Further Analysis of Outlier Detection with Deep Generative Models

Ziyu Wang[1], **Bin Dai**[2], **David Wipf**[3] **and Jun Zhu**[1]
[1]Tsinghua University, Beijing, China
[2]Samsung Research China, Beijing, China
[3]AWS AI Lab, Shanghai, China
{wzy196,daib09physics,davidwipf}@gmail.com, dcszj@mail.tsinghua.edu.cn

Abstract

The recent, counter-intuitive discovery that deep generative models (DGMs) can frequently assign a higher likelihood to outliers has implications for both outlier detection applications as well as our overall understanding of generative modeling. In this work, we present a possible explanation for this phenomenon, starting from the observation that a model's typical set and high-density region may not conincide. From this vantage point we propose a novel outlier test, the empirical success of which suggests that the failure of existing likelihood-based outlier tests does not necessarily imply that the corresponding generative model is uncalibrated. We also conduct additional experiments to help disentangle the impact of low-level texture versus high-level semantics in differentiating outliers. In aggregate, these results suggest that modifications to the standard evaluation practices and benchmarks commonly applied in the literature are needed.

1 Introduction

Outlier detection is an important problem in machine learning and data science. While it is natural to consider applying density estimates from expressive deep generative models (DGMs) to detect outliers, recent work has shown that certain DGMs, such as variational autoencoders (VAEs [1]) or flow-based models [2], often assign similar or higher likelihood to natural images with significantly different semantics than the inliers upon which the models were originally trained [3, 4]. For example, a model trained on CIFAR-10 may assign higher likelihood to SVHN images. This observation seemingly points to the infeasibility of directly applying DGMs to outlier detection problems. Moreover, it also casts doubt on the corresponding DGMs: One may justifiably ask whether these models are actually well-calibrated to the true underlying inlier distribution, and whether they capture the high-level semantics of real-world image data as opposed to merely learning low-level image statistics [3]. Building on these concerns, various diagnostics have been deployed to evaluate the calibration of newly proposed DGMs [5–9], or applied when revisiting older modeling practices [10].

As we will review in Section 5, many contemporary attempts have been made to understand this ostensibly paradoxical observation. Of particular interest is the argument from *typicality*. Samples from a high-dimensional distribution will often fall on a *typical set* with high probability, but the typical set itself does not necessarily have the highest probability density at any given point. Per this line of reasoning, to determine if a test sample is an outlier, we should check if it falls on the typical set of the inlier distribution rather than merely examining its likelihood under a given DGM. However, previous efforts to utilize similar ideas for outlier detection have not been consistently successful [3, 11]. Thus it is unclear whether the failure of the likelihood tests studied in [3] should be attributed to the discrepancy between typical sets and high-density regions or instead, the miscalibration of the corresponding DGMs. The situation is further complicated by the recent discovery that certain

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

energy-based models (EBMs) do actually assign lower likelihoods to these outliers [5, 6], even though we present experiments indicating that the probability density function (pdf) produced by these same models at out-of-distribution (OOD) locations can be inaccurate.

In this work we will attempt to at least partially disambiguate these unresolved findings. To this end, We first present an outlier test generalizing the idea of the typical set test. Our test is based on the observation that applying the typicality notion requires us to construct an independent and identically distributed (IID) sequence out of the inlier data, which may be too difficult given finite samples and imperfect models. For this reason, we turn to constructing sequences satisfying weaker criteria than IID, and utilizing existing tests from the time series literature to check for these properties. Under the evaluation settings in previous efforts applying DGMs to outlier detection, our test is found to work well, suggesting that the previously-observed failures of outlier tests based on the DGM likelihood should not be taken as unequivocal evidence of model miscalibration per se. We further support this claim by demonstrating that even the pdf from a simple multivariate Gaussian model can mimic the failure modes of DGMs.

Beyond these points, our experiments also reveal a non-trivial shortcoming of the existing outlier detection benchmarks. Specifically, we demonstrate that under current setups, inlier and outlier distributions can often be differentiated by a simple test using linear autocorrelation structures applied in the original image space. This implies that contrary to prior belief, these benchmarks do not necessarily evaluate the ability of DGMs to capture semantic information in the data, and thus alternative experimental designs should be considered for this purpose. We present new benchmarks that help to alleviate this problem.

The rest of the paper is organized as follows: In Section 2 we review the typicality argument and present our new outlier dectection test. We then evaluate this test under a range of settings in Section 3. Next, Section 4 examines the difficulty of estimating pdfs at OOD locations. And finally, we review related work in Section 5 and present concluding discussions in Section 6.

2 From Typicality to a White Noise Test

2.1 OOD Detection and the Typicality Argument

It is well-known that model likelihood can potentially be inappropriate for outlier detection, especially in high dimensions. For example, suppose the inliers follow the d-dimensional standard Gaussian distribution, $p_{\text{in}}(x) \propto \exp(-\|x\|_2^2/2)$, and the test sample is the origin. By concentration inequalities, with overwhelming probability an inlier sample will fall onto an annulus with radius $\sqrt{d}(1 \pm o(1))$, the typical set, and thus the test sample could conceivably be classified as outlier. Yet the (log) pdf of the test sample is higher than most inlier samples by $O(d)$. This indicates that the typical set does not necessarily coincide with regions of high density, and that to detect outliers we should consider checking if the input falls into the former set. We refer to such a test as the *typicality test*.

Remark 2.1. *The relevance of the typicality argument can be demonstrated through the following variant of the Gaussian example: suppose the inlier distribution is $\mathcal{N}(\mu, \Sigma)$, where Σ has a fast-decaying eigenspectrum corresponding to latent factors of varying level of importance. E.g., in a natural image distribution the leading eigenvectors may correspond to semantic variables, while the trailing eigenvectors may correspond to textural or background information.*[1] *In this case an outlier distribution purely representing background texture may have a covariance spanned by the least significant eigenspaces of Σ, with similar eigenvalues; and similar to the above example, its samples usually have higher density when evaluated using p_{in}.*

However, the typicality test is not directly applicable to general distributions, since it is difficult to generalize the notion of typical set beyond simple cases such as component-wise independent distributions, while maintaining a similar concentration property.[2] One appealing proposal that generalizes this idea is to fit a deep latent variable model (LVM) on the inlier dataset using a

[1] [12] focuses on a similar scenario where the input can be decomposed into a foreground and a background part. However, the typicality argument is more general, and corresponding tests do not require an explicit decomposition.

[2] While several papers have referred to the typical set for general distributions (e.g. a natural image distribution) which can be defined using the notion of weak typicality [13], we are only aware of concentration results for log-concave distributions [14], or for stationary ergodic processes [13]. Neither setting describes general distributions encountered in many practical applications.

factorized prior, so that we can transform the inlier distribution back to the prior and invoke the typicality test in the latent space. This idea has been explored in [3], where the authors conclude that it is not effective. One possible explanation is that for such a test to work, we must accurately identify the LVM, which may be far more difficult than generating visually plausible samples, requiring a significantly larger sample size and/or better models. Overall, the idea of typicality has not yet been successfully applied to single-sample outlier detection for general inlier distributions.

