# Contrastive Self-supervision Defines General-Purpose Similarity Functions

**Charles Guille-Escuret**[1,2], **Pau Rodríguez**[2], **David Vázquez**[2],
**Ioannis Mitliagkas**[1], **João Monteiro**[2]
1-Mila, Université de Montréal
2-ServiceNow Research
guillech@mila.quebec

## Abstract

Handling out-of-distribution (OOD) and adversarial inputs has become a major stake in the real-world deployment of machine learning systems. In this work, we explore the use of maximum mean discrepancy (MMD) two-sample test in conjunction with self-supervised contrastive learning to verify whether two sets of samples have been drawn from a same distribution. In particular, we find that the similarity functions defined on top of models trained with contrastive learning lead to high testing power on different types of distributional shifts. Our approach is able to differentiate CIFAR10 from CIFAR10.1 with much higher probability and using less samples than previous methods. Moreover, when trained on ImageNet, our approach shows great efficiency in detecting both adversarial attacks and OOD data on challenging benchmarks, using only 3 to 20 samples.

## 1 Introduction

While modern machine learning systems now have countless successful real-world applications, robustness to out-of-distribution (OOD) and adversarial inputs remain tough challenges of significant importance. The issue is especially acute for high-dimensional problems like image classification. Models are typically trained in a close-world setting but inevitably faced with novel input classes when deployed in the real world. The impact can range from displeasing customer experience to dire consequences in the case of safety-critical applications such as autonomous driving [8] or medical analysis [14]. Although achieving high accuracy against all meaningful distributional shifts is the most desirable solution, it is particularly challenging. Anomaly detection methods [18] aim to mitigate the consequences of unexpected inputs by allowing the system to anticipate its inability to process unusual inputs and react adequately. However, different distributional shifts (e.g., natural or adversarial) will unpredictably affect statistics used for anomaly detection. Accordingly, detection systems either achieve good performance on specific types of out-distributions or require tuning on OOD samples. In both cases, their practical use is severely limited. Motivated by these issues, recent work has tackled the challenge of designing detection systems for unseen classes without assuming access to OOD samples [20, 16, 18].

In this work, we push this reasoning further and define approaches able to distinguish sets of samples regardless of whether they correspond to unseen classes or in-distribution samples perturbed by adversarial attacks. A few pieces of work have addressed these tasks simultaneously. They either focus on particular in-distribution data such as medical imaging for specific diseases [17] or evaluate their performance on datasets with very distant classes such as CIFAR10 [9], SVHN [12], and LSUN [22], resulting in simple OOD benchmarks that do not translate to general real world applications [10]. In our work, we focus on popular two-sample tests based on the Maximum Mean

NeurIPS 2022 Self-Supervised Learning - Theory and Practice Workshop. New Orleans.

Discrepancy (MMD) [6, 11, 19, 5, 15, 3, 7], and show that a similarity function trained via contrastive learning can be used as a general-purpose similarity function to discriminate data sources.

**Contributions:** **1**-We show that one can use similarity functions learned by self-supervised contrastive learning with MMD [6] to define a general approach that discriminates sources of data. In particular, we show that the test sets of CIFAR10 and CIFAR10.1 [13] have different distributions with greater confidence than supervised alternatives. **2**-We propose improvements to MMD-based tests and show they can also be used to confidently detect distributional shifts when given a small number of samples from the same out-distribution. Namely, we show that our proposed method is able to discriminate very subtle changes on the test set of ImageNet, including adversarial perturbations.

