Enhancing Multiple Reliability Measures via Nuisance-extended Information Bottleneck

Jongheon Jeong[†] Sihyun Yu[†] Hankook Lee[‡*] Jinwoo Shin[†]
[†]Korea Advanced Institute of Science and Technology (KAIST) [‡]LG AI Research
{jongheonj,sihyun.yu,jinwoos}@kaist.ac.kr hankook.lee@lgresearch.ai

Abstract

In practical scenarios where training data is limited, many predictive signals in the data can be rather from some biases in data acquisition (i.e., less generalizable), so that one cannot prevent a model from co-adapting on such (so-called) "shortcut" signals: this makes the model fragile in various distribution shifts. To bypass such failure modes, we consider an adversarial threat model under a mutual information constraint to cover a wider class of perturbations in training. This motivates us to extend the standard information bottleneck *to additionally model the* nuisance information. *We propose an autoencoder-based training to implement the objective, as well as practical encoder designs to facilitate the proposed hybrid discriminative-generative training concerning both convolutional- and Transformer-based architectures. Our experimental results show that the proposed scheme improves robustness of learned representations (remarkably without using any domain-specific knowledge), with respect to multiple challenging reliability measures. For example, our model could advance the state-of-the-art on a recent challenging OBJECTS benchmark in novelty detection by* $78.4\% \rightarrow 87.2\%$ *in AUROC, while simultaneously enjoying improved corruption, background and (certified) adversarial robustness. Code is available at* https://github.com/jh-jeong/nuisance_ib.

1. Introduction

Despite the recent breakthroughs in computer vision in aid of deep learning, *e.g.*, in image/video recognition [9, 20, 109], synthesis [42, 55, 96, 125], and 3D scene rendering [85, 86, 104], deploying deep learning models to the real-world still places a burden on contents providers as it affects the *reliability* of their services. In many cases, *deep neural networks* make substantially fragile predictions for *out-of-distribution* inputs, *i.e.*, samples that are not likely from the training distribution, even when the inputs are semantically close enough to in-distribution samples for humans [35, 102]. Such a vulnerability can be a significant threat in risk-sensitive systems, such as autonomous driving, medical imaging, and health-care applications, to name a few [2]. Overall, the phenomena highlight that deep neural networks tend to extract "shortcut" signals [26] from given limited (or potentially biased) training data in practice.

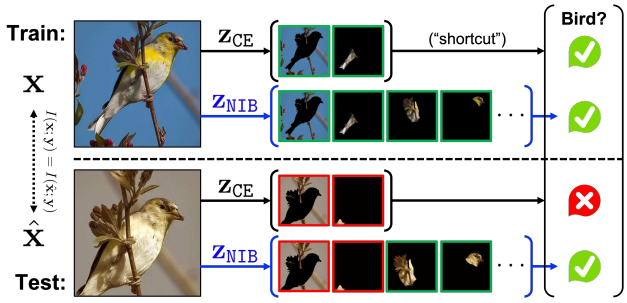

Figure 1. A high-level illustration of our method, *nuisance-extended information bottleneck* (NIB). In this paper, we focus on scenarios when the input $\mathbf{x}$ can be corrupted $\mathbf{x} \rightarrow \hat{\mathbf{x}}$ in test-time while preserving its semantics. Unlike the standard cross-entropy training (CE), NIB aims to encode *every* target-correlated signal in $\mathbf{x}$, some of which can be more reliable under distribution shifts.

To address such concerns, multiple literatures have been independently developed based on different aspects of model reliability. Namely, their methods use different *threat models* and benchmarks, depending on how a shift in input distribution happens in test-time, and how to evaluate model performance against the shift. For example, in the context of *adversarial robustness* [11, 17, 82, 128], a typical threat model is to consider the *worst-case* noise inside a fixed ℓ_p-ball around given test samples. Another example of *corruption robustness* [34, 35, 39, 116] instead assumes predefined types of common corruptions (*e.g.*, Gaussian noise, fog, *etc.*) that applies to the test samples. Lastly, *novelty detection* [36, 72, 74, 76] usually tests whether a model can detect a specific benchmark dataset as out-of-distribution from the (in-distribution) test samples.

*Work done at KAIST.

Due to discrepancy between each of "ideal" objectives and its practical threat models, however, the literatures have commonly found that optimizing under a certain threat model often hardly generalizes to other threat ones: *e.g.*, (a) several works [16, 62, 121] have observed that standard adversarial training [82] often harms other reliability measures such as corruption robustness or uncertainty estimation, as well as its classification performance; (b) Hendrycks et al. [34] criticize that none of the previous claims on corruption robustness could consistently generalize on a more comprehensive benchmark. This also happens even for threat models targeting the same objective: *e.g.*, (c) Yang et al. [122] show that state-of-the-arts in novelty detection are often too sensitive, so that they tend to also detect "near-in-distribution" samples as out-of-distribution and perform poorly on a benchmark regarding this. Overall, these observations suggest that one should avoid optimizing reliability measures assuming a specific threat model or benchmark, and motivate to find a new threat model that is generally applicable for diverse scenarios of reliability concerns.

Contribution. In this paper, we propose *nuisance-extended information bottleneck* (NIB), a new training objective targeting model reliability without assuming a prior on domain-specific tasks. Our method is motivated by rethinking the *information bottleneck* (IB) principle [107, 108] under presence of distribution shifts. Specifically, we argue that a "robust" representation $\mathbf{z} := f(\mathbf{x})$ should always encode *every* signal in the input $\mathbf{x}$ that is correlated with the target $\mathbf{y}$, rather than extracting only a few shortcuts (*e.g.*, Figure 1). This motivates us to consider an *adversarial* form of threat model on distribution shifts in $\mathbf{x}$, under a constraint on the mutual information $I(\mathbf{x}, \mathbf{y})$. To implement this idea, we propose a practical design by incorporating a *nuisance representation* $\mathbf{z}_n$ alongside $\mathbf{z}$ of the standard IB so that $(\mathbf{z}, \mathbf{z}_n)$ can reconstruct $\mathbf{x}$. This results in a novel synthesis of *adversarial autoencoder* [83] and *variational IB* [1] into a single framework. For the architectural side, we propose (a) to utilize the *internal feature statistics* for convolutional network based encoders, and (b) to incorporate *vector-quantized* patch representations for Transformer-based [24] encoders to model $\mathbf{z}_n$, mainly to efficiently encode the nuisance $\mathbf{z}_n$ (as well as $\mathbf{z}$) in a scalable manner.

We perform an extensive evaluation on the representations learned by our scheme, showing comprehensive improvements in modern reliability metrics: including (a) novelty detection, (b) corruption (or natural) robustness, (c) background robustness and (d) certified adversarial robustness. The results are particularly remarkable as the gains are not from assuming a prior on each of specific threat models. For example, we obtain a significant reduction in CIFAR-10-C error rates of the highest severity, *i.e.*, by $26.5\% \rightarrow 19.5\%$, without extra domain-specific prior as assumed in recent methods [39, 40]. Here, we also show that the effectiveness of our method is scalable to larger-scale (ImageNet) datasets. For novelty detection, we could advance AUROCs in recent OBJECTS [122] benchmarks by a large margin of $78.4\% \rightarrow 87.2\%$ in average upon the state-of-the-art, showing that our representations can provide a more semantic information to better discriminate out-of-distribution samples. Finally, we also demonstrate how the representations can further offer enhanced robustness against adversarial examples, by applying randomized smoothing [17] on them.

