LEARNING ROBUST REPRESENTATIONS VIA NUISANCE-EXTENDED INFORMATION BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

The *information bottleneck* (IB) is a principled approach to obtain a succinct representation $\mathbf{x} \to \mathbf{z}$ for a given downstream task $\mathbf{x} \to \mathbf{y}$: 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 practical scenarios where the training data is limited, however, many predictive-yet-compressible signals in the data can be rather from some biases in data acquisition (*i.e.*, less generalizable), so that even the IB objective cannot prevent $\mathbf{z}$ from co-adapting on such (so-called) "shortcut" signals. To bypass such a failure mode, we consider an adversarial threat model of $\mathbf{x}$ under constraint on the mutual information $I(\mathbf{x}; \mathbf{y})$. This motivates us to extend IB to additionally model the *nuisance information* against $\mathbf{z}$, namely $\mathbf{z}_n$, so that $(\mathbf{z}, \mathbf{z}_n)$ can reconstruct $\mathbf{x}$. To enable the idea, we propose an auto-encoder based training upon the variational IB framework, as well as practical encoder designs to facilitate the proposed hybrid discriminative-generative training considering 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 modern security measures including novelty detection, corruption (or natural) robustness and certified adversarial robustness.

1 INTRODUCTION

Despite the recent breakthroughs with the development of deep learning, *e.g.*, in vision and language processing (He et al., 2016; Vaswani et al., 2017; Brown et al., 2020; Mildenhall et al., 2020), reinforcement learning (Vinyals et al., 2019; Jang et al., 2021), and scientific discovery (Davies et al., 2021; Jumper et al., 2021), deploying current deep learning models to the real-world still places a significant burden on contents providers as the models are likely to affect 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: *e.g.*, for samples perturbed by an imperceptable, adversarially-crafted noise (Szegedy et al., 2014; Goodfellow et al., 2015), or natural corruptions such as "fog" (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021).

Generally speaking, a neural network, say f, is a parametric mapping of a given random variable $\mathbf{x}$ into its representation $\mathbf{z} := f(\mathbf{x})$, that encodes "useful" features in $\mathbf{x}$ to predict a target random variable $\mathbf{y}$: *i.e.*, a simpler (*e.g.*, linear) mapping can recover $\mathbf{y}$ from $\mathbf{z}$. In other words, 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* (IB) principle (Tishby et al., 1999; Tishby & Zaslavsky, 2015) is a simple and natural implementation of this idea, which sets the *mutual information* $I(\mathbf{x}; \mathbf{z})$ as the complexity measure of $\mathbf{z}$. Specifically, it aims to maximize the following objective:

$$\max_{f} R_{\texttt{IB}}(f), \qquad \text{for} \quad R_{\texttt{IB}}(f) := I(\mathbf{z}; \mathbf{y}) - \beta I(\mathbf{x}; \mathbf{z}), \tag{1}$$

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

However, the brittleness of neural networks for out-of-distribution samples can still persist even with the IB objective (1): in other words, a "good" model f from the objective can work poorly under a certain *distribution shift* in $\mathbf{x}$, say $\hat{\mathbf{x}}$, so that $I(\mathbf{x}; \mathbf{y}) = I(f(\mathbf{x}); \mathbf{y}) \gg I(f(\hat{\mathbf{x}}); \mathbf{y})$. In practice, this can

occur especially when the (hard-to-compute) mutual information terms in (1) are approximated based on limited, and potentially biased data: for example, many well-curated datasets commonly used in research (Krizhevsky, 2009; Russakovsky et al., 2015) are likely to be processed prior to release for quality control, *e.g.*, by filtering out some severely corrupted samples from its original collection. Such a bias can make the computation of $I(\mathbf{z};\mathbf{y})$ to be also biased, *i.e.*, toward over-estimating a "shortcut" signal (Geirhos et al., 2020) in the data that is not generalizable for $\hat{\mathbf{x}}$. Even worse, by jointly minimizing $I(\mathbf{x};\mathbf{z})$ in (1), it can further compress out other useful signal in $\mathbf{x}$ if the shortcuts are already predictive enough.

Contribution. In this paper, we rethink the implementation of the information bottleneck (IB) principle under presence of distribution shifts (between training and test data). Specifically, we argue that a "robust" representation $\mathbf{z}$ should always encode *every* signal in $\mathbf{x}$ that is correlated with $\mathbf{y}$, rather than extracting only a few shortcuts; the capacity constraint in IB (1) can still be applied for the *nuisance* information which is not related to predict $\mathbf{y}$ at all. This motivates us to consider an *adversarial* form of threat model of distribution shifts in $\mathbf{x}$, under a constraint on the mutual information $I(\mathbf{x},\mathbf{y})$. To enable this idea, we propose a practical design by incorporating a *nuisance representation* $\mathbf{z}_n$ alongside $\mathbf{z}$ of the standard IB framework so that $(\mathbf{z},\mathbf{z}_n)$ can reconstruct $\mathbf{x}$. This results in a novel synthesis of *adversarial autoencoder* (Makhzani et al., 2015) and *variational information bottleneck* (Alemi et al., 2017) into a single framework. For more details on the neural architectural side, we propose (a) to utilize the *internal feature statistics* for convolutional network based encoders, *i.e.*, the collection of mean and variance of each feature map, and (b) to incorporate *vector-quantized* patch representations for Transformer-based (Dosovitskiy et al., 2021) encoders to model $\mathbf{z}_n$ along with continuous $\mathbf{z}$, mainly to efficiently encode the nuisance representation $\mathbf{z}_n$ (as well as $\mathbf{z}$) in a scalable manner.

We perform an extensive evaluation on the representations learned by our scheme, particularly focusing on their generalization ability on out-of-distribution inputs. Overall, we demonstrate that our framework can now successfully address possible future corruptions in the input, making consistent improvements in all the modern robustness measures considered compared to the standard (*e.g.*, cross-entropy based) training. The results are particularly remarkable as the gains are not from assuming a prior on out-of-distribution. 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 any domain-specific priors as assumed in recent methods, *e.g.*, AugMix and PixMix (Hendrycks et al., 2020; 2022). Here, we also show that the effectiveness of our method is scalable to larger-scale (ImageNet) datasets. For novelty detection, we show that our representations can provide a more semantic information to better discriminate out-of-distribution samples: *e.g.*, we could advance AUROCs in recent OBJECTS (Yang et al., 2022) benchmarks by $78.4\% \rightarrow 87.2\%$ in average upon the previous best results. Finally, we also demonstrate how the representations can further offer enhanced certified robustness against adversarial examples "for free", by applying randomized smoothing (Cohen et al., 2019) on them.

2 Nuisance-extended information bottleneck

Notation. Given two random variables $\mathbf{x} \in \mathcal{X}$, the input, and $\mathbf{y} \in \mathcal{Y}$, the target, we consider a general problem of *representation learning* (Bell & Sejnowski, 1995; Kohonen, 1990; Dinh et al., 2014; Alemi et al., 2017; Donahue et al., 2017; Oord et al., 2018), where the goal 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. We assume that the encoder f is parametrized by a neural network, and the mapping is *stochastic* to adopt an information theoretic view of neural networks (Tishby & Zaslavsky, 2015), *i.e.*, the encoder output is a random variable defined as $p_f(\mathbf{z}|\mathbf{x})$ rather than a constant. In practice, such a modeling can be done through the *reparametrization trick* (Kingma & Welling, 2014), *i.e.*, by allowing an independent random variable $\boldsymbol{\epsilon}$ to the (deterministic) mapping f as an additional input, namely $\mathbf{z} := f(\mathbf{x}, \boldsymbol{\epsilon})$. For example, a popular design of *Gaussian decoder* parametrizes f by:

$$f(\mathbf{x}, \boldsymbol{\epsilon}) := f^{\mu}(\mathbf{x}) + \boldsymbol{\epsilon} \cdot f^{\sigma}(\mathbf{x}), \tag{2}$$

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.

[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 $\mathbf{x}$ (Oord et al., 2018; Chen et al., 2020).

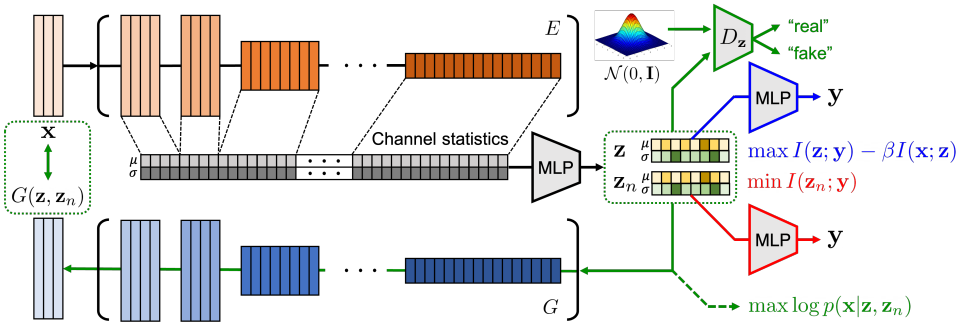


Figure 1: An overview of our proposed framework, *nuisance-extended information bottleneck* (NIB), instantiated by an autoencoder-based design. Here, the illustration is based on convolutional architectures, while we also propose a similar instantiation for Transformer-based models in Appendix C. Overall, the training incorporates adversarial autoencoder into the variational information bottleneck framework by introducing a *nuisance* $\mathbf{z}_n$ with respect to $\mathbf{y}$ in representation learning.

Conventionally, the given data $\mathcal{D}$ is usually assumed to consist of *i.i.d.* samples from a certain *data generating distribution* $(x_i, y_i) \sim p_d(\mathbf{x}, \mathbf{y})$, and one expects that f learned from $\mathcal{D}$ could generalize well 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 additionally 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*: where the current neural networks commonly fail under the standard training practices.

Nuisance-extended IB. The standard information bottleneck (IB) objective (1) obtains a representation $\mathbf{z} := f(\mathbf{x})$ on premise that the future inputs will be also from the data generating distribution $p_d(\mathbf{x}, \mathbf{y})$. In this paper, we aim to extend the IB objective under assumption that the input $\mathbf{x}$ can possibly 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 a given input $\mathbf{x}$ 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 practical deep neural networks, *e.g.*, Ilyas et al. (2019) empirically observe that adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2015) crafted from a given neural network are sufficient for the 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:

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

given that one has no information on how the channel would behave *a priori*. This minimax 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 $I(\mathbf{x}; \mathbf{z}, \mathbf{z}_n)$, while compressing out 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 (1) becomes even more important: not only for regularizing $\mathbf{z}$ to attain simpler representation, it additionally 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 objective (1), we define *nuisance-extended IB* (NIB) as the following:

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

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.*, towards $H(\mathbf{x}|\mathbf{z}, \mathbf{z}_n) = 0$ and $I(\mathbf{z}_n; \mathbf{y}) = 0$, respectively. The following observation highlights that having these conditions, additionally with the independence $\mathbf{z} \perp \mathbf{z}_n$, leads f that can recover the original information of $I(\mathbf{x}; \mathbf{y})$ from the noisy channel $I(\hat{\mathbf{z}}; \mathbf{y})$:

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 a representation $[\hat{\mathbf{z}}, \hat{\mathbf{z}}_n] := f(\hat{\mathbf{x}})$ 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 objective based on an autoencoder-based architecture. Section 2.1 and 2.2 detail out its losses and architectures, respectively, and Section 2.3 summarizes the overall training. Figure 1 illustrates an overview of our framework.

