Understanding Likelihood of Normalizing Flow and Image Complexity through the Lens of Out-of-Distribution Detection

Genki Osada[1], **Tsubasa Takahashi**[1], **Takashi Nishide**[2]
[1]LINE Corporation*
[2]University of Tsukuba
{genki.osada, tsubasa.takahashi}@lycorp.co.jp, nishide@risk.tsukuba.ac.jp

Abstract

Out-of-distribution (OOD) detection is crucial to safety-critical machine learning applications and has been extensively studied. While recent studies have predominantly focused on classifier-based methods, research on deep generative model (DGM)-based methods have lagged relatively. This disparity may be attributed to a perplexing phenomenon: DGMs often assign higher likelihoods to unknown OOD inputs than to their known training data. This paper focuses on explaining the underlying mechanism of this phenomenon. We propose a hypothesis that less complex images concentrate in high-density regions in the latent space, resulting in a higher likelihood assignment in the Normalizing Flow (NF). We experimentally demonstrate its validity for five NF architectures, concluding that their likelihood is untrustworthy. Additionally, we show that this problem can be alleviated by treating image complexity as an independent variable. Finally, we provide evidence of the potential applicability of our hypothesis in another DGM, PixelCNN++.

1 Introduction

Deep neural network (DNN) models deployed in real-world systems often encounter out-of-distribution (OOD) inputs — samples from a different distribution of the training set. DNN models often make incorrect predictions with high confidence for OOD inputs (Nguyen, Yosinski, and Clune 2015), making it crucial to distinguish OOD inputs from in-distribution (In-Dist) ones at test time, especially for safety-critical applications such as autonomous driving (Du et al. 2022) and medical diagnosis (Linmans, van der Laak, and Litjens 2020). Numerous methods have been proposed to improve empirical performance (Yang et al. 2022, 2021), and analytical approaches have also been studied (Morteza and Li 2022; Liu et al. 2020).

One major approach to OOD detection is using deep generative models (DGMs). Particularly, Normalizing Flows (NFs) and Autoregressive (AR) models are the two most commonly selected DGMs due to their ability to compute exact model likelihoods. The DGM-based approach is attractive for its strengths: it does not require labeled data to train detection models, and the performance of OOD detection is independent of the number of classification classes (Huang and Li 2021). However, most of the recent progress in this field has been based on another prevailing approach, the classifier-based method (Yang et al. 2022, 2021). The primary factor hindering the progress of DGM-based methods may be the observation presented by (Nalisnick et al. 2019; Choi, Jang, and Alemi 2018): DGMs were expected to assign a higher likelihood, i.e., probabilistic density, to In-Dist inputs than to OOD inputs, but it turned out that this is not the case in some particular cases. We refer to this phenomenon as the *failure of likelihood*. When an OOD detection method is deployed to real-world applications with no control over its inputs, this unreliability is unacceptable. Above all, correctly estimating the likelihood is a fundamental feature that DGMs are expected to fulfill. Hence, research on DGM-based methods has focused on addressing this issue before enhancing detection performance and tackling more challenging tasks.

Several hypotheses have been proposed to explain this phenomenon. One argues that image complexity has some influence on the likelihood (Nalisnick et al. 2019; Serrà et al. 2020), while another suggests the need to account for the notion of the *typical set* (Choi, Jang, and Alemi 2018; Nalisnick et al. 2020). However, these methods still fail in specific cases, as we will show later, and the mechanism behind the failure of likelihood remains unclear.

In this paper, we address the cause of the failure of likelihood. We focus on Normalizing Flows due to their analytical tractability. The contributions of this work are as follows:

- We present a hypothesis that explains how image complexity can arbitrarily control the likelihood of NFs and how the density concentration in latent distribution causes this phenomenon.
- Our hypothesis provides a unified explanation for two separate questions: why images with less complexity are assigned high likelihood and why OOD inputs are regarded as the typical set samples in some cases.
- We experimentally verify our argument with five different NF architectures and conclude that OOD detections based on the likelihood of NFs are untrustworthy.

The relationship between the likelihood and image complexity that we explain in this paper finds implicit support

*The company name was changed to LY Corporation after submission.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

in Independent Component Analysis (ICA) (Hyvärinen and Oja 2000), as well as in Shannon's source coding theorem (Shannon 1948), as discussed later in this section.

Additionally, we propose a countermeasure and argue that the failure of likelihood can be overcome by leveraging information about the root cause of the problem, namely image complexity. We demonstrate that a Gaussian mixture model (GMM) with two variables, image complexity and likelihood in latent space, can detect all of the OOD datasets we used in our evaluation, thereby highlighting the validity of our argument from another perspective.

Finally, given a recent study in which AR models can be seen as a specific type of NF (Nielsen and Winther 2020), we show the potential applicability of our hypothesis to AR models as well through the experiments using PixelCNN++.

Related works. Two main hypotheses have been proposed to explain the failure of likelihood-based OOD detection: the involvement of *image complexity* and the necessity of considering *typical set*. Experimental studies by (Serrà et al. 2020; Ren et al. 2019) demonstrated that the log-likelihood $\log p(\mathbf{x})$ increases with less complex images, but the underlying mechanism remained unclear. (Schirrmeister et al. 2020; Kirichenko, Izmailov, and Wilson 2020) attributed this phenomenon to the architecture of the NFs. However, the same phenomenon has also been observed in PixelCNN (Nalisnick et al. 2019; Ren et al. 2019), implying that it is not exclusively caused by the architecture of the NFs. (Nalisnick et al. 2019) suggested that the phenomenon could be theoretically explained. Their analysis implicitly assumed that the likelihood in the latent space and the determinant of the Jacobian matrix remain consistent across different datasets. However, as shown in Figs. 3 (right) and 4, in reality, these values differ considerably between datasets. We explain that this variation arises from differences in image complexities, ultimately serving as a factor that can arbitrarily influence $\log p(\mathbf{x})$. We experimentally show that our hypothesis likely applies to PixelCNN as well.

(Choi, Jang, and Alemi 2018; Nalisnick et al. 2020) introduced a testing method based on the typical set in latent space. However, (Wang et al. 2020; Choi, Jang, and Alemi 2018; Zhang, Goldstein, and Ranganath 2021) concluded that its performance was insufficient. Our hypothesis also explains *why* typicality-based testing fails. Additionally, in Appendix A, we present other DGM-based approaches and an overview of classifier-based detection methods.