2. Background

Notation. Given two random variables $\mathbf{x} \in \mathcal{X}$, the input, and $\mathbf{y} \in \mathcal{Y}$, the target, our goal is is to find a mapping (or an *encoder*) $f : \mathcal{X} \rightarrow \mathcal{Z}$ from data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$[1] so that $\mathbf{z} := f(\mathbf{x})$, the representation, can predict $\mathbf{y}$ with a simper (*e.g.*, linear) mapping [7, 64]. We assume that f is parametrized by a neural network, and is *stochastic* to adopt an information theoretic view [108], *i.e.*, the encoder output is a random variable defined as $p_f(\mathbf{z}|\mathbf{x})$ rather than a constant. Such a modeling can be done through the *reparametrization trick* [61] with an independent random variable $\boldsymbol{\epsilon}$ and (deterministic) f by assuming $\mathbf{z} := f(\mathbf{x}, \boldsymbol{\epsilon})$. For example, one of standard designs parametrizes f by:

$$f(\mathbf{x}, \boldsymbol{\epsilon}) := f^{\mu}(\mathbf{x}) + \boldsymbol{\epsilon} \cdot f^{\sigma}(\mathbf{x}), \tag{1}$$

where $f^{\mu} \in \mathbb{R}^{|\mathcal{Z}|}$ and $f^{\sigma} \in \mathbb{R}_{+}^{|\mathcal{Z}|}$ are deterministic mappings modeling μ and σ in $\mathcal{N}(\mathbf{x}; \mu, \sigma^2 I)$, respectively, so that they can still be learned through a gradient-based optimization.

The data $\mathcal{D}$ is usually assumed to consist of *i.i.d.* samples from a certain *data generating distribution* $(x_i, y_i) \sim p_d(\mathbf{x}, \mathbf{y})$. One expects that f learned from $\mathcal{D}$ could generalize to predict $p_d(\mathbf{y}|\mathbf{x})$ for unseen samples from $p_d(\mathbf{x}, \mathbf{y})$. The formulation, however, does not specify how f should behave for inputs that are not likely from p_d, say $\hat{x}$. This becomes problematic for those who expect that the decision making of f should be close to that of human being, at least when $\hat{x}$ differs from p_d only up to what humans regard as *nuisance*. This is where the current neural networks commonly fail under the standard training practices.

Information bottleneck. Intuitively, a "good" representation $\mathbf{z}$ should keep information of $\mathbf{x}$ that is correlated with $\mathbf{y}$, while preventing $\mathbf{z}$ from being too complex. The *information bottleneck* [107, 108] (IB) is a principled approach to obtain such a succinct representation $\mathbf{x} \rightarrow \mathbf{z}$ for a given downstream task $\mathbf{x} \rightarrow \mathbf{y}$: namely, it finds $\mathbf{z}$ that (a) maximizes the (task-relevant) mutual information $I(\mathbf{z}; \mathbf{y})$, while (b) minimizing $I(\mathbf{x}; \mathbf{z})$ to constrain the capacity of $\mathbf{z}$ for better generalization. In other words, it sets the *mutual information* $I(\mathbf{x}; \mathbf{z})$ as the complexity measure of $\mathbf{z}$. Specifically,

[1] Although we focus on *supervised learning*, the framework itself in general does not rule out more general scenarios, *e.g.*, when the target $\mathbf{y}$ can be *self-supervised* from $\mathbf{x}$ [14, 91].

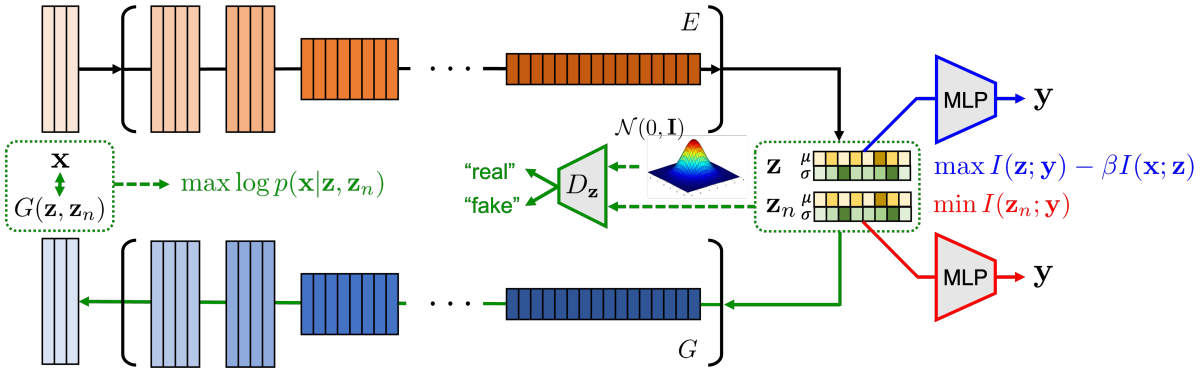

Figure 2. An overview of the proposed framework, the *autoencoder-based nuisance-extended information bottleneck* (AENIB). It illustrates the general pipeline, and Appendix C provides specific instantiations for convolutional and Transformer-based architectures. Overall, we incorporate adversarial autoencoder into variational information bottleneck by introducing a *nuisance* $\mathbf{z}_n$ (to $\mathbf{y}$) in representation learning.

it aims to maximize the following objective:

$$\max_f R_{\texttt{IB}}(f), \quad \text{for} \quad R_{\texttt{IB}}(f) := I(\mathbf{z};\mathbf{y}) - \beta I(\mathbf{x};\mathbf{z}), \quad (2)$$

where $\beta \geq 0$ controls the capacity constraint which ensures $I(\mathbf{x};\mathbf{z}) \leq I_\beta$ for some I_β.

3. Nuisance-extended information bottleneck

The standard information bottleneck (IB) objective (2) obtains a representation $\mathbf{z} := f(\mathbf{x})$ on premise that the future inputs will be also from $p_d(\mathbf{x},\mathbf{y})$. In this paper, we aim to extend IB under assumption that the input $\mathbf{x}$ can be corrupted through an *unknown* noisy channel in the future, say $\mathbf{x} \to \hat{\mathbf{x}}$, while $\hat{\mathbf{x}}$ still preserves the *semantics* of $\mathbf{x}$ with respect to $\mathbf{y}$: in other words, we assume $I(\mathbf{x};\mathbf{y}) = I(\hat{\mathbf{x}};\mathbf{y}) > 0$. Intuitively, one can imagine a scenario that $\mathbf{x}$ contains multiple signals that each is already highly correlated with $\mathbf{y}$, *i.e.*, filtering out the remainder from $\mathbf{x}$ does not affect its mutual information with $\mathbf{y}$. It may or may not be surprising that such signals are quite prevalent in deep neural networks, *e.g.*, [44] empirically observe that adversarial perturbations [29, 102] are sufficient for a model to perform accurate classification.