2.1 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 by approximating it with an autoencoder-based architecture upon *variational information bottleneck* (VIB) (Alemi et al., 2017), calling it *autoencoder-based nuisance-extended information bottleneck* (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 (Goodfellow et al., 2014):

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

$$L_{\texttt{recon}} := -C \cdot \log p(\mathbf{x}|\mathbf{z}, \mathbf{z}_n) = \tfrac{1}{\|\mathbf{x}\|_2^2} \|\mathbf{x} - g(\mathbf{z}, \mathbf{z}_n)\|_2^2 =: \mathrm{NMSE}(\mathbf{x}; g(\mathbf{z}, \mathbf{z}_n)). \tag{5}$$

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

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

where $\mathbb{CE}$ denotes the cross entropy loss. Here, it optimizes the cross-entropy towards the "uniform" distribution in $\mathcal{Y}$.

(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.*, $p(\mathbf{z}, \mathbf{z}_n) \sim \mathcal{N}(0, I)$, and performs a GAN-based training:

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

where $q_{\mathbf{z}}$ is an MLP that discriminates $[\mathbf{z}, \mathbf{z}_n]$ from $\mathcal{N}(0, I)$.

Lastly, to approximate the original IB objective $R_{\texttt{IB}}(f)$ in NIB (4), we instead maximize the *variational information bottleneck* (VIB) (Alemi et al., 2017) objective $L_{\texttt{VIB}}^{\beta}$, that can provide a lower bound on $R_{\texttt{IB}}$.[3] 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 $\mathcal{N}(\mathbf{z}|0, I)$. Recalling that we assume a Gaussian decoder (2) for $f(\mathbf{x}, \boldsymbol{\epsilon})$, we have:

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

[2] We also explore a SSIM-based (Wang et al., 2004) reconstruction loss as given in Appendix C, which we found beneficial for robustness particularly with Transformer-based models.

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

2.2 ARCHITECTURES FOR NUISANCE MODELING

In principle, our framework is generally compatible with existing any deep network 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 size of latent representation $\mathbf{z}$: here, each output head models the 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 the encoder f models representations $\mathbf{z}$ and $\mathbf{z}_n$ by simply taking outputs from a deep feed-forward representations following conventions, we observe that modeling nuisances $\mathbf{z}_n$ (which should be essentially "generative") as well as $\mathbf{z}$ in standard discriminative architectures can incur a bottleneck in training stability thus in performance compared to modeling without $\mathbf{z}_n$: the nuisance information often requires to model the fine details in a given inputs, which is available in early layers of f, but may not in the later layers for classification.

Here, we propose a simple architectural treatment to improve stability of nuisance modeling that are applicable for any convolutional networks: specifically, we encode $\mathbf{z}_n$ (as well as $\mathbf{z}$) from the collection of *internal features statistics*, rather than by a mapping from the last layer of f.

Motivation: Feature statistics discriminator (FSD) for GANs. Designing a stable discriminator has been crucial for GANs: a usual practice in the literature is to have a separate, carefully-designed network with a comparable generator, but with a significant overhead. We observe that the *internal feature statistics* of a convolutional encoder f can be a surprisingly effective representation to define a simple yet efficient discriminator. Concretely, for a given encoder f and an input $\mathbf{x}$, we consider L intermediate feature maps of $\mathbf{x}$, 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 channel-wise feature maps in $\mathbf{x}^{(l)}$, assuming that $\mathbf{x}^{(l)} \in \mathbb{R}^{HWC}$ follows the format of convolutional feature maps:

$$\mathbf{m}_c^{(l)} := \frac{1}{HW} \sum_{h,w} \mathbf{x}_{h,w,c}^{(l)}, \text{ and } \mathbf{s}_c^{(l)} := \frac{1}{HW} \sum_{h,w} (\mathbf{x}_{h,w,c}^{(l)} - \mathbf{m}_c^{(l)})^2. \tag{10}$$

The *features statistics discriminator* (FSD) we consider here is then simply a 3-layer MLP applied on $\Pi_f(\mathbf{x})$. In Appendix G.1, we empirically confirm that this simplest design of discriminator can dramatically accelerate GAN training, particularly when applied upon pre-trained discriminative encoders: similarly to Sauer et al. (2021) but with a simpler architecture.

Motivated by the observation that the features statistics based projection Π_f can better encode generative representations in discriminative models, we apply this to model the encoder representations $\mathbf{z}$ and $\mathbf{z}_n$: specifically, we encode $\mathbf{z}$ and $\mathbf{z}_n$ by simply applying MLPs to the feature statistics projection $\Pi_f(\mathbf{x})$ (9). Despite its simplicity, we observe this modeling indeed enables a faster and stable training of AENIB: in the context of autoencoder modeling, this architecture not only stabilizes the training but also opens up new ways to further improve their generation quality, *e.g.*, via *adversarial similarity* or by leveraging *pre-trained* representations, where the details can be found in Appendix G.2.

In Appendix C, we also present a nuisance-aware architecture applicable for ViT-based (Dosovitskiy et al., 2021) models, which is even simpler thanks to their patch-level representations available.

2.3 OVERALL TRAINING OBJECTIVE

Combining the proposed objectives as well as the original VIB loss, $L_{\texttt{VIB}}^{\beta}$ (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:

$$L_{\texttt{AENIB}} := L_{\texttt{VIB}}^{\beta} + \alpha \cdot L_{\texttt{recon}} + L_{\texttt{nuis}} + L_{\texttt{ind}}. \tag{11}$$

Algorithm 1 in Appendix A summarizes the overall procedure of AENIB training.

3 EXPERIMENTS

We verify the effectiveness of our proposed AENIB training for various aspects of out-of-distribution generalization compared to the standard training methods: specifically, we cover (a) novelty detection

Table 1: Comparison of AUROC (%; ↑) for various OOD detection methods trained on CIFAR-10 with five OOD datasets: SVHN, LSUN, ImageNet, CIFAR-100, and CelebA. Bolds indicate the best results.

Method	Score	SVHN	LSUN	ImageNet	C100	CelebA
JEM	$\log p(x)$	0.67	-	-	0.67	0.75
JEM	$\max_y p(y\|x)$	0.89	-	-	0.87	0.79
SupCon	$\max_y p(y\|x)$	0.97	0.93	0.91	**0.89**	-
Cross-entropy	$\max_y p(y\|x)$	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	$\max_y p(y\|x)$	0.95	0.94	0.92	0.88	0.76
VIB	$\log \mathrm{Dir}_{0.05}(y)$	0.97	0.96	0.94	0.88	0.78
AENIB (ours)	$\max_y p(y\|x)$	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 2: Comparison of test error rates (%; ↓) of ViT-S/4 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.

Method	C10	C10-C	C10.1	C10.2	CINIC
Cross-ent.	6.08	16.0	13.4	18.3	23.7
VIB	5.98	15.2	13.6	16.8	23.6
AugMix	6.52	15.1	14.2	17.2	24.2
PixMix	5.43	<u>10.3</u>	13.1	16.6	23.2
AENIB	<u>4.97</u>	12.3	<u>11.6</u>	<u>15.5</u>	<u>22.2</u>
+ AugMix	5.35	12.0	12.5	15.8	22.6
+ PixMix	**4.67**	**8.08**	**10.4**	**14.8**	**22.1**

Table 3: Comparison of OOD detection performances on the OBJECTS benchmark (Yang et al., 2022), 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.

FS-OOD: OBJECTS		AUROC (%; ↑) / AUPR (%; ↑) / FPR@TPR95 (%; ↓)			
Method	Score	MNIST	FashionMNIST	Texture	CIFAR-100-C
Cross-entropy	$\max_y p(y\|x)$	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	70.31 / 49.58 / 82.04	80.98 / 91.53 / **68.73**	70.14 / 89.97 / <u>72.91</u>	67.51 / 83.97 / 84.26
	Energy-based	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	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	75.69 / 76.61 / 99.70	79.40 / 93.14 / 93.72	<u>79.69</u> / <u>95.48</u> / 82.15	78.89 / 92.07 / 83.92
	$\log \mathrm{Dir}_{0.05}(y)$	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	$\max_y p(y\|x)$	80.23 / 73.50 / 80.69	76.35 / 91.22 / 84.75	74.67 / 94.09 / 87.22	76.12 / 91.03 / 84.99
	$\log \mathrm{Dir}_{0.05}(y)$	86.13 / 79.45 / 64.92	81.11 / 93.12 / 77.82	73.84 / 93.50 / 88.00	78.54 / 91.85 / <u>81.47</u>
AENIB (ours)	$\max_y p(y\|x)$	79.67 / 71.50 / 80.22	77.33 / 91.63 / 84.31	74.95 / 93.97 / 86.01	74.31 / 89.89 / 86.26
	$\log \mathrm{Dir}_{0.05}(y)$	<u>90.53</u> / <u>85.68</u> / <u>52.08</u>	<u>84.56</u> / <u>94.61</u> / <u>74.24</u>	75.04 / 93.83 / 86.01	<u>79.39</u> / <u>92.33</u> / 81.51
	$+\log \mathcal{N}(z_n; 0, I)$	**92.43 / 89.38 / 48.10**	**84.85 / 94.84** / 74.67	**88.91 / 97.49 / 48.44**	**82.66 / 93.62 / 74.14**

(Section 3.1), (b) corruption robustness (Section 3.2), and (c) adversarial robustness (Section 3.3) tasks which all have been challenging without assuming task-specific priors (Hendrycks et al., 2020; 2019a; Madry et al., 2018). We also present evaluations on the effectiveness of our proposed components in the context of unconditional generative modeling in Appendix G. We provide an ablation study in Appendix D for a component-wise analysis on the method. The full details on the experiments, *e.g.*, datasets, training details, and hyperparameters, can be found in Appendix B.

3.1 Out-of-distribution detection

We first show that our AENIB model can be 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 samples from in-distribution samples. A typical practice here is to assign a *score function* for each input based on the model, *e.g.*, the maximum confidence score (Hendrycks & Gimpel, 2017) as commonly used for supervised models, to threshold out samples as out-of-distribution when the score is low. To define a score function for our AENIB models, we first observe that the log-likelihood score of the nuisance representation $\mathbf{z}_n$, which is a unique information for AENIB, can be a strong score function especially for detecting novelties those are semantically far from in-distribution, *i.e.*, 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.[4]

We consider two evaluation benchmarks and compare ResNet-18 (He et al., 2016) models trained on CIFAR-10: (a) the *"standard"* benchmark, that has been actively adopted in the literature (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018b) assumes the test set of CIFAR-10 as in-distribution and measures the detection performance of other independent datasets; (b) the *OBJECTS*

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

Table 4: Comparison of average corruption error rates (%; ↓) per severity level on CIFAR-10/100-C (Hendrycks & Dietterich, 2019). Bold and underline denote the best and runner-up, respectively.

Architecture	Severity	CIFAR-10-C							CIFAR-100-C						
		Clean	1	2	3	4	5	Avg.	Clean	1	2	3	4	5	Avg.
ResNet-18	Cross-entropy	5.71	12.9	18.1	24.3	31.7	43.5	26.1	26.9	39.2	46.9	53.2	59.8	69.3	53.7
	VIB	**5.47**	**12.5**	17.5	23.6	30.7	42.5	25.4	**26.5**	39.7	47.5	53.8	60.5	70.1	54.3
	AENIB (ours)	7.07	13.2	**17.2**	**21.7**	**27.5**	**37.0**	**23.3**	28.0	**39.0**	**45.5**	**51.4**	**57.6**	**67.0**	**52.1**
ViT-S/4	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	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	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	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	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	**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 5: Comparison of test error rates (%; ↓) or mean corruption error (mCE, %; ↓) on ImageNet and its variants, namely ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-R (Hendrycks et al., 2021), and ImageNet-Sketch (Wang et al., 2019). Bold indicate the best results.