Beyond the context of OOD detection, the credibility of likelihood has been discussed. (Theis, van den Oord, and Bethge 2016) has shown that high likelihood does not necessarily indicate high sample quality produced by the model. (Perkiö and Hyvärinen 2009) proposed that image complexity can be estimated by $\log p(\mathbf{x})$ of ICA (Hyvärinen and Oja 2000), which can be regarded as a linear NF (Dinh, Krueger, and Bengio 2015). Their study implies that less complex image $\mathbf{x}$ is assigned a higher $\log p(\mathbf{x})$ in ICA, aligning with our assertion. Moreover, Shannon's source coding theorem (Shannon 1948) states that the expected code length of a symbol $\mathbf{x}$ is bounded by its entropy: $\mathbb{E}_{\mathbf{x}\sim P_{\mathrm{x}}}[-\log p(\mathbf{x})] \leq \mathbb{E}_{\mathbf{x}\sim P_{\mathrm{x}}}[L(\mathbf{x})]$. Here, $L(\mathbf{x})$ represents the length of the encoded message for $\mathbf{x}$ using a lossless compression algorithm, and P_{x} denotes a data distribution, which is generally unknown. Minimizing $\mathbb{E}_{\mathbf{x}\sim P_{\mathrm{x}}}[L(\mathbf{x})]$ is achieved by assigning a smaller $L(\mathbf{x})$ to $\mathbf{x}$ with a larger $\log p(\mathbf{x})$ and a larger $L(\mathbf{x})$ to $\mathbf{x}$ with a smaller $\log p(\mathbf{x})$, ensuring efficient encoding (Bishop and Nasrabadi 2006). The complexity of an image $\mathbf{x}$ is defined in the next section as $L(\mathbf{x})/d$, where $\mathbf{x}$ with less complexity have a smaller $L(\mathbf{x})$. Therefore, the theorem implies that images with less complexity are assigned higher likelihoods, aligning with our claim.

2 Background and Problem Statement

This paper shows that the complexity of input images can control the likelihood of Normalizing Flows (NFs). We first define the image complexity and describe the NFs. Then, we discuss existing approaches and their failures, leading us to articulate the problem statement addressed in this paper.

2.1 Preliminary

Image complexity and entropy coding. Estimating the complexity of images lacks a definitive method; however, two common options are available: Kolmogorov complexity and Shannon's entropy (Rigau, Feixas, and Sbert 2005). While a connection is suggested between the two (Grünwald and Vitányi 2003), the former is uncomputable. Thus, the prevalent method is based on entropy, which is directly used in lossless compression (coding) (Sayood 2017; Chen et al. 2022). In this work, we use the following definition for image complexity used in previous studies (Serrà et al. 2020; Ahmadian and Lindsten 2021).

Definition 1 (Image complexity)**.** *Let $L(\mathbf{x})$ be the length of the bit string obtained after compressing $\mathbf{x} \in \{0, 1, \ldots, 255\}^d$ using a lossless compression algorithm,* comp. *Image complexity for $\mathbf{x}$ is defined as $C(\mathbf{x}) = \frac{1}{d}L(\mathbf{x})$. The more complex $\mathbf{x}$ is, the larger $C(\mathbf{x})$ is, and vice versa.*

We use the JPEG2000 compression, which is based on entropy coding, as the comp in all our experiments.

Normalizing Flow. We focus our investigation on NFs (Rezende and Mohamed 2015; Dinh, Krueger, and Bengio 2015; Dinh, Sohl-Dickstein, and Bengio 2017) among DGMs. NFs offer distinct advantages over other types of DGMs, as they enable the exact computation of likelihoods, unlike VAEs, and allow for the separation of the volume term, facilitating analysis (Nalisnick et al. 2019). NFs learn an invertible mapping $f : \mathcal{X} \to \mathcal{Z}$ that maps observable data $\mathbf{x}$ to the latent vector $\mathbf{z} = f(\mathbf{x})$, where $\mathcal{X} \in \mathbb{R}^d$ is the data space and $\mathcal{Z} \in \mathbb{R}^d$ is the latent space. We denote a distribution on $\mathcal{Z}$ as P_{z} with probability density $p(\mathbf{z})$. NFs learn an approximate model distribution $\widehat{P}_{\mathrm{x}}$ to match the unknown true distribution P_{x} on $\mathcal{X}$. Under the change of variable rule, the log density of $\widehat{P}_{\mathrm{x}}$ is expressed as:

$$\log p(\mathbf{x}) = \log p(\mathbf{z}) + \log \left|\det J_f(\mathbf{x})\right| \tag{1}$$

where $J_f(\mathbf{x}) = df(\mathbf{x})/d\mathbf{x}$ is the Jacobian matrix of f at $\mathbf{x}$, and $\log \left|\det J_f(\mathbf{x})\right|$ is referred to as the *volume*. By maximizing $\log p(\mathbf{z})$ and $\log \left|\det J_f(\mathbf{x})\right|$ simultaneously with respect to samples $\mathbf{x} \sim P_{\mathrm{x}}$, the NF model f is trained to match $\widehat{P}_{\mathrm{x}}$

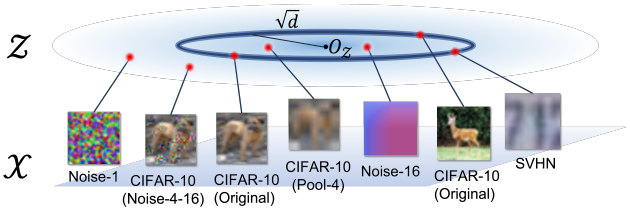

Figure 1: DCAS (Remark 1) attracts less complex images to the high-density region in latent space. $\mathcal{Z}$ represents a Gaussian latent space trained on CIFAR-10. $O_{\mathcal{Z}}$ represents the origin of $\mathcal{Z}$. The dark blue circle represents the typical set in $\mathcal{Z}$, identified as In-Dist by the typicality test. Complex OOD images like Noise-1 and CIFAR-10 (Noise-4-16) are mapped far from $O_{\mathcal{Z}}$. However, due to DCAS, less complex images like Noise-16 and CIFAR-10 (Pool-4) are mapped closer to $O_{\mathcal{Z}}$ than the circle. We hypothesize that SVHN (an OOD image) should be mapped far beyond the circle. However, due to its less complexity, it is attracted towards $O_{\mathcal{Z}}$ and coincidentally falls on the circle, leading to the misidentification of SVHN as In-Dist by the typicality test.

with P_{x}. In our work, P_{x} represents In-Distribution (In-Dist), and we train an NF model using samples from In-Dist.

