

## Article

# Perfect Density Models Cannot Guarantee Anomaly Detection

Charline Le Lan <sup>1,2,\*</sup> and Laurent Dinh <sup>2</sup>

<sup>1</sup> Department of Statistics, University of Oxford, Oxford OX1 3LB, UK

<sup>2</sup> Google Research, Montreal, QC H3B 2Y5, Canada; laurentdinh@google.com

\* Correspondence: charline.lelan@stats.ox.ac.uk

**Abstract:** Thanks to the tractability of their likelihood, several deep generative models show promise for seemingly straightforward but important applications like anomaly detection, uncertainty estimation, and active learning. However, the likelihood values empirically attributed to anomalies conflict with the expectations these proposed applications suggest. In this paper, we take a closer look at the behavior of distribution densities through the lens of reparametrization and show that these quantities carry less meaningful information than previously thought, beyond estimation issues or the curse of dimensionality. We conclude that the use of these likelihoods for anomaly detection relies on strong and implicit hypotheses, and highlight the necessity of explicitly formulating these assumptions for reliable anomaly detection.

**Keywords:** deep generative modeling; probabilistic modeling; anomaly detection



**Citation:** Le Lan, C.; Dinh, L. Perfect Density Models Cannot Guarantee Anomaly Detection. *Entropy* **2021**, *23*, 1690. <https://doi.org/10.3390/e23121690>

Academic Editors: Eric Nalisnick and Boris Ryabko

Received: 30 September 2021

Accepted: 13 December 2021

Published: 16 December 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Several machine learning methods aim at extrapolating a behavior observed on training data in order to produce predictions on new observations. However, every so often, such extrapolation can result in wrong outputs, especially on points that we would consider infrequent with respect to the training distribution. Faced with unusual situations, whether adversarial [1,2] or just rare [3], a desirable behavior from a machine learning system would be to flag these *outliers* so that the user can assess if the result is reliable and gather more information if it should be necessary [4,5]. This can be critical for applications like medical decision making [6] or autonomous vehicle navigation [7], where such outliers are ubiquitous.

What are the situations that are deemed unusual? Defining these *anomalies* [8–12] manually can be laborious if not impossible, and so generally applicable, automated methods are preferable. In that regard, the framework of *probabilistic reasoning* has been an appealing formalism because a natural candidate for outliers are situations that are *improbable*. Since the true *probability distribution density*  $p_X^*$  of the data is mostly not provided, one would instead use an estimator  $p_X^{(\theta)}$  from this data to assess the regularity of a point.

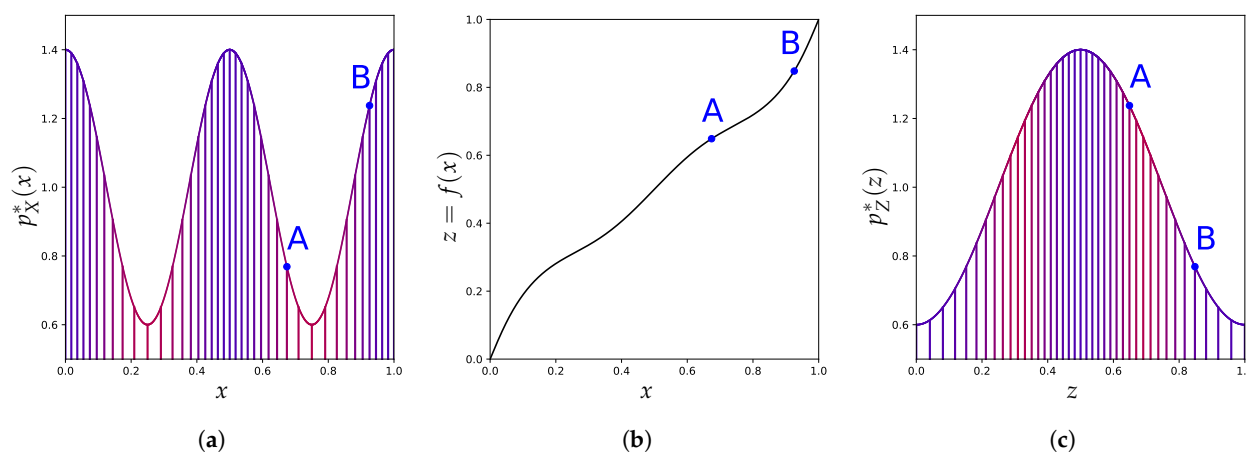
Density estimation has been a particularly challenging task on high-dimensional problems. However, recent advances in *deep probabilistic models*, including variational auto-encoders [13–15], deep autoregressive models [16–18], and flow-based generative models [19–24], have shown promise for density estimation, which has the potential to enable accurate *density-based methods* [25] for anomaly detection.

Yet, several works have observed that a significant gap persists between the potential of density-based anomaly detection and empirical results. For instance [26–28] noticed that generative models trained on a benchmark dataset (e.g., CIFAR-10, [29]) and tested on another (e.g., SVHN, [30]) are not able to identify the latter as an outlier with current methods. Different hypotheses have been formulated to explain that discrepancy, ranging from the *curse of dimensionality* [31] to a significant *mismatch between*  $p_X^{(\theta)}$  and  $p_X^*$  [26,32–36].

In this work, we propose a new perspective on this discrepancy and challenge the expectation that density estimation should always enable anomaly detection. We show that

the aforementioned discrepancy persists even with perfect density models, and therefore goes beyond issues of estimation, approximation, or optimization errors [37]. We highlight that this issue is pervasive as it occurs even in low-dimensional settings and for a variety of density-based methods for anomaly detection. Focusing on the continuous input case, we make the following contributions:

- Similar to classification, we propose in Section 3 a principle of invariance to formalize the underlying assumptions behind the current practice of (deep) density-based methods.
- We use the well-known change of variables formula for probability density to show in Section 4 how these density-based methods are not invariant to reparametrization (see Figure 1) and contradict this principle. We demonstrate the extent of the issues with current practices by building adversarial cases, even under strong distributional constraints.
- Given the resulting tension between the use of these anomaly detection methods and their lack of invariance, we focus in Section 5 on the importance of explicitly incorporating prior knowledge into (density-based) anomaly detection methods as a more promising avenue to reconcile this tension.



**Figure 1.** An invertible change of representation can affect the relative density between two points A and B, which has been interpreted as their relative regularity. (a) An example of a distribution density  $p_X^*$ . (b) Example of an invertible function  $f$  from  $[0, 1]$  to  $[0, 1]$ . (c) Resulting density  $p_Z^*$  as a function of the new axis  $z = f(x)$ . In (a,c) points with high original density  $p_X^*(x)$  are in blue and red for low original density.

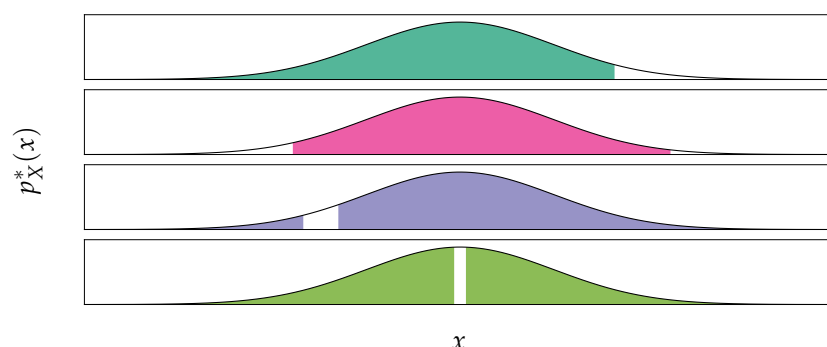
## 2. Density-Based Anomaly Detection

In this section, we present existing density-based anomaly detection approaches that are central to our analysis. Seen as methods without explicit prior knowledge, they aim at unambiguously defining outliers and inliers.

### 2.1. Unsupervised Anomaly Detection: Problem Statement

Unsupervised anomaly detection is a classification problem [38–40], where one aims at distinguishing between regular points (*inliers*) and irregular points (*outliers*). However, as opposed to the usual classification task, labels distinguishing inliers and outliers are not provided for training, if outliers are even provided at all. Given an input space  $\mathcal{X} \subseteq \mathbb{R}^D$ , the task can be summarized as partitioning this space between the subset of outliers  $\mathcal{X}_{out}$  and the subset of inliers  $\mathcal{X}_{in}$ , i.e.,  $\mathcal{X}_{out} \cup \mathcal{X}_{in} = \mathcal{X}$  and  $\mathcal{X}_{out} \cap \mathcal{X}_{in} = \emptyset$ . When the training data is distributed according to the probability measure  $P_X^*$  (with density  $p_X^*$ , that we assume in the rest of the paper to be such that  $\forall x \in \mathcal{X}, p_X^*(x) > 0$ ) one would usually pick the set of regular points  $\mathcal{X}_{in}$  such that this set contains the majority (but not all) of the mass (e.g., 95%) of this distribution [39], i.e.,  $P_X^*(\mathcal{X}_{in}) = 1 - \alpha \in (\frac{1}{2}, 1)$ . However, for any given  $\alpha$ , there exists in theory an infinity of corresponding partitions into  $\mathcal{X}_{in}$  and  $\mathcal{X}_{out}$  (see Figure 2). How are these partitions picked to match our intuition of inliers and outliers? In particular, how can we *learn* from data to discriminate between inliers and outliers (without

of course predefining them)? We will focus in this paper on recently used methods based on probability density.