2.2 A White Noise Test for Outlier Detection

As we focus on the high-dimensional case, it is natural to take a longitudinal view of data, and interpret a d-dimensional random variable x as a sequence of d random variables. From this perspective, the aforementioned LVM test essentially transforms x to another sequence $T(x)$, so that when $x \sim p_{\text{in}}$, $T(x)$ is IID.[3] Given a new sample x', the test evaluates whether $T(x')$ is still IID by checking the value of $\sum_{i=1}^{d} T_i(x')^2$. The statistical power of the test is supported by concentration properties.

Of course IID is a strong property characterizing the lack of any dependency structure in a sequence, and transforming a long sequence back to IID may be an unreasonable objective. Thus it is natural to consider alternative sequence mappings designed to achieve a weaker criteria, and then subsequently test for that criteria. In the time series literature, there are two such weaker possibilities: the martingale difference (MD) and white noise (WN). A sequence x is said to be a MD sequence if $\mathbb{E}(x_t|x_{<t}) = 0$ for all t; x is said to be WN if for all $s \neq t$, $\text{Cov}(x_t, x_s) = 0, \text{Var}(x_s) = 1$. It is thus clear that for sequences with zero mean and unit variance, MD is a weaker property than IID, and WN is weaker than MD.

While IID sequences are automatically MD and WN, we can also construct WN or MD sequences from inlier samples using residuals from autoregressive models per the following:

Claim 2.1. *Let $\tilde{R}_t(x) := x_t - \mathbb{E}_{p_{\text{in}}}(x_t|x_{<t})$ and $R(x) := \tilde{R}_t(x)/\sqrt{\text{Var}_{p_{\text{in}}}(\tilde{R}_t(x))}$; let $W_t(x) := x_t - \sum_{s=1}^{t-1} a_{ts}x_s$, where the lower triangular matrix $A = (a_{ts})$ is the inverse of the Cholesky factor of $\text{Cov}_{x\sim p_{\text{in}}}(x)$. Assume $\text{Var}_{p_{\text{in}}}(\tilde{R}_t) > 0$ for all t. Then when $x \sim p_{\text{in}}$, $\tilde{R}(x), R(x)$ are both MD, and $R(x), W(x)$ are both WN.*

The first claim above follows from definition. For the second, R is WN because it is MD and has unit variance. Also, W is WN since $\text{Cov}_{x\sim p_{\text{in}}}[W_t(x)] = I$.

The conditional expectations in R can be estimated with deep autoregressive models. For convenience we choose to estimate them with existing autoregressive DGMs in literature (e.g. PixelCNN). However, even though we are fitting generative models, we only need to estimate the mean of the autoregressive distributions $\{p(x_t|x_{<t})\}$ accurately, as opposed to estimating the entire probability density function. For this reason, tests using R should be more robust against estimation errors than tests based on model likelihood.

As testing for the MD property is difficult, we choose to test the weaker WN property. This can be implemented using the classical Box-Pierce test statistics [15]

$$Q_{\text{BP}} := d \textstyle\sum_{l=1}^{L} \hat{\rho}_l^2, \tag{1}$$

where $\hat{\rho}_l$ is the l-lag autocorrelation estimate of a test sequence $(T_t(x))_{t=1}^{d}$. In practice, we can use either W or R as the test sequence, which are both WN when constructed from inliers. When (T_t) has zero mean and unit variance, we have $\hat{\rho}_l = \frac{1}{d-l}\sum_{t=1}^{d-l} T_t T_{t+l}$. We consider a data point x_{test} more likely to be outlier when $Q_{\text{BP}}(x_{\text{test}})$ is larger. Under the context of hypothesis testing where a binary decision (whether x_{test} is an outlier) is needed, we can determine the threshold using the distribution of Q_{BP} evaluated on inlier data.

In high dimensions, formally characterizing the power of a outlier test can be difficult; as illustrated in Section 2.1, it is difficult to even find a proper definition of outlier that is simultaneously practical. Nonetheless, the following remark provides some intuition on the power of our test, when the test sequence derived from outliers has non-zero autocorrelations. This is a natural assumption for image data, where the residual sequence from outlier data could contain more unexplained semantic

[3] Note that such a transformation is possible as long as p_{in} is absolutely continuous w.r.t. the Lebesgue measure; it does not require x to represent truly temporal data.

information, which subsequently contributes to higher autocorrelation; see Appendix A for empirical verification and further discussion on this matter.

Remark 2.2 (Connection with the concentration-of-measure phenomenon)**.** *The power of the Box-Pierce test is supported by a concentration-of-measure phenomenon: When $\{T_t(x)\}$ is IID Gaussian,*[4] *Q_{BP} will approximately follow a χ^2_L distribution [15], and Q_{BP}/L will concentrate around 1. On the other hand, if the null hypothesis does not hold and there exists a non-zero ρ_l, Q_{BP}/L will be at least $d\rho_l^2/L$, which is much larger than 1 when d is large.*

It should be noted, however, that our test benefits from the concentration phenomenon in a different way comparing to the typicality test. As an example, consider the following outlier distribution, which connects to the colored noise example in [16]: for $x \sim p_{\text{ood}}$, $(T_1(x), T_2(x))$ follow the uniform distribution on the circle centered at origin with radius $\sqrt{2}$, and $T_j(x) = T_{j-2}(x)$ for $j > 2$. Then $\frac{1}{d}\sum_{j=1}^d T_j^2(x) = 1$, and thus the typicality test cannot detect such outliers. In contrast, our test will always detect the lag-2 autocorrelation in T, and, as described above, reject the null hypothesis.

2.3 Implementation Details

Incorporating prior knowledge for image data: When applied to image data, the power of the proposed test can be improved by incorporating prior knowledge about outlier distributions. Specifically, as the test sequence $T(x)$ is obtained by stacking residuals of natural images, ρ_l is likely small for the lags l that do not align with fixed offsets along the two spatial dimensions. As the corresponding finite-sample estimates $\hat{\rho}_l$ are noisy (approximately normal), they constitute a source of independent noise that has a similar scale in both inlier and outlier data, and removing them from (1) will increase the gap between the distributions of the test statistics computed from inlier and outlier data, consequently improving the power of our test. For this reason, we modify (1) to only include lags that correspond to vertical autocorrelations in images. When the data sequence is obtained by stacking an image with channel-last layout (i.e., for $x_{3(W(i-1)+j)+c}$ refers to the c-th channel of the (i,j) pixel of a $H \times W$ RGB image), we will only include lags that are multiples of $3W$. For empirical verifications and further discussion on this issue, see Appendix A.

Testing on transformed data: Instead of fitting autoregressive models directly in the input space, we may also fit them on some transformed domain, and use the resulting residual for the WN test. Possible transformations include residuals from VAEs and lower-level latent variables from hierarchical generative models (e.g. VQ-VAE).[5] This can be particularly appealing for the test using (W_t), as linear autoregressive models have limited capacity and cannot effectively remove nonlinear dependencies from data, yet the lack of dependency seems important for the Box-Pierce test, as suggested by Remark 2.2.

3 Evaluating the White Noise Test

In this section we evaluate the proposed test, with the goal of better understanding the previous findings in [3]. We consider three implementations of our white noise test, which use different sequences to compute the test statistics (1):

- the residual sequence R, estimated with autoregressive DGMs (denoted as **AR-DGM**);
- the residual sequence W from a linear AR model, directly fitted on the input space (**Linear**);
- the sequence W constructed from a linear model fitted on the space of VAE residuals (**VAE+linear**).