We investigate five architectures: Glow (Kingma and Dhariwal 2018), CV-Glow (Nalisnick et al. 2019), iResNet (Behrmann et al. 2019), ResFlow (Chen et al. 2019), and IDF (Hoogeboom et al. 2019). iResNet and ResFlow improve the stability of NFs by controlling the Lipschitz constant of f to be less than one while maintaining high expressive power. IDF uses a categorical distribution for P_{z}, while the other four architectures use the standard Gaussian distribution. CV-Glow, as the abbreviation for *Glow with the constant volume*, has a fixed volume that depends only on the weight matrix of the 1×1 convolutions, resulting in a constant volume across all inputs $\mathbf{x}$. Similarly, the volume of IDF is fixed at 0 to build a latent space with integer values. It is worth noting that CV-Glow, IDF, and even PixelCNN++ (Salimans et al. 2017) exhibit a similar behavior due to the commonality of the constant volume, as we will see later. The implementation of these models are described in Appendix D.3.

2.2 Failure of Existing Approaches

Likelihood test. The likelihood test, introduced by (Bishop 1995), is an OOD detection method that relies a density estimation model. Treating the probabilistic density of input as a likelihood, it assumes that OOD examples would be assigned lower likelihoods compared to In-Dist examples. However, counter-evidence presented by (Nalisnick et al. 2019; Choi, Jang, and Alemi 2018) showed that DGMs trained on CIFAR-10 assigned higher likelihoods to samples from SVHN (OOD) than to samples from CIFAR-10 (In-Dist), for instance. This has sparked controversy and led to the proposal of two approaches described below as methods for improvement.

Complexity-aware likelihood test. Several studies have linked the failure of the likelihood test to image complexity (Nalisnick et al. 2019; Serrà et al. 2020; Ren et al. 2019; Schirrmeister et al. 2020; Kirichenko, Izmailov, and Wilson 2020). These studies experimentally showed that low-complexity regions in images contribute to an increase in the likelihood $\log p(\mathbf{x})$. One such study by (Serrà et al. 2020) introduced a complexity-aware likelihood test (CALT), which directly incorporates image complexity $C(\mathbf{x})$. Specifically, CALT computes the score $S_{\mathrm{CALT}}(\mathbf{x}) = \log p(\mathbf{x}) + C(\mathbf{x})$. However, our experiments have revealed that CALT still exhibits failures in certain cases, indicating that merely adding $C(\mathbf{x})$ to $\log p(\mathbf{x})$ is insufficient for effective OOD detection. (See Appendix B.2 and Section 5.1.)

Typicality test. (Choi, Jang, and Alemi 2018; Nalisnick et al. 2020) attributed the failure of the likelihood test to an ignorance of the notion of the *typical set*. In a d-dimensional isotropic Gaussian $\mathcal{N}(0, \mathbf{I}_{\mathrm{d}})$, the typical set is expected to reside in a hypersphere with a radius of $\sqrt{d}$ with high probability, rather than around its mean (See Appendix B.1). Utilizing this property, (Choi, Jang, and Alemi 2018; Nalisnick et al. 2020) have proposed a method that identifies a test input $\mathbf{x}$ as OOD when it fall outside of the distribution's typical set, which we refer to as the typicality test in latent space (TTL). TTL computes the score $S_{\mathrm{TTL}}(\mathbf{x}) = \mathtt{abs}(\|\mathbf{z}\| - \sqrt{d})$, where $\mathbf{z} = f(\mathbf{x})$ and f is a trained NF model. If $\mathbf{x}$ is OOD, $S_{\mathrm{TTL}}(\mathbf{x})$ is expected to be large. However, (Zhang, Goldstein, and Ranganath 2021; Wang et al. 2020; Choi, Jang, and Alemi 2018) concluded that the performance of TTL was insufficient, and our experiments also revealed failure cases. We observed that $\|\mathbf{z}\|$ for In-Dist samples is concentrated around the theoretical value, $\sqrt{d}$ (i.e., $\sqrt{32 \times 32 \times 3} \simeq 55.4$ for CIFAR-10 and SVHN). This indicates that NFs can correctly perceive test samples from In-Dist as the typical set. However, in the case where the In-Dist is CIFAR-10, $\|\mathbf{z}\|$ for SVHN (OOD) is also highly concentrated around 55.4, indicating the failure of TTL to detect SVHN samples as OOD. This situation is depicted in Fig. 1. (For experimental results, see Fig. 5 (bottom two) in Appendix B.2.) It becomes evident that even OOD inputs can reside within the typical set, challenging the effectiveness of the typicality test.

2.3 Problem Statement

The aforementioned issues can be summarized in the following two questions:

- Why does less image complexity result in the failure of the likelihood test?
- Why are OOD inputs often misclassified as typical set samples?

We address these questions in this study.

3 Hypothesis and Experimental Validation

We present Hypothesis 1, which comprehensively explains the questions posed in Section 2.3. Subsequently, we provide experimental results that support the hypothesis.

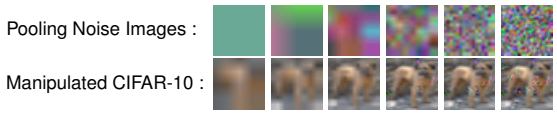

Figure 2: Complexity controlled images. Image complexity increases from left to right in both rows. Top: *Pooling noise images* with pooling size κ decreases as 32, 16, 8, 4, 2, and 1 from left to right. Bottom: *Manipulated CIFAR-10* with Pool-8, Pool-4, Pool-2, Noise-4-4, Noise-4-8, and Noise-4-16 from left to right.

