Flexible Distribution Shift and Outlier Detection with Self-Supervised Kernels

Anonymous CVPR 2021 submission

Paper ID 10871

Abstract

Existing methods for detecting anomalies and distribution shifts require training on reference data from the nominal distribution, which can be difficult to estimate from small datasets or in non-stationary conditions. To avoid any training on the reference dataset, we propose generic kernels for distribution shift and outlier detection obtained from self-supervised features on ImageNet. Surprisingly, even though it does not specialize by training on any particular reference dataset, our method still significantly outperforms state-of-the-art methods as well as supervised kernels on a variety of datasets for both tasks.

1. Introduction

Changes in the input distribution of machine learning systems can significantly degrade performance [34, 33] and represent a key challenge in safety-critical applications [40, 13]. Many techniques have thus been introduced to detect distribution shifts from a reference distribution (typically the one used to train the model), either based on a group of samples (*i.e.*, two-sample tests) [12, 25] or just a single one (*i.e.*, outlier/anomaly/OOD detection) [18, 19, 39].

Current approaches for two-sample test and anomaly detection are typically trained on data from the reference distribution. Intuitively, they attempt to identify properties (*e.g.*, mean, variance, etc) and representations (*e.g.*, class structure) of the reference data that can be useful to detect shifts or anomalies. This process, however, can be challenging for high-dimensional data (such as images) due to the curse of dimensionality. In practice, if the size of the reference dataset is small, performance can drastically degrade, as we demonstrate empirically in Section 5. This can be especially problematic in non-stationary settings where the reference dataset is constantly changing, as additional training is required to adapt to the new reference.

In a departure from existing approaches, we propose to detect distribution shifts using *generic* representations obtained from a *separate but potentially much larger* visual dataset (Figure 1). Inspired by the excellent transfer learning performance of self-supervised features [16, 3, 2, 14], we pre-train a deep network on a large and diverse dataset (ImageNet) in a self-supervised manner. We then use these *fixed* representations to detect shifts by comparing a reference dataset to one or more query points. Specifically, we use the pre-trained network as a feature extractor and define kernels for two-sample test and outlier detection tasks. Self-supervised learning allows us to use large *unlabeled* datasets for pre-training, enabling kernels parameterized by deeper and more expressive models without further tuning.

Empirically we find that for two sample test our approach performs significantly better than existing works on challenging benchmarks like CIFAR-10 vs CIFAR-10.1 and ImageNet vs ImageNet v2. Using just 500 samples for each of CIFAR-10 and CIFAR-10.1 we get average test power of 1.0, while the prior state-of-the-art achieves a test power of 0.71 even after using the entire CIFAR-10.1 data (1k samples). We also compare against features from the same backbone network, but one which has been trained in a supervised fashion over ImageNet, and find that the self-supervised features outperform the supervised features.

For OOD detection our method outperforms existing supervised/unsupervised approaches over 5 different benchmark datasets using only $5,000$ samples per in-distribution dataset. We also compare against prior works that make use of even larger datasets (ImageNet-22K) for OOD detection and find that our approach outperforms them too even when using only a relatively smaller dataset (ImageNet-1k). On datasets significantly different from ImageNet, such as CelebA and Chest X-Ray, our approach still achieves good OOD detection performance. Despite the significant gains in terms of performance, our approach is more computationally and time efficient than the prior works as it does not require training on new reference datasets.

To summarize, our main contributions are listed below.

- We propose novel approaches to two-sample tests and anomaly detection, where no training is needed over the reference datasets.
- We establish new state-of-the-art distributional shift detection performance over challenging benchmarks.

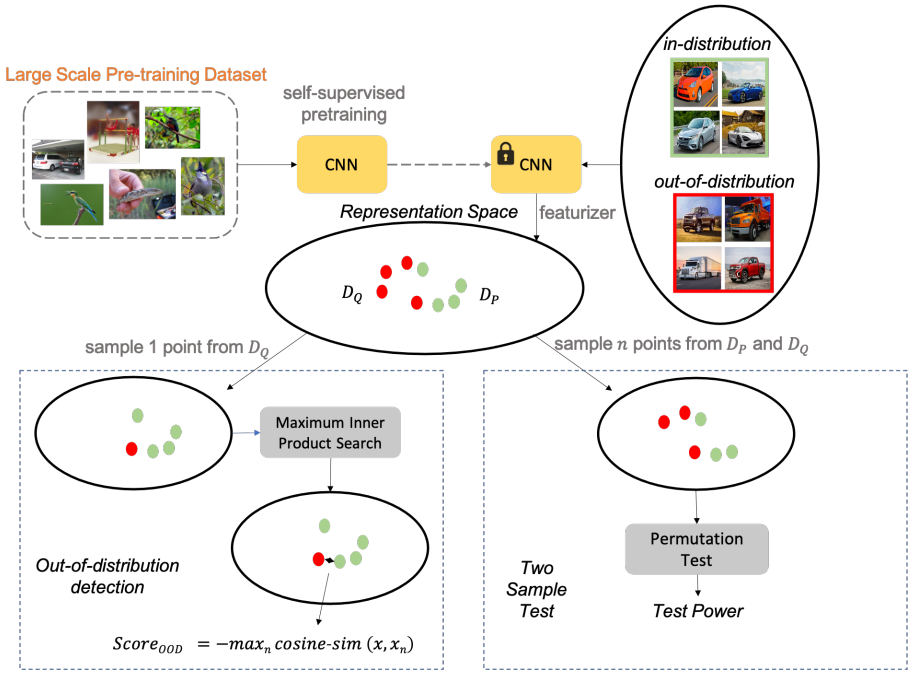

Figure 1: A summary of our approach. Methods for two-sample tests and out-of-distribution detection typically requires training on the reference datasets; instead, we propose to adopt kernels obtained from self-supervision on large vision datasets that transforms images into representations (dots in this figure); this approach avoids having to train on references datasets and achieves much superior performance. (**Bottom right**) We perform two sample tests over self-supervised features (green and red dots); (**Bottom left**) We perform out-of-distribution detection by measuring the maximum inner product between the test feature (red) and the reference features (green).