## 2 Contrastive model

We build our model on SimCLRv2 [1] for its simplicity and efficiency. It is composed of an encoder backbone network $f_\theta$ as well as a 3-layer contrastive head $h_{\theta'}$. Given an in-distribution sample $\mathcal{X}$, a similarity function *sim*, and a distribution of training transformations $\mathcal{T}_{train}$, the goal is to simultaneously maximize $\mathbb{E}_{x\sim\mathcal{X};t_0,t_1\sim\mathcal{T}_{train}}\left[sim(h_{\theta'}\circ f_\theta(t_0(x)), h_{\theta'}\circ f_\theta(t_1(x)))\right]$ and minimize $\mathbb{E}_{x,y\sim\mathcal{X};t_0,t_1\sim\mathcal{T}_{train}}\left[sim(h_{\theta'}\circ f_\theta(t_0(x)), h_{\theta'}\circ f_\theta(t_1(y)))\right]$, i.e., we want to learn representations in which random transformations of the same exemplar are close while random transformations of different exemplars are distant. To achieve this, given an input batch $\{x_i\}_{i=1,\ldots,N}$, we compute the set $\{x_i^{(j)}\}_{j=0,1;i=1,\ldots,N}$ by applying two transformations independently sampled from $\mathcal{T}_{train}$ to each $x_i$. We then compute the embeddings $z_i^{(j)} = h_{\theta'}\circ f_\theta(x_i^{(j)})$ and apply the following contrastive loss:

$$L(\mathbf{z}) = \sum_{i=1,\ldots,N} -\log\frac{e^{sim(z_i^{(0)},z_i^{(1)})/\tau}}{\sum_{j\in\{1,\ldots,N\}} e^{sim(z_i^{(0)},z_j^{(1)})/\tau} + \sum_{j\in\{1,\ldots,N\}\setminus i} e^{sim(z_i^{(0)},z_j^{(0)})/\tau}}, \tag{1}$$

where $\tau$ is the temperature hyperparameter and $sim(x,y) = \frac{\langle x|y\rangle}{\|x\|_2\|y\|_2}$ is the *cosine* similarity.

**Hyperparameters:** We follow as closely as possible the setting from SimCLRv2 with a few modifications to adapt to hardware limitations. In particular, we use the LARS optimizer [21] with learning rate $1.2$, momentum $0.9$, and weight decay $10^{-4}$. We scale up the learning rate for the first $40$ epochs linearly, then use an iteration-wise cosine decaying schedule until epoch $800$, with temperature $\tau = 0.1$. We train on $8$ $V100$ GPUs with a total batch size of $1024$. We compute the contrastive loss on all batch samples by aggregating the embeddings computed by each GPU. We use synchronized BatchNorm and fp32 precision and do not use a memory buffer. We use the same set of transformations, i.e., Gaussian blur and horizontal flip with probability $0.5$, color jittering with probability $0.8$, random crop with scale uniformly sampled in $[0.08, 1]$, and grayscale with probability $0.2$. For computational simplicity and comparison with previous work, we use a ResNet50 encoder architecture with final features of size $2048$. Following SimCLRv2, we use a three-layer fully connected contrastive head with hidden layers of width $2048$ using ReLU activation and batchNorm and set the last layer projection to dimension $128$. For evaluation, we use the features produced by the encoder without the contrastive head. *We do not use supervised fine-tuning.*

## 3 MMD two-sample test

The **Maximum Mean Discrepancy (MMD)** is a statistic used in two-sample tests to assess whether two sets of samples $S_\mathbb{P}$ and $S_\mathbb{Q}$ are drawn from the same distribution. It estimates the expected difference between the intra-set distances and the across-sets distances.

**Definition 3.1** (Gretton et al. [6])**.** Let $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be the kernel of a reproducing Hilbert space $\mathcal{H}_k$, with feature maps $k(\cdot,x)\in\mathcal{H}_k$. Let $X, X'\sim\mathbb{P}$ and $Y, Y'\sim\mathbb{Q}$. Under mild integrability conditions,

$$MMD(\mathbb{P},\mathbb{Q};\mathcal{H}_k) := \sup_{f\in\mathcal{H}_k,\|f\|_{\mathcal{H}_k}\leq 1} \left|\mathbb{E}\left[f(X)\right] - \mathbb{E}\left[f(Y)\right]\right| \tag{2}$$

$$= \sqrt{\mathbb{E}\left[k(X,X') + k(Y,Y') - 2k(X,Y)\right]}. \tag{3}$$

Given two sets of $n$ samples $S_{\mathbb{P}} = \{X_i\}_{i \leq n}$ and $S_{\mathbb{Q}} = \{Y_i\}_{i \leq n}$, respectively drawn from $\mathbb{P}$ and $\mathbb{Q}$, we can compute the following unbiased estimator [11]:

$$\widehat{MMD}_u^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k) := \frac{1}{n(n-1)} \sum_{i \neq j} (k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(Y_i, X_j)). \quad (4)$$

