

---

# Density of States Estimation for Out-of-Distribution Detection

---

Warren R. Morningstar  
Google Research

Cusuh Ham  
Georgia Institute of Technology

Andrew G. Gallagher  
Google Research

Balaji Lakshminarayanan  
Google Research

Alexander A. Alemi  
Google Research

Joshua V. Dillon  
Google Research

## Abstract

Perhaps surprisingly, recent studies have shown probabilistic model likelihoods have poor specificity for out-of-distribution (OOD) detection and often assign higher likelihoods to OOD data than in-distribution data. To ameliorate this issue we propose DoSE, the *Density of States Estimator*. Drawing on the statistical physics notion of “density of states,” the DoSE decision rule avoids direct comparison of model probabilities, and instead utilizes the “probability of the model probability,” or indeed the frequency of any reasonable statistic. The frequency is calculated using nonparametric density estimators (e.g., KDE and one-class SVM) which measure the typicality of various model statistics given the training data and from which we can flag test points with low typicality as anomalous. Unlike many other methods, DoSE requires neither labeled data nor OOD examples. DoSE is modular and can be trivially applied to any existing, trained model. We demonstrate DoSE’s state-of-the-art performance against other unsupervised OOD detectors on previously established “hard” benchmarks.

## 1 Introduction

An important assumption behind the success of machine learning methods is that the data seen at test time follows a similar distribution to the training data. When a model encounters an anomalous, or out-of-distribution (OOD) input, it can output incorrect pre-

dictions with high confidence. Therefore, it is important to the reliability and safety of these systems to be able to recognize distributional shifts that are often present in real-world applications, such as autonomous driving and medical diagnoses.

The many proposed approaches to OOD detection can be broadly categorized into supervised and unsupervised methods. In a supervised setting, models have access to class labels and/or specific OOD examples, and are either calibrated *post hoc* to flatten the predictive distribution as the distance from the training set increases [Liang et al., 2018] or directly trained to distinguish in- and out-of-distribution examples [Hendrycks et al., 2019, Meinke and Hein, 2020, Dhamija et al., 2018].

In an unsupervised setting, generative models are often employed because of their ability to approximate or calculate the density  $q(X)$  that describes the distribution of the training set, which can then be used to determine when to trust the prediction  $q(Y|X)$ . Historically, this approach centers around interpreting this density as a probability of the input  $x$ , and therefore assuming OOD inputs would be assigned lower probability than in-distribution inputs, making them “less likely” to be in-distribution. However, Nalisnick et al. [2019a], Hendrycks et al. [2019] exposed some egregious failure modes of this methodology, such as OOD inputs being assigned higher likelihoods than in-distribution examples. Concurrent work by Choi et al. [2018] and follow-up work from Nalisnick et al. [2019b] showed that this failure occurs because the typical set of the data may not intersect with the region of high density.

Consider a high-dimensional isotropic Gaussian distribution with zero mean and unit variance. In Figure 1(a), we show a two-dimensional slice of this distribution. The mean of this distribution has the highest likelihood (red), but it is clearly not *typical* since the likelihood of draws concentrate on lower likelihoods, as shown in Figure 1(c). This phenomenon is a conse-

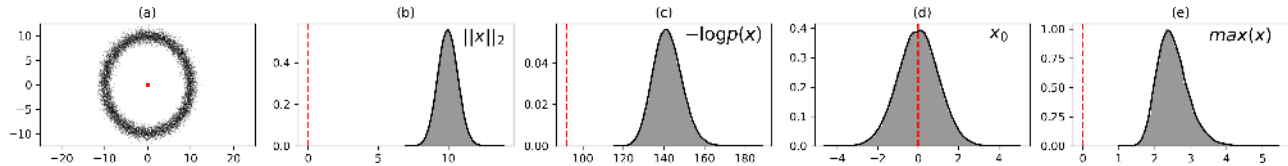


Figure 1: (a) A two-dimensional projection of a 100-dimensional multivariate normal distribution. The origin is shown in red. We show histogrammed measurements of 100,000 random draws from this distribution: (b) The observed norm of the draws, (c) the negative log-likelihood, (d) the value of the first coordinate, and (e) the max over the coordinates of the draws. The dashed vertical line denotes the corresponding measurement for the origin.

quence of the norm’s sensitivity to one large variance dimension. The Gaussian Annulus Theorem [Blum et al., 2020] formalizes this idea and stipulates that samples concentrate on the spherical shell of radius  $\sqrt{d}$ , as depicted in Figure 1(b).

These observations have parallels to physical systems. In statistical mechanics, the probability of observing a particle in a given state is governed not only by the probability of the state, but also by the geometry of the system. The density of states codifies this idea; it describes the number of configurations in the system which take on particular values of a given statistic. Figure 1(d, e) show how different statistics convey different information about the state. From this we hypothesize that the density of states—as measured by different statistics—might be potentially useful as a tool for identifying OOD data.