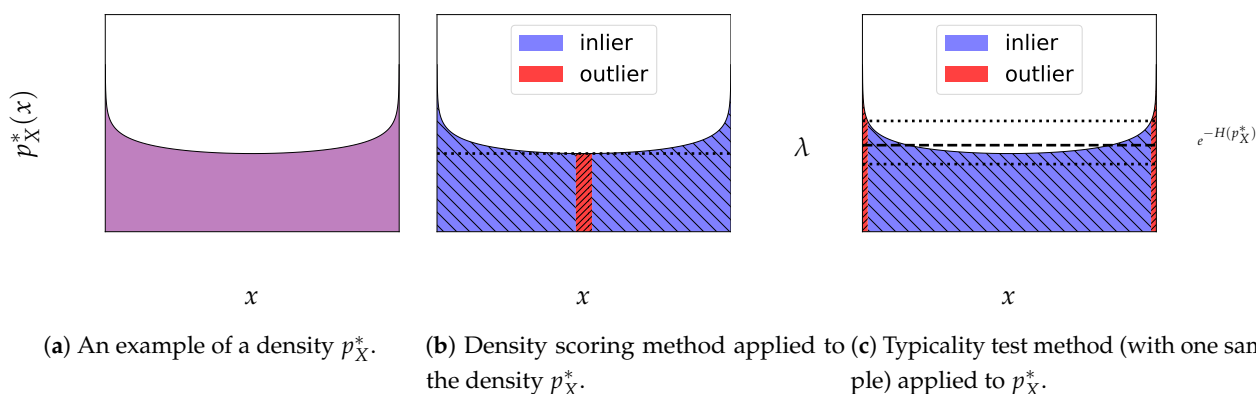


**Figure 2.** There is an infinite number of ways to partition a distribution in two subsets,  $\mathcal{X}_{in}$  and  $\mathcal{X}_{out}$  such that  $P_X^*(\mathcal{X}_{in}) = 0.95$ . Here, we show several choices for a standard Gaussian  $p_X^* = \mathcal{N}(0, 1)$ .

## 2.2. Density Scoring Method

When talking about outliers—infrequent observations—the association with probability can be quite intuitive. For instance, one would expect an anomaly to happen rarely and be unlikely. Since the language of statistics often associate the term *likelihood* with quantities like  $p_X^{(\theta)}(x)$ , one might consider an unlikely sample to have a low “likelihood”, that is, a low probability density  $p_X^*(x)$ . Conversely, regular samples would have a high density  $p_X^*(x)$  following that reasoning. This is an intuition that is not only prevalent in several modern anomaly detection methods [25,28,34,41–43] but also in techniques like low-temperature sampling [44] used for example in Kingma and Dhariwal [21] and parmar et al. [45].

The associated approach, described in Bishop [25], consists in defining the inliers as the points whose density exceed a certain threshold  $\lambda > 0$  (for example, chosen such that inliers include a predefined amount of mass, e.g., 95%), making the modes the most regular points in this setting.  $\mathcal{X}_{out}$  and  $\mathcal{X}_{in}$  are then respectively the lower-level and upper-level sets  $\{x \in \mathcal{X}, p_X^*(x) \leq \lambda\}$  and  $\{x \in \mathcal{X}, p_X^*(x) > \lambda\}$  (see Figure 3b).



**Figure 3.** Illustration of different density-based methods applied to a particular one-dimensional distribution  $p_X^*$ . Outliers are in red and inliers are in blue. The thresholds are picked so that inliers include 95% of the mass. In (b), inliers are considered as the points with density above the threshold  $\lambda > 0$  while in (c), they are the points whose log-density are in the  $\epsilon$ -interval around the negentropy  $-H(p_X^*)$ .

## 2.3. Typicality Test Method

The *Gaussian Annulus theorem* [46] generalized in [47] attests that most of the mass of a high-dimensional standard Gaussian  $\mathcal{N}(0, \mathbb{I}_D)$  is located close to the hypersphere of radius  $\sqrt{D}$ . However, the mode of its density is at the center 0. A natural conclusion is that the

*curse of dimensionality* creates a discrepancy between the density upper-level sets and what we expect as inliers [26,31,48,49]. This motivated Nalisnick et al. [31] to propose another method for testing whether a point is an inlier or not, relying on a measure of its *typicality*. This method relies on the notion of *typical set* [50] defined by taking as inliers points whose average log-density is close to the average log-density of the distribution (see Figure 3c).

**Definition 1** ([50]). Given independent and identically distributed elements  $(x^{(n)})_{n \leq N}$  from a distribution with density  $p_X^*$ , the typical set  $A_\epsilon^{(N)}(p_X^*) \subset \mathcal{X}^N$  is made of all sequences that satisfy

$$\left| H(p_X^*) + \frac{1}{N} \sum_{n=1}^N \log p_X^*(x^{(n)}) \right| \leq \epsilon,$$

where  $H(p_X^*) = -\mathbb{E}[\log p_X^*(X)]$  is the (differential) entropy and  $\epsilon > 0$  a constant.

This method matches the intuition behind the Gaussian Annulus theorem on the set of inliers of a high-dimensional standard Gaussian. Indeed, using a concentration inequality, we can show that  $\lim_{N \rightarrow +\infty} P_{(X_i)_{1 \leq n \leq N}}(A_\epsilon^{(N)}) = 1$ , which means that with  $N$  large enough,  $A_\epsilon^{(N)}(p_X^*)$  will contain most of the mass of  $(p_X^*)^N$ , justifying the name *typicality*.

### 3. The Role of Reparametrization

Density-based anomaly detection is applied in practice [25,28,34,41–43] as follows: first, learn a density estimator  $p_X^{(\theta)}$  to approximate the data density  $p_X^*$ , and then plug that estimate in the density-based methods from Sections 2.2 and 2.3 to discriminate between inliers and outliers. Recent empirical failures [3,26,27] of this procedure applied to density scoring have been attributed to the discrepancy between  $p_X^{(\theta)}$  and  $p_X^*$  [28,33–35,48]. Instead, we choose in this paper to question the fundamental assumption that these density-based methods should result in a correct classification between outliers and inliers.

#### 3.1. A Binary Classification Analogy

We start by studying the desired behavior of a classification method under infinite data and capacity, a setting where the user is provided with a perfect density model  $p_X^{(\theta)} = p_X^*$ .

In Magritte [51], the author reminds us that the input  $x$  we use is merely an arbitrary representation of the studied object (in other words, “a map is not the territory” [52]), standardized here to enable the construction of a large-scale homogeneous dataset to train on [53]. This is after all the definition of a random variable  $x = X(\omega)$ , which is by definition a function from the underlying outcome  $\omega$  to the corresponding observation  $x$ . For instance, in the case of object classification,  $\omega$  is the object while  $X(\omega)$  is the image (produced as a result of lighting, camera position and pose, lenses, and the image sensor). For images, a default representation is the bitmap one. However, this choice of representation remains arbitrary and practitioners have also trained their classifier using pretrained features instead [54–56], JPEG representation [57], encrypted version of the data [58,59], or other resulting transformations  $f(x) = f(X(\omega))$ , without modifying the associated labels. In particular, if  $f$  is invertible,  $f(x) = f(X(\omega))$  contains the same information about  $\omega$  as  $x = X(\omega)$ . Therefore both representations should be classified the same, as we associate the label with the underlying outcome  $\omega$ . If  $c^*$  is the perfect classifier on  $X$ , then  $c^* \circ f^{-1}$  should be the perfect classifier on  $f(X) = (f \circ X)$  to assess the label of  $\omega$ , since  $(c^* \circ f^{-1})(f(x)) = c^*(x)$ .

As an illustration, we can consider the transition from of a cartesian coordinate system  $(x_i)_{i \leq D}$  to a hyperspherical coordinate system, consisting of a radial coordinate  $r > 0$  and  $(D - 1)$  angular coordinates  $(\phi_i)_{i < D}$ ,

$$\forall d < D, x_d = r \left( \prod_{i=1}^{d-1} \sin(\phi_i) \right) \cos(\phi_d)$$

$$x_D = r \left( \prod_{i=1}^{D-2} \sin(\phi_i) \right) \sin(\phi_{D-1}),$$

where for all  $i \in \{1, 2, \dots, D - 2\}$ ,  $\phi_i \in [0, \pi)$  and  $\phi_{D-1} \in [0, 2\pi)$ . While significantly different, those two systems of coordinates (or representations) describe the same vector and are connected by an invertible map  $f$ . In a Cartesian coordinate system, an optimal classifier  $c^*(x) = \mathbb{1}(\sum x_i^2 > 1)$  would become in a hyperspherical representation  $(c^* \circ f^{-1})(r, \phi) = \mathbb{1}(r^2 > 1)$ .

The ability to learn the correct classification rule from infinite data and capacity, regardless of the representation used (or with minimal requirement), is a fundamental requirement (albeit weak) for a machine learning algorithm, and hence an interest in universal approximation properties, see [60–62]. While we do not dismiss the important role of the input representation as an inductive bias (e.g., using pretrained features as inputs), its influence should in principle dissipate entirely in the infinite data and capacity regime and the resulting solution from this ideal setting should be unaffected by this inductive bias. In ideal conditions, solutions to classification should be invariant to any invertible change of representation.

We consider this is in fact one of key tenets behind deep learning [63] and feature engineering/learning in general [64].