Under the null hypothesis $\mathfrak{h}_0 : \mathbb{P} = \mathbb{Q}$, this estimator follows a normal distribution of mean 0 [6]. Its variance can be directly estimated [5], but it is simpler to perform a permutation test as suggested in Sutherland et al. [15], which directly yields a $p$-value for $\mathfrak{h}_0$. The idea is to use random splits $X, Y$ of the input sample sets to obtain $n_{perm}$ different (though not independent) samplings of $\widehat{MMD}_u^2(X, Y; k)$, which approximate the distribution of $\widehat{MMD}_u^2(S_{\mathbb{P}}, S_{\mathbb{P}}; k)$ under the null hypothesis. Liu et al. [11] train a deep kernel to maximize the test power of the MMD two-sample test on a training split of the sets of samples to test. We propose instead to use our similarity function learned from data, without any fine-tuning. Additionally, we propose an improvement of MMD called MMD-CC (MMD with Clean Calibration). Instead of computing $p_i$ (see Algorithm 1 for a precise notation definition) based on random splits of $S_{\mathbb{P}} \bigcup S_{\mathbb{Q}}$, we require as input two disjoint sets of samples drawn from $\mathbb{P}$ and compute $p_i$ based on random splits of $S_{\mathbb{P}}^{(1)} \bigcup S_{\mathbb{P}}^{(2)}$. This change requires to use twice as many samples from $\mathbb{P}$, but reduces the variance induced by the random splits of $S_{\mathbb{P}} \bigcup S_{\mathbb{Q}}$, which is significant when the number of samples is small.

We also highlight that Algorithm 1 returns the $p$-value $\frac{1}{n_{perm}+1} \left(1 + \sum_{i=1}^{n_{perm}} \mathbb{1}\left(p_i \geq est\right)\right)$ instead of $\frac{1}{n_{perm}} \sum_{i=1}^{n_{perm}} \mathbb{1}\left(p_i \geq est\right)$. Indeed, under the null hypothesis $\mathbb{P} = \mathbb{Q}$, $est$ and $p_i$ are drawn from the same distribution, so for $j \in \{0, 1, \dots, n_{perm}\}$, the probability for $est$ to be smaller than exactly $j$ elements of $\{p_i\}$ is $\frac{1}{n_{perm}+1}$. Therefore, the probability that $j$ elements or less of $\{p_i\}_i$ are larger than $est$ is $\sum_{i=0}^{j} \frac{1}{n_{perm}+1} = \frac{j+1}{n_{perm}+1}$. While this change has a small impact for large values of $n_{perm}$, it is essential to guarantee that we indeed return a correct $p$-value. Notably, the algorithm of Liu et al. [11] has a probability $\frac{1}{n_{perm}} > 0$ to return an output of $0.00$ even under the null hypothesis.

### 3.1 Distribution shift between CIFAR-10 and CIFAR-10.1 test sets

After years of evaluation of popular supervised architectures on the test set of CIFAR-10 [9], modern models may overfit it through their hyperparameter tuning and structural choices. CIFAR-10.1 [13] was collected to verify the performances of these models on a *truly* independent sample from the training distribution. The authors note a consistent drop in accuracy across models and suggest it could be due to a distributional shift, though they could not demonstrate it. Recent work [11] leveraged the two-sample test to provide strong evidence of distributional shifts between the test sets of CIFAR-10 and CIFAR-10.1. We run MMD-CC and MMD two-sample tests for 100 different samplings of $S_{\mathbb{P}}^{(1)}, S_{\mathbb{P}}^{(2)}, S_{\mathbb{Q}}$, always using $n_{perm} = 500$, and rejecting $\mathfrak{h}_0$ when the obtained $p$-value is below the threshold $\alpha = 0.05$. We also report results using cosine similarity applied to the features of supervised models as a comparative baseline. We report the results in Table 1 for a range of sample sizes. We compare the results to three competitive methods reported in Liu et al. [11]: Mean embedding (ME) [3, 7], MMD-D [11], and C2ST-L [2]. Finally, we show in Figure 1 the ROC curves of the proposed model for different sample sizes.