Our approach for identifying samples as being in- or out-of-distribution is to produce an estimator of this density of states on several summary statistics of the in-distribution data, and then to evaluate the density of states estimator (DoSE) on new trial points, marking those that have low support under the observed densities of the measurements as out-of-distribution. In general we expect and observe that a relatively small set of reasonable measurements of the samples works well at OOD detection. We summarize our contributions as follows:

1. We propose a novel OOD detection method, DoSE, inspired by ideas from statistical physics and the notion of typicality, that jointly leverages multiple summary statistics from generative models to differentiate between in-distribution and out-of-distribution data.
2. We show that two variants of DoSE can be easily applied to any pre-trained, generative model. Specifically, we evaluate DoSE with  $\beta$ -VAEs [Higgins et al., 2017] and Glow [Kingma and Dhariwal, 2018].
3. We evaluate our method on OOD detection benchmarks and demonstrate state-of-the-art performance among unsupervised methods, and com-

parable performance to state-of-the-art supervised methods.

## 2 Related Work

Given that modern large-scale neural networks can both be fooled by very small perturbations to their inputs [Szegedy et al., 2013], as well as make poorly-calibrated predictions [Guo et al., 2017], it is increasingly important for neural networks used in applications to be able to identify when it is asked to make predictions for out-of-distribution inputs.

**Unsupervised OOD Detection.** Bishop [1994] first proposed that generative models may be a useful tool for OOD detection, using a one-sided threshold on the log-likelihood as a decision rule. The underlying idea behind this approach is that the likelihood represents the “probability of the data,” and therefore a high likelihood means that the data is “good,” and a low likelihood is “bad.” They found that this approach, applied to a model trained on 4 classes was successful at detecting OOD data generated from a fifth class using 16-dimensional feature vectors. However, the success of this early approach may have been merely coincidental, or due to the fact that the model operated on a low-dimensional feature space. Notably Choi et al. [2018] pointed out that in extremely high dimensions, the previously held assumption that in-distribution inputs should have high likelihoods does *not* hold. This was concurrently validated empirically by Nalisnick et al. [2019a], Hendrycks et al. [2019], who showed that the log-likelihood returned by deep generative models can often be higher for OOD data than it is for in-distribution data. Serrà et al. [2019] suggested that the input complexity of the data may be responsible for this effect.

Subsequent work on unsupervised OOD detection has focused largely on ways to correct this pathology. For example, Choi et al. [2018] proposed that OOD data may receive higher likelihoods because of epistemic errors in the likelihood computation, and instead proposed to use the Watanabe-Akaike Information Criterion (WAIC), thereby leveraging multiple generative models trained in parallel to identify OOD data. Al-

ternatively, Ren et al. [2019] argue that the reason models assign high likelihoods to OOD data is instead because they are confounded by background information present in the dataset. Thus, they propose to use the likelihood ratio of an autoregressive model trained on in-distribution data with a heavily regularized model trained on mutated pixel data to try to normalize the likelihood by removing contributions from the “background pixels.” However, neither of these approaches address the issue with high-dimensional likelihoods, and therefore may be unreliable in broader applications. More recently, efforts have been made to attempt to directly measure the typicality of the input data. Nalisnick et al. [2019b] propose a simple typicality test by flagging a batch of data  $X$  if the mean of the generative model log-likelihood ( $\log q(\hat{X}|\theta_n)$ ) for that batch disagrees with the mean of  $q(X|\theta_n)$  in the training set by a user-specified threshold. There are two shortcomings to this approach: First, their test operates on an entire batch, for which all examples are assumed to be either jointly in-distribution or jointly OOD. Performance noticeably degrades as the batch size decreases to 1. For practical purposes we require a decision rule that can reliably operate on individual samples. Second, for both VAEs and flow-based models, the likelihood may not be the most informative metric, while its constituents or an alternative might (see Figure 1 or Appendix B). Finally, Li et al. [2019] appear to have created a special case of DoSE for malignant cancer detection. We believe that our methodology, specifically Theorem 3.1 and extensive empirical study, further justifies their domain-specific application of DoSE-like ideas. Additionally, we think the ideas presented in this paper could further improve their results but defer the evaluation of this hypothesis to future work.

**Supervised OOD detection.** There are also many proposed approaches to OOD detection that use labeled in-distribution inputs [Alemi et al., 2018, Hsu et al., 2020, Lakshminarayanan et al., 2017] and/or known OOD examples [e.g. Liang et al., 2018, Hendrycks et al., 2019, Stutz et al., 2019, Meinke and Hein, 2020]. All of these methods have demonstrated successful performance, but are trained with class labels and/or specific outlier examples.

In this work we do not use class labels or any exposure of OOD data to the model during training. This presents a significant practical advantage over these methods for several reasons: First, in many settings one may need to identify OOD data without being given class labels. Second, training specifically to predict class labels may otherwise discard information that may be useful when identifying OOD data (though it also may highlight information which *is* useful). Third,

models trained using specific instances of OOD data are overly tuned to attributes in the OOD dataset, and therefore may suffer from overconfident yet incorrect predictions when given inputs from a separate OOD set.

### 3 Approach

We first establish notation. Assume access to data generated according to  $\{X_i = x_i\}_i^n \stackrel{\text{iid}}{\sim} p(X)$  for  $X \in \mathcal{X}$  and that our task is to construct a summary statistic  $T_n$  suitable for evaluation on unseen data. Example summary statistics include  $T_n^{(\text{nl})}(X) \stackrel{\text{def}}{=} -\log q(X|\theta_n)$ ,  $T_n^{(L2)}(X) \stackrel{\text{def}}{=} \|X - \mu_n\|_2$ , or  $T^{(\text{ml})}(\{X, Y\}) \stackrel{\text{def}}{=} \max_Y q(Y|X, \theta_n)$ . Suppose, however, that each unseen sample datum is drawn from the mixture  $\hat{X} \sim \alpha p(\hat{X}) + (1 - \alpha)\tilde{p}(\hat{X})$  where  $\alpha$  and  $\tilde{p}$  are unfixed and unknown confounders and  $\tilde{X} \sim \tilde{p}$  has  $\tilde{X} \in \mathcal{X}$ . Our task is to devise a decision rule for identifying when  $T_n(\hat{X})$  is not to be trusted.

