MEAN-SHIFTED CONTRASTIVE LOSS FOR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep anomaly detection methods learn representations that separate between normal and anomalous samples. It was previously shown that the most accurate anomaly detectors can be obtained when powerful externally trained feature extractors (e.g. ResNets pre-trained on ImageNet) are fine-tuned on the training data which consists of normal samples and no anomalies. Although contrastive learning is currently the state-of-the-art in self-supervised anomaly detection, we show that it achieves poor results when used to fine-tune pre-trained feature extractors. We investigate the reason for this collapse, and find that pre-trained feature initialization causes poor conditioning for standard contrastive objectives, resulting in bad optimization dynamics. Based on our analysis, we provide a modified contrastive objective named the *Mean-Shifted Contrastive Loss*. Our method is highly effective and achieves a new state-of-the-art anomaly detection performance on multiple benchmarks including 97.2% ROC-AUC on the CIFAR-10 dataset.

1 INTRODUCTION

Anomaly detection is a fundamental task for intelligent agents that aims to detect if an observed pattern is normal or anomalous (unusual or unlikely). Anomaly detection has broad applications in scientific and industrial tasks such as detecting new supernovae or genetic mutations, as well as production line inspection and video surveillance. Due to the significance of the task, many efforts have been focused on automatic anomaly detection, particularly on statistical and machine learning methods. A common paradigm used by many anomaly detection methods is measuring the probability of samples and assigning high-probability samples as normal and low-probability samples as anomalous. The quality of the density estimators is closely related to the quality of features used to represent the data. Classical methods used statistical estimators such as K-means, K nearest-neighbors (kNN) or Gaussian mixture models (GMMs) on raw features, however this often results in sub-optimal results on high-dimensional data such as images.

Anomaly detection on high-dimensional data requires high quality features. Many recent methods learn features in a self-supervised way and use them in order to detect anomalies. Unfortunately, anomaly detection datasets are typically small and do not include anomalous samples, resulting in weak features. An alternative is transferring features learned from auxiliary tasks on large-scale external datasets such as ImageNet classification. It was found that fine-tuning the pre-trained features on the normal training data can result in significant performance improvements. Although it may appear natural that this can simply be done by initializing standard anomaly detection techniques with the pre-trained features, it is quite challenging. Reiss et al. (2021) proposed PANDA that combined the DeepSVDD objective (Ruff et al., 2018) with pre-trained features. As the top self-supervised anomaly detection methods use contrastive learning rather than DeepSVDD, we hypothesize that combining pre-trained feature with contrastive methods would achieve the best of both worlds.

We begin with the surprising result that standard contrastive methods, initialized with pre-trained weights, do not improve anomaly detection accuracy at all. An analysis of the learning dynamics reveals that this occurs due to the fact that the standard contrastive loss is poorly suited for data that are concentrated in a compact subspace (which the normal data under strong pre-trained features are). We propose an alternative objective, the mean-shifted contrastive (MSC) loss. The MSC loss is found to achieve better One-Class Classification (OCC) performance than the center-loss (used in DeepSVDD and PANDA), and sets a new anomaly detection state-of-the-art.

Our contributions:

1. We analyze the standard contrastive loss for fine-tuning pre-trained representations for OCC and show that it is poorly initialized and achieves poor performance.
2. Proposing an alternative objective, named the *Mean-Shifted Contrastive Loss* and providing analysis that it is crucial for achieving strong performance for adapting features for OCC.
3. Extensive experiments demonstrating that our method is able to outperform the state-of-the-art anomaly detection performance (e.g. 97.2% ROC-AUC on CIFAR-10).

2 Related Work

Classical anomaly detection methods: Detecting anomalies in images has been researched for several decades. The methods follow three main paradigms: i) Reconstruction - characterizing the normal data by a set of basis functions and then attempts to reconstruct a new example using these basis functions (with sparsity or norm constraints). Anomalies typically have High reconstruction errors. Notable methods include: principal component analysis (Jolliffe, 2011) and K nearest neighbors (kNN) (Eskin et al., 2002). ii) Density estimation - test samples are denoted as anomalous if their estimated density is low. Methods include Ensembles of Gaussian Mixture Models (EGMM) (Glodek et al., 2013), and kernel density estimation (Latecki et al., 2007). iii) OCC - fitting a classifier to discriminate between normal samples and all others. It is then used to classify new samples as normal or anomalous. Such methods include one-class support vector machine (OCSVM) (Scholkopf et al., 2000) and support vector data description (SVDD) (Tax & Duin, 2004).

Self-supervised deep learning methods: Instead of using supervision for learning deep representations, self-supervised methods train neural networks to solve an auxiliary task for which obtaining data is free or at least very inexpensive. Auxiliary tasks for learning high-quality image features include: video frame prediction (Mathieu et al., 2016), image colorization (Zhang et al., 2016; Larsson et al., 2016) and puzzle solving (Noroozi & Favaro, 2016). RotNet (Gidaris et al., 2018) used a set of image processing rotations around the image axis, and predicted the true image orientation to learn high-quality image features. Golan & El-Yaniv (2018) have used similar image-processing task prediction for detecting anomalies in images. This method was improved by Hendrycks et al. (2019), and extended to tabular data by Bergman & Hoshen (2020). Another commonly used self-supervised paradigm is contrastive learning (Chen et al., 2020a), which learns representations by distinguishing similar views of the same samples from other data samples. Recently, variants of contrastive learning were also introduced to OCC. CSI (Tack et al., 2020) treats augmented input as positive samples and the distributionally-shifted input as negative samples. DROC (Sohn et al., 2020) shares a similar technical formulation as CSI without any test-time augmentation nor ensemble of models.