3.1 Our Hypothesis

Definition 2 (Local Lipschitz continuity)**.** *For a subset $\mathcal{A} \subset \mathcal{Z}$, we define an invertible function $f : \mathcal{X} \to \mathcal{Z}$ as locally $L_{\mathcal{A}}$-Lipschitz as follows:*

$$\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\| \leq L_{\mathcal{A}} \|\mathbf{x}_1 - \mathbf{x}_2\|, \forall f(\mathbf{x}_1), f(\mathbf{x}_2) \in \mathcal{A}. \quad (2)$$

Assumption 1. *We consider that f is implemented by an NF and assume that the latent space $\mathcal{Z}$ is semantically continuous (Dinh, Sohl-Dickstein, and Bengio 2017).*

Hypothesis 1. *Let $f : \mathcal{X} \to \mathcal{Z}$ be an invertible function locally $L_{\mathcal{A}}$-Lipschitz for $\mathcal{A} \subset \mathcal{Z}$. For all $\mathbf{z}' \in \mathcal{A}$, let $\mathbf{x}' = f^{-1}(\mathbf{z}')$. Also, let $\mathcal{B}^{\epsilon}_{\mathbf{z}} = \{\mathbf{z}' \in \mathcal{A} : \|\mathbf{z}' - \mathbf{z}\| < \epsilon\}$ and $\mathcal{B}^{\epsilon}_{\mathbf{z}} \subset \mathcal{A}$ for a constant $\epsilon \geq 0$. Then, letting $C(\mathbf{x})$ be the image complexity of $\mathbf{x}$ (Definition 1) and C_1 be a constant, we have*

$$\frac{\epsilon^2}{{L_{\mathcal{A}}}^2} \left(1 - \mathbb{P}(\mathcal{B}^{\epsilon}_{\mathbf{z}})\right) \leq C_1 \exp(C(\mathbf{x})). \quad (3)$$

Considering $\mathcal{A}$ to be a very small region in $\mathcal{Z}$ and based on Assumption 1, we posit that samples $\mathbf{x} = f^{-1}(\mathbf{z})$ and $\mathbf{x}' = f^{-1}(\mathbf{z}')$ for $\mathbf{z}, \mathbf{z}' \in \mathcal{A}$ share common semantics. This allows us to treat their complexities as approximately equal, i.e., $C(\mathbf{x}) \approx C(\mathbf{x}')$, which we use in the derivation. We derive Eq. (3) from experimental observations and present it in Appendix C.2. Assuming $\mathbf{x}$ follows a diagonal Gaussian distribution, we can analytically derive it as stated in Appendix C.1, resulting in the right-hand side being $\frac{d}{2\pi e}\delta_x^{\frac{2}{d}} \exp\left(2C(\mathbf{x})\right)$ where δ_x is the volume of bins used in discretization. Next, we present an observation necessary to state the following Remarks.

Observation 1. *A positive correlation exists between $\log p(\mathbf{z})$ and the volume $\log|det\ J_f(\mathbf{x})|$ in response to the variation of input $\mathbf{x}$.*

We defer the explanation to Section 3.2, Experiment 3. In the following Remarks, we consider that $\mathbf{z} = f(\mathbf{x})$ is determined for a given input $\mathbf{x}$ such that $\mathbb{P}(\mathcal{B}^{\epsilon}_{\mathbf{z}})$ satisfies the inequality in Eq. (3).

Remark 1. As $\mathbf{x}$ becomes less complex, i.e., as $C(\mathbf{x})$ becomes small, $\mathbb{P}(\mathcal{B}^{\epsilon}_{\mathbf{z}})$ becomes large. In order to increase $\mathbb{P}(\mathcal{B}^{\epsilon}_{\mathbf{z}})$, $\mathbf{z}$ needs to be in a high-density region in $\mathcal{Z}$, because the volume of $\mathcal{B}^{\epsilon}_{\mathbf{z}}$, $\text{vol}(\mathcal{B}^{\epsilon}_{\mathbf{z}})$, is fixed by ϵ. In other words, $\mathbf{z}$ for a less complex $\mathbf{x}$ will concentrate on a high-density region in $\mathcal{Z}$, meaning that $\log p(\mathbf{z})$ for them will be large. Specifically, when the distribution of $\mathcal{Z}$ is $\mathcal{N}(0, \mathbf{I}_{\text{d}})$, $\mathbf{z}$ is mapped to the region close to the origin, meaning that $\|\mathbf{z}\|$ will be small for a less complex $\mathbf{x}$ (See Fig. 1). We refer to this effect as Density Concentration Attraction for Simpleness, DCAS.

Remark 2. As the input becomes less complex, $L_{\mathcal{A}}$ becomes large. Since $|\text{det}\ J_f(\mathbf{x})| < L_{\mathcal{A}}^d$ (Federer 2014), the volume term, $\log|\text{det}\ J_f(\mathbf{x})|$, for a less complex input is allowed to become large.

Remark 3. From Observation 1 and Remarks 1, 2, the decrease in image complexity increases both $\log p(\mathbf{z})$ and $\log|\text{det} J_f(\mathbf{x})|$. Consequently, according to Eq. (1), $\log p(\mathbf{x})$ increases for a less complex input.

3.2 Experimental Results

To support the above Remarks, we conducted experiments using a dataset in which image complexity was systematically controlled.

Datasets. We constructed two datasets with controlled image complexity: *pooling noise images* and *manipulated CIFAR-10*. The details are provided in Appendix D.1, and samples are shown in Fig. 2. The pooling noise images are generated by applying average pooling with different filter sizes, denoted as κ, to random noise images. Each set of images is labeled as Noise-κ. For the manipulated CIFAR-10 dataset, the complexity increases from Noise-4-1 to Noise-4-32 and decreases from Pool-2 to Pool-16, compared to the original image. The NF models used in this section were trained on CIFAR-10 (Krizhevsky, Hinton et al. 2009).

Experiment 1: Complexity vs. $\log p(\mathbf{z})$. We examined Remark 1 using the five NF models introduced in Section 2.1. We used $\|\mathbf{z}\|$ $(\propto -\sqrt{\log p(\mathbf{z})})$ as a proxy for $\log p(\mathbf{z})$ for the convenience of considering the typical set (for the four NFs with Gaussian latent distribution). The plot for ResFlow on the manipulated CIFAR-10 is depicted in Fig. 3 (left). It shows that $\log p(\mathbf{z})$ increases (equivalently, $\|\mathbf{z}\|$ decreases) as the complexity decreases, providing support for Remark 1. Similar plots were observed for all other nine cases, as shown in Figs. 9 and 10 in Appendix E.1.