In the context of IB framework, where the goal is to obtain a succinct encoder f, it is now reasonable to presume that the noisy channel $\hat{\mathbf{x}}$ acts like an *adversary*, *i.e.*, it minimizes:

$$\min_{\hat{\mathbf{x}}} I(\hat{\mathbf{z}} := f(\hat{\mathbf{x}});\mathbf{y}) \;\; \text{subject to} \;\; I(\mathbf{x};\mathbf{y}) = I(\hat{\mathbf{x}};\mathbf{y}), \quad (3)$$

given that one has no information on how the channel would behave *a priori*. This optimization thus would require f to extract *every* signal in $\mathbf{x}$ whenever it is highly correlated with $\mathbf{y}$, to avoid the case when $\hat{\mathbf{x}}$ filters out all the signal except one that f has missed. We notice that, nevertheless, directly optimizing (3) with respect to $\hat{\mathbf{x}}$ is computationally infeasible in practice, considering that (a) it is in many cases an unconstrained optimization in a high-dimensional $\mathcal{X}$, (b) with a constraint on (hard-to-compute) mutual information.

In this paper, to make sure that f still exhibits the adversarial behavior without (3), we propose to let f to model the *nuisance representation* $\mathbf{z}_n$ as well as $\mathbf{z}$. Specifically, $\mathbf{z}_n$ aims to model the "remainder" information from $\mathbf{z}$ needed to reconstruct $\mathbf{x}$, *i.e.*, it maximizes $I(\mathbf{x};\mathbf{z},\mathbf{z}_n)$. At the same time, $\mathbf{z}_n$ compresses out any information that is correlated with $\mathbf{y}$, *i.e.*, it also minimizes $I(\mathbf{z}_n;\mathbf{y})$. Therefore, every information that is correlated with $\mathbf{y}$ should be encoded into $\mathbf{z}$ in a complementary manner. Here, we remark that now the role of the capacity constraint in (2) becomes more important: not only for regularizing $\mathbf{z}$ to be simpler, it also penalizes $\mathbf{z}_n$ from pushing out unnecessary information to predict $\mathbf{y}$ into $\mathbf{z}$, making the objective competitive again between $\mathbf{z}$ and $\mathbf{z}_n$ as like in (3). Combined with the original IB (2), we define *nuisance-extended IB* (NIB) as the following:

$$\max_f R_{\texttt{NIB}}(f) := R_{\texttt{IB}}(f) - I(\mathbf{z}_n;\mathbf{y}) + \alpha I(\mathbf{x};\mathbf{z},\mathbf{z}_n), \quad (4)$$

where $\alpha \geq 0$. The proposed NIB objective can be viewed as a regularized form of IB by introducing a nuisance $\mathbf{z}_n$. Specifically, this optimization additionally forces $I(\mathbf{x};\mathbf{z},\mathbf{z}_n)$ and $I(\mathbf{z}_n;\mathbf{y})$ in (4) to be maximized and minimized, *i.e.*, $H(\mathbf{x}|\mathbf{z},\mathbf{z}_n) = 0$ and $I(\mathbf{z}_n;\mathbf{y}) = 0$, respectively. The following highlights that having these conditions, also with the independence $\mathbf{z} \perp \mathbf{z}_n$, leads f that can recover the original $I(\mathbf{x};\mathbf{y})$ from $I(\hat{\mathbf{z}};\mathbf{y})$ (see Appendix E for the derivation):

Lemma 1. *Let $\mathbf{x} \in \mathcal{X}$, and $\mathbf{y} \in \mathcal{Y}$ be random variables, $\hat{\mathbf{x}}$ be a noisy observation of $\mathbf{x}$ with $I(\mathbf{x};\mathbf{y}) = I(\hat{\mathbf{x}};\mathbf{y})$. Given that $[\hat{\mathbf{z}},\hat{\mathbf{z}}_n] := f(\hat{\mathbf{x}})$ of $\hat{\mathbf{x}}$ satisfies (a) $H(\hat{\mathbf{x}}|\hat{\mathbf{z}},\hat{\mathbf{z}}_n) = 0$, (b) $I(\hat{\mathbf{z}}_n;\mathbf{y}) = 0$, and (c) $\hat{\mathbf{z}} \perp \hat{\mathbf{z}}_n$, it holds $I(\hat{\mathbf{z}};\mathbf{y}) = I(\mathbf{x};\mathbf{y})$.*

In the following sections, we provide a practical design of the proposed NIB based on an autoencoder-based architecture. Section 3.1 and 3.2 detail out its losses and architectures, respectively, and Section 3.3 summarizes the overall training. Figure 2 illustrates an overview of our framework.

3.1. AENIB: A practical autoencoder-based design

Based on the NIB objective defined in (4) and Lemma 1, we design a practical training objective to implement the proposed framework. Here, we present a simple instantiation of NIB with an autoencoder-based architecture upon *variational information bottleneck* (VIB) [1], calling it *autoencoder-based NIB* (AENIB).

Overall, Lemma 1 states that a robust encoder f demands for a "good" nuisance model that achieves generalization on $\hat{\mathbf{z}}$ in three aspects: (a) a *good reconstruction*, (b) *nuisance-ness*, and (c) the *independence between* $\mathbf{z}$ *and* $\mathbf{z}_n$. To model these behaviors, we consider a decoder $g : \mathcal{Z} \to \mathcal{X}$ as well as the encoder $f : \mathcal{X} \to \mathcal{Z}$, and adopt the following practical training objectives which incorporates an autoencoder-based loss and two adversarial losses [28]:

(a) We first pose a reconstruction loss to maximize $\log p(\mathbf{x}|\mathbf{z}, \mathbf{z}_n)$; standard designs assume the decoder output to follow $\mathcal{N}(\mathbf{x}, \sigma^2\mathbf{I})$, which is equivalent to the *mean-squared error* (MSE). Here, we use the *normalized MSE* to efficiently balance with other losses:[2]

$$L_{\texttt{recon}} := \tfrac{1}{\|\mathbf{x}\|_2^2}\|\mathbf{x} - g(\mathbf{z}, \mathbf{z}_n)\|_2^2 \tag{5}$$

(b) To force the nuisance-ness of $\mathbf{z}_n$ with respect to $\mathbf{y}$, we approximate $p(\mathbf{y}|\mathbf{z}_n)$ with a multi-layer perceptron (MLP), say q_n, and perform an adversarial training:

$$L_{\texttt{nuis}} := \mathbb{E}_{\mathbf{x}}[\mathbb{CE}(q_n^*(\mathbf{z}_n), \tfrac{\mathbf{1}}{|\mathcal{Y}|})], \quad \text{where } q_n^* := \min_{q_n} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbb{CE}(q_n(\mathbf{z}_n), \mathbf{y})], \tag{6}$$

and $\mathbb{CE}$ denotes the cross entropy loss. Here, it optimizes $\mathbb{CE}$ towards the uniform distribution in $\mathcal{Y}$.[3]