Feature adaptation for one-class classification: Similarly to previous work in multi-class image classification, these OCC methods are first initialized using pre-trained features. Features are then adapted on OCC objectives to improve their accuracy. DeepSVDD (Ruff et al., 2018) suggested to first train an auto-encoder on the normal training data, and then using the encoder as the initial feature extractor. Moreover, since the features of the encoder are not specifically fitted to anomaly detection, DeepSVDD adapts on the encoder training data. However, this naive training procedure leads to catastrophic collapse. An alternative direction, is to use features learned from auxiliary tasks on large-scale external datasets such as ImageNet classification. Deep features representations trained on the ImageNet dataset have been shown by Huh et al. (2016) to significantly boost performance on other datasets that are only vaguely related to some of the ImageNet classes. Transferring ImageNet pre-trained features for out-of-distribution detection has been proposed by Hendrycks et al. (2019). Analogous pre-training for OCC has been proposed by Perera & Patel (2019), where they jointly train anomaly detection with the original task, which achieves only limited adaptation success. PANDA (Reiss et al., 2021) proposed techniques based on early stopping and EWC (Kirkpatrick et al., 2017), a continual learning method, to mitigate catastrophic collapse.

3 Background: Learning Representations for One-Class Classification

3.1 Preliminaries

In the one-class classification task, we are given a set of training samples $x_1, x_2..x_N \in \mathcal{X}_{train}$ that are all normal (and contain no anomalies). The objective is to classify a new sample x as being normal or anomalous. The methods considered here learn a deep representation of a sample parametrized by the neural network function $\phi : \mathcal{X} \to \mathbb{R}^d$, where $d \in \mathbb{N}$ is the feature dimension. In several methods, ϕ is initialized by pre-trained weights ϕ_0, which can be learned either using external datasets (e.g. ImageNet classification) or using self-supervised tasks on the training set. The representation is further tuned on the training data to form the final representation ϕ. Finally, an anomaly scoring function $s(\phi(x))$ determines the anomaly score of sample x. The binary anomaly classification can be predicted by applying a threshold on $s(x)$. In Sec. 3.2 and Sec. 3.3, we review the most relevant methods for learning the representation ϕ.

3.2 Self-supervised Objectives for Anomaly Detection

We review two deep self-supervised objectives relevant to this work:

Center Loss: This loss uses the simple idea, that features should be learned so that normal data lie within a compact region of feature space, whereas anomalous data lie outside it. As we focus on the OCC setting, there are no examples of anomalies in training. Instead, the center loss encourages the features of the normal samples to lie as near as possible to a predetermined center. Specifically, the center loss for an input sample $x \in \mathcal{X}_{train}$ can be written as follows:

$$\mathcal{L}_{center}(x) = \|\phi(x) - c\|^2 \tag{1}$$

This objective suffers from a trivial solution - the features $\phi(x)$ collapse to a singular point c for all samples, normal and anomalous. This is often called "catastrophic collapse". Such a collapsed representation cannot, of course, discriminate between normal and anomalous samples.

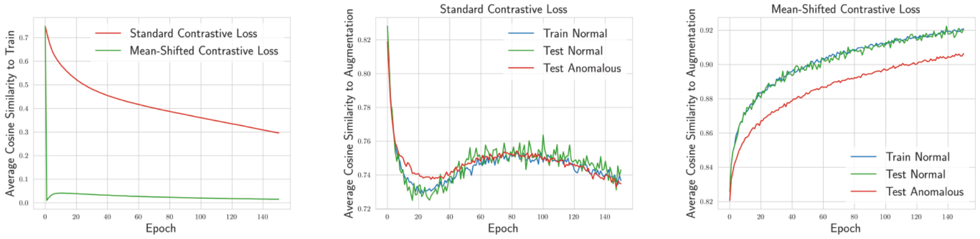
Contrastive Loss: Recently, contrastive learning was responsible for much progress in self-supervised representation learning (Chen et al., 2020a). In the contrastive training procedure a mini-batch of size B is randomly sampled and the contrastive prediction task is defined on pairs of augmented examples derived from the mini-batch, resulting in $2B$ data points. For anomaly detection in the one-class classification setting, the contrastive objective simply states that: i) the angular distance between the features of any positive pair (x'_i, x''_i) should be small ii) the distance between the features of a normal sample x_i and other normal samples x_m should be large. The typical contrastive loss for a positive pair (x'_i, x''_i), where x'_i and x''_i are augmentations of $x_i \in \mathcal{X}_{train}$, is written below:

$$\mathcal{L}_{con}(x'_i, x''_i) = -\log \frac{\exp(sim(\phi(x'_i), \phi(x''_i))/\tau)}{\sum_{m=1}^{2B} \mathbb{1}[i' \neq m'] \cdot \exp(sim(\phi(x'_i), \phi(x'_m))/\tau)} \tag{2}$$

where $\forall m \in [2B] : x'_m$ is an augmented view of some $x_m \in \mathcal{X}_{train}$, τ denotes a temperature hyper-parameter and sim is the cosine similarity. Augmentations include crops, flips, color jitter, grayscale and Gaussian blurs. Contrastive methods currently achieve the top performance for anomaly detection without utilization of externally trained network weights.

3.3 Initialization with Pre-trained Weights

Self-supervised representation learning methods have high sample complexity and in many cases do not outperform supervised representation learning methods. It is common practice in deep learning to transfer the weights of classifiers pre-trained on large, some-what related, labeled datasets to the task of interest. Previous methods used pre-trained weights for anomaly detection (Perera & Patel, 2019; Reiss et al., 2021). It was found that fine-tuning the pre-trained weights of ϕ_0 on the normal data, results in a stronger feature extractor ϕ. The latest approach, PANDA, simply used the center loss (Eq. 1) for fine-tuning the pre-trained weights. Several attractive properties of methods based on ImageNet pre-trained features were established: i) they outperform self-supervised anomaly detection methods by a wide margin, without using any labeled examples of anomalies or outlier exposure. ii) they generalize to datasets that are very different from ImageNet including aerial and medical images.



(a) (b) (c)

Figure 1: CIFAR-10 "Airplane" class. Average cosine similarity between features on training set vs. training epoch. ***(a)*** Similarity between pairs of images. Similarity between images and their augmentation for ***(b)*** Contrastive objective ***(c)*** Mean-shifted contrastive objective.

