No Shifted Augmentations (NSA): compact distributions for robust self-supervised Anomaly Detection

Mohamed Yousef[1], Marcel Ackermann[1], Unmesh Kurup[1*]; Tom Bishop[2*]

[1]Intuition Machines, Inc.
[2]Glass Imaging, Inc. [‡]
{myb,marcel,unmk}@imachines.com, tom@glass-imaging.com

Abstract

Unsupervised Anomaly detection (AD) requires building a notion of normalcy, distinguishing in-distribution (ID) and out-of-distribution (OOD) data, using only available ID samples. Recently, large gains were made on this task for the domain of natural images using self-supervised contrastive feature learning as a first step followed by kNN or traditional one-class classifiers for feature scoring. Learned representations that are non-uniformly distributed on the unit hypersphere have been shown to be beneficial for this task. We go a step further and investigate how the geometrical compactness *of the ID feature distribution makes isolating and detecting outliers easier, especially in the realistic situation when ID training data is polluted (i.e. ID data contains some OOD data that is used for learning the feature extractor parameters). We propose novel architectural modifications to the self-supervised feature learning step, that enable such compact distributions for ID data to be learned. We show that the proposed modifications can be effectively applied to most existing self-supervised objectives, with large gains in performance. Furthermore, this improved OOD performance is obtained without resorting to tricks such as using strongly augmented ID images (e.g. by 90 degree rotations) as proxies for the unseen OOD data, as these impose overly prescriptive assumptions about ID data and its invariances. We perform extensive studies on benchmark datasets for one-class OOD detection and show state-of-the-art performance in the presence of pollution in the ID data, and comparable performance otherwise. We also propose and extensively evaluate a novel feature scoring technique based on the angular Mahalanobis distance, and propose a simple and novel technique for feature ensembling during evaluation that enables a big boost in performance at nearly zero run-time cost compared to the standard use of model ensembling or test time augmentations. Source code is available Here*

1. Introduction

Anomaly detection (AD) or out-of-distribution (OOD) detection requires using only available in-distribution (ID) samples for training a classifier to decide upon the relative normalcy of samples at test time, without knowledge of the nature of the OOD data. OOD detection is an important problem with practical applications in, for example, industrial defect detection, fraud detection, autonomous driving, biometrics, spoofing detection and many other domains [38].

1.1. Background

For natural images (which according to the manifold hypothesis lie in a compact set in a suitable space), OOD detection translates to finding as tight a decision boundary as possible around the normal set, while excluding unseen samples from other classes or distributions. This detection was traditionally done with either generative [49] or discriminative models [34] on top of shallow features. Deep representations subsequently provided a large boost in performance. However, the density of deep generative image models have often proved to be ineffective [17], with poorly calibrated likelihoods away from observed data. There have instead been two recent directions to learn suitable deep features used for OOD evaluation: a) supervised pre-training on an external dataset. b) Self-supervised learning (SSL) pre-training on either the normal set only or also on an external dataset. A variety of learned metrics, scoring functions, or one-class classifiers have then been employed on top of these learned features and this general paradigm has shown to be highly effective in many cases [36].

The best recent results with SSL-based training for OOD has been with contrastive learning [38, 36, 35, 43]. Con-

[*] Equally contributing last authors.
[‡] Work done while at Intuition Machines, Inc.

trastive learning has been shown to distribute ID data uniformly on the hypersphere [42]. While this helps general multi-class SSL training, it hurts OOD detection, as it makes isolating outliers from the single class harder [36]. This uniformity also makes OOD detection much more sensitive to pollution in the inlier training data [13, 36]. See Appendix H (supplementary material) for more background on related methods.

If some labeled OOD data are available as negatives, semi-supervised learning maybe be used [31, 13]. If no such labeled negatives are available, one way to soften the effect of a contrastively learned uniform representation is to introduce artificial negative samples as proxies for these outliers. Using hard augmentations (e.g. 90 degree rotations) has been termed *Distributional Shifting* [38, 36, 24, 39]. Such augmentations are intended to make the in-distribution data less uniform, and thus easier to isolate from OOD data. However, they also make the significant assumption that the data is not (fully or partially) invariant to those augmentation(s), and that the augmentations are a good proxy for the true negative distribution.

The other direction is using features from a model pretrained on a large external dataset, with the hope of producing universal features that can work in any OOD detection scenario. This can be done in either supervised [29] or self-supervised manner [46]. However, the assumption that such representative labeled samples will be available in the latter case, or that learned image features from general datasets such as ImageNet will transfer well may be restrictive at best.

1.2. Contribution

In this work, we first investigate the training dynamics of contrastive SSL methods, and show that their performance decays significantly over long term training. We find that uniformity or non-compactness of the learned ID representation is the main reason for this decay. We study this effect on positive-pairs-only SSL using SimSiam [6], and show that in such cases, the decay does not happen.

