# OOD Aware Supervised Contrastive Learning

Soroush Seifi
soroush.seifi@external.toyota-europe.com

Daniel Olmeda Reino
daniel.olmeda.reino@toyota-europe.com

Nikolay Chumerin
nikolay.chumerin@toyota-europe.com

Rahaf Aljundi
rahaf.al.jundi@toyota-europe.com

Toyota Motor Europe

## Abstract

*Out-of-Distribution (OOD) detection is a crucial problem for the safe deployment of machine learning models identifying samples that fall outside of the training distribution, i.e. in-distribution data (ID). Most OOD works focus on the classification models trained with Cross Entropy (CE) and attempt to fix its inherent issues. In this work we leverage powerful representation learned with Supervised Contrastive (SupCon) training and propose a holistic approach to learn a classifier robust to OOD data. We extend SupCon loss with two additional contrast terms. The first term pushes auxiliary OOD representations away from ID representations without imposing any constraints on similarities among auxiliary data. The second term pushes OOD features far from the existing class prototypes, while pushing ID representations closer to their corresponding class prototype. When auxiliary OOD data is not available, we propose feature mixing techniques to efficiently generate pseudo-OOD features. Our solution is simple and efficient and acts as a natural extension of the closed-set supervised contrastive representation learning. We compare against different OOD detection methods on the common benchmarks and show state-of-the-art results.*

## 1. Introduction

Modern deep learning architectures have demonstrated great generalization performance, surpassing human baselines on different tasks [6, 19, 57]. However, these models are often trained and evaluated in a closed-set setting, where both train and test sets are assumed to be drawn from the same distribution (*i.e*., in-distribution data).

When encountered with examples coming from any other distribution (*i.e*., out-of-distribution data), these models tend to give predictions that are highly confident but not reliable [54]. In an autonomous driving scenario, OOD samples might include new object classes, road signs or traffic conditions that the model has not seen during training. Therefore, for such safety-critical applications, it is vital to detect the OOD samples, avoid making predictions on them and possibly ask for human intervention instead.

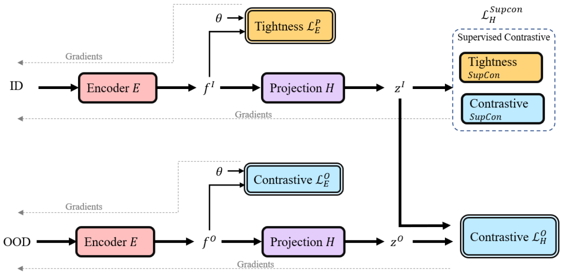

Figure 1. An illustration of our OOD-aware Prototypical Supervised Contrastive Learning method. We consider an encoder network extracting features from the input samples. The features are projected to a vector to which a supervised contrastive learning loss $\mathcal{L}^{\text{SupCon}}$ is applied. Rather than using Cross-Entropy to learn a classifier on top of the features, we learn class prototypes $\boldsymbol{\theta}$ by applying a tightness term $\mathcal{L}_E^P$ to the in-distribution samples. This penalizes features that are far from others of the same class. In addition, we propose to use a contrastive term $\mathcal{L}_H^O$ to push apart projections coming from ID and OOD samples respectively. Likewise, we minimize the maximal similarity of OOD features with the closest class prototypes using loss $\mathcal{L}_E^O$. The proposed new terms are marked with double outline.

Cross-Entropy is a popular choice to train classification models under the closed-set assumption. Popular datasets used in the closed-set setting [13, 14, 28] have mutually exclusive classification labels that can be one-hot encoded. A perfect fit for CE. However, when a model is trained to always select an object class with a confidence close to 1 for any input, it will likely produce highly confident predictions for OOD data as well [49]. Besides, CE is shown to be sensitive to noise and susceptible to overfitting [4].

Supervised contrastive training [26] has been shown to improve the performance of a classification model by learning a rich representation of the samples. The core idea is to leverage a large number of sample pairs and push sample representations of the same class to lie close together and far from the others in the embedding space. This has been recently demonstrated to improve the OOD detection performance [43, 48, 50] although it remains open how to explicitly employ the learned embedding for a better OOD detection.

Recent OOD detection methods, exposing the model to auxiliary OOD data during training, do not leverage the strength of representations learned with contrastive learning as they are tailored for the Softmax Cross-Entropy loss [23, 32]. In fact, jointly minimizing Softmax Cross-Entropy and Contrastive losses has been shown to lead to sub-optimal performance [26].

In this work, we propose an OOD-aware contrastive training objective. We start from SupCon as a basis to learn the embedding. Instead of relying on CE to learn the classifier weights, we learn prototypes, vectors lying in the same embedding as the feature extractor. These prototypes are learnt by forming positive pairs of samples belonging to the same class and then maximizing their similarity. We show that this prototype-based classifier provides less overconfident predictions on OOD data. Next, we enrich SupCon with two loss terms that exploit any available auxiliary or synthesized OOD data. The first loss term is applied at the projection head, similar to SupCon, but targets minimizing the pairwise similarities of ID and OOD features. Note that SupCon takes care of grouping ID features according to their classes while our first loss term pulls OOD features away from ID features. The second loss term is applied at the feature extractor level and minimizes the likelihood of the OOD data, as per the prototype classifier, by pulling OOD features away from all the learned classes prototypes. When auxiliary OOD data is unavailable, we propose ID features augmentation techniques to synthesize OOD-like features leveraged to regularize the training. Figure 1 illustrates our proposed OOD-aware contrastive training.

We evaluate our model in supervised and unsupervised settings where the OOD data is either available for training or it is synthesized using the available ID data. Our model improves the OOD detection performance achieving state-of-the-art results, while maintaining or improving the classification accuracy on the ID data. In summary, the contributions of this work are:

- We propose an OOD-aware training scheme that, in combination with the representation learning loss, learns ID class prototypes and forces OOD features to lie further from the ID features and prototypes.
- When auxiliary OOD data is not available, we propose a simple and very efficient feature augmentation technique to generate OOD-like features.
- Our experiments show the effectiveness of our training method compared to CE-based models. We compare against existing OOD works and show state-of-the-art results. We show an especially significant reduction in the false positive rate (FPR), an important metric in safety-critical applications.

The rest of this paper is organized as follows. We provide a background on existing methods in section 2. In section 3 we detail our methodology and evaluate our proposed ideas in section 4. We present conclusions in section 5.

## 2. Related Work

OOD detection methods can be divided based on whether they operate on a fixed pretrained model, adapt the model parameters for the OOD detection task, or leverage auxiliary OOD data to fine-tune the model.