### 3.2. A Principle for Anomaly Detection Methods

Current practices of deep anomaly detection commonly include the use of deep density models on either default input feature [26–28,31,34,42] or features learned independently from the anomaly detection task [6,48,65,66]. The process of picking a particular input representation is rarely justified in the context of density-based anomaly detection, which suggests that a similar implicit assumption is being used: *the status of inlier/outlier corresponds to the underlying outcome  $\omega$  behind an input feature  $x = X(\omega)$ , whose only role is to inform us on  $\omega$* . As described in Section 2.1, the goal of anomaly detection is, like classification, to discriminate (although generally in an unsupervised way) between inliers and outliers. Similarly to classification, the label of inlier/outlier of an underlying outcome should remain invariant to reparametrization in an infinite data and capacity setting, especially since information about  $\omega$  (and whether the outcome is anomalous or not) is conserved under an invertible transformation up to numerical instabilities, see [67]. We consider the following principle:

**Principle.** *In an infinite data and capacity setting, the result of an anomaly detection method should be invariant to any continuous invertible reparametrization  $f$ .*

This principle is coherent with the fact that, with  $f$  invertible, the set of outliers  $\mathcal{X}_{out}$  remains a low probability subset as  $P_X(\mathcal{X}_{out}) = P_{f(X)}(f(\mathcal{X}_{out}))$  and  $\forall x \in \mathcal{X}, x \in \mathcal{X}_{out} \iff f(x) \in f(\mathcal{X}_{out})$ . However, density-based methods do not follow this principle as densities are not representation-invariant. In particular, the change of variables formula [68], also used in Dinh et al. [69], Tabak and Turner [19], Rezende and Mohamend [70], formalizes a simple intuition of this behavior: where points are brought closer together the density increases whereas this density decreases when points are spread apart.

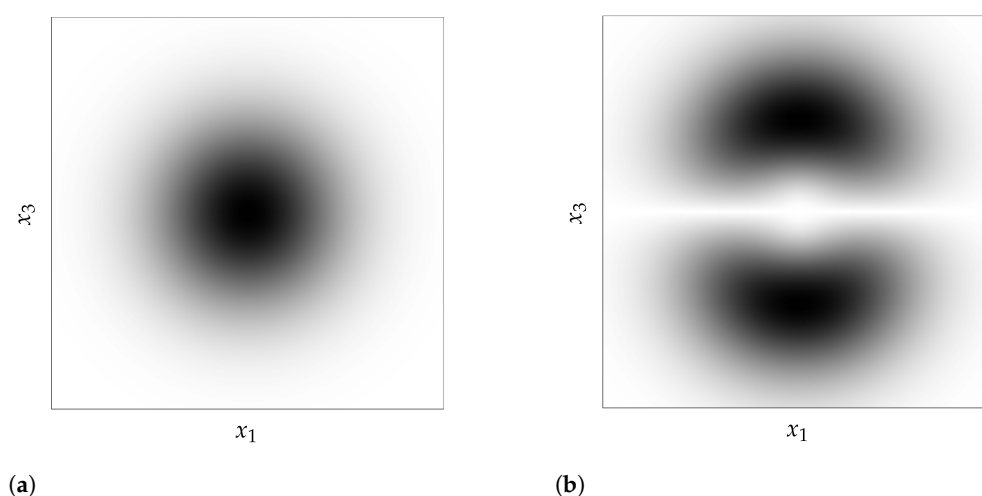
The formula itself is written as:

$$p_{f(X)}^*(f(x)) = p_X^*(x) \left| \frac{\partial f}{\partial x^T}(x) \right|^{-1}$$

where  $\left| \frac{\partial f}{\partial x^T}(x) \right|$  is the Jacobian determinant of  $f$  at  $x$ , a quantity that reflects a local change in volume incurred by  $f$ . Figure 1 already illustrates how the function  $f$  Figure 1b can spread apart points close to the extremities to decrease the corresponding density around  $x = 0$  and  $x = 1$ , and, as a result, turns the density on the left Figure 1a into the density on the right Figure 1c. Figure 4 shows how much a simple change of coordinate system, from Cartesian (Figure 4a) to hyperspherical (Figure 4b), can significantly affect the resulting density associated with a point. This comes from the Jacobian determinant of this change of coordinates:

$$r^{D-1} \left( \prod_{d=1}^{D-1} (\sin(\phi_d))^{D-d-1} \right).$$

With these examples, one can wonder to which degree an invertible change of representation can affect the density and thus the anomaly detection methods presented in Sections 2.2 and 2.3 that use it. This is what we explore in Section 4.



**Figure 4.** Illustration of the change of variables formula for a three-dimensional standard Gaussian distribution with a change of coordinate system, from Cartesian to hyperspherical (where density follows the intuition of the Gaussian Annulus Theorem) (a) A three-dimensional standard Gaussian distribution density in Cartesian coordinates on the hyperplane defined by  $x_2 = 0$ . (b) A three-dimensional standard Gaussian distribution density in hyperspherical coordinates (plotted in Cartesian coordinates) on the hyperplane defined by  $x_2 = 0$ .

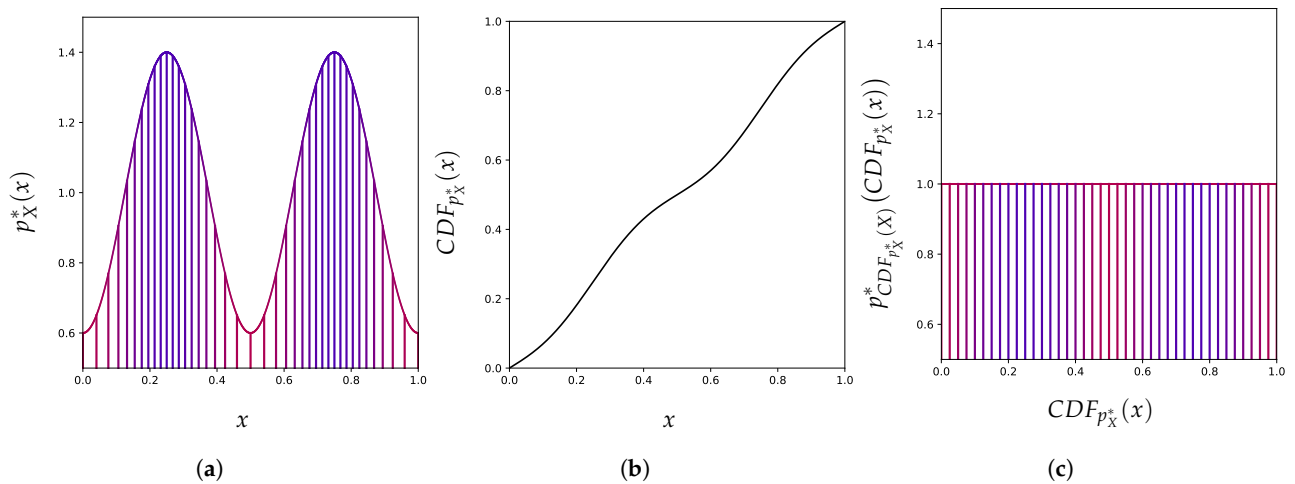
## 4. Leveraging the Change of Variables Formula

### 4.1. Uniformization

We start by showing that unambiguously defining outliers and inliers with any density-based approach becomes impossible when considering a particular type of invertible reparametrization of the problem, irrespective of dimensionality.

Under weak assumptions, one can map any distribution to a uniform distribution using an invertible transformation [71]. This is in fact a common strategy for sampling from complicated one-dimensional distributions [72]. Figure 5 shows an example of this where a bimodal distribution (Figure 5a) is pushed through an invertible map (Figure 5b) to obtain a uniform distribution (Figure 5c).





**Figure 5.** Illustration of the one-dimensional case version of a Knothe-Rosenblatt rearrangement, which is just the application of the cumulative distribution function  $CDF_{p_X^*}$  on the variable  $x$ . Points  $x$  with high density  $p_X^*(x)$  are in blue and points with low density  $p_X^*(x)$  are in red. (a) An example of a distribution density  $p_X^*$ . (b) The corresponding cumulative distribution function  $CDF_{p_X^*}$ . (c) The resulting density from applying  $CDF_{p_X^*}$  to  $X \sim p_X^*$  is  $p_{CDF_{p_X^*}(X)}^* = \mathcal{U}([0, 1])$ .

To construct this invertible uniformization function, we rely on the notion of Knothe-Rosenblatt rearrangement [73,74]. A Knothe-Rosenblatt rearrangement notably used in [71] is defined for a random variable  $X$  distributed according to a strictly positive density  $p_X^*$  with a convex support  $\mathcal{X}$ , as a continuous invertible map  $f^{(KR)}$  from  $\mathcal{X}$  onto  $[0, 1]^D$  such that  $f^{(KR)}(X)$  follows a uniform distribution in this hypercube. This rearrangement is constructed as follows:  $\forall d \in \{1, \dots, D\}, f^{(KR)}(x) = CDF_{p_{X_d}^*|_{X_{<d}}}(x_d | x_{<d})$  where  $CDF_p$  is the cumulative distribution function corresponding to the density  $p$ .