- For two sample test we find self-supervised features to outperform the supervised features. We experimentally justify possible reasons for this gain.
- Using only 5k samples per reference dataset we outperform existing supervised/unsupervised approaches for OOD detection over 5 datasets, even outperforming the approaches that use additional data.

2. Related Work

Self-Supervised Learning. Self-supervised learning is used to learn good representations from raw data (without requiring any expensive and laborious annotations) that can be useful for downstream tasks such as image classification, image segmentation, etc. Several prior works define novel pre-text tasks [11, 42, 31] in an effort to learn good representations from the data. [22] showed that the performance of these simple approaches can be further boosted by using deeper and wider networks.

Most of the recent approaches to self-supervised learning are based on contrastive learning [15, 2]. These methods are based on maximizing a lower bound of the mutual information using InfoNCE loss [32]. SimCLR [2] proposed the use of a non-linear projection head on top of representations before applying the InfoNCE loss. Additionally, they showed that using stronger augmentations than those used for supervised learning drastically improved performance. BYOL [14] is another new algorithm for self-supervised learning of image representations which learns its representation by predicting previous versions of its outputs, without using negative pairs.

Two Sample Tests. Given two sets of samples each drawn independently from a separate distribution, two sample tests determines if difference between the two distributions is statistically significant. Kernel [37] based methods construct a mean kernel embedding [1, 28] for each distribution and then measure the difference between the two embeddings given by the Maximum Mean Discrepancy (MMD) [12] distance. Classifier based approaches [26, 4], instead of learning kernels, train a binary classifier over the training set, and then use the classification performance over the validation set as a metric to perform the two sample test.

[25] propose paramterizing kernels by deep-neural networks, so that complex kernels can be learned for high dimensional data. Liu et al. [25] show the importance of learning good kernels, where they show that simple kernels such as Gaussian or Laplace map often map distinct distri-

butions to nearby mean embeddings. Such simple kernels do well in small datasets, but not so much on high dimensional data typically used in vision tasks. Another class of two-sample tests are based on leveraging a graphical structure from the data samples of a distribution. [8] leverage such graphical tests as a loss function to learn implicit generative models.

In our work also, we use the MMD distance [12] to measure the distance between the two mean kernel embeddings. However, in contrast to all the existing works, we do not learn a kernel over the visual data, but rather use a pretrained network's features to define a kernel. We also leverage the graph tests in [8] to compute distance between distributions and do not use them to train a generative model.

Out-of-distribution detection. Out-of-distribution detection refers to the task of identifying anomalous data points. Several prior works address out-of-distribution detection by assigning anomaly or OOD scores to inputs. Recent works can be categorized as:

Classifier-Based Methods. [17] show that a deep, pretrained classifier has a lower maximum softmax probability on anomalous examples than in-distribution examples. [7] define a new OOD score based on an auxiliary branch attached onto a pre-trained classifier. Outlier exposure [18] propose a simple technique to use out-of-distribution samples (from an auxiliary dataset) during training of classifier, which achieves a very high OOD detection score. [24] present a method to make the maximum softmax probability approach more discriminative between in-distribution and OOD samples by pre-processing input data with adversarial perturbations. [23] train a classifier concurrently with a GAN, and the classifier is trained to have lower confidence on GAN samples.

Density-Based Methods. Density-based methods are a class of classical methods for OOD detection which directly use the likelihood of the sample as the detection score. However, recent studies reveal that the likelihood is often not the best metric - especially for deep neural networks with complex datasets [29]. Several work thus proposed modified scores, e.g., typicality [30], WAIC [5], likelihood ratio [35], input complexity [38], or unnormalized likelihood (i.e., energy) [9].

Self-Supervised Methods. Recent self-supervised approaches show outstanding results on various OOD detection benchmark datasets. [19] supplement current supervised methods with an auxiliary rotation loss and find that such self-supervision can drastically improve out-of-distribution detection on difficult, near-distribution anomalies. [39] propose an effective method named contrasting shifted instances (CSI), which extends the power of contrastive learning for out-of-distribution (OOD) detection problems. Most of the above works in OOD detection focus on the supervised setting, where the detector is trained on in-distribution samples while required to identify unseen OOD samples. However, our approach is flexible and mitigates the need to train the model on in-distribution dataset yet being able to effectively detect OOD samples. We remark that our unsupervised setting is the more practical for challenging scenario since there are infinitely many cases for in-distribution, and it is often not possible to train the model for each reference dataset.

3. Methodology

We investigate the efficacy of features obtained from a network pretrained on a large dataset using self-supervised learning, for two tasks: (a) two-sample test, and (b) OOD detection. We explain our techniques below.

3.1. Learning Kernels with Self-supervision

The idea of self-supervised learning is to learn an encoder, parameterized by f_θ, to extract the necessary information to distinguish similar samples from the others without requiring any labels. We pre-train the encoder f_θ on a large dataset like Imagenet using self-supervised learning. We choose SimCLR-v2 [2] (with ResNet-50 backbone) as the self-supervised learning method based on its recent success in learning good representations for various downstream tasks. After self-supervised pretraining, we use f_θ as the featurizer for in-distribution and out-of-distribution datasets. The features are then used to define a kernel for our two tasks by taking the dot product, L_1 distance or cosine similarity between the feature vectors. We discuss them in detail in the following sections.