Note that both R and W can be viewed as constructed from generative models: for the sequence W, the corresponding model is a simple multivariate normal distribution. Therefore, we can always gain insights from comparing our test to other tests based on the corresponding generative model.

3.1 Evaluation on Standard Image Datasets

We first evaluate our white noise test following the setup in [3], where the outlier data comes from standard image datasets, and can be different from inlier data in terms of both low-level details

[4] It is common to use the B-P test in the more general, non-IID case, so long as we are interested in alternative hypotheses where autocorrelation structure exist. Also recall that $\{T_t\}$ are residuals from an autoregressive model, so this condition is much weaker than requiring x to be IID.

[5] Note this is different from testing with the sequence R, which is constructed from autoregressive models.

(textures, etc) as well as high-level semantics. In Appendix B we present additional experiments under a similar setup, in which we compare with more baselines.

Evaluation Setup: We use CIFAR-10, CelebA, and TinyImageNet images as inliers, and CIFAR-10, CelebA and SVHN images as outliers. All colored images are resized to 32×32 and center cropped when necessary. For deep autoregressive models, we choose PixelSNAIL [17] when the inlier dataset is TinyImageNet, and PixelCNN++ [18] otherwise. We use the pretrained unconditional models from the respective papers when possible; otherwise we train models using the setups from the paper.[6] For the VAE-based tests, we use an architecture similar to [19], and vary the latent dimension n_z as it may have an influence on the likelihood-based outlier test. See Appendix C.1 for more details.

We compare our test (**WN**) with three baselines that have been suggested for generative-model-based outlier detection: a single-sided likelihood test (**LH**), a two-sided likelihood test (**LH-2S**), and, for the DGM-related tests, the likelihood-ratio test proposed in [20] (**LR**). The LH test classifies samples with lower likelihood as outliers. The LH-2S test classifies samples with model likelihood deviated from the inlier median as outliers. It can be viewed as testing if the input falls into the *weakly typical set* [13];[7] while there is no concentration guarantee in the case of general inlier distributions, it is natural to include such a baseline. The LR test is a competitive approach to single-sample OOD detection; it conducts a single-sided test using the statistics $\log \frac{p_{model}(x)}{p_{generic}(x)}$, where $p_{generic}$ refers to the distribution corresponding to some generic image compressor (e.g., PNG). Samples with a lower value of this statistics is considered outlier. The test is based on the assumption that outlier samples with a higher model likelihood may have inherently lower complexity, as measured by $\log p_{generic}$. The test statistics, having the form of a Bayes factor, and can also be viewed that comparing two competing hypotheses (p_{model} and $p_{generic}$) without assuming either is true [22].

Table 1: AUROC values for the single-sample test, and average ranks within each group. **Boldface** indicates best results; underline indicates notable failures (AUC < 0.5).

Inlier Dist.		CIFAR-10		CelebA		TinyImageNet		Avg.
Outlier Dist.		CelebA	SVHN	CIFAR-10	SVHN	CIFAR-10	SVHN	Rank
AR-DGM	LH	0.88	0.16	0.82	0.15	0.28	0.05	3.67
	LH-2S	0.77	0.69	0.84	0.78	0.55	0.93	2.50
	LR	0.86	0.86	0.99	1.00	0.39	0.56	2.00
	WN	0.97	0.83	0.85	0.93	0.85	0.62	**1.67**
VAE+Linear $n_z = 64$	LH	0.64	0.09	0.88	0.26	0.28	0.04	3.33
	LH-2S	0.47	0.81	0.85	0.69	0.51	0.87	3.00
	LR	0.39	0.90	0.98	0.99	0.64	0.91	1.83
	WN	0.64	0.67	0.93	0.99	0.92	0.99	**1.50**
VAE+Linear $n_z = 512$	LH	0.76	0.04	0.81	0.09	0.19	0.01	3.33
	LH-2S	0.61	0.85	0.76	0.81	0.59	0.90	2.67
	LR	0.56	0.86	0.97	0.99	0.55	0.90	2.50
	WN	0.61	0.88	0.88	1.00	0.94	0.99	**1.33**
Linear	LH	0.77	0.02	0.72	0.03	0.11	0.00	2.50
	LH-2S	0.69	0.76	0.70	0.80	0.64	0.81	2.17
	WN	0.67	0.95	0.90	0.99	0.92	0.99	**1.33**

Results and Discussion: We compare the distribution of the test statistics on the inlier test data and outlier datasets, and report the AUROC values. The results are shown in Table 1, where we observe that our WN proposal outperforms all the others in terms of the average ranking across testing conditions; see rightmost column. (We have deferred to Appendix C.1 the results of likelihood-based tests based on multivariate normal models fitted on VAE residuals, as those tests did not work well.)

Drilling further into details, we can see that our WN test generally outperforms the likelihood-based tests, and the single-side likelihood test exhibits pathological behaviors. This happens across all choices of generative models, including the simple Gaussian model corresponding to the linear test. Therefore, it is reasonable to doubt whether the previously observed failures of likelihood-based tests

[6] This choice is made to maximize model capacity within the limit of computational resources we have.

[7] It can also be viewed as the single-sample version of [21].

should be attributed to some undesirable properties of DGMs. Alternatively, those results may be better explained by the counter-intuitive properties of high-dimensional probability, as in Section 2.1. Furthermore, the fact that we can always construct a principled test statistics out of generative models suggests that these models have in some sense calibrated behavior on such outliers. In other words, under these settings the models do know what they don't know. Our result is to be compared with the recent discovery that EBMs assign lower likelihood to outliers under this setting [5, 6], which naturally leads to the question of whether a calibrated DGM should always have a similar behavior. However, our findings are not necessarily inconsistent with theirs, as we explain in Section 4.

Comparison between our test and the LR test is more nuanced, as the latter is also competitive in many cases. Still, the LR test consistently produces a slightly higher average rank, and also has two cases of notable failures.

Finally, note that the simple linear generative model, especially when combined with the WN test, works well in most cases. This challenges the intuition that the inflexibility of a linear model would hamper outlier-detection performance, and has two-fold implications. First, these results indicate that the linear white-noise test could be useful in practice, as it is easy to implement, and does not have unexpected failures like the likelihood tests. Hence, it could be applied as a cheap, first test in a detection pipeline. And secondly, the success of the linear test shows that the current benchmarks leave a lot to be desired, since it implies that the differences between the inlier and outlier distributions being exploited for outlier detection are mostly low-level. Consequently, it remains unclear if these benchmarks are adequate for showcasing tests that are sensitive to semantic differences. Such a semantics-oriented evaluation is arguably more important for downstream applications. Moreover, it better reflects the ability of DGMs to learn high-level semantics from data, as was the intent of [3]. To address this issue, in the following subsection we conduct additional experiments that are more focused on semantics.

