Out-of-Distribution Detection with Reconstruction Error and Typicality-based Penalty

Genki Osada[1,2], Tsubasa Takahashi[1], Budrul Ahsan[3], and Takashi Nishide[2]

[1]LINE Corporation, Japan
[2]University of Tsukuba, Japan
[3]IBM Japan

Abstract

The task of out-of-distribution (OOD) detection is vital to realize safe and reliable operation for real-world applications. After the failure of likelihood-based detection in high dimensions had been shown, approaches based on the typical set *have been attracting attention; however, they still have not achieved satisfactory performance. Beginning by presenting the failure case of the typicality-based approach, we propose a new reconstruction error-based approach that employs normalizing flow (NF). We further introduce a typicality-based penalty, and by incorporating it into the reconstruction error in NF, we propose a new OOD detection method, penalized reconstruction error (PRE). Because the PRE detects test inputs that lie off the in-distribution manifold, it effectively detects adversarial examples as well as OOD examples. We show the effectiveness of our method through the evaluation using natural image datasets, CIFAR-10, TinyImageNet, and ILSVRC2012.*

1. Introduction

Recent works have shown that deep neural network (DNN) models tend to make incorrect predictions with high confidence when the input data at the test time are significantly different from the training data [1, 2, 3, 4, 5, 6] or adversely crafted [7, 8, 9]. Such anomalous inputs are often referred to as out-of-distribution (OOD). We refer to a distribution from which expected data, including training data, comes as the in-distribution (In-Dist) and to a distribution from which unexpected data we should detect comes as the OOD. We tackle OOD detection [1, 3, 10, 11] which attempts to distinguish whether an input at the test time is from the In-Dist or not.

Earlier, [12] introduced the likelihood-based approach: detecting data points with low likelihood as OOD using a density estimation model learned on training data. However, recent experiments using the deep generative models (DGMs) showed that the likelihood-based approach often fails in high dimensions [10, 13] (Section 3.1). This observation has motivated alternative methods [14, 15, 16, 17, 18, 19]. Among them, [13, 20] argued the need to account for the notion of *typical set* instead of likelihood, but those typicality-based methods still did not achieve satisfactory performance [13, 21, 22]. We first argue in Section 3.2 the failure case of the typicality-based detection performed on an isotropic Gaussian latent distribution proposed by [13, 20], which we refer to as the typicality test in latent space (TTL). Because the TTL reduces the information in the input vector into a single scalar as the L_2 norm in latent space, the TTL may lose the information that distinguishes OOD examples from In-Dist ones.

To address this issue, we first propose a new reconstruction error-based approach that employs normalizing flow (NF) [23, 24, 25]. We combined the two facts that the previous studies have shown: 1) Assuming the manifold hypothesis is true [26, 27, 28], the density estimation model, including NFs, will cause very large Lipschitz constants in the regions that lie off the data manifold [29]. 2) The Lipschitz constants of NFs can be connected to its reconstruction error [30]. On the premise that the In-Dist examples lie on the manifold yet the OOD examples do not, we detect a test input that lies off the manifold as OOD when its reconstruction error is large. Unlike the TTL, our method uses the information of latent vectors as-is, enabling the preservation of the information that distinguishes OOD examples from In-Dist ones. Second, to boost detection performance further, we introduce a typicality-based penalty. By applying controlled perturbation (we call a penalty) in latent space according to the *atypicality* of inputs, we can increase the reconstruction error only when inputs are likely to be

OOD, thereby improving the detection performance. The overview of our method is shown in Fig. 1.

Contribution. The contributions of this paper are the following three items:

- An OOD detection method based on the reconstruction error in NFs. Based on the property between the Lipschitz constants of NFs and its reconstruction error given by [30], the proposed method detects test inputs that lie off the manifold of In-Dist as OOD.
- We further introduce a typicality-based penalty. It enhances OOD detection by penalizing inputs atypical in the latent space, on the premise that the In-Dist data are typical. Incorporating this into the reconstruction error, we propose penalized reconstruction error (PRE).
- We demonstrate the effectiveness of our PRE in extensive empirical observations on CIFAR-10 [31] and Tiny-ImageNet. The PRE consistently showed high detection performance for various OOD types. Furthermore, we show on ILSVRC2012 [32] that the proposed methods perform better than 95% detection in AUROC on average, even for large-size images. When an OOD detector is deployed for real-world applications with no control over its input, having no specific weakness is highly desirable.

Our PRE also effectively detects adversarial examples. Among several explanations about the origin of adversarial examples, [33, 34, 35, 36] hypothesized that adversarial examples exist in regions close to, but lie off, the manifold of normal data (i.e., In-Dist data), and [37, 38, 39, 40] provided experimental evidence supporting this hypothesis. Thus, our PRE should also detect adversarial examples as OOD, and we demonstrate it in the evaluation experiments. Historically, the studies of OOD detection and detecting adversarial examples have progressed separately, and thus few previous works used both samples in their detection performance evaluation, but this work addresses that challenge.

2. Related Work

For the likelihood-based methods and typicality test, we refer the reader to Section 3.1 and 3.2. We introduce two other approaches: 1) Reconstruction error in Auto-Encoder (AE) has been widely used for anomaly detection [41, 42, 43, 44, 45] and also has been employed for detecting adversarial examples [40]. Using an AE model that has been trained to reconstruct normal data (i.e., In-Dist data) well, this approach aims to detect samples that fail to be reconstructed accurately as anomalies (or adversarial examples). Since the basis of our proposed method is the reconstruction error in NF models, we will evaluate

