\!\!\perp}(\mathbf{z})}$. When $\mathbf{z}$ is low-dimensional, we estimate the mutual information by training a model to predict $\mathbf{z}$ from $r(\mathbf{x}), \mathbf{y}$ under the reweighted distribution $p_{\perp\!\!\!\perp}$:

$$\text{I}_{p_{\perp\!\!\!\perp}}(\mathbf{z}; r(\mathbf{x}), \mathbf{y}) \tag{2}$$

$$= \int p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}, \mathbf{z}) \log \frac{p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}, \mathbf{z})}{p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}) p_{\perp\!\!\!\perp}(\mathbf{z})} d\mathbf{x} d\mathbf{y} d\mathbf{z} \tag{3}$$

$$= \mathbb{E}_{p_{\perp\!\!\!\perp}} \Big[\log \frac{p_{\perp\!\!\!\perp}(\mathbf{z} \mid r(\mathbf{x}), \mathbf{y})}{p_{\perp\!\!\!\perp}(\mathbf{z})} \Big]. \tag{4}$$

Other neural network-based mutual information estimators can also be used (Belghazi et al. 2018; Poole et al. 2019). Zero mutual information implies independence; otherwise, there is still dependence, and we add the mutual information as a penalty to the loss when training the main classifier.

Why Naive Independence Doesn't Work. Why must we ensure independence of $r(\mathbf{x})$ and $\mathbf{z}$ under $p_{\perp\!\!\!\perp}$ instead of under the original training distribution p_D? In cases where the nuisance and label are strongly correlated, forcing independence of the penultimate representations and the nuisance will force the representation to ignore information that is predictive of the label, simply because it is also predictive of the nuisance. At one extreme, if the nuisance and label are nearly perfectly correlated under p_D, then $r(\mathbf{x}) \perp\!\!\!\perp_{p_D} \mathbf{z}$ will force $r(\mathbf{x})$ to contain almost no information which could predict $\mathbf{y}$. This situation is avoided when label and nuisance are independent.

Summarizing Nuisance-Aware OOD Detection

We propose the following for nuisance-aware OOD detection:

1. When performing output-based detection, train a classifier with reweighting: $\mathcal{L}_{p_{\perp\!\!\!\perp}}(f(\mathbf{x}); \mathbf{y})$.
2. When performing feature-based detection, train a classifier with reweighting and a joint independence penalty: $\mathcal{L}_{p_{\perp\!\!\!\perp}}(f(\mathbf{x}); \mathbf{y}) + \lambda \mathrm{I}_{p_{\perp\!\!\!\perp}}(\mathbf{z}; r(\mathbf{x}), \mathbf{y})$.

Reweighting is performed based on Equation (1) by estimating $p_D(\mathbf{y})$ and $p_D(\mathbf{y} \mid \mathbf{z})$ from the data. When $\mathbf{z}$ is a discrete label, both terms can be estimated by counting the sizes of groups defined by $\mathbf{y}$ and $\mathbf{z}$. When $\mathbf{z}$ is continuous or high-dimensional, as in a masked image, an additional reweighting model $p_D(\mathbf{y} \mid \mathbf{z})$ can be estimated prior to training the main model.

To implement the joint independence penalty, Equation (2) is estimated after every gradient step of the classifier. As described in the above section, one can fit a critic model $p_{\perp\!\!\!\perp}(\mathbf{z} \mid r(\mathbf{x}), \mathbf{y})$ when $\mathbf{z}$ is low-dimensional or employ the density ratio trick when $\mathbf{z}$ is high-dimensional. We re-estimate the critic model after each step of the main model training.

5 Why Domain Generalization Methods Fail

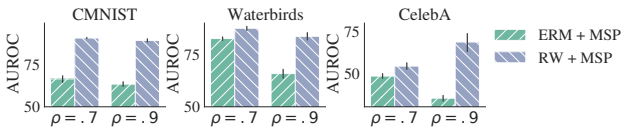
Domain generalization algorithms utilize multiple training environments in order to generalize to new test environments. Using knowledge of nuisance to create environments, Ming, Yin, and Li (2022) do not see improved OOD detection when training a classifier using domain generalization algorithms implemented in Gulrajani and Lopez-Paz (2021). Here, we explain how, despite taking into account nuisance information, these algorithms can fail to improve output-based SN-OOD detection because the resulting models can fail to be robust to spurious correlations.

