

RODD: A Self-Supervised Approach for Robust Out-of-Distribution Detection

Umar Khalid, Ashkan Esmaeili, Nazmul Karim, and Nazanin Rahnavard

Department of Electrical and Computer Engineering

University of Central Florida, USA

umarkhalid@knights.ucf.edu, ashkan.esmaeili@ucf.edu, nazmul.karim18@knights.ucf.edu,
nazanin.rahnavaard@ucf.edu

Abstract

Recent studies have started to address the concern of detecting and rejecting the out-of-distribution (OOD) samples as a major challenge in the safe deployment of deep learning (DL) models. It is desired that the DL model should only be confident about the in-distribution (ID) data which reinforces the driving principle of the OOD detection. In this paper, we propose a simple yet effective generalized OOD detection method independent of out-of-distribution datasets. Our approach relies on self-supervised feature learning of the training samples, where the embeddings lie on a compact low-dimensional space. Motivated by the recent studies that show self-supervised adversarial contrastive learning helps robustify the model, we empirically show that a pre-trained model with self-supervised contrastive learning yields a better model for uni-dimensional feature learning in the latent space. The method proposed in this work, referred to as RODD, outperforms SOTA detection performance on extensive suite of benchmark datasets on OOD detection tasks. On the CIFAR-100 benchmarks, RODD achieves a 26.97 % lower false positive rate (FPR@95) compared to SOTA methods. Our code is publicly available.¹

1. Introduction

In a real-world deployment, machine learning models are generally exposed to the *out-of-distribution* (OOD) objects that they have not experienced during the training. Detecting such OOD samples is of paramount importance in safety-critical applications such as health-care and autonomous driving [7, 19]. Therefore, the researchers have started to address the issue of OOD detection more recently [1, 2, 13–15, 26, 32, 39]. Most of the recent studies [11, 22, 23, 38] on OOD detection use OOD data for the model regularization such that some distance metric between the ID and OOD distributions is maximized. In recent studies [28, 30], generative models and auto-encoders have

been proposed to tackle OOD detection. However, they require OOD samples for hyper-parameter tuning. In the real-world scenarios, OOD detectors are distribution-agnostic. To overcome this limitation, some other methods that are independent of OOD data during the training process have been proposed [6, 13, 14, 17, 31, 36, 39]. Such methods either use the membership probabilities [6, 13, 14, 31] or a feature embedding [36, 39] to calculate an uncertainty score. In [36], the authors proposed to reconstruct the samples to produce a discriminate feature space. Similarly, [6] proposed synthesizing virtual outliers to regularize the model's decision boundary during training. Nevertheless, the performance of the methods that rely on either reconstruction or generation [6, 28, 36] degrades on large-scale datasets or video classification scenarios.

In this work, we claim that if the feature vectors belonging to each known class lie on a low-dimensional subspace, a representative singular vector can be calculated for each class that can be used to calculate uncertainty scores [39]. In order to achieve such a compact representation of the features belonging to each class, we have leveraged contrastive learning as a pre-training tool that has improved the performance of the proposed robust out-of-distribution detector (RODD) as it has helped the better feature mapping in the latent space during the downstream fine-tuning stage [18, 34]. Self-supervised pre-training, where we use adversaries as a form of data augmentation, helps to raise the RODD's performance in the settings with corrupted samples. This concept has been established by [3, 12, 16, 20, 35] that a self-supervised contrastive adversarial learning can generate an adversarially robust model during the fine-tuning. The overall architecture of the RODD is shown in Fig. 1.

In summary, we make the following contributions in this study. First, we propose that OOD detection test can be designed using the features extracted by self-supervised contrastive learning that reinforce the uni-dimensional projections of the ID set. Second, we have theoretically proved that such uni-dimensional projections, boosted by the contrastive learning, can be characterized by the prominent first