3.2 Semantics-Oriented Evaluation

In this section we evaluate the OOD tests in scenarios where the inlier and outlier distributions have different semantics, but the influence from background or textual differences is minimized. We consider two setups:

- **CIFAR**, in which we use CIFAR-10 images as inliers and a subset of CIFAR-100 as outliers. In this setup the inlier and outlier distributions have significantly different semantics, as we have removed from CIFAR-100 all classes that overlap with CIFAR-10, namely, non-insect creatures and vehicles. Furthermore, this setup also reduces textual differences contributed by inconsistent data collection processes; note that both CIFAR datasets have been created from the 80 Million Tiny Images dataset [23].
- **Synthetic**, in which we further reduce the background and textual differences between image classes by using synthesized images from BigGAN [24]. The outliers are class-conditional samples corresponding to two semantically different ImageNet classes; the inlier distribution is obtained by interpolating between these two classes using the GAN model. In this case, the semantic difference between inlier and outlier distributions is smaller, although in most cases it is still noticeable. We construct three benchmarks under this setting. Detailed settings and sample images are postponed to Appendix C.2.

Table 2: Results for the semantics-oriented experiments. Boldface indicates the best result.

	CIFAR, AUROC↑				Synthetic, Avg. Rank↓			
	LH	LH-2S	LR	WN	LH	LH-2S	LR	WN
AR-DGM	0.49	0.57	**0.61**	0.58	**2**	3.5	2.5	**2**
Linear	0.56	0.59	-	**0.60**	2.33	**1.67**	-	2
VAE+Linear, 64	0.51	0.55	0.64	**0.84**	**1.67**	3.33	2.67	2.33
VAE+Linear, 512	0.59	0.58	0.73	**0.80**	**2**	3.67	**2**	2.33

The results are summarized in Table 2, with full results for the synthetic experiments deferred to Appendix C.2. In the CIFAR setup, none of the tests that are based on the AR DGM or the vanilla Gaussian model works well, which is consistent with the common belief that these models cannot capture the high-level semantics. When using VAEs, the WN test works well. This experiment reaffirms that DGMs such as VAEs are able to distinguish between distributions with significantly different semantics, even though they may assign similar likelihood to samples from both distributions.

However, as we move to the synthetic setup where the semantic difference is smaller but still evident, the outcome becomes quite different. The LH test performs much better, and our test no longer consistently outperforms the others. It is also interesting to note that the LR test does not work well on the second synthetic setup (see Appendix C.2), and completely fails to distinguish between inliers and outliers when using an autoregressive DGM. To understand this failure, we plot the distributions of model likelihood and test statistics in Appendix C.2. We can see that the outlier distribution has a slightly higher complexity as measured the generic image compressor, contrary to the assumption in [20] that the lower input complexity of outliers causes the failure of likelihood-based OOD test.

The difference in outcome between these experiments and Section 3.1 demonstrates the difficulty in developing a universally effective OOD test. It is thus possible that in the purely unsupervised setting we have investigated, OOD tests are best developed on a problem-dependent basis. Compared with Section 3.1, we can also see that the previous evaluation setups do not adequately evaluate the ability of each test to measure semantic differences. For this purpose, our approach may be more appropriate.[8]

4 On the Difficulty of Density Estimation in OOD Regions

While DGMs such as GANs, VAEs, autoregressive models, and flow-based models tend to assign higher likelihoods to certain OOD images, high-capacity energy-based models have been shown at times to have the opposite behavior [5, 6]. This observation naturally leads to the question of whether calibrated generative models trained on natural image datasets should always assign lower likelihood to such outliers. In this section, we argue that such a question is unlikely to have a clear-cut answer, by showing that given the relatively small sample size of typical image datasets compared to the high dimensionality of data, density estimation on OOD regions is intrinsically difficult, and even models such as EBMs can make mistakes.

Specifically, we train a PixelCNN++ and the high-capacity EBM in [5] on samples generated by a VAE. Since by design we have access to (lower bounds of) the true log probability density of the inlier distribution, we can check if a test model's density estimation in OOD regions is correct, simply by comparing it to the ground truth.

Our ground truth VAE has the same architecture as in Section 3, with $n_z = 64$; training is conducted on CIFAR-10. The DGMs to be tested are trained using 80000 samples from the VAE, under the same setup as in the original papers. See Appendix C.3 for details. We generate outliers by setting half of the latent code in the VAE to zero. Such outliers are likely to have a higher density under the ground truth model, per the reasoning from Section 2.1. Therefore, a DGM that correctly estimates the ground-truth data pdf should also assign higher likelihood to them.

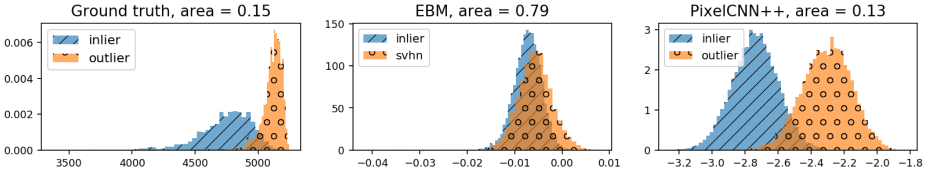


Figure 1: Distribution of log likelihood approximations from the ground-truth VAE (left), EBM (center), and PixelCNN++ (right). The intersection area of the two histograms is reported at the top.

The distributions of density estimates are shown in Figure 1. We can see that while both the EBM and PixelCNN++ models being tested assign a higher relative likelihood to the outliers (note that the absolute likelihoods between different models are not comparable because of different scaling and offset factors), the inlier and outlier density estimates from the EBM overlap significantly (middle plot) as compared to analogous overlap within the ground-truth VAE (left plot). Such behavior may be attributed to the inductive bias of the EBM, which has a stronger influence than data on the estimated pdf in OOD regions given the relatively small sample size.

[8] To balance the discussion, note that in some cases it may be desirable to have a benchmark outlier dataset with low-level differences, as such differences could be detrimental to down-stream applications. An example is the low-level differences of radiographs taken from different medical sites, which can influence diagnostics models [25]. Detection of such differences can be of practical interest in this context.

While we conjecture that VAEs or deep AR models can exhibit similar failures due to a different type of inductive bias, we cannot reverse the above experiment and train these models on EBM samples, as sampling from EBMs rely on ad hoc processes such as premature termination of MCMC chains [5, 6, 26]. Nonetheless, our experiment has demonstrated the intrinsic difficulty of density estimation in OOD regions under the finite-sample, high-dimensional setting. For this reason, it is difficult to draw a definitive conclusion as to whether real-world outliers should be assigned higher likelihoods, and alternative explanations, such as the typicality argument in Section 2, deserve more attention. The hardness of density estimation in OOD regions also suggests that OOD tests based on DGM likelihood should be used with caution, as is also suggested by the results in Section 3.1.

5 Related Work

Several works have explored the use of DGMs in outlier detection under settings similar to [3], some of which also provided possible explanations to the findings in [3]. For example, [11] presents a heuristic test using the Watanebe-Akaike Information Criterion; however, the efficacy of this test remains poorly understood. As another alternative, [12] proposes to compute the likelihood ratio between the inlier model and a background model, based on the intuition that background can be a confounding factor in the likelihood test. In Appendix B we present evaluations for the two tests, showing that they do not always work across all settings. In Section 3 we have introduced the work of [20], and demonstrated that its assumption does not always hold. In summary then, to date there has not been a comprehensive explanation of the peculiar behavior of generative models on semantically different outliers, although previous works can be illuminating and practically useful in certain scenarios.

For the general problem of high-dimensional outlier detection, methods have also been developed under different settings. For example, [21] proposes a typicality test assuming input contains a batch of IID samples, while [4] assumes a few outlier samples are available before testing. There is also work on outlier detection in supervised learning tasks, where auxiliary label information is available; see, e.g. [27–33].