As contrastive objectives typically perform better than the center loss, it is natural to assume that replacing PANDA's center loss by the contrastive loss would be advantageous. Unfortunately, the representation collapses immediately and this modification achieves poor OCC results. In Sec. 4 we will analyze this phenomenon and present an alternative objective which overcomes this issue.

4 Modifying the Contrastive Loss for Anomaly Detection

In this section, we introduce our new approach for OCC feature adaptation. In Sec 4.1 we analyze the mechanism that prevents standard contrastive objectives from benefiting from pre-trained weights for OCC. In Sec 4.2 we present our new objective function, the mean-shifted contrastive (MSC) loss. In Sec 4.3 we analyze the the proposed mean-shifted contrastive loss for OCC transfer learning.

4.1 Adaptation Failure of the One-Class Classification Contrastive Loss

While contrastive methods have achieved state-of-the-art performance on visual recognition tasks, they are not apriori designed for feature adaptation for OCC. In this section, we analyze the following phenomenon: when optimizing a contrastive objective for the OCC setting of anomaly detection with ImageNet pre-trained features, the representations do not only fail to improve, but degrade quickly.

To understand this phenomenon, we present in Fig. 1 plots of two metrics as a function of training epoch: i) uniformity: the average cosine similarity between the features of pairs of examples in the training set (more uniform = close to zero) ii) augmentation distance: the average cosine similarity between features of train samples and their augmentation (higher generally means better ordering of feature space). Wang & Isola (2020) showed the contrastive loss optimizes two properties i) uniform distribution of $\{\phi(x)\}_{x \in \mathcal{X}_{train}}$ across the unit sphere. ii) different augmentations of the same images mapping to the same representation. We can see that in our OCC setting, contrastive training significantly improved the uniformity of the distribution of training images but failed to increase the similarity between the features of images and their augmentation. Results for other temperature values is presented in Appendix A.3. This shows that contrastive training in this case did not make features more discriminative, suggesting the training objective is not well specified.

We provide an intuitive explanation for the empirical observation. It is common that the normal data occupy a compact region in the ImageNet pre-trained feature space. When viewed in the spherical coordinate system having its center at the origin, normal images span only a small, bounded region of the sphere. As one of the objectives of contrastive learning is to have features that occupy the entire sphere, the optimization would be focused on changing the features accordingly, putting far less emphasis on improving the features so that they are invariant to augmentations. This is not good for anomaly detection as this uniformity actually makes anomalies harder to detect (as they become less likely to occupy a sparse region of the feature space). Additionally, such drastic changes of the features cause the loss of the useful properties of the ImageNet pre-trained feature space. This is counter to the objective of transferring strong auxiliary features.

4.2 The Mean-Shifted Contrastive Loss for better adaptation

To overcome the limitations of contrastive learning explained above, we propose a simple modification of its objective for OCC feature adaptation. In our modified objective, we compute the angles between the features of images with respect to the **center** of the normal features rather than the (Wang &

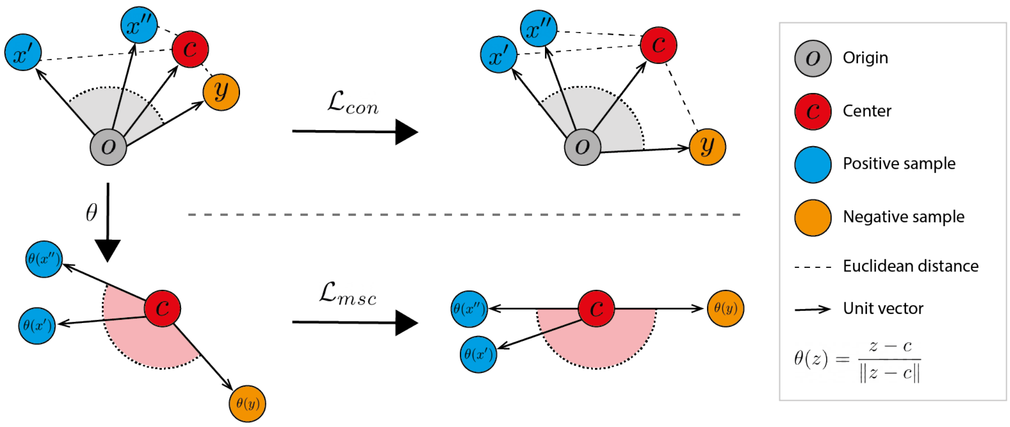


Figure 2: **Top:** The angular representation in relation to the origin. $\mathcal{L}_{con}$ enlarging the angles between positive and negative samples, thus increasing their Euclidean distance to c. **Bottom:** The mean-shifted representation. $\mathcal{L}_{msc}$ does not affect the Euclidean distance between c and the mean-shifted representations while maximizes the angles between the negative pairs.

Isola, 2020) (as done in the original contrastive loss). Although this can be seen as a simple shift of the original objective, we will show that it resolves the critical issues highlighted above and allows contrastive learning to benefit from the powerful, pre-trained feature initialization (See Sec. 4.3). We name this new objective, the *Mean-Shifted Contrastive* (MSC) loss.

Let us denote the center of the normalized feature representations of the training set by c:

$$c = \mathbb{E}_{x \in \mathcal{X}_{train}}[\frac{\phi_0(x)}{\|\phi_0(x)\|}] \tag{3}$$

where ϕ_0 is the initialized pre-trained model. For each image x, we create two different augmentations of the image, denoted x', x''. All the augmented images are first passed through a feature extractor ϕ. They are then scaled to the unit sphere by ℓ_2 normalization (see Sec. 5.2 for the motivation of using ℓ_2 normalization). We mean-shift each representation, by subtracting the center c from each normalized feature representation. The mean-shifted contrastive loss for two augmentations x_i', x_i'' of image x_i from an augmented mini-batch of size $2B$ is defined as follows:

$$\mathcal{L}_{msc}(x_i', x_i'') = -\log \frac{\exp(sim(\frac{\phi(x_i')}{\|\phi(x_i')\|} - c, \frac{\phi(x_i'')}{\|\phi(x_i'')\|} - c))/\tau)}{\sum_{i=1}^{2B} \mathbb{1}[i' \neq m'] \cdot \exp(sim(\frac{\phi(x_i')}{\|\phi(x_i')\|} - c, \frac{\phi(x_m')}{\|\phi(x_m')\|} - c))/\tau)} \tag{4}$$

where τ denotes a temperature hyper-parameter and sim is the cosine similarity.