3.2. Two Sample Tests

For two sets $D_P = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and $D_Q = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{x}'_N\}$ that are *i.i.d.* samples from unknown distributions P and Q respectively, we wish to test the *null hypothesis*,

$$h_0 : P = Q$$

and its corresponding *alternative hypothesis*

$$h_1 : P \neq Q.$$

Here P might represent the training distribution, and Q the one encountered at test time. Any decision procedure based on finite samples will incur false positive and false negatives errors. Here we consider decision procedures with a guaranteed false positive error rate (significance level). Specifically, we use permutation tests [10] with a significance level $\alpha = 0.05$, which we briefly review next.

Let T be the test statistic of choice; we first compute the test statistic on the original sets, which we denote $T(D_P, D_Q)$. For example, this might be the difference on

the sample means of D_P and D_Q. Then we shuffle the set $D = D_P \cup D_Q$, randomly split the union to obtain D_1 and D_2 where $|D_1| = |D_2| = |D_P| = |D_Q|$ and calculate the test statistic on D_1 an D_2. Note that if indeed $P = Q$, then D_1 an D_2 are still distributed like D_P and D_Q. By repeating the procedure above M times, we obtain a set $\tilde{T}$:

$$\tilde{\mathcal{T}} = \{T(D_1^{(i)}, D_2^{(i)})| \\ D_1^{(i)}, D_2^{(i)} \text{ are random splits of } D, \forall i \in [M]\}$$

Finally we calculate the p-value,

$$\hat{p} = \Pr(T' \geq T(D_P, D_Q))$$

where $T' \sim \text{Uniform}(\tilde{\mathcal{T}})$, which is the probability of the statistic $T(D_1, D_2)$ being greater than $T(D_P, D_Q)$; we accept h_0 if $\hat{p} \geq \alpha$ and otherwise reject. This procedure is guaranteed to have a false positive error rate no larger than α, regardless of the statistic T.

We evaluate the capability of our pre-trained features f_θ with the *test power*, which is the probability to correctly reject the null hypothesis when the two sets come from different distribution (*i.e.*, null hypothesis is indeed false). A higher test power is better. In our approach we compute the statistic across any two sets using two different kernel-based approaches based on features f_θ:

MMD. We compute the maximum mean discrepancy (MMD, [12]) between the mean kernel embeddings of two empirical distributions D_P and D_Q, using a kernel defined by f_θ. This distance is used as the statistic T for the two sets of samples. For two sets of samples D_P and D_Q drawn from P and Q respectively, a kernel k, let $\mathbf{x}, \mathbf{x}' \sim D_P$ and $\mathbf{y}, \mathbf{y}' \sim D_Q$, we can use the following test statistic:

$$T_{\text{MMD}}(D_P, D_Q) = \sqrt{\sum_{\mathbf{x},\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + \sum_{\mathbf{y},\mathbf{y}'} k(\mathbf{y}, \mathbf{y}') - 2\sum_{\mathbf{x},\mathbf{y}} k(\mathbf{x}, \mathbf{y})}$$

where $k(\mathbf{x}, \mathbf{y}) = ||f_\theta(\mathbf{x}) - f_\theta(\mathbf{y})||_2^2$ is the kernel defined by the features. A higher MMD distance suggests that the D_P and D_Q are further than each other according to the kernel k. Therefore, if the alternative hypothesis is true, then we expect the permuted MMD scores to become lower.

Nearest Neighbor Graph Tests. Graph tests are a classical way of performing two-sample tests, and we consider one based on nearest neighbors in particular [20]. For two sets of samples D_P and D_Q sampled *i.i.d.* from two distributions P and Q respectively, we connect each sample $\mathbf{x}$ with the m-nearest neighbors around it (*e.g.*, according to some kernel k based on our features f_θ). Intuitively, if many points $\mathbf{x}$ from D_P are connected to many points $\mathbf{y}$ from D_Q (or vice versa), then it suggests that the two distributions P and Q are more likely to be similar. We can use the following test statistic:

$$T_G(D_P, D_Q) = \sum_{\mathbf{x} \in D_P} \sum_{\mathbf{y} \in D_Q} h_m(\mathbf{x}, \mathbf{y}, k) + h_m(\mathbf{y}, \mathbf{x}, k)$$

where$h_m(\mathbf{x}, \mathbf{y}, k) = 0$ if $\mathbf{y}$ is one of the m-nearest neighbors of $\mathbf{x}$ according to the kernel k; 1 otherwise. If T_m is high, then it suggests that the D_P and D_Q are relatively far from each other, since there are few nearest neighbors between the two sets. Therefore, when the alternative hypothesis is true, we expect the T_G statistic to become large.

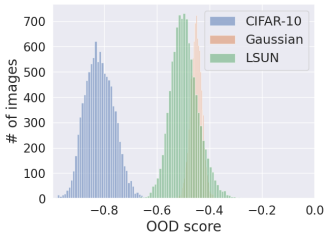

Figure 2: Histogram showing OOD scores for CIFAR-10 test data, LSUN and Gaussian noise with CIFAR-10 training data as D_P. The scores of the OOD data (LSUN, gaussian) are very different from the score of in-distribution data (CIFAR-10), making them easy to distinguish.

3.3. Out-of-distribution Detection

Out-of-distribution detection refers to the task of identifying whether a particular data sample $\mathbf{y}$ comes from a reference data distribution (in-distribution) P or not. Given a reference distribution P, we define a scoring function $s_{\text{ood}}(\mathbf{y})$ that measure how "out-of-distribution" is the test sample $\mathbf{y}$. Our goal is to have $s_{\text{ood}}(\mathbf{y})$ take a high value when $\mathbf{y}$ is out-of-distribution and $s_{\text{ood}}(\mathbf{y})$ take a low value when $\mathbf{y}$ is in-distribution.