Finally, it is worth mentioning the formulation of atypicality [16], as motivated by the possible mismatch between the typical set and the high-density regions. The atypicality test considers a test sequence to be OOD when there exists an alternative model leading to a smaller description length [34]. Our test shares a similar spirit in the use of autocorrelation estimation, which can be viewed as fitting a linear alternative model. However, our test works on the residual sequence from a flexible generative model, which benefits from information in the inlier distribution; in contrast, they opt to estimate $p(x_t|x_{<t})$ *directly* on a test point x. This becomes problematic when x cannot be viewed as a stationary process, and prevents the use of flexible models such as DGMs.

6 Discussion

The recent discovery that DGMs may assign higher likelihood to natural image outliers casts into doubt the calibration of such models. In this work, we present a possible explanation based on an OOD test that generalizes the notion of typicality. In evaluations we have found that our test is effective under the previously used benchmarks, and that such peculiar behaviors of model likelihood are not restricted to DGMs. We have also demonstrated that certain DGMs cannot accurately estimate pdfs at OOD locations, even if at times they may correctly differentiate outliers. These findings suggest that it may be premature to judge the merits of a model by its (in)ability to assign lower likelihood to outliers.

Further investigation of the behavior of DGMs on outliers will undoubtedly continue to provide useful insights. However, our analyses suggest a change of practice in such investigations, such as considering alternatives to simply the model likelihood as our proposed test has exemplified. Likewise, the observation that a simple linear test performs well under current evaluation settings also suggests that care should be taken in the design and diversity of benchmark datasets, e.g., inclusion of at least some cases where low-level textures cannot be exclusively relied on.

And finally, from the perspective of unsupervised outlier detection, our experiments also revealed the intrinsic difficulty in designing universally effective tests. It is thus possible that future OOD tests are best developed on a problem-dependent basis, with prior knowledge of potential outlier distributions taken into account. [12] provides an example of such practice.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[3] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?," in *International Conference on Learning Representations*, 2019.

[4] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[5] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," in *Advances in Neural Information Processing Systems 32*, pp. 3608–3618, 2019.

[6] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," *arXiv preprint arXiv:1912.03263*, 2019.

[7] L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther, "BIVA: A very deep hierarchy of latent variables for generative modeling," in *Advances in neural information processing systems*, pp. 6548–6558, 2019.

[8] C. Louizos, X. Shi, K. Schutte, and M. Welling, "The functional neural process," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché Buc, E. Fox, and R. Garnett, eds.), pp. 8746–8757, Curran Associates, Inc., 2019.

[9] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Hybrid models with deep and invertible features," in *International Conference on Machine Learning*, pp. 4723–4732, 2019.

[10] J. Bütepage, P. Poklukar, and D. Kragic, "Modeling assumptions and evaluation schemes: On the assessment of deep latent variable models," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[11] H. Choi, E. Jang, and A. A. Alemi, "WAIC, but why? generative ensembles for robust anomaly detection," 2018.

[12] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *arXiv preprint arXiv:1906.02845*, 2019.

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

[14] S. Bobkov, M. Madiman, *et al.*, "Concentration of the information in data with log-concave distributions," *The Annals of Probability*, vol. 39, no. 4, pp. 1528–1543, 2011.

[15] G. E. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *Journal of the American statistical Association*, vol. 65, no. 332, pp. 1509–1526, 1970.

[16] E. Sabeti and A. Host-Madsen, "Data discovery and anomaly detection using atypicality for real-valued data," *Entropy*, vol. 21, no. 3, p. 219, 2019.

[17] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "PixelSNAIL: An improved autoregressive generative model," in *International Conference on Machine Learning*, pp. 864–872, 2018.

[18] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[19] B. Dai and D. Wipf, “Diagnosing and enhancing VAE models,” in *International Conference on Learning Representations*, 2018.

[20] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, “Input complexity and out-of-distribution detection with likelihood-based generative models,” in *International Conference on Learning Representations*, 2020.

[21] E. Nalisnick, A. Matsukawa, Y. W. Teh, and B. Lakshminarayanan, “Detecting out-of-distribution inputs to deep generative models using typicality,” *arXiv preprint arXiv:1906.02994*, 2019.

[22] A. Ly, J. Verhagen, and E.-J. Wagenmakers, “Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology,” *Journal of Mathematical Psychology*, vol. 72, pp. 19–32, 2016.

[23] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[24] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in *International Conference on Learning Representations*, 2018.

[25] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study,” *PLoS medicine*, vol. 15, no. 11, 2018.

[26] E. Nijkamp, M. Hill, T. Han, S.-C. Zhu, and Y. N. Wu, “On the anatomy of mcmc-based maximum likelihood learning of energy-based models,” *arXiv preprint arXiv:1903.12370*, 2019.

[27] A. A. Alemi, I. Fischer, and J. V. Dillon, “Uncertainty in the variational information bottleneck,” *arXiv preprint arXiv:1807.00906*, 2018.

[28] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” *arXiv preprint arXiv:1706.02690*, 2017.

[29] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” in *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

[30] L. Bergman and Y. Hoshen, “Classification-based anomaly detection for general data,” *arXiv preprint arXiv:2005.02359*, 2020.

[31] I. Golan and R. El-Yaniv, “Deep anomaly detection using geometric transformations,” in *Advances in Neural Information Processing Systems*, pp. 9758–9769, 2018.

[32] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, “Using self-supervised learning can improve model robustness and uncertainty,” in *Advances in Neural Information Processing Systems*, pp. 15663–15674, 2019.

[33] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” in *International conference on machine learning*, pp. 4393–4402, 2018.

[34] P. D. Grünwald and A. Grunwald, *The minimum description length principle*. MIT press, 2007.

[35] C. Cortes and M. Mohri, “Confidence intervals for the area under the roc curve,” in *Advances in neural information processing systems*, pp. 305–312, 2005.

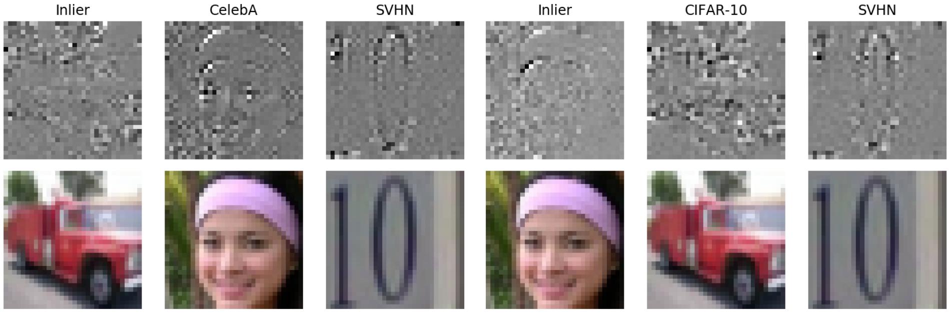
Appendix

A On the Assumptions and Efficacy of the White Noise Test

In this section we provide visualizations to better understand the statistical power of our test, and to verify the claims in Section 2.3.