**Algorithm 1** MMD-CC two-sample test

```
Input: $S_{\mathbb{P}}^{(1)}, S_{\mathbb{P}}^{(2)}, S_{\mathbb{Q}}, n_{perm}, sim$
  $est \leftarrow \widehat{MMD}_u^2(S_{\mathbb{P}}^{(1)}, S_{\mathbb{Q}}; sim)$
  for $i = 1, 2, \dots, n_{perm}$ do
      Randomly split $S_{\mathbb{P}}^{(1)} \bigcup S_{\mathbb{P}}^{(2)}$ into two disjoint sets $X, Y$ of equal size
      $p_i \leftarrow \widehat{MMD}_u^2(X, Y; sim)$
  end for
Output: $p$: $\frac{1}{1+n_{perm}} \left(1 + \sum_{i=1}^{n_{perm}} \mathbb{1}\left(p_i \geq est\right)\right)$
```

### 3.2 Detection of OOD and adversarial distributions

Given a small set of samples with potential unknown classes or adversarial attacks, we can similarly use the two-sample test with our similarity function to verify whether these samples are in-distribution [4]. In particular, we test for samples drawn from ImageNet-O, iNaturalist, and PGD perturbations,

Table 1: Average rejection rates of $\mathfrak{h}_0$ on CIFAR-10 vs CIFAR-10.1 for $\alpha = 0.05$ across different sample sizes $n$, using a ResNet50 backbone.

| | n=2000 | n=1000 | n=500 | n=200 | n=100 | n=50 |
|---|---|---|---|---|---|---|
| ME [3] | 0.588 | - | - | - | - | - |
| C2ST-L [2] | 0.529 | - | - | - | - | - |
| MMD-D [11] | 0.744 | - | - | - | - | - |
| MMD + SimCLRv2 (ours) | **1.00** | **1.00** | **0.997** | **0.702** | **0.325** | **0.154** |
| MMD-CC + SimCLRv2 (ours) | **1.00** | **1.00** | **0.997** | 0.686 | 0.304 | 0.150 |
| MMD + Supervised (ours) | **1.00** | **1.00** | 0.884 | 0.305 | 0.135 | 0.103 |
| MMD-CC + Supervised (ours) | **1.00** | **1.00** | 0.870 | 0.298 | 0.131 | 0.096 |

Table 2: AUROC for detection using two-sample test on 3 to 20 samples drawn from ImageNet and from ImageNet-O, iNaturalist or PGD perturbations, with a ResNet50 backbone.

| | ImageNet-O | | | | iNaturalist | | | | PGD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n_samples | 3 | 5 | 10 | 20 | 3 | 5 | 10 | 20 | 3 | 5 | 10 | 20 |
| MMD + SimCLRv2 | 64.3 | 72.4 | 86.9 | 97.6 | 88.3 | 97.6 | **99.5** | **99.5** | 35.2 | 53.8 | 86.6 | 98.8 |
| MMD-CC + SimCLRv2 | **65.3** | **73.2** | **88.0** | **97.7** | 95.4 | 99.2 | **99.5** | **99.5** | **70.5** | **84.0** | **96.6** | **99.5** |
| MMD + Supervised | 62.7 | 69.7 | 83.2 | 96.4 | 91.8 | 98.7 | **99.5** | **99.5** | 20.0 | 22.5 | 33.0 | 57.5 |
| MMD-CC + Supervised | 62.6 | 71.0 | 85.5 | 97.2 | **98.0** | **99.5** | **99.5** | **99.5** | 57.4 | 61.3 | 70.5 | 85.8 |

with sample sizes ranging from 3 to 20. For these experiments, we sample $S_{\mathbb{P}}^{(1)}$ and $S_{\mathbb{P}}^{(2)}$ 5000 times across all of ImageNet's validation set and compare their MMD and MMD-CC estimators to the one obtained from $S_{\mathbb{P}}$ and $S_{\mathbb{Q}}$. We report in Table 2 the AUROC of the resulting detection and compare it to the ones obtained with a supervised ResNet50 as the baseline. We discuss it further in Section 3.3.