We propose an architectural modification that can be applied generally across such networks, and show extensive analysis that this modification improves performance and always encourages learning a more compact ID representations. In doing so we are able to learn high quality One-class classifiers without resorting to distributionally shifted augmented samples as negatives, hence we term the resulting methodology *No Shifted Augmentations* (NSA). We summarize our contributions as follows:

- We investigate and empirically verify and quantify that the the non-uniformity and compactness of learned ID is a main factor of the final OOD detection performance, independent from the quality of the learned features.
- We propose an assumption-free, simple and novel architectural modification for inducing a non-uniform learned ID representation, and show that this works very well with both SimSiam and SimCLR and produces solid performance improvements.
- We identify, investigate, and solve a gradient problem in SimSiam (and also BYOL) that greatly affects the proper propagation of the norms inside the network; we solve it and notice much higher stability, especially in the low batch size training regime.
- We consider improved feature scoring methods for OOD detection, including in our proposed solution a Mahalanobis Cosine score on nearest neighbors, related to methods in open-set metric learning. We then present a computationally efficient method of feature ensembling that also boosts performance.
- We show unexpected case(s) (e.g. SVHN) where the usual ImageNet-Pretrained ResNet methods fail catastrophically on One-Class Classification Anomaly Detection tasks, even with feature adaptation. We show that training from scratch without using shifted augmentations avoids this.
- We extensively evaluate and ablate the proposed models with a wide variety of different datasets and scenarios, separating the contributions of representation learning, scoring, data augmentation, and additional variations like ensembling. We show our solutions have comparable performance against more complex methods. More importantly they show state-of-the-art performance, by a wide margin, achieving robustness for small batch sizes and in the presence of polluted data.

2. Is representation quality the only important factor for OOD detection ?

Recent work on OOD detection using pretrained networks has suggested that OOD detection performance is dependent on stronger representations for ID data [8, 41, 18]. We here conduct an experiment to examine the interaction between quality of representation on ID data and OOD performance.

2.1. Experiment

We train SimCLR on one class of CIFAR10 and evaluate both the quality of learned representation and OOD detection against all other classes throughout training; this is the standard one-class protocol used in e.g. [35, 38, 29]. However, in order to examine all modes of the model during training, we train for much longer than usual. We repeat

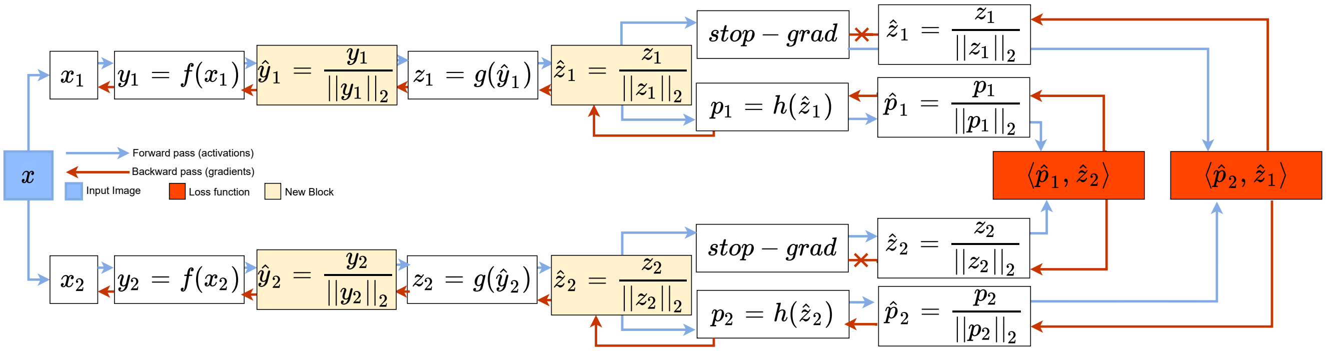

x
x_1
$y_1 = f(x_1)$
$\hat{y}_1 = \frac{y_1}{||y_1||_2}$
$z_1 = g(\hat{y}_1)$
$\hat{z}_1 = \frac{z_1}{||z_1||_2}$
$stop - grad$
$\hat{z}_1 = \frac{z_1}{||z_1||_2}$
$p_1 = h(\hat{z}_1)$
$\hat{p}_1 = \frac{p_1}{||p_1||_2}$
$\langle \hat{p}_1, \hat{z}_2 \rangle$
$\langle \hat{p}_2, \hat{z}_1 \rangle$
Forward pass (activations)
Backward pass (gradients)
Input Image
Loss function
New Block
x_2
$y_2 = f(x_2)$
$\hat{y}_2 = \frac{y_2}{||y_2||_2}$
$z_2 = g(\hat{y}_2)$
$\hat{z}_2 = \frac{z_2}{||z_2||_2}$
$stop - grad$
$\hat{z}_2 = \frac{z_2}{||z_2||_2}$
$p_2 = h(\hat{z}_2)$
$\hat{p}_2 = \frac{p_2}{||p_2||_2}$