Let D_P be the training samples from the distribution P. Using features from our network, f_θ, we define a score function for some test sample $\mathbf{y}$ as the negative cosine similarity to the nearest training sample in D_P:

$$s_{\text{ood}}(\mathbf{y}) = -\max_{\mathbf{x} \in D_P} \text{cosine-sim}(f_\theta(\mathbf{x}), f_\theta(\mathbf{y})) \quad (1)$$

where

$$\text{cosine-sim}(f_\theta(\mathbf{x}), f_\theta(\mathbf{y})) = \frac{f_\theta(\mathbf{x})^\top f_\theta(\mathbf{y})}{\|f_\theta(\mathbf{x})\|_2 \|f_\theta(\mathbf{y})\|_2} \quad (2)$$

is the cosine similarity between the two feature vectors.

We find that with self-supervised features (trained on Imagenet) as f_θ, our score function based on cosine similarity to the nearest sample in D_P is a simple yet highly effective approach for detecting OOD samples for a variety of

data distributions. In Section 4.2.3, we demonstrate the effectiveness of cosine similarity for OOD detection. Intuitively, a successful self-supervised learning method should be able to cluster semantically similar images. We posit that an in-distribution sample will be closer to samples in D_P, thus resulting in a high maximum cosine similarity (and low OOD score). On the contrary, OOD samples should be farther from the reference dataset D_P, implying that the cosine similarity is an effective feature to detect OOD samples. As can be seen from Figure 2, OOD datasets indeed get a higher OOD score via our approach.

To make our approach more scalable, instead of computing cosine similarity with all points in the training set, we can randomly select a small number of training samples (such as $5,000$) and compute the cosine similarity of the query sample $\mathbf{y}$ over this subset. In Section 5, we show that we almost maintain the same performance in OOD detection even with only 2% of the data in D_P.

4. Experiments and Results

For our method, we use the same ResNet50 network trained on ImageNet with SimCLR-v2 for all experiments.

4.1. Two Sample Test

As discussed in 3.2, we use the test power to measure if two sets of samples come from the same distribution. Following [25], we compute the average test power by repeating permutation test $M = 100$ times across different subsets of the two data. Finally we repeat the entire experiment 10 times to get mean and variance of the average test power. A high value of average test power across two sets which come from different distribution, means that the model is more confident that the distributions are actually disjoint.

4.1.1 Baselines

MMD-D [25]: Uses a kernel parameterized by a deep network to estimate the MMD distance. The deep network is trained to maximize the test power across a training set of the two sets.

MMD-O [25]: Uses a Gaussian Kernel to estimate MMD. The kernel is again learnt over a training set.

C2ST-S [26] and **C2ST-L** [4]: Train a binary classifier with the two distributions, then use the classification accuracy (C2ST-S) or the difference of the logit functions (C2ST-L) as the test statistic.

Supervised: We use a ResNet50 network pre-trained over ImageNet using supervised learning as our featurizer. We use the activations from penultimate layer of this network as features for our task. All baselines except for **Supervised** are based on MMD tests only.

4.1.2 Results

We perform our experiments over 4 different benchmark datasets with different sizes of the pair of datasets.

CIFAR-10 vs CIFAR-10.1 CIFAR-10.1 [33] is an additional test set for CIFAR-10 produced from Tiny ImageNet with vertain post-processing. The dataset was created to test the generalization performance of models trained over CIFAR-10 under small distribution shifts. The plot showing average test power for different sample sizes for MMD based test is shown in Figure 3 (a). While the prior state-of-the-art MMD-D [25] achieves average test power of 0.71 using 1000 samples, we reach to a value of 1.0 using only 350 samples. We also see that our approach even outperforms the Supervised baseline which has been pre-trained using the same dataset (ImageNet) and network architecture (ResNet50) as used for Self-Supervised pre-training in our approach. We justify this gain in Section 5. For graph test our approach again outperforms the Supervised baseline, and the corresponding plot is shown in Figure 3 (c).

CIFAR-10 vs CINIC-10 CINIC-10 is an extension of CIFAR-10 formed by downsampling the ImageNet images. The result has been shown in Figure 3 (b). Our approach performs better than the Supervised baseline for different sample sizes, while the remaining baselines do not perform nearly as well (with a test power near 0).

ImageNet vs ImageNet-v2 ImageNet-v2 [34] is a new test set for ImageNet created using a seperate pipeline than the original dataset pieline. The results for MMD based test is shown in Figure 4 (a). We again find our approach to perform much better than the other baselines and we reach an average test power of 1.0 using only 350 samples. For the graph test we again perform much better and the corresponding plot is shown in Figure 4 (c).

LSUN vs LSUN Fake We use the validation set of the data used in Wang *et al.* [41] where the fake images have been generated by Karras *et al.* [21]. The results are shown in Figure 4 (b), and we find that using only 100 samples we can perfectly discriminate between the real and fake distributions (with a test power of 1).

4.2. Out-of-distribution detection

We evaluate OOD detectors on a wide range of datasets. Each evaluation consists of an in-distribution reference set D_P, an in-distribution test set D_P^{test}, and an out-of-distribution dataset D_Q. We compute the OOD score over each data point in D_P^{test}, and similarly over each data point in D_Q, and then use these OOD scores to compute our metrics. We use only 5k samples from D_P in all our experiments for computing the OOD score.