Since we presume  $\alpha, \tilde{p}$  are unfixed and unknown we can neither access OOD samples  $\{\tilde{X}_i\}_i^m$  nor make assumptions of  $\alpha, \tilde{p}$  when constructing  $T_n$ . Our only option is to devise a rule based solely on  $T_n$  and  $\{x_i\}_i^n$ . Our proposal—and indeed an obvious idea—is to fit a distribution to  $\{T(x_i)\}_i^n$  and use that probability as a threshold for classifying whether a sample is OOD. For example, assuming the statistic  $T$  is multivariate ( $T : \mathcal{X} \rightarrow \mathbb{R}^D$ ) one could use a product-of-experts (POE) kernel density estimator (KDE) of the form,

$$q(X = x|T, \{x_i\}_i^n, h) = \prod_d \frac{1}{nh_d} \sum_i \phi_j \left( \frac{[T(x)]_d - [T(x_i)]_d}{h_d} \right), \quad (1)$$

one-class SVM [Schölkopf et al., 2000], or any other similarly constructed density.

#### 3.1 What is a good $q$ in theory?

What makes for a good OOD distribution,  $q$ ? How do we choose the statistic  $T$  and associated density estimator hyperparameters when we have neither  $\tilde{p}$  nor samples from it? Similarly, why is direct use of the maximum likelihood distribution a generally poor OOD detector? [Nalisnick et al., 2019a]

To answer these questions, consider a slightly generalized notion of the information theoretic typical set,

$$\mathcal{A}_{p,q}^{(s,\epsilon)} = \left\{ \{X_i\}_i^s \in \mathcal{X}_p^s : \left| -\frac{1}{s} \sum_i \log q(X_i) - \mathbb{H}[p] \right| \leq \epsilon \right\}, \quad (2)$$

where  $\mathbb{H}[p]$  is the entropy of the true process  $p$ ,  $\epsilon$  governs the permissible entropic gap,  $s$  is the sequence

length, and  $q$  is any distribution over  $\mathcal{X}_p$ . Equation 2 generalizes the standard definition,  $\mathcal{A}_{p,p}^{(s,\epsilon)}$  [Cover and Thomas, 2012] by considering the typicality coverage on  $p$  by a possibly different distribution  $q$ . To make an effective OOD classifier, we are concerned with identifying the  $q$  which maximizes the expected typicality of  $q$  on  $p$ , i.e.,  $\max_{q \in \mathcal{Q}} p(X^s \in \mathcal{A}_{p,q}^{(s,\epsilon)})$ . Notably, our work is concerned with the  $s = 1$  case, i.e., capability for singleton OOD designation. Theorem 3.1 clarifies this objective by way of bound.

**Theorem 3.1.** *Bias/Variance Tradeoff for Typicality.*

$$p(\{X_i\}_i^s \notin \mathcal{A}_{p,q}^{(s,\epsilon)})\epsilon^2 \leq \text{KL}[p, q]^2 + \frac{1}{s} \text{Var}_p[\log q(X)] \quad (3)$$

*Proof.* Write  $Y = -\frac{1}{s} \sum_i \log q(X_i) - \text{H}[p]$ . From Markov’s inequality,  $p(|Y| \geq \epsilon)\epsilon^2 \leq \mathbb{E}_p[Y^2]$ . Making substitutions based on  $\frac{1}{i^2} \sum_i \text{Var}_p[\log q(X_i)] = \text{Var}_p[\frac{1}{s} \sum_i \log q(X_i)] = \mathbb{E}_p[(\frac{1}{s} \sum_i \log q(X_i))^2] - \text{H}[p, q]^2$  and  $\text{KL}[p, q]^2 = (\text{H}[p, q] - \text{H}[p])^2$  completes the proof.  $\square$

Through the lens of Theorem 3.1, we understand the MLE-fitted distribution’s shortcomings as an OOD probability measure. When  $q$  is chosen solely to minimize  $\text{KL}[p, q]$ , it will generally be a looser bound on the  $s = 1$  typical set—the case of interest when making single sample OOD evaluations. Likewise, many choices of  $T$  are also apparently sub-optimal. For example,  $T^{(42)}(x) = 42$  would generally be useless because any density  $q$  built solely from this statistic would have an infinite  $\text{KL}[p, q]$  unless  $p(X) = \delta(X - 42)$ . Also, we can generally rule out degenerate KDEs ( $h = 0$ ) because of their lack of smoothness, i.e., disregard for  $\text{Var}_p[\log q]$ .

### 3.2 What is a good $q$ in practice?

Although Theorem 3.1 is cognitively appealing, it is not directly computable owing to its nonlinear dependency on  $\text{H}[p]$  (an unknown). We now describe a heuristic procedure for minimizing the right-hand side of Theorem 3.1 and justify this procedure both by exploring a plugin estimate to Theorem 3.1 and by appealing to rationale from statistical physics.