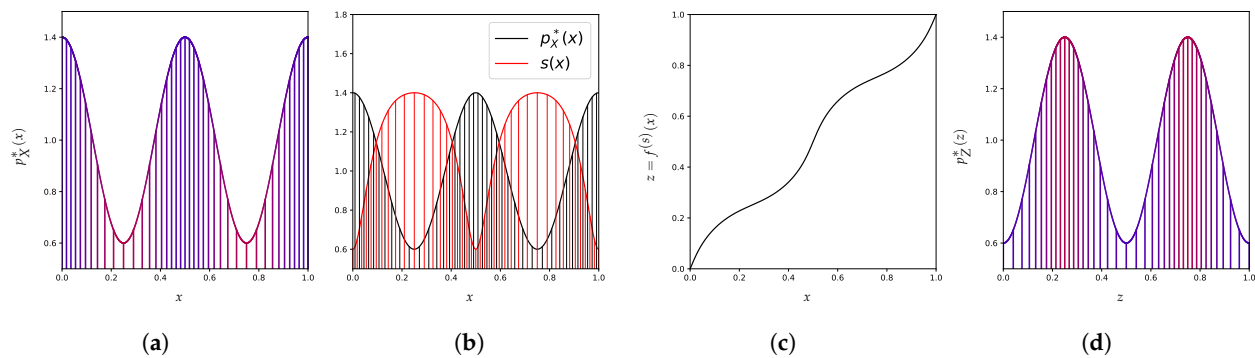
In these new coordinates, neither the density scoring method nor the typicality test approach can discriminate between inliers and outliers in this uniform  $D$ -dimensional hypercube  $[0, 1]^D$ . Since the resulting density  $p_{f^{(KR)}(X)}^* = 1$  is constant, the density scoring method attributes the same regularity to every point or set of points. Moreover, a typicality test on  $f^{(KR)}(X)$  will always succeed as

$$\forall \epsilon > 0, N \in \mathbb{N}^*, \forall (x^{(n)})_{n \leq N}, T_{typ}(p_{f^{(KR)}(X)}^*, (x^{(n)})_{n \leq N}) = 0 \leq \epsilon.$$

However, these uniformly distributed points are merely a different representation of the same initial points. Therefore, if the identity of the outliers is ambiguous in this uniform distribution, then anomaly detection in general should be as difficult.

#### 4.2. Arbitrary Scoring

We find that it is possible to build a reparametrization of the problem to impose to each point an arbitrary density level in the new representation. To illustrate this idea, consider some points from a distribution whose density is depicted in Figure 6a and a score function indicated in red in Figure 6b. In this example, high-density regions correspond to areas with low score value (and vice-versa), such that the ranking from the densities is reversed with this new score. Given that desired score function, we show how to systematically build a reparametrization (depicted in Figure 6c) such that the density in this new representation (Figure 6d) now matches the desired score, which can be designed to mislead density-based methods into a wrong classification of anomalies by modifying a single dimension (in a potentially high-dimensional input vector).



**Figure 6.** Illustration of how we can modify the space with an invertible function so that each point  $x$  follows a predefined score. In (a,b) points with high original density  $p_X^*(x)$  are in blue and red for low original density. (a) An example of a distribution density  $p_X^*$ . (b) The density  $p_X^*$  (in black) and the desired density scoring  $s$  (in red). (c) A continuous invertible reparametrization  $z = f^{(s)}(x)$  such that  $p_Z^*(z) = s(x)$ . (d) Resulting density  $p_Z^*$  from applying  $f^{(s)}$  to  $X \sim p_X^*$  as a function of  $z = f^{(s)}(x)$ .

**Proposition 1.** For any random variable  $X \sim p_X^*$  with  $p_X^*$  strictly positive (with  $\mathcal{X}$  convex) and any measurable continuous function  $s : \mathcal{X} \rightarrow \mathbb{R}_+$  bounded below by a strictly positive number, there exists a continuous bijection  $f^{(s)}$  such that for any  $x \in \mathcal{X}$ ,  $p_{f^{(s)}(X)}^*(f^{(s)}(x)) = s(x)$ .

**Proof.** We write  $x$  to denote  $(x_1, \dots, x_{D-1}, x_D)$  and  $(x_{<D}, t)$  for  $(x_1, \dots, x_{D-1}, t)$ . Let  $f^{(s)} : \mathcal{X} \rightarrow \mathcal{Z} \subset \mathbb{R}^D$  be a function such that

$$(f^{(s)}(x))_D = \int_0^{x_D} \frac{p_X^*((x_{<D}, t))}{s((x_{<D}, t))} dt,$$

and  $\forall d \in \{1, \dots, D-1\}$ ,  $(f^{(s)}(x))_d = x_d$ . As  $s$  is bounded below,  $f^{(s)}$  is well defined and invertible.  $\square$

By the change of variables formula,

$$\forall x \in \mathcal{X}, p_{f^{(s)}(X)}^*(f^{(s)}(x)) = p_X^*(x) \cdot \left| \frac{\partial f^{(s)}}{\partial x^T}(x) \right|^{-1} = p_X^*(x) \cdot \left( \frac{p_X^*(x)}{s(x)} \right)^{-1} = s(x).$$

If  $\mathcal{X}_{in}$  and  $\mathcal{X}_{out}$  are respectively the true sets of inliers and outliers, we can pick a ball  $A \subset \mathcal{X}_{in}$  such that  $P_X^*(A) = \alpha < 0.5$ , we can choose  $s$  such that for any  $x \in (\mathcal{X} \setminus A)$ ,  $s(x) = 1$  and for any  $x \in A$ ,  $s(x) = 0.1$ . With this choice of  $s$  (or a smooth approximation) and the function  $f^{(s)}$  defined earlier, both the density scoring and the (one-sample) typical set methods will consider the set of inliers to be  $(\mathcal{X} \setminus A)$  while  $\mathcal{X}_{out} \subset (\mathcal{X} \setminus A)$ , making their results completely wrong. While we can also reparametrize the problem so that these methods may succeed, e.g., a parametrization where anomalies have low density for the density scoring method, such a reparametrization requires knowledge of  $(p_X^*/s)(x)$ . Without any constraint on the space considered, individual densities can be arbitrarily manipulated, which reveals how little this quantity says about the underlying outcome in general.

#### 4.3. Canonical Distribution

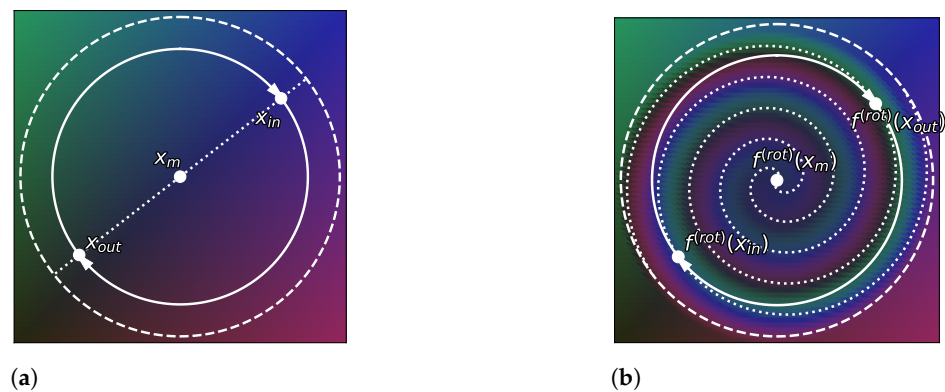
Our analysis from Section 4.2 revealing that densities or low typicality regions are not sufficient conditions for an observation to be an anomaly, whatever its distribution or its dimension, we are now interested in investigating whether additional stronger assumptions can lead to some guarantees for anomaly detection. Motivated by several representation learning algorithms which attempt to learn a mapping to a predefined distribution, e.g., a standard Gaussian, see [13,14,19,65,75], we consider the more restricted setting of a fixed distribution of our choice, whose regular regions could for instance be known. Surprisingly,



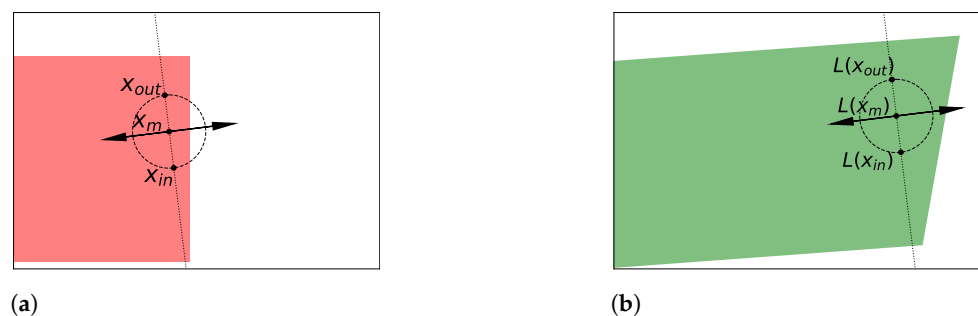
we find that it is possible to exchange the densities of an inlier and an outlier even within a canonical distribution.

**Proposition 2.** For any strictly positive density function  $p_X^*$  over a convex space  $\mathcal{X} \subseteq \mathbb{R}^D$  with  $D \geq 2$ , for any  $x_{in}, x_{out}$  in the interior  $\mathcal{X}^o$  of  $\mathcal{X}$ , there exists a continuous bijection  $f : \mathcal{X} \rightarrow \mathcal{X}$  such that  $p_X^* = p_{f(X)}^*$ ,  $p_{f(X)}^*(f(x^{(in)})) = p_X^*(x^{(out)})$ , and  $p_{f(X)}^*(f(x^{(out)})) = p_X^*(x^{(in)})$ .