Figure 1: An illustration of modified SimSiam with the yellow boxes showing our proposed changes described in Sec. 3.2, 3.3. In the original version the gradients of $\hat{z}_i = z_i/||z_i||_2$ are prevented from flowing backwards in *all possible paths* by the stop-grad operation. In our modified version, thanks to the added operations, we can mimic the flow of the gradients of $\hat{z}_i = z_i/||z_i||_2$ that were blocked.

this experiment for 4 different classes and report the average. Figure 2a (blue curve, baseline) shows OOD detection performance, and it is evident that performance reaches a peak very fast (within few hundred epochs) then deteriorates significantly as training progresses. Appendix A (supplementary material) shows the exact same phenomena using two other OOD evaluation metrics, with detailed graphs for each class.

If the representation quality hypothesis was correct, then we should see a proportional drop in feature quality. In figure 2c we use linear evaluation [50] to evaluate the quality of learned representation. We can see that the quality of the representation is maintained throughout training, i.e. there is only the expected small decrease after the peak and then just small fluctuations. Weighted k-NN [44] results supporting the same fact are in Appendix A. Putting all these data together, both OOD performance and feature quality reach a peak together in a few hundred epochs, then for the rest of training, feature quality is maintained, but OOD performance deteriorates significantly. This is validated using 5 different metrics in 4 different classes, all show the same trend.

2.2. Analysis using von Mises-Fisher distribution

To study what is happening, we use the von Mises-Fisher (vMF) distribution which is a fundamental probability distribution on the $(n-1)$-dimensional hyper-sphere $S^{d-1} \subset R^d$. Its probability density function is $f_n(z, \mu, \kappa) = C_n(\kappa)e^{\kappa\mu^T z}$, where μ is the mean direction, and κ is the concentration parameter. The vMF shape depends on κ: for high values, the distribution has a mode at the mean direction μ; for $\kappa = 0$ it is uniform on the hyper-sphere S^{d-1}. vMF has been successfully used in both analyzing [42, 21, 23] and learning [14, 7, 20] deep neural networks.

We fit a vMF [1, 37] to the learned normalized embeddings of SimCLR, and study how κ changes during training. Figure 2b shows how SimCLR starts with a relatively large κ (high concentration) then reduces monotonically (low concentration, more uniform) as training progresses – this is true for both ID and OOD.

In Appendix A, Figure 2b we notice the same behaviour using another tool: the Maximum Mean Discrepancy (MMD) [11] between the learned representation and samples from a uniform distribution on a unit hypersphere, as suggested in [36]. MMD measures the distance between two probability distributions, so a high MMD means a less uniform and more concentrated distribution, and a low MMD the opposite. We note that this perfectly aligns with findings in [42], that SimCLR (and generally contrastive) features converge to a uniform distribution.

This gives an explanation of the decaying performance seen in Figure 2a (blue curve). While the quality of the features is maintained, their distribution changes dramatically. As the ID features' distribution gets more uniform, the probability p to find an inlier sample $x \in$ ID arbitrarily close to an outlier query sample $x' \in$ OOD increases, i.e. ID and OOD get more and more indistinguishable. This is exactly the intuition behind many methods for AD, for example [31, 10, 32]

Given this analysis, we take the natural step and perform the same analysis on SimSiam, a non-contrastive SSL method. Results are shown in Figure 2c, 2d, where we can see a big difference compared to SimCLR. OOD performance stays nearly constant after reaching the peak. In Figure 2e, 2f although the representation is also becoming more uniform, it is changing at a much slower rate, and is able to maintain high density (high κ and MMD values) even after thousands of epochs of training.

Figures also show that the number of training epochs is an important hyperparameter for contrastive methods used for OOD. However, non-contrastive methods are much less sensitive to the number of epochs. Other works that noticed importance of early stopping in this context include [29, 13, 30]

3. Methodology

3.1. Self-supervised Learning (SSL) setup

A general SSL setup resembling SimSiam [6] is depicted in Figure 1. SimCLR is a very similar architecture that omits a prediction network, so we use SimSiam for this illustration. The figure shows both the forward pass and backward pass of two random augmentations x_1 and x_2 of an input image x. Both are encoded by the shared encoder (the convolutional backbone network) f into y_1 and y_2. Both augmentations are projected into z_1 and z_2 by a projection network g. A prediction network h transforms z_i into $p_i = h(z_i)$ and the whole network learns to match the output of the prediction network fed with the projection of one view $p_i = h(z_i)$ to the projection of the other view $z_j = g(y_j)$ and vice versa, by minimizing their negative cosine similarity:

$$\mathcal{D}(p_i, z_j) = -\frac{p_i}{\|p_i\|_2} \cdot \frac{z_j}{\|z_j\|_2}. \tag{1}$$

Most current SSL algorithms use a projection head g on top of a convolutional feature extractor, as this was empirically found to help learn a much better final representation [4, 5, 12, 6]. However, it was also found that the learned projection is much worse in downstream tasks [4] including for OOD [36]. Therefore, the output of the encoder backbone f used during feature evaluation in most OOD methods is based on SSL, as it learns a much better representation. We call the outputs of f the learned embedding.