Experiment 2: $\log p(\mathbf{z})$ vs. volume. We examined Remark 2 using Glow, iResNet, and ResFlow, excluding CV-Glow and IDF, which have a constant (or zero) volume. Fig. 3 (right) presents the plot for ResFlow on the manipulated CIFAR-10. It shows that a decrease in complexity (from Noise-4-16 to Pool-32) causes an increase in the volume, providing support for Remark 2. Similar plots were observed for all other five cases, as shown in Fig. 11 in Appendix E.2. Notably, Remark 2 asserts the opposite of the speculation presented in a previous study (Ahmadian and Lindsten 2021) (that higher image complexity corresponds to a larger volume).

Experiment 3: Correlation between $\log p(\mathbf{z})$ and volume. Fig. 3 (right) also shows that the positive correlation between $\log p(\mathbf{z})$ $(\propto -\|\mathbf{z}\|^2)$ and the volume $\log|\text{det}\ J_f(\mathbf{x})|$. It indicates that the two increase simultaneously in a balanced manner, which we have already presented as Observation 1. The balance between the two is a result of the learning process of NFs. By maximizing Eq. (1), an NF is trained

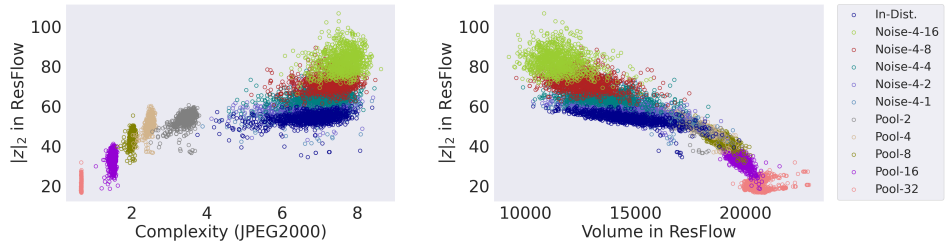

Figure 3: Plots for manipulated CIFAR-10. Left plot shows complexity vs. $\|\mathbf{z}\|$, supporting Remark 1. Right plot shows volume vs. $\|\mathbf{z}\|$, supporting Remark 2 and Observation 1. We note that $\|\mathbf{z}\| \propto -\sqrt{\log p(\mathbf{z})}$.

to reach an equilibrium between two opposing forces: 1) attracting $\mathbf{z}$ to high-density region in $\mathcal{Z}$ to maximize $\log p(\mathbf{z})$, and 2) scattering $\mathbf{z}$ to expand the volume $|\det J_f(\mathbf{x})|$ (Dinh, Krueger, and Bengio 2015; Behrmann et al. 2021). In other words, training an NF model involves optimizing the ratio between $\log p(\mathbf{z})$ and the volume to partition the variation in $\log p(\mathbf{x})$. The correlation strength between the two, as discussed in Appendix D.4, is dependent on the architecture of NFs. We utilize this knowledge in the selection of architectures for OOD detection in Section 5.1.

4 Likelihood is Untrustworthy

With Remarks presented in the previous section, we address the questions posed in Section 2.3 and assert the untrustworthiness of the likelihood of Normalizing Flows.

Why does less image complexity result in the failure of the likelihood test? Remark 3 directly answers this question. Inputs with less image complexity, even OOD ones, cause an increase in $\log p(\mathbf{x})$, leading to their misidentification as In-Dist. While some studies have experimentally shown that a decrease in the image complexity causes an increase in $\log p(\mathbf{x})$, we have formulated the underlying mechanism in Eq. (3) and identified that the culprit is the effect caused by a density concentration in latent distribution, which we refer to as DCAS in Remark 1.

Why are OOD inputs often misclassified as typical set samples? Based on Remark 1, we explain the mechanism underlying the failure of TTL. We provide an illustration in Fig. 1 for better intuition. We posit that an NF trained on In-Dist data inherently assigns a smaller $\log p(\mathbf{z})$ to an OOD input $\mathbf{x}_{\text{ood}}$ compared to an In-Dist input $\mathbf{x}_{\text{in}}$. In the case of NFs with a Gaussian latent space, the NF, f, attempts to map $\mathbf{z}_{\text{ood}} = f(\mathbf{x}_{\text{ood}})$ to a region farther than the distance $\sqrt{d}$ from the origin, where the typical set samples, including $\mathbf{z}_{\text{in}} = f(\mathbf{x}_{\text{in}})$, concentrate. However, when the complexity of $\mathbf{x}_{\text{ood}}$ is lower, the effect of DCAS comes into play: $\mathbf{z}_{\text{ood}}$ is attracted to high-density region on $\mathcal{Z}$, i.e., the origin of the standard Gaussian, causing $\|\mathbf{z}_{\text{ood}}\|$ to become smaller. Consequently, $\|\mathbf{z}_{\text{ood}}\|$ can be as small as $\sqrt{d}$, rendering the TTL unable to differentiate such $\mathbf{x}_{\text{ood}}$ from the typical set, i.e., the In-Dist examples. Therefore, the failure of TTL is caused by the balance between the *original* $\|\mathbf{z}_{\text{ood}}\|$ assuming the effect of DCAS could be removed, and the extent to which the DCAS shrinks $\|\mathbf{z}_{\text{ood}}\|$. Both factors depend on the combination of In-Dist dataset and OOD inputs. The combination of CIRAR-10 for In-Dist and SVHN for OOD is a case where $\|\mathbf{z}_{\text{ood}}\|$ is coincidentally shrunk to approximately $\sqrt{d}$. On the other hand, inputs with even less complexity than SVHN, such as Pool-16/32, are more intensely affected by the DCAS, resulting in $\|\mathbf{z}_{\text{ood}}\| < \sqrt{d}$. In this case, $\mathbf{z}_{\text{ood}}$ falls out of the annulus where the typical set samples reside, and thus the TTL can successfully identify such $\|\mathbf{x}_{\text{ood}}\|$ as OOD.