¹<https://github.com/UmarKhalidcs/RODD>

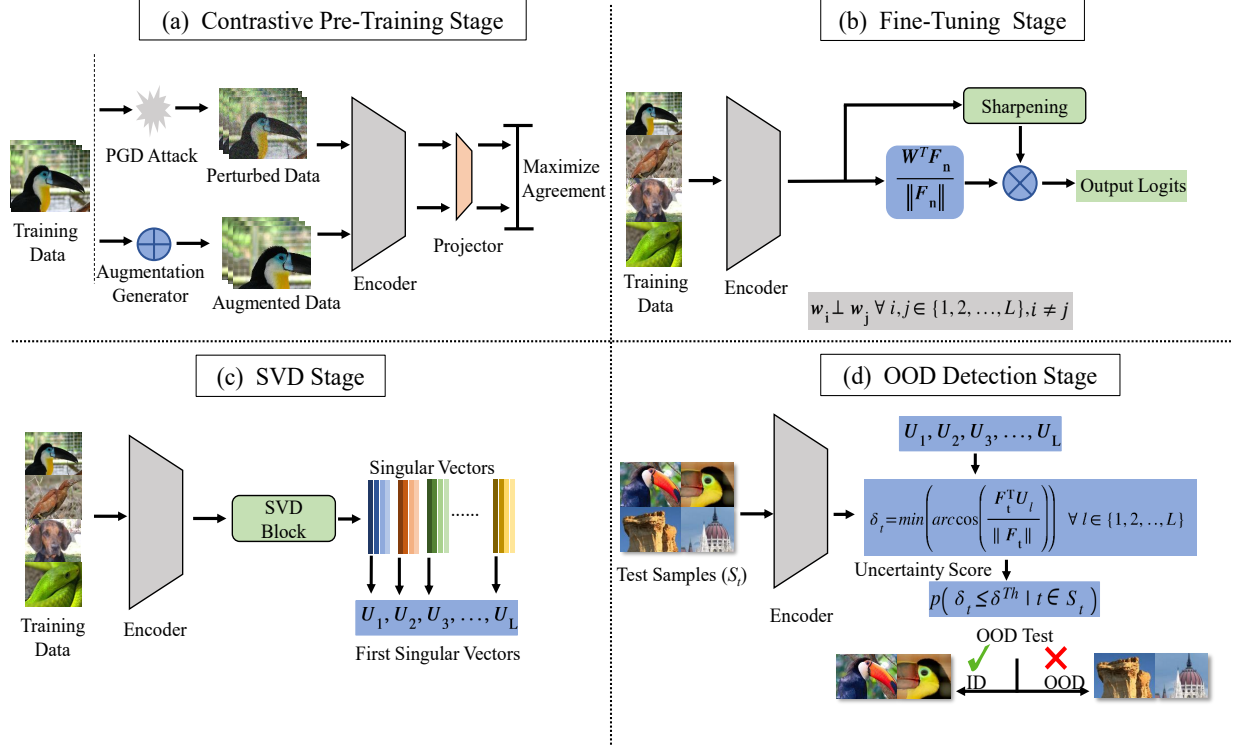


Figure 1. Overall architecture of the proposed OOD detection method. (a) In the first step, self-supervised adversarial contrastive learning is performed. (b) Secondly, the encoder is fine-tuned by freezing the weights ($\mathbf{W}$) of the penultimate layer. The columns of $\mathbf{W}$ are initialized to be orthonormal. (c) Thirdly, employing singular value decomposition (SVD), we calculate the first singular vector of each class using its features. (d) The final step is the OOD detection, where an uncertainty score is estimated using cosine similarity between the feature vector ($\mathbf{F}_t$) representing the test sample t and first singular vector of each ID class. Here, BN represents Batch Normalization, L is the number of classes, and δ^{th} is the threshold for the uncertainty score.

singular vector that represents its corresponding class attributes. Furthermore, the robustness of the proposed OOD detector has been evaluated by introducing corruptions in both OOD and ID datasets. Extensive experiments illustrate that the proposed OOD detection method outperforms the state-of-the-art (SOTA) algorithms.

2. Approach

Our proposed OOD detection approach builds upon employing a self-supervised training block to extract robust features from the ID dataset. This is carried out by training a contrastive loss on ID data as shown in Fig. 1 (a). Next, we utilize the concept of *union of one-dimensional-embeddings* to project the deep features of different classes onto one-dimensional and mutually orthogonal predefined vectors representing each class to obtain logits. At the final layer's output, we evaluate the cross-entropy between the logit probability output and the labels to form the supervised loss as shown in Fig. 1 (b). The uni-dimensional mapping is carried out to guarantee that intra-class distribution consists of samples aligning the most with the uni-dimensional vector characterizing its samples. To this end, the penultimate layer of the model is modified by using cosine similarity and introducing a sharpening layer as shown in Fig. 1 (b), where

$$\text{output logits are calculated as, } P(\mathbf{F}_n) = \frac{Z(\mathbf{F}_n)}{G(\mathbf{F}_n)}, \text{ where}$$

$$Z(\mathbf{F}_n) = \frac{\mathbf{W}^T \mathbf{F}_n}{\|\mathbf{F}_n\|}, G(\mathbf{F}_n) = \sigma(\text{BN}(\mathbf{W}_g^T \mathbf{F}_n)) \quad (1)$$