3.2. Learning a dense representation

[33] show that the sample complexity of robust learning can be significantly larger than that of standard learning. While some works tried to address this difference with extra positive or negative data, [27] propose the interesting idea of manipulating the local sample distribution of the training data via appropriate training objectives such that by inducing high-density feature regions, there would be locally sufficient samples to train robust classifiers and return reliable predictions. We propose pursuing the same direction with the OOD detection problem, and propose an architectural modification that can help induce high density feature space.

We propose adding a differentiable l_2-normalization operation after the encoder f and before the projection head g. As such, the output $y_i = f(x_i)$ is transformed into $\hat{y_i} = y_i/||y_i||_2$ (as in Figure 1). The intuition is that adding the normalization step would deprive the model of any gradients when the norm of the embedding is changed, thus enforcing the model to learn directional transformations as those would be the only way to actually decrease the loss function. Regularizing the model this way should give a finer directional control on the learned embedding (the cosine distance makes more sense), and yield a more efficient use of volume and thus a denser representation.

We perform a series of experiments to empirically verify our intuition. Figure 2a (red curve) shows evaluation logs for SimCLR after adding normalization. It is evident the training is much more regularized now, the decrease in OOD performance after the peak is much smaller and performance is maintained for thousands of epochs. Figure 2b (red curve) shows a denser learned representation when compared to original SimCLR. We can see the same behaviour for SimSiam in Figure 2d, 2e.

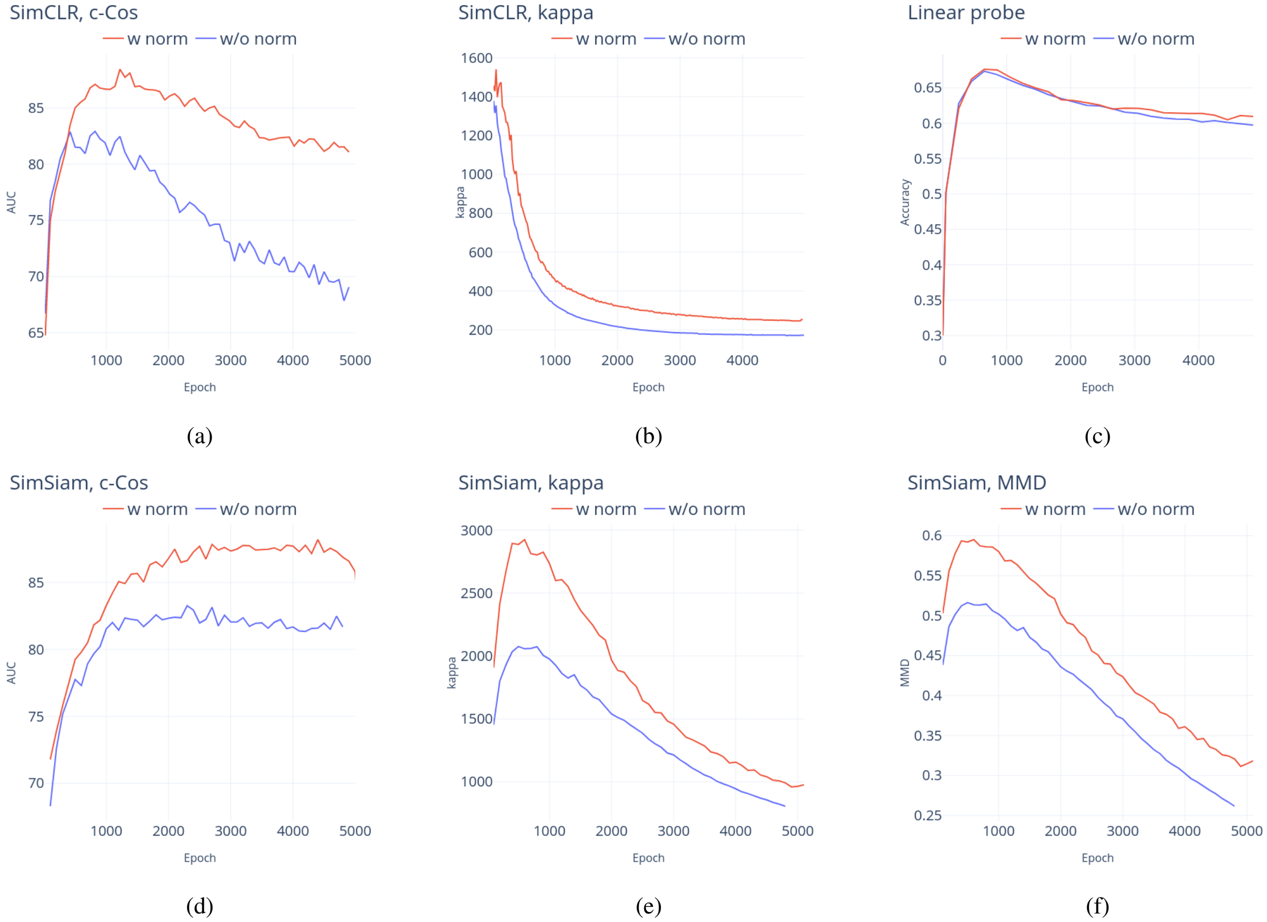
One thing to note here that greatly strengthens the analysis made in the previous section, is the linear probe accuracy with and w/o norm (Figure 2c), they are essentially the same, even after 5000 epochs, on the other hand there is big deference between their OOD performance at that point (see Figure 2a). Further showing that it is a problem of feature distribution not feature quality.