Following [18], we use three evaluation metrics (with D_P^{test} labeled 0 and D_Q labeled 1): Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the

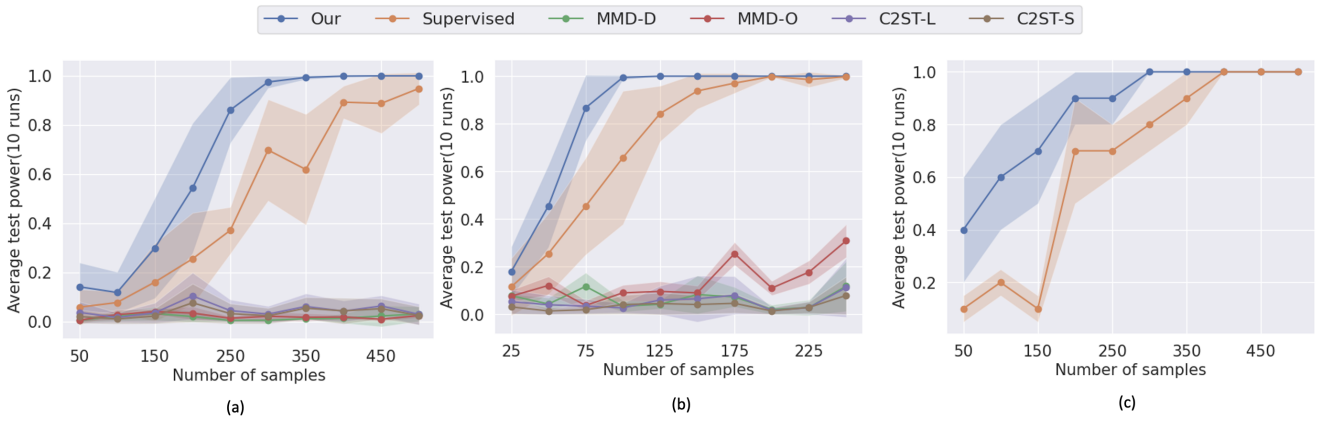

Figure 3: Average Test Power for different sample sizes across different pairs of dataset. (a) CIFAR-10 vs CIFAR-10.1 using MMD test, (b) CIFAR-10 vs CINIC-10 using MMD test, (c) CIFAR-10 vs CIFAR-10.1 using Graph Test.

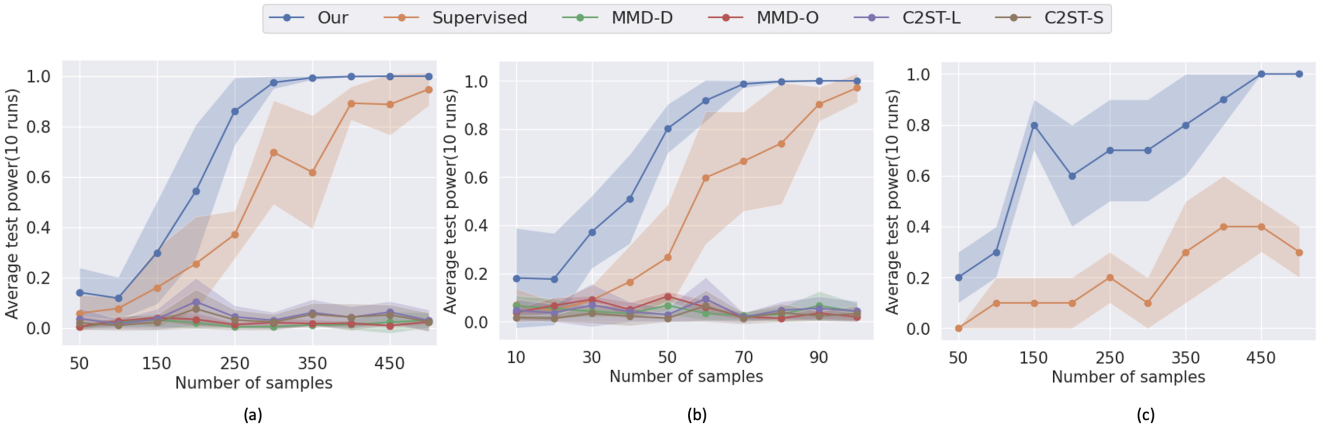

Figure 4: Average Test Power for different sample sizes across different pairs of dataset. (a) ImageNet vs ImageNet v2 using MMD test, (b) LSUN vs LSUN Fake using MMD test, (c) ImageNet vs ImageNet v2 using Graph Test.

Precision-Recall curve (AUPR), and the False Positive Rate at N% true positive rate (FPRN).

4.2.1 Datasets

The in-distribution and corresponding OOD datasets used in our experiments for OOD detection are shown in Table 1. We use the same set of datasets as used in [18]. Additional details are mentioned in **Appendix**.

4.2.2 Baselines

Maximum Softmax Probability [17] (MSP). Let g be the neural network model trained over in-distribution dataset D_P for classification task. MSP uses the maximum softmax probability output by g for an input query as the OOD score.

Table 1: In-distribution and corresponding OOD datasets used in our experiments.

In-distribution	Out-of-distribution
CIFAR-10	Gaussian, Blob, Rademacher, CIFAR-100, Textures, SVHN, Places365, LSUN
CIFAR-100	Gaussian, Blob, Rademacher, CIFAR-10, Textures,SVHN, Places365, LSUN
SVHN	Gaussian, Blob, Rademacher, Textures, Places365, LSUN, CIFAR-10
Tiny ImageNet	Gaussian, Blob, Rademacher, Textures, SVHN, Places365, LSUN, ImageNet
Places-365	Gaussian, Blob, Rademacher, Textures, SVHN, Places19, ImageNet

Outlier Exposure [18] (OE). In addition to the reference dataset, OE uses an auxiliary dataset: in-distribution

outliers, to train the supervised classifier. The maximum softmax score output by the model for an input query is used as its OOD score. While in our work we use only one single auxiliary dataset (ImageNet-1k) as the outlier set for all other smaller datasets, [18] uses different outlier dataset for each indivisual datasets, 80 million Tiny Images for CIFAR-10/100, SVHN and ImageNet-22k for Places and Tiny ImageNet.