**Post-hoc methods** operate on the output of a pretrained model with different scoring functions for OOD detection. Maximum Softmax Probability [22] is among the most commonly used scoring functions. However, the softmax function is known to contribute to the highly confident predictions in DNNs and usually exhibits weak OOD detection performance. Liang [31] proposed to enhance the separation of the softmax scores between ID and OOD inputs by temperature scaling and applying small perturbations to the input. Similarly, with the goal of providing a more robust scoring function, a variety of different techniques were proposed, *e.g.*, Mahalanobis distance to class centroids [30], predictions energy [32] or maximum-logit [21]. These measures produce a wider range of confidence values compared to softmax and are easier to threshold for OOD detection.

More recently, building on the observation that ID and OOD inputs display highly distinctive signature patterns of network's internal activations, [42] showed that clipping the activations of the penultimate layer of a pretrained model makes the output distributions for ID and OOD data better separated. Targeting a similar phenomena, [49] proposed to normalize the logits to unit vectors before applying the CE loss during training instead, and showed strong OOD detection performance. In this work, we use maximum-logit [21] as our scoring function and overcome the logits-norm issue outlined in [49] by operating only on normalized features and normalized prototypes during training.

**Training-based methods** are broadly split into generative or self-supervised methods. Building on the assumption of an underlying distribution shift between ID and OOD data, there is a large body of literature on leveraging generative models [18, 27, 45] to capture the distribution of ID examples and discard OOD data [5, 11, 23, 35, 38, 41]. Coupled with a generative model, reconstruction-based methods train an autoencoder on the existing ID data and classify a sample as OOD if the reconstruction error is high [37, 39, 59]. Generative models remain however difficult to train and optimize in large scale and [34] challenged some of the assumptions on their feasibility for OOD detection.

Lately, self-supervised methods improve the OOD detection performance by training the model to predict geometric transformations applied on the ID data [3, 17, 24]. More recently, [10, 40, 43, 50] revealed that different variants of contrastive training [9] on the ID data improves the OOD detection performance.

In this work, we propose an OOD-aware Supervised Contrastive (OPSupCon) training approach combining the supervised contrastive loss with additional tightness/contrastive losses to increase the OOD robustness.

**OOD-leveraging methods**. As an alternative to post-hoc approaches or training-based methods relying solely on ID data, more powerful OOD detection can be obtained by explicitly leveraging auxiliary OOD data. Such works use supervision from OOD samples collected from another mutually exclusive large dataset [23]. In [23] CE loss is applied on auxiliary OOD data with a uniform target distribution. [32] fine-tunes the model to explicitly create an energy gap by assigning low energy values to ID and high energy values to OOD training data.

When auxiliary OOD data is not available, works have instead synthesized outlier examples using ID data *e.g.*, by applying strong augmentations [43] or by sampling outliers assuming that features follow normal distribution in the penultimate layer [15]. Adversarial Reciprocal Point Learning (ARPL) [7] constructs reciprocal points modeling the empty space between clusters of different classes samples. In addition to generating confusing and diverse samples, a training scheme with adversarial margin constraint on the reciprocal points is proposed. However, this method is complex and requires intense hyperparameter tuning. In this work we propose a generic training scheme that includes OOD exemplars in the contrastive training scheme. Besides, we consider the case where no representative OOD data is available and alternatively propose feature manipulation techniques for generating pseudo OOD data.

## 3. Methodology

We consider a neural network that encodes each sample $\mathbf{x}$ with an encoder $E$, acting as a feature extractor, $E(\mathbf{x}) = \mathbf{f}$. The projection head (*e.g.*, a Multi Layer Perceptron) $H(\mathbf{f}) = \mathbf{z}$, maps the encoder feature $\mathbf{f}$ into the corresponding projection head feature $\mathbf{z}$. Eventually, the SupCon [26] loss is used to train both networks. Typically, an additional linear classifier is trained on top of the encoder features $\mathbf{f}$ using CE. However, in order to avoid its argued short-comings, we replace the CE-based training with *learning* randomly initialized class prototypes $\boldsymbol{\theta}$ using a *tightness* term that penalizes features $\mathbf{f}$ falling far from others of the sample class. In addition, *contrastive* terms push OOD samples far from ID samples and their prototypes.

We consider the setting where pairs of datum are available, one being an ID sample together with its label $D^I = (X_i^I, y_i)$, the other an auxiliary OOD sample $D^O = X_i^O$. In section 3.2 we extend this concept to the case where no auxiliary OOD data is available.

### 3.1. Loss Terms

Here we provide details of each loss term. An illustration of our OOD-aware Contrastive Learning method can be found in Fig. 1.

**Losses on ID data: SupCon loss on the head features**. The SupCon learning encourages samples of the same class to be pushed close together and pulled away from the samples of other classes. For a given ID sample embedding $\mathbf{z}_i^I$, we consider the embeddings of all other samples in the batch $\mathbf{z}_p^I$, belonging to the same class, as a set of positives $P_i$. The SupCon loss, given the embedding and its set of positives is:

$$\begin{aligned}&\mathcal{L}^{\text{SupCon}}_{H,i}(\mathbf{z}_i^I, P_i) = \\ &-\frac{1}{|P_i|}\sum_{\mathbf{z}_p^I\in P_i}\log\frac{\exp(\text{sim}(\mathbf{z}_i^I,\mathbf{z}_p^I)/\tau)}{\sum_{j\neq i}\exp(\text{sim}(\mathbf{z}_i^I,\mathbf{z}_j^I)/\tau)} = \\ &\frac{1}{|P_i|}\sum_{\mathbf{z}_p^I\in P_i}\left(\underbrace{-(\mathbf{z}_i^{I\top}\mathbf{z}_p^I)/\tau}_{\text{Tightness}} + \underbrace{\log\sum_{j\neq i}\exp\left((\mathbf{z}_i^{I\top}\mathbf{z}_j^I)/\tau\right)}_{\text{Contrast}}\right),\end{aligned} \tag{1}$$

where the index $j$ iterates over all (original and augmented) samples. The SupCon loss is expressed as the average of the loss defined on each positive pair, where (in this supervised setting) the positive pairs are formed of augmented views and other samples of the same class. Note, that the SupCon loss can be expressed as a combination of a tightness and a contrast term where positive pairs similarities are maximized via the tightness term and negative pairs similarities are minimized with the contrast term.

The total SupCon loss is the mean of the losses for the $N^I$ ID samples considered.

$$\mathcal{L}^{\text{SupCon}}_{H} = \frac{1}{N^I}\sum_{i=1}^{N^I}\mathcal{L}^{\text{SupCon}}_{H,i}. \tag{2}$$

**A tightness loss on the encoder features** serves to learn class prototypes $\boldsymbol{\theta}_c$ (in the encoder feature space) by maximizing their similarities with the corresponding class features:

$$\mathcal{L}^P_E = \frac{1}{N^I}\sum_{i=1}^{N^I}\mathcal{L}^{\text{tt}}(\mathbf{f}_i^I, \boldsymbol{\theta}_{y_i}) = \underbrace{\frac{1}{N^I}\sum_{i=1}^{N^I} -\mathbf{f}_i^{I\top}\boldsymbol{\theta}_{y_i}}_{\text{Tightness}}. \tag{3}$$

We assume that all sample features $\mathbf{f}_i$ and class prototypes $\boldsymbol{\theta}_k$ are normalized to have unit length ($\|\mathbf{f}_i^I\| = \|\boldsymbol{\theta}_k\| = 1$) and that the classifier is linear with no bias term.

