THE SIGNAL IN THE NOISE: OOD DETECTION THROUGH GOODNESS-OF-FIT TESTING IN FACTORISED LATENT SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep generative models offer a natural foundation for out-of-distribution (OOD) detection, yet prior work has shown that their assigned likelihoods are notoriously unreliable indicators for in- vs out-of-distribution data. In this paper, we address this problem by leveraging the diffeomorphic and mass-preserving properties of continuous normalising flows. Our analysis shows that OOD samples are mapped to noise samples that are highly atypical under the noise prior in ways not captured by the likelihood. Based on this observation, we propose a new method—Signal in the Noise (SITN)—for OOD detection on the single-sample level. SITN requires no access to OOD data, incurs minimal computational overhead, and provides strict control of false positive rates. Comprehensive evaluations through standard benchmarks and synthetic perturbations highlight the method's effectiveness and the absence of the complexity bias inherent to likelihood-based methods.

[removed for review]

1 INTRODUCTION

The deployment of machine learning models in high-risk domains like healthcare is fundamentally limited by their reliability. Models are prone to failing catastrophically—and with high confidence—when confronted with out-of-distribution (OOD) data that deviates substantially from the training distribution (Nguyen et al., 2015). To mitigate this risk, OOD detection can enable systems to abstain from decision-making or defer to a human expert when a model's prediction is deemed unreliable.

Deep generative models provide a natural foundation for OOD detection by explicitly modelling the data distribution. In particular, Continuous Normalising Flows (CNFs) (Chen et al., 2018), often trained via flow matching methods (Lipman et al., 2023), are a logical choice for this task because they enable the tractable computation of exact likelihoods. However, prior work has shown that these likelihoods are unreliable indicators for in- vs out-of-distribution data in high dimensions (Nalisnick et al., 2019b). For instance, widely replicated experiments have demonstrated that generative models trained on CIFAR-10 consistently assign higher likelihoods to samples from MNIST or SVHN (Nalisnick et al., 2019b; Serrà et al., 2020; Schirrmeister et al., 2020; Kirichenko et al., 2020).

Previous work has investigated the theoretical underpinnings of this limitation of likelihood-based OOD detection and proposed various corrections to mitigate it. Notable approaches include typicality tests (essentially a two-sided thresholding of the likelihood) (Nalisnick et al., 2019a), complexity penalties (Serrà et al., 2020), likelihood ratios (Schirrmeister et al., 2020; Ren et al., 2019; Xiao et al., 2020), and fine-grained analyses of likelihood components (Nalisnick et al., 2019b; Morningstar et al., 2021). While these methods succeed in partially alleviating the pronounced failure mode observed between datasets like CIFAR-10 and SVHN, our analysis of representative approaches (Typicality (Nalisnick et al., 2019a), DoSE (Morningstar et al., 2021)) reveals that they still exhibit undesirable biases linked to inherent limitations of the likelihood. Furthermore, these approaches frequently impose practical constraints, as they may require domain-specific compression algorithms, depend on (typically unavailable) out-of-distribution data to construct background models, or incur significant computational overhead.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We depart from these likelihood-centric approaches and propose a new framework. Leveraging the mass-conserving, diffeomorphic mappings of CNFs, we first show that OOD samples map to noise latents with atypical empirical properties under the noise prior. Notably, this holds independently of the divergence of the vector field, which does affect likelihoods. We then quantify this atypicality at the single-sample level by exploiting the completely factorised nature of the Gaussian noise prior. Specifically, this factorisation implies that the elements of a single high-dimensional latent vector must behave as i.i.d. samples from a univariate standard normal distribution. This allows us to test an individual sample based on two defining properties: the marginal normality of its elements and their dimensional independence. We present a new method termed Signal in the Noise (SITN) as a specific instantiation of this framework, and validate it through comprehensive evaluations on standard benchmarks and synthetic perturbations.

In summary, our main contributions are as follows:

- We introduce a new, theoretically grounded framework that casts OOD detection as single-sample goodness-of-fit testing against the factorised Gaussian noise prior of CNFs.
- We propose a specific method (SITN) for OOD detection. SITN incurs minimal computational overhead, requires no access to OOD data, and provides strict Type I error control. We demonstrate the method's effectiveness through standard benchmarks and detailed ablations.
- We present visualisations and analyses that shed light on the differential behaviour of SITN and various baseline methods and highlight promising directions for future work.

2 Related Work

Limitations of the Likelihood for OOD Detection Several works have identified that the likelihood assigned by deep generative models is not a reliable indicator for OOD detection (Nalisnick et al., 2019b; Serrà et al., 2020; Schirrmeister et al., 2020; Kirichenko et al., 2020; Havtorn et al., 2021; Kamkari et al., 2024). Nalisnick et al. (2019a) highlighted an inherent limitation of the likelihood: regions of high probability *density* can lie outside the typical set of the distribution that contains almost all probability *mass*. This is well illustrated by the concentration of measure behaviour of a high-dimensional ($d \gg 1$) standard Gaussian distribution, where the maximum density is attained at the mode, but virtually all probability mass is concentrated in the annulus with radius $\sqrt{d}$ (Blum et al., 2020). Zhang et al. (2021) challenged the typical set hypothesis as an explanation for the observed failures, attributing them instead to model estimation error. Other works have argued that likelihoods obtained through generative models are ineffective for OOD detection because they are dominated by low-level features that are similar across datasets, rather than semantic content (Schirrmeister et al., 2020; Kirichenko et al., 2020; Havtorn et al., 2021).

Likelihood Corrections To address these limitations, several works have attempted to correct or augment the raw likelihoods. Based on their concentration of measure observations, Nalisnick et al. (2019a) introduced typicality tests, arguing that in-distribution (ID) samples should not simply lie in a high density region, but rather reside in the typical set. Extending this geometric intuition, Kamkari et al. (2024) proposed pairing likelihoods with estimates of Local Intrinsic Dimension (LID) to flag OOD samples that reside on low-dimensional manifolds—regions that contain negligible probability mass despite potentially high likelihoods. However, estimating LID for normalising flows relies on explicitly constructing the full Jacobian matrix. While this is tractable for heavily constrained architectures (e.g., Glow), it is computationally prohibitive for modern, highly expressive flow matching models. Serrà et al. (2020) observed a strong correlation between likelihood and sample complexity (e.g., PNG compression size) and proposed a complexity correction that can be interpreted as a likelihood ratio test. This requires a specific complexity estimate, which can be arbitrary and severely limiting in non-standard data domains where suitable compression algorithms are unavailable. Other works have also proposed likelihood ratio tests (Ren et al., 2019; Schirrmeister et al., 2020). A downside of these approaches is the requirement of a separate background model trained on OOD data, which is not always straightforward to construct and adds computational overhead. A special case that bypasses the need for OOD data is the Likelihood Regret method proposed by Xiao et al. (2020). Their method fine-tunes the generative model on the individual test sample (through additional optimisation steps during inference) and then computes the log ratio between the likelihoods obtained from the fine-tuned and the baseline model. As a downside, this

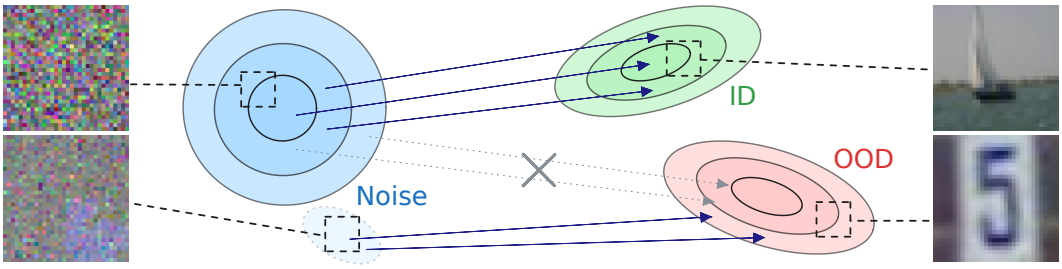


Figure 1: Illustration of OOD detection in the noise space. CNF models are trained to map noise samples from a standard normal distribution to the target distribution (ID). Since the resulting flow is a probability mass-conserving diffeomorphism, OOD samples with low probability mass under the target distribution will not correspond to typical Gaussian noise samples in noise space (crossed dotted lines), such that OOD detection can be cast as a goodness of fit test in the noise space.

approach complicates the inference process and still incurs computational overhead. Alternatively, Choi et al. (2018) proposed using the Watanabe-Akaike Information Criterion (WAIC) to penalize epistemic uncertainty by calculating the variance of likelihood evaluations across an ensemble of generative models, albeit at the cost of training multiple networks. Ding et al. (2025) computed likelihoods on embeddings of pre-trained encoders rather than raw samples. A practical limitation for this approach is the requirement of a good encoder.

Beyond Likelihoods More recently, methods have moved beyond purely data-space likelihoods to probe the latent representations and generative mappings of generative models. Most similar in spirit to our method is the work by Jiang et al. (2022), which also explored latent-space goodness-of-fit testing. Their work is however focused on group-OOD detection and not directly applicable to single samples. Density of States Estimation (DoSE) measures the empirical density of multiple statistics, including the log-probability of the latent representation and the log-determinant of the Jacobian of the mapping (Morningstar et al., 2021). Heng et al. (2024) proposed DiffPath, which measures the rate-of-change and curvature of diffusion paths. Interestingly, their approach relies on the ability of diffusion models to map even OOD samples approximately to the noise prior. Our method leverages the exact opposite: using deterministic and diffeomorphic CNFs instead of stochastic diffusion models, we exploit that OOD and ID samples are not mapped to the same regions in the noise space.

Discriminative OOD Detection Methods A lot of work in the OOD detection literature has focused on discriminative methods, where a classifier is trained via supervised learning on ID data and OOD detection methods directly operate on that model's softmax or logit outputs (Hendrycks & Gimpel, 2018; Liu et al., 2020; Liang et al., 2025) or features (Lee et al., 2018; Sun et al., 2022). Such methods are particularly well suited to semantic-shift benchmarks, where the ID/OOD distinction is closely aligned with the supervised classification task (Guille-Escuret et al., 2024). Our focus, in contrast, is on OOD detection as a broader safety and robustness mechanism, including distribution shifts that need not align with class semantics. Li et al. (2025b;a) have furthermore outlined fundamental theoretical limitations of discriminative approaches. While our work is therefore focused on generative methods, we include comparisons to discriminative methods in Appendix J.

3 Methods

3.1 Out-of-distribution Detection in Noise Space

We formalise the problem of OOD detection as classifying a novel observation, X, as being drawn from one of two possible distributions: P_{ID} (In Distribution) or P_{OOD} (Out Of Distribution). During training time, it is assumed one only has access to data drawn from P_{ID}.

We begin by considering a CNF trained via, e.g., Flow Matching (Lipman et al., 2023). A perfectly trained CNF defines a vector field whose corresponding Ordinary Differential Equation (ODE) admits a flow $\phi : \mathbb{R}^D \to \mathbb{R}^D$ that maps samples from a prior P_0 to samples from P_{ID}; i.e., the push-forward

measure of P_0 by ϕ is P_{ID}. P_0 is usually taken to be $\mathcal{N}(0, I)$—we will assume this choice going forward. The reversibility of ODEs means one also has access to ϕ^{-1}, and is able to map real data points back to samples from P_0. Consequently, mapping ID data points back to the noise space will result in samples from $\mathcal{N}(0, I)$. Conversely, OOD samples from P_{OOD} will yield samples that are not distributed according to the standard multivariate Gaussian. That is, for $Z = \phi^{-1}(X)$ we have that $Z \sim P_0$ if and only if $X \sim P_{\text{ID}}$. We quantify the extent to which Z conforms with the prior through test statistics, reducing OOD detection to a statistical goodness-of-fit test in the noise space.

3.2 Goodness-of-Fit Testing Against the Prior

To determine the goodness-of-fit of some Z against the theoretical standard multivariate Gaussian, we probe its two defining properties: marginal normality (each element Z_i follows a standard univariate normal distribution) and dimensional independence (elements Z_i and Z_j are mutually independent for all $i \neq j$). Formally, we treat this as a goodness-of-fit test with a composite null hypothesis, H_0, and decompose it into a conjunction of two constituent hypotheses H_{MN} and H_{DI}, for marginal normality and dimensional independence, respectively: $H_0 \Rightarrow H_{\text{MN}} \wedge H_{\text{DI}}$. Under this decomposition, each constituent hypothesis poses a necessary condition for H_0. That is, if we reject one of the constituent hypotheses, then we should also reject the null hypothesis.

Crucially, we are able to test these hypotheses entirely at the single-sample level by exploiting the fact that the multivariate standard Gaussian prior factorises completely. Specifically, a single observation, $Z \sim \mathcal{N}(0, I)$, can be equivalently viewed as a collection of D i.i.d. samples from a univariate standard Gaussian, $\mathcal{N}(0, 1)$. Because the dimensionality D of many standard data modalities (such as images) is large, this perspective allows us to compute robust empirical statistics from an individual high-dimensional sample, rather than requiring a batch of multiple samples.