We first note that it is possible to make coarse tuning to the OOD detecting distribution  $q$  via crude approximations to Theorem 3.1’s implications. The empirical approximation of the right-hand of Theorem 3.1 over the held-out distribution  $\{x_i\}_i^m$  is,

$$\begin{aligned} \text{KL}[p, q]^2 + \text{Var}_p[\log q] &\approx \frac{1}{m} \sum_j^m (\log q(x_j | \{x_i\}_i^n, T, \gamma))^2 \\ &+ 2 \text{H}[p] \frac{1}{m} \sum_j^m \log q(x_j | \{x_i\}_i^n, T, \gamma) + c \quad (4) \end{aligned}$$

where  $m$  is the size of the evaluation set,  $\gamma$  represents the parameters of our density (e.g.,  $h$  for a KDE and  $\nu$  for one-class SVM), and  $c$  is constant for any choice of  $q$ . A general strategy to minimize 4 is to consider several different choices for  $\text{H}[p]$  and explore different choices of  $T$  under this range. Alternatively, one can consider using  $\text{H}[q(X|\theta_n^{(ml)})]$  as a plugin estimate for  $\text{H}[p]$ . This is the “resubstitution estimator” introduced by Beirlant et al. [1997] and used by Nalisnick et al. [2019b]. Assuming the OOD distribution has the same discrete support, one can additionally explore use of entropy bounds like  $\text{H}[p] \in [0, h \cdot w \cdot c \cdot \log k]$  (for image height  $h$ , width  $w$ , channels  $c$ , and discrete pixel intensity levels  $k$ ) or use or use known estimates of  $\text{H}[p]$ , e.g., Parmar et al. [2018]. We emphasize that discrete entropy is only reasonable if  $q$  also has the same support; failing this requirement may introduce an inconsistent sign in Equation 4.

We use equation 4 and the resubstitution estimator for the entropy to evaluate the tightness of the bound for all different statistics. For a statistic which is completely informative about the typicality (i.e., it minimizes the bound from 3.1), one need only evaluate that statistic to evaluate the typicality of trial points and identify those which are out-of-distribution. In practice we find that multiple different statistics get indistinguishable values for this bound, and therefore we do not know which statistic is the most informative. We therefore construct our estimator based on the KDE of multiple different statistics evaluated on the same data. The procedure is straightforward: If we interpret the KDE estimates as probabilities of typicality, then the product of the KDEs gives the probability that a given input is typical for all metrics jointly (assuming no correlation between statistics). We can further relax the assumption of independence by jointly evaluating the DoSE using an alternative density estimator, such as a one-class Support Vector Machine [SVM; Schölkopf et al., 2001] in our case. We show in the experiments that both of these approaches outperform alternative methods, which only query a single statistic. We further show in Section 5 that both of these approaches are robust against the inclusion of uninformative, or even obfuscatory statistics.

Our procedure to construct DoSE for OOD detection is as follows:

1. Train a deep probabilistic model  $q(X|\theta_n)$  using training set  $\{x_i\}_i^n$  where  $n$  is the size of a training set from which  $m$  samples (chosen randomly) are excluded from training and used as a validation set.
2. Evaluate summary statistics  $T_n(x)$  on the training data.

3. Construct DoSE using a KDE or SVM on each set of statistics from the training set.
4. Evaluate the DoSE score by computing the sum of the log-probabilities from the KDE on each statistic for each example in the training set  $\{x_i\}_i^n$  and validation set  $\{x_i\}_i^m$ . Alternatively, compute the scores for both sets using the SVM.
5. Check the DoSE calibration between the training and validation DoSE scores using the expected calibration error (ECE) [Guo et al., 2017].
6. Determine threshold for OOD rejection, by choosing a number of examples to discard from the validation set, and identifying the corresponding threshold to place on the DoSE score.

We now establish intuition to further explain the underpinnings of our empirical methodology. In statistical physics, a system contains particles  $x$ . For each particle, a measurement or statistic  $T_n$  represents a physical property of that particle. Our challenge is to determine if any given particle is atypical, using only the physical properties of that particle, along with the physical properties of  $n$  particles from the system. Atypicality here means that the particle should not be found having these properties assuming that the system is in equilibrium (i.e. the particle is an anomaly). For any physical property (e.g., energy), the probability of occurrence in a physical system is determined by the *density of states*:  $g(T) = \int dX \delta(T'(X) - T)$ . This quantity describes the number of occupied configurations in the system which have a given value of  $T$ .

One can often calculate the statistical physics notion of density of states from first principles. Since this is not possible in our problem setup, we instead simply approximate the density of states using a *Density of States Estimator* (DoSE): a nonparametric density estimator trained to measure the density of states of a statistic  $T$  evaluated on an input  $\check{X}$  using the sample particles from the system. We can apply this approach towards any statistic  $T$  evaluated on the data  $\{x_i\}_i^n$  to construct the DoSE of that statistic. DoSE then measures the empirical density of the statistic  $T$  evaluated at some new point  $\hat{x}$  using nearby points in the training set. Note that we need not offer any interpretation for  $T$ , and even if the statistic is not interpretable, we can still measure its typicality.