Here, $\mathbf{F}_n$ represents the encoder output for the training sample n , σ is the sigmoid function, and $\mathbf{W}_g$ is the weight matrix for the sharpening layer, represented by $G(\mathbf{F}_n)$, which essentially maps $\mathbf{F}_n$ to a scalar value. In the sharpening layer, batch normalization (BN) is used for faster convergence as proposed by [13]. It is worth mentioning that during the fine-tuning stage, we do not calculate the *bias* vector for the penultimate and sharpening layers.

The *orthogonality* comes with wide angles between the uni-dimensional embeddings of separates classes creating a large and expanded rejection region for the OOD samples if they lie in the vast inter-class space. To achieve this, we initialize the weight matrix $\mathbf{W} = [\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_L]$ of the penultimate layer with orthonormal vectors as in [29] and then freeze it during the fine-tuning stage. Here, $\mathbf{w}_l$ represents the weights of the last fully connected layer corresponding to class l . During fine-tuning, the features are projected onto the predefined set of orthogonal vectors $\mathbf{w}_l$ for $l = 1, 2, \dots, L$, where L is the number of ID classes.

After training, OOD testing can be done by evaluating the inner products between the calculated first singular

vectors ($\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_L$) representing their corresponding classes as shown in Fig. 1 (c), and the extracted feature for the sample of interest. To perform OOD inspection on the test sample $t \in S_t$, where S_t is the test set, the uncertainty score is calculated as,

$$\delta_t = \min(\arccos\left(\frac{\mathbf{F}_t^T \mathbf{U}_l}{\|\mathbf{F}_t\|}\right)), \forall l \in \{1, 2, \dots, L\} \quad (2)$$

Here, $\mathbf{F}_t$ is the output of the encoder for the test sample t . The measured uncertainty is then used to calculate the probability that if t belongs to ID or OOD using the probability function $p(\delta_t \leq \delta^{Th} | t \in S_t)$ as RODD is a probabilistic approach where sampling is performed during the test time. In an ideal scenario, features of ID class l have to be aligned with the corresponding $\mathbf{w}_l$, where $\mathbf{w}_l$ is the l^{th} column of matrix $\mathbf{W}$. In that case, $\delta^{Th} = 0$. However, in practice, all class features are not exactly aligned with their respective column in $\mathbf{W}$, that further strengthens the idea of using the first singular vector of each class feature matrix, separately.

Next, we will explain how the contrastive learning pre-training and sharpening module, $G(\mathbf{F}_n)$, boosts the performance of our approach. Firstly, contrastive learning has been beneficial because we do not freeze the weights of the encoder after the self-supervised learning and keep fine-tuning them along the training procedure using the cross-entropy loss. In other words, the features are warm-started with initialized values derived from the contrastive loss pre-training, yet the final objective function to optimize is composed of two terms $\mathcal{L}_{CL} + \mu\mathcal{L}_{LL}$, where $\mathcal{L}_{CL}$ and $\mathcal{L}_{LL}$ denote the contrastive and cross-entropy losses, respectively. In addition, the cross-entropy loss imposes the orthogonality assumption infused by the choice of orthogonal matrix containing union of $\mathbf{w}_l \forall l \in \{1, 2, \dots, L\}$ each of which represent one class. By feeding the inner products of features with $\mathbf{W}$ into $\mathcal{L}_{LL}$, the features are endorsed to get reshaped to satisfy orthogonality and rotate to align $\mathbf{w}_l$.

Furthermore, augmenting the data of each class with the adversarial perturbations can improve classification performance on ID perturbed data while still detecting the OOD data [3, 20]. Moreover, prior to feeding the optimizer with the inner products for supervised training, we modify the uni-dimensional mappings using $G(\mathbf{F}_n)$ to optimally benefit from the self-supervised learned features. To compensate for the uni-dimensional confinement which can downgrade the classifier's performance, we use the *sharpening* concept, where we enhance the confidence of the obtained logit vector by scaling the inner products with a factor denoted with the sharpening function $G(\mathbf{F}_n)$ explained above.