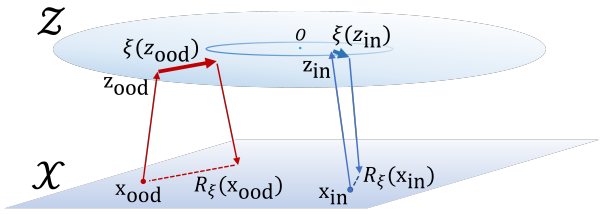

Figure 1: Illustration of our method, PRE. The red dot ($\mathbf{x}_{\text{ood}}$) and blue dot ($\mathbf{x}_{\text{in}}$) represent an OOD and In-Dist samples, respectively. The cyan circle in latent space $\mathcal{Z}$ centered at the origin O represents the Gaussian Annulus, on which the typical set (i.e., In-Dist examples) concentrates. The $\mathbf{z}_{\text{ood}}$ and $\mathbf{z}_{\text{in}}$ are subject to controlled perturbations ξ as a penalty (bold arrows), according to L_2 distance to the Gaussian Annulus. While $\mathbf{z}_{\text{in}}$ will be close to the Gaussian Annulus, $\mathbf{z}_{\text{ood}}$ will be away from it (Section 3.2), so $|\xi(\mathbf{z}_{\text{ood}})| > |\xi(\mathbf{z}_{\text{in}})|$. The reconstruction errors measured in data space $\mathcal{X}$ (the length of the dashed lines) are increased according to $|\xi|$, which leads to $R_\xi(\mathbf{x}_{\text{ood}}) > R_\xi(\mathbf{x}_{\text{in}})$ and allows us to detect $\mathbf{x}_{\text{ood}}$.

the AE-based reconstruction error method as a baseline. 2) Classifier-based methods that use the outputs of a classifier network has also been taken in many previous works [1, 2, 3, 46, 4, 6]. This approach has also been taken in the works of detecting adversarial examples [47, 48, 49]. However, the limitation of this approach is that label information is required for training classifiers. Furthermore, the dependence on the classifier's performance is their weakness. We show that later in the experimental results.

3. Preliminary

3.1. Likelihood-based OOD Detection

As an approach to OOD detection, [50] introduced a method that uses a density estimation model learned on In-Dist samples, i.e., training data. By interpreting the probabilistic density for an input $\mathbf{x}$, $p(\mathbf{x})$, as a likelihood, it assumes that OOD examples would be assigned a lower likelihood than the In-Dist ones. Based on this assumption, [38] has proposed a method detecting adversarial examples using DGMs, specifically PixelCNN [51]. However, [10, 13] have presented the counter-evidence against this assumption: DGMs trained on a particular dataset often assigns higher log-likelihood, $\log p(\mathbf{x})$, to OOD examples than the samples from its training dataset (i.e., the In-Dist) in high dimensions. [13, 20] argued that this failure of the likelihood-based approach is due to the lack of accounting for the notion of *typical set*: a set, or a region, that contains the enormous probability mass of distribution (see [20, 52] for formal definitions). Samples drawn from a DGM will come from its typical set; however, in

high dimensions, the typical set may not necessarily intersect with high-density regions, i.e., high-likelihood regions. While it is difficult to formulate the region of the typical set for arbitrary distributions, it is possible for an isotropic Gaussian distribution, which is the latent distribution of NFs [23, 24, 25]. It is well known that if a vector $\mathbf{z}$ belongs to the typical set of the d-dimensional Gaussian $\mathcal{N}(\mu, \sigma^2\mathbf{I}_d)$, $\mathbf{z}$ satisfies $\|\mathbf{z}-\mu\| \simeq \sigma\sqrt{d}$ with a high probability, i.e., concentrates on an annulus centered at μ with radius $\sigma\sqrt{d}$, which is known as Gaussian Annulus [53]. As dimension d increases, the regions on which the typical samples (i.e., In-Dist samples) concentrate move away from the mean of Gaussian, where the likelihood is highest. That is why In-Dist examples are often assigned low likelihood in high dimensions.

3.2. Typicality-based OOD Detection

[13, 20] suggested flagging test inputs as OOD when they fall outside of the distribution's typical set. The deviation from the typical set (i.e., *atypicality*) in a standard Gaussian latent distribution $\mathcal{N}(0, \mathbf{I}_d)$ of NF is measured as $\texttt{abs}(\|\mathbf{z}\| - \sqrt{d})$, where $\|\mathbf{z}\|$ is L_2 norm of the latent vector corresponding to test input and $\sqrt{d}$ means the radius of the Gaussian Annulus. Their proposed method measures the atypicality as an OOD score, which we refer to as the typicality test in latent space (TTL). The authors however concluded that the TTL was not effective [13, 21, 22]. We have the following views regarding the failure of TTL. As the latent distribution is fixed in NF, In-Dist examples will be highly likely to fall into the typical set, i.e., $\texttt{abs}(\|\mathbf{z}\| - \sqrt{d}) \approx 0$. However, the opposite is not guaranteed: there is no guarantee that OOD examples will always be out of the typical set. It is because the TTL reduces the information of a vector of test input to a single scalar as its L_2 norm. For example, in a 5-dimensional space with the probability density $\mathcal{N}(0, \mathbf{I}_5)$, the typicality test will judge a vector $\mathbf{z}$ as In-Dist when $\|\mathbf{z}\| \simeq \sqrt{5}$. At the same time, however, even a vector such as $[\sqrt{5}, 0, 0, 0, 0]$, which has an extremely low occurrence probability (in the first element) and thereby possibly should be detected as OOD, will judge as In-Dist as well because it belongs to the typical set in terms of its L_2 norm ($\sqrt{5}$). Indeed, we observed such an example of this case in the experiments and will show it in Section 6.2. We propose a method that addresses this issue.

3.3. Normalizing Flow

The normalizing flow (NF) [23, 24, 25] has been becoming a popular method for density estimation. In short, the NF learns an invertible mapping $f : \mathcal{X} \to \mathcal{Z}$ that maps the observable data $\mathbf{x}$ to the latent vector $\mathbf{z} = f(\mathbf{x})$ where $\mathcal{X} \in \mathbb{R}^d$ is a data space and $\mathcal{Z} \in \mathbb{R}^d$ is a latent space. A distribution on $\mathcal{Z}$, which is denoted by $P\text{z}$, is fixed to an isotropic Gaussian $\mathcal{N}(0, \mathbf{I}_d)$, and thus its density is $p(\mathbf{z}) = (2\pi)^{-\frac{d}{2}}\exp(-\frac{1}{2}\|\mathbf{z}\|^2)$. An NF learns an unknown target, or true, distribution on $\mathcal{X}$, which is denoted by $P\text{x}$, by fitting an approximate model distribution $\widehat{P\text{x}}$ to it. Under the change of variable rule, the log density of $\widehat{P\text{x}}$ is $\log p(\mathbf{x}) = \log p(\mathbf{z}) + \log|\det J_f(\mathbf{x})|$ where $J_f(\mathbf{x}) = df(\mathbf{x})/d\mathbf{x}$ is the Jacobian matrix of f at $\mathbf{x}$. Through maximizing $\log p(\mathbf{z})$ and $\log|\det J_f(\mathbf{x})|$ simultaneously w.r.t. the samples $\mathbf{x} \sim P\text{x}$, f is trained so that $\widehat{P\text{x}}$ matches $P\text{x}$. In this work, $P\text{x}$ is In-Dist, and we train an NF model using samples from the In-Dist.