**Proof.** The proof is given in Appendix A. It relies on the transformation depicted in Figure 7, which can swap two points while acting in a very local area. If the distribution of points is uniform inside this local area, then this distribution will be unaffected by this transformation. To come to this, we use the uniformization method presented in [71], along with a linear function to fit this local area inside the support of the distribution (see Figure 8). Once those two points have been swapped, we can reverse the functions preceding this swap to recover the original distribution overall.  $\square$



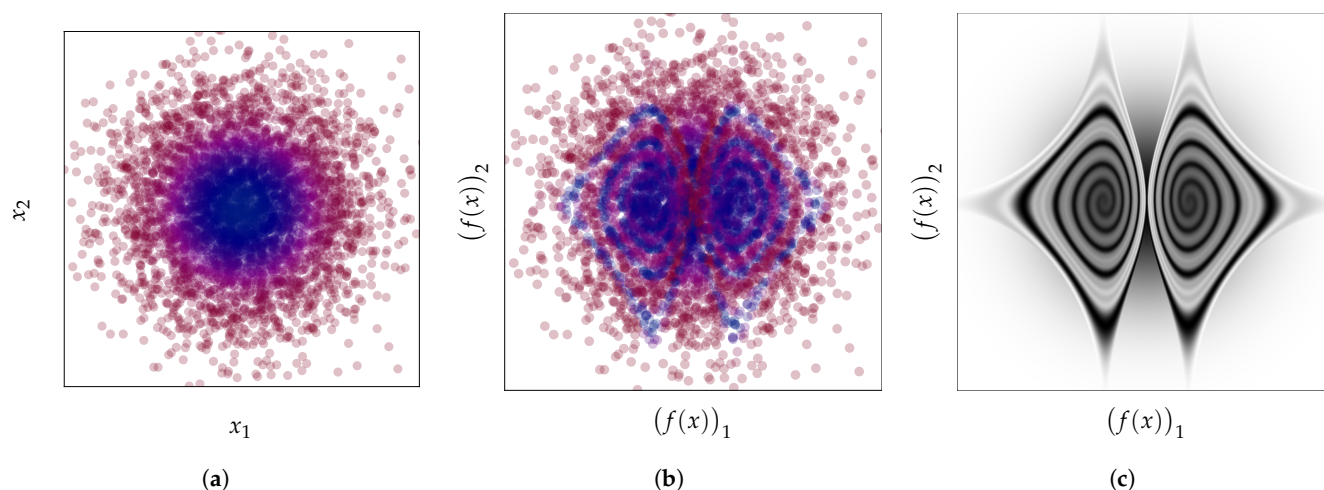
**Figure 7.** Illustration of the norm-dependent rotation, a locally-acting bijection that allows us to swap two different points while preserving a uniform distribution (as a volume-preserving function). (a) Points  $x_{in}$  and  $x_{out}$  in a uniformly distributed subset.  $f^{(rot)}$  will pick a two-dimensional plane and use the polar coordinate using the mean  $x_m$  of  $x_{in}$  and  $x_{out}$  as the center. (b) Applying a bijection  $f^{(rot)}$  exchanging the points  $x_{in}$  and  $x_{out}$ .  $f^{(rot)}$  is a rotation depending on the distance from the mean  $x_m$  of  $x_{in}$  and  $x_{out}$  in the previously selected two-dimensional plane.



**Figure 8.** We illustrate how, given  $x_{in}$  and  $x_{out}$  in a uniformly distributed hypercube  $[0, 1]^D$ , one can modify the space such that  $f^{(rot)}$  shown in Figure 7 can be applied without modifying the distribution. (a) When taking two points  $x_{in}$  and  $x_{out}$  inside the hypercube  $[0, 1]^D$ , there is sometimes no  $L_2$ -ball centered in their mean  $x_m$  containing both  $x_{in}$  and  $x_{out}$ . (b) However, given  $x_{in}$  and  $x_{out}$ , one can apply an invertible linear transformation  $L$  such that there exists a  $L_2$ -ball centered in their new mean  $L(x_m)$  containing both  $L(x_{in})$  and  $L(x_{out})$ . If the distribution was uniform inside  $[0, 1]^D$ , then it is now also uniform inside  $L([0, 1]^D)$ .

Since the resulting distribution  $p_{f(X)}^*$  is identical to the original distribution  $p_X^*$ , their entropies are the same  $H(p_{f(X)}^*) = H(p_X^*)$ . Hence, when  $x_{in}$  and  $x_{out}$  are respectively

an inlier and an outlier, whether in terms of density scoring or typicality, there exists a reparametrization of the problem conserving the overall distribution while still exchanging their status of inlier/outlier. We provide an example applied to a standard Gaussian distribution in Figure 9.



**Figure 9.** Application of the bijection from Figure 7 to a standard Gaussian distribution  $\mathcal{N}(0, \mathbb{I}_2)$  leaving it an overall invariant. (a) Points sampled from  $p_X^* = \mathcal{N}(0, \mathbb{I}_2)$ . (b) Applying a bijection  $f$  that preserves the distribution  $p_{f(X)}^* = \mathcal{N}(0, \mathbb{I}_2)$  to the points in Figure 9a. (c) The original distribution  $p_X^*$  with respect to the new coordinates  $f(x)$ :  $p_X^* \circ f^{-1}$ .

This result is important from a representation learning perspective and a complement to the general non-identifiability result in several representation learning approaches [71,76]. It means that learning a representation with a predefined, well-known distribution and knowing the true density  $p_X^*$  are not sufficient conditions to control the individual density of each point and accurately distinguish outliers from inliers.

#### 4.4. Practical Consequences for Anomaly Detection

We showed that the choice of representation can heavily influence the output of the anomaly detection methods described in Sections 2.2 and 2.3.

##### 4.4.1. Learning a Representation by Applying Explicit Transformations $f$

Surprisingly, this problem can persist even when the learned representation is lower-dimensional, contains *only* the relevant information for the task, and is axis-aligned with semantic variables, since a reasoning similar to Section 4.2 can be applied using axis-aligned bijections to tamper with densities. If a recent review [12] has highlighted the importance of the choice of representation in the context of low-level/high-level anomalies, our result goes further and shows that a problem still persists as even high-level information can be invertibly reparametrized to impose an arbitrary density-based ranking. This leads us to believe that characterizing which representations are suitable for density-based methods (to conform with human expectations) cannot be answered in the absence of prior knowledge (see Section 4.2), e.g., on the distribution of anomalies.

##### 4.4.2. Arbitrary Input Representation Result from Implicit Transformations $f$

While (to our knowledge) input features are rarely designed or heavily tampered with to obfuscate density-based methods in practice, input features can often be the result of a system not fully understood end-to-end, that is of some *implicit transformations*  $f$ , as to how they influence the task of anomaly detection. For instance, cameras used can be tuned to different tasks and the spectral response of film and image sensors has been tuned to maximize performance on the “Shirley Card” [77,78]. Images can also go through

processing techniques like high-dynamic range imaging [79] or arbitrary downsampling as in [29,30,80].

It is well-understood in *representation learning* [64] that the default input features handed to the learning algorithm are rarely well-tuned to the task it tries to solve, e.g., euclidean distance rarely follows a notion of semantic distance, see [81]. Figure 10 provides an example where these methods fail in pixel space despite being endowed with a perfect density model. Details about its construction and analysis are provided below.



**Figure 10.** We generated  $5^6$  pixels according to the procedure described and concatenated them in a single  $125 \times 125$  RGB bitmap image for an easier visualization. While, visual intuition would suggest that white pixels are the outliers in this figure, density-based definitions of anomalies described Section 2.2 (density scoring) and Section 2.3 (typicality) would consider a specific dark shade of gray to be the outlier.

We generate  $5^6$  individual pixels as three-dimensional vectors according to a distribution built as follows: let  $p_w = \mathcal{U}([255, 256]^3)$  (corresponding to the color white),  $p_b = \mathcal{U}([0, 10]^3)$  (corresponding to shades of black), and  $p_{out} = \mathcal{U}([10, 11]^3)$  (corresponding to a dark shade of grey) be distributions with disjoint supports. We consider pixels following the distribution

$$p_X(x) = \beta \cdot p_{out}(x) + (1 - \beta)(\alpha \cdot p_w(x) + (1 - \alpha) \cdot p_b(x)),$$

where  $\alpha = 1001^{-3}$  and  $\beta = 10^{-4}$ . Once generated, we concatenate these pixels in a  $125 \times 125$  RGB bitmap image in Figure 10 for a more convenient visualization.

Visually, a common intuition would be to consider white pixels to be the anomalies in this figure. However, following a construction similar to Section 4.2, the final densities corresponding to pixels from  $p_w$  (equal to  $\alpha(1 - \beta)$ ) and  $p_b$  (equal to  $(1 - \alpha)(1 - \beta)10^{-3}$ ) are equal to  $1001^{-3}(1 - 10^{-4}) \approx 10^{-3}$ , and the final density corresponding to pixels from  $p_{out}$  (equal to  $\beta$ ) is  $10^{-4}$ . Therefore, none of the methods presented in Section 2.2 (density scoring) and Section 2.3 (one-sample typicality) would consider the white pixels

(in  $[255, 256]^3$ ) as outliers. They would only classify the pixels of a particular dark shade of gray in  $[10, 11]^3$  as outliers.

Given the considerable influence, the choice of input representation has on the output of even the true data density function  $p_X^*$ , one should question the strong but understated assumption behind current practices that (density-based) anomaly detection methods applied on *default input representations* decontextualized from their design process [82], *representations orthogonally learned from the task*, or even obtained by filtering noise variables (non-semantic) ought to result in proper outlier classification.