Architecture	Method	ImageNet	Corruption (mCE)	Rendition	Sketch
ViT-S/16	Cross-entropy	**25.1**	65.9	70.3	80.3
	AENIB (ours)	**25.1**	**65.2** (−0.7)	**67.1** (−3.2)	**77.7** (−2.6)
ViT-B/16	Cross-entropy	**21.8**	58.6	66.3	76.5
	AENIB (ours)	21.9	**57.5** (−1.1)	**64.4** (−1.9)	**74.4** (−2.1)

benchmark, recently proposed by Yang et al. (2022), further extends the CIFAR-10 benchmark to also consider "near" in-distribution in OOD evaluation. Specifically, OBJECTS assumes CIFAR-10-C (Hendrycks & Dietterich, 2019) and ImageNet-10 as in-distribution in test-time as well as CIFAR-10, making the detection task much more challenging as shown by Yang et al. (2022).

The results are reported in Table 1 and 3 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 1, the combined score achieves near-perfect AUROCs for detecting SVHN, LSUN and ImageNet datasets. Regarding Table 3, on the other hand, our method of AENIB shows even more significant improvements here: *e.g.*, AENIB improves the previous best AUROC (of Mahalanobis (Lee et al., 2018b)) on OBJECTS *vs.* MNIST from $77.04 \rightarrow 92.43$. This shows that both representation and score obtained from AENIB help to better discriminate in- *vs.* out-of-distribution in a more semantic sense compared to prior arts.

3.2 Robustness against natural corruptions

Next, we evaluate corruption robustness of our method, namely, the generalization ability of a representation in the situation that the given input can be distorted with natural corruptions (*e.g.*, fog, brightness, etc.) those are still semantic to humans. To this end, we consider a wide range of benchmarks those are constructed from CIFAR-10 and ImageNet for the purpose of measuring generalization. Namely, for CIFAR-10 models we test on (a) CIFAR-10/100-C (Hendrycks & Dietterich, 2019), a corrupted version of CIFAR-10/100 simulating 15 common corruptions in 5 severity levels, respectively, as well as (b) CIFAR-10.1 (Recht et al., 2018), CIFAR-10.2 (Lu et al., 2020), and CINIC-10 (Darlow et al., 2018), *i.e.*, three re-generations of the CIFAR-10 test set. For ImageNet models, on the other hand, we test ImageNet-C (Hendrycks & Dietterich, 2019), a corrupted version of ImageNet validation set, as well as ImageNet-R (Hendrycks et al., 2021), a collection of rendition images for 200 ImageNet classes, and ImageNet-Sketch (Wang et al., 2019). We test two different encoder architectures for CIFAR-10, namely ResNet-18 (He et al., 2016) and ViT-S (Dosovitskiy et al., 2021; Touvron et al., 2021), to also investigate the effect of architectures in AENIB. For the ImageNet experiments, on the other hand, we consider ViT-S and ViT-B (Touvron et al., 2021) to further examine the scalability of our method.

Table 2, 4, and 5 summarize the results. In Table 4, we observe that AENIB significantly and consistently improves corruption errors upon VIB in both architectures tested, and these gains are strong even compared with state-of-the-art methods: *e.g.*, AENIB can solely outperform a strong

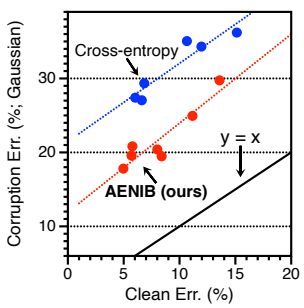
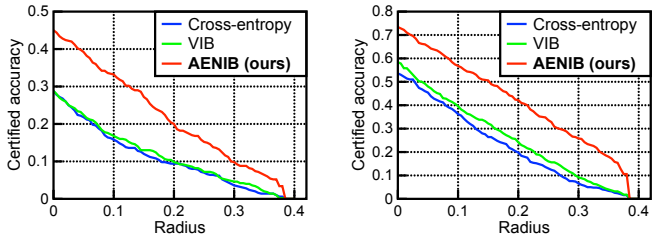


Figure 2: Comparison of trends in clean vs. corruption errors (against Gaussian) on ViT-S/4.



(a) ResNet-18 (b) ViT-S

Figure 3: Comparison of certified adversarial robust accuracy at various radii on CIFAR-10. The sharp drops in the plots are due to the statistical upper bound in radius the current certification protocol (Cohen et al., 2019) can output for a given σ.

baseline of AugMix (Hendrycks et al., 2020). Although a more recent method of PixMix (Hendrycks et al., 2022) 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. Interestingly, we observe that the impact of AENIB in the clean error can be different depending on the encoder architecture: with the ViT-S, AENIB could even further improve the clean errors compared to both Cross-entropy and VIB. This is possibly due to that the representation induced via AENIB can be extracted better with non-local (attention-based) operations.

Next, Table 5 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. Lastly, Figure 2 compares the linear trends made by Cross-entropy and AENIB across different data augmentations and hyperparameters, confirming that AENIB exhibits a better operating points even in terms of *effective robustness* (Taori et al., 2020), given the strong performance correlations recently observed between in- *vs.* many out-of-distribution benchmarks across models (Taori et al., 2020; Hendrycks et al., 2021; Miller et al., 2021).

3.3 Robustness against adversarial examples

We also evaluate adversarial robustness (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018) adopting the *randomized smoothing* framework (Cohen et al., 2019) that can measure a *certified* robustness for a given representation: specifically, any classifier can be "robustified" by averaging its predictions under Gaussian noise $\mathcal{N}(0, \sigma^2 I)$, where the robustness at input x depends on how consistent the classifier is on classifying $\mathcal{N}(x, \sigma^2 I)$ (Jeong & Shin, 2020). We adopt such a certified (or provable) protocol since it better aligns with our focus of testing robustness of representations that are not adversarially-trained (Madry et al., 2018): empirical robustness, *i.e.*, that reports the worst-case accuracy after directly attacking a classifier with diverse adversarial attacks, is usually hard to get a non-trivial accuracy without a thorough adversarial training. The randomized smoothing based evaluation, on the other hand, provides a more meaningful metric for classifiers even for the "Cross-entropy" baseline, while still representing a lower-bound in robustness that a given classifier can achieve (with an aid of randomized smoothing) against *every* adversarial attack method.

We follow the standard certification protocol (Cohen et al., 2019) 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 both ResNet-18 and ViT-S architectures on CIFAR-10, and assume $\sigma = 0.1$ for this experiment. The results summarized in Figure 3 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 $\varepsilon = 0.1$ with ViT-S. 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.

4 RELATED WORK

Out-of-distribution robustness. Since the seminal works (Szegedy et al., 2014; Nguyen et al., 2015; Amodei et al., 2016) revealing the fragility of neural networks for out-of-distribution inputs, there have been significant attempts on identifying and improving various notions of robustness: *e.g.*, detecting novel inputs (Hendrycks & Gimpel, 2017; Lee et al., 2018b;a; Tack et al., 2020), robustness against corruptions (Hendrycks & Dietterich, 2019; Geirhos et al., 2019; Hendrycks et al., 2020), and adversarial noise (Madry et al., 2018; Athalye et al., 2018; Cohen et al., 2019; Carlini et al., 2019), 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 (Hendrycks et al., 2019a; 2020), extra information from test-time samples (Wang et al., 2021), or specific knowledge in threat models (Tramer & Boneh, 2019; Kang et al., 2019). 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 (Lee et al., 2018b; Schott et al., 2019; Grathwohl et al., 2020; Yang & Ji, 2021). For example, Lee et al. (2018b; 2019) 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) (Grathwohl et al., 2020; Yang & Ji, 2021) 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* (Dinh et al., 2016; Kingma & Dhariwal, 2018; Behrmann et al., 2019; Grathwohl et al., 2019), where the nuisance can be defined by splitting the (full-information) encoding $\mathbf{z}$ for a given subspace of interest as explored by Jacobsen et al. (2019); Ardizzone et al. (2020). 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 Jacobsen et al. (2019). Other related works (Jaiswal et al., 2018; 2019; Pan et al., 2021) instead introduce a separate encoder for nuisance factors, although the notion of nuisance-ness has been focused as the independence to $\mathbf{z}$ (mostly for the purpose of disentangling), rather than to $\mathbf{y}$ as we focus in this work (for the purpose of robustness): *e.g.*, DisenIB (Pan et al., 2021) applies FactorVAE (Kim & Mnih, 2018) between semantic and nuisance embeddings to force their independence.[5] Yet, the literature has been also questioned on that the idea can be scaled-up beyond, *e.g.*, MNIST, and our work does explore and establish a practical design recent architectures and datasets addressing modern security metrics.

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

5 CONCLUSION

In this work, we suggest that having a good *nuisance model* can be a tangible approach to induce a robust representation. Specifically, we develop a practical method of learning deep nuisance representation from data, and show its effectiveness to improve various notions of model robustness under a challenging setup of assuming no prior (Taori et al., 2020). We believe our work can be a useful step towards better understanding of the robustness in deep neural networks. Although the scope of this paper currently focuses on a particular design of autoencoder based models, our framework of *nuisance-extended IB* is not limited to it and future works could consider a more diverse class of implementations, *e.g.*, a bi-directional GAN (Donahue et al., 2017) based design. Ultimately, we aim to approximate a challenging form of adversarial training with a mutual information constraint, which we believe will be a promising future direction to explore.

[5] We provide a more direct empirical comparison with DisenIB (Pan et al., 2021) with AENIB in Appendix H.

ETHICS STATEMENT

Securing reliable deep learning based models is arguably essential for *AI safety* (Amodei et al., 2016), especially for security-concerned systems (Caruana et al., 2015; Yurtsever et al., 2020). Nevertheless, one should also recognize that current techniques for assessing robustness in deep learning have a clear gap to the real-world, which should be considered deliberately to avoid any potentially biased, false sense of security in use.

REPRODUCIBILITY STATEMENT

We provide all the details to reproduce our experimental results in Appendix B (for training details, hyperparameters, and datasets) and Appendix C (for architectural details). We plan to publicly release our code and models upon publication of our manuscript.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

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

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.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7828–7840. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/593906af0d138e69f49d251d3e7cbed0-Paper.pdf.

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 (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

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

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

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.

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

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a89cf525e1d9f04d16ce31165e139a4b-Paper.pdf.

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.

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.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=RLRXCV6DbEJ.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

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, pp. 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1e0X3C9tQ.

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.

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74, 2021.

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.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP, 2016.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning, 2017.

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. URL https://openreview.net/forum?id=YicbFdNTTy.

Ethan Fetaya, Joern-Henrik Jacobsen, Will Grathwohl, and Richard Zemel. Understanding the limitations of conditional generative models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1lPleBFvH.

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. URL https://openreview.net/forum?id=Bygh9j09KX.

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.

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 (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

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.

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. URL https://openreview.net/forum?id=Hkxzx0NtDB.

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*, pp. 10696–10706, 2022.

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*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

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. URL https://openreview.net/forum?id=Hkg4TI9xl.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=HyxCxhRcY7.

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, 2019b.

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. URL https://openreview.net/forum?id=S1gmrxHFvB.

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*, pp. 8340–8349, October 2021.

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*, pp. 16783–16792, 2022.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

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. URL https://proceedings.neurips.cc/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

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

Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkfbpsAcF7.

Jörn-Henrik Jacobsen, Arnold W.M. Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJsjkMb0Z.

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

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

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-z: Zero-shot task generalization with robotic imitation learning. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=8kbp23tSGYv.

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

Jongheon Jeong and Jinwoo Shin. Training GANs with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eo6U4CAwVmg.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

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

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*, pp. 4401–4410, 2019.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12104–12114. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf.

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*, pp. 8110–8119, 2020b.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf.

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.

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.

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

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

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

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*, pp. 3581–3590. PMLR, 2019.

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. ISSN 1558-2256. doi: 10.1109/5.726791.

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*, 2018a. URL https://openreview.net/forum?id=ryiAv2xAZ.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf.

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 (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3763–3772. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19f.html.

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. URL https://openreview.net/forum?id=H1VGkIxRZ.

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. URL https://openreview.net/forum?id=1Fqg133qRaI.

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.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

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.