Anomaly criterion: To classify a sample as normal or anomalous, we use the cosine similarity from a set of K suitably selected training exemplars $N_k(x)$. The set $N_k(x)$ can be selected by K nearest-neighbors (more accurate) or K-means (faster). We compute the cosine similarity between the features of the target image x and the K exemplars $N_k(x)$. The anomaly score is given by:

$$s(x) = \sum_{\phi(y) \in N_k(x)} 1 - sim(\phi(x), \phi(y)) \tag{5}$$

where sim is the cosine similarity. By checking if the anomaly score $s(x)$ is larger than a threshold, we determine if the image x is normal or anomalous. A comparison between the different exemplar selection methods is presented in Sec. 5.2.

4.3 Understanding the Mean-Shifted Contrastive Loss

Here, we compare the mean-shifted contrastive loss and the standard contrastive loss.

Uniformity: Optimizing pre-trained weights with the standard contrastive loss focuses on optimizing uniformity around the origin-centered sphere but hurts feature semantic similarity (Sec. 4.1). The mean-shifted loss proposes a simple but very effective solution - evaluating uniformity in the coordinate frame around the data-center. In this frame the features are already roughly uniform, making the

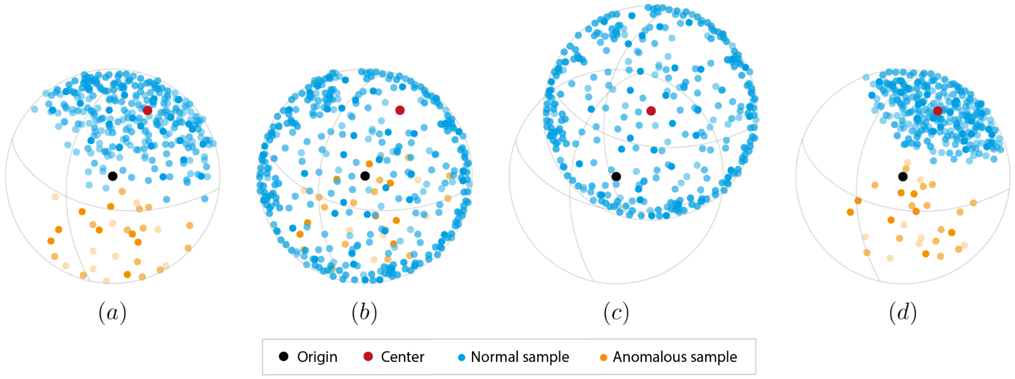


Figure 3: An illustration of our feature adaptation. ***(a)*** The initialized feature space derived by ϕ_0. ***(b)*** $\mathcal{L}_{con}$ forces $\{\phi(x)\}_{x \in \mathcal{X}_{train}}$ to be equally distributed across the unit sphere, resulting: i) the loss of the useful properties of the pre-trained model features space. ii) that every anomalous sample $\hat{x} \notin \mathcal{X}_{train}$ will have a nearby normal sample ***(c)*** $\mathcal{L}_{msc}$ operates in the space of angles around the center in which the features are scattered across the unit sphere surrounding the center, thus focusing on improving the features. ***(d)*** Projecting the mean-shifted features to the unit sphere after optimizing $\mathcal{L}_{msc}$ yields an informative compact representation of normal samples features around the center.

optimization focus on improving the semantic similarity of features. In Fig.1 we see that the features are uniform right from initialization according to our objective (low cosine similarity between normal examples). The optimization can thus focuses on improving the features.

Compactness around center: The standard contrastive loss maximizes the angles between representations of negative pairs even when they are both normal training images. By maximizing these angles, the distance to the center increases as well, as illustrated in Fig. 2 (top). This behaviour is in contrast to the optimization of the center loss (Eq. 1), which learns representations by minimizing the Euclidean distance between normal representations and the center. Reiss et al. (2021) showed that optimizing the center loss results in high anomaly detection performance. Our proposed loss does not suffer from this issue. Instead of measuring the angular distance between samples in relation to the origin, we measure the angular distance in relation to the center of the normal features. As can be seen in Fig. 2 (bottom), our proposed mean-shifted contrastive loss maximizes the angles between the negative pairs while preserving their distance to the center.

A further illustration of the above analysis is presented in Fig. 3. We can see that contrastive learning forces normal features further away from the center and makes them uniform around the origin. This in fact increases the overlap between normal and anomalous samples. On the other hand, with our mean-shifted contrastive, the normal features are encouraged to lie in a compact region around the center rather than around the origin. This makes the normal features lie in a more compact region and decreases the overlap between normal and anomalous samples.

5 Experiments

In this section, we extensively evaluate our method and demonstrate that it outperforms the state-of-the-art. In Sec. 5.1, we report our OCC results with a comparison to previous works on the standard benchmark datasets. In Sec.5.2 we further analyze our objective and we present an ablation study.

Building up on the framework suggested in (Reiss et al., 2021), we use ResNet152 pre-trained on ImageNet classification task as ϕ_0, and adding an additional final ℓ_2 normalization layer - this is our initialized feature extractor ϕ. By default, we fine-tune our model with $\mathcal{L}_{msc}$ (as in Eq. 4). For inference we use the criterion described in Sec. 4.2. We adopt the ROC-AUC metric as detection performance score. Full training and implementation details are in Appendix A.1

5.1 Main Results

We evaluated our approach on a wide range of anomaly detection benchmarks. Following (Golan & El-Yaniv, 2018; Hendrycks et al., 2019) we run our experiments on commonly used datasets: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 coarse-grained version that consists of 20 classes

Table 1: Anomaly detection performance (mean ROC-AUC %, ours is averaged over five runs)

Dataset	Self-supervised				Pre-trained	
	DeepSVDD	MHRot	DROC	CSI	PANDA	Ours
CIFAR-10	64.8	90.1	92.5	94.3	96.2	**97.2**±0.1
CIFAR-100	67.0	80.1	86.5	89.6	94.1	**96.4**±0.1
CatsVsDogs	50.5	86.0	89.6	86.3	97.3	**99.3**±0.0

(Krizhevsky et al., 2009), and CatsVsDogs (Elson et al., 2007). Following standard protocol, multi-class dataset are converted to anomaly detection by setting a class as normal and all other classes as anomalies. This is performed for all classes, in practice turning a single dataset with C classes into C datasets. Full dataset descriptions are in Appendix A.1.1 We compare our approach with the top current self-supervised and pre-trained feature adaptation methods (Ruff et al., 2018; Hendrycks et al., 2019; Tack et al., 2020; Sohn et al., 2020; Reiss et al., 2021). Results that were reported in the original papers were copied. When the results were not reported, we ran the experiments ourselves.