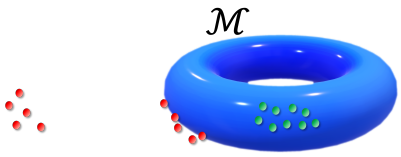

Figure 2: llustration of the manifold $\mathcal{M}$ (represented by torus as an example), In-Dist examples (green dots), and OOD examples close to and far from $\mathcal{M}$ (red dots).

4. Method

We propose a method that flags a test input as OOD when out of the manifold of In-Dist data (Fig. 2). Our method is based on the reconstruction error in NFs, combined with a typicality-based penalty for further performance enhancement, which we call the penalized reconstruction error (PRE) (Fig. 1). Before describing the PRE in Section 4.2, We first introduce the recently revealed characteristics of NFs that form the basis of the PRE in Section 4.1: the mechanism of how reconstruction errors occur in NFs, which are supposed to be invertible, and how they become more prominent when inputs are OOD.

4.1. Motivation to Use Reconstruction Error in NF for OOD Detection

Reconstruction error occurs in NFs. Recent studies have shown that when the supports of $P\text{x}$ and $P\text{z}$ are topologically distinct, NFs cannot perfectly fit $\widehat{P\text{x}}$ to $P\text{x}$ [54, 55, 56, 57, 58]. As $P\text{z}$ is $\mathcal{N}(0, \mathbf{I}_d)$, even having 'a hole' in the support of $P\text{x}$ makes them topologically different. This seems inevitable since $P\text{x}$ is usually very complicated (e.g., a distribution over natural images). In those works, the discrepancy between $\widehat{P\text{x}}$ and $P\text{x}$ is quantified with the Lipschitz constants. Let us denote the Lipschitz constant of f by $\text{Lip}(f)$ and that of f^{-1} by $\text{Lip}(f^{-1})$. When $P\text{x}$ and $P\text{z}$ are topologically distinct, in order to make $\widehat{P\text{x}}$ fit into $P\text{x}$ well, $\text{Lip}(f)$ and $\text{Lip}(f^{-1})$ are required to be significantly large [30, 58]. Based on this connection, the inequalities to assess the discrepancy between $P\text{x}$ and $\widehat{P\text{x}}$ were presented by [56] which uses Total Variation (TV) distance and by [55]

which uses the Precision-Recall (PR) [59] as well. [30] has analyzed it locally from the aspect of numerical errors and introduced another inequality:

$$\left\|\mathbf{x} - f^{-1}(f(\mathbf{x}))\right\| \leq \left\|\mathrm{Lip}(f^{-1})\right\| \|\delta_{\mathbf{z}}\| + \|\delta_{\mathbf{x}}\| . \quad (1)$$

See Appendix A in this paper or Appendix C in the original paper for the derivation. The $\delta_{\mathbf{z}}$ and $\delta_{\mathbf{x}}$ represent numerical errors in the mapping through f and f^{-1}. When $\mathrm{Lip}(f)$ or $\mathrm{Lip}(f^{-1})$ is large, $\delta_{\mathbf{z}}$ and $\delta_{\mathbf{x}}$ grow significantly large and cause `Inf/NaN` values in $f^{-1}(f(\mathbf{x}))$, which is called *inverse explosions*. Note that Eq. (1) considers the *local* Lipschitz constant, i.e., $\mathrm{Lip}(f^{-1})^{-1} \|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\| \leq \mathrm{Lip}(f) \|\mathbf{x}_1 - \mathbf{x}_2\|, \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{A}$, and thus it depends on the region $\mathcal{A}$ where the test inputs exist. While computing $\mathrm{Lip}(f^{-1})$, $\delta_{\mathbf{z}}$, and $\delta_{\mathbf{x}}$ directly in a complex DNN model is hard [60], Eq. (1) suggests that the reconstruction error for $\mathbf{x}$ (the LHS) can approximately measure the discrepancy between $P\mathrm{x}$ and $\widehat{P\mathrm{x}}$ locally. Even though the NF is theoretically invertible, [30] has demonstrated that it is not the case in practice and the reconstruction error is non-zero. Another example of the numerical error in an invertible mapping has been observed in [61].

Connection between OOD and reconstruction errors. [29] further connected this discussion to the manifold hypothesis, i.e., that the high-dimensional data in the real world tend to exist on a low-dimensional manifold [26, 27, 28]. Assuming the manifold hypothesis is true, the density $p(\mathbf{x})$ would be very high only if $\mathbf{x}$ is on the manifold, $\mathcal{M}$, while $p(\mathbf{x})$ would be close to zero otherwise. Thus, the value of $p(\mathbf{x})$ may fluctuate abruptly around $\mathcal{M}$. It means that the local Lipschitz constants of NFs, i.e., $\mathrm{Lip}(f^{-1})$ in Eq. (1), become significantly large, if not infinity. As a result, the reconstruction error, which is the lower bound of Eq. (1), will be large. By contrast, since In-Dist examples should be on $\mathcal{M}$, abrupt fluctuations in $p(\mathbf{x})$ are unlikely to occur. Thus $\mathrm{Lip}(f^{-1})$ will be smaller, and the reconstruction error will be smaller for In-Dist examples. Thus, this argument allows us to consider an input with a large reconstruction error to be OOD.