Lastly, we would like to emphasize that while having a dense ID representation can be important for OOD detection with clean ID data, it is much more important in the presence of pollution in the ID data. In this polluted setup, some OOD data are mixed in during training and considered ID by the loss. In this case, a compact ID data distribution decreases chance of other OOD data to be considered as ID as much as possible. This is empirically verified in the experimental section.

3.3. SimSiam and BYOL gradient flow problem

To avoid the degenerate solution of a collapsed representation, the authors of SimSiam [6] found it crucial to have a gradient blocking operation (`stop-grad`) that blocks gradients starting from $\hat{z_i}$ from flowing back to other parts of the network. This lack of gradients acts as a regularizer and makes it hard for the optimizer to reach the trivial solution of a collapsed representation. Note that the same analysis presented here equally applies to BYOL [12]: the existence of the momentum encoder enforces an implicit `stop-grad`.

However both [6, 12] do not study the effect `stop-grad` may have on proper gradient flow in the network. Studying Figure 1, we can see that the l_2-norm of the output of the prediction network $\hat{p_i} = p_i/||p_i||_2$ gives proper gradients that flows back to other parts of the network. The exact opposite happens for the l_2-norm of the encoder network's output: all gradients of that operation gets blocked by the `stop-grad`, though it is an integral part of the loss function.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2: Analysis of training and evaluation of SimCLR and SimSiam on CIFAR10 for OOD with and without the proposed normalization. Every point on these plots represents the average from 4 independent classes. First row, SimCLR evaluated with (a) *c-Cos* (b) κ from fitting a vMF (c) linear probe accuracy. Second row, SimSiam evaluated with (d) *c-Cos* (e) κ from fitting a vMF (f) Maximum Mean Discrepancy (MMD) with a uniform distribution.

Proposed solution Simply removing the `stop-grad` operation converges quickly to a collapsed representation [6]. Another apparent straightforward solution is to minimize the norm of the output representation of the encoder f, but this limits the representation ability of the encoder and also rapidly collapses. One last trial would be removing normalization altogether in $\mathcal{D}$, however this converges to a sub-optimal representation with an unbounded norm [12].

Our proposed solution is instead based on a simple observation. The missing gradient from $\hat{z}_i$ carries two pieces of information: (a) moving the output of the encoder z_i to be closer to the output of the predictor p_i (which is the information we want to hide from the optimizer). (b) encouraging the projector g to learn a l_2-norm invariant representation; this facilitates training the predictor network and thus also the encoder, and is what we want to maintain.

In order to maintain (b) while discarding (a), we propose a small modification (see Figure 1): we apply a differentiable l_2-normalization to the projection z_i, such that the new projection is $\hat{z}_i = z_i/||z_i||_2$. This gives the network proper gradients for learning an l_2-normalized representation satisfying the loss function $\mathcal{D}$ while still having the `stop-grad` operation and avoiding related collapse.

3.4. Feature evaluation for out-of-distribution detection

Many recent works on OOD detection are based on scoring a given test sample using its distance to the nearest training sample [38, 35, 22]. This provides a simple evaluation baseline with strong results. Many metrics have been used for this, the most commonly used are the Mahalanobis distance [22, 35] and the Cosine distance [38] evaluated at the output of the network after the last convolutional block.

To take the advantages of both of these distances, we go a step further and score features with the arccosine of the Mahalanobis Cosine similarity (i.e. angular Mahalanobis distance) [3], which have been proposed and studied well previously in the field of face recognition [3, 40]. Mahalanobis Cosine is the Cosine similarity between vectors after projection into the Mahalanobis space, where the famous Mahalanobis distance is the Euclidean distance computed between vectors after also projection into the Mahalanobis space.