Untrustworthiness of $\log p(\mathbf{z})$ and $\log p(\mathbf{x})$. We explained the reasons why existing OOD detection methods fail above. Now, we present the main claim of this study. The impact of image complexity on the volume term can be mitigated by the selection of the NF architecture, and indeed, in fixed-volume architectures such as CV-Glow and IDF, the effect described in Remark 2 is nullified (Appendix D.4). However, the effect of DCAS described in Remark 1 affects $\log p(\mathbf{z})$ whenever the latent distribution $P_{\mathbf{z}}$ has a density concentration, regardless of its distributional form. Since there is no way to disable the DCAS, $\log p(\mathbf{z})$ is unreliable. This not only undermines the trustworthiness of the TTL, but also renders $\log p(\mathbf{x})$, which incorporates $\log p(\mathbf{z})$, untrustworthy. Therefore, we conclude that the likelihood of Normalizing Flows is untrustworthy.

5 OOD Detection with Complexity Awareness

In the previous section, we have shown that the OOD detection based on the likelihood of Normalizing Flows is untrustworthy. In this section, we show that this situation can be overcome by exploiting the information on the root cause of the problem, i.e., image complexity. We aim to further validate Hypothesis 1 through the presented demonstration.

5.1 Experiment 4: Normalizing Flows

We aim to isolate the influence of image complexity $C(\mathbf{x})$ on the likelihood. To achieve this, we treat $C(\mathbf{x})$ as an independent variable and train a multivariate detection model that takes two variables, $C(\mathbf{x})$ and $\log p(\mathbf{z})$. While the volume term $\log|\det J_f(\mathbf{x})|$ could also be used as an input variable, we choose not to include it due to its dependence on the NF architecture and higher computational cost for iResNet and ResFlow. Among the various methods available for multivariable detection, we employ one of the simplest models, a

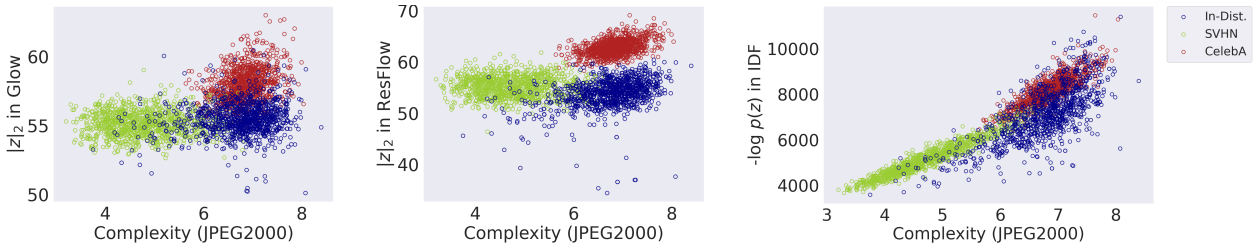

Figure 4: Complexity vs. $\|\mathbf{z}\|$ ($\propto -\sqrt{\log p(\mathbf{z})}$) for OOD datasets. Blue is In-Dist (CIFAR-10), green is SVHN, and Red is CelebA. From left to right, Glow, ResFlow, and IDF. Glow and ResFlow exhibit more pronounced separation between datasets compared to IDF. The GMM is trained to capture the in-distribution in these two-dimensional spaces.

	SVHN	CelebA	TIN	Bed	Living	Tower	N-1	N-2	N-4	N-8	N-16	N-32
TTL	47.86	89.46	84.62	90.77	91.70	89.76	100	57.32	36.88	91.70	96.41	99.97
CALT	94.08	52.78	36.01	54.28	36.12	61.35	97.74	100	100	100	100	100
LRB	54.16	65.69	38.33	34.56	32.81	28.59	0.79	0.26	6.47	15.31	14.87	11.81
LRG	75.08	94.08	81.41	67.92	61.96	68.97	100	73.51	99.21	99.23	97.51	96.37
WAIC	74.78	33.97	78.31	78.43	85.97	83.60	100	100	88.3	90.79	98.08	99.60
Ours (Glow)	91.47	83.15	88.59	86.97	89.25	88.42	100	99.23	100	100	100	100
Ours (ResFlow)	93.51	99.92	93.53	94.04	94.10	91.92	100	99.99	100	99.96	100	100
Ours (iResNet)	87.51	62.27	61.98	81.41	86.04	80.52	100	99.94	10.80	99.47	100	100

	CIFAR-10	CelebA	TIN	Bed	Living	Tower	N-1	N-2	N-4	N-8	N-16	N-32
TTL	97.59	99.98	99.83	99.99	100	99.85	100	100	99.98	78.34	81.69	99.89
CALT	8.12	4.12	8.80	14.14	9.17	21.58	100	100	100	100	100	100
LRB	1.45	5.05	7.01	14.10	9.89	14.67	53.86	0.00	0.11	6.29	16.20	11.68
LRG	94.99	99.96	100	100	100	100	100	100	99.72	99.04	96.71	96.24
WAIC	99.56	99.45	99.86	99.96	99.99	99.74	100	100	100	98.25	95.87	98.04
Ours (Glow)	98.01	99.96	99.91	99.98	100	99.86	100	100	100	100	100	100
Ours (ResFlow)	65.96	63.76	85.76	68.96	84.86	57.04	100	92.12	100	100	100	100
Ours (iResNet)	89.60	98.86	98.99	99.69	99.89	99.03	100	100	99.97	65.61	99.16	99.22

Table 1: AUROC (%)↑. In-Dist datasets are CIFAR-10 (top) and SVHN (bottom). Due to space constraints, Noise-κ is abbreviated as N-κ. Failure cases (lower than 60%) are underlined.

Gaussian mixture model (GMM), for demonstration.

NF models. We use Glow, iResNet, and ResFlow in this demonstration because in these architectures, $\log p(\mathbf{z})$ is insensitive to $C(\mathbf{x})$, which we believe is advantageous for distinguishing between different datasets. For details regarding this architecture selection, refer to Appendices D.5 and D.4. Fig. 4 visually demonstrates the superior capability of Glow and ResFlow to separate and differentiate each dataset, outperforming IDF.