2.1. Theoretical Analysis

In this section, we provide theoretical analyses on how pre-training with contrastive loss promotes the uni-dimensional embeddings approach utilized in RODD by promoting one prominent singular vector (with a dominant singular value) in the deep feature extraction layer.

The objective function used in our optimization is composed of a contrastive loss and a softmax cross entropy. For simplicity, we use a least squared loss measuring the distance between linear prediction on a sample's extracted feature to its label vector $\|\mathbf{W}^T \mathbf{F}_n - \mathbf{y}_n\|_2^2$ as a surrogate for the softmax cross entropy ($\mathcal{L}_{LL}$)². This is justified in [34].

Let $\mathbf{A} = [a_{i,j}]$ denote the adjacency matrix for the augmentation graph of training data formally defined as in [34]. In general, two samples are connected through an edge on this graph if they are believed to be generated from the same class distribution. Without loss of generality, we assume that the adjacency matrix is block-diagonal, i.e., different classes are well-distinguished. Therefore, the problem can be partitioned into data specific to each class. Let $\mathbf{F}$ and $\mathbf{Y}$ denote the matrix of all features and label vectors, i.e., $\mathbf{F}_n$ and $\mathbf{y}_n$, where n denotes the n^{th} sample, respectively. The training loss including one term for contrastive learning loss and one for the supervised uni-dimensional embedding matching can be written as:³

$$\mathcal{L}(\mathbf{F}) = \underbrace{\|\mathbf{A} - \mathbf{F}\mathbf{F}^T\|_F^2}_{\mathcal{L}_{CL}(\mathbf{F})} + \mu \underbrace{\|\mathbf{W}^T \mathbf{F} - \mathbf{Y}\|_F^2}_{\mathcal{L}_{LL}(\mathbf{F})}. \quad (3)$$

$\mathbf{Y}$ and $\mathbf{A}$ are given matrices, and $\mathbf{W}$ is fixed to some orthonormal predefined matrix. The optimization variable is therefore the matrix $\mathbf{F}$. Thus, the optimization problem can be written as:

$$\min_{\mathbf{F}} \|\mathbf{A} - \mathbf{F}\mathbf{F}^T\|_F^2 + \mu \|\mathbf{W}^T \mathbf{F} - \mathbf{Y}\|_F^2. \quad (4)$$

Before bringing the main theorem, two assumptions are made on the structure of the adjacency matrix arising from its properties [34]: **1:** For a triple of images $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_s$, we have $\frac{a_{i,j}}{a_{j,s}} \in [\frac{1}{1+\delta}, 1 + \delta]$ for small δ , i.e., samples of the same class are similar. **2:** For a quadruple of images $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_s, \mathbf{x}_t$, where $\mathbf{x}_i, \mathbf{x}_j$ are from different classes and $\mathbf{x}_s, \mathbf{x}_t$ are from the same classes, $\frac{a_{i,j}}{a_{s,t}} \leq \eta$ for small η .

Lemma 1. Let $\mathbf{F}^*$ denote the solution to $\min_{\mathbf{F}} \mathcal{L}_{CL}$ (first loss term in (4)). Assume $\mathbf{F}^*$ can be decomposed as $\mathbf{F}^* = \mathbf{U}\Sigma\mathbf{V}^T$. Under Assumptions 1,2 (above), for $\mathbf{F}^*$ with singular values σ_i , we have $\sum_{i=2}^{N_l} \sigma_i^2 \leq \sqrt{6((1+\delta)^{\frac{3}{2}} - 1)}$ for some small δ , where $\sigma_i = \Sigma_{ii}$, and N_l is the number training samples of class l .

Proof. In [34], it is shown that $\sum_{i=2}^{N_l} \sigma_i^4 \leq 2((1+\delta)^{\frac{3}{2}} - 1)$. The proof is straightforward powering $\sum_{i=2}^{N_l} \sigma_i^2$ by two and applying Cauchy-Schwartz inequality.

Theorem 1. Let $\mathbf{F}^*$ denote the solution to (4). Assume $\mathbf{F}^*$ is decomposed as $\mathbf{F}^* = \mathbf{U}\Sigma\mathbf{V}^T$. There exist a μ_{min} such that $\sum_{i=2}^{N_l} \sigma_i^4 \leq 2((1+\delta)^{\frac{3}{2}} - 1)$, if $\mu < \mu_{min}$ in P (4).