### 3.3 Discussion

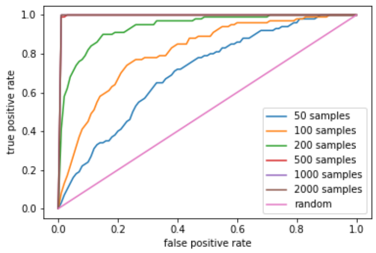

Figure 1: ROC curves of MMD-CC two-sample test on CIFAR-10.1 against CIFAR-10 for different sample sizes.

**CIFAR-10 vs. CIFAR10.1**: Other methods do not use external data for pre-training, as we do with ImageNet, rendering a fair comparison difficult. However, it is noteworthy that our learned similarity can very confidently distinguish samples from the two datasets, even if fewer samples are available. Furthermore, while we achieve excellent results even with a supervised network, our model trained with contrastive learning outperforms the supervised alternative significantly. We note however that with such high number of samples available, MMD-CC performs slightly worse than MMD. Finally, we believe the confidence obtained with our method decisively concludes that CIFAR10 and CIFAR10.1 have different distributions, which is likely the primary explanation for the significant drop in performances across models on CIFAR10.1, as conjectured by Recht et al. [13]. The difference in distribution between CIFAR10 and CIFAR10.1 is neither based on label set nor adversarial perturbations, making it an interesting task.

**ImageNet-O, iNaturalist, and PGD**:

Despite using very few samples ($3 \leq n \leq 20$), our method can detect OOD samples with high confidence. We observe particularly outstanding performances on iNaturalist, which is easily explained by the fact that the subset we are using only contains plant species, logically inducing an abnormally high similarity within its samples. Furthermore, we observe that MMD-CC performs significantly better than MMD, especially on detecting samples perturbed by PGD. Although our method attains excellent detection rates for sufficient numbers of samples, the requirement to have a set of samples all drawn from the same distribution to perform the test makes it unpractical for real-world applications, which we seek to address in future work.

# References

[1] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[2] X. Cheng and A. Cloninger. Classification logit two-sample testing by neural networks. *arXiv preprint arXiv:1909.11298*, 2019.

[3] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. *arXiv preprint arXiv:1506.04725*, 2015.

[4] R. Gao, F. Liu, J. Zhang, B. H. 0003, T. Liu, G. N. 0001, and M. Sugiyama. Maximum mean discrepancy test is aware of adversarial attacks. In M. Meila and T. Z. 0001, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3564–3575. PMLR, 2021. URL `http://proceedings.mlr.press/v139/gao21b.html`.

[5] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, 2009.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[7] W. Jitkrittum, Z. Szabo, K. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. *arXiv preprint arXiv:1605.06796*, 2016.

[8] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In *2010 IEEE Intelligent Vehicles Symposium*, pages 486–492, 2010. doi: 10.1109/IVS.2010.5548123.

[9] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[10] K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

[11] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, 2020.

[12] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[13] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

[14] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *arXiv preprint arXiv:1703.05921*, 2017.

[15] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

[16] J. Tack, S. Mo, J. Jeong, and J. Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, 2020.

[17] A. Uwimana1 and R. Senanayake. Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis. *arXiv preprint arXiv:2107.04882*, 2021.

[18] H. Wang, Z. Li, L. Feng, and W. Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[19] L. Wenliang, D. Sutherland, H. Strathmann, and A. Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, 2019.

[20] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[21] Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[22] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.