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. URL https://openreview.net/forum?id=rJzIBfZAb.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

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

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. In *European Conference on Computer Vision*, 2020.

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*, pp. 7721–7735. PMLR, 2021.

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

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. URL https://openreview.net/forum?id=H1xwNhCcYm.

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.

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

Ziqi Pan, Li Niu, Jianfu Zhang, and Liqing Zhang. Disentangled information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9285–9293, 2021.

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*, pp. 823–832, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018. URL https://arxiv.org/abs/1806.00451.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/1e79596878b2320cac26dd792a6c51c9-Paper.pdf.

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. doi: 10.1007/s11263-015-0816-y.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=fUxqIofPPi.

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. URL https://openreview.net/forum?id=S1EHOsC9tX.

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. URL https://openreview.net/forum?id=SyxIWpVYvr.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

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.

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.

R Thobaben, M Skoglund, et al. The convex information bottleneck lagrangian. *Entropy*, 22(1), 2020.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop*, pp. 1–5, 2015. doi: 10.1109/ITW.2015.7133169.

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*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

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. doi: 10.1109/TPAMI.2008.128.

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*, pp. 10347–10357. PMLR, 2021.

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

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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*, pp. 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

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. URL https://openreview.net/forum?id=uXl3bZLkr3c.

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*, pp. 10506–10518, 2019.

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.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20685–20696. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/eddea82ad2755b24c4e168c5fc2ebd40-Paper.pdf.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2204.05306*, 2022.

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

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. URL https://openreview.net/forum?id=pfNyExj7z2.

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*, pp. 6023–6032, 2019.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.

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*, pp. 7472–7482. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

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

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.

A TRAINING PROCEDURE OF AENIB

Algorithm 1 Autoencoder-based nuisance-extended information bottleneck (AENIB)

Require: encoder f, decoder g, discriminators d, prior $p_0(\mathbf{z})$, $\alpha, \beta, \tau > 0$.

for # training iterations **do**
  Sample $(x_i, y_i)_{i=1}^m \sim p_d(\mathbf{x}, \mathbf{y})$
  $\mathbf{z}^{(i)}, \mathbf{z}_n^{(i)} \leftarrow f(x_i)$, and sample $z^{(i)}, z_n^{(i)} \sim \mathbf{z}^{(i)}, \mathbf{z}_n^{(i)}$
  $\hat{x}_i \leftarrow g(z^{(i)}, z_n^{(i)})$
  // UPDATE DISCRIMINATORS
  $L_{\texttt{ind}} \leftarrow \mathbb{E}_{\mathbf{z},\mathbf{z}_n \sim \mathcal{N}(0,I)}[\log d_{\mathbf{z}}(\mathbf{z}, \mathbf{z}_n)] + \frac{1}{m}\sum_i \log(1 - d_{\mathbf{z}}(z, z_n))$
  $L^D_{\texttt{nuis}} \leftarrow \frac{1}{m}\sum_i \mathbb{CE}(q_n(\mathbf{y}|z_n^{(i)}), y_i)$
  $L_D \leftarrow L^D_{\texttt{nuis}} - L_{\texttt{ind}}$
  $d_{\mathbf{z}}, q_n \leftarrow$ Update $d_{\mathbf{z}}, q_n$ to minimize L_D
  // UPDATE ENCODER AND DECODER
  $L^{\beta}_{\texttt{VIB}} \leftarrow \frac{1}{m}\sum_i [-\log q(y_i|z_i) + \beta \mathrm{KL}(p(\mathbf{z}|x_i)\|p_0(\mathbf{z}))]$
  $L_{\texttt{recon}} \leftarrow \frac{1}{m}\sum_i \frac{1}{2}\|x_i - \hat{x}_i\|_2^2$
  $L_{\texttt{nuis}} \leftarrow \frac{1}{m}\sum_i \mathbb{CE}(q_n^*(\mathbf{y}|z_n^{(i)}), \frac{1}{|\mathcal{Y}|})$
  $L_{\texttt{AENIB}} \leftarrow L^{\beta}_{\texttt{VIB}} + \alpha L_{\texttt{recon}} + L_{\texttt{nuis}} + L_{\texttt{ind}}$
  $f, g, \Pi_f \leftarrow$ Update f, g, Π_f to minimize $L_{\texttt{AENIB}}$
end for

B EXPERIMENTAL DETAILS

B.1 TRAINING DETAILS

Unless otherwise noted, we train each model for 200K updates for CIFAR-10 models, and 1M updates for ImageNet models. For training AENIB models, we use $\alpha = 10.0, \beta = 0.0001$ unless otherwise noted. We use different training configurations depending on the encoder architecture, *i.e.*, whether is it ResNet or ViT: (a) For ResNet-based models, we train the encoder part (f) via stochastic gradient descent (SGD) with batch size of 64 using Nesterov momentum of weight 0.9 without dampening. We set a weight decay of 10^{-4}, and use the cosine learning rate scheduling (Loshchilov & Hutter, 2016) from the initial learning rate of 0.1. For the remainder parts of our AENIB architecture, *e.g.*, the decoder g and discriminator MLPs, on the other hand, we follow the training practices of GAN instead: specifically, we use Adam (Kingma & Ba, 2015) with $(\alpha, \beta_1, \beta_2) = (0.0002, 0.5, 0.999)$, following the hyperparameter practices explored by Kurach et al. (2019). (b) For ViT-based models, on the other hand, we train both (transformer-based) encoder and decoder models via AdamW (Loshchilov & Hutter, 2019) with a weight decay of 10^{-4}, using batch size 128 and $(\alpha, \beta_1, \beta_2) = (0.0002, 0.9, 0.999)$ with the cosine learning rate scheduling (Loshchilov & Hutter, 2016). We use 2K and 100K steps of a linear warm-up phase in learning rate for CIFAR and ImageNet models, respectively. Overall, we observe that a stable training of ViT (even for CIFAR-10) requires much stronger regularization compared to ResNets, otherwise they often significantly suffer from overfitting. In this respect, we apply the regularization practices those are now widely used for ViTs on ImageNet, namely mixup (Zhang et al., 2018), CutMix (Yun et al., 2019), and RandAugment (Cubuk et al., 2020), following those established in Beyer et al. (2022).

B.2 DATASETS

CIFAR-10/100 datasets (Krizhevsky, 2009) consist of 60,000 images of size 32×32 pixels, 50,000 for training and 10,000 for testing. Each of the images is labeled to one of 10 and 100 classes, and the number of data per class is set evenly, *i.e.*, 6,000 and 600 images per each class, respectively. By default, we use the random translation up to 4 pixels as a data pre-processing. We normalize the images in pixel-wise by the mean and the standard deviation calculated from the training set. The full dataset can be downloaded at https://www.cs.toronto.edu/~kriz/cifar.html.

CIFAR-10/100-C, and ImageNet-C datasets (Hendrycks & Dietterich, 2019) are collections of 75 replicas of the CIFAR-10/100 test datasets (of size 10,000) and ImageNet validation dataset (of size 50,000), respectively, which consists of 15 different types of common corruptions each of which contains 5 levels of corruption severities. Specifically, the datasets includes the following corruption types: (a) *noise*: Gaussian, shot, and impulse noise; (b) *blur*: defocus, glass, motion, zoom; (c) *weather*: snow, frost, fog, bright; and (d) *digital*: contrast, elastic, pixel, JPEG compression. In our experiments, we evaluate test errors on CIFAR-10/100-C for models trained on the "clean" CIFAR-10/100 datasets, where the error values are averaged across different corruption types per severity level. For ImageNet-C, on the other hand, we compute and compare the mean Corruption Error (mCE) proposed by Hendrycks & Dietterich (2019). Specifically, mCE is the average of Corruption Error (CE) over corruption types, where CE is defined by the error rates normalized by those from AlexNet (Krizhevsky et al., 2012) to adjust varying difficulties across corruption types. Formally, for a classifier f, CE for a specific corruption type c is defined by:

$$\mathrm{CE}_c^f := \left(\sum_{s=1}^{5} \texttt{error}_{c,s}^{f}\right) \Bigg/ \left(\sum_{s=1}^{5} \texttt{error}_{c,s}^{\text{AlexNet}}\right), \tag{12}$$

where s denotes the severity level ($1 \leq s \leq 5$). The full datasets, as well as the information on the pre-computed AlexNet error rates on ImageNet-C (to compute mCE), can be downloaded at https://github.com/hendrycks/robustness.

CIFAR-10.1/10.2 datasets (Recht et al., 2018; Lu et al., 2020) are reproductions of the CIFAR-10 test set that are separately collected from Tiny Images dataset (Torralba et al., 2008). Both datasets consist 2,000 samples for testing, and designed to minimize distribution shift relative to the original CIFAR-10 dataset in their data creation pipelines. The datasets can be downloaded at https://github.com/modestyachts/CIFAR-10.1 (for CIFAR-10.1; we use the "v6" version) and https://github.com/modestyachts/cifar-10.2 (for CIFAR-10.2).

CINIC-10 dataset (Darlow et al., 2018) is an extension of the CIFAR-10 dataset generated via addition of down-sampled ImageNet images. The dataset consists of 270,000 images in total of size 32×32 pixels, those are equally distributed for train, validation and test splits, *i.e.*, the test dataset (that we use for our evaluation) consists of 90,000 samples. Due to the discrepancy in distributions between CIFAR and ImageNet, CINIC-10 by design contains a more significant distribution shift compared to CIFAR-10.1/10.2, thus is more challenging when considered as a generalization benchmark from CIFAR-10. The full datasets can be downloaded at https://github.com/BayesWatch/cinic-10.

ImageNet dataset (Russakovsky et al., 2015), also known as ILSVRC 2012 classification dataset, consists of 1.2 million high-resolution training images and 50,000 validation images, which are labeled with 1,000 classes. As a data pre-processing step, we perform a 256×256 resized random cropping and horizontal flipping for training images. For testing images, on the other hand, we apply a 256×256 center cropping for testing images after re-scaling the images to have 256 in their shorter edges. Similar to CIFAR-10, all the images are normalized by the pre-computed mean and standard deviation. A link for downloading the full dataset can be found in http://image-net.org/download.

ImageNet-R dataset (Hendrycks et al., 2021) consists of 30,000 images of various artistic renditions for 200 (out of 1,000) ImageNet classes: *e.g.*, art, cartoons, deviantart, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, video game renditions, and so on. To perform an evaluation of ImageNet classifiers on this dataset, we apply masking on classifier logits for the 800 classes those are not in ImageNet-R. The full dataset can be downloaded at https://github.com/hendrycks/imagenet-r.

ImageNet-Sketch dataset (Wang et al., 2019) consists of 50,000 sketch-like images, 50 images for each of the 1,000 ImageNet classes. The dataset is constructed with Google image search, using queries of the form "`sketch of [CLS]`" within the "black and white" color scheme, where `[CLS]` is the placeholder for class names. The full dataset as well as the scripts to collect the dataset can be accessed at https://github.com/HaohanWang/ImageNet-Sketch.

CelebFaces Attributes (CelebA) dataset (Liu et al., 2015) consists of 202,599 face images, where each is labeled with 40 attribute annotations. We follow the standard train/validation/test splits of the dataset as provided by Liu et al. (2015), and use the train split for training and computing FID scores following the protocol of other baselines (Parmar et al., 2021; Aneja et al., 2021). We also follow the pre-processing procedure of (Liu et al., 2015) to fit in the images into the size of 64×64:

namely, we first perform a center crop into size 140×140 to the images, followed by a resizing operation into 64×64. The full dataset can be downloaded at https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

B.3 COMPUTING INFRASTRUCTURE

Unless otherwise noted, we use a single NVIDIA Geforce RTX-2080Ti GPU to execute each of the experiments. For experiments based on StyleGAN2 architecture (reported in Table 8), we use two NVIDIA Geforce RTX-2080Ti GPUs per run. For the ImageNet experiments (reported in Table 5 in the main text), we use 8 NVIDIA Geforce RTX-3090 GPUs per run.