(c) To induce the independence between $\mathbf{z}$ and $\mathbf{z}_n$, we assume that the joint prior of $\mathbf{z}$ and $\mathbf{z}_n$ is the isotropic Gaussian, *i.e.*, $p(\mathbf{z}, \mathbf{z}_n) \sim \mathcal{N}(0, \mathbf{I})$, and performs a GAN training with a 2-layer MLP discriminator $q_{\mathbf{z}}$:

$$L_{\texttt{ind}} := \max_{q_{\mathbf{z}}} \mathbb{E}_{\mathbf{x}}[\log(q_{\mathbf{z}}(f(\mathbf{x})))] + \mathbb{E}_{\mathbf{z},\mathbf{z}_n \sim \mathcal{N}(0,\mathbf{I})}[\log(1 - q_{\mathbf{z}}(\mathbf{z}, \mathbf{z}_n))]. \tag{7}$$

Lastly, to approximate the original IB objective $R_{\texttt{IB}}(f)$ in NIB (4), we instead maximize the *variational information bottleneck* (VIB) [1] objective $L_{\texttt{VIB}}^{\beta}$, that can provide a lower bound on $R_{\texttt{IB}}$.[4] Specifically, it makes variational approximations of: (a) $p(\mathbf{y}|\mathbf{z})$ by a (parametrized) decoder neural network $q(\mathbf{y}|\mathbf{z})$, and (b) $p(\mathbf{z})$ by an "easier" distribution $r(\mathbf{z})$, *e.g.*, isotropic Gaussian $\mathcal{N}(\mathbf{z}|0, \mathbf{I})$. Assuming a Gaussian decoder (1) for $f(\mathbf{x}, \boldsymbol{\epsilon})$, we have:

$$L_{\texttt{VIB}}^{\beta} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\epsilon}}[-\log q(y_i|f(x_i, \boldsymbol{\epsilon}))] + \beta\, \mathrm{KL}\,(p(\mathbf{z}|x_i)\|r(\mathbf{z})). \tag{8}$$

3.2. Architectures for nuisance modeling

In principle, our framework is generally compatible with any encoder architectures: *e.g.*, say an encoder $f : \mathcal{X} \to \mathcal{Z}$ and decoder $g : \mathcal{Z} \to \mathcal{X}$, respectively. In order to apply VIB, we assume that the encoder has two output heads of dimension $2K$, where K denotes the dimension of $\mathbf{z}$. Here, each output head models a Gaussian random variable by reparametrization, *i.e.*, by modeling (μ, σ) as the encoder output for both $\mathbf{z} \in \mathbb{R}^K$ and $\mathbf{z}_n \in \mathbb{R}^{K_n}$.

Although it is possible that f models $\mathbf{z}$ and $\mathbf{z}_n$ by simply taking deep feed-forward representations following conventions, we observe that modeling nuisances $\mathbf{z}_n$ (which is essentially "generative") in standard (discriminative) architectures incur a training instability thus in performance: the nuisance information often requires to model finer details in a given inputs, which may be available rather in early layers of f, but not in the later layers for classification.

In this paper, we propose simple architectural treatments to improve the stability of nuisance modeling concerning both convolutional networks and Vision Transformers [24] (ViTs). This section focuses on introducing the design for convolutional networks, and we refer the readers for the ViT-based design to Appendix C: which is even simpler thanks to their patch-level representations available.

Given a convolutional encoder f, we encode $\mathbf{z}_n$ (as well as $\mathbf{z}$) from the collection of *internal features statistics*, rather than directly using the output of f. Specifically, we extract L intermediate feature maps of a given input $\mathbf{x}$, namely $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(L)}$ from $f(\mathbf{x})$, and define the *projection* of $\mathbf{x}$ by:

$$\Pi_f(\mathbf{x}) := \begin{bmatrix} \mathbf{m}^{(1)} & \mathbf{m}^{(2)} & \cdots & \mathbf{m}^{(L)} \\ \mathbf{s}^{(1)} & \mathbf{s}^{(2)} & \cdots & \mathbf{s}^{(L)} \end{bmatrix}, \tag{9}$$

where $\mathbf{m}^{(l)}$ and $\mathbf{s}^{(l)}$ are the first and second moment of feature maps in $\mathbf{x}^{(l)}$, assuming that $\mathbf{x}^{(l)} \in \mathbb{R}^{HWC}$: $\mathbf{m}_c^{(l)} := \frac{1}{HW}\sum_{hw} \mathbf{x}_{hwc}^{(l)}$, and $\mathbf{s}_c^{(l)} := \frac{1}{HW}\sum_{hw}(\mathbf{x}_{hwc}^{(l)} - \mathbf{m}_c^{(l)})^2$.

In Appendix I, we demonstrate that this simple projection can sufficiently encode a *generative* representation of $\mathbf{x}$: *viz.*, we show that one can successfully and efficiently train GANs with a discriminator defined upon Π_f. Motivated by this observation, we adopt Π_f in modeling the encoder representations $\mathbf{z}$ and $\mathbf{z}_n$. We encode $\mathbf{z}$ and $\mathbf{z}_n$ by simply applying MLPs to $\Pi_f(\mathbf{x})$ (9). Despite its simplicity, we observe this treatment enables a stable training of AENIB.

[2] We also explore a SSIM-based [118] reconstruction loss as given in Appendix C, which we found beneficial for robustness particularly with Transformer-based models.

[3] Alternatively, one can directly maximize $\mathbb{CE}(q_n^*(\mathbf{z}_n), \mathbf{y})$; we use the current design to avoid potential instability of the maximization-based loss.

[4] A more detailed description on the VIB framework (as well as on GAN) can be found in Appendix F.2.

	CIFAR-10-C							CIFAR-100-C						
Severity	Clean	1	2	3	4	5	Avg.	Clean	1	2	3	4	5	Avg.
Cross-entropy	6.08	8.89	11.1	14.0	19.7	26.5	16.0	25.1	31.4	35.1	39.3	46.8	54.0	41.3
VIB [1]	5.98	8.68	10.7	13.4	18.6	24.9	15.2	26.0	31.9	35.9	40.4	47.8	55.2	42.2
AugMix [39]	6.52	8.97	10.8	13.4	18.4	23.9	15.1	24.9	29.9	33.3	37.1	43.6	51.1	39.0
PixMix† [40]	5.43	7.10	8.14	9.40	12.1	14.9	10.3	23.2	26.7	28.7	30.8	35.0	39.0	32.0
AENIB (ours)	4.97	7.49	8.96	11.0	14.8	19.5	12.3	22.6	27.6	30.5	34.1	39.8	47.1	35.8
+ AugMix [39]	5.35	7.65	8.99	11.0	14.2	18.4	12.0	21.9	26.4	29.1	32.4	37.8	44.3	34.0
+ PixMix† [40]	**4.67**	**5.90**	**6.55**	**7.45**	**9.12**	**11.4**	**8.08**	**21.2**	**24.4**	**26.0**	**27.8**	**31.1**	**34.8**	**28.8**

Table 1. Comparison of average corruption error rates (%; ↓) per severity level on CIFAR-10/100-C [35]. Bold and underline denote the best and runner-up, respectively. †PixMix [40] utilizes an external dataset consisting of pattern- and fractal-like images.