## 5. Promising Avenues for Unsupervised Density-Based Anomaly Detection

While anomaly detection can be an ill-posed problem as mentioned in [26,31,48] without *prior knowledge*, several approaches are more promising by making this prior knowledge more explicit. We highlighted the strong dependence of density-based anomaly detection methods on a choice of representation, which needs to be justified as it is crucial to the success of the approach. This was proven by using the change of variables formula, which describes how the density function varies with respect to a reparametrization. If we consider the fundamental definition of a density as a Radon-Nikodym derivative  $p_X^* = \frac{dP_X^*}{d\mu_X}$  with respect to a base measure (here the Lebesgue measure  $\mu_X$  in  $\mathcal{X}$ ), we notice that this variation stems from a change of “denominator”: the Lebesgue measure corresponding to  $\mathcal{X}$  is different to the one corresponding to another space  $\mathcal{Z}$  (the Jacobian determinant accounting for this mismatch  $\mu_X \circ X \neq \mu_Z \circ Z$ ).

A way to incorporate more transparently the choice of representation is to consider a similar fraction. For example, *density ratio methods* [83] score points using a ratio  $p_X^*/p_{BG}$  between two densities. The task is then to figure out whether a point comes from a regular source (the foreground distribution in the numerator) or an anomalous source (the background distribution in the denominator). The concurrent work [84] also draws a similar conclusion showing that no test can distinguish between a given source distribution and an unspecified outlier distribution better than random chance. In Bishop [25], the density scoring method has been interpreted as a density ratio method with a default uniform density function. More refined methods can be used as a background distribution, e.g.,  $p_X^*$  convolved with a noise distribution [85], the implicit distribution of a compressor [86], or a mixture including  $p_X^*$  as a component, i.e., a “superset”, see [87]. In addition to being more transparent with respect to its underlying assumptions, density ratio methods are invariant to invertible reparametrization.

While appealing in their property, density ratio methods still require the explicit definition of a background distribution, an explicit guess on how the anomalies should be distributed. It is actually possible in some cases to be more intentional in the definition of this denominator. For example, for exploration in reinforcement learning, Houthoofd et al. [88] and Bellemare et al. [89] use an (invertible) reparametrization-invariant proxy for potential information gain.

## 6. Discussion and Limitations

We discussed the ill-defined (and arguably subjective) notion of *outlier* or *anomaly*, which several works attempted to characterize through a seemingly clearer notion of probability density used in the density scoring and typicality test methods. We show in this paper that an undesirable degree of freedom persists in how density functions can be manipulated by an arbitrary choice of representation, rarely set to fit the task. We consider that the lack of attention paid to this crucial element has undermined the foundations of these off-the-shelf methods, potentially providing a simpler explanation to their empirical failures studied in [26–28,32–34] as a discrepancy with unstated prior assumptions.

We conclude that being more intentional about integrating prior knowledge explicitly in density-based anomaly detection algorithms then becomes essential to their success.

Although a similar issue persists in practice for *discrete spaces* as noted in [49], where outputs with highest probability are atypical, the same reparametrization trick used

throughout this paper to formalize this issue for continuous inputs is not directly applicable for discrete input spaces. However, similar adversarial constructions can be made in an analogous way: semantically close inputs can be considered distinct or identical depending on arbitrary choices of discretization/categorization [90], resulting in different probability values. Arbitrary choices of discretization include tokenization, lemmatization, or encoding see [91] for language modeling but also choice of language [92]. Figure 10 provides a similar construction in discrete pixel space.

Similarly, while approaches involving *probability masses* are unaffected by invertible reparametrizations, they explicitly rely on a deliberate choice in partitioning the input space, which is why we consider such approaches coherent with a more explicit incorporation of prior knowledge.

We make in the paper the assumption that the data distribution density  $p_X^*$  is *strictly positive* everywhere in the set of possible instances  $\mathcal{X}$  since in practice deep density models spread probability over all the input space. Arguably, an instance occurring outside the support of the data distribution would be considered an anomaly. An example would be CIFAR-10 and SVHN, which can be assumed to be disjoint. However, considering even the slightest Gaussian noise on either data distribution is sufficient to have *non-disjoint supports* as it makes the densities non-zero everywhere in the pixel space. Since Section 2.3 highlighted a failure of our geometrical intuition of density through the Gaussian Annulus theorem, we advocate for some skepticism on the assumption that these data distributions ought to be completely disjoint. In the general case, it is unknown whether anomalies lie outside of the distribution support and not uncommon to consider the probability of an anomaly happening to be non-zero with respect to the data distribution (i.e.,  $P_X^*(\mathcal{X}_{out}) > 0$ ), which is coherent with this strict positivity assumption. On the contrary, the concurrent work [84] chooses to assume a disjoint support for the inlier and outlier distributions, leading them to conclude that the model misestimation is the source of the observations made by Nalisnick et al. [27].

## 7. Broader Impact

Anomaly detection is commonly proposed as a fundamental element to safely deploy machine learning models in the real world. Its applications range from medical diagnostics and autonomous driving to cyber security and financial fraud detection. The use of such models on outlier points can result in dangerous behaviors but also discriminatory outcomes. Our paper aims at questioning current density-based anomaly detection methods, which is essential to mitigate the risks associated with their use in the real-world.

More broadly, our study also leads to reconsider the role of density as a standalone quantity and practices built around it, e.g., temperature sampling [21,44,45] and evaluating density models on anomaly detection, e.g., as in [34,93–95].

Finally, a common opinion in machine learning [96] has been that, given enough data and capacity, machine learning bias generally has a vanishing influence over the resulting bias in the learned solution. On the contrary, scale can obfuscate [82] misspecifications in the task and/or data collection design [97,98]. Here, we focused on how misspecifications in the algorithm design for anomaly detection can result in gross failure even in the ideal theoretical settings of infinite data and capacity.

However, this study provides a constructive proof in Section 4.2 that bad actors can use to arbitrarily manipulate the results of currently used anomaly detection algorithms, without modifying a learned model  $p_X^{(\theta)}$ . This opens the door to potential negative impacts if unreasonable trust in these methods are maintained in practice.

**Author Contributions:** Conceptualization, L.D.; methodology, L.D.; formal analysis, C.L.L. and L.D.; investigation, C.L.L. and L.D.; writing—original draft preparation, C.L.L. and L.D.; writing—review and editing, C.L.L. and L.D.; visualization, L.D.; supervision, L.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.



**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank Kyle Kastner, Johanna Hansen, Ben Poole, Arthur Gretton, Durk Kingma, Samy Bengio, Jascha Sohl-Dickstein, Adam Foster, Polina Kirichenko, Pavel Izmailov, Ross Goroshin, Hugo Larochelle, Jörn-Henrik Jacobsen, and Kyunghyun Cho for initial discussions for this paper. We also thank Eric Jang, Alex Alemi, Jishnu Mukhoti, Jannik Kossen, Sebastian Schmon, Francisco Ruiz, David Duvenaud, Luke Vilnis, and the anonymous reviewers for useful feedback on this paper. We would also like to thank the Python community [99,100] for developing tools that enabled this work, including NumPy [101–103], SciPy [104], and Matplotlib [105].

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proof of Proposition 2

**Proposition A1.** For any strictly positive density function  $p_X^*$  over a convex space  $\mathcal{X} \subseteq \mathbb{R}^D$  with  $D \geq 2$ , for any  $x_{in}, x_{out}$  in the interior  $\mathcal{X}^o$  of  $\mathcal{X}$ , there exists a continuous bijection  $f : \mathcal{X} \rightarrow \mathcal{X}$  such that  $p_X^* = p_{f(X)}^*, p_{f(X)}^*(f(x^{(in)})) = p_X^*(x^{(out)})$ , and  $p_{f(X)}^*(f(x^{(out)})) = p_X^*(x^{(in)})$ .

**Proof.** Our proof will rely on the following non-rigid rotation  $f^{(rot)}$ . Working in a hyper-spherical coordinate system consisting of a radial coordinate  $r > 0$  and  $(D - 1)$  angular coordinates  $(\phi_i)_{i < D}$ ,

$$\begin{aligned} \forall d < D, x_d &= r \left( \prod_{i=1}^{d-1} \sin(\phi_i) \right) \cos(\phi_d) \\ x_D &= r \left( \prod_{i=1}^{D-2} \sin(\phi_i) \right) \sin(\phi_{D-1}), \end{aligned}$$

where for all  $i \in \{1, 2, \dots, D - 2\}$ ,  $\phi_i \in [0, \pi)$  and  $\phi_{D-1} \in [0, 2\pi)$ , given  $r_{max} > r_0 > 0$ , we define the continuous mapping  $f^{(rot)}$  as:

$$\begin{aligned} f^{(rot)}((r, \phi_1, \dots, \phi_{D-2}, \phi_{D-1})) \\ = (r, \phi_1, \dots, \phi_{D-2}, \\ \phi_{D-1} + \pi \frac{(r_{max} - r)_+}{r_{max} - r_0} [\text{mod } 2\pi]). \end{aligned}$$

where  $(\cdot)_+ = \max(\cdot, 0)$ . This mapping only affects points inside  $\mathcal{B}_2(0, r_{max})$ , and exchanges two points corresponding to  $(r_0, \phi_1, \dots, \phi_{D-2}, \phi_{D-1})$  and  $(r_0, \phi_1, \dots, \phi_{D-2}, \phi_{D-1} + \pi)$  in a continuous way (see Figure 7). Since the Jacobian determinant of the hyperspherical coordinates transformation is not a function of  $\phi_{D-1}$ ,  $f^{(rot)}$  is volume-preserving in cartesian coordinates.