If x_m, y are a training and test sample respectively, and the projection of their features $f(x_m), f(y)$ into the Mahalanobis space is u, v using the sample covariance matrix Σ_m and mean μ_m of the training data $\{x_m\}_{m=1}^{M}$, then the Mahalanobis Cosine distance $\mathcal{D}_{MC}$ and our scoring $\mathcal{S}_{\textit{k-Cos}}$ are

$$\begin{aligned} u =& \Sigma_m^{-1/2}(x_m - \mu_m), \\ \mathcal{D}_{MC}(x_m, y) =& \frac{u}{\|u\|_2} \cdot \frac{v}{\|v\|_2}, \\ \mathcal{S}_{\textit{k-Cos}}(y) =& arccos(\max_m \mathcal{D}_{MC}(x_m, y)) \end{aligned} \tag{2}$$

$\mathcal{S}_{\textit{k-Cos}}$ considers only the distance to the nearest training sample and is used for most of our experiments. In Appendices C and G (supplementary material) we study a variant $S_{\textit{c-Cos}}$, that is especially robust to pollution; it computes distance to μ_m, the mean vector of the training data $\{x_m\}$. $\mathcal{S}_{\textit{c-Cos}}(y) = arccos(\mathcal{D}_{MC}(\mu_m, y))$.

Feature ensembling We also propose another evaluation scheme where all the intermediate feature maps of the network are scored independently, then their scores are all summed together. The idea is to get both high level (from final layers) and low level (from initial layers) OOD scores using different feature maps. Note that this is different from [22] that learn a weighted sum of all the feature scores, where the weights are learned on a validation set; our proposed method is a simple sum and doesn't need a validation set. It also attains a huge runtime saving when compared to test-time augmentation (TTA) used in [38] or model ensembling used in [36], as it requires a single model forward pass on a single instance. It consists of three steps:

1. Computing a score for each feature map, using either $\mathcal{S}_{\textit{k-Cos}}$ or $\mathcal{S}_{\textit{c-Cos}}$ or both.
2. Normalizing the range of the scores to be between 0 and 1, using the range of training data scores.
3. Summing the normalized scores.

Due to space limitations we show detailed feature evaluation results in Appendix C, along with comparing different evaluation metrics. We then show extensive ablations in Appendix G. In Section 4.2, we firstly consider experimental results without ensembling, before comparing to methods that use ensembling in Section 4.2.4. Any result that includes ensembling strictly has the $_E$ suffix, which indicates using the *Ens.* combination whose components are precisely described in Appendix C.

4. Empirical evaluation

4.1. Experimental setup

We perform a thorough evaluation of our proposed normalization modifications on SimSiam, BYOL, and SimCLR (with and without negative / shifted augmentations). The problem discussed in Section 3.3 doesn't apply to SimCLR (there is no stop-grad), so while both modifications are applied to SimSiam, our normalized variant of SimCLR only includes the change proposed in Section 3.2 for $\hat{y}$.

We evaluate in the one-vs-all OOD detection setting for CIFAR-10, CIFAR-100 super-classes [19], Fashion-MNIST [45], and SVHN [26]. In this setting one class is treated as the normal class and the rest are treated as outliers.

For all results presented we use a ResNet-18, trained with Adam [16] with a learning rate of 0.0001 and a cosine learning rate decay [25]. For SimCLR we used 2-layer MLP as a projection head with an architecure similar to [36]. For SimSiam we used a 3-layer MLP for both the projector and the predictor. SimCLR is trained for 500 epochs, and SimSiam and BYOL for 4000 epochs. All models are trained on Nvidia V100 16GB GPUs and written in Pytorch [28]. All our reported results are without test-time augmentation.

For brevity, we often refer to SimSiam as **SS** and SimCLR as **SC**. Also **SS(n)**, **SC(n)** and **BYOL(n)** indicate inclusion of the proposed normalization, and **SC(-)** is with negative augmentations. Unless otherwise stated, our ensemble-free results use k-Cos Scoring. More details of the datasets, and additional comparisons/descriptions of competing methods can be found in Appendices B and F (supplementary material). An extensive ablation study including training 640 different models is provided in Appendix G.

4.2. Results

4.2.1 One-Class Classification

Table 1 compares ensemble-free versions of our baselines BYOL, SimSiam, and SimCLR with and without normalization against current state of the art (without ensembling) DROC [36], RotNet [9], and GOAD [2]. Extended comparison of our results (without ensembling) with many other published methods can be found in Appendix F.

We see that adding normalization is consistently effective across BYOL, SS, and SC in all scenarios. Also, SS(n) and BYOL(n) always achieve very competitive results compared to methods that use negative augmentations (SC(-) and DROC) while making much less assumptions about the underlying training data.