Method	C10	C10-C	C10.1	C10.2	CINIC
Cross-entropy	6.08	16.0	13.4	18.3	23.7
VIB [1]	5.98	15.2	13.6	16.8	23.6
NLIB [65]	6.81	17.0	14.6	17.5	24.3
sq-NLIB [106]	6.02	15.5	13.0	17.1	23.7
DisenIB [92]	5.76	15.2	13.2	17.2	23.7
AugMix [39]	6.52	15.1	14.2	17.2	24.2
PixMix [40]	5.43	10.3	13.1	16.6	23.2
AENIB (ours)	4.97	12.3	11.6	15.5	22.2
+ AugMix [39]	5.35	12.0	12.5	15.8	22.6
+ PixMix [40]	**4.67**	**8.08**	**10.4**	**14.8**	**22.1**

Table 2. Comparison of test error rates (%; ↓) on CIFAR-10 and its variants: CIFAR-10-C/10.1/10.2, and CINIC. Bold and underline indicate the best and runner-up results, respectively.

3.3. Overall training objective

Combining the proposed objectives as well as the VIB loss, $L^{\beta}_{\texttt{VIB}}$ (8) leads us to the final objective. Although combining multiple losses in practice may introduce additional hyperparameters, we found most of the proposed losses can be added without scaling except for the reconstruction loss $L_{\texttt{recon}}$ and the β in the original VIB loss. Hence, we get:

$$L_{\texttt{AENIB}} := L^{\beta}_{\texttt{VIB}} + \alpha \cdot L_{\texttt{recon}} + L_{\texttt{nuis}} + L_{\texttt{ind}}. \qquad (10)$$

Algorithm 1 in Appendix A summarizes the procedure.

4. Experiments

We verify the effectiveness of our proposed AENIB training for various aspects of model reliability: specifically, we cover (a) corruption and natural robustness (Section 4.1), (b) novelty detection (Section 4.2), and (c) certified adversarial robustness (Section 4.3) tasks which all have been challenging without task-specific priors [37, 39, 82]. We provide an ablation study in Appendix D for a component-wise analysis. We also present an evaluation on our proposed components in the context of generative modeling in Appendix I. The full details on the experiments, *e.g.*, datasets, training details, and hyperparameters, can be found in Appendix B.

4.1. Robustness against natural corruptions

We first evaluate corruption robustness of our method, *i.e.*, its generalization ability under natural corruptions (*e.g.*, fog, brightness, *etc.*) and distribution shifts those are still semantic to humans. To this end, we consider a wide range of benchmarks those are derived from CIFAR-10 and ImageNet for the purpose of measuring generalization. Namely, for CIFAR-10 models we test on (a) CIFAR-10/100-C [35], a corrupted version of CIFAR-10/100 simulating 15 common corruptions in 5 severity levels, as well as (b) CIFAR-10.1 [94], CIFAR-10.2 [81], and CINIC-10 [21], *i.e.*, three re-generations of the CIFAR-10 test set. For ImageNet models, on the other hand, we test (a) ImageNet-C [35], a corrupted version of ImageNet validation set, (b) ImageNet-R [34], a collection of rendition images for 200 ImageNet classes, and ImageNet-Sketch [117], as well as (c) the Background Challenge [119] benchmark to evaluate model bias against background changes. This section mainly reports the results from ViT [24, 111] based architectures, but we also report the results with ResNet-18 [33] in Appendix H.

Table 1 and 2 summarize the results on CIFAR-based models. In Table 1, we observe that AENIB significantly and consistently improves corruption errors upon VIB, and these gains are strong even compared with state-of-the-art methods: *e.g.*, AENIB can solely outperform a strong baseline of AugMix [39]. Although a more recent method of PixMix [40] could achieve a lower corruption error by utilizing extra (pattern-like) data, we remark that (a) AENIB also benefit from PixMix (*i.e.*, the extra data) as given in "AENIB + PixMix", and (b) the results on Table 2 show that the generalization capability of AENIB is better than PixMix on CIFAR-10.1, 10.2 and CINIC-10, *i.e.*, in beyond common corruptions, by less relying on domain-specific data.

Next, Table 3 highlights that the effectiveness of AENIB can generalize to a more larger-scale, higher-resolution dataset of ImageNet: we still observe that AENIB can consistently improve robust accuracy for diverse corruption types, again without leveraging any further data augmentation during training. Figure 3 compares the linear trends made by Cross-entropy and AENIB across different data augmenta-

Dataset	ViT-S/16		ViT-B/16	
	Baseline	**AENIB (ours)**	Baseline	**AENIB (ours)**
IN-1K	**25.1**	**25.1**	21.8	**21.9**
IN-C (mCE)	65.9	**65.2** (−0.7)	58.6	**57.5** (−1.1)
IN-R	70.3	**67.1** (−3.2)	66.3	**64.4** (−1.9)
IN-Sketch	80.3	**77.7** (−2.6)	76.5	**74.4** (−2.1)

Table 3. Comparison of error rates (%; ↓) or mean corruption error (mCE, %; ↓) on ImageNet (IN) and its variants, namely IN-C [35], IN-R [34], and IN-Sketch [117]. Bold indicates the best results.

BG-Challenge	ViT-S/16		ViT-B/16	
Dataset	Baseline	**AENIB (ours)**	Baseline	**AENIB (ours)**
ORIGINAL (IN-9; ↑)	95.3	**95.5**	96.0	**96.1**
ONLY-BG-T (↓)	20.3	**17.8** (−2.5)	24.2	**21.1** (−3.1)
MIXED-SAME (↑)	86.3	**88.3** (+2.0)	87.4	**88.9** (+1.5)
MIXED-RAND (↑)	77.8	**80.5** (+2.7)	80.1	**81.8** (+0.7)
BG-gap (↓)	8.5	**7.8** (−0.7)	7.3	**7.1** (−0.2)

Table 4. Evaluation of AENIB on Backgrounds Challenge [119] compared to the cross-entropy baseline. All the models are trained on ImageNet, and warped to perform classification on ImageNet-9.

Method	Score	SVHN	LSUN	ImageNet	C100	CelebA
JEM [31]	$\log p(x)$	0.67	-	-	0.67	0.75
JEM [31]	$\max_y p(y\|x)$ [36]	0.89	-	-	0.87	0.79
SupCon [56]	$\max_y p(y\|x)$ [36]	0.97	0.93	0.91	**0.89**	-
Cross-entropy	$\max_y p(y\|x)$ [36]	0.94	0.94	0.92	0.86	0.64
Cross-entropy	$\log \mathrm{Dir}_{0.05}(y)$	0.96	0.95	0.94	0.86	0.61
VIB [1]	$\max_y p(y\|x)$ [36]	0.95	0.94	0.92	0.88	0.76
VIB [1]	$\log \mathrm{Dir}_{0.05}(y)$	0.97	0.96	0.94	0.88	0.78
AENIB (ours)	$\max_y p(y\|x)$ [36]	0.88	0.88	0.86	0.84	**0.81**
AENIB (ours)	$\log \mathrm{Dir}_{0.05}(y)$	0.90	0.95	0.92	0.86	0.80
AENIB (ours)	$+\log \mathcal{N}(z_n; 0, I)$	**0.98**	**0.99**	**0.99**	0.86	0.79

Table 5. Comparison of AUROC (%; ↑) for OOD detection from CIFAR-10 with five OOD datasets: SVHN, LSUN, ImageNet, CIFAR-100, and CelebA. Bolds indicate the best results.

tions and hyperparameters, confirming that AENIB exhibits a better operating points even in terms of *effective robustness* [105], given the recent observations on the correlation between in- *vs.* out-of-distribution performances across different models [34, 87, 105].