## 4 Experimental Setup

To evaluate the empirical performance of DoSE, we follow the procedures outlined in Choi et al. [2018], Nalisnick et al. [2019b], Ren et al. [2019]. To summarize, we first train an ensemble of deep generative

models on a given in-distribution dataset. We then evaluate statistics on examples from the training set and construct our DoSEs using the measured statistics. We validate that our models are not memorizing using a heldout set of examples from the training set. We finally compute the DoSE scores on the in-distribution test set, and several OOD datasets. We measure the success of OOD identification using the Area Under the ROC Curve (AUROC).

We compare our performance against several established unsupervised baselines:

1. A single-sided threshold on the log-likelihood  $q(X|\theta_n)$  [Bishop, 1994].
2. The *single-sample* typicality test (TT) from Nalisnick et al. [2019b]. To evaluate the AUROC using this method, we simply use the raw typicality score  $\text{TT}(\check{X}) = |\log q(\check{X}|\theta_n) - \mathbb{H}[q(X|\theta_n)]|$ . Similar to Nalisnick et al. (2019), we calculate  $\mathbb{H}[q(X|\theta_n)]$  as an empirical average over the training set.
3. The *Watanabe-Akaike Information Criterion* (WAIC) from Choi et al. [2018]. For this, we use 5 models trained separately and measure  $\text{WAIC}(\check{X}) = \mathbb{E}_\theta[\log q(\check{X}|\theta_n)] - \text{Var}_\theta[\log q(\check{X}|\theta_n)]$
4. The likelihood ratio method (LLR) from Ren et al. [2019]. To compute LLR, we train a background model using their proposed method of mutations, using a mutation rate of 0.15, the center of the range in which they found successful results. The LLR score is then simply  $\text{LLR}(\check{X}) = \log q_s(\check{X}|\theta_n) - \log q_b(\check{X}|\theta_n)$ , where the subscripts  $s$  and  $b$  indicate the semantic and background models, respectively.

For all of these methods, we use the *same* models to evaluate the OOD scores. This highlights the difference in performance caused by the methodology, rather than due to differences in the training procedure. To quantify the uncertainty in performance resulting from the parameters  $\theta$  found during an individual training run, we train 5 separate models in parallel, and evaluate the performance of *all* methods using all models.

For DoSE on  $\beta$ -VAEs, we used 5 statistics:

1. Posterior/prior cross-entropy,

$$T_n^{(\text{xent})}(X) = \mathbb{H}[q(Z|X, \theta_n), q(Z)],$$

2. Posterior entropy:

$$T_n^{(\text{ent})}(X) = \mathbb{H}[q(Z|X, \theta_n)],$$

3. Posterior/prior KL divergence:

$$T_n^{(\text{rate})}(X) = \text{KL}[q(Z|X, \theta_n), q(Z)],$$

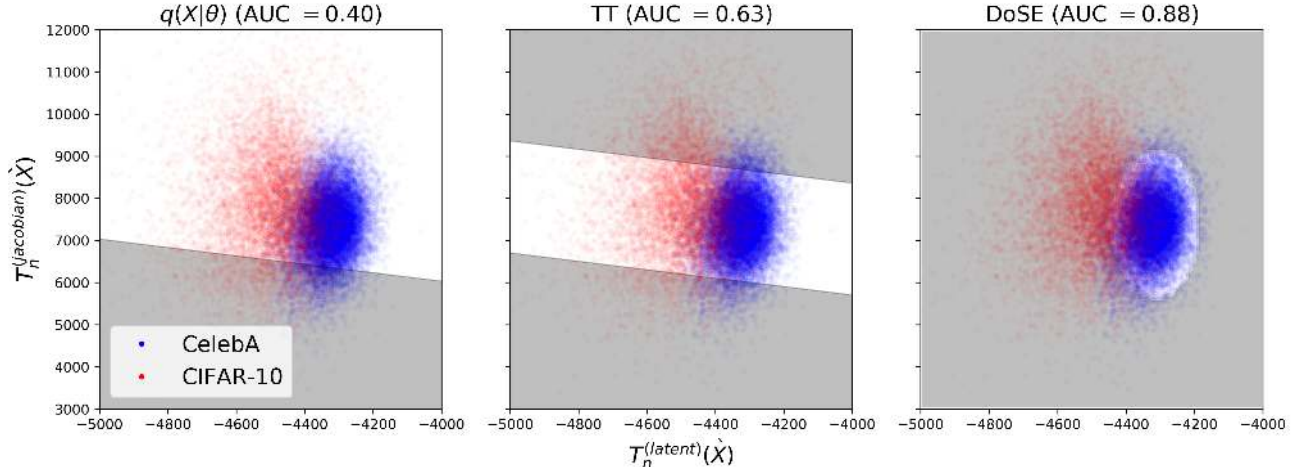


Figure 2: Decomposition of  $q(X|\theta_n)$  for a Glow model trained on CelebA. The blue points show the test data, in coordinates of  $T_n^{(\text{latent})}(\dot{X})$  and  $T_n^{(\text{jac})}(\dot{X})$ . Red points show the same coordinates observed for CIFAR10, an OOD dataset. We show the decision boundaries that exclude 10% of the in-distribution data for  $q(X|\theta_n)$  (left), TT (middle) and DoSE (right). The shaded (gray) region is classified as out-of-distribution, and the non-shaded region is classified as in-distribution.