Detection with GMM. Using the samples in the training portion of the In-Dist datasets, $\mathbf{x}_{\text{train}}$, we obtain a set of $C(\mathbf{x}_{\text{train}})$ and $\log p(f(\mathbf{x}_{\text{train}}))$ where f is a trained NF model. Then, we train a GMM $\mathcal{G}$ to capture the distribution of two-dimensional vectors composed of $C(\mathbf{x}_{\text{train}})$ and $\log p(\mathbf{z}_{\text{train}})$. At the testing phase, given a test sample $\mathbf{x}_{\text{test}}$, we input $C(\mathbf{x}_{\text{test}})$ and $\log p(\mathbf{z}_{\text{test}})$ into the trained $\mathcal{G}$, obtaining the likelihood score $S_{\text{GMM}}(\mathbf{x}_{\text{test}})$ from $\mathcal{G}$. A larger $S_{\text{GMM}}(\mathbf{x}_{\text{test}})$ indicates a higher likelihood that $\mathbf{x}_{\text{test}}$ belongs to the In-Dist dataset. For more details, see Appendix D.5.

Datasets. When the In-Dist dataset is CIFAR-10 and SVHN, we use CelebA (Liu et al. 2015), TinyImageNet (TIN) (Russakovsky et al. 2015), LSUN (Yu et al. 2015), and the pooling noise images (Section 3.2) as OOD datasets. For LSUN, which comprises various scene categories, we select the Bedroom (*Bed*), Living room (*Living*), and *Tower* categories, treating each of them as individual OOD datasets. When the In-Dist dataset is MNIST and FMNIST, we use OMNIGLOT (Lake, Salakhutdinov, and Tenenbaum 2015) and NotMNIST (Bulatov 2011) as OOD datasets. In experiments using ImageNet of size 224×224 as the In-Dist dataset, we use the pooling noise images as OOD datasets.

Evaluation metrics. We evaluate our method using two standard metrics in OOD detection literature: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR). These metrics provide an overall assessment of performance by varying the detection threshold. Higher values for both metrics indicate better performance. Given that the chance level for AUROC is 50%, we set a minimum threshold of 60% to evaluate detectability. Our primary focus is on the method's capability

	SVHN	CelebA	TIN	Bed	Living	Tower	N-2	N-4	N-8	N-16	N-32
PixelCNN++	83.38	14.37	23.20	15.53	10.33	24.62	0.03	36.47	98.92	100	100
Ours	95.34	75.71	74.91	76.74	78.77	80.68	100	100	100	100	100
	CIFAR-10	CelebA	TIN	Bed	Living	Tower	N-2	N-4	N-8	N-16	N-32
PixelCNN++	1.86	0.02	0.75	0.18	0.05	0.76	0.00	0.00	66.80	99.30	96.96
Ours	91.48	96.65	96.94	98.42	99.31	97.61	100	100	100	100	100

Table 2: AUROC (%)↑. 'Ours' indicates complexity-aware PixelCNN++. 'PixelCNN++' indicates the likelihood test with PixelCNN++ as a baseline. In-Dist datasets are CIFAR-10 (top) and SVHN (bottom). Due to space constraints, Noise-κ is abbreviated as N-κ. Failure cases (lower than 60%) are underlined. The significant improvements shown here suggest that our hypothesis may be applicable not only to NFs but also to autoregressive models.

to detect any type of OODs. Thus, we evaluate the methods based on whether the AUROC scores exceed 60% for each OOD dataset used in the evaluation. In the tables, scores below 60% are underlined to indicate that they do not meet the detectability threshold.

Competitors. In addition to TTL and CALT, we compare our method to three other existing NF-based methods on CIFAR-10 and SVHN: the Watanabe-Akaike Information Criterion (WAIC) (Choi, Jang, and Alemi 2018), the likelihood-ratio to background model (LRB) (Ren et al. 2019), and the likelihood-ratio to general model (LRG) (Schirrmeister et al. 2020). Appendix D.5 provides detailed descriptions of each method.

Results on MNIST and FMNIST. The AUROC and AUPR are shown in Appendix E.2 (Tables 3 and 4). Our methods detected all cases, with Glow achieving the highest performance, followed by ResFlow and iResNet.

Results on CIFAR-10 and SVHN. The AUROC and AUPR are shown in Tables 1 and 5 in Appendix E.2, respectively. Once again, our method with Glow demonstrated the most superior performance overall. Our method with ResFlow performed best on CIFAR-10 but showed weaker performance on SVHN, scoring below 60% on Tower. Among the other methods, only LRG scored higher than 60% in all test cases, excluding our methods, our speculation on which is included in Appendix E.3.

Results on ImageNet. The AUROC and AUPR are shown in Table 7 in Appendix E.2. Our method achieved detection accuracy of 100% or close to it across all evaluated pooling noise images, validating our claims for large images.

Benefits of using two variables are visually shown in Fig. 4 (left and center). Relying solely on $C(\mathbf{x})$ results in the inability to differentiate between CIFAR-10 (In-Dist.) and CelebA. Similarly, using only $\log p(\mathbf{z})$ fails to differentiate between CIFAR-10 (In-Dist.) and SVHN. Effective separation is achieved by utilizing both variables.

6 Applicability to Autoregressive Model

We now shift our focus from Normalizing Flows (NFs) to Autoregressive (AR) models, which are another type of DGM where instances of the failure of the likelihood test have also been observed (Nalisnick et al. 2019). While our Hypothesis 1 is developed based on a function invertible between latent space and data space, AR models are designed without the concept of latent space. However, a recent interpretation has proposed considering AR models as a single-layer NF, thus implying the presence of an implicit latent space (Nielsen and Winther 2020). In this interpretation, the latent distribution of PixelCNN++ is a discretized mixture of logistics. With this perspective, we argue that our hypothesis is also applicable to AR models and can provide an explanation for the failure of the likelihood tests with AR models. To substantiate this claim, we present the following experimental results.

6.1 Experiment 5: PixelCNN++

First, we present the results of the same experiments using GMM, previously done, but this time applied to PixelCNN++. Similar to CV-Glow and IDF, the volume term in PixelCNN++ is fixed (zero) as $\log p(\mathbf{x}) = \log p(\mathbf{z})$, so the GMM was constructed on $\log p(\mathbf{x})$ and $C(\mathbf{x})$. The AUROC results are provided in Table 2, while the AUPR results are available in Table 6 in Appendix E.3. The results clearly show that our complexity-aware method significantly improves the OOD detection performance for PixelCNN++ as well. These outcomes suggest the existence of the DCAS described in Remark 1 in PixelCNN++.