We first plot samples of the residual sequence R in Figure 2,[9] under varying choices of inlier and outlier distributions. We can see that R constructed from outlier images generally include a higher proportion of unexplained semantic information: comparing the CelebA residual in Fig.2(a) (second column) where the model is trained on CIFAR-10, to Fig.2(b) (first column) where CelebA is inlier, we can see that the facial structure in CelebA residual is more evident when the model is trained on CIFAR-10. Similarly, comparing the CIFAR-10 residual from both models, we can see that the structure of the vehicle (e.g. front window and car frame) is more evident when the model is trained on CelebA. As the residual sequences constructed from outliers tend to have more natural image-like structures, they will also have stronger spatial autocorrelations, compared with residuals from inlier samples that should in principle be white noise.

Note that while the residual sequences constructed from inliers also contain unexplained semantic information, this is due to estimation error of the deep AR model, and should not happen should we have access to the ground truth model, as we have shown in Section 2.2. Moreover, the estimation error should have a small impact on the efficacy of the white noise test, as it is very easy to learn the correct linear autocorrelation structure of the inlier distribution, and thus the deviation of R from WN is usually small, as we show in Figure 3 right.



(a) Inlier: CIFAR-10 (b) Inlier: CelebA

Figure 2: Samples of normalized residual R on different datasets, and the corresponding input images. The left 3 columns are generated from a PixelCNN++ trained on CIFAR-10; the right 3 columns corresponds to CelebA.

We now turn to the verification of our prior belief about the autocorrelation structure in $T(x_{\mathrm{test}})$, when x_{test} comes from the outlier distribution. Specifically, we plot the *average* ACFs on inlier and outlier data in Figure 3. We can see that the ACF estimates on outlier residuals peaks at lags that are multiples of 96, which corresponds to the vertical spatial autocorrelations in $32 \times 32 \times 3$ images. Moreover, on inlier and outlier distributions, the ACF estimates at other lags have approximately equal variances. When aggregated, these estimates will constitute a noticeable source of noise which reduces the gap between the distributions of inlier and outlier test statistics, and thus excluding them from the statistics will improve the power of the WN test.

Finally, we remark that it is also possible use spatial correlations directly in the construction of test statistics. However, our main focus in this work is to understand previous findings in generative outlier detection (instead of improving the state-of-the-art of OOD tests), and our choice to include only the vertical spatial autocorrelations is good enough for this purpose.

[9] With a slight abuse of notation, we use R to refer to both the MD sequence constructed from true conditional expectation $\mathbb{E}_{p_{\mathrm{in}}}(x_t|x_{<t})$, and the sequence constructed with DGM-based estimation to the conditional expectation.

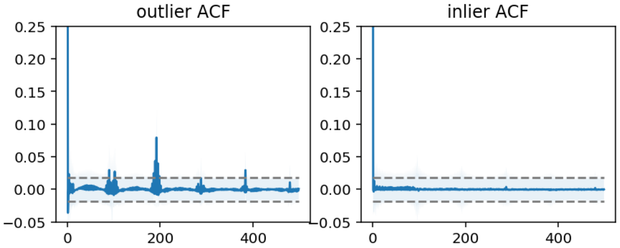


Figure 3: Averaged ACF estimates and their standard deviations (over sample images) on PixelCNN++ residuals. Left: residual generated from outlier (SVHN); right: from inlier (CIFAR-10) test set. Shaded area indicates the standard deviation of $\hat{\rho}_l$ *where the randomness is from the input data* x. Gray dashed line indicates the standard deviation of $\hat{\rho}_l$ *under the null hypothesis of IID residuals.*

B More Experiments on Standard Image Datasets

In this section we conduct additional experiments, and evaluate a variety of generative outlier detection methods under a common setting. As we will see, while several tests are in general more competitive than others, no single test achieves the best performance across all settings. This experiment strengthens our argument in the main text that unsupervised OOD tests should be developed on a problem-dependent basis.

Evaluation Setup: We use CIFAR-10 as inlier data. For outliers we consider two setups. The first setup is taken from [20], and consists of 9 generic image datasets and 2 synthetic datasets, const and random; see Appendix A in [20] for details. The second setup controls for low-level differences by using the CIFAR-100 subset constructed in Section 3.2. The tests to be evaluated include those considered in Section 3.1, as well as the WAIC test [11] and the background likelihood ratio (BLR) test [12]. We base these tests on two DGMs: the VAE-512 model used in Section 3.1, and a smaller-capacity PixelCNN++ model as in [12].[10] For the BLR test, a noise level of the background model needs to be determined. Following the recommendations of the authors, we search for the optimal parameter in the range of $\{0.1, 0.2, 0.3\}$ using the grayscaled CIFAR-10 dataset as outlier. We found the optimal noise level to be 0.1, which is consistent with [12].

Results: Results are shown in Table 3-4. When using VAEs, neither of the newly added baselines are very competitive, suggesting that these methods are more prone to model misspecification. Notably, the WAIC test does not work with SVHN as outlier. This is also observed in [12, 21] using different generative models (autoregressive and flow-based models, respectively). For this reason we drop it in the PixelCNN++ experiment.

When we switch to PixelCNN++, the BLR test performs much better under the setting of [20]. However, in either case it does not work well with the subset-of-CIFAR-100 dataset, despite the dataset's clear semantic difference from the inlier dataset. Such results are not surprising since the difference in background or low-level details is much smaller for CIFAR-100 compared with the other datasets, as we have discussed in Section 3.2. Again, the difference in outcome between the two different settings demonstrates the difficulty of constructing universally effective OOD tests in the unsupervised setup.

C Experiment Details and Additional Results

C.1 Details for Section 3.1

Experiment Setup: For the AR-DGM experiments, we use the pretrained unconditional models from official repositories for CIFAR-10 and TinyImageNet. For CelebA we train a PixelCNN++ model using the authors' setup for unconditional CIFAR-10 generation. Both PixelCNN++ and

[10] Using the standard hyperparameters in [18] results in the BLR test rejecting inlier test data as outlier with high confidence (AUROC>0.9). As such a failure mode can be detected without access to outlier samples, we modify the model hyperparameters to follow [12] and train for 20 epochs. The BPD on inlier test set is 3.15.

Table 3: Results for the Experiment in Appendix B, using VAE. Notable failures (with AUC< 0.5) are underlined.

	LH	LH-2S	LR [20]	WAIC[11]	BLR[12]	WN
celeba	0.76	0.61	0.57	0.64	0.22	0.62
svhn	0.04	0.85	0.86	0.14	0.96	0.88
imagenet	0.90	0.88	0.85	0.84	0.08	0.91
facescrub	0.65	0.47	0.52	0.55	0.33	0.60
mnist	0.28	0.42	1.00	0.68	0.81	0.83
fashion	0.45	0.36	1.00	0.64	0.61	0.83
omniglot	0.53	0.46	1.00	0.81	0.57	0.85
trafficsign	0.44	0.71	0.82	0.41	0.55	0.80
random	1.00	1.00	1.00	1.00	0.00	0.96
const	0.16	0.79	1.00	0.62	0.84	1.00
avg. rank	3.8	4.1	**2.2**	3.7	4.3	**2.2**
cifar100’	0.58	0.58	**0.73**	0.58	0.40	**0.80**
(inlier test)	0.45	0.47	0.43	0.63	0.44	0.49

Table 4: Results for the Experiment in Appendix B, using PixelCNN++. Notable failures (with AUC< 0.5) are underlined.