4. Posterior expected log-likelihood:

$$T_n^{(\text{distortion})}(X) = \mathbb{E}_{q(Z|X, \theta_n)}[\log q(X|Z, \theta_n)],$$

5. IWAE [Burda et al., 2015]:

$$T_n^{(\text{iwae})}(X) = \log \mathbb{E}_{q(Z|X, \theta_n)}[q(X, Z, \theta_n)/q(Z|X, \theta_n)].$$

In all cases, the intractable expectation  $\mathbb{E}_{q(Z|X, \theta_n)}[f(Z)]$  was replaced with a seeded Monte Carlo approximation,  $\frac{1}{16} \sum_t f(Z_t)$  with  $Z_t \stackrel{\text{iid}}{\sim} q_{\text{post}}(Z|X, \theta_n, \text{seed}=\text{hash}(X, t))$ . By seeding, we ensure the statistics’ reproducibility yet preserve the logic of the approximation. For Glow models, we used 3 statistics:

1. Log-likelihood  $T_n^{(\text{like})}(X) = q(X|\theta_n)$ ,
2. Log-probability of the latent variable  $T_n^{(\text{latent})}(X) = q(Z|X, \theta_n)$ ,
3. Log-determinant of the Jacobian between  $X$ ; the input space, and  $Z$ ; the transformed space (i.e.,  $T_n^{(\text{jac})}(X) \stackrel{\text{def}}{=} \log |J(X)|$ ).

Additional model and training details are in Appendix C.

## 5 Results

A summary of all quantitative results on all baselines is presented in Table 1. We show the AUROC computed between all pairs of in- and out-of-distribution data, measured using our method as well as alternative techniques.

**DoSE vs related methods** We find that, for all “hard” dataset pairings, both variants of DoSE either outperform or significantly outperform all competing methods. Note that this same result is observed for either DoSE evaluated on an individual model or on a full ensemble of models. For an individual model, we observe that all 5 runs of DoSE<sub>KDE</sub> outperform all 5 runs of all competing techniques. This corresponds to a probability of 0.003 that our result was observed due to random chance, compared against any competing technique. We further find that our method generally outperforms competing techniques on most easy dataset pairings as well, with a few exceptions (e.g., SVHN→CIFAR10), which are typically found by a one-sided threshold on the likelihood  $q(X|\theta_n)$ . While DoSE may not then be the highest performing technique in all dataset pairings, it is important to note that it is the highest performing overall, with an average ranking of 1.2 for both DoSE<sub>KDE</sub> and DoSE<sub>SVM</sub> against other competing techniques (we exclude the other when computing the ranking). For reference,  $q(X|\theta_n)$  has rank of 2.2, TT 2.7, WAIC 3.06, and LLR 4.19.

In general, we find that TT achieves more reliable performance than the alternatives. This is, on some level, to be expected because TT also attempts to directly measure the typicality of a datum. However, we also find several situations where TT is vulnerable because it relies exclusively on the likelihood. In particular, we find that TT achieves only AUROC  $\approx 0.6$  when trying to identify CIFAR10 or CIFAR100 when trained on CelebA. For GLOW, the log-likelihood  $T_n^{(\text{like})}(X)$  is itself a sum of two different statistics  $T_n^{(\text{latent})}(X)$  and  $T_n^{(\text{jac})}(X)$ . Figure 2 shows scatter

Dataset/OOD Dataset	Model	$q(X \theta_n)$	WAIC	TT	LLR	DoSE <sub>KDE</sub>	DoSE <sub>SVM</sub>
MNIST	VAE						
Omniglot		1.000	1.000	1.000	0.470	1.000	<b>1.000</b>
FashionMNIST		0.998	0.988	0.997	0.404	<b>0.999</b>	0.996
Uniform		1.000	1.000	1.000	0.277	<b>1.000</b>	<b>1.000</b>
Gaussian		1.000	1.000	1.000	0.228	<b>1.000</b>	<b>1.000</b>
HFlip		0.839	<b>0.861</b>	0.776	0.473	0.760	0.812
VFlip		<b>0.838</b>	0.821	0.837	0.499	0.818	0.830
FashionMNIST	VAE						
Omniglot		0.995	0.893	0.991	0.508	<b>1.000</b>	0.998
MNIST		0.931	0.950	0.901	0.503	<b>0.998</b>	0.997
Uniform		0.998	0.878	0.998	0.573	<b>1.000</b>	0.998
Gaussian		0.997	0.852	0.997	0.501	<b>1.000</b>	0.998
HFlip		0.658	0.503	0.599	0.479	<b>0.658</b>	0.625
VFlip		0.702	0.473	0.635	0.485	<b>0.748</b>	0.728
CIFAR10	Glow						
CIFAR100		0.520	0.532	0.548	0.520	0.569	<b>0.571</b>
CelebA		0.914	0.928	0.848	0.914	0.976	<b>0.995</b>
SVHN		0.064	0.143	0.870	0.064	<b>0.973</b>	0.955
ImageNet32		0.794	0.870	0.754	0.795	0.914	<b>0.930</b>
Uniform		1.000	1.000	1.000	1.000	<b>1.000</b>	<b>1.000</b>
Gaussian		1.000	1.000	1.000	1.000	<b>1.000</b>	<b>1.000</b>
HFlip		0.501	0.499	0.500	0.501	<b>0.507</b>	0.502
VFlip		0.505	0.505	0.501	0.505	<b>0.533</b>	0.523
SVHN	Glow						
CelebA		1.000	0.991	1.000	0.912	<b>1.000</b>	<b>1.000</b>
CIFAR10		<b>0.990</b>	0.802	0.970	0.819	0.988	0.962
CIFAR100		<b>0.989</b>	0.831	0.965	0.779	0.986	0.965
ImageNet32		0.998	0.980	0.994	0.916	0.999	<b>0.999</b>
Uniform		1.000	1.000	1.000	1.000	<b>1.000</b>	<b>1.000</b>
Gaussian		1.000	1.000	1.000	1.000	<b>1.000</b>	<b>1.000</b>
HFlip		0.504	0.502	0.499	0.502	<b>0.520</b>	0.512
VFlip		0.502	0.504	0.500	0.501	0.510	<b>0.511</b>
CelebA	Glow						
CIFAR10		0.404	0.507	0.634	0.323	0.861	<b>0.949</b>
CIFAR100		0.427	0.535	0.671	0.357	0.867	<b>0.956</b>
SVHN		0.008	0.139	0.982	0.028	0.993	<b>0.997</b>
ImageNet32		0.705	0.837	0.775	0.596	0.995	<b>0.998</b>
Uniform		1.000	0.961	1.000	1.000	<b>1.000</b>	<b>1.000</b>
Gaussian		1.000	1.000	1.000	1.000	<b>1.000</b>	<b>1.000</b>
HFlip		0.600	0.754	0.526	0.529	0.945	<b>0.985</b>
VFlip		0.706	0.734	0.602	0.606	0.983	<b>0.998</b>