Note that the number of samples in (3) might differ from $N^I$ (*e.g.*, due to augmentation), in which case $N^I$ should be replaced by the corresponding number of samples. With that assumption, we use a nearest prototype classifier *i.e.*, assigning a test sample to the class of the nearest prototype in the feature space. With this formulation, features of the same class are forced to become closer and each class prototype is learned as the closest to its class features. Besides, features of different classes are forced to become further apart. A similar loss term was introduced in [1] to train only the linear classifier (the prototypes) for supervised classification. However, as discussed below, we use the tightness term in our work to also train the encoder which enhances the robustness when OOD data are present during training.

**Losses on OOD data.** Auxiliary OOD data are additional samples that do not belong to the concerned task's distribution. No other information such as specific class labels or samples similarities are either provided or can be assumed. Here we try to answer the question on how to increase the robustness of Supervised Contrastive training against OOD data without assuming any specific additional information about these auxiliary data. We propose two additional loss terms to be combined with the aforementioned losses on ID data, one at the level of the projection head and the other at the encoder level.

**Contrastive term on the projection head features**. When the SupCon loss is applied on ID samples, it is composed of the tightness term operating on positive pairs and the contrast term operating on other pairs. For OOD samples, we do not want to impose any superficial similarity to any other sample, the target is simply to learn how to project those samples as far from ID samples as possible. We thus propose to only deploy a contrast or a pull term to pairs of OOD/ID samples:

$$\mathcal{L}_H^O = \frac{1}{N^O} \sum_{i=1}^{N^O} \underbrace{\log \sum_{j=1}^{N^I} \exp\left((\mathbf{z}_i^{O\top} \mathbf{z}_j^I)/\tau\right)}_{\text{Contrast}}. \tag{4}$$

**Contrastive term on the encoder features**. The OOD samples can be considered as coming from a new (w.r.t. the $K$ known ID classes) category. Similarly to the projection head case (4), the contrastive term for OOD data can be defined on the encoder feature level. In this case, instead of using *all* ID features $\mathbf{f}_i^I$, only their class prototypes $\boldsymbol{\theta}_k$ are deployed:

$$\mathcal{L}_E^O = \frac{1}{N^O} \sum_{i=1}^{N^O} \underbrace{\frac{1}{K} \log \sum_{k=1}^{K} \exp\left((\mathbf{f}_i^{O\top} \boldsymbol{\theta}_k)/\tau\right)}_{\text{Contrast}}. \tag{5}$$

We propose to minimize the Log Sum Exponential (LSE) of each OOD feature similarity to all prototypes, which would lead to the desired minimization of the *maximal* similarity of the OOD feature to the closest class prototype.

**The objective function**. Our training objective on both ID and OOD samples is composed of losses operating on both the projection head and the encoder level. At the projection head, we operate on pairwise similarities. At the encoder level, however, the prototype-based proxy-similarities are optimized. ID samples and their class prototypes are encouraged to be close together, whereas the similarities between ID prototypes and OOD samples are minimized. We hypothesize that such treatment of the OOD samples would generalize better to other OOD data, as opposed to imposing a specific clustering of OOD data.

$$\mathcal{L} = \mathcal{L}_H^{\text{SupCon}} + \gamma \mathcal{L}_H^O + \alpha \left(\mathcal{L}_E^P + \mathcal{L}_E^O\right), \tag{6}$$

where the parameters $\alpha$ and $\gamma$ control the contribution of the additional. The minimization of our final loss optimizes jointly the representations and the class prototypes while attempting at increasing the OOD robustness by contrasting auxiliary OOD features from both ID samples and their class prototypes.

### 3.2. Pseudo-OOD features generation

Our method leverages auxiliary OOD data to improve the OOD robustness of the learned model. Here, we propose a simple alternative that generates pseudo OOD features when auxiliary OOD data cannot be provided. We suggest to transform ID features to produce pseudo-OOD features that mimic realistic and challenging OOD cases. In many real-life applications, such as autonomous driving, there is a high chance of encountering OOD inputs that lie in between class categories in the embedding space. For instance, a model which has not seen any examples belonging to the class "Motorcyclist", may assign an internal representation to such examples close to both the pedestrian and vehicle features. Based on the observation that OOD data are commonly projected in between ID clusters and in areas where different ID classes overlap [2], our idea is to generate features spanning this space between ID samples of different classes.

Inspired by the Manifold Mixup [47] technique, where Mixup [56] is applied at feature level (any hidden state) instead of at the input images, we suggest to perform a Mixup of the ID features extracted from the encoder being trained. Differently from existing Mixup techniques, we consider the generated features as OOD and apply our proposed loss using this OOD features at the projection head level. Given an ID feature we generate a pseudo-OOD feature:

$$\mathbf{f}_i^O = \lambda \mathbf{f}_i^I + (1-\lambda)\mathbf{f}_j^I,\ j = \underset{j, y_j \neq y_i}{\arg\max} (\mathbf{f}_i^{I\top} \mathbf{f}_j^I), \tag{7}$$