	LH	LH-2S	LR[20]	BLR[12]	WN
imagenet	0.86	0.82	0.88	0.92	0.84
svhn	0.11	0.79	0.80	0.79	0.86
celeba32	0.81	0.64	0.75	0.89	0.97
mnist	0.00	1.00	1.00	0.91	0.98
fashion	0.00	1.00	0.97	0.82	0.96
omniglot	0.00	1.00	1.00	0.98	0.93
facescrub	0.80	0.69	0.82	0.93	0.82
trafficsign	0.55	0.59	0.90	0.90	0.77
random	1.00	1.00	1.00	1.00	1.00
const	0.09	0.87	1.00	0.04	1.00
avg. rank	4.44	3.22	**1.89**	2.78	**2.56**
cifar100’	0.50	0.57	**0.63**	0.45	**0.58**
(inlier test)	0.51	0.50	0.51	0.52	0.51

PixelSNAIL use the discretized mixture-of-logistics (DMOL) likelihood parameteriation. To calculate its expectation, we first calculate the expectation of the continuous mixture of logistics distribution, and then *clip the result to the range of* $[0, 1]$. This is needed because the definition of the DMOL likelihood include a similar truncation [18]: extra probability mass for the interval $(1, +\infty)$ (or $(-\infty, 0)$) are assigned to the discretization bin $[1 - 1/256, 1]$ (or $[0, 1/256]$, respectively), so that the distribution is always supported on $[0, 1]$.

For the VAE experiments, we use the discretized logistics likelihood as the observation model. The network architecture is adapted from [19]; we vary the capacity of the model by increasing the number of filters in convolutional layers by k times, where k may be in $\{1, 2, 4, 8\}$. We train for at most 8×10^5 iterations using a learning rate of 10^{-4}, and perform early stopping based on the validation ELBO. We choose k to maximize validation ELBO. This leads to $k = 1$ for CIFAR-10, 4 for CelebA and 8 for TinyImageNet. This step is needed, because when k is further increased, the reconstruction error will start to have different distributions between training and held-out set. Such a difference would be undesirable for all tests, as they will start to find false differences between the inlier training set and the test set. Note that this difference is not due to overfitting, as we have performed early stopping based on validation ELBO; instead, it is simply due to the fact that the model is exposed to training samples and not validation samples, and the gap appears very early in training. We use ELBO to approximate model likelihood in likelihood-related tests. The discrepancy between ELBO and true model likelihood is likely to have little impact on test performance, since we have also experimented with IWAE_{100} which led to very similar results.

We compare the distributions of the test statistics evaluated on the inlier test set and outlier test set, and report the AUROC value. We verified that the four tests used in this section do not falsely distinguish between inlier training samples and test samples: the AUROC value for such a comparison is always in the range of $(0.42, 0.53)$. For outlier datasets with more than 50000 test samples, we sub-sample 50000 images for evaluation. Using the formula in [35], we can thus show that the maximum possible 95% confidence interval for the AUROC values is ± 0.011. For a description of the four datasets used in this section, please refer to, e.g., Table 3 in [20].

Choice of L and Sensitivity: For our test, we use $L = 1200$ when computing the Box-Pierce statistics (1). This is because while in principle we should include all lags that are known *a priori* to be informative, in practice we only have $d - l$ samples to estimate $\hat{\rho}_l$, so the most distant lags can be difficult to estimate. Nonetheless, the impact of L on the test outcome is relatively small: as is shown in Figure 4, using different L does not lead to qualitatively different outcome. We also note that our purpose in the experiments is not to build new state-of-the-art in OOD detection, but is to use the proposed test to validate our explanation to previous findings. Still, if it is desirable to further improve the performance of the test, we can consider tuning L on "validation outlier datasets" that is known *a priori* to be similar to the outliers that will be encountered in practice, as is done in e.g. [12].

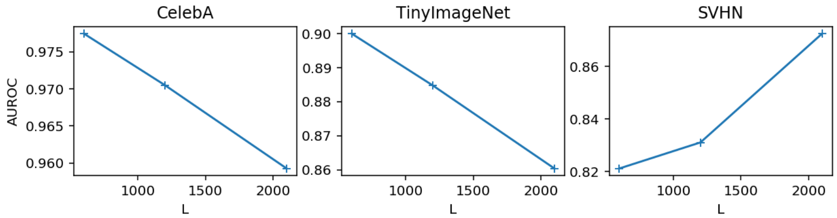


Figure 4: Sensitivity to the maximum number of lags L of the proposed WN test using AR-DGM. Inlier is CIFAR-10.

Results for the Normal Likelihood Test on VAE Residuals: In Table 5 we present results for the likelihood tests using a multivariate normal model fitted on VAE residual, denoted with a prefix of "LN". We also consider both single-side and two-side tests. Overall the performance is similar to DGM likelihood, and the single-side likelihood test still manifests catastrophic failures.

Table 5: Full results for the VAE-related experiment in Section 3.1.

Inlier Dist.		CIFAR-10		CelebA		TinyImageNet		Rank
Outlier Dist.		CelebA	SVHN	CIFAR-10	SVHN	CIFAR-10	SVHN	
VAE-64	DGM-LH	0.64	0.09	0.88	0.26	0.28	0.04	4.50
	DGM-LH-2S	0.47	0.81	0.85	0.69	0.51	0.87	3.67
	LN-LH	0.98	0.10	0.72	0.09	0.08	0.00	4.83
	LN-LH2S	0.98	0.69	0.67	0.74	0.68	0.80	3.17
	LR	0.39	0.90	0.98	0.99	0.64	0.91	2.33
	WN	0.64	0.67	0.93	0.99	0.92	0.99	**2.17**
VAE-512	DGM-LH	0.76	0.04	0.81	0.09	0.19	0.01	4.50
	DGM-LH-2S	0.61	0.85	0.76	0.81	0.58	0.90	3.17
	LN-LH	0.95	0.07	0.68	0.05	0.10	0.00	4.83
	LN-LH2S	0.95	0.72	0.65	0.79	0.64	0.79	3.67
	LR	0.56	0.86	0.97	0.99	0.55	0.90	3.00
	WN	0.61	0.88	0.88	1.00	0.94	0.99	**1.83**

C.2 Details for Section 3.2

The CIFAR Experiment: We use the trained models from Section 3.1. We remove from CIFAR-100 the superclasses 1,2,9,12-17,19,20. For reference, the class names of CIFAR-10 and CIFAR-100 can be found in `https://www.cs.toronto.edu/~kriz/cifar.html`.

The Synthetic Experiments: We use a pretrained BigGAN model on ImageNet 128×128,[11] and down-sample the generated images to 32×32. To generate the outliers, recall the BigGAN generator takes as input a noise vector $z \in \mathbb{R}^{128}$ and the one-hot class encoding vector $c \in \mathbb{R}^{1000}$. Therefore, we interpolate between two classes i and j by setting $c_k = 0.5 \cdot \mathbf{1}_{k \in \{i,j\}}$. There are two tunable parameters in our generation process: the truncation parameter σ that determines the truncated normal prior, and a crop parameter τ. Before down-sampling the generated samples, we apply center-cropping to retain a proportion of $(1-2\tau)^2$ pixels, to reduce the amount of details lost in the down-sampling process. The classes and generation parameters used are listed in Table 6; they are hand-picked to ensure the background is similar in inlier and outlier classes. In each setup we generate 200000 samples and use 80% for training.