In the following, we detail the specific statistics and combination mechanism that form our proposed SITN method as an instantiation of this framework.

Marginal Normality To evaluate whether the individual elements of Z conform to the expected univariate normal distribution, we employ the Anderson-Darling (AD) statistic. By treating the D dimensions of Z as an empirical sample, the AD statistic quantifies the discrepancy between the empirical cumulative distribution function (CDF) of the sample and the theoretical CDF of the standard normal distribution, Φ. Sorting the elements of Z in ascending order such that $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(D)}$, the statistic is computed as

$$S_{\text{AD}} = -D - \sum_{i=1}^{D} \frac{2i-1}{D} \left[\ln \Phi(Z_{(i)}) + \ln(1 - \Phi(Z_{(D+1-i)})) \right]. \tag{1}$$

A high S_{AD} indicates that ϕ^{-1} mapped the data sample to a noise sample with anomalous mean, incorrect variance, or distorted distributional shape violating the prior's marginal normality expectation.

Dimensional Independence Acknowledging that no single statistical test can comprehensively detect all forms of dependence (Hutter, 2022), we specifically test for autocorrelation. Independence between random variables implies there is no autocorrelation, so this still represents a necessary condition for the null hypothesis. We compute the coefficient of variation (CV) of the empirical power spectrum (PSD) of Z, which is defined as the ratio of the standard deviation of the power spectrum to its arithmetic mean,

$$S_{\text{CV}} = \frac{\sigma_{\text{PSD}}}{\mu_{\text{PSD}}} = \frac{\sqrt{\frac{1}{D}\sum_{k=1}^{D}(\text{PSD}(k) - \mu_{\text{PSD}})^2}}{\mu_{\text{PSD}}}, \tag{2}$$

where $\mu_{\text{PSD}} = \frac{1}{D}\sum_{k=1}^{D} \text{PSD}(k)$. For standard Gaussian white noise, the empirical power spectrum follows an exponential distribution ($\sigma = \mu$), yielding a theoretical S_{CV} of exactly 1. Conversely, structural correlations between the components of Z manifest as dominant frequencies, inflating the spectral variance and driving S_{CV} significantly above 1. This statistic serves as a global test for autocorrelation: by the Wiener–Khinchin theorem, the power spectrum and autocorrelation function form a Fourier transform pair, meaning that an inflated S_{CV} mathematically implies non-zero autocorrelation. We prefer this spectral approach over time-domain alternatives like the Ljung-Box test, as it captures global dependencies without the need to select a max-lag hyperparameter.

Combining Statistics To produce a single scalar OOD score, we combine the normality and independence statistics using a max-quantile strategy. The empirical cumulative distribution functions (CDFs) for both statistics, denoted as $\hat{F}_{\text{AD}}$ and $\hat{F}_{\text{CV}}$, are estimated using an ID calibration set. The combined score is then defined as the maximum of the quantile-transformed statistics

$$S_{\text{SITN}} = \max\left(\hat{F}_{\text{AD}}(S_{\text{AD}}), \hat{F}_{\text{CV}}(S_{\text{CV}})\right). \tag{3}$$

The use of quantile transforms ensures that the statistics are mapped into a space where they have the same range and marginal distribution (i.e., uniform on the $[0, 1]$ interval), thus allowing them to be meaningfully compared. By taking the maximum, we identify a sample as OOD if it appears as an outlier in either its marginal distribution or its structural independence, achieving our goal of rejecting H_0 if any of the constituent hypotheses should be rejected.

3.3 Theoretical Properties

Provable Control of False Positive Rate Determining a suitable threshold for OOD scores is difficult in practice. Ideally, one would have access to a set of known OOD and ID data, and use this to choose a threshold that trades off between the false positive and false negative rates of the OOD detector. While this is not possible without access to OOD data, ID data is sufficient to calibrate the detector, allowing one to select a tolerable false positive rate. In Appendix K, we describe how S_{SITN} can be calibrated and prove that this provides false positive rate (i.e., Type I error) control.

Robustness to CNF Approximation Error CNFs typically do not learn an exact model of the target P_{ID}, but a reasonably good approximation. Our framework remains applicable when the CNF is not a perfect representation of P_{ID}, but one must interpret the constituent hypotheses, H_{MN} and H_{DI}, differently. They should be understood as statements about the maximum acceptable deviation from normality and independence, respectively. One should expect poorer quality CNFs to result in increased false negative (i.e., Type II error) rates, while the false positive rate is not impacted.

Computational Complexity For a noise sample Z of dimensionality D, the computational cost of S_{AD} is dominated by the sorting of the elements, resulting in a complexity of $\mathcal{O}(D \log D)$. Similarly, the computation of S_{CV} relies on obtaining the empirical power spectrum, which is dominated by the Fast Fourier Transform (FFT) and also scales as $\mathcal{O}(D \log D)$. The combined overhead of computing these statistics is negligible compared to the computational cost of the neural vector field evaluations required for the inverse flow. Compared to likelihood-based methods, SITN does not require the computationally expensive evaluations of the vector field's divergence.

4 Experimental Design

For each dataset described in Section 4.1, we train a single unconditional CNF to evaluate the OOD detection performance of the baseline methods described in Section 4.2 and SITN. For cross-dataset experiments, we use the test split of the training dataset as ID data and the test split of a different dataset as OOD data. For the CIFAR-10 model, we furthermore evaluate OOD detection performance using the synthetic perturbations from the CIFAR-10-C benchmark (Hendrycks & Dietterich, 2019).

Importantly, for comparisons across datasets with different original resolutions (e.g. CIFAR-10 and CelebA), we downsample the higher resolution dataset. This is to prevent introducing blurriness that leads to trivial OOD detection as reported by Heng et al. (2024). In their experiments, the OOD detection AUROC on CIFAR-10 vs CelebA of their model dropped from 0.999 (when the lower resolution dataset was upsampled) to 0.502 (when the higher resolution dataset was downsampled).

We use the UNet architecture following Ho et al. (2020), as implemented in Dhariwal & Nichol (2021). Following the flow matching objective of Lipman et al. (2023), we construct a linear interpolant between a Gaussian noise sample $Z \sim \mathcal{N}(0, I)$ and a data sample $X \sim P_{\text{ID}}$: $X_t = (1-t)Z + tX, \quad t \sim \mathcal{U}[0, 1]$, and minimise the mean-squared error between the network output and the target vector field $X - Z$. Full architectural and training details appear in Appendix L.

The adaptive Dormand-Prince ODE solver (dopri5) (Dormand & Prince, 1980) is used to compute ϕ and ϕ^{-1}. OOD detection scores are then derived using the baseline methods and SITN.

4.1 DATASETS

CIFAR-10 CIFAR-10 (Krizhevsky et al., 2009) comprises 60,000 32×32 images across 10 categories. We use 10% of the training set for validation, yielding 45,000 training, 5,000 validation, and 10,000 test images.

SVHN The Street View House Numbers (SVHN) dataset (Netzer et al., 2011) contains 32×32 images from Google Street View and is available for non-commercial use. We use the official train (73,257 images) and test (26,032 images) splits, reserving 10% of the training set for validation.

CelebA CelebA (Liu et al., 2015) comprises over 200,000 celebrity images with attribute annotations. It is available for non-commercial research purposes. We adopt the predefined train, validation, and test splits, downsampling all images to 32×32 pixels for consistency with the other datasets.

4.2 BASELINE METHODS

Log-likelihood The log-likelihood is evaluated through the instantaneous change of variables formula as the sum of the latent's log-probability and the integral of the vector field's divergence.

Typicality Typicality was proposed by Nalisnick et al. (2019a) and is defined as the absolute difference between the sample's negative log-likelihood and the resubstitution estimator of the entropy (i.e., the average negative log-likelihood over the training data; experiments using validation data instead are in Appendix I). We use the single-sample version as in (Morningstar et al., 2021).

DoSE Density of States Estimation (DoSE) (Morningstar et al., 2021) measures the empirical density of multiple statistics using ID training data (experiments using validation data are in Appendix I). Specifically, we implement the DoSE_{KDE} variant for flow-based models, which uses three statistics: the log-likelihood, the latent's log-probability, and the log-determinant of the Jacobian. Because we operate in the continuous-time regime of CNFs, the log-determinant is replaced with the vector field's divergence integral. A Kernel Density Estimator (KDE) is fit to each statistic and the OOD score is computed by summing the individual KDEs' log-probabilities. As in the original paper, we use SciPy's Gaussian KDE with Scott's rule for bandwidth selection. For efficiency, KDEs are fit on a subset of 10,000 samples; we empirically verified that this has minimal impact on the results.

5 RESULTS

5.1 OOD DETECTION PERFORMANCE: CROSS-DATASET EXPERIMENTS

Table 1 reports the Area Under the Receiver Operating Characteristic (AUROC) curve for the cross-dataset experiments. Models were trained on one dataset and evaluated for their OOD detection performance using the training dataset's test split as ID and another dataset's test split as OOD data. Additional comparisons to complexity correction, WAIC, and DiffPath can be found in Appendix E.

Table 1: Cross-dataset OOD detection performance in terms of AUROC. Values in brackets denote bootstrapped 95% CIs and results for the best method are in **bold**. SITN consistently outperforms baselines, demonstrating that noise samples contain useful information for OOD detection.

Train (ID)	Test (OOD)	Log-likelihood	Typicality	DoSE	SITN (Ours)
CIFAR-10	SVHN	0.105 [.100, .110]	0.790 [.785, .796]	0.802 [.796, .808]	**0.946** [.943, .949]
	CelebA	0.496 [.488, .503]	0.421 [.414, .428]	0.432 [.425, .439]	**0.671** [.665, .678]
SVHN	CIFAR-10	0.980 [.978, .981]	0.954 [.951, .957]	0.984 [.982, .985]	**0.986** [.984, .987]
	CelebA	**0.999** [.999, .999]	0.995 [.994, .995]	0.997 [.997, .998]	0.996 [.995, .996]
CelebA	CIFAR-10	0.728 [.721, .734]	0.688 [.681, .694]	0.717 [.710, .723]	**0.910** [.907, .914]
	SVHN	0.179 [.175, .183]	0.656 [.651, .661]	0.776 [.771, .780]	**0.986** [.985, .987]

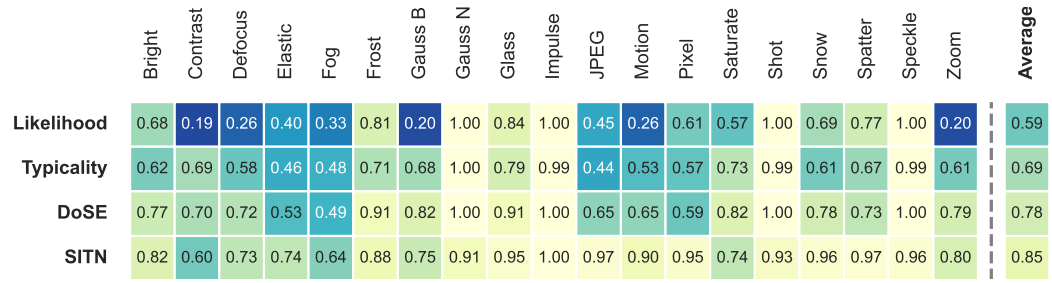


Figure 2: OOD detection performance across the CIFAR-10-C corruptions in terms of AUROC. For each corruption, samples with all severity levels were pooled. SITN performs best overall with the highest average performance across corruptions and a minimum AUROC of 0.6 on the contrast perturbation, whereas other methods drop to (or below) chance level in multiple conditions.

In line with prior work, the log-likelihood results in high AUROC values when trained on the simpler SVHN dataset and evaluated against CIFAR-10 or CelebA, but fails consistently when the setup is reversed. Typicality improves on the raw log-likelihoods in setups where their distributions across the two datasets are well separated, with consistently higher likelihoods across the OOD dataset (e.g., CIFAR-10→SVHN). However, it can fail when the likelihood distributions are strongly overlapping, particularly if the OOD dataset's likelihood distribution is more concentrated (e.g., CIFAR-10→CelebA; see Appendix C for likelihood distribution plots). DoSE shows similar behaviour to Typicality, albeit with a small improvement. Our method, SITN, consistently outperforms these baselines, even in conditions where the baselines exhibit relatively low performance.

Overall, these results provide empirical evidence that noise samples contain useful information for OOD detection and position SITN as an effective method to leverage that information.

5.2 OOD Detection Performance: Synthetic Perturbations

Figure 2 shows OOD detection performance for the CIFAR-10-trained model across the synthetic perturbations from CIFAR-10-C. SITN shows the most consistent result with a macro average across conditions of 0.85 compared to 0.59 for the Log-likelihood, 0.69 for Typicality, and 0.78 for DoSE.

SITN achieves its lowest performance for the contrast and fog corruptions. Both Typicality and DoSE perform significantly better on the former, but even worse on the latter. Conversely, SITN performs particularly well for JPEG, snow, or spatter where other methods struggle.

Overall, these results provide further evidence for the effectiveness of goodness-of-fit testing against the factorised noise prior, showing that it is sensitive to a broad range of perturbations.