Tab. 1 shows that our proposed approach surpasses the previous state-of-the-art on the common OCC benchmarks. This establishes the superiority of our approach, resulted by our new objective, over previous self-supervised and pre-trained methods. Full class-wise results are in Appendix A.1.5.

Table 2: Anomaly detection accuracy (mean ROC-AUC %) on small dataset. Self-supervised methods fail while adapting pre-trained features achieves strong results. Bold denotes the best results.

	DIOR	MvTec	CIFAR-10 (200 Train samples)	CIFAR-10 (500 Train samples)
CSI	78.5	63.6	81.8	88.1
PANDA	94.3	86.5	95.4	95.6
Ours	**97.2**	**87.2**	**96.5**	**96.7**

5.2 Further Analysis & Ablation Study

Small datasets. In order to demonstrate different challenges in image anomaly detection, we further extend our results on small datasets following the standard protocol. We tested our method on: MVTec (Bergmann et al., 2019) and DIOR (Li et al., 2020). Furthermore, we used the CIFAR-10 dataset with different amount of training data. In Tab. 2 we present a comparison between (i) top self-supervised contrastive-learning based method - CSI (ii) top OCC feature adaptation method - PANDA (iii) our method. We see that the self-supervised method does not perform well on such small datasets, whereas our method achieves very strong performance. The reason for the poor performance of self-supervised methods on small datasets, is due to the fact that the only training data they see is the small dataset, and they cannot learn strong features using such a small amount of data. This is particularly severe for contrastive methods (but is also the case for all other self-supervised methods). As pre-trained methods transfer features from external datasets, they do not have this failure mode.

Optimization from scratch. The mean-shifted objective assumes that relative distance to the center of the features is correlated with high detection performance. When initializing the center as a random Gaussian vector we lose this strong prior, as a result, the detection capabilities are drastically degraded. Therefore when training a model from scratch without any strong initialization that comes from a pre-trained model, our objective does not improve over standard contrastive losses. The mean-shited contrastive loss is therefore a directed contribution to anomaly detection from pre-trained features.

Does the superiority of pre-trained features extend to very different domains? It has already been established in Reiss et al. (2021) that anomaly detection methods based on ImageNet pre-trained features perform very well on distant domains (medical, aerial, industrial datasets). Our results on DIOR and MVTec that are significantly different from ImageNet provide further evidence.

Catastrophic collapse & Early Stopping. Similarly to other OCC pre-trained feature adaptation methods (e.g. PANDA), our method suffers from catastrophic collapse for a very large number of training epochs. However, our method is less sensitive than PANDA, as we dominate PANDA at any point in the curve and collapse much more slowly.See Appendix A.2 for more details.

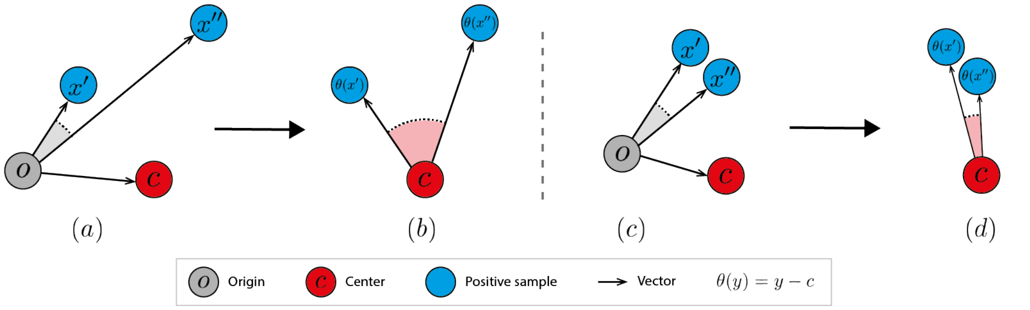


Figure 4: Sensitivity of the mean-shifted loss to class confidence. (**a**): The angular representation in relation to the origin without confidence normalization. (**b**): The mean-shifted representation enlarges the angle between the positive samples. (**c**): The angular representation after confidence normalization. (**d**): The angle between the positive samples is approximately preserved after mean-shifting.

Why do self-supervised OCC models not suffer from catastrophic collapse? pre-trained methods start from highly discriminative features and can therefore lose accuracy whereas self-supervised features start from random features and therefore have nothing to forget. Another way of looking at it, is that pre-trained initialization creates a useful inductive bias that may erode as a function of training. But this useful bias is only present for pre-trained and not for self-supervised methods.

The Angular Representation. Our initial feature extractor ϕ_0 is pre-trained on a classification task (specifically ImageNet classification). To obtain class probabilities from the features $\phi_0(x)$, which are subsequently multiplied by classifier matrix C and passed through a softmax layer. The logits are therefore given by $C\phi_0(x)$. As softmax is a monotonic function, scaling of the logits does not change the order of probabilities. However, scaling does determine the degree of confidence in the decision. We propose to disambiguate the representation $\phi_0(x)$ into two components: i) the semantic class $\frac{\phi_0(x)}{\|\phi_0(x)\|}$, and the confidence $\|\phi_0(x)\|$. The confidence acts as a per-sample temperature that determines how confident the discrimination between the classes is. A thorough investigation that we conducted, showed that the confidence of an ImageNet pre-trained feature representation did not help the anomaly detection performance. In Fig. 5, we compare the histogram of confidence values between the normal and anomalous values on a particular class of the CIFAR-10 dataset ("Bird"). We observe that confidence does not discriminate between normal and anomalous images in this dataset. In Fig. 4 we demonstrate the sensitivity of the mean-shifted representation to the class confidence. This emphasizes the importance of confidence normalization for the mean-shifted contrastive optimization.