The VAEs are trained using the same setting as in Section 3.1. For PixelCNN++ we use the hyperparameters of the unconditional CIFAR-10 experiment in the original paper. As the synthetic datasets contain more samples, we train for 80 epochs.

The full AUROC values for the synthetic experiments are shown in Table 7. We plot the distributions of various statistics related to the LR tests using AR-DGM in the second synthetic experiment in Figure 5. We also plot additional inlier and outlier samples in Figure 6.

Table 6: Generation parameters for the synthetic experiment in Section 3.2.

No.	Class 1	Class 2	σ	τ
1	Sea Snake	Electric Ray	0.8	0.25
2	Bus	Vending Machine	0.7	0.125
3	Elephant	Magpie	0.8	0.25

Table 7: AUROC scores for the synthetic experiments.

	Synthetic 1				Synthetic 2				Synthetic 3			
	LH	LH-2S	LR	WN	LH	LH-2S	LR	WN	LH	LH-2S	LR	WN
AR-DGM	0.65	0.57	0.68	0.59	0.61	0.58	0.48	0.76	0.57	0.57	0.56	0.64
Linear	0.62	0.64	-	0.61	0.64	0.57	-	0.76	0.60	0.67	-	0.62
VAE+Linear, 64	0.62	0.57	0.62	0.66	0.81	0.70	0.62	0.74	0.85	0.78	0.93	0.76
VAE+Linear, 512	0.65	0.60	0.69	0.65	0.77	0.65	0.65	0.71	0.70	0.64	0.83	0.71

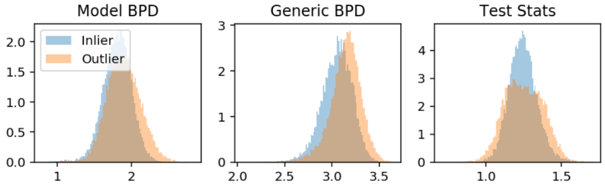


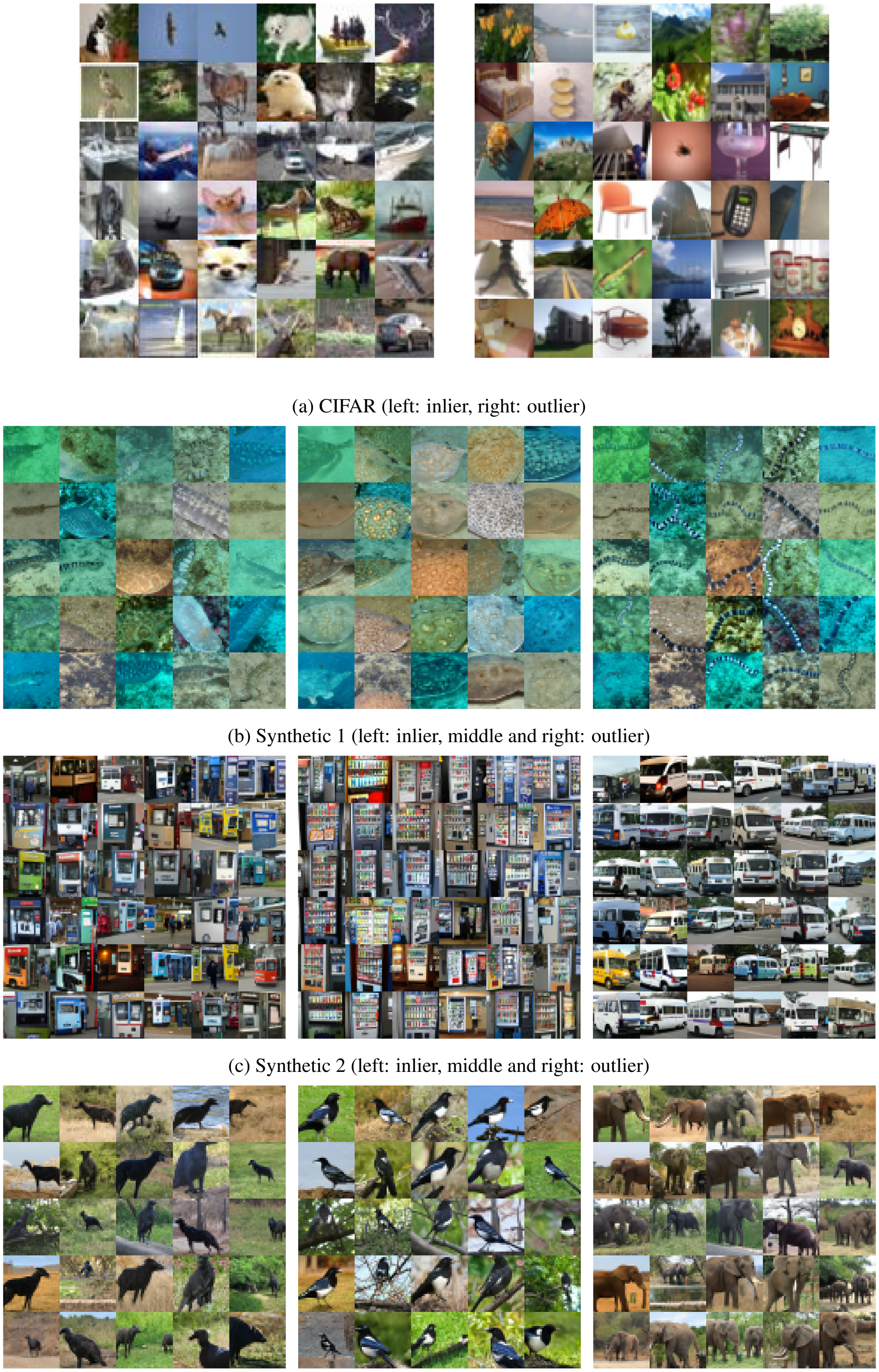
Figure 5: Distribution of various statistics related to the LR test using AR-DGM on the second synthetic experiment.

C.3 Details for Section 4

The ground truth VAE has the same architecture as in Section 3.1, but with a continuous normal likelihood. We use $n_z = 64$. The VAE (log) likelihood is lower bounded by IWAE_{200}. For EBM and PixelCNN++, we use the authors' hyperparameters and training setup for the unconditional CIFAR-10 experiments. After training, we verified that the distributions of energy values of training and held-out samples have small differences, so the models do not appear to overfit.

As the OOD test results in [5, 6] are obtained with conditional models, we perform the single-sided likelihood test with the unconditional model (trained on the real CIFAR-10 dataset) to check if its behavior on the SVHN dataset is similar to the conditional model. The AUROC value from the single-side likelihood test is 0.529, meaning that the EBM assigns similar or lower likelihood to

[11] `https://github.com/huggingface/pytorch-pretrained-BigGAN`



(a) CIFAR (left: inlier, right: outlier)

(b) Synthetic 1 (left: inlier, middle and right: outlier)

(c) Synthetic 2 (left: inlier, middle and right: outlier)

(d) Synthetic 3 (left: inlier, middle and right: outlier)

Figure 6: More sample images for the setups in Section 3.2.

SVHN compared with the inliers. This is still significantly different from the results using other generative models, justifying our use of an unconditional model.