5.3 Metric Extremes

Figure 3 shows the images with highest and lowest OOD scores per method in the CIFAR-10→SVHN setup (trained on CIFAR-10, evaluated on the CIFAR-10 test split for ID and on the SVHN test split for OOD). For SITN, 18 samples were tied with a score of 1, from which 5 were selected, approximately preserving the ID/OOD ratio in the 18-way tie. Appendix D shows the same visualisation for CIFAR-10→CelebA, and Appendix O shows the 150 images with highest OOD scores per method.

The complexity-bias of the likelihood reported by Serrà et al. (2020) is very conspicuous, as the highest likelihood scores are assigned to simple images with constant backgrounds. Typicality flags these samples as the most OOD, although they are technically considered in-distribution in this setting. This points to an inherent limitation of Typicality for datasets with a broad distribution of likelihoods like CIFAR-10, which contains both the highest and lowest likelihood samples in this case. DoSE inherits some of the same failure cases and additionally flags some of the lowest likelihood samples as OOD. This makes sense given that one of the components considered in DoSE is a KDE of the likelihood, such that it will be sensitive to extremely high and low likelihoods. These results show that while Typicality and DoSE mitigate some of the catastrophic failures of using the raw

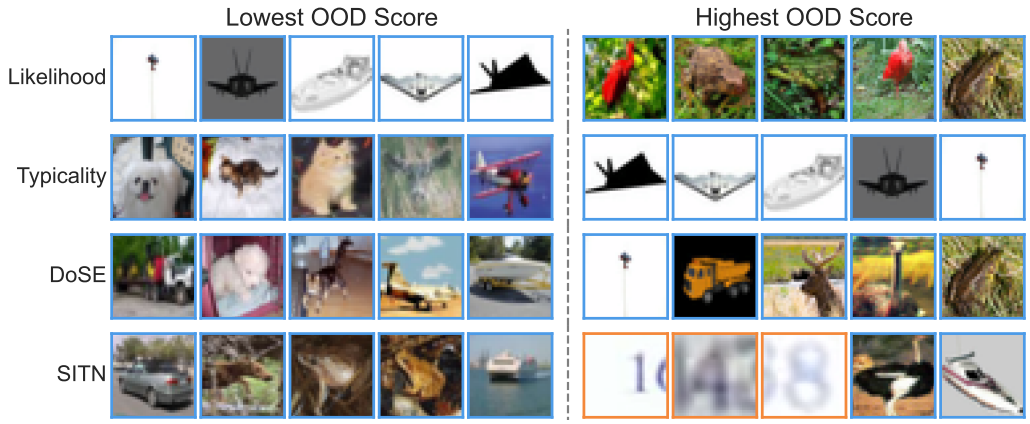


Figure 3: Images with the highest and lowest OOD scores for each method in the CIFAR-10→SVHN setup, i.e. trained on CIFAR-10 and evaluated on the test split of CIFAR-10 as ID and the test split of SVHN as OOD. ID images from CIFAR-10 are marked with a blue border and OOD images from SVHN with an orange border. All baseline methods falsely flag in-distribution samples as most OOD, and both Typicality and DoSE inherit failure cases from the complexity-biased likelihood scores.

likelihood for OOD detection, they are still affected by its inherent complexity bias. We illustrate this point further in Appendix B and show that SITN does not exhibit this bias.

Overall, SITN is robust to the failure modes of likelihood-based methods and the only method with actual OOD images among the most OOD-scored samples. We discuss its failure cases in Section 5.5.

5.4 Ablations

Appendix A presents an ablation study, showing the performance of the individual metric components in SITN (AD for normality testing and CV for independence testing) across the corruptions from CIFAR-10-C. The two statistics show complementary value across the different perturbations. For instance, normality testing was more sensitive to brightness perturbations, whereas motion artefacts were picked up much better through independence testing.

In Table 2, we further show an ablation for the metric aggregation approach. We compare the max-quantile approach described in Section 3.2 with the KDE approach used in DoSE. Furthermore, as both Typicality and DoSE use the training data to fit the entropy estimate and KDE, respectively, we also evaluate SITN with the max-quantile or KDE approach fitted on the training data.

Table 2: Ablation study of SITN metric aggregation. The table shows AUROC performance with 95% bootstrapped confidence intervals for the CIFAR-10-trained model. Max-quantile consistently outperforms the KDE approach, while the difference between fitting it on training vs validation data is negligible. The first row shows the default SITN configuration.

Metric aggregation	**Fit split**	**SVHN**	**CelebA**
Max Quantile	Val	0.946 [.943, .949]	0.671 [.665, .678]
Max Quantile	Train	0.948 [.946, .951]	0.672 [.666, .679]
KDE	Val	0.915 [.911, .918]	0.615 [.608, .621]
KDE	Train	0.913 [.909, .917]	0.617 [.610, .624]

The max-quantile strategy outperforms the KDE approach, while the difference between fitting it on training vs validation data is negligible. We prefer validation data to preclude overfitting and enable proper calibration as detailed in Appendix K. We note that our max-quantile strategy cannot be trivially applied to combine metrics in DoSE, as it requires an inherent OOD direction for all metrics.

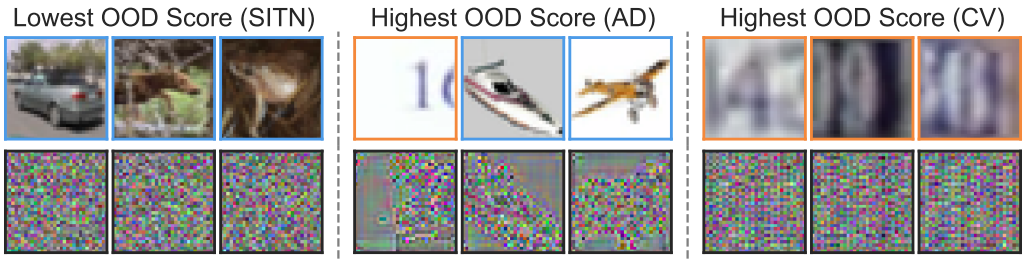


Figure 4: Visualisation of images along with their corresponding noise samples. Images from the CIFAR-10→SVHN setup are shown with ID CIFAR-10 images marked by a blue border and OOD SVHN images by an orange border. **Left:** samples with lowest SITN OOD score, where the noise is in line with standard Gaussian noise in terms of both the Anderson Darling (AD) statistic and the coefficient of variation (CV) of the power spectrum. **Middle:** samples with highest OOD score driven by the AD statistic. **Right:** samples with highest OOD score driven by the CV statistic.

5.5 Noise Visualisations

Figure 4 shows images from the CIFAR-10→SVHN setup, alongside corresponding noise samples obtained via backwards integration along the probability flow ODE. The left panel displays samples with the lowest SITN scores—i.e., samples where both normality (assessed via AD) and independence (assessed via CV) are consistent with the standard Gaussian noise prior. The middle and right panels show samples with high SITN scores driven by violations of normality and independence, respectively.

As expected, ID samples with low OOD scores map seamlessly to typical Gaussian noise. Furthermore, OOD samples flagged by the CV component of SITN exhibit strong checkerboard-like patterns, clearly violating dimensional independence. Interestingly, samples flagged by the AD component contain large patches of constant background, a structure that is visibly retained in the noise space. Because this phenomenon occurs not only in OOD samples but also in CIFAR-10, it reveals a failure mode. While AD accurately identifies a deviation from normality, the flow matching model struggles to map constant background to typical Gaussian noise even for ID samples. It is worth emphasising that this behaviour reveals an edge case while the method performs well overall as evidenced by the strong AUROC results in Table 1 and the visualisation of a larger set of samples in Appendix O.

6 Discussion and Conclusions

In this work, we introduced a framework that casts OOD detection as a goodness-of-fit test in the noise space of CNFs. As a concrete instantiation of this framework, we proposed SITN, which probes noise samples against a standard Gaussian prior. SITN exploits the complete factorisation of this prior to operate on individual samples and test them for deviations in marginal normality and dimensional independence. Crucially, SITN offers significant practical and statistical advantages: it requires no access to OOD data, adds minimal computational overhead—requiring neither additional training nor likelihood evaluations—and enables the formal derivation of p-values for strict Type I error control.

We rigorously validated the effectiveness of SITN across standard cross-dataset benchmarks and synthetic perturbations, where it consistently outperformed baselines. Moreover, detailed visualisations and empirical analyses revealed that the likelihood-based baselines Typicality and DoSE retain a strong complexity bias inherent to the raw likelihood. SITN, by contrast, does not exhibit this bias.

Our findings also highlight avenues for future research. While SITN yields highly competitive overall performance, we observed a specific edge case where the model failed to map large, constant background patches to Gaussian noise, causing the AD component to flag the retained spatial structure as atypical. Future work will naturally focus on mitigating this artefact. More broadly, we see great promise in exploring statistics other than AD and CV—such as those that identify other forms of dependence—and in studying and optimising the interplay between the OOD detection method and the underlying generative model. Together, these directions offer a clear path to refine and better understand the design space for OOD detection through goodness-of-fit testing in the noise space.

AI USE STATEMENT

In this work, generative AI tools were used to assist in the implementation of methods and to support data analysis by assisting in the implementation of visualisation and plotting code. Generative AI tools were not used to generate synthetic data sets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, design experiments, clean and reformat datasets, formulate mathematical claims, assist with proofs, or interpret results. Translating text is not applicable to this work. All AI-assisted work was extensively reviewed; specifically, AI-assisted code was manually verified, refined, and tested for correctness by the authors. The authors take full responsibility for the final content of this work, including text, claims, or artefacts produced with the aid of generative AI.

REFERENCES

Avrim Blum, John E. Hopcroft, and Ravindran Kannan. *Foundations of data science*. 2020. ISBN 978-1-108-48506-7 978-1-108-75552-8. doi: 10.1017/9781108755528.

Rania Briq, Michael Kamp, Ohad Fried, Sarel Cohen, and Stefan Kesselheim. The Amazing Stability of Flow Matching, April 2026. URL http://arxiv.org/abs/2604.16079. arXiv:2604.16079 [cs].

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proc. of NeurIPS*, pp. 6572–6583, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Hyunsun Choi, Eric Jang, and Alexander A. Alemi. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. *ArXiv preprint*, abs/1810.01392, 2018. URL https://arxiv.org/abs/1810.01392.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Proc. of NeurIPS*, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Yifan Ding, Arturas Aleksandraus, Amirhossein Ahmadian, Jonas Unger, Fredrik Lindsten, and Gabriel Eilertsen. Revisiting Likelihood-Based Out-of-Distribution Detection by Modeling Representations, April 2025. URL https://arxiv.org/abs/2504.07793v3.

J. R. Dormand and P. J. Prince. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. ISSN 0377-0427. doi: 10.1016/0771-050X(80)90013-3. URL https://www.sciencedirect.com/science/article/pii/0771050X80900133.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

Charles Guille-Escuret, Pierre-André Noël, Ioannis Mitliagkas, David Vazquez, and Joao Monteiro. Expecting The Unexpected: Towards Broad Out-Of-Distribution Detection. *Advances in Neural Information Processing Systems*, 37:130953–130976, December 2024. doi: 10.52202/079017-4162. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ec80d18205e39d42a27192d5f3ddd688-Abstract-Datasets_and_Benchmarks_Track.html.

Jakob Drachmann Havtorn, Jes Frellsen, Søren Hauberg, and Lars Maaløe. Hierarchical vaes know what they don't know. In *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4117–4128, 2021. URL http://proceedings.mlr.press/v139/havtorn21a.html.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. of ICLR*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, October 2018. URL http://arxiv.org/abs/1610.02136. arXiv:1610.02136 [cs.NE].

Alvin Heng, Alexandre H. Thiery, and Harold Soh. Out-of-distribution detection with a single unconditional diffusion model. In *Proc. of NeurIPS*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/4dc37a7bc61057252ce043fa3b83aac2-Abstract-Conference.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Marcus Hutter. Testing Independence of Exchangeable Random Variables, October 2022. URL http://arxiv.org/abs/2210.12392. arXiv:2210.12392 [math].

Dihong Jiang, Sun Sun, and Yaoliang Yu. Revisiting flow generative models for out-of-distribution detection. In *Proc. of ICLR*, 2022. URL https://openreview.net/forum?id=6y2KBh-0Fd9.

Hamidreza Kamkari, Brendan Leigh Ross, Jesse C. Cresswell, Anthony L. Caterini, Rahul G. Krishnan, and Gabriel Loaiza-Ganem. A geometric explanation of the likelihood OOD detection paradox. In *Proc. of ICML*, 2024. URL https://openreview.net/forum?id=EVMzCKLpdD.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Proc. of NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ecb9fe2fbb99c31f567e9823e884dbec-Abstract.html.

Alex Krizhevsky, Geoffrey Hinton, and others. Learning multiple layers of features from tiny images. 2009.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Out-of-Distribution Detection Methods Answer the Wrong Questions, July 2025a. URL https://arxiv.org/abs/2507.01831v1.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised Classifiers Answer the Wrong Questions for OOD Detection. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 81689–81713. PMLR, October 2025b. URL https://proceedings.mlr.press/v267/li25ec.html.