| Dataset/Method<br>Metrics | CE | | | PSupCon | | | CE + Energy | | | PSupCon + Energy | | | OPSupCon-R | | | OPSupCon-P | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FPR↓ | AUROC↑ | AUPR↑ | FPR↓ | AUROC↑ | AUPR ↑ | FPR↓ | AUROC↑ | AUPR↑ | FPR↓ | AUROC↑ | AUPR ↑ | FPR↓ | AUROC↑ | AUPR ↑ | FPR↓ | AUROC↑ | AUPR ↑ |
| DTD | 25.01 | 95.02 | 98.81 | 14.09 | 97.44 | 99.44 | 7.83 | 98.43 | 99.67 | **2.71** | **99.43** | **99.87** | 4.95 | 99.04 | 99.80 | 16.57 | 96.69 | 99.22 |
| SVHN | 3.08 | 99.19 | 99.84 | 3.16 | 99.39 | 99.87 | 1.55 | 99.47 | 99.90 | 1.92 | 99.57 | 99.91 | **0.85** | **99.75** | **99.95** | 5.41 | 98.46 | 99.70 |
| Places365 | 28.56 | 94.07 | 98.52 | 26.96 | 94.88 | 98.79 | 20.61 | 95.70 | 98.94 | 36.85 | 92.11 | 98.07 | 21.17 | 95.63 | 98.91 | **14.48** | **96.76** | **99.21** |
| LSUN-C | 12.10 | 97.68 | 99.54 | 3.74 | 99.22 | 99.84 | 5.28 | 98.75 | 99.75 | 64.98 | 81.89 | 95.42 | **1.33** | **99.60** | **99.92** | 2.39 | 99.34 | 99.87 |
| LSUN-R | 8.98 | 98.15 | 99.63 | 6.43 | 98.65 | 99.73 | 8.69 | 98.37 | 99.67 | **4.40** | **99.11** | **99.81** | 9.52 | 98.16 | 99.64 | 6.62 | 98.57 | 99.72 |
| iSUN | 11.54 | 97.86 | 99.58 | 6.29 | 98.71 | 99.74 | 7.24 | 98.59 | 99.72 | **2.48** | **99.44** | **99.88** | 7.71 | 98.40 | 99.69 | 7.24 | 98.52 | 99.70 |
| iNaturalist | 37.24 | 94.10 | 98.77 | 10.70 | 98.18 | 99.63 | 18.49 | 96.40 | 99.21 | **7.53** | **98.61** | **99.70** | 9.87 | 98.11 | 99.63 | 12.48 | 97.70 | 99.53 |
| CIFAR-100 | 40.73 | 91.85 | 98.03 | 41.03 | 92.30 | 98.19 | 37.04 | 93.00 | 98.34 | 51.07 | 89.59 | 97.57 | 36.42 | **93.25** | **98.51** | **36.04** | 93.15 | 98.41 |
| Mnist | 30.88 | 95.76 | 99.17 | **1.62** | **99.50** | **99.90** | 32.55 | 94.93 | 98.97 | 45.78 | 92.85 | 98.53 | 2.79 | 99.42 | 99.89 | 8.10 | 98.55 | 99.72 |
| TIN | 32.05 | 93.22 | 98.30 | 30.95 | 93.86 | 98.50 | 27.00 | 94.36 | 98.58 | 31.80 | 93.45 | 98.47 | 25.83 | 94.39 | 98.61 | **25.55** | **94.61** | **98.64** |
| Average | 23.02 | 95.69 | 99.02 | 14.49 | 97.20 | 99.36 | 16.63 | 96.80 | 99.27 | 24.95 | 94.60 | 98.72 | **12.01** | **97.56** | **99.44** | 13.52 | 97.24 | 99.38 |

Table 1. **OOD detection performance on Cifar-10:** a) comparison of CE and PSupCon (1, 2 columns) and, b) comparison of OOD training with our method compared to energy finetuning. Our method outperforms performance energy finetuning even with pseudo OOD.

where the new pseudo-OOD sample $\mathbf{f}_i^O$ is a linear combination of the concerned ID feature $\mathbf{f}_i^I$ and the most similar ID feature $\mathbf{f}_i^I$ of a different class. The selection of the closest feature of a different class is to ensure that the generated OOD feature indeed lies between two close ID samples, of different classes, and to avoid generating redundant and easy OOD features. The $\lambda$ is drawn at each iteration from a normal distribution centered at $0.5$ with $0.3$ standard deviation. Here ID features come from raw and augmented ID samples. The proposed pseudo-OOD features generation technique is extremely efficient and adds minimal computational cost. Further, as the pseudo features are generated at the encoder level, we can remove the term $\mathcal{L}_E^O$ from our full loss function (6) and rely on $\mathcal{L}_H^O$ (5) applied at the projection head to train the model.

# 4. Experiments

In this section, we evaluate the effectiveness of our method and its components and compare it to state-of-the-art OOD detection methods on various datasets.

## 4.1. Experimental Settings

We employ a ResNet18 [20] backbone, following [53], and use CIFAR-10 and CIFAR-100 [28] as the in-distribution datasets. We train our models with a batch size of 512 using SGD as the optimizer and a cosine annealing scheduler [33]. We use the same data augmentation as in SupCon [26], namely AutoAugment. Training is performed for 500 epochs. In the supplementary material we provide further results, using ResNet50 as the feature extractor.

We extensively evaluate and report our results on Describable Textures Dataset (DTD) [12], SVHN [36], Places365 [58], LSUN-Crop, LSUN-Resize [55], iSUN [51], iNaturalist [46], Mnist [14] and Tiny Imagenet (TIN) [29] and CIFAR datasets. We use the following metrics to evaluate our experiments.

**FPR@95** (↓), measures the false positive rate when true positive rate is set to 95%, and referred to as FPR.

**AUROC** (↑), the area under the Receiver Operating Characteristic (ROC) curve; denoting TPR/FPR relationship.

**AUPR** (↑), the area under the Precision-Recall (PR) curve. We consider ID samples as positives.

## 4.2. OOD Training Dataset

The selection of the dataset to be used as the OOD data highly depends on whether it can comprehensively represent all other possible OOD data or not, which in turn would depend on the distribution of each ID dataset. Many works such as [23,32] opt for a large diverse dataset, *i.e.*, TinyImages [44] which acts as an extensive set of all other possible objects. Such extensive OOD dataset would likely present overlaps with many ID datasets, and would require careful curation each time. Additionally, this dataset is not publicly available anymore due to ethical related issues.

We use instead a more scalable approach: the much smaller DTD dataset [12], a collection of various textures that can be synthetically generated using state-of-the-art generation techniques. This way, we show that our proposed contrastive training scheme can generalize to other OOD datasets without accessing a huge collection of various types of OOD objects. We further ablate the effect of the choice of auxiliary OOD dataset in the supplementary material.

Moreover, as we have seen, an alternative to using a real OOD dataset is to synthesize the auxiliary examples from the accessible ID dataset [7, 16, 25, 43, 49]. Similarly, we provide a simple alternative strategy to mimic OOD like features as described in Section 3.2. We show that, using our method, the generalization performance is close to that of using a real OOD dataset.

## 4.3. Compared methods

We compare the proposed method with representative state-of the-art methods.

**Post hoc methods.** MSP [22] uses the maximum Softmax probability as a scoring function, a standard baseline in OOD detection literature. ODIN [31] performs perturbation in the input image and uses the MSP score on the perturbed image. ReAct [42] performs a rectification on the logits for computing the OOD detection score (energy score).