Lastly, Table 4 further evaluates the ImageNet classifiers on *Background Challenge* [119], a benchmark established to test the model robustness against background shifts: specifically, it constructs variants of ImageNet-9 (that combines 370 subclasses of ImageNet; ORIGINAL) with different combinations of backgrounds. Our AENIB-based models still consistently improve upon the cross-entropy baseline on the benchmark, showing that AENIB indeed tends to learn less-biased features against background changes.

4.2. Novelty detection

Next, we show that our AENIB model can be also a good detector for *out-of-distribution samples* (OODs), *i.e.*, to solve the *novelty detection* task. In general, the task is defined by a binary classification problem that aims to discriminate novel

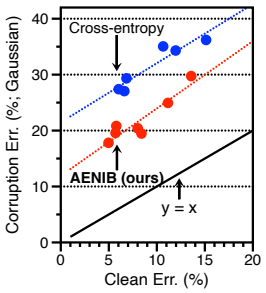

Figure 3. Comparison of trends in clean vs. corruption errors against Gaussian on ViT-S/4.

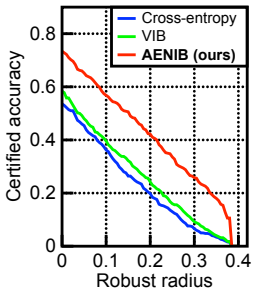

Figure 4. Comparison of certified adversarial robust accuracy at various radii on CIFAR-10.

samples from in-distribution samples. A typical practice here is to define a *score function* for each input, *e.g.*, the maximum confidence score [36], to threshold out samples as out-of-distribution when the score is low. To define a score function for AENIB models, we first observe that the *log-likelihood* of $\mathbf{z}_n$, which is only available for AENIB (and not for standard models), can be a strong score to detect novelties those are semantically far from in-distribution. Specifically, we use $\log \mathcal{N}(\mathbf{z}_n; 0, I) = -\frac{1}{2}\|\mathbf{z}_n\|^2$, as we assume that $\mathbf{z}$ follows isotropic Gaussian $\mathcal{N}(0, I)$. For detecting so-called "harder" novelties, we propose to use the log-likelihood score of $\mathbf{y}$ under a *symmetric Dirichlet distribution* of parameter $\alpha > 0$, namely $\mathrm{Dir}_\alpha(\mathbf{y}) \in \Delta^{|\mathcal{Y}|-1}$, rather than simply using $\max_y p(y|x)$: *i.e.*, $\log \mathrm{Dir}_\alpha(\mathbf{y}) = (\alpha - 1)\sum_i \log y_i$. Note that the distribution gets closer to the symmetric (discrete) one-hot distribution as $\alpha \to 0$, which makes sense for most classification tasks, and here we simply use $\alpha = 0.05$ throughout experiments.[5]

We consider two evaluation benchmarks: (a) the *"standard"* benchmark, that has been actively adopted in the literature [36, 72, 74], assumes the CIFAR-10 test set as in-distribution and measures the detection performance of other independent datasets; (b) a recent *OBJECTS* benchmark [122], on the other hand, extends the CIFAR-10 benchmark to also consider "near" in-distribution in OOD evaluation. Specifically, OBJECTS assumes CIFAR-10-C [35] and ImageNet-10 as in-distribution in test-time as well as CIFAR-10, making the detection much more challenging. In this experiment, we compare ResNet-18 [33] models trained on CIFAR-10 following the setup of [122].

The results are reported in Table 5 and 6 for the standard and OBJECTS benchmarks, respectively. Overall, we confirm that the score function combining the information of $\mathbf{z}_n$ and $\mathbf{y}$ of AENIB significantly improves novelty detection in a complementary manner over strong baselines, showing the effectiveness of modeling nuisance. For example, in Table 5, the combined score achieves near-perfect AUROCs for detecting SVHN, LSUN and ImageNet datasets.

[5] In practice, we observe that other choices in a moderate range of α near 0 do not much affect performance.

FS-OOD: OBJECTS		AUROC (%; ↑) / AUPR (%; ↑) / FPR@TPR95 (%; ↓)			
Method	Score	MNIST	FashionMNIST	Texture	CIFAR-100-C
Cross-entropy	$\max_y p(y\|x)$ [36]	66.98 / 52.66 / 93.54	73.78 / 90.15 / 88.08	74.18 / 93.34 / 85.64	74.12 / 89.74 / 87.26
	ODIN [74]	70.31 / 49.58 / 82.04	80.98 / 91.53 / **68.73**	70.14 / 89.97 / 72.91	67.51 / 83.97 / 84.26
	Energy-based [76]	54.55 / 34.14 / 92.23	76.50 / 89.80 / 72.40	68.63 / 89.51 / 75.57	68.37 / 85.54 / 83.64
	Mahalanobis [72]	77.04 / 65.31 / 84.59	80.33 / 92.28 / 77.17	72.02 / 88.46 / 72.98	68.13 / 82.97 / 85.53
	SEM [122]	75.69 / 76.61 / 99.70	79.40 / 93.14 / 93.72	79.69 / 95.48 / 82.15	78.89 / 92.07 / 83.92
	$\mathbf{\log \mathrm{Dir}_{0.05}(y)}$	76.75 / 66.26 / 83.51	82.88 / 93.97 / 77.19	70.69 / 92.68 / 91.35	78.80 / 92.21 / 82.50
VIB [1]	$\max_y p(y\|x)$ [36]	80.23 / 73.50 / 80.69	76.35 / 91.22 / 84.75	74.67 / 94.09 / 87.22	76.12 / 91.03 / 84.99
	$\mathbf{\log \mathrm{Dir}_{0.05}(y)}$	86.13 / 79.45 / 64.92	81.11 / 93.12 / 77.82	73.84 / 93.50 / 88.00	78.54 / 91.85 / 81.47
AENIB (ours)	$\max_y p(y\|x)$ [36]	79.67 / 71.50 / 80.22	77.33 / 91.63 / 84.31	74.95 / 93.97 / 86.01	74.31 / 89.89 / 86.26
	$\mathbf{\log \mathrm{Dir}_{0.05}(y)}$	90.53 / 85.68 / 52.08	84.56 / 94.61 / 74.24	75.04 / 93.83 / 86.01	79.39 / 92.33 / 81.51
	$\mathbf{+ \log \mathcal{N}(z_n; 0, I)}$	**92.43 / 89.38 / 48.10**	**84.85 / 94.84** / 74.67	**88.91 / 97.49 / 48.44**	**82.66 / 93.62 / 74.14**

Table 6. Comparison of OOD detection performances on the OBJECTS benchmark [122], which considers CIFAR-10-C and ImageNet-10 as in-distribution as well as the training in-distribution of CIFAR-10. Bold and underline denote the best and runner-up results, respectively.