C ARCHITECTURAL DETAILS

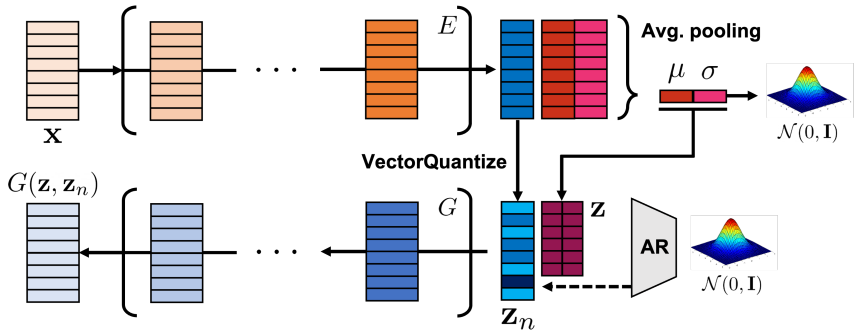


Figure 4: An overview of our proposed framework, *nuisance-extended information bottleneck* (NIB), instantiated by an autoencoder-based design with Transformer-based architectures.

Recall that our proposed AENIB architecture consists of (a) an encoder f, (b) a decoder g, and (c) MLP-based discriminators $d_{\mathbf{y}}$, $d_{\mathbf{z}}$, and an MLP for feature statistic projection Π_f. We set 128 as the nuisance dimension $\mathbf{z}_n$, and use hidden layer of size 1,024 for MLP-based discriminators, *e.g.*, $d_{\mathbf{y}}$, $d_{\mathbf{z}}$, and MLPs for projection Π_f.

ConvNet-based architectures. We mainly consider ResNet-18 (He et al., 2016) as a ConvNet-based encoder. For this encoder, we consider the generator architecture of FastGAN (Liu et al., 2021) as the decoder, but with a modification on normalization layers: specifically, we replace the standard batch normalization (Ioffe & Szegedy, 2015) layers in the architecture with adaptive instance normalization (AdaIN) (Karras et al., 2019) so that the affine parameters can be modulated by $\mathbf{z}$ and $\mathbf{z}_n$ as well as the decoder input: we observe a consistent gain in FID from this modification.

Transformer-based architectures. We consider ViT-S and ViT-B (Dosovitskiy et al., 2021; Touvron et al., 2021) in our experiments. When Transformer-based encoder is used, we use the same Transformer architecture as the decoder model where it is preceded by linear layers that maps both $\mathbf{z}$ and $\mathbf{z}_n$ into the space of patch embedding. We assume the patch size of ViT to be 4 for CIFAR-10 and 16 for ImageNet, *i.e.*, we denote it as ViT-S/4 and ViT-S/16, respectively, so that the outputs from the models have similar numbers of patch embeddings (8×8 and 16×16, respectively) to those of ResNet-18. To model $\mathbf{z}$ and $\mathbf{z}_n$ in the ViT architecture, we simply split the output patch embedding into two separate embeddings (of reduced embedding dimensions): one of these embeddings is average-pooled to define $\mathbf{z}$, and the remaining one is vector-quantized (Yu et al., 2022) to define a nuisance representation $\mathbf{z}_n$, as described in the next paragraph in more details. Figure 4 illustrates an overview of our proposed AENIB for ViT-based architectures.

VQ-based nuisance modeling for ViT-AENIB. Remark that our current ViT-based design allocates *different* numbers of feature dimension for $\mathbf{z}$ and $\mathbf{z}_n$: specifically, we only apply global average pooling for $\mathbf{z}$ (not for $\mathbf{z}_n$), so that its dimensionality becomes independent to the input resolution, while the nuisance $\mathbf{z}_n$ would still get an increasing dimensionality for higher-resolution inputs. In practice, this difference in feature dimension may cause some training difficulties in AENIB training: (a) it makes harder to balance between the objectives given that AENIB is essentially a "competition" between two information channels, *i.e.*, $\mathbf{z}$ and $\mathbf{z}_n$; also, (b) it becomes increasingly difficult to force $\mathbf{z}_n$ to follow the independent Gaussian marginal for a tractable sampling as $\mathbf{z}_n$ gets higher

dimensions, unlike our ConvNet-based design. To alleviate these issues, we propose to apply the *vector quantization* (VQ) (van den Oord et al., 2017; Yu et al., 2022) to the nuisance embedding: namely, we train the output of the nuisnace head, say $\hat{\mathbf{z}}_n$, to have one of (discrete) vectors in a learned dictionary $\mathbf{e}$. With this, now $\mathbf{z}_n$ becomes independent to the dimensionality of per-patch embeddings, which allows a more scalable balancing with $\mathbf{z}$, as well as offering a tractable marginal distribution to sample: given that $\mathbf{z}_n$ now gets a sequence of discrete distribution, one can apply a post-hoc generative modeling (*e.g.*, with an autoregressive prior (van den Oord et al., 2017), or with a diffusion model (Gu et al., 2022)) to allow an efficient sampling. Specifically, we add the following objective upon our proposed AENIB objective (11) to enable VQ-based nuisance modeling:

$$L_{\texttt{VQ}}(\mathbf{x};\mathbf{e}) := \|\texttt{sg}[\hat{\mathbf{z}}_n(\mathbf{x})] - e\|_2^2 + \beta\|\hat{\mathbf{z}}_n(\mathbf{x}) - \texttt{sg}[e]\|_2^2, \tag{13}$$

where $e := \min_i \|\hat{\mathbf{z}}_n(\mathbf{x}) - e_i\|_2^2$, and $\texttt{sg}(\cdot)$ denotes the stop-gradient operator defined by $\texttt{sg}(x) \equiv x$ and $\frac{d}{dx}\texttt{sg}(x) \equiv 0$. Here, the commitment hyperparameter β is set to 0.25 following (van den Oord et al., 2017; Yu et al., 2022). We allocate 32 per-patch dimensions for $\mathbf{z}_n$, with an embedding dictionary $\mathbf{e}$ of size 256. We adopt the embedding normalization (Yu et al., 2022) as we found it consistently improves the stability of VQ-based training.

SSIM-based D_2^2 reconstruction loss. Recall that our proposed AENIB is based on minimizing reconstruction loss (5) to implement $I(\mathbf{x};\mathbf{z},\mathbf{z}_n)$ in NIB (4). Although we introduce the *normalized mean-squared error* (NMSE) as a default design choice, the choice may not be limited to that: here, we demonstrate a SSIM-based (Wang et al., 2004) reconstruction loss as an alternative, and show its effectiveness on improving corruption robustness. Specifically, for a given pair of images[6] (x, y), the *structural similarity index measure* (SSIM) defines a similarity metric between x and y considering differences in luminance (represented by S_1) and structures (represented by S_2):

$$\text{SSIM}(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \cdot \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} =: S_1 \cdot S_2, \tag{14}$$

where $c_1 := 0.01^2$ and $c_2 := 0.03^2$ are small constants for numerical stability, as well as to simulate the saturation effects of visual system under low luminance (and contrast) (Brunet et al., 2011). Given that SSIM itself is not a distance metric (*e.g.*, it often allows the value to be negative), however, we instead consider the following modification of SSIM, the *squared-D_2* (D_2^2), as our reconstruction loss, which is originally defined by Brunet et al. (2011), and shown to be a distance

$$D_2^2(x,y) := (1 - S_1) + (1 - S_2) = 2 - S_1 - S_2. \tag{15}$$

In Table 6, we compare the effect of having different reconstruction losses in AENIB between the default choice of NMSE and D_2^2: the results on CIFAR-10/100-C with ViT-S/4 show that D_2^2-based reconstruction loss can reliable improve corruption robustness of the AENIB models over NMSE. This confirms that the choice of reconstruction loss impacts the final robustness of AENIB, and also suggests that a more perceptually-aligned similarity metric could possibly make the model less biased toward spurious features that are not necessary to build a robust representation.

In this respect, we adopt the D_2^2-based loss in AENIB for ViT-based models in our experiments: somewhat interestingly, we found the objective becomes much harder to be minimized for ConvNet-based models, where we keep the default choice of NMSE. This is possibly because that there can be a discrepancy between what ConvNets typically extract and those from a D_2^2-based reconstruction.

Table 6: Comparison of average per-corruption error rates (%; ↓) on CIFAR-10/100-C (Hendrycks & Dietterich, 2019). We use ViT-S/4 for this experiment. All the models reported here are trained via AENIB but with different reconstruction losses.

AENIB (ViT-S/4)	Loss	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*JPEG*	*Average*
CIFAR-10-C	NMSE	20.0	15.8	19.2	**10.2**	**18.3**	**12.7**	**11.9**	9.76	10.0	11.0	6.33	11.3	**10.5**	**13.4**	14.7	13.0
	D_2^2	**17.8**	**14.0**	**17.6**	**10.2**	19.7	13.2	12.4	**9.12**	**8.87**	**9.56**	**5.86**	**8.29**	10.7	14.3	**13.5**	**12.3**
CIFAR-100-C	NMSE	48.2	42.8	43.2	32.4	48.6	36.1	35.6	31.4	32.3	35.2	25.3	33.7	33.2	36.2	40.9	36.9
	D_2^2	**45.7**	**40.1**	**42.1**	**32.2**	**47.9**	**35.8**	**35.1**	**31.1**	**31.7**	**34.1**	**24.4**	**31.5**	**32.4**	**34.5**	**38.4**	**35.8**

[6]In practice, SSIM is often computed in per-patch basis for a sliding window of a certain kernel size, *e.g.*, 8. The values are then averaged to define the metric. In our experiments, we also follow this implementation.

D ABLATION STUDY

Figure 5: Reconstructions under random nuisance $\mathbf{z}_n$. The leftmost per row shows the original reconstruction.

Table 7: Comparison of the test error rate (Err.; %, ↓), corruption error (C-Err.; %, ↓) and FID on CIFAR-10 across ablations.

β	$L_{\texttt{recon}}$	$L_{\texttt{nuis}}$	$L_{\texttt{ind}}$	$L_{\texttt{sim}}$	Err.	C-Err.	FID
1e-4	✓	✓	✓	✓	7.07	23.3	33.3
1e-3	✓	✓	✓	✓	7.32	24.5	31.0
1e-2	✓	✓	✓	✓	7.38	26.3	30.8
1e-4	✗	✓	✓	✓	8.29	29.2	33.8
1e-4	✓	✗	✓	✓	8.01	24.1	29.2
1e-4	✓	✓	✗	✓	7.31	22.4	78.3
1e-4	✓	✓	✓	✗	7.95	28.6	83.1

We further perform an ablation study on CIFAR-10 for a detailed analysis of the proposed AENIB:

Effect of β. As also introduced in the original IB objective, $\beta \geq 0$ plays the key role in AENIB training as it controls the information balance between the semantic $\mathbf{z}$ and the nuisance $\mathbf{z}_n$. Here, Figure 5 examine how using different value of β affect the actual representations, by comparing the reconstructed samples for a fixed input while randomizing the nuisance $\mathbf{z}_n$. Indeed, we observe a clear trend from this comparison demonstrating the effect of β: having larger β makes the model to push more "semantic" information into $\mathbf{z}_n$ regarding it as the nuisance. Without information bottleneck, *i.e.*, in case when $\beta = 0.0$, we qualitatively observe that the network rather encodes most information in $\mathbf{z}$, due to the minimax loss applied to the nuisance $\mathbf{z}_n$. Quantitatively, this behavior is further evidenced in Table 7 as an increase in the corruption errors when using larger β.