**Training-based methods.** CSI [43], close to our work, leverages supervised contrastive learning. However, as a proxy for OOD data, the method leverages strongly augmented samples. SSD [40] combines contrastive learning

| Dataset/Method | CE | | | PSupCon | | | CE + Energy | | | PSupCon + Energy | | | OPSupCon-R | | | OPSupCon-P | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | FPR↓ | AUROC↑ | AUPR↑ | FPR↓ | AUROC↑ | AUPR ↑ | FPR↓ | AUROC↑ | AUPR↑ | FPR↓ | AUROC↑ | AUPR ↑ | FPR↓ | AUROC↑ | AUPR ↑ | FPR↓ | AUROC↑ | AUPR ↑ |
| DTD | 73.53 | 82.29 | 95.81 | 66.14 | 80.14 | 94.53 | 57.44 | 88.57 | 97.41 | **27.65** | **93.75** | **98.44** | 51.22 | 88.44 | 97.28 | 54.23 | 84.77 | 95.89 |
| SVHN | 34.38 | 93.89 | 98.71 | 47.74 | 91.22 | 98.12 | **17.21** | **96.93** | **99.34** | 29.13 | 94.94 | 98.93 | 44.26 | 92.39 | 98.39 | 49.49 | 90.89 | 98.04 |
| Places365 | 80.83 | 77.96 | 94.52 | 76.89 | 78.24 | 94.59 | 78.94 | 78.81 | 94.79 | 75.86 | 78.82 | 94.56 | 74.52 | 79.30 | 94.79 | **74.45** | **79.71** | **94.95** |
| LSUN-C | 54.84 | 90.19 | 97.93 | 27.64 | 95.03 | 98.92 | 53.70 | 91.81 | 98.33 | 52.26 | 91.62 | 98.25 | 20.38 | 96.48 | 99.27 | **18.10** | **96.71** | **99.30** |
| LSUN-R | 62.42 | 88.00 | 97.42 | 47.64 | 90.54 | 97.92 | 46.04 | 91.41 | 98.10 | **13.46** | **96.98** | **99.28** | 38.54 | 93.01 | 98.49 | 37.85 | 92.78 | 98.43 |
| iSUN | 64.57 | 87.38 | 97.29 | 55.10 | 88.29 | 97.38 | 50.00 | 90.33 | 97.85 | **14.38** | **96.92** | **99.29** | 46.45 | 91.33 | 98.13 | 46.38 | 90.82 | 97.97 |
| iNaturalist | 85.23 | 79.19 | 95.33 | **43.67** | 89.11 | 97.24 | 79.15 | 83.25 | 96.32 | 70.28 | 85.19 | 96.71 | 47.71 | 89.87 | 97.63 | 45.38 | **89.97** | **97.64** |
| CIFAR-10 | **76.84** | **79.16** | **94.88** | 83.45 | 73.12 | 92.66 | 80.03 | 78.40 | 94.70 | 89.84 | 71.60 | 92.75 | 84.74 | 71.01 | 91.50 | 84.08 | 73.11 | 92.73 |
| Mnist | 88.81 | 75.01 | 94.34 | 33.93 | 94.24 | 98.79 | 95.33 | 67.34 | 92.36 | 96.27 | 77.41 | 95.21 | 33.89 | **94.38** | **98.83** | **33.78** | 94.37 | **98.83** |
| TIN | 75.20 | 80.56 | 95.08 | 72.15 | 81.09 | 95.22 | 71.70 | 82.91 | 95.68 | **62.50** | **85.02** | **96.17** | 68.00 | 82.67 | 95.52 | 69.23 | 82.12 | 95.44 |
| Average | 69.67 | 83.36 | 96.13 | 55.43 | 86.10 | 96.53 | 62.95 | 84.98 | 96.49 | 53.16 | 87.22 | 96.96 | **50.97** | **87.89** | **96.98** | 51.29 | 87.53 | 96.92 |

Table 2. **OOD detection performance on Cifar-100:** a) comparison of CE and PSupCon (1, 2 columns) and, b) comparison of OOD training with our method compared to energy finetuning. Our method outperforms performance energy finetuning even with pseudo OOD.

on ID data with k-means clustering for OOD detection. ARPL [8] proposes the concept of reciprocal points as representatives for OOD data, and train the neural network such that the features of ID classes lie within a margin distance from those points. VOS [16] synthesizes outliers in the penultimate layer, by assuming that ID features follow a normal distribution within each class. LNorm [49] trains the network such that the logits norm is constrained to be a constant.

**OOD-leveraging methods.** OE [23] makes use of CE for ID data. For OOD, they set an uniform distribution as the target. UDG [52] leverages unsupervised data for both OOD training and enhancing ID performance. Two heads are proposed. The first one minimizes CE loss on ID labeled data and maximizes the entropy on OOD data. The second performs deep clustering.

Energy [32] maximizes an energy gap between the ID and OOD samples. First, the mean energy values on the ID ($m_{in}$) and OOD ($m_{out}$) datasets are calculated. Next, the model is fine-tuned to produce energy values lower than $m_{in}$ for the ID samples and higher than $m_{out}$ for OOD samples. It achieves the best OOD detection performance compared to previous OOD training methods. However, an extra step to calculate the thresholds $m_{in}$ and $m_{out}$ is required.

**Our method.** We refer to our method as OPSupCon: OOD-aware Prototypical Supervised Contrastive learning. Throughout the experiments we consider different variants of our method. **PSupCon** refers to the combination of supervised contrastive training loss (1) with the prototypes learning loss (3), OOD regularization is not applied here. **OPSupCon-R** refers to models trained with our complete loss (6) based on **r**eal auxiliary OOD data. **OPSupCon-P** refers to models trained with $\mathcal{L}_H^O$ in loss (6). We generate **p**seudo OOD-like features from ID examples using (7), as described in Section 3.2.

We use Maximum Logit [21] as our scoring function. Logit here refers to the dot product between a sample representation and a given prototype. We ablate the choice of different scoring functions in the supplementary materials.

## 4.4. PSupCon OOD detection performance

We first compare the OOD detection performance of the prototype classifier, trained with SupCon [26] and the tightness term (PSupCon), to a classifier trained with a Cross-Entropy (CE) loss. The purpose is to observe the inherent OOD detection capability of each model without explicit OOD fine-tuning.

Detailed results can be found in tables 1 and 2 (columns 1 and 2). We see that PSupCon consistently outperforms CE on most datasets and metrics with a a significant reduction on the FPR. This is a especially relevant metric for the purpose of rejecting OOD samples, as it is calculated where the rejection rate is fixed to $5\%$. Achieving a lower FPR is crucial for real life applications, where it is not possible to know the threshold in advance.

Therefore, the prototype classifier based on SupCon is shown to be more robust for OOD detection.

## 4.5. OOD-Aware Supervised Contrastive Learning

The formulation of our proposed loss function (6) permits its different components to be applied at different stages, as required by different use cases. We first train our model with SupCon [26] and fine-tune it with our full objective function (6) for additional $50$ epochs. This allows learning a good initial representation of the task at hand and improving these representation for a stronger separation of ID and OOD data. It also permits fine-tuning of any pre-trained model when OOD data becomes available.

In our loss function (6), we use $\gamma = 1$ for training with real OOD and $\gamma = 0.5$ for synthesized OOD. The weight of the encoder losses $(\mathcal{L}_E^P + \mathcal{L}_E^O)$ is set to $\alpha = 0.1$.