OOD examples close to and far from manifold. In the above, we considered OOD examples $\mathbf{x}_{\text{ood}}$ that lie off but are close to $\mathcal{M}$ of the In-Dist $P\mathrm{x}$. We depict it as red dots near the blue torus in Fig. 2. Adversarial examples are considered to be included in such $\mathbf{x}_{\text{ood}}$. On the other hand, there are also $\mathbf{x}_{\text{ood}}$ far away from $\mathcal{M}$, as depicted by the cluster on the left in Fig. 2. Random noise images may be an example of this. In the region far from $\mathcal{M}$, $p(\mathbf{x})$ should be almost constant at 0, so $\mathrm{Lip}(f^{-1})$ may not become large. Nevertheless, as we will explain in Section 6.1, the reconstruction error will be large even for such $\mathbf{x}_{\text{ood}}$ far from $\mathcal{M}$, as long as $\mathbf{x}_{\text{ood}}$ are regarded as atypical in $P\mathrm{z}$. We defer explaining this mechanism to Section 6.1. In a nutshell, atypical samples are assigned minimal probabilities in $P\mathrm{z}$, which causes $\|\delta_{\mathbf{z}}\|$ and $\|\delta_{\mathbf{x}}\|$, as opposed to $\mathrm{Lip}(f^{-1})$, to be larger, resulting in a larger reconstruction error.

4.2. Our Method: Penalized Reconstruction Error (PRE)

For OOD inputs that lie off the manifold of In-Dist, $\mathcal{M}$, regardless of whether they are close to or far from $\mathcal{M}$, the reconstruction error in NFs will increase. Thus we can judge whether a test input is OOD or not by measuring the magnitude of the reconstruction error, written as

$$R(\mathbf{x}) := \left\|\mathbf{x} - f^{-1}\left(f(\mathbf{x})\right)\right\| . \quad (2)$$

Contrary to the PR- and TV-based metrics that need a certain amount of data points to compare $P\mathrm{x}$ and $\widehat{P\mathrm{x}}$, the reconstruction error works on a single data point, and thus $R(\mathbf{x})$ is suited for use in detection.

Typicality-based penalty. To further boost detection performance, we add a penalty ξ to a test input $\mathbf{x} \in \mathbb{R}^d$ in the latent space $\mathcal{Z}$ as $\hat{\mathbf{z}} = \mathbf{z} + \xi$ where $\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^d$. Since an NF model provides a one-to-one mapping between $\mathbf{x}$ and $\mathbf{z}$, the shift by ξ immediately gains the reconstruction error. We conducted the controlled experiments to confirm its validity and saw that the degree of the reconstruction error is proportional to the intensity of ξ, regardless of the direction of ξ. (See Appendix B.) We want to make ξ large only when $\mathbf{x}$ is OOD. To this end, we use the typicality in $P\mathrm{z}$ described in Section 3.2, and specifically, we design ξ as

$$\xi(\mathbf{z}) = -\mathrm{sign}\left(\|\mathbf{z}\| - \sqrt{d}\right) \left(\frac{\|\mathbf{z}\| - \sqrt{d}}{\sqrt{d}}\right)^2 . \quad (3)$$

There may be several possible implementations of $\xi(\mathbf{z})$, but we chose to emulate the elastic force in the form of an attractive force proportional to the square of the distance from the center, $\sqrt{d}$. The larger the deviation of $\mathbf{z}$ from the typical set in $P\mathrm{z}$, the larger the value of $\mathtt{abs}(\|\mathbf{z}\| - \sqrt{d})$ and thus the larger the value of $\xi(\mathbf{z})$.

Penalized Reconstruction Error (PRE) Incorporating ξ, the score we use is:

$$R_\xi(\mathbf{x}) := \left\|\mathbf{x} - f^{-1}\left(\mathbf{z} + \lambda\xi(\mathbf{z})\frac{\mathbf{z}}{\|\mathbf{z}\|}\right)\right\| \quad (4)$$

where λ is a coefficient given as a hyperparameter. We call the test based on R_ξ the penalized reconstruction error (PRE). Unlike the TTL that uses the information of $\mathbf{z} = f(\mathbf{x})$ in a reduced form as $\|\mathbf{z}\|$, the computation of R_ξ uses $\mathbf{z}$ as-is without reducing. Therefore, the PRE works well even for cases where the TTL fails.

PRE as the OOD detector. The larger $R_\xi(\mathbf{x}_{\text{test}})$, the more likely a test input $\mathbf{x}_{\text{test}}$ is OOD. With using the threshold τ, we flags $\mathbf{x}_{\text{test}}$ as OOD when

$$R_\xi(\mathbf{x}_{\text{test}}) > \tau. \tag{5}$$

The overview of how the PRE identifies the OOD examples is shown in Fig. 1. As the NF model is trained with samples from $P\text{x}$, if a test input $\mathbf{x}_{\text{test}}$ belongs to $P\text{x}$ (i.e., $\mathbf{x}_{\text{test}}$ is In-Dist), $\mathbf{z}_{\text{test}} = f(\mathbf{x}_{\text{test}})$ would be also typical for $P\text{z}$. Then, as $P\text{z}$ is fixed to $\mathcal{N}(0, \mathbf{I})$, $\|\mathbf{z}_{\text{test}}\|$ would be close to $\sqrt{d}$ as described in Section 3.1, and as a result, $\xi(\mathbf{z}_{\text{test}})$ becomes negligible. In contrast, if $\mathbf{x}_{\text{test}}$ is OOD, $\mathbf{z}_{\text{test}}$ would be atypical for $P\text{z}$, and thus $\|\mathbf{z}_{\text{test}}\|$ deviates from $\sqrt{d}$, which makes $\xi(\mathbf{z}_{\text{test}})$ large and consequently enlarges $R_\xi(\mathbf{x}_{\text{test}})$. Thus, we can view ξ as a deliberate introduction of numerical error $\delta_{\mathbf{z}}$ in Eq. (1).

We emphasize that the ξ depends only on the typicality of $\mathbf{x}_{\text{test}}$ and not on its distance from the manifold $\mathcal{M}$. Therefore, whenever $\mathbf{x}_{\text{test}}$ is atypical, R_ξ will be large, regardless of whether it is close to or far from $\mathcal{M}$, i.e., for any OOD data points in Fig. 2.