Contrasting Shifted Instances [39] (CSI). CSI leverages features from a self-supervised network to define a score function based on cosine similarity and norm of the representation. However, the model requires additional training on D_P with their modified SimCLR training scheme that contrasts a sample with distributionally-shifted augmentations of itself in addition to contrasting with other instances.

Self-Supervision using Rotation Angle Prediction [19] (Rot). This work trains a supervised model on D_P along with an auxiliary rotation loss. Finally they use the maximum softmax score and the rotation loss as the OOD score.

Detection using Gram Matrices [36] (Gram). It uses the patterns in Gram matrices of the supervised model trained on D_P to identify OOD samples.

Supervised We use the same baseline as in 4.1.1. Note that this baseline is not comparable with other methods that use auxiliary data such as [18], because this requires additional label information.

Table 2: Mean AUROC scores averaged over all the OOD datasets. All the numbers are in percentage.

Method	C10	C100	SVHN	Tiny	Places
MSP [17]	89.27	73.11	97.95	64.86	66.51
OE [18]	**97.81**	87.89	99.96	92.18	90.57
Gram [36]	96.36	86.57	99.9	-	-
Rot. [19]	96.20	-	-	-	-
CSI [39]	95.88	70.47	-	-	-
Supervised	97.09	**93.1**	99.98	**99.17**	92.31
Ours	97.39	92.95	**99.99**	98.89	**93.55**

4.2.3 Results

The mean AUROC scores averaged across all the OOD datasets for each in-distribution data is shown in table 2. Per OOD data scores as well as results for the other two metrics are given in Appendix. Our approach beats all other existing works over all the datasets except for CIFAR-10, where it is slightly lower than OE (where they used 80 million Tiny Images). We find that in spite of using bigger size of additional data (ImageNet-22k and 80 million Tiny Images), OE performs worse than our approach, which uses ImageNet-1k as the additional data.

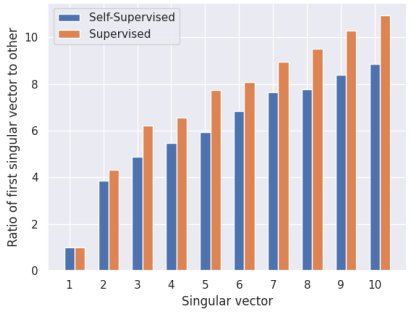

Figure 5: Plot of ratio of first singular value to other singular values over CIFAR-10 feature space.

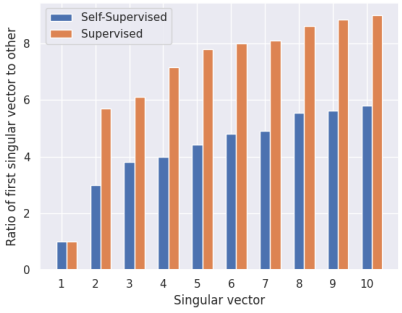

Figure 6: Plot of ratio of first singular value to other singular values over ImageNet feature space.

5. Ablation Studies

Self-Supervised vs Supervised Feat. In Section 4.1 we found that using a self-supervised trained network as feature extractor achieved better two-sample test performance compared to the supervised counterpart. In the SimCLR self-supervised learning framework, the query image is contrasted against several other images, either belonging to the same class or other class, while for a supervised task all images of a particular class are mapped to the same class; so the features learned from self-supervision may be better conditioned to tasks other than classification. We empirically verify this via performing SVD on the feature matrix obtained over the test set of CIFAR-10 and ImageNet and then plot the ratio of 1st singular value to the top 10 singular values. Fig. 5 and 6 show the corresponding plots for CIFAR-10 and ImageNet respectively. We find that the self-supervised features have a lower ratio compared to supervised features, which suggests that they are less centered on certain directions.

OOD detection vs In-distribution data size. Since our approach does not require any training over the in-

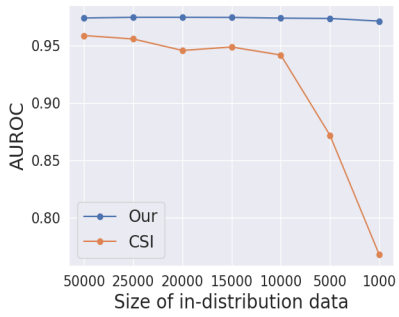

Figure 7: Variation of AUROC score with in-distribution data size over CIFAR-10.

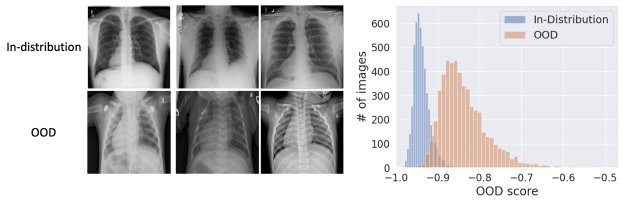

Figure 8: **Left:** Images from the two datasets. **Right:** Histogram of the OOD score by our approach.

distribution data, it can maintain its performance even when the size of in-distribution data is small. The changes of AUROC score with repsect to in-distribution data size for CIFAR-10 is shown in Fig. 7. While CSI [39] suffers a significant drop in performance as the data size reduces, the performance degradation of our method is minimal.

Table 3: OOD detection performance over CelebA.

Approach	**OOD Data**		
	SVHN	CIFAR-10	CIFAR-100
Ours	**99.98**	**99.98**	**99.97**
Blurring [6]	99.70	99.60	99.60

OOD over samples significantly different than ImageNet. In Section 4.2.3 we demonstrate OOD detection performance across 5 datasets. Arguably, these datasets have relatively small domain shifts from ImageNet. In order to test the efficacy of our approach we conduct experiments over two additional datasets: (a) CelebA and (b) Chest X-Rays. The two datasets were chosen because of significant visual differences with ImageNet. The results over CelebA are shown in Table 3. We find that we achieve near perfect detection performance over CelebA. We compare against [6], which uses OOD data to optimize certain parameters, unlike us. We also use two publicly available Chest X-ray datasets and treat one of them[1] as in-distribution and the other[2] as OOD. Using our approach we get an AUROC of 98.6%. The in-distribution and corresponding OOD data, along with the OOD scores have been shown in Fig. 8.