### 4.5.1 Comparison with SOTA OOD Training method

First we extensively compare our proposed OOD training scheme with energy fine-tuning [32]. This method shows state-of-the-art performance when fine-tuning a pretrained model with real auxiliary OOD data. Energy fine-tuning was originally introduced for models trained with Cross-Entropy loss. Here we compare our proposed method with energy fine-tuning on top of both CE and PSupCon models.

Tables 1 and 2 summarise our results for this purpose for

Cifar-10 and Cifar-100 datasets respectively. We observe that OPSupCon-R and OPSupCon-P (columns 5 and 6) outperform the models fine-tuned with Energy [32] (columns 3 and 4) on most datasets, achieving a better average for all metrics. Moreover, energy fine-tuning on top of a PSupCon model improves the results on some datasets while significantly worsening the results on some others, proving *unreliable* for OOD detection.

Note that Energy fine-tuning [32] requires an extra step to determine thresholds $m_{in}$ and $m_{out}$ for each model, before fine-tuning. Our method achieves a better performance without any extra model-dependant hyper-parameters.

### 4.5.2 Comparison with other methods

Tables 3 and 4 compare our method to different lines of literature described in Sec 4.3 on CIFAR-10 and CIFAR-100 datasets[1]. We follow the evaluation protocol of [53] by training the model for 100 epochs only and using a ResNet-18 architecture. OOD fine-tuning is set to 10 epochs.

The values reported for previous work might slightly differ from those in the original papers due to the architecture change (ResNet-18) and decreased number of training epochs (100 epochs). Note that this is suboptimal for our approach as well since supervised contrastive training usually requires more epochs to converge. However, [53] enables the community to have a fair, complete and model independent comparison of different lines of work.

Our OPSupCon-R outperforms state-of-the-art methods from different families reducing the average FPR rate by 22.81 on CIFAR-10 and 10.48 on CIFAR-100. These results suggests the effectiveness of our training pipeline when leveraging real auxiliary OOD data.

We further show with OPSupCon-P that our training scheme can benefit from synthesized OOD-like features and supersede the rival state-of-the-art methods, even those which leverage real-OOD data, thanks to the powerful representation training mechanism. On CIFAR-100, SSD [40] achieves an overall better performance for the FPR and AUROC metrics compared to OPSupCon-P. While OPSupCon-P does better on majority of the datasets, the difference on SVHN biases this comparison. More extensive experimentation on a larger number of datasets shows that our method achieves a better overal result to this work by a margin (supplementary material).

Finally, in the supplementary materials we show that further combining pseudo and real OOD data can provide an extra boost to the detection performance.

| Method | Metric | DTD | SVHN | Places365 | CIFAR-100 | MNIST | TIN | Average |
|---|---|---|---|---|---|---|---|---|
| CE + Energy | FPR↓ | 39.45 | 20.41 | 34.12 | 60.42 | 51.02 | 49.75 | 42.52 |
| | AUROC↑ | 93.61 | 96.78 | 92.97 | 88.41 | 93.36 | 90.47 | 92.60 |
| | AUPR↑ | 98.68 | 99.37 | 98.37 | 97.40 | 98.73 | 97.79 | 98.05 |
| OPSupCon R | FPR↓ | **8.27** | **3.27** | **21.98** | **43.70** | 6.46 | **33.12** | **19.46** |
| | AUROC↑ | **98.48** | **99.26** | **95.37** | **91.20** | 98.58 | 93.40 | **96.04** |
| | AUPR↑ | **99.68** | **99.85** | **98.83** | **97.87** | 99.72 | **98.36** | **99.05** |
| OPSupCon P | FPR↓ | 18.65 | 4.88 | 25.02 | 46.43 | **4.48** | 34.23 | 22.28 |
| | AUROC↑ | 96.11 | 99.00 | 95.00 | 90.48 | **98.97** | 93.16 | 95.45 |
| | AUPR↑ | 99.07 | 99.80 | 98.79 | 97.78 | **99.80** | 98.30 | 98.92 |
| MSP [22] | FPR↓ | 59.89 | 51.87 | 57.64 | 62.01 | 58.59 | 60.69 | 58.44 |
| | AUROC↑ | 88.72 | 90.88 | 89.03 | 87.11 | 89.91 | 86.62 | 88.71 |
| | AUPR↑ | 91.28 | 78.19 | 70.24 | 85.92 | 66.95 | 83.07 | 79.27 |
| ODIN [31] | FPR↓ | 51.10 | 67.92 | 50.51 | 59.09 | 36.23 | 59.06 | 53.98 |
| | AUROC↑ | 80.70 | 73.32 | 82.55 | 77.68 | 90.91 | 77.33 | 80.41 |
| | AUPR↑ | 82.25 | 42.13 | 50.27 | 73.24 | 64.74 | 70.07 | 63.78 |
| ReAct [42] | FPR↓ | 49.98 | 49.23 | 44.21 | 53.72 | 50.94 | 47.00 | 47.68 |
| | AUROC↑ | 88.18 | 89.50 | 90.09 | 86.35 | 88.34 | 88.90 | 88.56 |
| | AUPR↑ | 89.91 | 75.36 | 69.28 | 83.15 | 50.88 | 86.53 | 75.85 |
| CSI [43] | FPR↓ | 53.63 | 33.26 | 58.01 | 61.92 | 32.07 | 55.27 | 49.02 |
| | AUROC↑ | 91.04 | 95.22 | 88.57 | 88.08 | 95.09 | 90.18 | 91.36 |
| | AUPR↑ | 95.00 | 92.42 | 77.57 | 89.87 | 86.30 | 92.12 | 88.88 |
| ARPL [7] | FPR↓ | 69.86 | 73.41 | 66.20 | 69.81 | 68.99 | 68.46 | 69.45 |
| | AUROC↑ | 87.36 | 87.77 | 88.40 | 86.68 | 88.48 | 87.70 | 87.73 |
| | AUPR↑ | 92.85 | 82.92 | 77.63 | 88.65 | 74.27 | 89.87 | 84.36 |
| VOS [16] | FPR↓ | 37.38 | 29.92 | 45.37 | 52.94 | 42.22 | 45.85 | 42.28 |
| | AUROC↑ | 91.26 | 93.82 | 88.73 | 86.08 | 89.83 | 88.89 | 89.76 |
| | AUPR↑ | 92.72 | 83.73 | 63.93 | 83.52 | 52.37 | 87.15 | 77.23 |
| LNorm [49] | FPR↓ | 30.94 | 5.30 | 31.17 | 46.99 | 4.75 | 36.34 | 25.91 |
| | AUROC↑ | 94.30 | 98.86 | 94.76 | 91.13 | 98.82 | **93.90** | 95.29 |
| | AUPR↑ | 96.32 | 97.70 | 88.11 | 91.89 | 96.24 | 94.84 | 94.18 |
| OE [23] | FPR↓ | 79.49 | 84.59 | 84.69 | 82.14 | 94.32 | 78.44 | 83.90 |
| | AUROC↑ | 78.90 | 82.40 | 72.06 | 75.35 | 67.31 | 77.37 | 75.56 |
| | AUPR↑ | 85.78 | 73.96 | 41.59 | 75.35 | 35.09 | 77.80 | 65.22 |
| UDG [52] | FPR↓ | 43.97 | 61.91 | 42.44 | 55.33 | 39.32 | 42.48 | 47.57 |
| | AUROC↑ | 93.56 | 92.50 | 93.58 | 90.38 | 93.81 | 93.33 | 92.86 |
| | AUPR↑ | 96.55 | 90.85 | 87.89 | 91.67 | 82.67 | 94.66 | 90.71 |
| SSD SupCon [40] | FPR↓ | 24.29 | **1.67** | 29.52 | 49.18 | 8.38 | 44.06 | 26.18 |
| | AUROC↑ | 95.97 | **99.65** | 94.09 | 89.34 | 98.13 | 90.52 | 94.61 |
| | AUPR↑ | 93.12 | **99.86** | **99.76** | 88.43 | 97.23 | 89.77 | 94.69 |