Jiachen Liang, RuiBing Hou, Minyang Hu, Hong Chang, Shiguang Shan, and Xilin Chen. Revisiting Logit Distributions for Reliable Out-of-Distribution Detection. October 2025. URL https://openreview.net/forum?id=FLdLPUqnsP.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proc. of ICLR*, 2023. URL https://openreview.net/pdf?id=PqvMRDCJT9t.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow Matching Guide and Code. *ArXiv preprint*, abs/2412.06264, 2024. URL https://arxiv.org/abs/2412.06264.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. of ICCV*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425. URL https://doi.org/10.1109/ICCV.2015.425.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers, September 2024. URL http://arxiv.org/abs/2401.08740. arXiv:2401.08740 [cs.CV].

Warren R. Morningstar, Cusuh Ham, Andrew G. Gallagher, Balaji Lakshminarayanan, Alexander A. Alemi, and Joshua V. Dillon. Density of states estimation for out of distribution detection. In *Proc. of AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3232–3240, 2021. URL http://proceedings.mlr.press/v130/morningstar21a.html.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. *ArXiv preprint*, abs/1906.02994, 2019a. URL https://arxiv.org/abs/1906.02994.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *Proc. of ICLR*, 2019b. URL https://openreview.net/forum?id=H1xwNhCcYm.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, and others. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4, 2011.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc of CVPR*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640. URL https://doi.org/10.1109/CVPR.2015.7298640.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of NeurIPS*, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Proc. of NeurIPS*, pp. 14680–14691, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html.

Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In *Proc. of NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f106b7f99d2cb30c3db1c3cc0fde9ccb-Abstract.html.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *Proc. of ICLR*, 2020. URL https://openreview.net/forum?id=SyxIWpVYvr.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-Distribution Detection with Deep Nearest Neighbors, December 2022. URL http://arxiv.org/abs/2204.06507. arXiv:2204.06507 [cs.LG].

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-Of-Distribution with Virtual-logit Matching, March 2022. URL http://arxiv.org/abs/2203.10807. arXiv:2203.10807 [cs.CV].

Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *Proc. of NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddea82ad2755b24c4e168c5fc2ebd40-Abstract.html.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection, October 2022. URL http://arxiv.org/abs/2210.07242. arXiv:2210.07242 [cs.CV].

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection, December 2024. URL http://arxiv.org/abs/2306.09301. arXiv:2306.09301 [cs.LG].

Lily H. Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12427–12436, 2021. URL http://proceedings.mlr.press/v139/zhang21g.html.

A METRIC ABLATIONS

	Bright	Contrast	Defocus	Elastic	Fog	Frost	Gauss B	Gauss N	Glass	Impulse	JPEG	Motion	Pixel	Saturate	Shot	Snow	Spatter	Speckle	Zoom
SITN	0.82	0.60	0.73	0.74	0.64	0.88	0.75	0.91	0.95	1.00	0.97	0.90	0.95	0.74	0.93	0.96	0.97	0.96	0.80
AD	0.85	0.60	0.52	0.61	0.49	0.79	0.58	0.69	0.85	1.00	0.78	0.52	0.67	0.75	0.93	0.89	0.98	0.95	0.53
CV	0.52	0.57	0.76	0.76	0.67	0.84	0.79	0.92	0.95	0.99	0.98	0.93	0.96	0.60	0.89	0.96	0.73	0.91	0.86

Figure 5: Ablation showing the OOD detection performance of the individual statistics **Anderson Darling (AD)** and the **coefficient of variation of the power spectrum (CV)** against the **full method (SITN)**, which combines them. OOD detection performance is shown in terms of AUROC values across the synthetic perturbations from CIFAR-10-C. For each corruption, samples with all severity levels were pooled. The two statistics show complementary value where for example AD is more sensitive to brightness perturbations, whereas CV shows better discrimination for pixellation or motion artefacts.

B COMPLEXITY BIAS

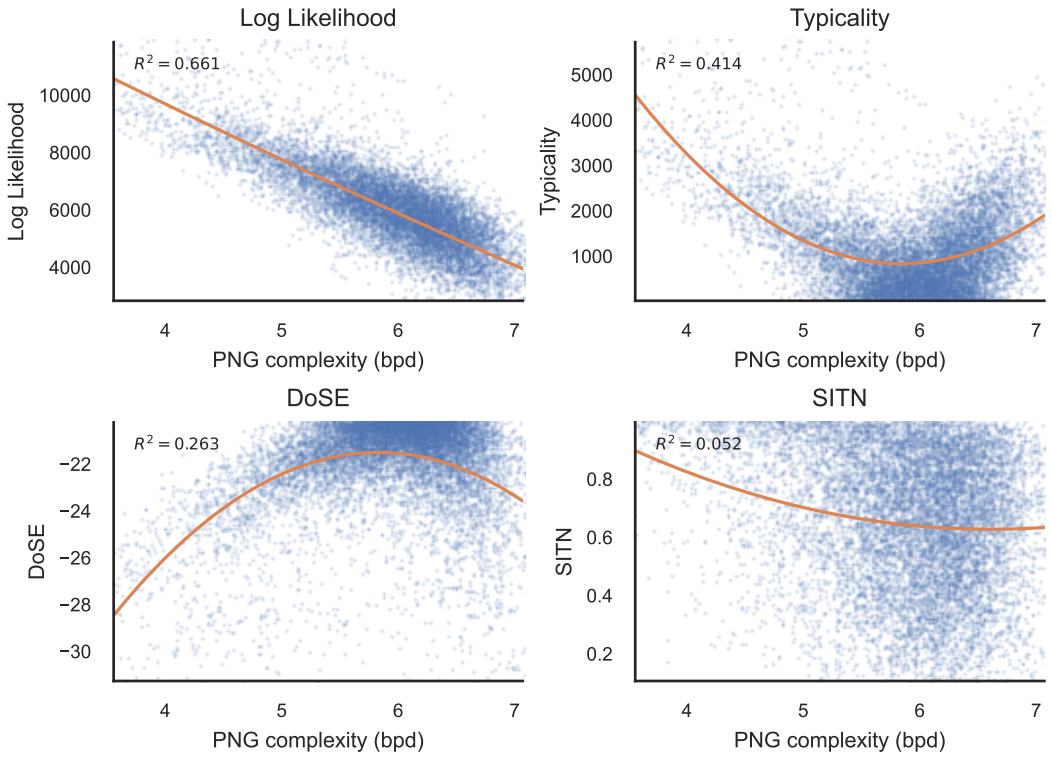


Figure 6: Complexity bias of OOD detection metrics. Scatter plots showing the relationship between image complexity (measured by PNG compression size) and the different OOD detection scores for a model trained and evaluated on the CIFAR-10 train and test splits, respectively. The lines show quadratic fits and R^2 values are reported alongside. Except for SITN, all metrics show a strong complexity bias.

C LIKELIHOOD DISTRIBUTIONS

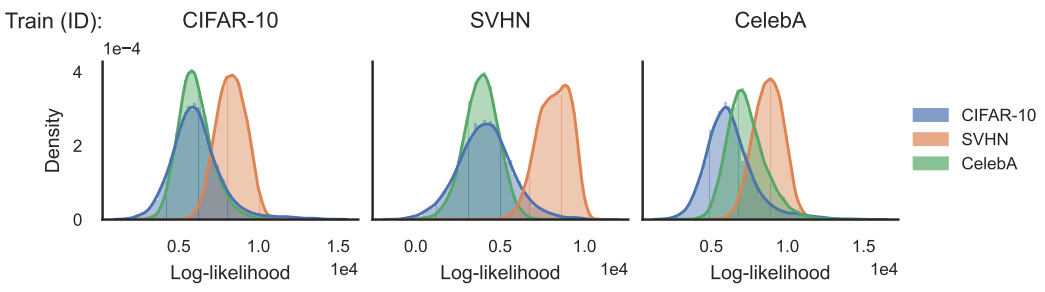


Figure 7: Log-Likelihood distributions across the different datasets. Plot titles indicate what dataset the corresponding model was trained on, e.g., the left plot shows the log-likelihood distributions for the model trained on CIFAR-10. In line with prior work, OOD samples from SVHN are consistently assigned higher likelihoods than ID CIFAR-10 (test) samples for the CIFAR-10 model.

D METRIC EXTREMES ON CELEBA

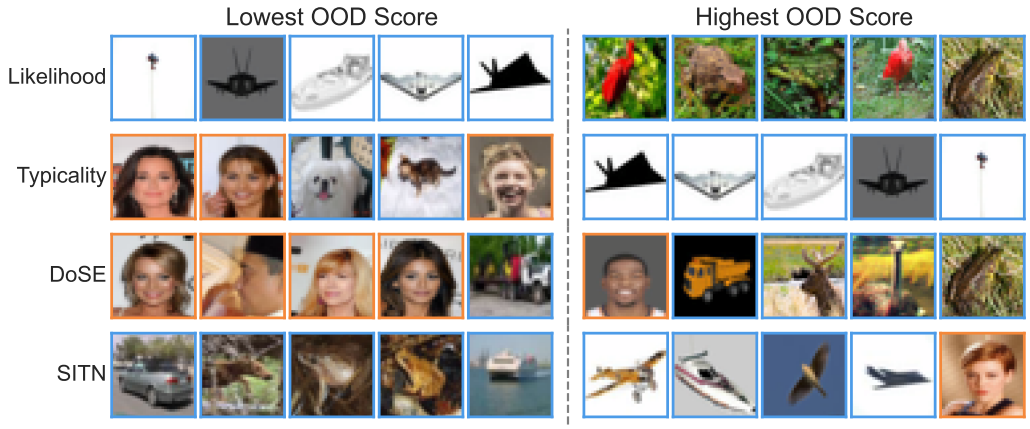


Figure 8: Images with the highest and lowest OOD scores for each method in the CIFAR-10→CelebA setup, i.e. trained on CIFAR-10 and evaluated on the test split of CIFAR-10 as ID and the test split of CelebA as OOD. ID images from CIFAR-10 are marked with a blue border and OOD images from CelebA with an orange border. Both Typicality and DoSE inherit failure cases from the complexity-biased likelihood scores and assign very low OOD scores to OOD images from CelebA.

E ADDITIONAL BASELINES

In addition to the baseline methods described in Section 4.2, we also evaluated the complexity correction method by Serrà et al. (2020), the ensemble method WAIC by Choi et al. (2018), and DiffPath by Heng et al. (2024).

For the complexity correction, we used PNG compression as the complexity metric. Specifically, PNG complexity $c(X)$ was computed in units of bits per dimension (bpd) for each image (after resizing to 32x32 for CelebA and without channel-wise normalisation by the dataset mean and standard deviation). As in the original publication, the log-likelihoods $L(X)$ were then also expressed in bpd and the complexity-corrected OOD score was computed as $S(X) = -L(X) - c(X)$.

We see this method as a less relevant benchmark for SITN due to its additional requirements of a suitable complexity metric that may not be available in less standard domains. Nevertheless, SITN consistently outperforms the complexity correction approach across all conditions (see Table 3). Our evaluations also include results for models trained on SVHN and CelebA, which were not previously reported by Serrà et al. (2020) and reveal severe failures of the method for the CelebA→CIFAR-10 and CelebA→SVHN setups.

Table 3: Comparison to additional baselines for cross-dataset OOD detection performance in terms of AUROC. Values in brackets denote bootstrapped 95% CIs and results for the best method are highlighted in **bold**.

Train (ID)	Test (OOD)	Complexity	WAIC	DiffPath	SITN (Ours)
CIFAR-10	SVHN	0.782 [.777, .787]	0.332 [.326, .338]	0.925 [.922, .928]	**0.946** [.943, .949]
	CelebA	0.615 [.608, .623]	0.450 [.444, .458]	0.573 [.566, .580]	**0.671** [.665, .678]
SVHN	CIFAR-10	0.729 [.723, .734]	0.745 [.739, .750]	0.961 [.959, .963]	**0.986** [.984, .987]
	CelebA	0.884 [.881, .887]	0.736 [.732, .741]	0.979 [.978, .980]	**0.996** [.995, .996]
CelebA	CIFAR-10	0.224 [.218, .230]	0.609 [.603, .616]	0.766 [.760, .771]	**0.910** [.907, .914]
	SVHN	0.077 [.075, .080]	0.449 [.444, .454]	0.984 [.983, .985]	**0.986** [.985, .987]

For WAIC, we followed the original publication by training an ensemble of five models with different parameter initialisations. The WAIC score is then computed as the mean log-likelihood minus the variance of the log-likelihoods across the five models, with lower scores interpreted as more out-of-distribution. As explained in (Choi et al., 2018), the variance term is intended to act as a correction term to penalise high epistemic uncertainty in the ensemble.

Compared to SITN, WAIC introduces substantial computational overhead, requiring the training of an entire ensemble of generative models along with likelihood evaluations for every model in the ensemble during inference time. As shown in Table 3, SITN nevertheless outperforms WAIC across all conditions. In fact, we find that WAIC performs quite poorly throughout our experiments, which we investigated further in Appendix F.