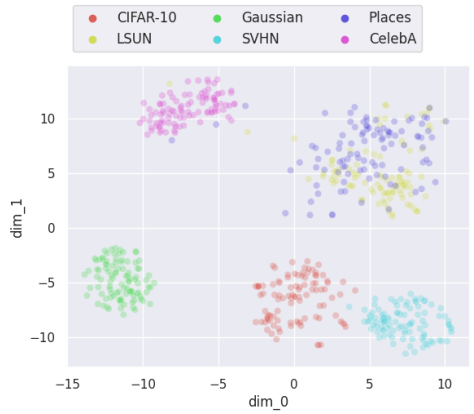

Figure 9: t-SNE [27] visualization of different datasets in our feature space.

Self-Supervised pre-training over in-distribution dataset. Instead of using ImageNet pre-trained network as the featurizer, we pre-train the network (ResNet50) over CIFAR-10 using SimCLR and anlysze its efficacy. We find that using this feature extractor and our approach of leveraging maximum cosine similarity for OOD score, reduces the AUROC value to 88.70.

t-SNE [27] visualization. We visualize the feature space of our featurizer across different datasets in Fig 9. Different datasets are mapped to separate clusters in this space, which facilitates us for the two tasks.

6. Conclusion

In this work, we propose a simple yet effective method for detecting outliers and distribution shifts. We advocate the use of generic kernels from self-supervised features on ImageNet for the two tasks as it mitigates any training on the reference dataset unlike other state-of-the-art methods, while still outperforming them on several benchmarks. We support our method with several ablation studies that suggest why self-supervised features perform better than supervised ones, and how that they are robust to changes in reference dataset sizes and dataset domains. Our work provides additional evidence that supports training on large unlabeled datasets instead of labeled ones. In future work, we hope to investigate additional benefits of self-supervised methods, such as better generative modeling.

[1] https://www.kaggle.com/nih-chest-xrays/data
[2] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

References

[1] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 3

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[4] Xiuyuan Cheng and Alexander Cloninger. Classification logit two-sample testing by neural networks. *arXiv preprint arXiv:1909.11298*, 2019. 2, 5

[5] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018. 3

[6] Sungik Choi and Sae-Young Chung. Novelty detection via blurring. *arXiv preprint arXiv:1911.11943*, 2019. 8

[7] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3

[8] Josip Djolonga and Andreas Krause. Learning implicit generative models using differentiable graph tests. arxiv e-prints, art. *arXiv preprint arXiv:1709.01006*, 2017. 3

[9] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019. 3

[10] Meyer Dwass. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pages 181–187, 1957. 3

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 1, 2, 3, 4

[13] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. 1

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 3, 6, 7

[18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 1, 3, 5, 6, 7

[19] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15663–15674, 2019. 1, 3, 7

[20] Norbert Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783, 1988. 4

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[22] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019. 2

[23] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 3

[24] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 3

[25] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Dougal J Sutherland. Learning deep kernels for non-parametric two-sample tests. *arXiv preprint arXiv:2002.09116*, 2020. 1, 2, 5

[26] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016. 2, 5

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8, 11

[28] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016. 2

[29] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018. 3

[30] 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*, 5, 2019. 3

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 2

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[33] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018. 1, 5

[34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019. 1, 5

[35] 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 14707–14718, 2019. 3

[36] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *arXiv preprint arXiv:1912.12510*, 2019. 7

[37] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. 2

[38] 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. 3

[39] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 3, 7, 8

[40] Shuo Wang, Yunfei Zha, Weimin Li, Qingxia Wu, Xiaohu Li, Meng Niu, Meiyun Wang, Xiaoming Qiu, Hongjun Li, He Yu, et al. A fully automatic deep learning system for covid-19 diagnostic and prognostic analysis. *European Respiratory Journal*, 2020. 1

[41] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 7, 2020. 5

[42] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

A. Dataset

We use following in-distribution datasets to evaluate the proposed method for OOD detection.

SVHN. It contains 32 × 32 color images of house numbers. There are ten classes comprised of the digits 0-9. The training set has 604, 388 images, and the test set has 26, 032 images.

CIFAR-10. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

CIFAR-100. The CIFAR-100 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

Tiny ImageNet. The Tiny ImageNet dataset is a 200-class subset of the ImageNet dataset where images are resized and cropped to 64 × 64 resolution. The dataset's images were cropped using bounding box information so that cropped images contain the target. The training set has 100, 000 images and the test set has 10, 000 images.

Places365. The Places365 training dataset consists of 1,803,460 large-scale photographs of scenes. Each photograph belongs to one of 365 classes.

B. OOD detection results

Here we present per OOD data scores as well as results for the other two metrics i.e. FPR95 and AUPR. The results are given in Table 4.

C. t-SNE visualization

We visualize the feature space of our featurizer across different datasets in Fig 10. The corresponding plot with the feature space of Supervised baseline is shown in Fig 11. As can be seen different clusters are more seperated in our approach as compared to the Supervised baseline, thereby achieving better performance across Two Sample Test and OOD detection.

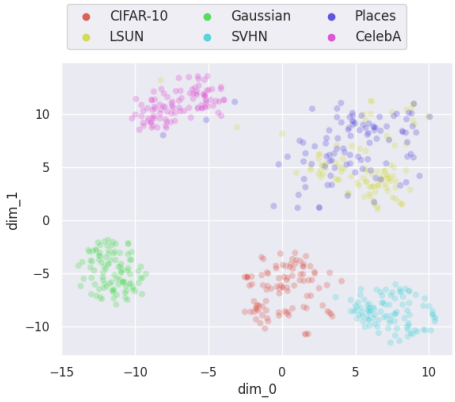

Figure 10: t-SNE [27] visualization of different datasets in our feature space.