Reconstruction loss. The reconstruction loss $L_{\texttt{recon}}$ is one of essential part to make AENIB work as a "nuisance modeling": in Table 7, we provide an ablation when this loss is omitted, showing a significant degradation in the final accuracy, and more crucially in the corruption error. This confirms the necessity of reconstruction loss to obtain a robust representation in AENIB. Nevertheless, due to the adversarial similarity loss $L_{\texttt{sim}}$ that can also work (while not perfectly) as a reconstruction loss, one can still observe that the FID of the model can be moderately preserved.

Nuisance loss. From the ablation of $L_{\texttt{nuis}}$ given in Table 7, we observe not only a considerable degradation in clean accuracy but also in its corruption robustness. This shows that strictly forcing the nuisance-ness to $\mathbf{z}_n$ (against $\mathbf{y}$) indeed helps $\mathbf{z}$ to learn a more robust representation, possibly from encouraging $\mathbf{z}$ to extract more diverse class-related information in a faithful manner by keeping the remainder information in $\mathbf{z}_n$ sufficient to infer $\mathbf{x}$.

Independence loss. The independence loss $L_{\texttt{ind}}$ in our current design, which essentially performs a GAN training toward $p(\mathbf{z}, \mathbf{z}_n) \sim \mathcal{N}(0, I)$, not only forces $\mathbf{z} \perp \mathbf{z}_n$ but also leads $\mathbf{z}$ and $\mathbf{z}_n$ to have a tractable marginal distribution: so that one could efficiently perform a sampling from the learned decoder. In a practical aspect, therefore, omitting $L_{\texttt{ind}}$ in AENIB can directly harm its generation quality as given in Table 7. Nevertheless, it is still remarkable that the ablation could rather improve the corruption error: this suggests that our current design of forcing the full Gaussian may be restrictive. An alternative design for the future work could assume a weaker condition for $\mathbf{z}$ and $\mathbf{z}_n$, instead with a more sophisticated sampling to obtain a valid generative model from AENIB.

Adversarial similarity. When the $L_{\texttt{sim}}$ is omitted, we observe a significant degradation in FID rather than accuracy, showing the effectiveness of our proposed adversarial similarity based guidance to improve decoder performance while affecting less to the accuracy compared to the case when $L_{\texttt{recon}}$ is ablated. It is quite remarkable that there is still a degradation in both clean and corruption accuracies compared to the case when $L_{\texttt{sim}}$ is jointly minimized: we observe that in this scenario of missing $L_{\texttt{sim}}$, the overall reconstruction loss $L_{\texttt{recon}}$ is often also less optimized, which could eventually affect the quality of $\mathbf{z}$.

E PROOF OF LEMMA 1

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 a representation $[\hat{\mathbf{z}}, \hat{\mathbf{z}}_n] := f(\hat{\mathbf{x}})$ 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 $I(\hat{\mathbf{z}};\mathbf{y}) = I(\mathbf{x};\mathbf{y})$.*

Proof. Given that f is invertible for the random variable $\hat{\mathbf{x}}$, the statement follows from the chain rule of mutual information and that of conditional mutual information, as well as by applying (b) and (c):

$$
\begin{aligned}
I(\mathbf{x};\mathbf{y}) = I(\hat{\mathbf{x}};\mathbf{y}) &= I(\mathbf{y};\hat{\mathbf{z}},\hat{\mathbf{z}}_n) = I(\mathbf{y};\hat{\mathbf{z}}_n) + I(\mathbf{y};\hat{\mathbf{z}}|\hat{\mathbf{z}}_n) && (16)\\
&= I(\mathbf{y};\hat{\mathbf{z}}) + H(\hat{\mathbf{z}}_n|\mathbf{y}) + H(\hat{\mathbf{z}}_n|\hat{\mathbf{z}}) - H(\hat{\mathbf{z}}_n|\mathbf{y},\hat{\mathbf{z}}) - H(\hat{\mathbf{z}}_n) && (17)\\
&= I(\mathbf{y};\hat{\mathbf{z}}) = I(\hat{\mathbf{z}};\mathbf{y}). && (18)
\end{aligned}
$$

□

F ADDITIONAL BACKGROUND

F.1 DETAILED SURVEY ON RELATED WORK

Out-of-distribution robustness. Since the seminal works (Szegedy et al., 2014; Nguyen et al., 2015; Amodei et al., 2016) revealing the fragility of neural networks for out-of-distribution inputs, there have been significant attempts on identifying and improving various notions of robustness: *e.g.*, detecting novel inputs (Hendrycks & Gimpel, 2017; Lee et al., 2018b; Hendrycks et al., 2019a;b; Lee et al., 2018a; Tack et al., 2020; Xiao et al., 2020), robustness against corruptions (Hendrycks & Dietterich, 2019; Geirhos et al., 2019; Hendrycks et al., 2020; Wang et al., 2021; Diffenderfer et al., 2021), and adversarial noise (Goodfellow et al., 2015; Madry et al., 2018; Athalye et al., 2018; Zhang et al., 2019; Cohen et al., 2019; Carlini et al., 2019), 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.*, *Outlier Exposure* (Hendrycks et al., 2019a) and *AugMix* (Hendrycks et al., 2020) assume an external dataset or a pipeline of data augmentations to improve the performances in novelty detection and corruption robustness, respectively; *Tent* (Wang et al., 2021) leverages extra information available from a batch of samples in test-time to adapt a given neural network; Tramer & Boneh (2019); Kang et al. (2019) observe that neural networks robust to a certain type of adversarial attack (*e.g.*, an ℓ_∞-constrained adversary) do not necessarily robust to other types of adversary (*e.g.*, an ℓ_1 adversary), *i.e.*, adversarial robustness hardly generalizes from the adversary assumed *a priori* for training. 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.

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 (Lee et al., 2018b; Schott et al., 2019; Grathwohl et al., 2020; Yang & Ji, 2021): for example, Lee et al. (2018b; 2019) have shown that assuming a simple Gaussian mixture model on the deep discriminative representations can improve novelty detection and robustness to noisy labels, respectively; Schott et al. (2019) develop an empirical defense against adversarial examples via generative classifiers; A line of research on *Joint Energy-based Models* (JEM) (Grathwohl et al., 2020; Yang & Ji, 2021) assumes the entire discriminative model as a joint generative model by interpreting the logits of $p(\mathbf{y}|\mathbf{x})$ as unnormalized log-densities of $p(\mathbf{x}|\mathbf{y})$, and shows that modeling $p(\mathbf{x}|\mathbf{y})$ as well as $p(\mathbf{y}|\mathbf{x})$ can improve out-of-distribution generalization of the classifier. Nevertheless, it is still an unexplored and open question that how to "better" incorporate generative representation into discriminative models: in case of novelty detection, for example, several recent works (Nalisnick et al., 2019; Ren et al., 2019; Serra et al., 2020; Xiao et al., 2020) observe that existing likelihood-based generative models are not accurate enough to detect out-of-distribution datasets, suggesting that relying solely on (likelihood-based) deep generative representation may not enough for robust classification (Fetaya et al., 2020). In case of JEM, on the other hand, it has been shown that directly assuming a joint generative-discriminative representation often makes a significant training instability. In this work, we propose to introduce an autoencoder-based model to avoid the training instability, and consider a design that the *nuisance* can succinctly supplement the given discriminative representation to be generative.

Invertible representations and nuisance modeling. The idea of incorporating nuisances can be also considered in the context of *invertible* modeling, or as known as *flow-based models* (Dinh et al., 2016; Kingma & Dhariwal, 2018; Jacobsen et al., 2018; Behrmann et al., 2019; Chen et al., 2019; Grathwohl et al., 2019),[7] which maps a given input $\mathbf{x}$ into a representation $\mathbf{z}$ of the same dimension so that one can construct an inverse of $\mathbf{z}$ to $\mathbf{x}$: here, the nuisance can be naturally defined as the remainder information of $\mathbf{z}$ for a given subspace of interest, *e.g.*, to model $\mathbf{y}$. For example, Jacobsen et al. (2019) adopt a fully-invertible variant of i-RevNet (Jacobsen et al., 2018) to analyze *excessive invariance* in neural networks, *i.e.*, the existence of pairs of completely different samples with the same representation in a neural network, and proposes to maximize the cross-entropy for the nuisances in a similar manner to our proposed minimax-based nuisance loss ((6) in the main text); Ardizzone et al. (2020), on the other hand, leverages invertible neural network to model a Gaussian mixture based generative classifier in the representation space, so that nuisance information can be preserved until its representation. Compared to such approaches relying on invertible neural networks, our autoencoder-based nuisance modeling does not guarantee the "full" invertibility for arbitrary inputs: instead, it only focuses on inverting the data manifold given, and this enables (a) a much flexible encoder design in practice, *i.e.*, other than flow-based designs, and (b) a more scalable generative modeling of nuisance representation $\mathbf{z}_n$ while forcing its *independence* to the semantic space $\mathbf{z}$. This is due to that it works on a compact space rather than those proportional to the input dimension, which is an important benefit of our modeling in terms of the scalability of nuisance-aware training, *e.g.*, beyond an MNIST-scale as done by Jacobsen et al. (2019). More closer related works (Jaiswal et al., 2018; 2019; Pan et al., 2021) in this respect instead introduce a separate encoder for nuisance factors, where the nuisanceness is induced by the independence to $\mathbf{z}$: *e.g.*, DisenIB (Pan et al., 2021) applies FactorVAE (Kim & Mnih, 2018) between semantic and nuisance embeddings to force their independence.[8] Yet, similarly to the invertible approach, the literature has been questioned on that the idea can be scaled-up beyond, *e.g.*, MNIST, and our work does explore and establish a practical design that is applicable for recent architectures and datasets addressing modern security metrics, *e.g.*, corruption robustness. On the technical side, for example, we find that the "nuisanceness to $\mathbf{y}$" is more important for $\mathbf{z}_n$ than the "independence with $\mathbf{z}$" (as usually done in the previous works (Jaiswal et al., 2018; 2019; Pan et al., 2021)) to induce a robust representation, as verified in our ablation study in Appendix D, which can be a useful practice for the future research concerning robust representation learning.

Autoencoder-based generative models. There have been steady advances in generative modeling based on autoencoder architectures, especially since the development in *variational autoencoders* (VAEs) (Kingma & Welling, 2014): due to its ability of estimating data likelihoods, and its flexibility to implement various statistical assumptions (Louizos et al., 2015; Kingma et al., 2016; Kim & Mnih, 2018). With the advances in its training objectives (Vincent et al., 2008; Makhzani et al., 2015; Higgins et al., 2016) as well as the architectural improvements (Vahdat & Kautz, 2020; Child, 2021), VAE-based models are currently considered as one of state-of-the-art approaches in likelihood based generative modeling: *e.g.*, a state-of-the-art *diffusion models* (Ho et al., 2020; Song et al., 2021) is built upon the denoising autoencoders under Gaussian perturbations, and recently-proposed *hierarchical VAEs* (Vahdat & Kautz, 2020; Child, 2021) have shown that VAEs can benefit from scaling up its architectures into deeper encoder networks. In perspectives of viewing our method as a *generative modeling*, AENIB is based on *adversarial autoencoders* (Makhzani et al., 2015) that replaces the KL-divergence based regularization in standard VAEs with a GAN-based adversarial loss, with a novel encoder architecture that is based on the *internal feature statistics* of discriminative models: so that the model can better encode lower-level features without changing the backbone architecture. We observe that this design enables autoencoder-based modeling even from a large, pre-trained discriminative models, and this "projection" of internal features can significantly benefit the generation quality, as well as for generative adversarial networks (GANs) as observed in Table 8.

F.2 Technical background

Variational information bottleneck. Although the information bottleneck (IB) principle given in (1) (Tishby et al., 1999) suggests a useful definition on what we mean by a "good" representation, computing mutual information of two random variables is generally hard and this makes the IB

[7] A more complete survey on flow-based models can be found in (Kobyzev et al., 2020).