Data	**p**	**BYOL**	**BYOL(n)**	**SS**	**SS(n)**	**SC**	**SC(n)**	**SC(-)**	**SC(n-)**	**RotNet**	**DROC**	**GOAD**
C10	0	88.5	90.5	89.5	91.7	86.3	88.9	90.7	**92.9**	89.3	92.5	88.2
	0.1	82.9	85.3	83.2	**86.3**	65.6	79.8	79.6	83.9	78.5	80.5	83.0
C100	0	79.6	80.2	81.4	84.3	80.5	84.0	84.7	**87.0**	81.9	86.5	74.5
	0.1	76.2	77.8	78.9	80.3	75.9	80.0	80.5	**82.8**	-	-	-
fMNIST	0	95.3	95.1	**95.9**	95.0	94.6	94.9	94.7	95.7	94.6	94.5	94.1
	0.1	61.3	73.2	63.0	75.3	46.5	53.1	78.7	**80.9**	-	76.6	-

Table 1: Results of baselines and proposed variants compared to state of the art, without ensembling. **p** is the ratio of outlier pollution data inside the training set. **SS** is SimSiam, **SC** is SimCLR.

		BYOL(n)$_E$	SS(n)$_E$	SC(n)$_E$	SC(n-)$_E$	CSI	STOC
# Inference steps / example		1	1	1	1	160	1
# Trained models / class		1	1	1	1	1	60
Requires a good approximation of p		✗	✗	✗	✗	✗	✓
Assumes rotation-variant data		✗	✗	✗	✓	✓	✓
Data	*p*						
CIFAR10	0.0	91.9	92.5	90.3	93.0	94.3	92.1
	0.1	88.3	88.5	86.7	87.8	84.5	89.9
fMNIST	0.0	96.2	96.1	96.3	95.9	-	95.5
	0.1	87.9	87.8	87.5	90.9	-	85.7
CIFAR100	0.0	83.4	86.6		89.4	89.6	-
	0.1	80.7	82.5	83.0	85.7	-	-

Table 2: Performance compared to state of the art under different ensembling setups. p is the ratio of outliers inside the training data.

4.2.2 Performance under pollution

Table 1 also shows a comparison of the performance under the very realistic scenario that some $p\%$ of the training data are polluted with OOD data. It can be seen that adding the proposed normalization drastically reduces the effect of polluted data, and achieves state-of-the-art results (without ensembling).

One important take-away from the table is that, as predicted by our analysis in Section 3.2, positive only SSL (e.g. SS or BYOL) is much more suited to real world OOD detection which can possibly contain a small subset of polluted data compared to contrastive SSL (without negative augmentations) which suffers big performance drops in this scenario. In the case where negative augmentations are a suitable assumption, proposed SC(n-) gets even better results.

4.2.3 Effect of normalization

Table 3 examines the effect of the two proposed normalizations on SS on different batch size to asses the stability of training. As we can see for both batch-sizes both proposed normalizations do help the performance, and their combinations (proposed) is the best. We can also see for low batch sizes SS without norm suffers a big performance loss, which the proposed normalization fixes; this is a very important property as large batch training is not always an option.

	CIFAR10		CIFAR100	
batch size	32	512	32	512
without norm	77.97	89.56	74.46	81.38
only normalize f	85.76	91.63	78.7	82.27
only normalize g	81.73	89.22	73.1	81.15
normalzie both	92.9	91.7	81.73	84.31

Table 3: An ablation of SimSiam showing the importance of both the normalization schemes proposed in Section 3.2 and 3.3, especially for in low batch-size training.

Trained from scratch		Pre-trained (1M images)				Pre-trained (1B images)	Pre-trained and adapted
		Supervised			Self-supervised		
SS(n)	SC(n)	PT R18	PT R50	PT R152	PT R50 (SC)	PT R50	PT R50
94.9	93.8	61	66	64	65	70	68

Table 4: OOD detection performance of various pre-trained backbones on SVHN vs. training from scratch

4.2.4 Effect of feature ensembling

Table 2 shows results of the feature ensembling proposed in Section 3.4, compared to state-of-the-art techniques that also utilize ensembling: CSI [38] and STOC [48]. For fair comparison, we also state the relative computational budget (training or inference) for each. The proposed feature ensembles offer big computational savings compared to CSI and STOC at roughly the same scores. Lastly, improvements by ensembling are more significant when there are outliers in the training data.

4.2.5 Pretrained backbones on SVHN

Some recent works [30, 8] claim that ImageNet pre-trained models can act as a universal OOD detector and can work well on nearly any in-domain distribution. [15] showed that for classes not present in the pre-training data, pre-trained models perform poorly at OOD detection.

We confirm this here, and show that even for very simple datasets (e.g. SVHN) that require different kinds of discriminative features than those required for natural images, pre-trained models perform very poorly. Table 4 shows that regardless of the model size (ResNet 18 to 152) or size of pre-training dataset (from 1M images to 1B images [47]), or using fully-supervised or SSL pre-training, or even with state-of-the-art feature adaptation after pre-training [30], all catastrophically fail at SVHN compared to training from scratch, which can get a nearly perfect score.