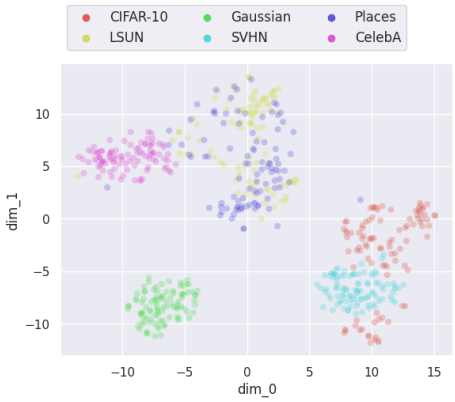

Figure 11: t-SNE [27] visualization of different datasets in the feature space of Supervised baseline.

Din	Dout	FPR95↓			AUROC↑			AUPR↑		
		MSP	OE	Ours	MSP	OE	Ours	MSP	OE	Ours
6*SVHN	Gaussian	5.4	0.0	0.0	98.2	100.0	100.0	90.5	100.0	100.0
	Blob	3.7	0.0	0.0	98.9	100.0	100.0	93.5	100.0	100.0
	Texture	7.2	0.2	0.0	97.5	100.0	99.9	90.9	99.7	100.0
	Places	5.6	0.1	0.0	98.1	100.0	100.0	92.5	99.9	100.0
	C10	6.0	0.1	0.0	98.0	100.0	99.9	91.2	99.9	99.9
	LSUN	6.4	0.1	0.0	97.8	100.0	100.0	91.0	99.9	100.0
	Mean	5.72	0.08	**0.00**	98.08	**100.0**	99.99	91.59	99.90	**99.99**
8*C10	Gaussian	14.4	0.7	0.0	94.7	99.6	100.0	70.0	94.3	100.0
	Rademacher	47.6	0.5	0.0	79.9	99.8	100.0	32.3	97.4	100.0
	Blob	16.2	0.6	0.0	94.5	99.8	100.0	73.7	98.9	100.0
	Texture	42.2	12.2	0.0	91.8	97.7	99.9	58.4	91.0	99.9
	SVHN	28.8	4.8	10.8	91.8	98.4	97.9	66.9	89.4	98.2
	Places	47.5	17.3	0.1	88.8	96.2	99.9	57.5	87.3	99.9
	LSUN	38.7	12.1	0.0	89.1	97.6	99.9	58.6	89.4	99.9
	C100	43.5	28.0	64.9	87.9	93.3	81.2	55.8	76.2	81.3
	Mean	34.94	9.50	**9.48**	89.27	**97.81**	97.39	59.16	90.48	**97.44**
8*C100	Gaussian	54.3	12.1	0.0	64.7	95.7	99.9	19.7	71.1	99.9
	Rademacher	39.0	17.1	0.0	79.4	93.0	100.0	30.1	56.9	100.0
	Blob	58.0	12.1	0.0	75.3	97.2	99.9	29.7	86.2	99.9
	Texture	71.5	54.4	0.0	73.8	84.8	99.9	33.3	56.3	99.9
	SVHN	69.3	42.9	59.1	71.4	86.9	89.2	30.7	52.9	91.2
	Places	70.4	49.8	0.3	74.2	86.5	99.8	33.8	57.9	99.8
	LSUN	74.0	57.5	0.5	70.7	83.4	99.8	28.8	51.4	99.8
	C10	64.9	62.1	95.2	75.4	75.7	54.8	34.3	32.6	56.9
	Mean	62.66	38.50	**19.39**	73.11	87.89	**92.95**	30.05	58.15	**93.47**
8*Tiny	Gaussian	72.6	45.4	0.0	33.7	76.5	99.9	12.3	28.6	99.9
	Rademacher	51.7	49.0	0.0	62.0	65.1	99.9	18.8	20.0	99.9
	Blob	79.4	0.0	0.0	48.2	100.0	99.9	14.4	99.9	99.9
	Texture	76.4	4.8	0.4	70.4	98.5	99.9	31.4	95.8	99.9
	SVHN	52.3	0.4	1.1	80.8	99.8	99.4	48.2	98.2	99.5
	Places	63.6	0.4	21.6	76.9	99.8	96.3	36.3	99.3	96.9
	LSUN	67.0	0.4	24.8	74.2	99.9	95.9	31.2	99.5	96.6
	ImageNet	67.3	11.6	2.6	72.8	97.9	99.4	30.0	92.9	99.4
	Mean	66.27	13.90	**6.30**	64.86	92.18	**98.86**	27.15	79.26	**99.04**
7*Places	Gaussian	37.1	9.4	0.0	72.2	93.5	99.9	23.5	54.1	99.9
	Rademacher	60.4	13.5	0.0	47.7	90.2	99.9	14.6	44.9	99.9
	Blob	73.7	0.1	0.0	41.9	100.0	99.9	13.0	99.4	99.9
	Texture	84.1	44.9	6.1	66.6	91.4	98.6	24.6	75.7	99.1
	SVHN	19.9	0.0	0.0	96.6	100.0	99.9	90.5	99.9	99.9
	ImageNet	86.3	65.3	7.1	63.0	86.5	98.4	25.1	69.7	98.3
	Places69	87.3	87.5	91.4	61.5	63.1	57.9	23.4	24.9	58.8
	Mean	63.46	28.21	**14.93**	66.51	90.57	**93.55**	33.08	71.04	**93.75**

Table 4: Per OOD data results in all three evaluation metrics for each in-distribution dataset.