²The least squared loss ($\mathcal{L}_{LL}$) measures the distance of the final layer predictions (assuming linear predictor in the deep feature space) from the one-hot encoded vector (alternatively logits if available)

³It is shown in [8] that the solution to the contrastive learning loss can be written as the following Cholesky decomposition problem, $\min_{\mathbf{F}} \|\mathbf{A} - \mathbf{F}\mathbf{F}^T\|_F^2$, which constitutes the first term of the loss in Eq. (3).

The purpose is to show that treating corrupted or adversarial ID data vs. OOD data, the uni-dimensional embedding is robust in OOD rejection. This mandates invariance and stability of the first singular vector for the features extracted for samples generated from each class. The goal of this theorem is to show that using the contrastive loss along certain values of μ regularizing the logit loss, the dominance of the first eigenvector of the adjacency matrix is also inherited to the first singular vector of the $\mathbf{F}$ and this is inline with the mechanism of proposed approach whose functionality depends on the stability and dominance of the first singular vector because we desire most of the information included in the samples belonging to each class can be reflected in uni-dimensional projections.

Assuming the dominance is held for the first singular value of each class data, the contrastive learning can therefore split them by summarizing the class-wise data into uni-dimensional separate representations. The $\mathbf{V}$ matrix is used to orthogonalize and rotate the uni-dimensional vectors obtained by contrastive learning to match the pre-defined orthogonal set of vectors $\mathbf{w}_l$ as much as possible.

Now the proof for the main theorem is provided.

Proof. $\mathbf{A}$ is Hermitian. Therefore, it can be decomposed as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$. The solution set to minimize $\mathcal{L}_{CL}$ is $\mathcal{S} = \{\mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T : \forall \text{ orthonormal matrix } \mathbf{V}\}$ ($\lambda_i = \Lambda_{ii} = \sigma_i^2$).

Let L_1 and L_2 be the minima for (4) obtained on the sets $\mathcal{S}$ and $\mathcal{S}^c$, i.e., the complementary set of $\mathcal{S}$. L_1 equals $\mu \min_{\mathbf{F} \in \mathcal{S}} \mathcal{L}_{LL}(\mathbf{F})$ as the first loss is 0 for elements in $\mathcal{S}$. Now, we consider L_2 . $\mathcal{S}^c$ can be partitioned into two sets $\mathcal{S}_1^c$ and $\mathcal{S}_2^c$, where elements in $\mathcal{S}_1^c$ set $\mathcal{L}_{LL}$ to zero and elements in $\mathcal{S}_2^c$ yield non-zero values for $\mathcal{L}_{LL}$. Therefore, L_2 is the minimum of the two partition's minima.

$$L_2 = \min \left\{ \underbrace{\min_{\mathbf{F} \in \mathcal{S}_1^c} \mathcal{L}_{CL}(\mathbf{F})}_{LHS}, \underbrace{\min_{\mathbf{F} \in \mathcal{S}_2^c} \mathcal{L}_{CL}(\mathbf{F}) + \mu \mathcal{L}_{LL}(\mathbf{F})}_{RHS} \right\} \quad (5)$$

It is obvious that for a small enough μ , L_2 equals the RHS above. This can be reasoned as follows. Let the LHS value be denoted with m_1 . $m_1 > 0$ since $\mathcal{S}$ and $\mathcal{S}_1^c$ are disjoint sets with no sharing boundaries. The RHS in (5) is composed of two parts. The first part can be arbitrarily small because although $\mathcal{S}$ and $\mathcal{S}_2^c$ are disjoint, they are connected sets with sharing boundaries. (For instance any small perturbation in $\mathbf{A}$ eigenvalues drags a matrix from $\mathcal{S}$ into $\mathcal{S}_2^c$. However, they are infinitesimally close due to the continuity property). The second term can also be shrunk with an arbitrarily small choice of $\mu = \mu_{min} = \frac{m_1}{\mathcal{L}_{LL}(\tilde{\mathbf{F}})}$ that guarantees the RHS takes the minimum in Eq. (5), where $\tilde{\mathbf{F}} = \arg \min_{\mathbf{F} \in \mathcal{S}_2^c} \mathcal{L}_{CL}(\mathbf{F})$ ⁴. Therefore, for $\mu < \mu_{min}$,