4.3. Reasons not to employ VAE

Variational Auto-Encoder (VAE) [62, 63] is another generative model that has a Gaussian latent space. Although VAE has been used in previous studies of likelihood-based OOD detection, we did not select it for our proposed method for the following reasons: 1) The primary goal of VAE is to learn a low-dimensional manifold, not to learn invertible transformations as in NF. Therefore, there is no guarantee that the mechanism described in Section 4.1 will hold. 2) It is known that the reconstructed image of VAE tends to be blurry and that the reconstruction error is large even for In-Dist samples [64, 65]; to be used for OOD detection, the reconstruction error must be suppressed for In-Dist samples, and VAE is not suited for this purpose. 3) Since the latent space of VAE is only an approximation of the Gaussian, $\|\mathbf{z}\|$ cannot correctly measure typicality; in previous studies that dealt with both VAE and NF, the typicality test in latent space (TTL) was applied only to NF, for the same reason [13, 20, 16]. In relation to 2) above, we evaluate in Section 5 the detection performance using the reconstruction error of Auto-Encoder, which is superior to VAE in terms of the small reconstruction error.

5. Experiments

We demonstrate the effectiveness of our proposed method. We measure the success rate of OOD detection using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR). The experiments run on a single NVIDIA V100 GPU.

5.1. Dataset

We utilize three datasets as In-Dist. The OOD datasets we use consists of two types: the different datasets from the In-Dist datasets and adversarial examples. The tests are performed on 2048 examples that consist of 1024 examples chosen at random from the In-Dist dataset and 1024 examples from the OOD dataset.

Dataset for In-Dist. We use widely used natural image datasets, CIFAR-10 (C-10), TinyImageNet (TIN), and ILSVRC2012, each of which we call the In-Dist dataset. They consist of 32×32, 64×64, and 224×224 pixel RGB images, and the number of containing classes is 10, 200, and 1000, respectively.

Different datasets from In-Dist. We use CelebA [66], TinyImageNet (TIN) [32], and LSUN [67] as OOD datasets. The LSUN dataset contains several scene categories, from which we chose *Bedroom*, *Living room*, and *Tower*, and treat each of them as a separate OOD dataset. See Appendix C.1 for processing procedures. As ILSVRC2012 contains an extensive range of natural images, including architectural structures such as towers and scenes of rooms, like those included in LSUN datasets, LSUN was excluded from the OOD datasets to be evaluated.[1] Noise images are also considered one type of OOD. Following [15], we control the noise complexity by varying the size of average-pooling (κ) to be applied. For detailed procedures, refer to Appendix C.1. Treating images with different κ as separate datasets, we refer to them as Noise-κ.

Adversarial examples. We generate adversarial examples with two methods, Projected Gradient Descent (PGD) [68, 69] and Carlini & Wagner's (CW) attack [70]. Following [71], we use PGD as L_∞ attack and CW as L_2 attacks, respectively. For descriptions of each method, the training settings for the classifiers, and the parameters for generating adversarial examples, we would like to refer the reader to Appendices C.2 and C.3. The strength of PGD attacks is controlled by ϵ, which is a parameter often called *attack budgets* that specifies the maximum norm of adversarial perturbation, while the strength of CW attacks is controlled by k, called *confidences*. We use $\frac{2}{256}$ and $\frac{8}{256}$ for ϵ in PGD and 0 and 10 for k in CW, which we refer to as PGD-2, PGD-8, CW-0, and CW-10, respectively. The classifier we used for both attacks is WideResNet 28-10 (WRN 28-10)[2] [72] for C-10 and TIN and ResNet-50 v2 [73][3] for

[1] A helpful site for viewing the contained images: cs.stanford.edu/people/karpathy/cnnembed/

[2] github.com/tensorflow/models/tree/master/research/autoaugment

[3] github.com/johnnylu305/ResNet-50-101-152

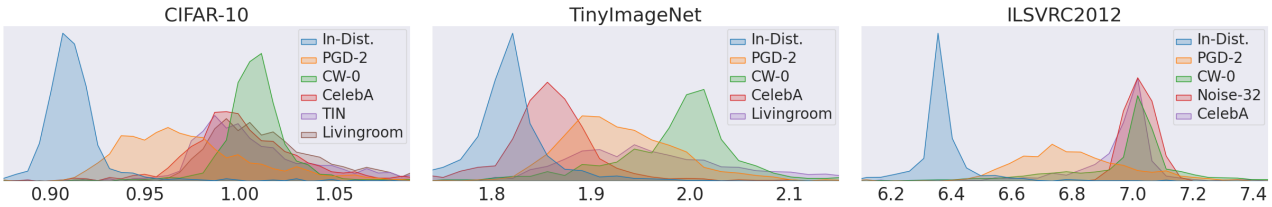

Figure 3: Histograms of the PRE (our method). The x-axis is R_ξ. The R_ξ for 'In-Dist' is lower and well separated than that for OOD datasets.

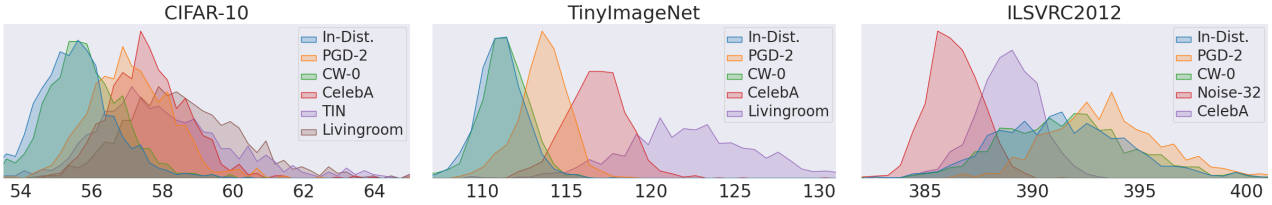

Figure 4: Histograms of L_2 norm in the latent space. The x-axis is $\|\mathbf{z}\|$. The 'In-Dist' is not separate from the OOD datases. In particular, 'CW-0' has almost completely overlapped 'In-Dist'.

ILSVRC2012. The classification accuracies before and after attacking are shown in Table 4 in Appendix C.1. For example, on CIFAR-10, the number of samples that could be correctly classified dropped to 57 out of 1024 samples (5.556%) by the PGD-2 attack, and the other three attacks had zero samples successfully classified (0.0%).