For DiffPath, which was formulated for diffusion models, we used the close relationship between diffusion- and flow-based generative models to derive the score function required by DiffPath from the velocity field predictions of our flow model as shown in (Ma et al., 2024). This allowed us to directly compare against DiffPath using the same underlying generative model. We specifically implemented the DiffPath6D version following the code from the official git repository: `https://github.com/clear-nus/diffpath`.

SITN outperformed DiffPath across all conditions. We note that DiffPath has the advantage that it could be used with a generative model trained on data from a different distribution than the ID data (using ID data only to fit the Gaussian mixture model on the extracted scores), although the original publication reports that the method works best when the generative model was trained on in-distribution data as evaluated here.

F Further Investigation of WAIC Performance

In Appendix E, we reported the performance of WAIC (Choi et al., 2018) and observed it to perform quite poorly overall. To further investigate these results, we visualised the distributions of the ensemble mean and variance of the likelihood evaluations. The variance term is intended to serve as a correction term to penalise epistemic uncertainty, which the authors expect to be larger for OOD samples, therefore providing an OOD signal. As shown in Figure 9 (right panel), we did not observe this behaviour. The distribution of the ensemble variance was highly overlapping between in- and out-of-distribution samples, explaining why WAIC underperformed in our experiments.

We hypothesise that the flow matching models used in our setup produce much more stable likelihood evaluations than the ensemble of Glow models used in the original work. This would be in line with recent findings about the stability of flow matching (Briq et al., 2026).

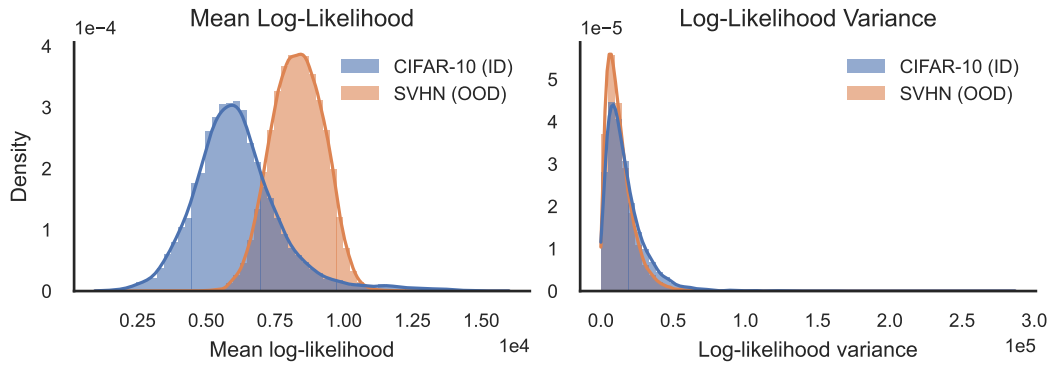


Figure 9: Distribution of the ensemble mean and variance of the log-likelihoods across the CIFAR-10 and SVHN test splits for a CIFAR-10-trained ensemble. WAIC is based on the idea of using the variance of the likelihoods (across ensemble models) as an OOD signal. However, in our experiments with flow matching models, we do not observe consistently higher variance for OOD samples.

G AD AND CV DISTRIBUTIONS

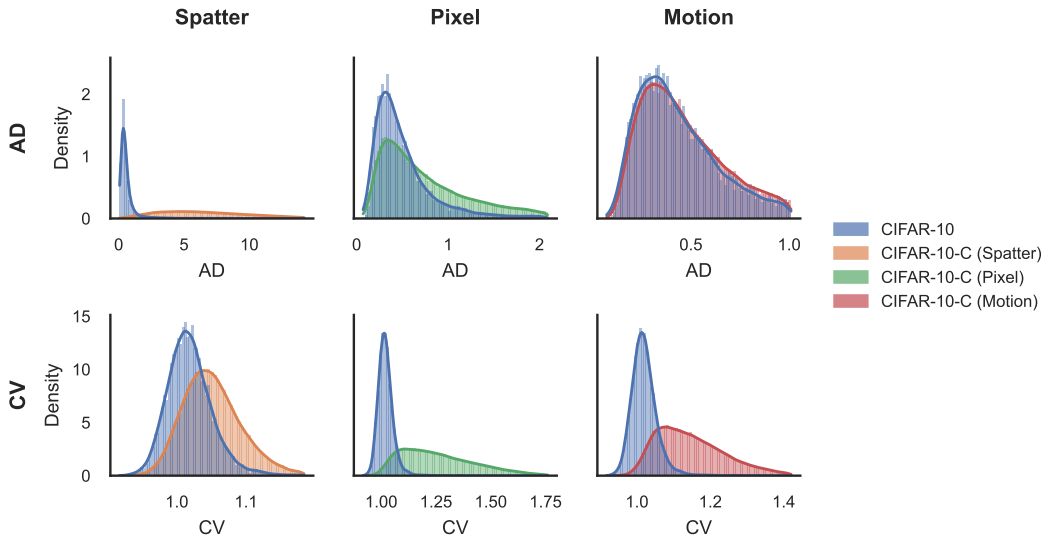


Figure 10: Distributions of the individual statistics **Anderson Darling (AD)** and the **coefficient of variation of the power spectrum (CV)** for different CIFAR-10-C conditions. As shown in Appendix A, the two statistics have complementary value. For instance, AD distributions are highly overlapping between ID and OOD data for the motion perturbation, but well separated for spatter. Conversely, the CV distributions are better separated for the motion perturbation.

H Analysis of Individual Likelihood Components

The instantaneous change of variables formula for the evaluation of the log-likelihood consists of the sum of two terms: the log-probability of the latent representation and the integral of the vector field's divergence. DoSE combines all three terms, i.e. the log-likelihood and its two constituents into an OOD score via a KDE. Typicality operates on the log-likelihood term alone with a simple absolute difference to the ID data average rather than a density estimate. SITN, on the other hand, operates on the latent representation itself and derives different statistics (AD and CV) from it rather than its log-probability.

Table 4 shows OOD detection results for the raw scores and KDE scores (same implementation details as for DoSE) using each of the individual components. Volume term denotes the divergence integral and the no-volume KDE condition denotes KDE scores using the likelihood and latent log-probability, but not the divergence integral. KDE likelihood-only can be seen as a refined version of Typicality, where a KDE is used instead of the simple absolute difference to the ID average. It results in higher performance in the CIFAR-10→SVHN condition, where it also outperforms DoSE, but slightly lower performance in the CIFAR-10→CelebA condition.

Table 4: OOD detection AUROC results for individual likelihood components in the instantaneous change of variables formula. Results for the full log-likelihood, Typicality, DoSE, and SITN (highlighted in gray) are shown for comparison.

Method	**CIFAR-10→SVHN**	**CIFAR-10→CelebA**
Log-likelihood	0.105 [.100, .110]	0.496 [.488, .503]
Latent Log-probability	0.518 [.511, .525]	0.489 [.482, .496]
Volume Term	0.104 [.099, .108]	0.496 [.488, .503]
KDE likelihood-only	0.828 [.822, .833]	0.418 [.410, .424]
KDE latent-only	0.488 [.481, .495]	0.499 [.492, .506]
KDE volume-only	0.832 [.826, .837]	0.417 [.410, .424]
KDE no-volume	0.763 [.757, .770]	0.443 [.436, .449]
Typicality	0.790 [.785, .796]	0.421 [.414, .428]
DoSE	0.802 [.796, .808]	0.432 [.425, .439]
SITN	0.946 [.943, .949]	0.671 [.665, .678]

I DoSE and Typicality Fit on Validation Data

Table 5: Control experiments using validation data to fit DoSE and the entropy estimate for Typicality. While both methods use training data in the original publications, these experiments test if instead using validation data meaningfully changes their performance. No consistent improvement is observed: both methods perform slightly better with validation data on SVHN, but worse on CelebA. Rows corresponding to the original configurations used in the main paper are highlighted in gray.

Train (ID)	**Test (OOD)**	**Fit Split**	**Typicality**	**DoSE**
CIFAR-10	SVHN	Val	0.803 [.798, .809]	0.812 [.806, .817]
		Train	0.790 [.785, .796]	0.802 [.796, .808]
	CelebA	Val	0.420 [.413, .427]	0.429 [.422, .436]
		Train	0.421 [.414, .428]	0.432 [.425, .439]

J Comparison to Discriminative OOD Detection Methods

J.1 Caveats for the Comparison of Discriminative and Generative OOD Detection Methods

Several factors complicate the comparisons between generative and discriminative approaches.

First, discriminative methods require labelled training data. This difference is consequential: in applications where labelled data are unavailable, discriminative methods that require such supervision are not applicable, whereas generative methods trained through unsupervised learning remain a viable option. Thus, even if discriminative methods achieve competitive or superior performance on particular benchmarks, the two approaches cannot be regarded as interchangeable, as their applicability differs fundamentally across problem settings.

Providing discriminative methods with class labels also results in an unfair advantage when pitched against generative methods that do not have access to that information. This is particularly important in widely used benchmarks like OpenOOD (Yang et al., 2022; Zhang et al., 2024), where OOD samples correspond to unseen class labels. With access to class labels, discriminative methods are optimised to discriminate precisely the semantic distinctions that are used to define whether a sample is OOD (Guille-Escuret et al., 2024).

Second, many discriminative methods require access to validation OOD data for the tuning of hyperparameters, a requirement that is accommodated by the OpenOOD benchmarks, but unrealistic in practice where the nature of distribution shifts is a priori unknown.

Finally, generative and discriminative methods are inherently based on different model architectures that are trained in different ways (supervised classification vs unsupervised learning). Performance comparisons between discriminative and generative approaches may therefore be obscured by differences in the underlying models.

J.2 OOD Detection on CIFAR-10-C

Figure 11 shows a comparison of generative and selected discriminative methods (MSP (Hendrycks & Gimpel, 2018), Energy scores (Liu et al., 2020), KNN (Sun et al., 2022), ViM (Wang et al., 2022)) across the 19 perturbations from CIFAR-10-C. For the discriminative methods, a pre-trained ResNet model from OpenOOD (Zhang et al., 2024) was used and results for the OpenOOD far-OOD detection dataset pairings were reproduced to validate implementations.

SITN outperformed all discriminative methods for 18 out of 19 corruptions.

	Bright	Contrast	Defocus	Elastic	Fog	Frost	Gauss B	Gauss N	Glass	Impulse	JPEG	Motion	Pixel	Saturate	Shot	Snow	Spatter	Speckle	Zoom	**Average**
MSP	0.52	0.63	0.63	0.64	0.57	0.69	0.70	0.82	0.83	0.80	0.68	0.68	0.67	0.54	0.78	0.67	0.63	0.78	0.71	0.68
Energy	0.52	0.61	0.62	0.62	0.56	0.68	0.70	0.80	0.82	0.79	0.66	0.66	0.67	0.54	0.77	0.66	0.62	0.76	0.69	0.67
KNN	0.53	0.66	0.65	0.68	0.59	0.73	0.73	0.85	0.87	0.85	0.71	0.72	0.69	0.57	0.81	0.70	0.68	0.80	0.74	0.71
ViM	0.51	0.65	0.64	0.64	0.57	0.71	0.72	0.85	0.84	0.86	0.66	0.70	0.65	0.54	0.81	0.66	0.65	0.80	0.72	0.69
Typicality	0.62	0.69	0.58	0.46	0.48	0.71	0.68	1.00	0.79	0.99	0.44	0.53	0.57	0.73	0.99	0.61	0.67	0.99	0.61	0.69
DoSE	0.77	0.70	0.72	0.53	0.49	0.91	0.82	1.00	0.91	1.00	0.65	0.65	0.59	0.82	1.00	0.78	0.73	1.00	0.79	0.78
SITN	0.82	0.60	0.73	0.74	0.64	0.88	0.75	0.91	0.95	1.00	0.97	0.90	0.95	0.74	0.93	0.96	0.97	0.96	0.80	0.85

Figure 11: Comparison between Generative and Discriminative OOD Detection Methods on CIFAR-10-C. OOD detection performance is shown in terms of AUROC values across the synthetic perturbations from CIFAR-10-C. For each corruption, samples with all severity levels were pooled.

J.3 Semantic Near- and Far-OOD Detection on OpenOOD

Tables 6 and 7 show far- and near-OOD detection results on the OpenOOD v1.5 benchmarks (Zhang et al., 2024) comparing discriminative and generative methods trained on CIFAR-10. As discussed in Section J.1, there are several caveats to this comparison. In particular, generative and discriminative methods are necessarily based on different model architectures and training setups. The results for all discriminative methods are directly reprinted from the extended results of the OpenOOD v1.5 publication and show standard deviations across three training runs rather than bootstrapped confidence intervals (Zhang et al., 2024). To respect the exact data partitioning used in the OpenOOD benchmarks, including the removal of certain images due to class overlaps, the generative baselines and SITN were re-trained and evaluated with the same splits.

SITN outperforms the best discriminative method (ViM) in the CIFAR-10→SVHN setup. On CIFAR-10→MNIST, SITN also outperforms the best discriminative method (MDSEns). Interestingly, all other discriminative methods (including ViM) perform considerably worse on this dataset pairing while performance of the generative approaches is saturated. On the other hand, several discriminative methods outperform the generative methods on Textures and Places365.