Second, as demonstrated in Appendix E.3, the response of $\log p(\mathbf{x})$ (or $\log p(\mathbf{z})$) in PixelCNN++ to image complexity closely resembles that observed in NFs with a fixed volume architecture, i.e., CV-Glow and IDF. This finding suggests that the discussion pertaining to Remark 2 (Appendix D.4) is also applicable to PixelCNN++. While further theoretical verification may be necessary to conclusively establish the applicability of Hypothesis 1 to AR models, our results provide plausible evidence supporting this assertion.

7 Conclusion

In this study, we have proposed a hypothesis that explains the failure of OOD detection methods based on the likelihood of Normalizing Flows and Autoregressive models and delineates how the likelihood in those models is affected by varying image complexity. We believe that the findings presented in this paper will contribute to the future advancement of DGM applications, particularly OOD detection.

References

Ahmadian, A.; and Lindsten, F. 2021. Likelihood-free Out-of-Distribution Detection with Invertible Generative Models. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2119–2125. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Behrmann, J.; Grathwohl, W.; Chen, R. T. Q.; Duvenaud, D.; and Jacobsen, J.-H. 2019. Invertible Residual Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 573–582. PMLR.

Behrmann, J.; Vicol, P.; Wang, K.-C.; Grosse, R.; and Jacobsen, J.-H. 2021. 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*, 1792–1800. PMLR.

Bishop, C. M. 1995. Training with noise is equivalent to Tikhonov regularization. *Neural computation*, 7(1): 108–116.

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Bulatov, Y. 2011. Notmnist dataset. *Google (Books/OCR), Tech. Rep.[Online]. Available: http://yaroslavvb. blogspot. it/2011/09/notmnist-dataset. html*, 2.

Cai, M.; and Li, Y. 2023. Out-of-distribution Detection via Frequency-regularized Generative Models. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Chen, R. T. Q.; Behrmann, J.; Duvenaud, D. K.; and Jacobsen, J.-H. 2019. Residual Flows for Invertible Generative Modeling. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chen, Z.; Gu, S.; Lu, G.; and Xu, D. 2022. Exploiting Intra-Slice and Inter-Slice Redundancy for Learning-Based Lossless Volumetric Image Compression. *IEEE Transactions on Image Processing*, 31: 1697–1707.

Choi, H.; Jang, E.; and Alemi, A. A. 2018. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection.

Cover, T. M.; and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. Nice: Nonlinear independent components estimation. In *International Conference on Learning Representations*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using real nvp. In *International Conference on Learning Representations*.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *Proceedings of the International Conference on Learning Representations*.

Federer, H. 2014. *Geometric measure theory*. Springer.

Fernandez-Granda, C. 2017. Optimization-based data analysis: Lecture Notes 3: Randomness.

Grünwald, P. D.; and Vitányi, P. M. B. 2003. Kolmogorov Complexity and Information Theory. With an Interpretation in Terms of Questions and Answers. *Journal of Logic, Language and Information*, 12(4): 497–529.

Ho, J.; Lohn, E.; and Abbeel, P. 2019. Compression with Flows via Local Bits-Back Coding. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hoogeboom, E.; Peters, J.; van den Berg, R.; and Welling, M. 2019. Integer Discrete Flows and Lossless Compression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8710–8719.

Hyvärinen, A.; and Oja, E. 2000. Independent Component Analysis: Algorithms and Applications. *Neural Netw.*, 13(4–5): 411–430.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems 31*, 10215–10224. Curran Associates, Inc.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why Normalizing Flows Fail to Detect Out-of-Distribution Data. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 20578–20589. Curran Associates, Inc.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Linmans, J.; van der Laak, J.; and Litjens, G. 2020. Efficient Out-of-Distribution Detection in Digital Pathology Using Multi-Head Convolutional Neural Networks. In Arbel, T.; Ben Ayed, I.; de Bruijne, M.; Descoteaux, M.; Lombaert, H.; and Pal, C., eds., *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, 465–478. PMLR.

Liu, J.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax Weiss, T.; and Lakshminarayanan, B. 2020. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 7498–7512. Curran Associates, Inc.

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

Morteza, P.; and Li, Y. 2022. Provable Guarantees for Understanding Out-of-distribution Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

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

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; and Lakshminarayanan, B. 2020. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 427–436.

Nielsen, D.; and Winther, O. 2020. Closing the Dequantization Gap: PixelCNN as a Single-Layer Flow. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3724–3734. Curran Associates, Inc.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Perkiö, J.; and Hyvärinen, A. 2009. Modelling Image Complexity by Independent Component Analysis, with Application to Content-Based Image Retrieval. In Alippi, C.; Polycarpou, M.; Panayiotou, C.; and Ellinas, G., eds., *Artificial Neural Networks – ICANN 2009*, 704–714. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood Ratios for Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rezende, D.; and Mohamed, S. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 1530–1538. Lille, France: PMLR.

Rigau, J.; Feixas, M.; and Sbert, M. 2005. An Information-Theoretic Framework for Image Complexity. In *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, Computational Aesthetics'05, 177–184. Goslar, DEU: Eurographics Association. ISBN 3905673274.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Salimans, T.; Karpathy, A.; Chen, X.; and Kingma, D. P. 2017. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. In *International Conference on Learning Representations*.

Sayood, K. 2017. *Introduction to data compression*. Morgan Kaufmann.

Schirrmeister, R.; Zhou, Y.; Ball, T.; and Zhang, D. 2020. Understanding Anomaly Detection with Deep Invertible Networks through Hierarchies of Distributions and Features. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 21038–21049. Curran Associates, Inc.

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

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Theis, L.; van den Oord, A.; and Bethge, M. 2016. A note on the evaluation of generative models. In *International Conference on Learning Representations*.

Wang, Z.; Dai, B.; Wipf, D.; and Zhu, J. 2020. Further Analysis of Outlier Detection with Deep Generative Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 8982–8992. Curran Associates, Inc.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; PENG, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Hendrycks, D.; Li, Y.; and Liu, Z. 2022. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *CoRR*, abs/1506.03365.

Zhang, L.; Goldstein, M.; and Ranganath, R. 2021. Understanding Failures in Out-of-Distribution Detection with Deep Generative Models. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 12427–12436. PMLR.