5.2. Implementation

Normalizing Flow. As with most previous works, we use Glow [74] for the NF model in our experiments. The parameters we used and the training procedure are described in Appendix C.4. We have experimentally confirmed that the application of data augmentation upon training the Glow is essential for high detection performance. We applied the data augmentation of random 2×2 translation and horizontal flipping on C-10 and TIN. For ILSVRC2012, the image is first resized to be 256 in height or width, whichever is shorter, and cropped to 224×224 at random.

Competitors. We compare the performance of the proposed methods to several existing methods. We implement nine competitors: the Watanabe-Akaike Information Criterion (WAIC) [13], the likelihood-ratio test (LLR) [14], the Complexity-aware likelihood test (COMP) [15], the typicality test in latent space (TTL) [13], the Maximum Softmax Probability (MSP) [1], the Dropout Uncertainty (DU) [47], the Feature Squeezing (FS) [48], the Pairwise Log-Odds (PL) [49], and the reconstruction error in Auto-Encoder (AE). See Appendix C.5 for a description of each method and its implementation. For the likelihood-based methods (i.e., WAIC, LLR, and COMP) and TTL, the same Glow model used in our method is used. For the classifier-based methods (i.e., MSP, DU, FS, and PL), the classifier is the WRN 28-10 or ResNet-50 v2 which is the same model we used to craft the adversarial examples in the previous section.

5.3. Results

We measure the success rate of OOD detection using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR). The higher is better for both. The results in AUPR are presented in Appendix D.1. We denote the reconstruction errors in NFs without the penalty ξ by RE. We also show the histograms in Fig. 3 for PRE (i.e., R_ξ) and Fig. 7 (in Appendix D.4) for RE (i.e., R). As for λ in Eq. (4), we empirically chose $\lambda = 50$ for CIFAR-10 and $\lambda = 100$ for TinyImageNet and ILSVRC2012. (The performance with different λ is presented in Appendix D.2.)

CIFAR-10 and TinyImageNet. Tables 1 and 2 show AUROC for C-10 and TIN. The PRE performed best for the majority of OOD datasets (the best scores are shown in bold). Importantly, unlike the other methods, the PRE exhibited high performance over all the cases: the columns of Avg. show that the PRE significantly outperformed the existing methods in average scores. When the detection method is deployed for real-world applications with no control over their input data, having no specific weakness is a strong advantage of PRE. The RE showed the second-best performance after the PRE on C-10. On TIN, while the RE performed well in the cases where the TTL failed

Table 1: AUROC (%) on CIFAR-10. The column labeled as 'Avg.' shows the averaged scores.

	CelebA	TIN	Bed	Living	Tower	PGD-2	PGD-8	CW-0	CW-10	Noise-1	Noise-2	Avg.
WAIC	50.36	77.94	77.25	84.76	79.35	41.82	73.23	47.90	46.13	**100.0**	**100.0**	70.79
LLR	59.77	38.05	33.33	31.42	46.01	59.10	76.09	52.28	54.36	0.80	0.61	41.07
COMP	72.00	82.01	78.87	87.82	5.03	60.69	98.81	51.27	53.02	**100.0**	**100.0**	71.77
TTL	84.87	84.19	90.39	91.36	89.68	75.22	98.99	51.14	54.15	**100.0**	54.80	79.52
MSP	79.32	91.00	93.83	91.67	82.05	23.85	0.0	**98.94**	5.17	98.25	96.27	69.12
PL	81.13	63.18	54.23	49.39	59.87	78.17	97.00	56.82	80.04	24.26	77.07	65.56
FS	83.35	88.90	89.16	88.23	94.71	90.86	72.47	93.76	94.43	91.99	96.34	89.47
DU	84.64	86.33	86.53	84.76	82.04	74.62	25.54	89.33	80.84	81.09	84.61	78.21
AE	67.36	80.28	73.71	87.01	7.83	50.69	61.80	50.02	50.07	**100.0**	99.52	66.21
RE (ours)	92.53	94.19	95.92	95.83	94.35	91.66	94.58	96.08	95.09	97.45	95.80	94.86
PRE (ours)	**93.62**	**95.74**	**97.43**	**97.52**	**95.88**	**92.23**	**99.93**	95.00	**95.21**	**100.0**	96.64	**96.29**

Table 2: AUROC (%) on TinyImageNet. The column labeled as 'Avg.' shows the averaged scores.

	CelebA	Bed	Living	Tower	PGD-2	PGD-8	CW-0	CW-10	Noise-1	Noise-2	Avg.
WAIC	11.92	63.54	67.75	72.87	40.49	49.78	48.44	46.05	**100.0**	**100.0**	60.08
LLR	92.76	69.95	70.19	78.27	58.45	96.78	51.66	53.86	50.57	0.0	62.25
COMP	39.83	48.02	61.61	46.72	55.98	95.01	50.54	52.11	**100.0**	**100.0**	64.98
TTL	**97.51**	98.78	99.36	**98.68**	83.47	**100.0**	51.89	57.73	**100.0**	**100.0**	88.74
MSP	76.88	77.51	73.38	73.91	6.91	0.0	69.09	4.20	68.34	74.76	52.50
PL	49.74	27.08	26.26	28.22	94.24	99.97	34.76	87.66	15.04	31.63	49.46
FS	29.21	27.26	29.36	25.70	71.98	16.86	50.64	82.21	42.17	50.27	42.57
DU	47.06	37.12	32.52	27.98	50.94	68.21	49.02	50.83	48.99	22.47	43.51
AE	14.92	20.90	32.94	23.74	49.89	52.12	50.00	50.03	95.45	70.37	46.04
RE (ours)	46.68	61.97	62.26	59.51	92.86	92.94	**92.43**	93.26	98.53	98.45	79.89
PRE (ours)	95.55	**99.01**	**99.46**	97.37	**95.04**	**100.0**	92.42	**94.93**	**100.0**	**100.0**	**97.38**

(i.e., CW, which we discuss in Section 6.2), the RE performed poorly in CelebA, Bed, Living, and Tower. Notably, however, the penalty ξ proved to be remarkably effective in those cases, and the PRE combined with ξ improved significantly. From the comparison of PRE and RE in the tables, we see that the performance is improved by ξ. It is shown that the performance of likelihood-based methods (WAIC, LLR, and COMP) is greatly dependent on the type of OOD. On C-10, the performance of classifier-based methods is relatively better than the likelihood-based methods, however, their performance was significantly degraded on TIN. In accordance with the increase in the number of classification classes from 10 (in C-10) to 200 (in TIN), the classifier's performance decreased, which caused the detection performance to decrease. It may be the weakness specific to the classifier-based methods. The performance of AE was at the bottom. Similar results were observed in AUPR as well.