In the near-OOD dataset pairings from OpenOOD, SITN achieves the best performance among generative methods. However, all generative methods perform substantially worse than discriminative approaches in this setting. As noted in Section J.1, discriminative methods have a considerable advantage in this setting as they have access to class labels during training, which represent exactly the semantic distinctions used to define ID and OOD data in semantic OOD detection benchmarks like OpenOOD.

Table 6: OpenOOD Far-OOD Detection. Results for the discriminative methods are from (Zhang et al., 2024).

Method	SVHN	MNIST	Textures	Places365
OpenMax	89.77 ± 0.45	90.50 ± 0.44	89.58 ± 0.60	88.63 ± 0.28
MSP	91.46 ± 0.40	92.63 ± 1.57	89.89 ± 0.71	88.92 ± 0.47
TempScale	91.66 ± 0.52	93.11 ± 1.77	90.01 ± 0.74	89.11 ± 0.52
ODIN	84.58 ± 0.77	95.24 ± 1.96	86.94 ± 2.26	85.07 ± 1.24
MDS	91.18 ± 0.47	90.10 ± 2.41	92.69 ± 1.06	84.90 ± 2.54
MDSEns	66.56 ± 0.58	99.17 ± 0.41	77.40 ± 0.28	52.47 ± 0.15
RMDS	91.84 ± 0.26	93.22 ± 0.80	92.23 ± 0.23	91.51 ± 0.11
Gram	91.52 ± 4.45	72.64 ± 2.34	62.34 ± 8.27	60.44 ± 3.41
EBO	91.79 ± 0.98	94.32 ± 2.53	89.47 ± 0.70	89.25 ± 0.78
OpenGAN	52.81 ± 27.60	56.14 ± 24.08	56.14 ± 18.26	53.34 ± 5.79
GradNorm	53.91 ± 6.36	63.72 ± 7.37	52.07 ± 4.09	60.50 ± 5.33
ReAct	89.12 ± 3.19	92.81 ± 3.03	89.38 ± 1.49	90.35 ± 0.78
MLS	91.69 ± 0.94	94.15 ± 2.48	89.41 ± 0.71	89.14 ± 0.76
KLM	84.99 ± 1.18	85.00 ± 2.04	82.35 ± 0.33	78.37 ± 0.33
VIM	94.50 ± 0.48	94.76 ± 0.38	**95.15** ± 0.34	89.49 ± 0.39
KNN	92.67 ± 0.30	94.26 ± 0.38	93.16 ± 0.24	**91.77** ± 0.23
DICE	90.02 ± 1.77	90.37 ± 5.97	81.86 ± 2.35	74.67 ± 4.98
RankFeat	68.15 ± 7.44	75.87 ± 5.22	73.46 ± 6.49	85.99 ± 3.04
ASH	73.46 ± 6.41	83.16 ± 4.66	77.45 ± 2.39	79.89 ± 3.69
SHE	86.38 ± 1.32	90.43 ± 4.76	81.57 ± 1.21	82.89 ± 1.22
GEN	91.97 ± 0.66	93.83 ± 2.14	90.14 ± 0.76	89.46 ± 0.65
Likelihood	10.5 [10.1, 11.0]	0.01 [0.00, 0.01]	61.0 [60.0, 62.0]	65.8 [65.2, 66.4]
Typicality	78.9 [78.4, 79.5]	**99.9** [99.9, 99.9]	64.2 [63.2, 65.1]	60.5 [59.8, 61.1]
DoSE	80.6 [80.0, 81.2]	**99.9** [99.8, 99.9]	74.9 [74.0, 75.7]	59.8 [59.2, 60.4]
SITN	**95.2** [94.9, 95.4]	**99.9** [99.8, 99.9]	89.3 [88.7, 89.8]	71.3 [70.7, 71.9]

Table 7: OpenOOD Near-OOD Detection. Results for the discriminative methods are from (Zhang et al., 2024).

Method	CIFAR-100	TIN
OpenMax	86.91 ± 0.31	88.32 ± 0.28
MSP	87.19 ± 0.33	88.87 ± 0.19
TempScale	87.17 ± 0.40	89.00 ± 0.23
ODIN	82.18 ± 1.87	83.55 ± 1.84
MDS	83.59 ± 2.27	84.81 ± 2.53
MDSEns	61.29 ± 0.23	59.57 ± 0.53
RMDS	88.83 ± 0.35	90.76 ± 0.27
Gram	58.33 ± 4.49	58.98 ± 5.19
EBO	86.36 ± 0.58	88.80 ± 0.36
OpenGAN	52.81 ± 7.69	54.62 ± 7.68
GradNorm	54.43 ± 1.59	55.37 ± 0.41
ReAct	85.93 ± 0.83	88.29 ± 0.44
MLS	86.31 ± 0.59	88.72 ± 0.36
KLM	77.89 ± 0.75	80.49 ± 0.85
VIM	87.75 ± 0.28	89.62 ± 0.33
KNN	**89.73** ± 0.14	**91.56** ± 0.26
DICE	77.01 ± 0.88	79.67 ± 0.87
RankFeat	77.98 ± 2.24	80.94 ± 2.80
ASH	74.11 ± 1.55	76.44 ± 0.61
SHE	80.31 ± 0.69	82.76 ± 0.43
GEN	87.21 ± 0.36	89.20 ± 0.25
Likelihood	51.7 [50.9, 52.5]	48.4 [47.5, 49.2]
Typicality	55.1 [54.3, 55.9]	48.3 [47.4, 49.2]
DoSE	56.0 [55.2, 56.8]	61.2 [60.4, 62.1]
SITN	57.2 [56.4, 58.0]	69.0 [68.2, 69.8]

K CALIBRATION

In the following, we demonstrate how S_{SITN} can be calibrated, allowing one to specify a tolerable false positive rate, α. This is accomplished by splitting the calibration set, C, into disjoint subsets, C_1 and C_2, then using C_1 to estimate the empirical CDFs of the constituent S_{AD} and S_{CV} statistics, and C_2 to estimate the empirical CDF, $\hat{F}_{S_{\text{SITN}}}$ of S_{SITN}. Finally, one defines the ID versus OOD classification rule as

$$h(X) = \begin{cases} \text{OOD} & \text{if } \hat{F}_{S_{\text{SITN}}}(S_{\text{SITN}}) \geq 1 - \alpha \\ \text{ID} & \text{otherwise.} \end{cases} \tag{4}$$

We show below how this controls false positive (i.e., Type I error) rate, and provide some analysis about false negative (i.e., Type II error) rates under some assumptions.

K.1 TYPE I ERROR CONTROL

A hypothesis test has a Type I error (i.e., false positive) rate of α if the probability of falsely rejecting the null hypothesis, H_0, is α. In the context of OOD detection, this can be stated formally as

$$P_{\text{ID}}(h(X) = \text{OOD}) = \alpha. \tag{5}$$

We prove Type I error control for a slightly more general class of composite hypotheses that can be decomposed into a conjunction over k constituent hypotheses. The approach generalises by simply estimating the empirical CDF for all k associated statistics, and taking a k-way max instead of a two-way max.

Let us write $M = \max_i F_{S_i}(S_i)$; then

$$P_{\text{ID}}(h(X) = \text{OOD}) = P_{\text{ID}}(F_M(M) \geq 1 - \alpha) \tag{6}$$
$$= 1 - P_{\text{ID}}(F_M(M) < 1 - \alpha) \tag{7}$$
$$= \alpha, \tag{8}$$

because $F_M(M)$ follows a uniform distribution. This tells us that controlling the Type I error rate for the composite hypothesis test can be accomplished by estimating F_M with bounded error. We demonstrate that our estimation procedure has bounded error in the theorem below.

Theorem K.1. *Let $\hat{M} = \max_i \hat{F}_{S_i}(S_i)$ be our estimate of M, where $\hat{F}_{S_i}$ are estimated using C_1 and $\hat{F}_{\hat{M}}$ is estimated using C_2. Then, with probability at least $1 - \delta$,*

$$\|F_M - \hat{F}_{\hat{M}}\|_\infty \leq 2k\sqrt{\frac{1}{2|C_1|}\log\frac{2k}{\delta}} + \sqrt{\frac{1}{2|C_2|}\log\frac{2}{\delta}}.$$

The proof is given in Section K.2. Taking $\hat{M} = S_{\text{SITN}}$ and $k = 2$ with the two statistics identified in the main paper, we obtain the following corollary for $|C| = n$ and $|C_1| = |C_2| = \frac{n}{2}$.

Corollary K.2. *The Type I error of the SITN test is bounded from above, with probability at least $1 - \delta$ over the draws of the calibration sets, by*

$$P_{ID}(h(X) = OOD) \leq \alpha + 4\sqrt{\frac{1}{n}\log\frac{4}{\delta}} + \sqrt{\frac{1}{n}\log\frac{2}{\delta}}.$$

This tells us that, as the size of the calibration sets increases, the control on the Type I error becomes more precise.

K.2 Proof of theorem K.1

The proof of this theorem will make use of the DKW inequality, given below.

Theorem K.3 (Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al., 1956))**.** *Let $\hat{F}$ be an empirical distribution function based on n i.i.d. samples from a distribution function F. Then the probability of maximum error between the distributions, $\sup_{x\in\mathbb{R}}(\hat{F}(x) - F(x))$ exceeding $\epsilon \in \mathbb{R}$ can be bounded by*

$$P(\sup_{x\in\mathbb{R}} |\hat{F}(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}.$$

We will also make use of the following Lemma.

Lemma K.4. *Let $\hat{F}$ be an empirical distribution function based on n i.i.d. samples from a continuous distribution with CDF F. Then, with probability at least $1 - \delta$,*

$$\|\hat{F}(x) - F(x)\|_\infty \leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

Proof of Lemma K.4. Let $\delta = P(\sup_{x\in\mathbb{R}} |\hat{F}(x) - F(x)| \geq \epsilon)$. Then, by the DKW Theorem (Theorem K.3)

$$\delta \leq 2e^{-2n\epsilon^2}$$
$$\implies \epsilon \leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$$

Therefore, with at least probability $1 - \delta, \sup_{x\in\mathbb{R}} |\hat{F}(x) - F(x)| \leq \epsilon$. It follows that, with probability $1 - \delta$, that $\|\hat{F}(x) - F(x)\|_\infty \leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$ □

This lemma can be easily extended to the case with k hypotheses as follows

Lemma K.5. *Let $\{F_i\}_{i=1}^k$ be a family of continuous distributions and $\{\hat{F}_i\}_{i=1}^k$ be a family of empirical distribution functions, where $\hat{F}_i$ is estimated based on n i.i.d. draws from the corresponding distribution. Then, with probability $1-\delta$;*

$$\max_i \|F_i(x) - \hat{F}_i(x)\|_\infty \leq \sqrt{\frac{1}{2n}\log\frac{2k}{\delta}}.$$

Proof of Lemma K.5. This result follows from a straightforward application of the union bound to the DKW in Lemma K.4. With probability at least $1-\delta$

$$P(\exists i \in \{1...k\} : \max_x |F_i(x) - \hat{F}_i(x)| \geq \epsilon) \leq \sum_{i=1}^k \mathbb{P}(|\max_x F_i(x) - \hat{F}_i(x)| \geq \epsilon) \leq 2ke^{-2n\epsilon^2}$$

If $\delta = P(\exists i \in \{1...k\} : \sup_x |F_i(x) - \hat{F}_i(x)| \geq \epsilon)$ then, for $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2k}{\delta}}$, with probability $1-\delta$, we have that

$$\max_i \sup_x |F_i(x) - \hat{F}_i(x)| \leq \epsilon.$$

□

The proof of Theorem K.1 will also make use of the following Lemma.

Lemma K.6. *Let F_i be the CDF over test statistic S_i, and $\hat{F}_i$ the empirical CDF of F based on n i.i.d. samples. If $M = \max_i F_i(S_i)$ and $\hat{M} = \max_i \hat{F}_i(S_i)$ then, with probability at least $1-\delta$,*

$$\|F_M - F_{\hat{M}}\|_\infty \leq 2k\sqrt{\frac{1}{2n}\log\frac{2k}{\delta}}.$$

Proof of Lemma K.6. Construct an upper bound on $|M - \hat{M}|$;

$$\begin{aligned}|M - \hat{M}| &= |\max_i F_i(S_i) - \max_i \hat{F}_i(S_i)| \\ &\leq \max_i |F_i(S_i) - \hat{F}_i(S_i)| \leq \epsilon\end{aligned}$$

with probability $1-\delta$ for $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2k}{\delta}}$. This result follows from the monotonicity of the max operation and application of Lemma K.5.