Let  $f^{(KR)}$  be a Knothe-Rosenblatt rearrangement of  $p_X^*$ ,  $f^{(KR)}(X)$  is uniformly distributed in  $[0, 1]^D$ . Let  $z^{(in)} = f^{(KR)}(x^{(in)})$  and  $z^{(out)} = f^{(KR)}(x^{(out)})$ . Since  $f^{(KR)}$  is continuous,  $z^{(in)}, z^{(out)}$  are in the interior  $(0, 1)^D$ . Therefore, there is an  $\epsilon > 0$  such that the  $L_2$ -balls  $\mathcal{B}_2(z^{(in)}, \epsilon)$  and  $\mathcal{B}_2(z^{(out)}, \epsilon)$  are inside  $(0, 1)^D$ . Since  $(0, 1)^D$  is convex, so is their convex hull.

Let  $r_0 = \frac{1}{2} \|z^{(in)} - z^{(out)}\|_2$  and  $r_{max} = r_0 + \epsilon$ . Given  $z \in (0, 1)^D$ , we write  $z_{\parallel}$  and  $z_{\perp}$  to denote its parallel and orthogonal components with respect to  $(z^{(in)} - z^{(out)})$ . We consider the linear bijection  $L$  defined by

$$L(z) = z_{\parallel} + \epsilon^{-1} r_{max} z_{\perp}.$$



Let  $f^{(z)} = L \circ f^{(KR)}$ . Since  $L$  is a linear function (i.e., with constant Jacobian),  $f^{(z)}(X)$  is uniformly distributed inside  $L([0, 1]^D)$ . If  $z^{(m)}$  is the mean of  $z^{(in)}$  and  $z^{(out)}$ , then  $f^{(z)}(X)$  contains  $\mathcal{B}_2(L(z^{(m)}), r_{max})$  (see Figure 8). We can then apply the non-rigid rotation  $f^{(rot)}$  defined earlier, centered on  $L(z^{(m)})$  to exchange  $L(z^{(in)})$  and  $L(z^{(out)})$  while maintaining this uniform distribution.

We can then apply the bijection  $(f^{(z)})^{-1}$  to obtain the invertible map  $f = (f^{(z)})^{-1} \circ f^{(rot)} \circ f^{(z)}$  such that  $p_{f(X)}^* = f_X^*$ ,  $p_{f(X)}^*(f(x^{(in)})) = p_X^*(x^{(out)})$ , and  $p_{f(X)}^*(f(x^{(out)})) = p_X^*(x^{(in)})$ .  $\square$

## References

1. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
2. Carlini, N.; Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 3–14.
3. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
4. Zhao, R.; Tresp, V. Curiosity-driven experience prioritization via density estimation. *arXiv* **2019**, arXiv:1902.08039.
5. Fu, J.; Co-Reyes, J.; Levine, S. Ex2: Exploration with exemplar models for deep reinforcement learning. *arXiv* **2017** arXiv:1703.01260.
6. Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7167–7177.
7. Filos, A.; Tigkas, P.; Mcallister, R.; Rhinehart, N.; Levine, S.; Gal, Y. Can Autonomous Vehicles Identify, Recover From, and Adapt to Distribution Shifts? In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; Volume 119, pp. 3145–3153.
8. Grubbs, F.E. Procedures for detecting outlying observations in samples. *Technometrics* **1969**, *11*, 1–21. [\[CrossRef\]](#)
9. Barnett, V.; Lewis, T. Outliers in statistical data. In *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*; John Wiley & Sons: Chichester, UK, 1984.
10. Hodge, V.; Austin, J. A survey of outlier detection methodologies. *Artif. Intell. Rev.* **2004**, *22*, 85–126. [\[CrossRef\]](#)
11. Pimentel, M.A.; Clifton, D.A.; Clifton, L.; Tarassenko, L. A review of novelty detection. *Signal Process.* **2014**, *99*, 215–249. [\[CrossRef\]](#)
12. Ruff, L.; Kauffmann, J.R.; Vandermeulen, R.A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T.G.; Müller, K.R. A unifying review of deep and shallow anomaly detection. *Proc. IEEE* **2021**. [\[CrossRef\]](#)
13. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
14. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of the International Conference on Machine Learning 2014, Beijing, China, 21–26 June 2014.
15. Vahdat, A.; Kautz, J. NVAE: A Deep Hierarchical Variational Autoencoder. *arXiv* **2020**, arXiv:2007.03898.
16. Uria, B.; Murray, I.; Larochelle, H. A deep and tractable density estimator. In Proceedings of the International Conference on Machine Learning 2014, Beijing, China, 21–26 June 2014; pp. 467–475.
17. van den Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In Proceedings of the International Conference on Machine Learning 2014, Beijing, China, 21–26 June 2014; pp. 1747–1756.
18. van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A. Conditional image generation with pixelcnn decoders. *arXiv* **2016**, arXiv:1606.05328.
19. Dinh, L.; Krueger, D.; Bengio, Y. Nice: Non-linear independent components estimation. *arXiv* **2014**, arXiv:1410.8516.
20. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using real nvp. *arXiv* **2016**, arXiv:1605.08803.
21. Kingma, D.P.; Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *arXiv* **2018**, arXiv:1807.03039.
22. Ho, J.; Chen, X.; Srinivas, A.; Duan, Y.; Abbeel, P. Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design. In Proceedings of the International Conference on Machine Learning, Beach, CA, USA, 9–15 June 2019; pp. 2722–2730.
23. Kobayev, I.; Prince, S.; Brubaker, M. Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [\[CrossRef\]](#)
24. Papamakarios, G.; Nalisnick, E.; Rezende, D.J.; Mohamed, S.; Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference. *J. Mach. Learn. Res.* **2021**, *22*, 1–64.
25. Bishop, C.M. Novelty detection and neural network validation. *IEE Proc. Vision Image Signal Process.* **1994**, *141*, 217–222. [\[CrossRef\]](#)

26. Choi, H.; Jang, E.; Alemi, A.A. Waic, but why? generative ensembles for robust anomaly detection. *arXiv* **2018**, arXiv:1810.01392.
27. Nalisnick, E.; Matsukawa, A.; Teh, Y.W.; Gorur, D.; Lakshminarayanan, B. Do Deep Generative Models Know What They Don't Know? In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
28. Hendrycks, D.; Mazeika, M.; Dietterich, T. Deep Anomaly Detection with Outlier Exposure. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
29. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Available online: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf> (accessed on 8 April 2009).
30. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. *Reading Digits in Natural Images with Unsupervised Feature Learning*; NIPS Workshop: Grenada, Spain, 2011.
31. Nalisnick, E.; Matsukawa, A.; Teh, Y.W.; Lakshminarayanan, B. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. *arXiv* **2019**, arXiv:1906.02994.
32. Just, J.; Ghosal, S. Deep Generative Models Strike Back! Improving Understanding and Evaluation in Light of Unmet Expectations for OoD Data. *arXiv* **2019**, arXiv:1911.04699.
33. Fetaya, E.; Jacobsen, J.H.; Grathwohl, W.; Zemel, R. Understanding the Limitations of Conditional Generative Models. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
34. Kirichenko, P.; Izmailov, P.; Wilson, A.G. Why Normalizing Flows Fail to Detect Out-of-Distribution Data. In *Advances in Neural Information Processing Systems*; Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.F.; Lin, H., Eds.; Curran Associates, Inc.: Nice, France, 2020; Volume 33, pp. 20578–20589.
35. Zhang, H.; Li, A.; Guo, J.; Guo, Y. Hybrid Models for Open Set Recognition. In Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020; Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J., Eds.; Springer: New York, NY, USA, 2020; Volume 12348, pp. 102–117. [[CrossRef](#)]
36. Wang, Z.; Dai, B.; Wipf, D.; Zhu, J. Further Analysis of Outlier Detection with Deep Generative Models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*. Available online: <http://proceedings.mlr.press/v137/wang20a.html> (accessed on 12 December 2021).
37. Bottou, L.; Bousquet, O. The tradeoffs of large scale learning. *Adv. Neural Inf. Process. Syst.* **2008**, *351*, 161–168.
38. Moya, M.M.; Koch, M.W.; Hostetler, L.D. One-class classifier networks for target recognition applications. *STIN* **1993**, *93*, 24043.
39. Schölkopf, B.; Platt, J.C.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural Comput.* **2001**, *13*, 1443–1471. [[CrossRef](#)]
40. Steinwart, I.; Hush, D.; Scovel, C. A Classification Framework for Anomaly Detection. *J. Mach. Learn. Res.* **2005**, *6*, 211–232.
41. Blei, D.; Heller, K.; Salimans, T.; Welling, M.; Ghahramani, Z. Presented at Panel: On the Foundations and Future of Approximate Inference. In Proceedings of the Advances in Approximate Bayesian Inference, Long Beach, CA, USA, 4–9 December 2017.
42. Rudolph, M.; Wandt, B.; Rosenhahn, B. Same Same but DifferNet: Semi-Supervised Defect Detection With Normalizing Flows. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 1907–1916.
43. Liu, W.; Wang, X.; Owens, J.; Li, Y. Energy-based Out-of-distribution Detection. *arXiv* **2020**, arXiv:2010.03759.
44. Graves, A. Generating sequences with recurrent neural networks. *arXiv* **2013**, arXiv:1308.0850.
45. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image Transformer. In Proceedings of the Machine Learning Research 2018, Stockholm, Sweden, 6–9 July 2018; pp. 4055–4064.
46. Blum, A.; Hopcroft, J.; Kannan, R. Foundations of data science. *Vorabversion Eines Lehrbuchs* **2016**, *5*, 21–23.
47. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science (Cambridge Series in Statistical and Probabilistic Mathematics)*; Cambridge University Press: Cambridge, UK, 2018; doi:10.1017/9781108231596. [[CrossRef](#)]
48. Morningstar, W.; Ham, C.; Gallagher, A.; Lakshminarayanan, B.; Alemi, A.; Dillon, J. Density of States Estimation for Out of Distribution Detection. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, online, 13–15 April 2021; Volume 130, pp. 3232–3240.
49. Dieleman, S. Musings on Typicality. 2020. Available online: <https://benanne.github.io/2020/09/01/typicality.html> (accessed on 12 December 2021).
50. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
51. Magritte, R. La trahison des images. *Oil Canvas Paint.* **1929**, *63*, 93.
52. Korzybski, A. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*; Institute of GS: New York, NY, USA, 1958.
53. Hanna, A.; Park, T.M. Against Scale: Provocations and Resistances to Scale Thinking. *arXiv* **2020**, arXiv:2010.08850.
54. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. In Proceedings of the 2nd International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
55. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
56. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
57. Gueguen, L.; Sergeev, A.; Kadlec, B.; Liu, R.; Yosinski, J. Faster neural networks straight from jpeg. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 3933–3944.