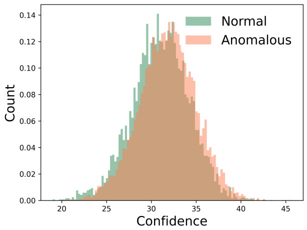


Figure 5: Confidence histogram of CIFAR-10 "Bird" class. The ℓ_2 norm confidence of the extracted features derived by ϕ does not differentiate between normal and anomalous samples.

We thus propose to use the angular center loss. The angular center loss encourages the angular distance between each sample and the center to be minimal. This contrasts with the standard center loss (used by PANDA and DeepSVDD), which uses the Euclidean distance. Although a simple change, the angular center loss achieves much better results than the regular center loss (see Tab. 3).

$$\mathcal{L}_{angular} = -\phi(x) \cdot c \tag{6}$$

Rotation-prediction methods do not benefit from pre-trained features. Self-supervised contrastive methods use rotation-prediction as a way to address the uniformity issue highlighted here (Tack et al., 2020; Sohn et al., 2020). Although it may appear that using pre-trained features might improve OCC methods that rely on rotation-prediction, this is in fact not the case. The reason is that features that generalize better, achieve better performance on rotation-prediction for both normal and anomalous data. Pre-training therefore decreases the gap between the performance of normal and anomalous images on rotation prediction than randomly-initialized networks. This gap is used for discriminating between normal and anomalous samples, and its decrease leads to degraded anomaly detection performance. Specifically, we found that CSI with ImageNet pre-trained features achieves 89.5% average result on CIFAR-10 compared to the standard version which results with 94.3%.

Table 3: Training objective ablation study (CIFAR-10, mean ROC-AUC %).

Dataset	DN2		PANDA		$\mathcal{L}_{msc}$	$\mathcal{L}_{msc} + \mathcal{L}_{angular}$
	Raw	Angular	$\mathcal{L}_{center}$	$\mathcal{L}_{angular}$		
CIFAR-10	92.5	95.8	96.2	96.8	97.2	**97.5**

Self-supervised methods do not benefit from large architectures. Pre-trained models can use large deep networks, a quality that OCC self-supervised methods lack. Since OCC benchmarks are not large, self-supervised methods do not benefit from bigger networks. We tested this by evaluating CSI with different ResNet backbone sizes (ResNet18, ResNet50, ResNet152). The CSI results were the same for all backbones sizes 94.3% ROC-AUC on CIFAR-10. This is in contrast to the effect of pre-trained feature adaptation in our method which benefits from bigger pre-trained models.

Training objective. An ablation of the objectives and of DN2 (kNN on unadapted ImageNet pre-trained ResNet features) is presented in Tab. 3. Note that both the confidence-invariant form of DN2 and PANDA outperform their Euclidean versions. We further notice that the mean-shifted loss outperforms the rest, and combining it with the angular center loss results in further improvements.

Multi-Modal Anomaly Detection. We evaluate the setting where all classes are designated as normal apart from a single class that is taken as anomalous. Note that we do not provide the class labels of the different classes that compose the normal class, rather we consider them to be a single multi-modal class. This setting is more challenging than the standard uni-modal setting as the normal class is complex and consists of many different unlabeled types of data. For each experiment, we denoted a single CIFAR-10 class as anomalous and all nine other CIFAR-10 classes as normal. We report the mean ROC-AUC% over the 10 experiments in Tab. 4. In this case PANDA does not improve results over the DN2 (with cosine distance) as its uni-modal assumption is no longer satisfied. This is because the normal set contains nine classes rather than one. On the other hand, our mean-shifted contrastive loss does not rely on the uni-modal assumption to the same extent leading to much better results. Moreover, self-supervised methods do not preserve their performance on a multi-modal distribution, and are outperformed by pre-trained deep features.

Table 4: Multi-Modal Anomaly detection accuracy (mean ROC-AUC %).

Dataset	DN2		PANDA		Self-Supervised		Ours
	Raw	Angular	$\mathcal{L}_{center}$	$\mathcal{L}_{angular}$	MHRot	CSI	$\mathcal{L}_{msc}$
CIFAR-10	76.2	80.4	78.5	78.0	76.7	79.0	**85.3**

Detection scoring functions. kNN has well established approximations that mitigate its inference time complexity. A simple, but effective solution is reducing the set of gallery samples via k-means. In Tab. 5 we present a comparison of performance of our method and its K-means approximations with the features of the normal training images compressed using different numbers of means (k). We use the sum of the distances to the nearest neighbor means as the anomaly score. We can see that significant inference time improvement can be achieved for a small loss in accuracy.

Table 5: CIFAR-10 Anomaly detection accuracy with K-means (mean ROC-AUC %)

$k = 1$	$k = 5$	$k = 10$	$k = 100$	Full train set
94.2	95.8	96.1	97.0	97.2

6 Conclusion

We presented a novel feature adaptation approach for deep anomaly detection. First, we conducted a thorough analysis of the standard contrastive loss and showed that it poorly initialized for OCC feature adaptation. Second, we introduced an alternative objective, the *Mean-Shifted Contrastive Loss*, that overcomes the limitations of the standard contrastive loss. Finally, we performed extensive experiments demonstrating that our method achieves the top anomaly detection performance.

REFERENCES

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *ICLR*, 2020.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

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

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

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

Jeremy Elson, John R Douceur, Jon Howell, and Jared Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. In *ACM Conference on Computer and Communications Security*, volume 7, pp. 366–374, 2007.

Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pp. 77–101. Springer, 2002.