Using this bound observe that, if $M \leq t - \epsilon$ then $\hat{M} \leq t$. Similarly, $\hat{M} \leq t \implies M \leq t + \epsilon$. Therefore $F_M(t-\epsilon) \leq F_{\hat{M}}(t) \leq F_M(t+\epsilon)$. It follows that

$$\begin{aligned}|F_{\hat{M}}(t) - F_M(t)| &\leq \max(F_M(t+\epsilon) - F_M(t), F_M(t) - F_M(t-\epsilon)) \\ &\leq F_M(t+\epsilon) - F_M(t-\epsilon) \\ \implies \sup_t |F_{\hat{M}}(t) - F_M(t)| &\leq \sup_t (F_M(t+\epsilon) - F_M(t-\epsilon)).\end{aligned}$$

Finally, observe that

$$\begin{aligned}F_M(t+\epsilon) - F_M(t-\epsilon) &= P(\max_i S_i \leq t+\epsilon) - P(\max_i S_i \leq t-\epsilon) \\ &\leq \sum_{i=1}^k P(U_i \in (t-\epsilon, t+\epsilon])\end{aligned}$$

by application of the union bound. Since $U_i = F_i(S_i)$ it is uniformly distributed in $[0,1]$, we have that $P(U_i \in (t-\epsilon, t+\epsilon]) = 2\epsilon$. And so, with probability at least $1-\delta$

$$\begin{aligned}\sup_t |F_M(t) - F_{\hat{M}}(t)| &\leq 2k\epsilon \\ &= 2k\sqrt{\frac{1}{2n}\log\frac{2k}{\delta}}\end{aligned}$$

completing the proof. □

Proof of Theorem K.1. Rewrite $\|F_M - \hat{F}_{\hat{M}}\|_\infty$ using the triangle inequality, as follows

$$\begin{aligned}\|F_M - \hat{F}_{\hat{M}}\|_\infty &= \|F_M - F_{\hat{M}} + F_{\hat{M}} - \hat{F}_{\hat{M}}\|_\infty \\ &\leq \|F_M - F_{\hat{M}}\|_\infty + \|F_{\hat{M}} - \hat{F}_{\hat{M}}\|_\infty.\end{aligned}$$

By application of Lemma K.6 and Lemma K.4

$$\|F_M - F_{\hat{M}}\|_\infty + \|F_{\hat{M}} - \hat{F}_{\hat{M}}\|_\infty \leq 2k\sqrt{\frac{1}{2|C_1|}\log\frac{2k}{\delta}} + \sqrt{\frac{1}{2|C_2|}\log\frac{2}{\delta}}$$

□

K.3 Empirical Demonstration of Type I Error Control

Figure 12 (left panel) shows calibration curves for the CIFAR-10-trained model evaluated on the CIFAR-10 test data. As expected, the raw max-quantile SITN scores are not calibrated (blue line). While the individual quantile-transformed statistics $\hat{F}_{\text{AD}}(S_{\text{AD}})$ and $\hat{F}_{\text{CV}}(S_{\text{CV}})$ could be interpreted as the fraction of validation samples with equal or lower scores, this interpretation no longer holds after their max-aggregation into the combined SITN OOD score. Using the method described above (Appendix K), we obtain a well-calibrated score (orange line). We also show Tippett's method (green line) for comparison, which assumes independence between the combined components, and observe that it produced only slightly more conservative scores in this case.

For comparison and illustration purposes, we also show calibration curves for the same calibration approaches for the combination of two statistics where the independence assumption for Tippett's method is clearly violated (Figure 12, right panel). In this example, we combine the Anderson-Darling statistic and the Kolmogorov-Smirnov statistic, both of which are sensitive to deviations from normality. It can be seen that Tippett's method is much too conservative in this case, whereas calibration on the split validation data, which does not assume independence of the statistics, still yields well-calibrated results. Consequently, when exploring new combinations of test statistics in future work, we caution against relying on blanket independence assumptions and recommend using the procedure described above.

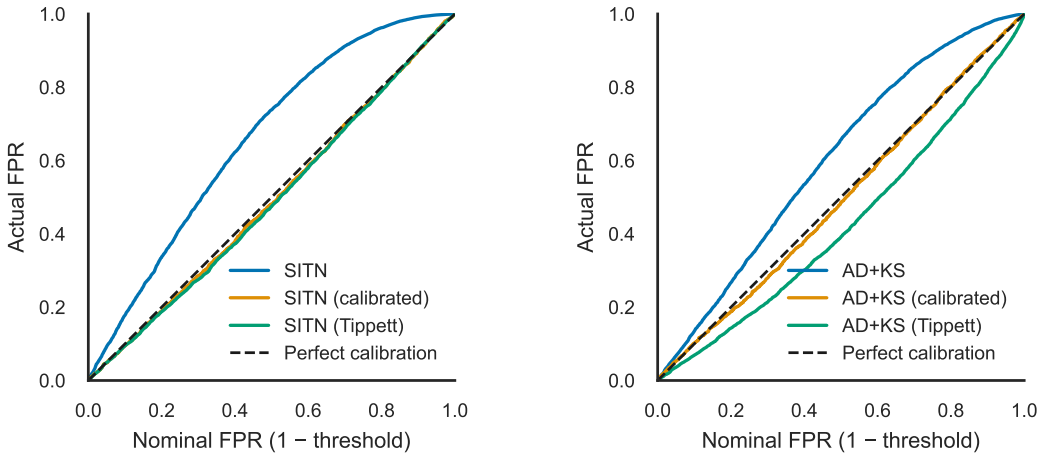


Figure 12: Calibration curves showing the actual false positive rate (FPR)—the fraction of in-distribution samples flagged as OOD—as a function of the nominal FPR set by the detection threshold on the CIFAR-10 test data. A perfectly calibrated method lies on the diagonal (dashed). **Left:** It can be seen that the raw max quantile scores (SITN) are uncalibrated, whereas Tippett's method (Tippett) and calibration on the split validation data (calibrated) provide similar and well-calibrated results. **Right:** For comparison, calibration with the split validation data (calibrated) remains effective if two statistics that clearly violate the independence assumption for Tippett's method (Tippett) are combined (here Anderson-Darling and Kolmogorov-Smirnov).

K.4 TYPE II ERROR CONTROL

The Type II error rate is given by

$$P_{\text{OOD}}(h_\gamma(X) = \text{ID}) = P_{\text{OOD}}(F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma \wedge F_{S_{\text{CV}}}(S_{\text{CV}}) \leq \gamma), \tag{9}$$

where we use γ to denote $F^{-1}_{S_{\text{SITN}}}(1-\alpha)$. We have no access to data from P_{OOD}, so it is not possible to control this quantity without making some assumptions about the relationship between P_{OOD} and P_{ID}. Because we will now make assumptions about the form of P_{OOD}, we take the liberty of making a few other assumptions that will ease the analysis without substantially changing the conclusions. In particular, we assume that ϕ perfectly models P_{ID}, that $\mathbb{E}[X] = 0$ for $X \sim P_{\text{ID}}$, and that the calibration set, C, is infinitely large. I.e., our empirical CDFs equal the true CDFs.

We define P_{OOD} in terms of how its density is related to that of P_{ID}. In particular, denoting density functions using lower case p, for some positive real a we consider

$$p_{\text{OOD}}(X) = a^{-D} e^{-\frac{1-a^2}{a^2}\|X\|_2^2} p_{\text{ID}}(X). \tag{10}$$

For $a = 1$ this sets $P_{\text{OOD}} = P_{\text{ID}}$, but for $a \neq 1$ it will result in more density being assigned to the tails or centre of the distribution. We choose this particular re-weighting because it results in a convenient distribution for ϕ^{-1} when $X \sim P_{\text{OOD}}$, as stated in the lemma below. However, we also feel that it matches the intuition that many people have about OOD data being more likely to be found in the tails of the distribution, and it is a natural way to parameterise the relationship between P_{OOD} and P_{ID} in terms of a single parameter a that controls how much more likely OOD data is to be found in the tails compared to ID data.

Lemma K.7. *Under the above definition of P_{OOD}, we have that*

$$\text{Law}(\phi^{-1}(X)) = \mathcal{N}(0, a^2 I_d), \qquad X \sim P_{OOD}. \tag{11}$$

Proof. We can use the change of variable formula to derive the density of $\phi^{-1}(X)$ in the situations where $X \sim P_{\text{ID}}$ and $X \sim P_{\text{OOD}}$. In the first case we have

$$p_0(\phi^{-1}(X)) = p_{\text{ID}}(X)|\text{det} J_{\phi^{-1}}(X)|, \tag{12}$$

and in the second we have that

$$p_{\text{A}}(\phi^{-1}(X)) = p_{\text{OOD}}(X)|\text{det} J_{\phi^{-1}}(X)|. \tag{13}$$

Taking the ratio of these, we obtain

$$\begin{aligned} p_{\text{A}}(\phi^{-1}(X)) &= \frac{p_{\text{OOD}}(X)}{p_{\text{ID}}(X)} p_0(\phi^{-1}(X)) \\ &= a^{-D} e^{-\frac{1-a^2}{a^2}\|X\|_2^2} \mathcal{N}(x; 0, I) \\ &= \mathcal{N}(x; 0, a^2 I). \end{aligned}$$

□

We can decompose the Type II error as

$$P_{\text{OOD}}(F_{S_{\text{CV}}}(S_{\text{CV}}) \leq \gamma | F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma) P_{\text{OOD}}(F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma) \tag{14}$$

$$= P_{\text{ID}}(F_{S_{\text{CV}}}(S_{\text{CV}}) \leq \gamma | F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma) P_{\text{OOD}}(F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma), \tag{15}$$

because the change in variance of the noise distribution for OOD compared to ID data does not result in a change in the CV statistic, due to the scale invariance of this statistic. Now we note that

$$P_{\text{ID}}(F_{S_{\text{CV}}}(S_{\text{CV}}) \leq \gamma | F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma) = \frac{1-\alpha}{P_{\text{ID}}(F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma)} \tag{16}$$

$$= \frac{1-\alpha}{\gamma}, \tag{17}$$

due to the marginal uniform distribution induced by the quantile transform. This yields a Type II error rate of

$$P_{\text{OOD}}(h_\gamma(X) = \text{ID}) = \frac{1-\alpha}{\gamma} P_{\text{OOD}}(F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma) \tag{18}$$

$$= \frac{1-\alpha}{F^{-1}_{S_{\text{SITN}}}(1-\alpha)} P_{\text{OOD}}(F_{S_{\text{AD}}}(S_{\text{AD}}) \leq \gamma). \tag{19}$$

From this, we can conclude that the Type II error rate of SITN for the defined P_{OOD} is dependent mainly on the Type II error of the Anderson-Darling test, premultiplied by a factor that can be interpreted as measuring the calibration of the S_{SITN} test statistic. I.e., if S_{SITN} is already perfectly calibrated it will follow a uniform distribution, and the premultiplication factor will be one. Prior work has shown that Anderson-Darling tests exhibit favourable Type II error rates in the setting where the alternative hypothesis are zero mean, non-unit variance normal distributions, such as the case of our conjectured P_{OOD}.

L Training Details

We used the UNet architecture following Ho et al. (2020), as implemented in Dhariwal & Nichol (2021) with the hyperparameters detailed in Table 8. Models were optimized with AdamW (learning rate 1e-4, weight decay 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon =$ 1e-8) and global gradient norm clipping (max norm 1.0) for 500,000 gradient steps (batch size 128), with weights restored to the checkpoint of lowest validation loss.

All images were normalised per-channel using statistics computed from the respective training set. No data augmentation was applied during training.

Table 8: UNet hyperparameters

Hyperparameter	Value
Channels	128
Depth	2
Channels multiple	1, 2, 2, 2
Heads	1
Attention resolution	16
Dropout	0.0

Models were trained and evaluated on an internal cluster with L40S GPUs. On this hardware, the training time for an individual model was around 20 hours and likelihood evaluations took about 6 minutes per 1,000 samples. The total compute time for the cross-dataset results (Section 5.1) is therefore around 100 hours (training 3 models, evaluating likelihoods on training and validation data for each dataset required for the fitting of Typicality, DoSE and SITN, and evaluating likelihoods on the test splits of all three datasets for each individual model). Additional evaluations for the perturbation experiments (Section 5.2) take roughly another 100 hours (50,000 samples for each of the 19 corruptions). The additional baseline experiments with the WAIC ensemble approach (Appendix E) require the training and evaluation of an additional four models (for a total of 5 models in the ensemble) for the aforementioned cross-dataset evaluations, resulting in another 400 hours of compute time.

M Statistical Analysis

Confidence intervals for AUROC values were computed through a stratified bootstrap resampling procedure with 10,000 iterations. In each iteration, data was sampled with replacement while preserving the original ratio of in-distribution to out-of-distribution samples. The confidence intervals were derived using the empirical 2.5th and 97.5th percentiles of the resulting metric distributions (i.e., the percentile bootstrap method).

N SOFTWARE

The research code for our experiments is publicly available at [removed for review]. Dependencies notably include PyTorch (Paszke et al., 2019), the flow matching library by Lipman et al. (2024) for implementations of the ODE solvers for the likelihood evaluations, and the UNet implementation by Dhariwal & Nichol (2021).