Table 1: A comparison of AUROC of our method against unsupervised baselines on the OOD detection task. We find that our method most reliably achieves SoTA performance across all datasets.

plot of the two components of the log-likelihood. Likelihood uses a one-sided test classifying the region  $T_n^{(\text{latent})}(X) + T_n^{(\text{jac})}(X) = T_n^{(\text{like})}(X) \leq \tau$  as OOD. TT uses a two-sided test on the likelihood but still cannot separate the distributions well as it relies on a single statistic. DoSE uses multiple statistics to identify OOD data, hence it achieves much higher AUROC.

We find that  $q(X|\theta_n)$ , WAIC, and LLR all exhibit performance that is much less consistent for different dataset pairings. In part, we attribute this to the fact that none of these methods attempt to measure the typicality of an input, and are therefore vulnerable to OOD datasets which are assigned anomalously high likelihoods. As such, all of these methods fail on CIFAR10→SVHN, CelebA→CIFAR10/100, CelebA→SVHN. For LLR, we may also violate the implicit assumptions underlying the methodology by using models such as VAEs, which may not be able to explicitly decompose the likelihood into semantic and background components in the same way autoregressive models do. We therefore speculate that LLR may be more successful if a different model were used, though we also note that it still would not measure typicality.

**Qualitative analysis** We also perform a qualitative examination of the attributes in the data which appear to be most indicative of the OOD score from each method. To do so, we take the 16 images with the highest and lowest OOD scores from a given in- and out-of-distribution pairing for a given method. These images correspond to the 4 elements of the confusion matrix for each method. We organize these images into their respective category in the confusion matrix, and show the results for TT, WAIC and DoSE in Figure 3 on CIFAR-10→SVHN. While it is difficult to provide an entirely objective assessment of these results, we speculate that DoSE identifies images with high color contrast as likely OOD candidates. TT appears to identify a split between images with uniform backgrounds, and images with noisy backgrounds as false positives. This makes sense, given that these are the images with highest and lowest log-evidence, respectively. WAIC appears to identify images with irregular colors as likely OOD candidates. Of course, despite its reasonable qualitative results, WAIC also gets 0.06 AUROC on this particular dataset pairing, undermining its utility as an OOD detection method.

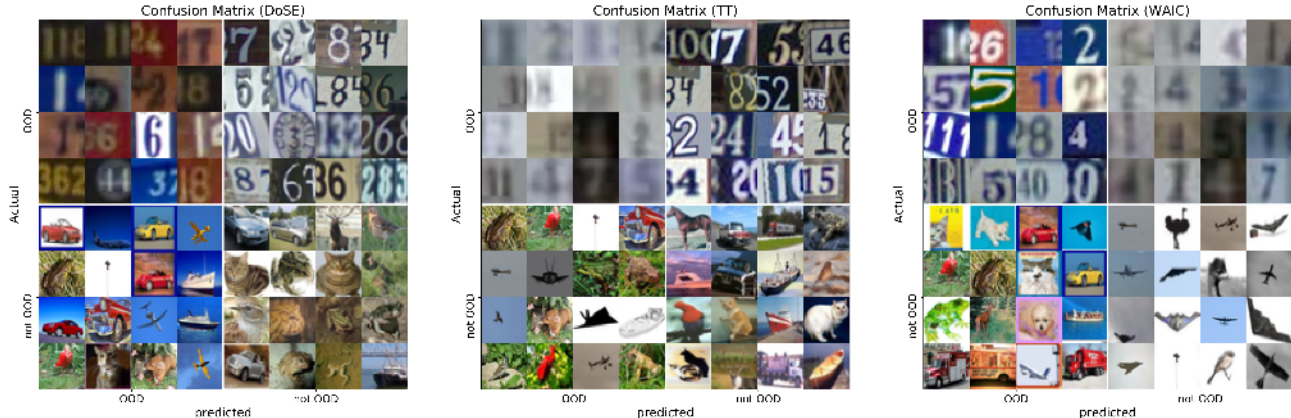


Figure 3: Confusion matrices for methods used in our OOD detection performance. The images in each quadrant of the matrix are in raster order, sorted by the confidence of the classification.