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

Michael Glodek, Martin Schels, and Friedhelm Schwenker. Ensemble gaussian mixture models for probability density estimation. *Computational Statistics*, 28(1):127–138, 2013.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *NeurIPS*, 2018.

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.

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

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019.

Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

Ian Jolliffe. *Principal component analysis*. Springer, 2011.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016.

Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 61–75. Springer, 2007.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

Pramuditha Perera and Vishal M Patel. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, 2019.

Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2806–2814, 2021.

Lukas Ruff, Nico Gornitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, 2018.

Bernhard Scholkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *NIPS*, 2000.

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.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *arXiv preprint arXiv:2007.08176*, 2020.

David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 2004.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2495–2504, 2021.

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

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

A Appendix

A.1 Experimental Details

A.1.1 Dataset Descriptions

Standard datasets: We evaluate our method on a set of commonly used datasets: *CIFAR-10* (Krizhevsky et al., 2009): Consists of RGB images of 10 object classes. *CIFAR-100* (Krizhevsky et al., 2009): We use the coarse-grained version that consists of 20 classes. *DogsVsCats*: High resolution color images of two classes: cats and dogs. The data were extracted from the ASIRRA dataset (Elson et al., 2007), we split each class to the first 10,000 images as train and the last 2,500 as test.

Small datasets: To further extend our results, we compared the methods on a number of small datasets from different domains. *MvTec* (Bergmann et al., 2019): This dataset contains 15 different industrial products, with normal images of proper products for train and $1 - 9$ types of manufacturing errors as anomalies. The anomalies in MvTec are in-class i.e. the anomalous images come from the same class of normal images with subtle variations. *DIOR* (Li et al., 2020): We pre-processed the DIOR aerial image dataset by taking the segmented object in classes that have more than 50 images with size larger than 120×120 pixels.

A.1.2 Baselines

DROC (Sohn et al., 2020): We used the numbers reported in the paper.

For the evaluation of the other competing method, we trained using the official repositories of their authors and make an effort to select the best configurations available.

DeepSVDD (Ruff et al., 2018): We resize all the images to 32×32 pixels and use the official pyTorch implementation with the CIFAR-10 configuration.

MHRot (Hendrycks et al., 2019): An improved version of the original RotNet approach. For high-resolution images we used the current GitHub implementation. For low resolution images, we modified the code to the architecture described in the paper, replicating the numbers in the paper on CIFAR-10.

CSI (Tack et al., 2020), *PANDA* (Reiss et al., 2021): We run the code and used the exact protocol as described in the official repositories.

A.1.3 Implementation details

We fine-tune the two last blocks of an ImageNet pre-trained ResNet152 with an additional ℓ_2 normalization layer for 25 epochs by minimizing $\mathcal{L}_{msc}$ where the temperature τ is set as 0.25. We use SGD optimizer with weight decay of $w = 5 \cdot 10^{-5}$, and no momentum. The size of the mini-batches is set to be 64. We adopt the data augmentation module proposed by Chen et al. (2020b); we sequentially apply a 224×224-pixel crop from a randomly resized image, random color jittering, random grayscale conversion, random Gaussian blur and random horizontal flip. Finally, for anomaly scoring we use kNN with $k = 2$ nearest neighbours.

A.1.4 Training Resources

Training each dataset class presented in this paper takes approximately 3 hours on a single NVIDIA RTX-2080 TI.

A.1.5 Per-class results

In Tab. 6, Tab. 7, Tab.8 we present the per-class results of CIFAR-10, CIFAR-100, CatsVsDogs respectively.

Table 6: CIFAR-10 anomaly detection performance (mean ROC-AUC %). Bold denotes the best results.

	DeepSVDD	MHRot	DROC	CSI	PANDA	Ours
0	61.7	77.5	90.9	89.9	**97.4**	97.0
1	65.9	96.9	98.9	**99.1**	98.4	98.7
2	50.8	87.3	88.1	93.1	93.9	**94.8**
3	59.1	80.9	83.1	86.4	90.6	**94.3**
4	60.9	92.7	89.9	93.9	**97.5**	96.9
5	65.7	90.2	90.3	93.2	94.4	**97.2**
6	67.7	90.9	93.5	95.1	97.5	**98.2**
7	67.3	96.5	98.2	**98.7**	97.5	98.3
8	75.9	95.2	96.5	97.9	97.6	**98.5**
9	73.1	93.3	95.2	95.5	97.4	**98.3**
Mean	64.8	90.1	92.5	94.3	96.2	**97.2**

Table 7: CIFAR-100 coarse-grained version anomaly detection performance (mean ROC-AUC %). Bold denotes the best results.

	DeepSVDD	MHRot	DROC	CSI	PANDA	Ours
0	66.0	77.6	82.9	86.3	91.5	**96.0**
1	60.1	72.8	84.3	84.8	92.6	**95.3**
2	59.2	71.9	88.6	88.9	98.3	**98.1**
3	58.7	81.0	86.4	85.7	96.6	**97.9**
4	60.9	81.1	92.6	93.7	96.3	**97.6**
5	54.2	66.7	84.5	81.9	94.1	**96.8**
6	63.7	87.9	73.4	91.8	96.4	**98.5**
7	66.1	69.4	84.2	83.9	91.2	**93.4**
8	74.8	86.8	87.7	91.6	94.7	**97.2**
9	78.3	91.7	94.1	95.0	94.0	**96.2**
10	80.4	87.3	85.2	94.0	96.4	**97.1**
11	68.3	85.4	87.8	90.1	92.6	**96.4**
12	75.6	85.1	82.0	90.3	93.1	**95.8**
13	61.0	60.3	82.7	81.5	89.4	**92.6**
14	64.3	92.7	93.4	94.4	98.0	**99.0**
15	66.3	70.4	75.8	85.6	89.7	**92.5**
16	72.0	78.3	80.3	83.0	92.1	**95.2**
17	75.9	93.5	97.5	97.5	97.7	**98.4**
18	67.4	89.6	94.4	95.9	94.7	**97.6**
19	65.8	88.1	92.4	95.2	92.7	**97.0**
Mean	67.0	80.1	86.5	89.6	94.1	**96.4**

Table 8: CatsVsDogs anomaly detection performance (mean ROC-AUC %). Bold denotes the best results.