58. Xie, P.; Bilenko, M.; Finley, T.; Gilad-Bachrach, R.; Lauter, K.; Naehrig, M. Crypto-nets: Neural networks over encrypted data. *arXiv* **2014**, arXiv:1412.6181.
59. Gilad-Bachrach, R.; Dowlin, N.; Laine, K.; Lauter, K.; Naehrig, M.; Wernsing, J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 201–210.
60. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **1989**, *2*, 303–314. [\[CrossRef\]](#)
61. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **1991**, *4*, 251–257. [\[CrossRef\]](#)
62. Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numer.* **1999**, *8*, 143–195. [\[CrossRef\]](#)
63. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press Cambridge: Cambridge, MA, USA, 2016.
64. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [\[CrossRef\]](#)
65. Krusinga, R.; Shah, S.; Zwicker, M.; Goldstein, T.; Jacobs, D. Understanding the (un) interpretability of natural image distributions using generative models. *arXiv* **2019**, arXiv:1901.01499.
66. Winkens, J.; Bunel, R.; Roy, A.G.; Stanforth, R.; Natarajan, V.; Ledsam, J.R.; MacWilliams, P.; Kohli, P.; Karthikesalingam, A.; Kohl, S. Contrastive Training for Improved Out-of-Distribution Detection. *arXiv* **2020**, arXiv:2007.05566.
67. Behrmann, J.; Vicol, P.; Wang, K.C.; Grosse, R.; Jacobsen, J.H. Understanding and Mitigating Exploding Inverses in Invertible Neural Networks. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, online, 13–15 April 2021, Volume 130, pp. 1792–1800.
68. Kaplan, W. *Advanced Calculus*; Pearson Education India: London, UK, 1952.
69. Tabak, E.G.; Turner, C.V. A family of nonparametric density estimation algorithms. *Commun. Pure Appl. Math.* **2013**, *66*, 145–164. [\[CrossRef\]](#)
70. Rezende, D.; Mohamed, S. Variational Inference with Normalizing Flows. In Proceedings of the Machine Learning Research, Cambridge, MA, USA, 16–18 November 2015.
71. Hyvärinen, A.; Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.* **1999**, *12*, 429–439. [\[CrossRef\]](#)
72. Devroye, L. Sample-based non-uniform random variate generation. In Proceedings of the 18th Conference on Winter Simulation, Washington, DC, USA, 8–10 December 1986; pp. 260–265.
73. Rosenblatt, M. Remarks on a multivariate transformation. *Ann. Math. Stat.* **1952**, *23*, 470–472. [\[CrossRef\]](#)
74. Knothe, H. Contributions to the theory of convex bodies. *Mich. Math. J.* **1957**, *4*, 39–52. [\[CrossRef\]](#)
75. Chen, S.S.; Gopinath, R.A. Gaussianization. In *Advances in Neural Information Processing Systems 13*; Leen, T.K., Dietterich, T.G., Tresp, V., Eds.; MIT Press: Cambridge, MA, USA, 2001; pp. 423–429.
76. Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In Proceedings of the International Conference on Machine Learning, Beach, CA, USA, 9–15 June 2019; pp. 4114–4124.
77. Roth, L. Looking at Shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Can. J. Commun.* **2009**, *34*. Available online: <https://pdfs.semanticscholar.org/e5e1/3351c49ae30baffe7339d085ed870b022e75.pdf> (accessed on 12 December 2021). [\[CrossRef\]](#)
78. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
79. Reinhard, E.; Heidrich, W.; Debevec, P.; Pattanaik, S.; Ward, G.; Myszkowski, K. In *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*; Morgan Kaufmann: Burlington, MA, USA, 2010.
80. Torralba, A.; Fergus, R.; Freeman, W.T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1958–1970. [\[CrossRef\]](#) [\[PubMed\]](#)
81. Theis, L.; van den Oord, A.; Bethge, M. A note on the evaluation of generative models. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
82. Raji, D.I.; Denton, E.; Hanna, A.; Bender, E.M.; Paullada, A. AI and the Everything in the Whole Wide World Benchmark. In *NeurIPS 2020 Workshop: ML-Retrospectives; Surveys & Meta-Analyses*, Vancouver, BC, Canada, 2020.
83. Griffiths, T.L.; Tenenbaum, J.B. From mere coincidences to meaningful discoveries. *Cognition* **2007**, *103*, 180–226. [\[CrossRef\]](#)
84. Zhang, L.; Goldstein, M.; Ranganath, R. Understanding Failures in Out-of-Distribution Detection with Deep Generative Models. In Proceedings of the International Conference on Machine Learning, Shenzhen, China, 18–21 July 2021; pp. 12427–12436.
85. Ren, J.; Liu, P.J.; Fertig, E.; Snoek, J.; Poplin, R.; Deprieto, M.; Dillon, J.; Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. *arXiv* **2019**, arXiv:1906.02845.
86. Serrà, J.; Álvarez, D.; Gómez, V.; Slizovskaia, O.; Núñez, J.F.; Luque, J. Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
87. Schirrmeister, R.T.; Zhou, Y.; Ball, T.; Zhang, D. Understanding Anomaly Detection with Deep Invertible Networks through Hierarchies of Distributions and Features. *arXiv* **2020**, arXiv:2006.10848.

88. Houthoofd, R.; Chen, X.; Chen, X.; Duan, Y.; Schulman, J.; De Turck, F.; Abbeel, P. VIME: Variational Information Maximizing Exploration. In *Advances in Neural Information Processing Systems*; Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; Garnett, R., Eds.; Curran Associates, Inc.: Nice, France, , 2016; Volume 29.
89. Bellemare, M.G.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; Munos, R. Unifying count-based exploration and intrinsic motivation. *arXiv* **2016**, arXiv:1606.01868.
90. Hanna, A.; Denton, E.; Smart, A.; Smith-Loud, J. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 27–30 January 2020; pp. 501–512.
91. Wang, C.; Cho, K.; Gu, J. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 7–12 February 2020; pp. 9154–9160.
92. de Vries, T.; Misra, I.; Wang, C.; van der Maaten, L. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, 16–17 June 2019; pp. 52–59.
93. Du, Y.; Mordatch, I. Implicit generation and modeling with energy based models. *Adv. Neural Inf. Process. Syst.* **2019**, 3608–3618. Available online: <https://openreview.net/forum?id=S1laPVSxIS> (accessed on 12 December 2021).
94. Grathwohl, W.; Wang, K.C.; Jacobsen, J.H.; Duvenaud, D.; Norouzi, M.; Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. In *Proceedings of the International Conference on Learning Representations*, Addis Ababa, Ethiopia, 26–30 April 2020.
95. Liu, H.; Abbeel, P. Hybrid Discriminative-Generative Training via Contrastive Learning. *arXiv* **2020**, arXiv:2007.09070.
96. Kurenkov, A. Lessons from the PULSE Model and Discussion. *Gradient* **2020**, 11.
97. Birhane, A.; Prabhu, V.U. Large Image Datasets: A Pyrrhic Win for Computer Vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikola, HI, USA, 5–9 January 2021; pp. 1537–1547.
98. Paullada, A.; Raji, I.D.; Bender, E.M.; Denton, E.; Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv* **2020**, arXiv:2012.05345.
99. Van Rossum, G.; Drake, F.L., Jr. *Python Reference Manual*; Centrum voor Wiskunde en Informatica: Amsterdam, The Netherlands, 1995.
100. Oliphant, T.E. Python for scientific computing. *Comput. Sci. Eng.* **2007**, 9, 10–20. [[CrossRef](#)]
101. Oliphant, T.E. *A guide to NumPy*; Trelgol Publishing: Spanish Fork, UT, USA, 2006.
102. Walt, S.V.d.; Colbert, S.C.; Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, 13, 22–30. [[CrossRef](#)]
103. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J. Array programming with NumPy. *Nature* **2020**, 585, 357–362. [[CrossRef](#)] [[PubMed](#)]
104. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, 17, 261–272. [[CrossRef](#)]
105. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, 9, 90–95. [[CrossRef](#)]