Regarding Table 6, on the other hand, AENIB improves the previous best AUROC (of Mahalanobis [72]) on OBJECTS *vs.* MNIST from $77.04 \rightarrow 92.43$. The improved results on OBJECTS imply that both of the representation and score obtained from AENIB help to better discriminate in- *vs.* out-of-distribution in more *semantic* senses.

4.3. Certified adversarial robustness

We also evaluate adversarial robustness [29, 82, 102] adopting the *randomized smoothing* framework [17, 70] that can measure a *certified* robustness for a given representation. Specifically, any classifier can be robustified by averaging its predictions under Gaussian noise, where the robustness at input x depends on how consistent the classifier is on classifying $\mathcal{N}(x, \sigma^2\mathbf{I})$ [50]. Under this evaluation protocol, we suggest that adversarially-robust representations can be a natural byproduct from AENIB when combined with the randomized smoothing technique, without using any thorough adversarial training methods [82] that often require significant training cost with no certification on the robustness.

We follow the standard certification protocol [17] to compare the *certified test accuracy at radius* r, which is defined by the fraction of the test samples that a smoothed classifier classifies correctly with its certified radius larger than r. We consider ViT-S models on CIFAR-10, and assume $\sigma = 0.1$ for this experiment. The results summarized in Figure 4 show that our proposed AENIB achieves significantly better certified robustness compared to the baselines at all radii tested: *e.g.*, it improves certified robust accuracy of VIB by $39.6\% \rightarrow 56.8\%$ at $\varepsilon = 0.1$. Again, the robustness obtained from AENIB is not from specific knowledge on the threat model, which implies that AENIB could offer *free* adversarial robustness when combined with randomized smoothing. This confirms that the robustness of AENIB is not only significant but also consistent *per input*, especially considering its high certified robustness at higher r's.

(a) DisenIB [92]

(b) AENIB (Ours)

Figure 5. Comparison of CIFAR-10 reconstructions when the nuisance $\mathbf{z}_n$ is swapped with those of another (random) sample.

4.4. Comparison with DisenIB [92]

In this section, we provide a comparison of AENIB with a related work of DisenIB [92], as well as other variants of VIB, namely Nonlinear-VIB [65] and Squared-VIB [106]. Here, DisenIB is a variant of IB which also considers a nuisance modeling (based on FactorVAE [57]), in a purpose of supervised disentangling. Specifically, DisenIB considers two independent encoders $\mathbf{z} := f(\mathbf{x})$ and $\mathbf{z}_n := g(\mathbf{x})$, and aims to optimize the following objective:

$$\max_{f,g} I(\mathbf{z}; \mathbf{y}) + I(\mathbf{x}; \mathbf{z}_n, \mathbf{y}) - I(\mathbf{z}_n; \mathbf{z}). \tag{11}$$

Compared to our proposed NIB (4), the most important difference between the two objectives is in their "reconstruction" terms: *i.e.*, $I(\mathbf{x}; \mathbf{z}_n, \mathbf{y})$ of (11) *vs.* $I(\mathbf{x}; \mathbf{z}, \mathbf{z}_n)$ of ours (4). Due to this difference, the DisenIB objective (11) cannot rule out the cases when $\mathbf{z}$ only encode few of "shortcut" signals in $\mathbf{x}$ (correlated to $\mathbf{y}$) even at optimum, in contrast to our key motivation of NIB (4) that aims to let $\mathbf{z}$ to encode *every* $\mathbf{y}$-correlated signal in $\mathbf{x}$ as much as possible.

We conduct experimental comparisons based on (a) our CIFAR-10 setups (Table 2), and (b) directly upon the official implementation[6] of DisenIB on MNIST [69] (Table 9 of Ap-

[6] https://github.com/PanZiqiAI/disentangled-information-bottleneck

pendix G). Here, we adopt and extend the benchmark to also cover *MNIST-C* [88], a corrupted version of the MNIST. For the latter comparison, we use a simple 4-layer convolutional network as the encoder architecture. We train every MNIST model here for 100K updates and follow the other training details from the CIFAR experiments (see Appendix B.1).

Overall, we observe that the effectiveness of AENIB still applies to these benchmarks. This is in contrast to DisenIB, given that the effectiveness from DisenIB, *e.g.*, its gain in AUROC on detecting Gaussian noise as an OOD (as conducted by [92]), could not be further generalized on CIFAR-10-C or MNIST-C, where AENIB still improves on as well as achieving the perfect score at the same OOD task. In Figure 5a, we further observe qualitatively that DisenIB often leaves highly semantic information in the nuisance $\mathbf{z}_n$: its reconstruction can be completely changed by swapping $\mathbf{z}_n$ with those of another sample. This is essentially what AENIB addresses, as compared in Figure 5b.

5. Related work

Out-of-distribution robustness. Since the seminal works [2, 90, 102] revealing the fragility of neural networks on out-of-distribution inputs, there have been significant attempts on identifying and improving various notions of robustness: *e.g.*, detecting novel inputs [36, 71, 72, 103], robustness against corruptions [27, 35, 39, 119], and adversarial noise [5, 11, 17, 82], to name a few. Due to its fundamental challenges in making neural network to extrapolate, however, most of the advances in the robustness literature has been made under assuming priors closely related to the individual problems: *e.g.*, an external data or data augmentations [37, 39], extra information from test-time samples [116], or specific knowledge in threat models [52, 112]. In this work, we aim to improve multiple notions of robustness without assuming such priors, through a new training scheme that extends the standard information bottleneck principle under noisy observations in test-time.

Hybrid generative-discriminative modeling. Our proposed method can be also viewed as a new approach of improving the robustness of discriminative models by incorporating a generative model, in the context that has been explored in recent works [31, 72, 99, 123]. For example, [72, 73] have incorporated a simple (but of low expressivity for generation) Gaussian mixture model into discriminative classifiers; a line of research on *Joint Energy-based Models* (JEM) [31, 123] assumes an energy-based model but with a notable training instability for the purpose. In this work, we propose an autoencoder-based model to avoid such training instability, and consider a design that the *nuisance* can succinctly supplement the given discriminative representation to be generative. We demonstrate that our approach can take the best of two worlds; it enables (a) stable training, while (b) attaining the high expressive generative performances.

Nuisance modeling. The idea of incorporating nuisances can be also considered in the context of *invertible* modeling, or as known as *flow-based models* [6, 23, 30, 59], where the nuisance can be defined by splitting the (full-information) encoding $\mathbf{z}$ for a given subspace of interest as explored by [4, 46]. Unlike such approaches, our autoencoder-based nuisance modeling does not focus on the "full" invertibility for arbitrary inputs, but rather on inverting the data manifold given, which enabled (a) a much flexible encoder design in practice, and (b) a more scalable generative modeling of nuisance $\mathbf{z}_n$, *e.g.*, beyond an MNIST-scale as done by [46]. Other related works [48, 49, 92] do introduce an encoder for nuisance factors, but the notion of nuisance-ness has been focused in terms of the independence to $\mathbf{z}$ (for the purpose of feature disentangling), rather than to $\mathbf{y}$ as we focus in this work (for the purpose of robustness): *e.g.*, DisenIB [92] applies FactorVAE [57] between semantic and nuisance embeddings to force their independence. Yet, the literature has been also questioned on whether the idea can be scaled-up beyond, *e.g.*, MNIST, and our work does explore and establish a practical design with recent architectures and datasets addressing diverse modern security metrics.