First, algorithms that rely on enforcing constraints across multiple environments (e.g. Invariant Risk Minimization (IRM) (Arjovsky et al. 2019), Risk Extrapolation (REX) (Krueger et al. 2021)) can fail to achieve robustness to spurious correlations if the environments do not have common support, which is typically the case if environments are denoted by nuisance values. In such scenarios, predicting well in one environment does not restrict the model in another environment. The set up in Ming, Yin, and Li (2022) is an example of this scenario, which has also been discussed in Guo et al. (2021) for IRM. Instead, to address spurious correlations, these algorithms rely on environments defined by different nuisance-label relationships. However, even then, an IRM solution can still fail to generalize to relevant test environments if the training environments available are not sufficiently diverse, numerous, and overlapping (Rosenfeld, Ravikumar, and Risteski 2021). In contrast, nuisance-aware OOD detection enables robustness across the family of distributions $\mathcal{F}$ given only one member in the family. Multi-environment objectives can also struggle with difficulty in optimization (Zhang, Lopez-Paz, and Bottou 2022), an additional challenge beyond that of training data availability.

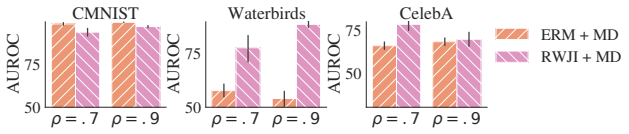
Next, algorithms that enforce invariant representations (e.g. Domain-Adversarial Neural Networks (DANN) (Ganin et al. 2016), Conditional Domain Adversarial Neural Networks (CDANN) (Li et al. 2018)) can hurt performance if the constraints imposed are too stringent. For instance, when environments are denoted by nuisance, the constraint posed by DANN is equivalent to the naive independence constraint $\mathbf{z} \perp\!\!\!\perp r(\mathbf{x})$ discussed earlier, which can actively remove semantic information helpful for prediction just because it is correlated with $\mathbf{z}$. CDANN enforces the distribution of representations to be invariant across environments conditional on the class label. In Appendix B, we show that a predictor based on the CDANN constraint is worse than the nuisance-aware predictor for certain distribution families $\mathcal{F}$. For such $\mathcal{F}$, a CDANN predictor is less robust to spurious correlations as measured by performance across the distribution family. Such a predictor will also be worse for SN-OOD detection that relies on strong performance over all regions of the in-distribution support in order for ID and SN-OOD outputs to be as distinct as possible.

Group Distributionally Robust Optimization (GDRO) (Sagawa et al. 2020) aims to minimize the worst-case loss over some uncertainty set of distributions, defined by mixtures of pre-defined groups. When the label and nuisance combined define the groups, their mixtures can be seen as distributions with varying spurious correlation strengths. However, other ways of defining groups do not yield the same interpretation; for instance, when the groups are defined by nuisance only,[1] the uncertainty set of distributions is one of varying mixtures of nuisance values while the nuisance-label relationship $p(\mathbf{y} \mid \mathbf{z})$ is held fixed. In other words, such a set up does not address robustness across varying spurious correlations. Even under an appropriate set up for robustness to spurious correlations, GDRO may still not reach optimal performance. Concretely, as the loss only depends on the outputs of the worst-performing group, GDRO does not favor solutions that continue to optimize performance on other groups if the loss for the worst-performing group is at an local optimum. As intuition, an example where such a situation could occur is one where the best losses possible across groups differs drastically. For an output-based detection method that relies on in-distribution outputs being as large or peaked as possible in order to separate them from OOD ones, GDRO's

[1] This is enforced by the code in Gulrajani and Lopez-Paz (2021) which requires that all classes are present in each environment.



(a) Output-based Detection (MSP)



(b) Feature-based Detection (MD)

Figure 3: Reweighting (RW) improves output-based SN-OOD detection (top), while reweighting and joint independence (RWJI) generally improves feature-based SN-OOD detection (bottom). ρ is the strength of the spurious correlation. Results on other methods follow this trend (Appendix Figure 9).

consideration of only worst group performance can hinder detection performance, especially relative to the proposed method which considers all inputs in the reweighting.