Table 3. **Comparison with the state-of-the-art on CIFAR-10 dataset.** OPSupCon improves significantly over state of the art methods.

### 4.5.3 Ablation

In this section we evaluate the effectiveness of the different components of our loss. We ablate this by fine-tuning the SupCon trained model with:

1) Prototype losses on the encoder level only ($\mathcal{L}_H^{\text{SupCon}} + \alpha\mathcal{L}_E^P$).
2) Auxiliary loss on the head level combined with the SupCon loss and the prototype loss ($\mathcal{L}_H^{\text{SupCon}} + \mathcal{L}_H^O + \alpha\mathcal{L}_E^P$).
3) Full loss function, with a contrast term applied at the encoder level as well ($\mathcal{L}_H^{\text{SupCon}} + \gamma\mathcal{L}_H^O + \alpha\left(\mathcal{L}_E^P + \mathcal{L}_E^O\right)$).

[1] The results for previous work (except for Energy and SSD) are taken from [53] and can be found here

| Method | Metric | DTD | SVHN | Places365 | CIFAR-10 | MNIST | TIN | Average |
|---|---|---|---|---|---|---|---|---|
| CE + Energy | FPR↓ | 77.45 | **24.79** | 76.25 | 87.67 | 93.71 | 74.45 | 72.38 |
| | AUROC↑ | 77.66 | **95.28** | 75.77 | 71.05 | 65.95 | 79.71 | 77.57 |
| | AUPR↑ | 94.24 | **98.96** | 93.60 | 92.23 | 92.06 | 94.89 | 94.33 |
| OPSupCon<br>R | FPR↓ | **43.54** | 79.73 | 77.59 | 87.21 | **9.75** | **73.63** | **61.90** |
| | AUROC↑ | **90.81** | 83.84 | 78.70 | 70.69 | **98.52** | 80.98 | **83.92** |
| | AUPR↑ | **97.78** | 96.59 | 94.61 | 91.83 | **99.70** | **95.15** | **95.94** |
| OPSupCon<br>P | FPR↓ | 57.25 | 83.20 | 76.75 | 84.70 | 20.61 | 74.12 | 66.10 |
| | AUROC↑ | 84.19 | 82.16 | 78.94 | 73.49 | 96.69 | 80.60 | 82.67 |
| | AUPR↑ | 95.71 | 96.20 | 94.68 | **92.97** | 99.30 | 95.00 | 95.64 |
| MSP [22] | FPR↓ | 83.83 | 83.69 | 81.24 | **81.82** | 87.78 | 76.22 | 82.43 |
| | AUROC↑ | 76.93 | 76.04 | 79.44 | **78.31** | 77.78 | 81.78 | 78.38 |
| | AUPR↑ | 85.24 | 60.76 | 62.39 | 79.58 | 54.19 | 86.30 | 71.41 |
| ODIN [31] | FPR↓ | 83.83 | 83.69 | 81.27 | 83.16 | 75.34 | 77.77 | 80.84 |
| | AUROC↑ | 79.39 | 71.08 | 79.83 | 78.18 | 83.71 | 81.39 | 78.93 |
| | AUPR↑ | 86.67 | 52.36 | 60.85 | 79.12 | 62.02 | 85.30 | 71.05 |
| ReAct [42] | FPR↓ | 76.76 | 77.41 | 79.18 | 82.89 | 89.32 | 75.81 | 80.22 |
| | AUROC↑ | 81.73 | 83.73 | 79.63 | 76.98 | 77.02 | **81.96** | 80.17 |
| | AUPR↑ | 89.01 | 76.43 | 59.44 | 77.78 | 52.01 | 85.89 | 73.42 |
| CSI [43] | FPR↓ | 89.27 | 67.96 | 87.91 | 88.23 | 92.38 | 85.30 | 85.17 |
| | AUROC↑ | 59.72 | 78.57 | 69.94 | 69.24 | 57.06 | 72.32 | 67.80 |
| | AUPR↑ | 68.86 | 60.24 | 48.53 | 71.03 | 27.43 | 78.18 | 59.04 |
| ARPL [7] | FPR↓ | 88.76 | 80.90 | 85.25 | 85.80 | 84.91 | 83.34 | 84.82 |
| | AUROC↑ | 69.50 | 78.97 | 74.57 | 73.48 | 72.94 | 76.31 | 74.29 |
| | AUPR↑ | 79.33 | 68.58 | 55.80 | 75.19 | 43.31 | 82.39 | 67.43 |
| VOS [16] | FPR↓ | 94.54 | 98.62 | 97.81 | 96.64 | 92.31 | 96.40 | 96.05 |
| | AUROC↑ | 68.33 | 68.99 | 68.21 | 71.74 | 82.17 | 72.08 | 71.92 |
| | AUPR↑ | 76.20 | 56.36 | 43.20 | 72.17 | 55.66 | 77.10 | 63.44 |
| LNorm [49] | FPR↓ | 87.06 | 79.16 | 80.20 | 83.77 | 53.07 | 77.19 | 76.74 |
| | AUROC↑ | 71.53 | 83.03 | **79.84** | 74.84 | 90.82 | 81.87 | 80.32 |
| | AUPR↑ | 79.08 | 75.57 | 63.10 | 73.56 | 76.09 | 86.28 | 75.61 |
| OE [23] | FPR↓ | 88.46 | 75.31 | 92.23 | 90.92 | 80.84 | 90.85 | 86.43 |
| | AUROC↑ | 64.70 | 77.43 | 64.91 | 63.23 | 76.89 | 64.14 | 68.55 |
| | AUPR↑ | 74.66 | 62.15 | 45.23 | 64.65 | 49.18 | 72.25 | 61.35 |
| UDG [52] | FPR↓ | 83.46 | 93.47 | **74.13** | 87.22 | 93.28 | 78.21 | 84.96 |
| | AUROC↑ | 72.15 | 53.38 | 78.61 | 72.88 | 66.63 | 78.79 | 70.40 |
| | AUPR↑ | 81.32 | 28.68 | 57.34 | 75.00 | 32.54 | 83.32 | 59.7 |
| SSD<br>SupCon [40] | FPR↓ | 48.10 | 28.46 | 81.00 | 86.66 | 52.62 | 76.22 | 62.17 |
| | AUROC↑ | 90.59 | 94.45 | 76.46 | 66.45 | 89.26 | 79.18 | 82.73 |
| | AUPR↑ | 83.99 | 98.70 | **98.83** | 64.32 | 88.04 | 76.30 | 85.03 |