We provide more extensive and detailed discussions on related works in Appendix F.

6. Conclusion

We suggest that having a good *nuisance model* can be a tangible approach to induce a reliable representation. We develop a practical method of learning deep nuisance representation from data, and show its effectiveness to improve diverse reliability measures under a challenging setup of assuming no prior [105]. We believe our work can be a useful step towards better understanding of out-of-distribution generalization in deep learning. Although the current scope is on a particular design of autoencoder based models, our framework of *nuisance-extended IB* is not limited to it and future works could consider more diverse implementations. Ultimately, we aim to approximate a challenging form of adversarial training with a mutual information constraint, which we believe will be a promising direction to explore.

Acknowledgments

This work was partly supported by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by Defense Acquisition Program Administration (DAPA) and Agency for Defense Development (ADD) (UD190031RD), and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)). We thank Subin Kim for the proofreading of our manuscript.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. 2, 4, 5, 6, 7, 20, 21, 22

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. 1, 8, 19

[3] Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. *Advances in Neural Information Processing Systems*, 34, 2021. 15, 22

[4] Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. Training normalizing flows with the information bottleneck for competitive generative classification. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7828–7840. Curran Associates, Inc., 2020. 8, 20

[5] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 10–15 Jul 2018. 8, 19

[6] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019. 8, 20

[7] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995. 2

[8] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain ViT baselines for ImageNet-1k. *arXiv preprint arXiv:2205.01580*, 2022. 14

[9] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning*, pages 1059–1071. PMLR, 2021. 1

[10] Dominique Brunet, Edward R Vrscay, and Zhou Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011. 17

[11] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. 1, 8, 19

[12] Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 20

[13] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 20

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020. 2

[15] Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021. 20

[16] Sanghyuk Chun, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. *arXiv preprint arXiv:2003.03879*, 2020. 2

[17] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019. 1, 2, 7, 8, 19

[18] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. 14

[19] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019. 22

[20] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 1

[21] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. CINIC-10 is not ImageNet or CIFAR-10. *arXiv preprint arXiv:1810.03505*, 2018. 5, 15

[22] James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. *Advances in Neural Information Processing Systems*, 34, 2021. 19

[23] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP, 2016. 8, 20

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 4, 5, 17

[25] Ethan Fetaya, Joern-Henrik Jacobsen, Will Grathwohl, and Richard Zemel. Understanding the limitations of conditional generative models. In *International Conference on Learning Representations*, 2020. 20

[26] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[27] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture;

increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 8, 19

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 4, 21

[29] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 3, 7, 19

[30] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. *International Conference on Learning Representations*, 2019. 8, 20

[31] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. 6, 8, 19

[32] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 17

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6, 16, 23

[34] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, October 2021. 1, 2, 5, 6, 15

[35] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 1, 5, 6, 8, 15, 18, 19, 21, 22

[36] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 1, 6, 7, 8, 19

[37] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 5, 8, 19

[38] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019. 19

[39] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, 2020. 1, 2, 5, 8, 19

[40] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. PixMix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792, 2022. 2, 5

[41] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016. 20

[42] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[43] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 20

[44] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[45] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 16, 21

[46] Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. In *International Conference on Learning Representations*, 2019. 8, 20

[47] Jörn-Henrik Jacobsen, Arnold W.M. Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018. 20

[48] Ayush Jaiswal, Rob Brekelmans, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. Discovery and separation of features for invariant representation learning, 2019. 8, 20

[49] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised adversarial invariance. *Advances in Neural Information Processing Systems*, 31, 2018. 8, 20

[50] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 33:10558–10570, 2020. 7

[51] Jongheon Jeong and Jinwoo Shin. Training GANs with stronger augmentations via contrastive discriminator. In *In-*

ternational Conference on Learning Representations, 2021. 22

[52] Daniel Kang, Yi Sun, Tom Brown, Dan Hendrycks, and Jacob Steinhardt. Transfer of adversarial robustness between perturbation types, 2019. 8, 19

[53] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12104–12114. Curran Associates, Inc., 2020. 22, 23

[54] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 16

[55] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 22, 23

[56] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 6

[57] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 7, 8, 20

[58] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 14

[59] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 8, 20

[60] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016. 20

[61] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 2, 20, 21

[62] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *Uncertainty in Artificial Intelligence*, pages 1012–1021. PMLR, 2022. 2

[63] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020. 20

[64] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990. 2

[65] Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019. 5, 7, 21

[66] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009. 14, 23

[67] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 15

[68] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in GANs. In *International Conference on Machine Learning*, pages 3581–3590. PMLR, 2019. 14

[69] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. 7, 21

[70] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019. 7

[71] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. 8, 19

[72] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1, 6, 7, 8, 19

[73] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3763–3772. PMLR, 09–15 Jun 2019. 8, 19

[74] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1, 6, 7

[75] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2021. 16, 22, 23

[76] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 1, 7

[77] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, December 2015. 15, 23

[78] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts, 2016. 14

[79] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 14

[80] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder, 2015. 20

[81] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. 5, 15

[82] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2, 5, 7, 8, 19

[83] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 2, 20, 23

[84] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 23

[85] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 1

[86] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[87] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021. 6

[88] Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019. 8, 21

[89] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. 20

[90] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 8, 19

[91] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[92] Ziqi Pan, Li Niu, Jianfu Zhang, and Liqing Zhang. Disentangled information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9285–9293, 2021. 5, 7, 8, 20, 21, 22

[93] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 823–832, 2021. 15, 22

[94] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018. 5, 15

[95] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 20

[96] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[97] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 15

[98] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 23

[99] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *International Conference on Learning Representations*, 2019. 8, 19

[100] Joan Serra, David Alvarez, Vicenc Gomez, Olga Slizovskaia, Jose F. Nunez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020. 20

[101] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 20

[102] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. 1, 3, 7, 8, 19

[103] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, 2020. 8, 19

[104] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 1

[105] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 6, 8

[106] R Thobaben, M Skoglund, et al. The convex information bottleneck lagrangian. *Entropy*, 22(1), 2020. 5, 7, 21

[107] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. 2, 20

[108] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop*, pages 1–5, 2015. 2

[109] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 1

[110] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008. 15

[111] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 5, 17

[112] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 8, 19

[113] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020. 20, 22

[114] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 17

[115] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. 20

[116] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1, 8, 19

[117] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 5, 6, 15

[118] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4, 17

[119] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. 5, 6, 8

[120] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20685–20696. Curran Associates, Inc., 2020. 19, 20

[121] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020. 2

[122] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2204.05306*, 2022. 2, 6, 7

[123] Xiulong Yang and Shihao Ji. JEM++: Improved techniques for training JEM. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6494–6503, October 2021. 8, 19

[124] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022. 17

[125] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022. 1

[126] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 14

[127] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 14

[128] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019. 1, 19

[129] Zijun Zhang, Ruixiang Zhang, Zongpeng Li, Yoshua Bengio, and Liam Paull. Perceptual generative autoencoders. In *International Conference on Machine Learning*, pages 11298–11306. PMLR, 2020. 22

[130] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020. 22, 23