6 Related Work

Our work is most closely related to Ming, Yin, and Li (2022) and Puli et al. (2022a). Ming, Yin, and Li (2022) first notice that OOD detection is harder on shared-nuisance OOD inputs; we expand on their analysis and provide a solution that makes progress on the issue where other approaches (e.g. domain generalization algorithms) do not. Our proposed solution for improving feature-based detection with high-dimensional nuisances, i.e. reweighting and enforcing joint independence, matches the algorithm proposed in Puli et al. (2022a) for OOD generalization, though the motivations for each component of the solution are different. Our joint independence algorithm is different for low-dimensional nuisances.

Fort, Ren, and Lakshminarayanan (2021) also demonstrate that classifier choice matters for detection, but they focus on large pretrained transformers, while we consider models that utilize domain knowledge of nuisances. The idea of removing non-discriminative features in the representations of ID and OOD inputs has also been explored in Kamoi and Kobayashi (2020), who modify the Mahalanobis distance to consider only a subset of the eigenvectors of the computed covariance matrix. This partial Mahalanobis score focuses on non-SN-OOD benchmarks and chooses principal components based on explained variation, whereas we remove shared-nuisance information to address SN-OOD failures. More broadly, nuisance-aware OOD detection changes the classifier used for detection and can be used alongside other detection methods which employ classifiers.

7 Experiments

We consider the following three tasks:

1. **CMNIST.** $\mathcal{Y}_{tr} = \{0, 1\}$, $\mathcal{Z}_{tr} = \{\text{red}, \text{green}\}$, $\mathcal{Y}_{\text{SN-OOD}} = \{2, ..., 10\}$.
2. **Waterbirds.** $\mathcal{Y}_{tr} = \{\text{waterbird}, \text{landbird}\}$, $\mathcal{Z}_{tr} = \{\text{water}, \text{land}\}$, $\mathcal{Y}_{\text{SN-OOD}} = \{\text{no bird}\}$.
3. **CelebA.** $\mathcal{Y}_{tr} = \{\text{blond}, \text{not blond}\}$, $\mathcal{Z}_{tr} = \{\text{male}, \text{female}\}$, $\mathcal{Y}_{\text{SN-OOD}} = \{\text{bald}\}$.

The open-source datasets from which these tasks are derived all have nuisance metadata available for all examples (e.g. environment label for Waterbirds, attributes for CelebA (Liu et al. 2015)) as well as a way to identify or construct SN-OOD examples, enabling us to control both the strength of the spurious correlation in the training set (we consider $\rho \in \{0.7, 0.9\}$) while also testing OOD detection. We ensure that the dataset sizes are the same across ρ. For non-SN-OOD datasets, we use blue MNIST digits (Deng 2012) for CMNIST and SVHN (Netzer et al. 2011) for Waterbirds and CelebA. For model architecture, we utilize a 4-layer convolutional network for CMNIST and ResNet-18 pretrained on ImageNet for Waterbirds and CelebA. Unless otherwise noted, we average all results over 5 random seeds, and error bars denote one standard error. We use AUROC as the metric for all detection results following previous literature (Hendrycks and Gimpel 2017). Code is available at https://github.com/rajesh-lab/nuisance-aware-ood-detection. See Appendix for more details.

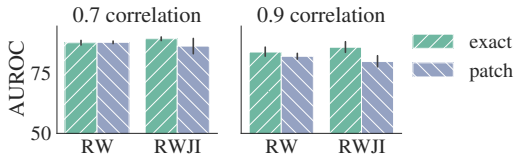
Main Results. Reweighting substantially improves output-based SN-OOD detection (Table 2 in Appendix), providing empirical evidence that increased robustness to spurious correlations (estimated by performance on a new distribution $p_{D'} \neq p_D$) correlates with improved output-based detection. Reweighting plus joint independence improves feature-based detection with statistically significant results, with the exception of CMNIST, likely due to the task construction where digit is only partially predictive of class (see Appendix D for details), and CelebA at $\rho = 0.9$, where joint independence results have high variance. Reversing training strategies and anomaly methods does not the same consistent positive benefit, highlighting the importance selecting the right nuisance-aware strategy for a given detection method (see Figure 8).

While Ming, Yin, and Li (2022) do not see success in combining nuisance information with domain generalization algorithms over their baseline ERM solution, Table 1 shows that nuisance-aware OOD detection succeeds over nuisance-unaware ERM. On non-SN-OOD inputs, reweighting yields consistent or better output-based detection performance, and reweighting plus joint independence generally yields comparable or better feature-based detection (Table 3 in Appendix).