5. Conclusion

We have considered a general framework to detection of Anomalies in images: using various SSL methods as deep feature extractors; followed by metric learning for outlier scoring. For each stage we have studied what does and doesn't work in a variety of scenarios, and proposed remedies and improvements that are also robust in the case of polluted training data and small batch sizes.

We have investigated and studied compactness of ID representation distributions as an important and very sensitive factor to the final OOD detection performance. Our experiments demonstrated that regardless of the quality of learned features, the ID representation compactness is critical. As its distribution gets closer to uniform, the OOD detection performance deteriorates significantly.

We have motivated, proposed, and studied an assumption-free, novel architectural modification for inducing this non-uniformity, and use it to solidly improve performance across contrastive and non-contrastive SSL-based OOD detection. We also studied several variants of feature scoring that work well across these different methods. More importantly, under the real world setting of totally unsupervised AD, where the ID training data can be polluted by some OOD outliers, our proposed modifications provide state-of-the-art performance among all competing methods.

Previous state-of-the-art literature for OOD detection was based on contrastive-based SSL with negative "distributionally-shifted" augmentations (e.g. 90 degree rotations). A big hurdle with the applicability of these methods is that the assumptions they make about the training data can be partially or fully invalid in real-world scenarios. Using our proposed "No Shifted Augmentations" (NSA) modifications, both contrastive and non-contrastive methods get a boost in their baseline performance, making them comparable to negative augmentation based techniques, but much more applicable to open world scenarios where little is controlled about ID data.

While model ensembling or Test-Time Augmentation is known in literature to be very effective for OOD, increased training/inference computational requirements can be often prohibitive. We went further and studied using light-weight multi-level feature ensembling for OOD. This enabled us to show state-of-the-art performance in terms of AUCROC, with huge savings in computational budget.

References

[1] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

[2] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

[3] J Ross Beveridge, David Bolme, Bruce A Draper, and Marcio Teixeira. The csu face identification evaluation system. *Machine vision and applications*, 16(2):128–138, 2005.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learn-

ing of visual representations. In *International Conference on Machine Learning*, 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

[7] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

[8] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *arXiv preprint arXiv:2106.03004*, 2021.

[9] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, 2018.

[10] Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *International Conference on Machine Learning*, pages 3711–3721. PMLR, 2020.

[11] 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.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[13] Sungwon Han, Hyeonho Song, Seungeon Lee, Sungwon Park, and Meeyoung Cha. Elsa: Energy-based learning for semi-supervised anomaly detection. *arXiv preprint arXiv:2103.15296*, 2021.

[14] Md Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, Liming Chen, et al. von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017.

[15] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Xiaodong Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.

[18] Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021.

[19] Alex Krizhevsky et al. Learning multiple layers of features from tiny images, 2009.

[20] Sachin Kumar and Yulia Tsvetkov. Von mises-fisher loss for training sequence to sequence models with continuous outputs. *arXiv preprint arXiv:1812.04616*, 2018.

[21] Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Directional analysis of stochastic gradient descent via von mises-fisher distributions in deep learning. *arXiv preprint arXiv:1810.00150*, 2018.

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

[23] Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *arXiv preprint arXiv:2109.12909*, 2021.

[24] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. *arXiv preprint arXiv:2104.04015*, 2021.

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[27] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv preprint arXiv:1905.10626*, 2019.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[29] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. PANDA: Adapting Pretrained Features for Anomaly Detection and Segmentation. Oct. 2020.

[30] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. *arXiv preprint arXiv:2106.03844*, 2021.

[31] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[32] Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Rethinking assumptions in deep anomaly detection. In *ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning*, 2021.

[33] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

[34] 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*, 2000.

[35] Vikash Sehwag, Mung Chiang, and Prateek Mittal. {SSD}: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021.

[36] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. *arXiv preprint arXiv:2011.02578*, 2020.

[37] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of i s (x). *Computational Statistics*, 27(1):177–190, 2012.

[38] 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*, 33:11839–11852, 2020.

[39] Yu Tian, Guansong Pang, Fengbei Liu, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, Gustavo Carneiro, et al. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. *arXiv preprint arXiv:2103.03423*, 2021.

[40] A Vinay, Vinay S Shekhar, KN Balasubramanya Murthy, and S Natarajan. Performance study of lda and kfa for gabor based face recognition system. *Procedia Computer Science*, 57:960–969, 2015.

[41] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022.

[42] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[43] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, and Simon Kohl. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[45] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[46] Zhisheng Xiao, Qing Yan, and Yali Amit. Do We Really Need to Learn Representations from In-domain Data for Outlier Detection? May 2021.

[47] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[48] Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O Arik, Chen-Yu Lee, and Tomas Pfister. Self-trained one-class classification for unsupervised anomaly detection. *arXiv preprint arXiv:2106.06115*, 2021.

[49] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning*, 2016.

[50] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, 2016.