O TOP 150 SAMPLES WITH HIGHEST OOD SCORES

In the following we visualise the top 150 samples with the highest OOD scores for the CIFAR-10→SVHN setup for each of the baseline methods and SITN. For all baseline methods, the highest OOD scores are predominantly assigned to in-distribution samples (failure cases), whereas the highest SITN OOD scores mostly correspond to actual OOD samples.

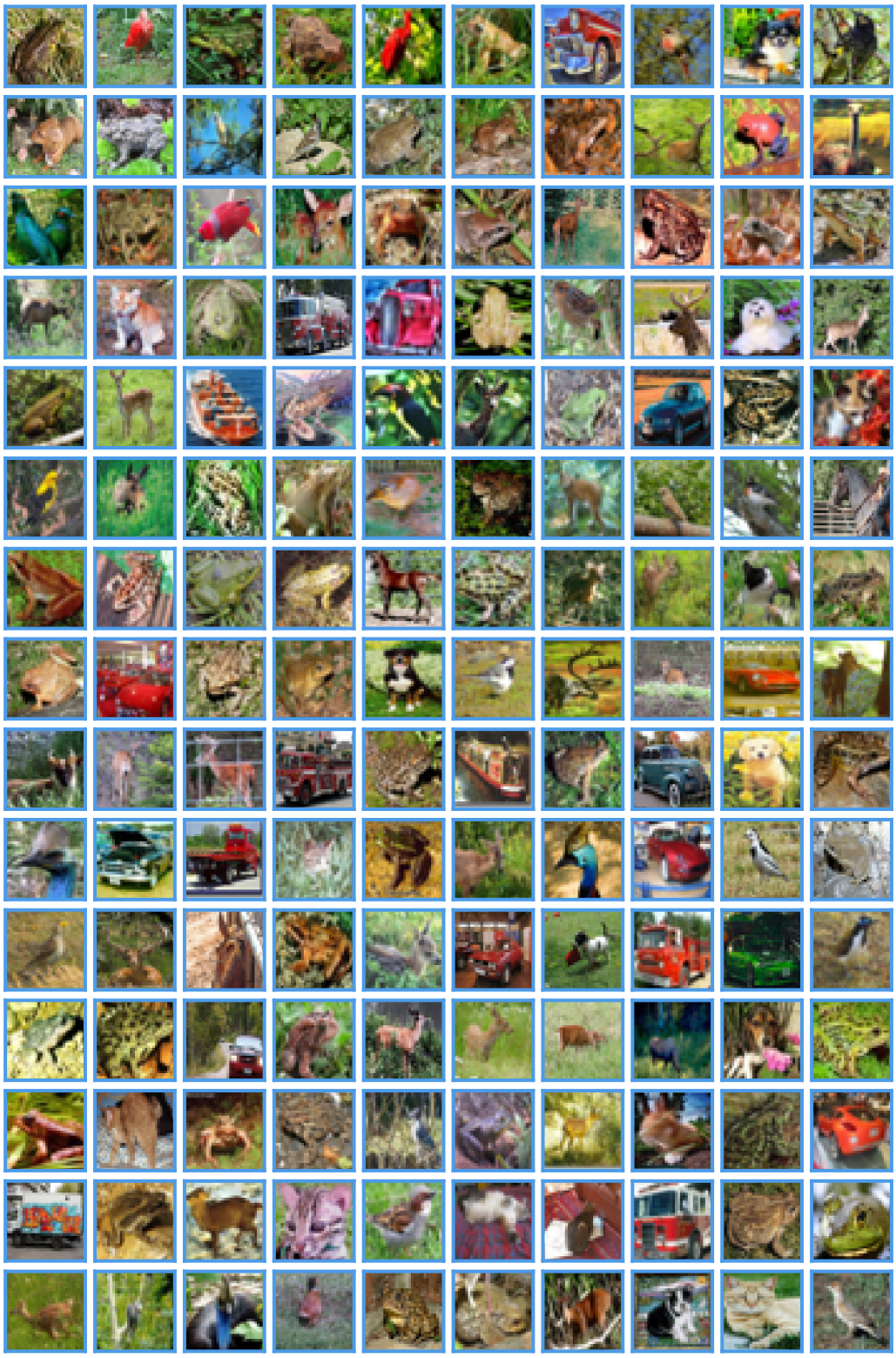


Figure 13: Top 150 samples with highest log-likelihood OOD score (lowest log-likelihoods) for the CIFAR-10→SVHN setup. ID images from CIFAR-10 and OOD images from SVHN are marked with blue and orange borders, respectively. All 150 samples flagged as most OOD by the likelihood metric are in-distribution (blue), illustrating a pronounced failure of the method in this setup.



Figure 14: Top 150 samples with highest Typicality OOD score for the CIFAR-10→SVHN setup. ID images from CIFAR-10 and OOD images from SVHN are marked with blue and orange borders, respectively. Samples flagged as most OOD by the Typicality metric are dominated by simple in-distribution images, illustrating the method's complexity bias and pronounced failure mode.

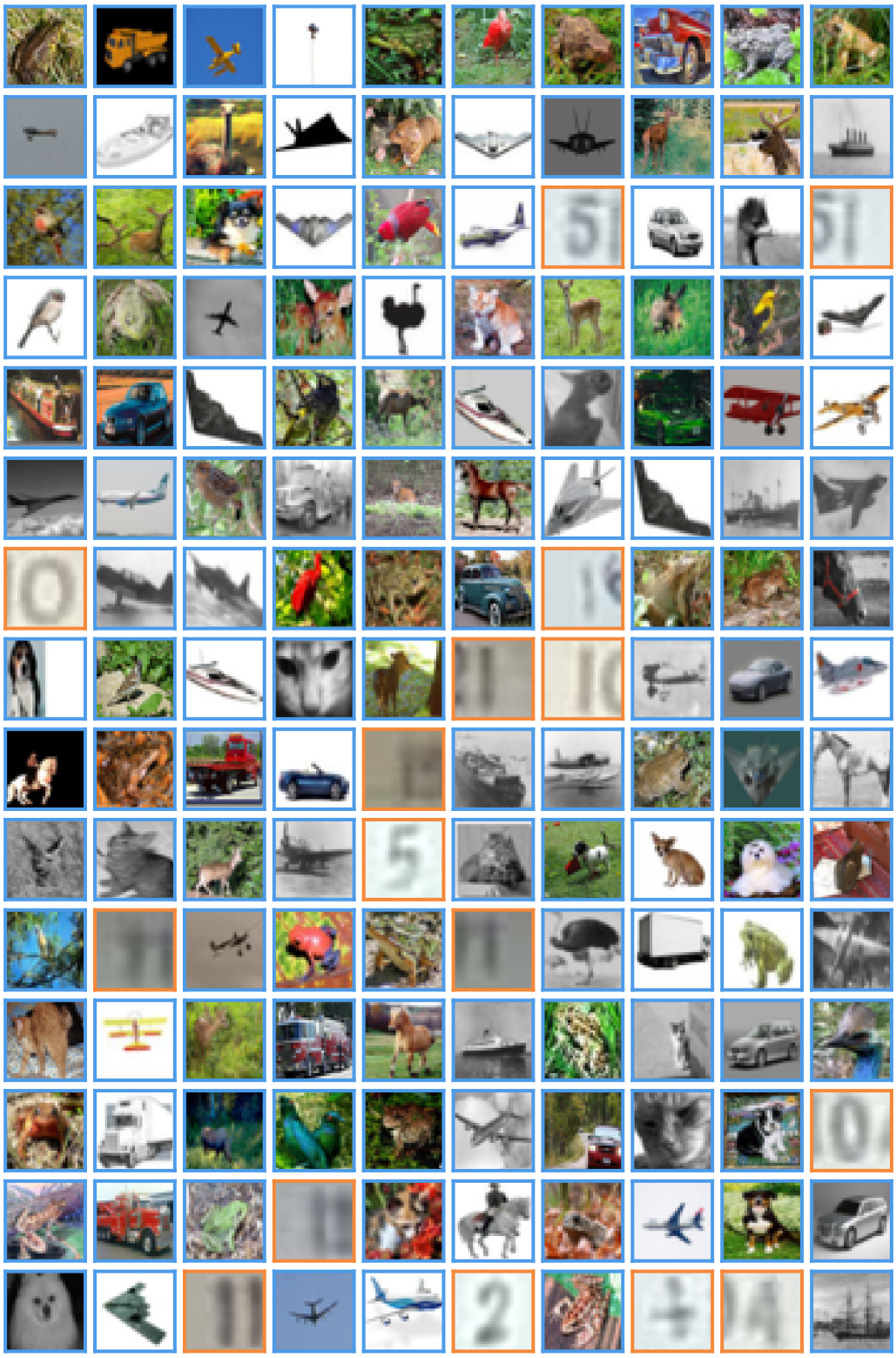


Figure 15: Top 150 samples with highest DoSE OOD score for the CIFAR-10→SVHN setup. ID images from CIFAR-10 and OOD images from SVHN are marked with blue and orange borders, respectively. Samples flagged as most OOD by DoSE are dominated by ID images. Shared failure cases with the likelihood and Typicality illustrate a persistent complexity bias and failure mode.

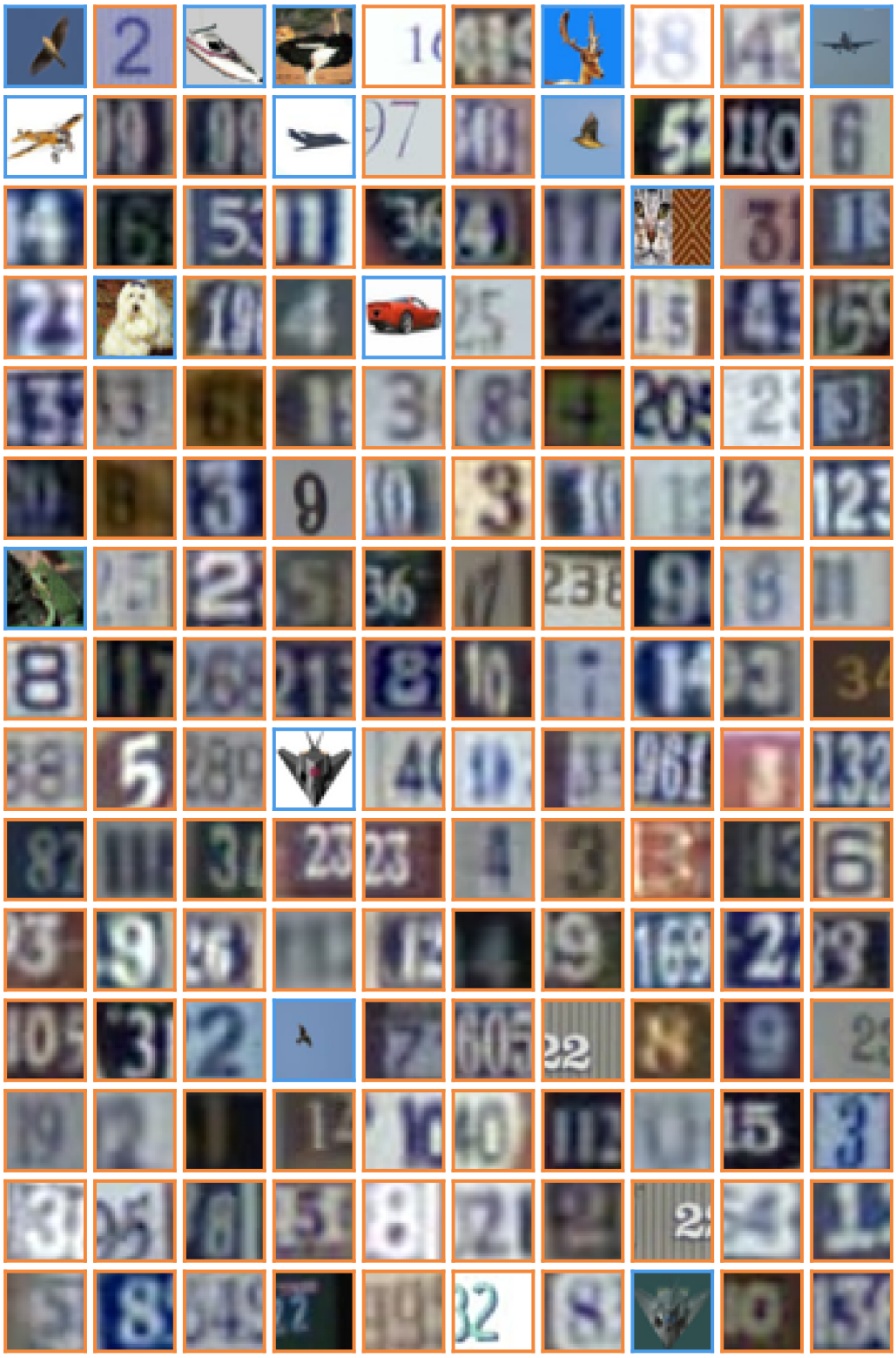


Figure 16: Top 150 samples with highest SITN OOD score for the CIFAR-10→SVHN setup. ID images from CIFAR-10 and OOD images from SVHN are marked with blue and orange borders, respectively. With a few exceptions, the highest SITN OOD scores successfully identify true out-of-distribution samples.