[8] We provide a more direct empirical comparison with DisenIB (Pan et al., 2021) to AENIB in Appendix H.

objective infeasible in practice. To overcome this, *variational information bottleneck* (VIB) (Alemi et al., 2017; Chalk et al., 2016) applies variational inference to obtain a lower bound on the IB objective (1). Specifically, it approximates: (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 $\mathcal{N}(\mathbf{z}|0, I)$. Having such (variational) approximations in computing (1) as well as the Markov chain property $\mathbf{y} - \mathbf{x} - \mathbf{z}$ of neural networks, one yields the following lower bound on the IB objective (1):

$$I(\mathbf{z};\mathbf{y}) - \beta I(\mathbf{z},\mathbf{x}) \geq \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\int dz \left(p(z|\mathbf{x}) \log q(\mathbf{y}|z) - \beta p(z|\mathbf{x}) \log \frac{p(z|\mathbf{x})}{r(z)}\right)\right]. \tag{19}$$

This bound can now be approximated with the empirical distribution $p(\mathbf{x},\mathbf{y}) \approx \frac{1}{n}\sum_i \delta_{x_i}(\mathbf{x})\delta_{y_i}(\mathbf{y})$ from data. By further assuming a Gaussian encoder $p(\mathbf{z}|\mathbf{x}) := \mathcal{N}(\mathbf{z}|f^{\mu}(\mathbf{x}), f^{\sigma}(\mathbf{x}))$ as defined in (2) and applying the reprarametrization trick (Kingma & Welling, 2014), we get the following VIB objective:

$$L^{\beta}_{\texttt{VIB}} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\epsilon}}[-\log q(y_i|f(x_i,\boldsymbol{\epsilon}))] + \beta \ \mathrm{KL}\left(p(\mathbf{z}|x_i)\|r(\mathbf{z})\right). \tag{20}$$

Generative adversarial networks. *Generative adversarial network* (GAN) (Goodfellow et al., 2014) considers the problem of learning a generative model p_g from given data $\{x_i\}_{i=1}^n$, where $x_i \sim p_d(\mathbf{x})$ and $\mathbf{x} \in \mathcal{X}$. Specifically, GAN consists of two neural networks: (a) a *generator* network $G : \mathcal{Z} \to \mathcal{X}$ that maps a latent variable $z \sim p(\mathbf{z})$ into $\mathcal{X}$, where $p(\mathbf{z})$ is a specific prior distribution, and (b) a *discriminator* network $D : \mathcal{X} \to [0,1]$ that discriminates samples from p_d and those from the implicit distribution p_g derived from $G(\mathbf{z})$. The primitive form of training G and D is the following:

$$\min_G \max_D V(G,D) := \mathbb{E}_{\mathbf{x}}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))]. \tag{21}$$

For a fixed G, the inner maximization objective (21) with respect to D leads to the following *optimal discriminator* D^*_G, and consequently the outer minimization objective with respect to G becomes to minimize the *Jensen-Shannon divergence* between p_d and p_g, namely $D^*_G := \frac{p_d}{p_d + p_g}$.

G Experiments on Image Generation

Table 8: Test FID and IS of GANs on CIFAR-10. Bold and underline indicate the best and runner-up, respectively. We note that the value of ADA* (Karras et al., 2020a) is taken after 2× longer training steps.

CIFAR-10, Unconditional	Augment.	FID (↓)	IS (↑)
StyleGAN2 (Karras et al., 2020b)	HFlip	11.1	9.18
+ DiffAug (Zhao et al., 2020)	Trans, CutOut	9.89	9.40
+ ContraD (Jeong & Shin, 2021)	SimCLR	9.80	9.47
+ ADA* (Karras et al., 2020a)	Dynamic	**7.01***	-
+ FSD (R-18; ours)	HFlip, Trans	8.43	9.68
+ FSD (R-50; ours)	HFlip, Trans	<u>7.39</u>	**10.0**
FastGAN (Liu et al., 2021)	HFlip, Trans	34.5	6.52
+ Proj-GAN (R-18)	HFlip, Trans	8.48	9.40
+ FSD (R-18; ours)	HFlip, Trans	**7.80**	**9.65**

Table 9: Test FID and IS of VAE models on unconditional generation of CIFAR-10 and CelebA. Bold and underline denote the best and runner-up, respectively.

Method	CIFAR-10		CelebA
	FID ↓	IS ↑	FID ↓
VAE (Parmar et al., 2021)	115.8	3.8	-
VAE/GAN (Parmar et al., 2021)	39.8	7.4	-
2s-VAE (Dai & Wipf, 2019)	72.9	-	44.4
Perceptual AE (Zhang et al., 2020)	51.5	-	13.8
NCP-VAE (Aneja et al., 2021)	24.1	-	**5.25**
NVAE (Vahdat & Kautz, 2020)	56.0	5.19	13.5
DC-VAE (Parmar et al., 2021)	<u>17.9</u>	<u>8.2</u>	19.9
L_{recon} (5) only	65.0	5.73	50.1
+ Adv. similarity (22)	46.8	6.29	25.1
+ Projection (R-18)	**12.6**	**8.86**	<u>6.91</u>

G.1 Feature statistics discriminator for GANs

We evaluate the effect of our proposed feature statistics discriminator (FSD; Section 2.2 to the generation quality of GANs: here, we consider ImageNet-pretrained ResNet-18 (R-18) and ResNet-50 (R-50) (He et al., 2016), and define GAN discriminators via FSD upon the pre-trained models. We adopt StyleGAN2 (Karras et al., 2020b) and FastGAN (Liu et al., 2021) for the generator architectures. For the StyleGAN2-based models, we follow the training details of DiffAug (Zhao et al., 2020) and ADA (Karras et al., 2020a) in their CIFAR experiments: specifically, we use Adam with $(\alpha, \beta_1, \beta_2) = (0.002, 0.0, 0.99)$ for optimization with batch size of 64. We use non-saturating loss for training, and use R_1 regularization (Mescheder et al., 2018) with $\gamma = 0.01$. We do not use, however, the path length regularization and the lazy regularization (Karras et al., 2020b) in training. We take exponential moving average on the generator weights with half-life of 500K samples. We

stop training after 800K generator updates, which is about the half of those conducted for the ADA baseline (Karras et al., 2020a). For the FastGAN baseline, on the other hand, we run the official implementation of FastGAN[9] (Liu et al., 2021) on CIFAR-10 for the length of 6.4M samples with batch size 16. For the "Projected GAN" baseline, we adapt the official implementation[10] (Sauer et al., 2021) onto the ImageNet pre-trained ResNet-18, and trained for 6.4M samples with batch size 64. Our results ("FSD") follows the same training details, but with a difference in its discriminator.

Table 8 summarizes the results. Overall, we observe that FSD can aid GAN training of given generator network surprisingly effectively: by leveraging pre-trained representations, FSD could achieve FID competitive with a state-of-the-art level approach of ADA (Karras et al., 2020a) even with using much weaker data augmentation. Compared to Projected GAN (Sauer et al., 2021) that also leverages pre-trained models to stabilize GANs, our approach offers a more simpler approach to leverage the given representations, *i.e.*, by just aggregating the features statistics, yet achieving a better FID.

G.2 Feature statistics encoder for Autoencoders

We also evaluate our proposed architecture and method as a *generative modeling*, especially focusing on the effectiveness of the *feature statistics encoder* (Section 2.2) and the *adversarial similarity* based training of autoencoders on CIFAR-10 (Krizhevsky, 2009) and CelebA (Liu et al., 2015) datasets. To this end, we consider an "unsupervised" version of AENIB which omits the VIB loss ($L^{\beta}_{\texttt{VIB}}$; (19)) and the nuisance loss ($L_{\texttt{nuis}}$; (6)) in training, so that the model can assume an unconditional setup. Here, we present an additional *adversarial objective* based on our feature statistics based encoder (see Section 2.2) in training AENIB models to enhance the generative modeling capability.

Adversarial similarity based guidance. We found that the *feature statistics* based encoder for ConvNet-based architectures can be further leveraged to provide the decoder g an extra guidance in minimizing the (pixel-level) reconstruction loss (5): specifically, we propose to additionally place a discriminator network, say $d_{\mathbf{x}} : \mathbb{R}^{|\Pi_E|} \to \mathbb{R}^e$, that computes similarity between $\Pi_f(\mathbf{x})$ and $\Pi_f(g(\mathbf{z}, \mathbf{z}_n))$ and performs adversarial training on it:

$$L_{\texttt{sim}} := \max_{d_{\mathbf{x}}} \log(1 - \sigma(\tfrac{1}{\tau} \cdot \mathrm{sim}(d_{\mathbf{x}}(\Pi_f(\mathbf{x})), d_{\mathbf{x}}(\Pi_f(g(\mathbf{z}, \mathbf{z}_n)))))), \tag{22}$$

where $\sigma(\cdot)$ is the sigmoid function and $\mathrm{sim}(\mathbf{x}, \mathbf{y}) := \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$ denotes the cosine similarity. Here, τ is a temperature hyperparameter, and we use $\tau = 0.2$ throughout our experiments. We apply this additional training objective for ConvNet-based AENIB models, which turns out to be helpful to improve the generation quality of the learned autoencoders.

Table 9 summarizes the quantitative generation results of our AENIB models optimized with different objectives.[11] Firstly, it confirms the effectiveness of adversarial similarity based training: when it is solely applied upon $L_{\texttt{recon}}$ ("$L_{\texttt{recon}}$ only"; equivalent to (Makhzani et al., 2015)) it makes a significant improvements in both FID and IS. To further investigate the effectiveness of our proposed feature statistics encoder, we also test a scenario that the encoder is *fixed* by ResNet-18 pre-trained on ImageNet, akin to the setup of Table 8: we observe that our encoder design can surprisingly benefit from using better representation, *e.g.*, "+ Projection (R-18)" in Table 9 further improves FID on CIFAR-10 from $46.8 \to 12.6$, better than the best results among considered VAE-based models, by only training an MLP upon the feature statistics of the (fixed) model. It is notable that the gain only appears when we apply the adversarial similarity based training: *i.e.*, even with the pre-trained model, it only achieves 67.5 in FID on CIFAR-10 without the training. This observation suggests an interesting direction to scale-up autoencoder-based models by leveraging large pre-trained representations, in a similar vein as (Sauer et al., 2021) as presented in the context of GANs.

[9] https://github.com/odegeasslbc/FastGAN-pytorch

[10] https://github.com/autonomousvision/projected_gan

[11] Following other baselines, we compute FIDs from 50,000 generated samples against the training dataset.