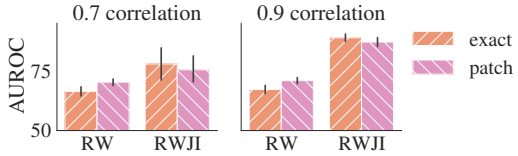
Exact vs. User-generated Nuisances. Figure 4 shows that that OOD detection and balanced ID classification results are comparable for Waterbirds whether the nuisance is the exact metadata label or denoted by the outer border of the image. These results illustrate the applicability of nuisance-aware OOD detection even when nuisance values are not provided as metadata. Other creative specifications of $\mathbf{z}$ are possible, e.g. an image with shuffled pixels if low-level pixel statistics are expected to be a nuisance (Puli et al. 2022b), the representations from a randomly initialized neural network if such representations are expected to cluster based on nuisance (see Badgeley et al. (2019) for an example).

	AUROC ↑
ERM (Ming)	80.98 ± 2.22
IRM	81.29 ± 2.62
GDRO	82.94 ± 2.29
REx	81.25 ± 2.49
DANN	81.11 ± 3.10
CDANN	82.13 ± 1.76
ERM (Ours)	82.14 ± 1.55
RW	**86.86 ± 1.59**

Table 1: On Waterbirds $\rho = 0.7$, nuisance-aware OOD detection (RW) yields statistically significant improvement over nuisance-unaware detection using ERM (energy score, mean ± standard error over 4 seeds). In contrast, Ming, Yin, and Li (2022) do not see benefit using nuisance information with domain generalization algorithms.



(a) Output-based Detection (MSP)



(b) Feature-based Detection (MD)

Figure 4: On Waterbirds, using the outer patch as nuisance (patch) yields comparable results to utilizing the exact environment label (exact), which requires external information beyond the image.

Other results. We also consider alternatives to reweighting and joint independence, namely undersampling and marginal independence respectively. We find that undersampling achieves comparable output-based detection but worse feature-based detection and ID classification accuracy, and marginal independence yields worse performance than joint independence (Figure 5). We also find that the independence penalty with coefficient $\lambda = 1$ yields representations that are less predictive but not completely independent of the nuisance (Table 4 in Appendix) and suspect that removing nuisance information further could improve results.

8 Discussion

Out-of-distribution detection based on predictive models can suffer from poor performance on shared-nuisance OOD inputs when the classifier relies on a spurious correlation or its feature representations encode nuisance information. To address these failures, we present nuisance-aware OOD detection: to improve output-based detection, we train a classifier via reweighting to estimate a distribution that is robust to

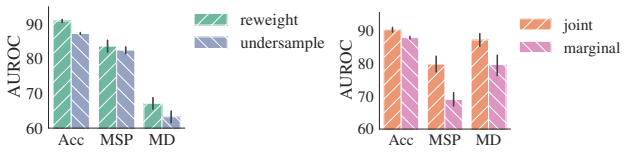


Figure 5: On Waterbirds $\rho = 0.9$, undersampling achieves comparable detection results to reweighting but worse balanced accuracy (left). Marginal independence performs worse than joint independence (right).

spurious correlations; to improve feature-based detection, we utilize reweighting and a joint independence constraint which encourages representations to be uninformative of the nuisance marginally or conditioned on the class label. Nuisance-aware OOD detection yields SN-OOD performance benefits for a wide array of existing detection methods while maintaining performance on non-SN-OOD inputs.

However, nuisance-aware OOD detection is not without its limitations. First, it requires *a priori* knowledge of nuisances and a way to specify them for a given task, though there are techniques that handle some missing values (Goldstein et al. 2022). In addition, implementing the joint independence penalty requires training a critic model for each gradient step of the main classifier, an expensive bi-level optimization. Fortunately, once the classifier is trained, prediction and thus detection time is no different from that of any other classifier, regardless of how it was trained. We also do not consider feature-based OOD detection methods which utilize all layers of a trained network, as such methods typically require validation OOD data.

Our work draws a connection between OOD generalization and OOD detection: classifiers that generalize well across spurious correlations also yield good output-based detection of SN-OOD inputs. Future work exploring other connections between OOD generalization and detection could be fruitful.

Ethical Statement

OOD detection is an important capability for reliable machine learning, spanning applications from robotics and transportation (e.g. novel object identification) to ecology and public health (e.g. novel species detection). SN-OOD detection focuses on a particularly difficult but relevant type of OOD input. Improved OOD detection can benefit these applications but also makes it easier for bad actors to deploy systems which detect anything that strays from the norm they define. Working towards intended applications while keeping in mind potential misuse can help usher in a future of more reliable machine learning systems with positive impact.