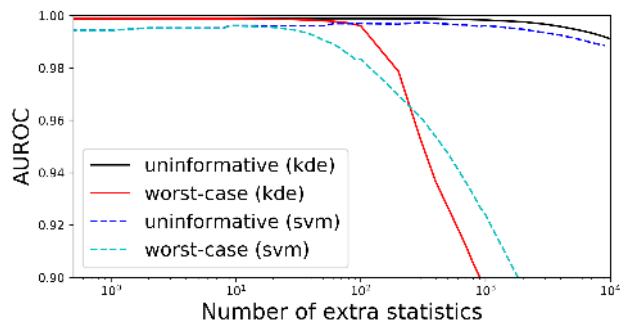


Figure 4: The AUROC observed when using DoSE on FashionMNIST→MNIST with the addition of extra statistics, which are either uninformative, or purposely obfuscatory. We find that uninformative statistics have little effect on performance, with only a 1% drop in the AUROC even after the addition of  $10^4$  uninformative statistics. Performance degrades by the same amount when roughly 80 obfuscatory statistics are used.

**Robustness to choice of statistics** In our experiments, we used statistics which were useful diagnostics of the model performance, and which we therefore expected to contain some degree of meaningful signal for OOD detection. When deploying DoSE on different types of models, one may not always have access to these same statistics or be tempted to choose as many statistics as are available. As we showed in Section 1, certain statistics are not be able to identify certain OOD datasets as atypical. A question we wanted to probe experimentally was then; “How hazardous are uninformative statistics for the OOD signal?” Since we do not have access to OOD data during training, answering this question will allow us to be slightly more liberal with choosing statistics.

For this experiment, we chose to use the FashionMNIST→MNIST pairing. We took the  $DoSE_{KDE}$  scores evaluated on the FashionMNIST and MNIST test sets. We then added “superfluous”

statistics, given by  $T^{(useless)} \sim \mathcal{N}(0, 1)$ , which was distributed identically for both the in-distribution and OOD data. We repeatedly drew more of these useless statistics, and added their DoSE scores to the test and OOD DoSE scores. We also further consider a worst-case statistic, for which OOD data is given maximally typical scores ( $-\log \sqrt{2\pi}$  for the unit-normal distribution) but in-distribution data is given  $T^{(useless)}$ .

We show the AUROC as a function of the number of superfluous statistics in Figure 4. We find that even after adding an extremely large number of superfluous statistics (at least  $3 \times 10^5$ ), the AUROC has only decayed by 0.04, meaning DoSE would still have higher AUROC than any competing technique. This phenomenon is observed using both a KDE and a SVM to evaluate the DoSE scores. In the worst case scenario, as expected we find that the number of statistics needed to degrade the OOD signal is much smaller, requiring only 100 statistics to produce noticeable degradation for the KDE, and only roughly 20 for the SVM. Even here, we find that roughly 300 statistics are necessary to drop the DoSE performance below alternative methods. Empirically this suggests that there may not be a strong need to carefully choose statistics. Further insight may be found in the literature of Approximate Bayesian Computation or Likelihood-Free Inference, which have an abundance of prior work dedicated towards choosing meaningful summary statistics [e.g., Fearnhead and Prangle, 2012].

## 6 Conclusion

We have presented a novel method, DoSE, for detecting out-of-distribution data, which can be easily applied to any pre-trained generative model or ensemble of generative models without any additional tuning or modification. We show that this approach is advantageous over likelihood-based approaches because it



provides multiple ways of evaluating the typicality of an input under the assumed generative model. DoSE does not require class labels or access to specific OOD examples. Leveraging the argument that likelihoods should not be interpreted as the probability that an input is in- or out-of-distribution as well as ideas from statistical physics, our method uses nonparametric density estimators to directly measure the typicality of various model statistics given the training data. We demonstrated state-of-the-art performance with DoSE among unsupervised methods on common OOD detection benchmarks.

## References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- Alexander A Alemi, Ian Fischer, and Joshua V Dillon. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.
- Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.
- Christopher M Bishop. Novelty detection and neural network validation. *IEEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994.
- Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.
- Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.
- Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.
- Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.
- Akshay Raj Dhamija, Manuel Günther, and Terrance Boulton. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*, pages 9157–9168, 2018.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.
- Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: Semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.
- Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=HyxCxhRcY7>.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vaе: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.
- Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. *arXiv preprint arXiv:2002.11297*, 2020.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.
- Xingyu Li, Marko Radulovic, Ksenija Kanjer, and Konstantinos N Plataniotis. Discriminative pattern mining for breast cancer histopathology image classification via fully convolutional autoencoder. *IEEE Access*, 7:36433–36445, 2019.

- Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=H1VGkIxRZ>.
- Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=ByxGkySKwH>.
- Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019a. URL <https://openreview.net/forum?id=H1xwNhCcYm>.
- Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 2019b.
- Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14680–14691, 2019.
- Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.
- Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, July 2001. ISSN 0899-7667. doi: 10.1162/089976601750264965. URL <https://doi.org/10.1162/089976601750264965>.
- 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. *arXiv preprint arXiv:1909.11480*, 2019.
- David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. *CoRR*, abs/1910.06259, 2019.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.
- Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: <https://doi.org/10.1038/s41592-019-0686-2>.