G.3 QUALITATIVE RESULTS

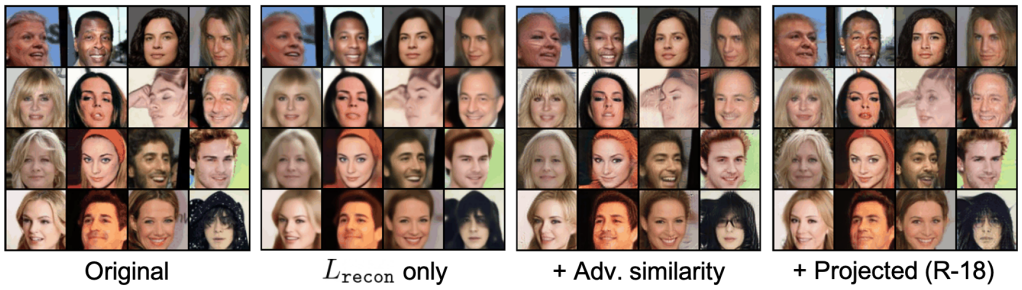
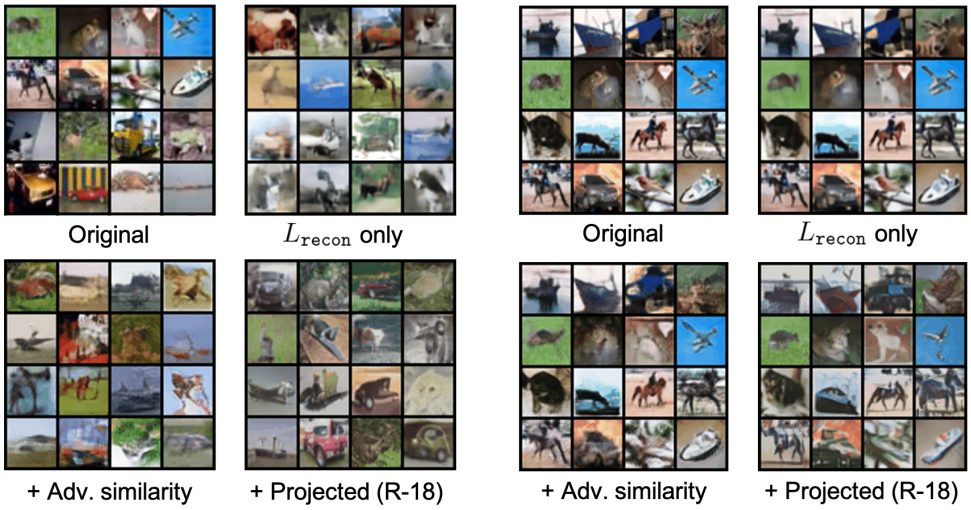


Figure 6: Qualitative comparison on reconstructed samples from fixed samples of unconditional AENIB model (and its ablations) trained on CelebA.

(a) $L_{\texttt{recon}}$ only (FID: 50.1) (b) + Adv. similarity (FID: 25.1) (c) + Projected (FID: 6.91)

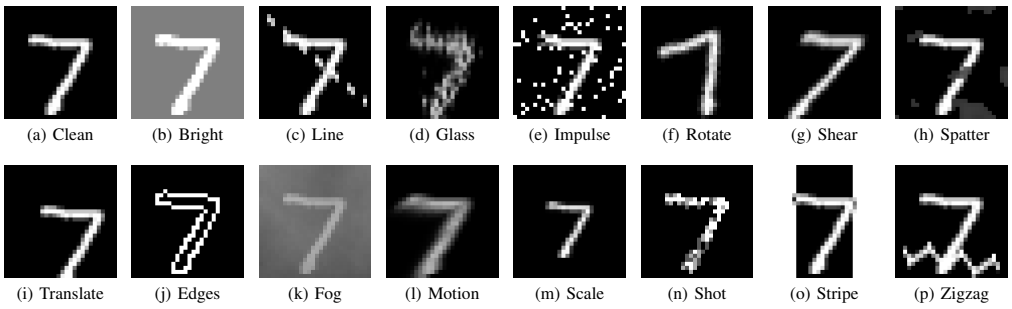
Figure 7: Qualitative comparison on uncurated random samples generated from unconditional AENIB model (and its ablations) trained on CelebA.



(a) Reconstruction (b) Sampling

Figure 8: Qualitative comparison on uncurated random samples generated from unconditional AENIB model (and its ablations) trained on CelebA.

H RESULTS ON MNIST-C



(a) Clean (b) Bright (c) Line (d) Glass (e) Impulse (f) Rotate (g) Shear (h) Spatter

(i) Translate (j) Edges (k) Fog (l) Motion (m) Scale (n) Shot (o) Stripe (p) Zigzag

Figure 9: Sample images in MNIST-C test dataset for different corruption types.

Table 10: Comparison of (a) clean error (%; ↓), (b) AUROC (↑) on detecting Gaussian noise (higher is better), and (c) corruption errors (%; ↓) per corruption type on MNIST-C (Mu & Gilmer, 2019). Each classifier is trained on MNIST with random translation as augmentation. We highlight our results as blue whenever the value improves the baselines more than 3% in absolute values.

Method	*Clean*	*AUROC (↑)*	*Shot*	*Impulse*	*Glass*	*Motion*	*Shear*	*Scale*	*Rotate*	*Brightness*	*Translate*	*Stripe*	*Fog*	*Spatter*	*Dotted line*	*Zigzag*	*Canny edges*	*Average*
Cross-entropy	0.45	0.987	4.69	69.6	60.3	46.5	1.41	2.97	**4.80**	88.7	2.45	76.6	88.7	27.3	5.64	**27.3**	44.1	34.5
VIB	**0.44**	0.988	4.52	73.5	73.8	71.8	1.73	2.84	5.85	90.1	2.15	78.1	89.8	28.4	5.85	28.5	44.0	37.6
sq-VIB	0.48	0.955	4.32	71.5	63.5	62.3	1.62	2.70	5.74	90.5	2.43	80.3	90.3	24.8	5.91	32.0	43.4	36.4
NLIB	1.15	0.974	7.13	67.9	62.5	57.9	2.15	4.00	7.06	**86.9**	3.28	81.8	88.7	30.1	8.97	31.0	41.8	36.4
sq-NLIB	3.19	0.908	9.90	73.3	66.7	64.7	4.25	6.19	9.21	88.7	6.43	**72.4**	89.8	32.4	9.69	36.2	72.5	40.3
DisenIB	0.54	0.997	4.60	68.8	56.4	50.4	1.11	**2.04**	4.84	88.7	2.01	74.3	**88.5**	20.1	4.75	27.4	69.0	35.2
AENIB (ours)	0.72	**1.000**	**3.71**	**48.8**	**44.0**	**27.1**	**0.99**	3.15	4.82	89.7	**0.88**	82.0	89.7	**16.4**	**4.14**	33.9	**25.9**	**29.8**

We also evaluate our proposed AENIB training on MNIST-C (Mu & Gilmer, 2019), a collection of corrupted versions of the MNIST (LeCun et al., 1998) test dataset of 15 corruption types (see Figure 9 for concrete examples) constructed in a similar manner to CIFAR-10/100-C (Hendrycks & Dietterich, 2019), to get a clearer view on the effectiveness of our method on a simpler setup. For this experiments, we use a simple 4-layer convolutional network (with batch normalization (Ioffe & Szegedy, 2015)) as the encoder architecture, and trained every model on the (clean) MNIST training dataset for 100K updates following other training details of the CIFAR experiments (see Appendix B.1): again, we notice that the training does not assume specific prior on the corruptions. We compare AENIB with the direct ablations of cross-entropy and VIB based models, as well as some variants of VIB, namely Nonlinear-VIB (Kolchinsky et al., 2019), Squared-VIB/NIB (Thobaben et al., 2020), and DisenIB (Pan et al., 2021). Especially, we compare with DisenIB as (a) it considers a nuisance modeling (based on FactorVAE (Kim & Mnih, 2018)) as AENIB does, while (b) also tackling some robustness concerns, *e.g.*, its claimed effectiveness on out-of-distribution detection for MNIST *vs.* Gaussian noise.

Table 10 summarizes the results: overall, we observe that the effectiveness of AENIB training still applies to MNIST-C, *e.g.*, our AENIB training improves the average corruption error from the baseline cross-entropy based training from $33.1\% \rightarrow 29.8\%$, which could not be obtained by simply sweeping on the baseline VIB training. Given that MNIST-C allows a visually clearer distinction between contents and corruptions compared to CIFAR-10/100-C, one can better interpret the behavior of given models on each corruption types: here, we observe that our training can dramatically improve robustness for certain types of corruptions where the baselines shows poor performances, *e.g.*, Impulse, Glass, and Motion, while still some types of corruptions are still remaining challenging even with AENIB, *e.g.*, especially for low-frequency biased corruptions such as Brightness and Stripe. Compared to DisenIB, on the other hand, we observe that the effectiveness from DisenIB, e.g., its gain in AUROC (as conducted by Pan et al. (2021)), could not be further generalized on MNIST-C, where AENIB still improves upon it as well as achieving the perfect score at the same OOD task.

I APPLICATION TO MODEL DEBUGGING

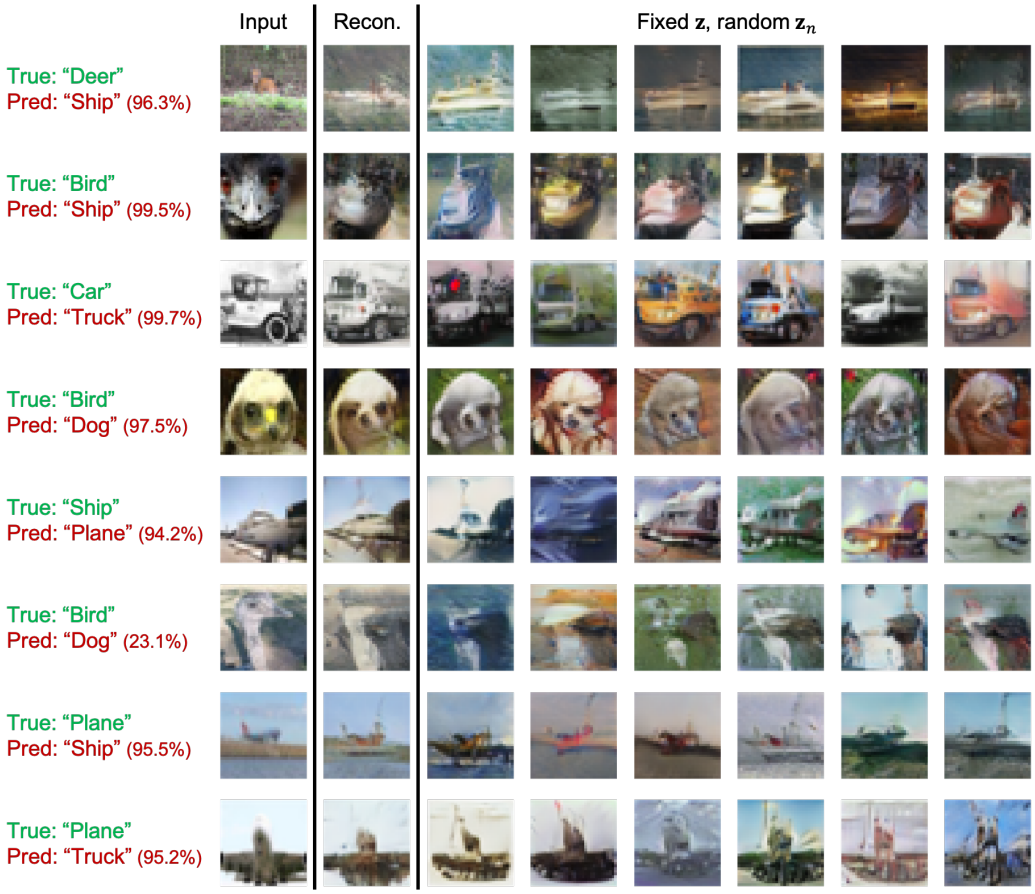


Figure 10: Qualitative comparisons between (a) the original input (the leftmost column), (b) its reconstruction (the second column), and (c) its further reconstructions with random nuisance $\mathbf{z}_n$ (the remaining columns), examined for test samples misclassified by a CIFAR-10 AENIB model.

To further understand how the proposed AENIB model internally works with its representation $\mathbf{z}$ and $\mathbf{z}_n$, we examine an AENIB model trained on CIFAR-10 to analyze how the model reconstruct given inputs when the model incorrectly classifies them. Specifically, Figure 10 illustrates a subset of CIFAR-10 test samples misclassified by an AENIB model by comparing the original input with its reconstructed samples from the model. Overall, we observe that such a qualitative comparison can provide a useful signal to interpret model errors: it effectively visualizes which visual cues of a given input negatively affected the decision making process of the given model, also visualizing the closest (misclassified) realizations that the model decodes for a given representation, *i.e.*, what the model actually perceived. For example, for the test input given at the first row of Figure 10, one can observe that the model essentially "ignored" the tiny part that represent the true semantic, *i.e.*, the "deer", and reconstructed the remaining part as a "ship".