Table 4. **Comparison with the state-of-the-art on CIFAR-100 dataset.** Our OPSupCon improves significantly over state of the art methods.

| Method | FPR↓ | AUROC↑ | AUPR↑ |
|---|---|---|---|
| $\mathcal{L}_H^{\text{SupCon}} + \alpha\mathcal{L}_E^P$ | 14.46 | 97.14 | 99.35 |
| $\mathcal{L}_H^{\text{SupCon}} + \mathcal{L}_H^O + \alpha\mathcal{L}_E^P$ | 12.45 | 97.43 | 99.41 |
| $\mathcal{L}_H^{\text{SupCon}} + \gamma\mathcal{L}_H^O + \alpha\left(\mathcal{L}_E^P + \mathcal{L}_E^O\right)$ | **12.01** | **97.56** | **99.44** |

Table 5. **Ablation study:** Investigating the effect of the components of our loss function on CIFAR-10 dataset. Average mean results over the different OOD datasets are reported.

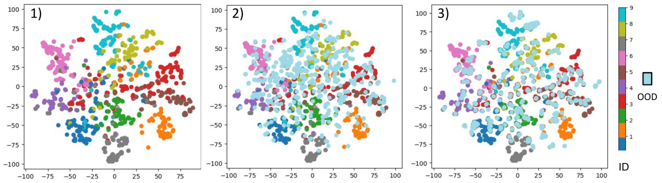

Figure 2. t-SNE 2D projection of the encoder features of 1) Cifar-10 ID samples, 2) real auxiliary OOD - DTD dataset - and, 3) pseudo OOD.

Table 5 reports the results of each variant in the case of training with real auxiliary OOD data with Cifar-10 as the ID dataset. While fine-tuning with the prototypes loss alone does not bring substantial improvements, minimizing our auxiliary loss on the head does. Adding the extra contrastive term at encoder level further enhances the quality of the prototypes, increasing the robustness of the OOD detection.

## 4.6. Pseudo Features Analysis

In the previous experiments, we show that our method can leverage synthetic OOD data and improve the OOD detection performance with a simple approach to generate pseudo OOD features, a mixup of ID features of different classes. Here we compare visually those synthesized features to the real OOD features of DTD dataset. Figure 2 shows the t-SNE 2D projection of ID features (Cifar-10) at the encoder level and the OOD features for both real OOD (DTD) and pseudo OOD cases. We refer to supplementary for more details. We can see that pseudo OOD examples act as perturbed ID samples, however denser at areas where different ID classes samples are overlapped. Training with those pseudo features encourages more compact ID clusters and hence stronger OOD detection capabilities.

# 5. Conclusion

Given the success of supervised contrastive representation learning (SupCon) in learning powerful representations and with the aim to overcome the known overconfidence problem by Softmax classifiers, we propose a new OOD-aware training regime tailored for representations trained with SupCon. We start with SupCon loss [26] and suggest to jointly learn classes prototypes as an alternative to the Softmax CE loss. The prototypes are optimized to be close to their corresponding class samples. We regularize the training on ID data with auxiliary or pseudo OOD data. We propose two losses operating on OOD data, one at the projection head level and one at the encoder level. The first operates on pairwise samples' similarities and pushes OOD head features away from ID head features and the later pushes OOD encoder features far from the prototypes. We perform experiments on a wide range of OOD datasets and show a significant reduction on FPR. Our approach does not rely on a large auxiliary OOD dataset and moves a step closer to deploying OOD detector in practice by providing more reliable OOD rejection.

## References

[1] Rahaf Aljundi, Yash Patel, Milan Sulc, Nikolay Chumerin, and Daniel Olmeda Reino. Contrastive classification and representation learning with probabilistic interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6675–6683, 2023. 4

[2] Rahaf Aljundi, Daniel Olmeda Reino, Nikolay Chumerin, and Richard E. Turner. Continual novelty detection. In Sarath Chandar, Razvan Pascanu, and Doina Precup, editors, *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 1004–1025. PMLR, 22–24 Aug 2022. 4

[3] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020. 3

[4] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Smooth loss functions for deep top-k classification. *International Conference on Learning Representations*, 2018. 1

[5] Christopher M Bishop. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994. 2

[6] Antoine Buetti-Dinh, Vanni Galli, Sören Bellenberg, Olga Ilie, Malte Herold, Stephan Christel, Mariia Boretska, Igor V Pivkin, Paul Wilmes, Wolfgang Sand, et al. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports*, 22:e00321, 2019. 1

[7] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3, 5, 7, 8

[8] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[10] Hyunsoo Cho, Jinseok Seol, and Sang-goo Lee. Masked contrastive learning for anomaly detection. *International Joint Conferences on Artificial Intelligence*, 2021. 3

[11] Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. 2018. 2

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[14] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 1, 5

[15] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, 2022. 3

[16] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022. 5, 6, 7, 8

[17] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[21] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *ICML*, 2022. 2, 6

[22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2, 5, 7, 8

[23] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 2, 3, 5, 6, 7, 8

[24] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[25] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 5

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 1, 2, 3, 5, 6, 8

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 5

[29] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[31] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018. 2, 5, 7, 8

[32] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 2, 3, 5, 6, 7

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2017. 5

[34] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? 2019. 3

[35] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019. 2

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[37] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2906, 2019. 2

[38] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[39] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 2

[40] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021. 3, 5, 7, 8

[41] Joan Serrà, David Alvarez, Vicencc Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. 2020. 2

[42] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 2, 5, 7, 8

[43] 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. 2, 3, 5, 7, 8

[44] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008. 5

[45] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[46] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 5

[47] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR, 09–15 Jun 2019. 4

[48] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pages 23446–23458. PMLR, 2022. 2

[49] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. 2022. 1, 2, 5, 6, 7, 8

[50] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020. 2, 3

[51] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015. 5

[52] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 6, 7, 8

[53] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022. 5, 7

[54] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[55] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5

[56] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 4

[57] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person reidentification. *arXiv preprint arXiv:1711.08184*, 2017. 1

[58] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 5

[59] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018. 2