Acknowledgements

This work was generously funded by NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, and NSF CAREER Award 2145542. We thank Aahlad Puli and the anonymous AAAI reviewers for their helpful comments and suggestions.

References

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

Badgeley, M. A.; Zech, J. R.; Oakden-Rayner, L.; Glicksberg, B. S.; Liu, M.; Gale, W.; McConnell, M. V.; Percha, B. L.; Snyder, T. M.; and Dudley, J. T. 2019. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Hjelm, R. D.; and Courville, A. C. 2018. Mutual Information Neural Estimation. In *ICML*.

Bishop, C. M. 1994. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4): 217–222.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Fort, S.; Ren, J.; and Lakshminarayanan, B. 2021. Exploring the limits of out-of-distribution detection. In *NeurIPS*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *JMLR*.

Goldstein, M.; Jacobsen, J.-H.; Chau, O.; Saporta, A.; Puli, A. M.; Ranganath, R.; and Miller, A. 2022. Learning invariant representations with missing data. In *Conference on Causal Learning and Reasoning*, 290–301. PMLR.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *ICLR*.

Guo, R.; Zhang, P.; Liu, H.; and Kıcıman, E. 2021. Out-of-distribution Prediction with Invariant Risk Minimization: The Limitation and An Effective Fix. *arXiv preprint arXiv:2101.07732*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hendrycks, D.; Basart, S.; Mazeika, M.; Mostajabi, M.; Steinhardt, J.; and Song, D. X. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. In *ICML*.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.

Kamoi, R.; and Kobayashi, K. 2020. Why is the Mahalanobis Distance Effective for Anomaly Detection? *arXiv preprint arXiv:2003.00402*.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Priol, R. L.; and Courville, A. C. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *ICML*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *ECCV*.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *ICLR*.

Liu, W.; Wang, X.; Owens, J. D.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *NeurIPS*.

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)*.

Ming, Y.; Yin, H.; and Li, Y. 2022. On the Impact of Spurious Correlation for Out-of-distribution Detection. In *AAAI*.

Nalisnick, E.; Matsukawa, A.; Teh, Y.; Görür, D.; and Lakshminarayanan, B. 2019. Do Deep Generative Models Know What They Don't Know? In *ICLR*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Poole, B.; Ozair, S.; van den Oord, A.; Alemi, A. A.; and Tucker, G. 2019. On Variational Bounds of Mutual Information. In *ICML*.

Puli, A.; and Ranganath, R. 2020. General Control Functions for Causal Effect Estimation from Instrumental Variables. In *NeurIPS*.

Puli, A.; Zhang, L. H.; Oermann, E. K.; and Ranganath, R. 2022a. Out-of-Distribution Generalization in the Presence of Nuisance-induced Spurious Correlations. In *ICLR*.

Puli, A. M.; Joshi, N.; He, H. Y.; and Ranganath, R. 2022b. Nuisances via Negativa: Adjusting for Spurious Correlations via Data Augmentation. *arXiv preprint arXiv:2210.01302*.

Rosenfeld, E.; Ravikumar, P. K.; and Risteski, A. 2021. The Risks of Invariant Risk Minimization. In *ICLR*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *ICLR*.

Salehi, M.; Mirzaei, H.; Hendrycks, D.; Li, Y.; Rohban, M. H.; and Sabokrou, M. 2021. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*.

Sudarshan, M.; Puli, A. M.; Tansey, W.; and Ranganath, R. 2023. DIET: Conditional independence testing with marginal dependence measures of residual information. *AISTATS*.

Sugiyama, M.; Suzuki, T.; and Kanamori, T. 2012. *Density Ratio Estimation in Machine Learning*. Cambridge University Press.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2021. Generalized Out-of-Distribution Detection: A Survey. *arXiv preprint arXiv:2110.11334*.

Zhang, J.; Lopez-Paz, D.; and Bottou, L. 2022. Rich Feature Construction for the Optimization-Generalization Dilemma. In *ICML*.

Zhang, L. H.; Goldstein, M.; and Ranganath, R. 2021. Understanding Failures in Out-of-distribution Detection with Deep Generative Models. In *ICML*.

Zhou, B.; Lapedriza, À.; Khosla, A.; Oliva, A.; and Torralba, A. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40: 1452–1464.