ILSVRC2012. Table 3 shows the AUROC for ILSVRC2012. It suggests that the proposed methods perform well even for large-size images. We found that the reconstruction error alone (i.e., RE) could achieve high performance and the effect of the penalty was marginal on ILSVRC2012.

6. Discussion

6.1. Analysis with Tail Bound

In Section 4.1 we explained how R (and R_ξ) increase for OOD examples $\mathbf{x}_{\text{ood}}$ close to the manifold $\mathcal{M}$. This section discusses how the proposed methods detect $\mathbf{x}_{\text{ood}}$ far from $\mathcal{M}$, using the tail bound.

OOD examples are assigned minimal probabilities in latent space. The intensity of the penalty for a particular input $\mathbf{x}$, $\xi(f(\mathbf{x}))$, depends on how its L_2 norm in the latent space $\mathcal{Z} \in \mathbb{R}^d$ (i.e.,$\|\mathbf{z}\| = \|f(\mathbf{x})\|$) deviates from $\sqrt{d}$, as Eq. (4). We show the histograms for $\|\mathbf{z}\|$ in Fig. 4. We note that $\sqrt{d}$ is about 55.43 for C-10, 110.85 for TIN, and 387.98 for ILSVRC2012. Thus, we see that the distribution modes for In-Dist examples are consistent approximately with the theoretical value, $\sqrt{d}$ (though it is slightly biased toward larger values on ILSVRC2012). At the same time, we see that $\|\mathbf{z}\|$ for $\mathbf{x}_{\text{ood}}$ deviate from $\sqrt{d}$, except for CW's examples. We assess the degree of this deviation with the Chernoff tail bound for the L_2 norm of i.i.d. standard Gaus-

Table 3: AUROC (%) on ILSVRC2012 with our methods. The column labeled as 'Avg.' shows the averaged scores.

	CelebA	PGD-2	PGD-8	CW-0	CW-10	Noise-2	Noise-32	Avg.
RE	94.65	93.96	96.42	**94.59**	94.87	97.24	**97.68**	95.63
PRE	**94.89**	**94.24**	**96.66**	94.58	**95.64**	**97.75**	**97.68**	**95.92**

sian vector $\mathbf{z} \in \mathbb{R}^d$: for any $\epsilon \in (0, 1)$ we have

$$\Pr\left[d(1-\epsilon) < \|\mathbf{z}\|^2 < d(1+\epsilon)\right] \geq 1 - 2\exp\left(-\frac{d\epsilon^2}{8}\right). \quad (6)$$

See Appendix E for the derivation of Eq. (6) and the analysis described below. When $d = 3072$ (i.e., on C-10) and we set $\epsilon = 0.32356413$, we have $\Pr[\|\mathbf{z}\| > 63.765108] \leq \frac{1}{2^{58}}$, for instance. It tells us that the probability that a vector $\mathbf{z} \in \mathbb{R}^{3072}$ with its L_2 norm 63.87, which corresponds to the median value of $\|\mathbf{z}\|$ over 1024 PGD-8 examples on C-10 we used, occurs in $P\mathrm{z}$ (i.e., $\mathcal{N}(0, \mathbf{I}_{3072})$) is less than $\frac{1}{2^{58}} = 3.4694\mathrm{e}{-18}$. As another example, the median value of $\|\mathbf{z}\|$ for 1024 CelebA (OOD) examples in $\mathcal{Z}$ built with TIN (In-Dist) is 116.81. The probability of observing a vector $\mathbf{z} \in \mathbb{R}^{12288}$ sampled from $\mathcal{N}(0, \mathbf{I}_{12288})$ with its L_2 norm 116.81 is less than $\frac{1}{2^{26}} = 1.4901\mathrm{e}{-08}$, as Eq. (6) gives $\Pr[\|\mathbf{z}\| > 116.700553] \leq \frac{1}{2^{26}}$ with $\epsilon = 0.108318603$. As such, the tail bound shows that those (OOD) examples are regarded as extremely rare events in $P\mathrm{z}$ built with the In-Dist training data.

Small probabilities increase the reconstruction error regardless of the distance to the manifold. The above observation implies the following: (OOD) examples not included in the typical set of $P\mathrm{z}$ are assigned extremely small probability. It leads to a decrease in the number of significant digits of the probabilistic density $p(\mathrm{z})$ in the transformation of the NF, and it may cause the rounding error of floating-point. Those rounding errors increase $\delta_{\mathbf{x}}$ and $\delta_{\mathbf{z}}$ in Eq. (1), increasing the reconstruction error in NFs, R (and hence R_ξ that the PRE uses). In other words, it suggests that atypical examples in $P\mathrm{z}$ will have larger R (and R_ξ), independent of the distance to $\mathcal{M}$. The Noise-κ dataset samples used in our experiments (Section 5) are possibly far from $\mathcal{M}$. Both PRE and RE showed high detection performance even against Noise-κ datasets. We understand that the mechanism described here is behind this success.