	DeepSVDD	MHRot	DROC	CSI	PANDA	Ours
Cat	49.2	87.7	91.7	85.7	99.2	**99.4**
Dog	51.8	84.2	87.5	86.9	95.4	**99.2**
Mean	50.5	86.0	89.6	86.3	97.3	**99.3**

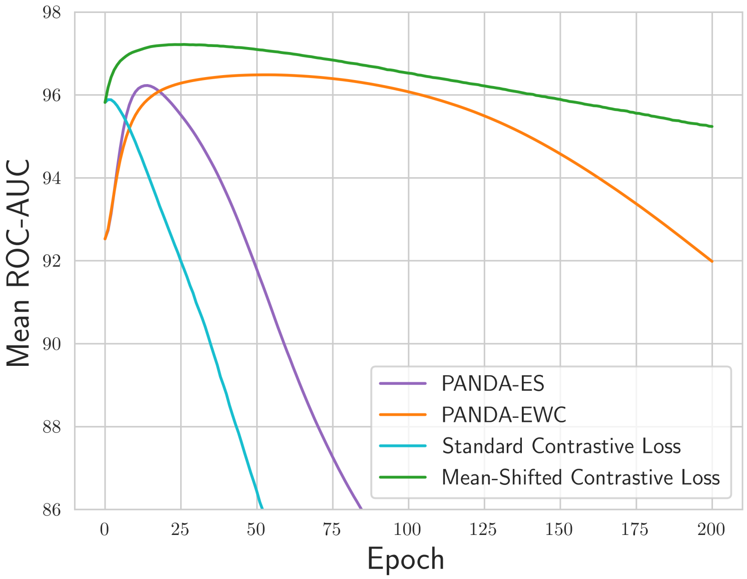


Figure 6: CIFAR-10 Mean ROC-AUC %. Catastrophic collapse of various objective functions.

A.2 Catastrophic Collapse

In Fig. 6, we evaluated the collapse of different training objectives averaged on all CIFAR-10 classes. We notice that the contrastive loss is unsuitable for OCC feature adaptation as it results in very fast catastrophic collapse. PANDA-ES (early-stopping) results in initial improvement in accuracy, but after few epochs the features degrade and become uninformative. PANDA-EWC postpones the collapse, but does not prevent it. Finally, we see that the mean-shifted contrastive loss dominates PANDA at any point in the curve and collapses much more slowly. We find that early stopping after 25 iterations typically gets very close to the optimal accuracy.

A.3 The Temperature parameter and Uniformity

The temperature τ has an important role in the contrastive objective. It was previously shown by Wang & Liu (2021) that it influences both the uniformity of sample distribution on the hypersphere and the weight given to hard negative samples. When the temperature approaches infinity, the model pays equal attention to the negative samples and when it approaches zero, the model ignores all the negative samples but the one with the maximum similarity. Based on this analysis, as the temperature increases, the feature space distribution tends to be less uniform, and when τ is small, the feature space distribution is closer to a uniform distribution. This suggests that using a small temperature parameter while optimizing the standard contrastive objective would solve the optimization dynamics failure that the above suffers from. This in fact not the case, in Fig. 7.a we present an ablation study of different temperature parameters while optimizing the standard contrastive loss. We observe that using a smaller τ slightly helps uniformity but not enough to make the optimization focus on improving the features so that they are invariant to augmentations, as catastrophic collapse still occurs (Fig. 7.b).

A.4 Negative samples

In additional to contrastive self-supervised learning methods such as SimCLR (Chen et al., 2020a) and MoCo (He et al., 2019), other non-contrastive methods have been proposed (e.g. BYOL (Grill et al., 2020) and Sim-Siam (Chen & He, 2020)) which only use positive pairs but no negative pairs. We evaluated our method with Sim-Siam (using the mean-shifted representations), which is the same as using our loss without negative examples. We found that the method experiences an immediate

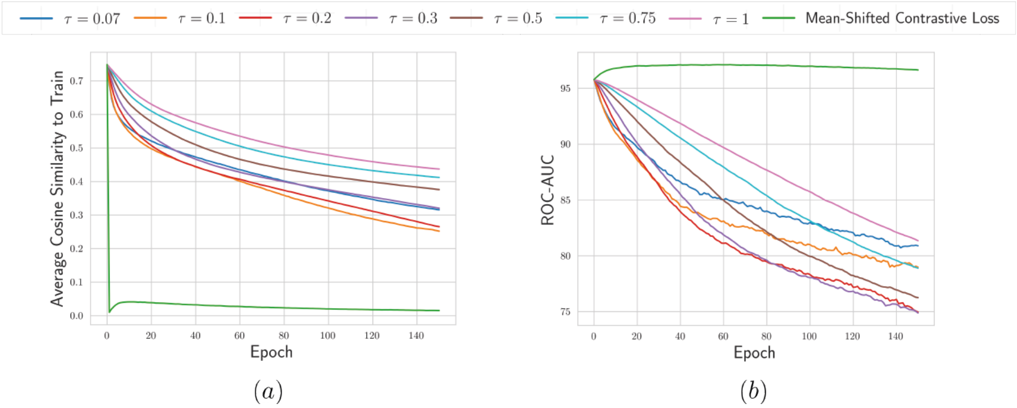


Figure 7: CIFAR-10 "Airplane" class. Ablation study of different temperature parameters while optimizing standard contrastive loss and mean-shifted contrastive loss with $\tau = 0.25$. ***(a)*** Similarity between pairs of images. ***(b)*** The standard contrastive objective is unsuitable for OCC feature adaptation as it results in very fast catastrophic collapse independently of the chosen τ.

catastrophic collapse. This indicates that negative examples are necessary for good performance when using mean-shifted representations. To give some intuition, note that the Sim-Siam objective (with or without mean-shifted representations), can in fact be optimized by having all representations mapped to a constant value. Although it does not happen when Sim-Siam is initialized from scratch, it appears that in the OCC case, it does degrade to the trivial solution. This establishes the need for a contrastive approach.