6.2. Analysis of Typicality Test Failures

Lastly, we analyze why the typicality test (TTL) failed to detect CW's adversarial examples (and some Noise datasets) (Tables 1 and 2). We addressed it by arbitrarily partitioning latent vectors $\mathbf{z} \in \mathbb{R}^{3072}$ on C-10 into two parts as $[\mathbf{z}_a, \mathbf{z}_b] = \mathbf{z}$ where $\mathbf{z}_a \in \mathbb{R}^{2688}$ and $\mathbf{z}_b \in \mathbb{R}^{384}$, and measuring the L_2 norm separately for each. (Due to space limitations, see Appendix D.3 for details of the experiment.) We then found that the deviation from the In-Dist observed in $\mathbf{z}_a$ and $\mathbf{z}_b$ cancels out, and as a result, $\|\mathbf{z}\|$ of CW's examples becomes indistinguishable from those of In-Dist ones. This is exactly the case described in Section 3.2 where the TTL fails. The typicality-based penalty in the PRE is therefore ineffective in these cases. However, in the calculation of the reconstruction error (RE), which is the basis of the PRE, the information of $\mathbf{z}$ is used as-is without being reduced to the L_2 norm, and the information on the deviation from the In-Dist in each dimension is preserved. Consequently, it enables a clear separation between In-Dist and OOD

7. Limitation

1) The penalty ξ we introduced is by design ineffective for OOD examples for which the TTL is ineffective. However, we experimentally show that it improves performance in most cases. 2) The adversarial examples have been shown to be generated off the data manifold [37]; however, it has also been shown that it is possible to generate the ones lying on the manifold deliberately[37, 34, 35]. Since the detection target of our method is off-manifold inputs, we left such on-manifold examples out of the scope in this work. If it is possible to generate examples that are on-manifold and at the same time typical in the latent space, maybe by 'adaptive attacks' [71], it would be difficult to detect them with the PRE. This discussion is not limited to adversarial examples but can be extended to OOD in general, and we leave it for future work.

8. Conclusion

We have presented PRE, a novel method that detects OOD inputs lying off the manifold. As the reconstruction error in NFs increases regardless of whether OOD inputs are close to or far from the manifold, PRE can detect them by measuring the magnitude of the reconstruction error. We further proposed a technique that penalizes the atypical inputs in the latent space to enhance detection performance. We demonstrated state-of-the-art performance with PRE on CIFAR-10 and TinyImageNet and showed it works even on the large size images, ILSVRC2012.

References

[1] 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.

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

[3] 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.

[4] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pages 7167–7177. Curran Associates, Inc., 2018.

[5] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[6] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[8] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

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

[10] 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.

[11] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey, 2021.

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

[13] Hyunsun Choi, Eric Jang, and Alexander A. Alemi. Waic, but why? generative ensembles for robust anomaly detection, 2018.

[14] 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 *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[15] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020.

[16] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3232–3240. PMLR, 13–15 Apr 2021.

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

[18] Amirhossein Ahmadian and Fredrik Lindsten. Likelihood-free out-of-distribution detection with invertible generative models. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2119–2125. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[19] Federico Bergamin, Pierre-Alexandre Mattei, Jakob Drachmann Havtorn, Hugo Sénétaire, Hugo Schmutz, Lars Maaløe, Soren Hauberg, and Jes Frellsen. Model-agnostic out-of-distribution detection using combined

statistical tests. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 10753–10776. PMLR, 28–30 Mar 2022.

[20] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality, 2020.

[21] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12427–12436. PMLR, 18–24 Jul 2021.

[22] Ziyu Wang, Bin Dai, David Wipf, and Jun Zhu. Further analysis of outlier detection with deep generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8982–8992. Curran Associates, Inc., 2020.

[23] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[24] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. In *International Conference on Learning Representations*, 2015.

[25] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

[26] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

[27] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems 23*, pages 1786–1794. Curran Associates, Inc., 2010.

[28] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[29] Chenlin Meng, Jiaming Song, Yang Song, Shengjia Zhao, and Stefano Ermon. Improved autoregressive modeling with distribution smoothing. In *International Conference on Learning Representations*, 2021.

[30] Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Joern-Henrik Jacobsen. Understanding and mitigating exploding inverses in invertible neural networks. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1792–1800. PMLR, 13–15 Apr 2021.

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[32] 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 (IJCV)*, 115(3):211–252, 2015.

[33] Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

[34] Ajil Jalal, Andrew Ilyas, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

[35] Justin Gilmer, Luke Metz, Fartash Faghri, Sam Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres, 2018.

[36] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021.

[37] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

[39] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against

adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

[40] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 135–147, New York, NY, USA, 2017. Association for Computing Machinery.

[41] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[42] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[43] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.

[44] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[45] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[46] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[47] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[48] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security (NDSS) Symposium*, NDSS 18, February 2018.

[49] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5498–5507, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[50] Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[51] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, volume 29, pages 4790–4798. Curran Associates, Inc., 2016.

[52] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[53] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[54] Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2133–2143. PMLR, 13–18 Jul 2020.

[55] Ugo Tanielian, Thibaut Issenhuth, Elvis Dohmatob, and Jeremie Mary. Learning disconnected manifolds: a no GAN's land. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9418–9427. PMLR, 13–18 Jul 2020.

[56] Alexandre Verine, Yann Chevaleyre, Fabrice Rossi, and benjamin negrevergne. On the expressivity of bi-lipschitz normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[57] Laurent Dinh, Jascha Sohl-Dickstein, Razvan Pascanu, and Hugo Larochelle. A RAD approach to deep mixture models. *CoRR*, abs/1903.07714, 2019.

[58] Cheng Lu, Jianfei Chen, Chongxuan Li, Qiuhao Wang, and Jun Zhu. Implicit normalizing flows. In *International Conference on Learning Representations*, 2021.

[59] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[60] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[61] Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The reversible residual network: Backpropagation without storing activations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[62] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[63] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

[64] Huaibo Huang, zhihang li, Ran He, Zhenan Sun, and Tieniu Tan. Introvae: Introspective variational autoencoders for photographic image synthesis. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[65] Genki Osada, Budrul Ahsan, Revoti Prasad Bora, and Takashi Nishide. Regularization with latent space virtual adversarial training. In *Computer Vision – ECCV 2020*, pages 565–581, Cham, 2020. Springer International Publishing.

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

[67] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015.

[68] 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.

[69] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[70] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

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

[72] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[74] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31*, pages 10215–10224. Curran Associates, Inc., 2018.

[75] Krzysztof Kolasinski. *An implementation of the GLOW paper and simple normalizing flows lib*, 2018.

[76] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[77] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[78] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[79] Carlos Fernandez-Granda. Optimization-based data analysis: Lecture notes 3: Randomness, Fall 2017.