⁴(As discussed, $\tilde{\mathbf{F}}$ makes the first term arbitrarily approach 0 due to continuity property holding between $\mathcal{S}$ and $\mathcal{S}_2^c$ and there is an element in

the minimum objective value in Eq. (4) ($\min\{L_1, L_2\}$) is, $\min \{ \min_{\mathbf{F} \in \mathcal{S}_2^c} \mathcal{L}_{CL}(\mathbf{F}) + \mu \mathcal{L}_{LL}(\mathbf{F}), \min_{\mathbf{F} \in \mathcal{S}} \mu \mathcal{L}_{LL}(\mathbf{F}) \}$.

The final aim is to show that μ can be chosen such that $\mathbf{F}^*$ inherits the dominance of first eigenvalue from $\mathbf{A}$. This is straightforward if the solution is RHS in (5) because the solution lies on $\mathcal{S}$ in that case and therefore, can be expressed as $\mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T$ inheriting the property in Lemma 1.

Thus, we first consider cases where $\min\{L_1, L_2\}$ is obtained by the RHS by explicitly writing when LHS > RHS. We assume the minimizers for the RHS and LHS differ in a matrix $\mathbf{R}$. Let $\mathbf{F}^*$ denote the minimizer for RHS. Then, the minimizer of LHS is $\mathbf{F}^* + \mathbf{R}$. We have

$$\begin{aligned} LHS &= \|\mathbf{A} - (\mathbf{F}^* + \mathbf{R})(\mathbf{F}^* + \mathbf{R})^T\|_F^2 + \mu \|\mathbf{W}^T \mathbf{F}^* + \mathbf{W}^T \mathbf{R} - \mathbf{Y}\|_F^2 \\ &= \underbrace{\|\mathbf{A} - \mathbf{F}^* \mathbf{F}^{*T}\|_F^2}_0 - \underbrace{(\mathbf{F}^* \mathbf{R}^T + \mathbf{R} \mathbf{F}^{*T} + \mathbf{R} \mathbf{R}^T)}_{\mathbf{E}} + \mu \|\mathbf{W}^T \mathbf{F}^* - \mathbf{Y} + \mathbf{W}^T \mathbf{R}\|_F^2 \\ &= \|\mathbf{E}\|_F^2 + \mu \|\mathbf{W}^T \mathbf{R}\|_F^2 + 2\mu \langle \mathbf{W}^T \mathbf{F}^* - \mathbf{Y}, \mathbf{W}^T \mathbf{R} \rangle, \end{aligned}$$

where the inner product of two matrices $\mathbf{A}, \mathbf{B}$ ($\langle \mathbf{A}, \mathbf{B} \rangle$) is defined as $Tr(\mathbf{A}\mathbf{B}^T)$. The RHS in (5) equates $\mu \|\mathbf{W}^T \mathbf{F}^* - \mathbf{Y}\|_F^2$ since $\mathbf{F}^*$ is its minimizer and the loss has only the logit loss term. Thus, the condition LHS > RHS reduces to $\|\mathbf{E}\|_F^2 + \mu \|\mathbf{W}^T \mathbf{R}\|_F^2 + 2\mu \langle \mathbf{W}^T \mathbf{F}^* - \mathbf{Y}, \mathbf{W}^T \mathbf{R} \rangle > 0$.

Using the fact that the matrix $\mathbf{W}$ is predefined to be an orthonormal matrix, multiplying it by $\mathbf{R}$ does not change the Frobenius norm. Hence, the condition reduces to $\|\mathbf{E}\|_F^2 + \mu \|\mathbf{R}\|_F^2 > 2\mu \langle \mathbf{Y} - \mathbf{W}^T \mathbf{F}^*, \mathbf{W}^T \mathbf{R} \rangle$. To establish this bound, the Cauchy-Schwartz inequality (C-S) and the Inequality of Arithmetic and Geometric Means (AM-GM) are used to obtain the upper bound for the inner product. The sufficient condition holds true if it is established for the obtained upper bound (tighter inequality). Applying (C-S) and (AM-GM) inequalities we have

$$\begin{aligned} \langle \mathbf{Y} - \mathbf{W}^T \mathbf{F}^*, \mathbf{W}^T \mathbf{R} \rangle &\stackrel{C-S}{\leq} \|\mathbf{Y} - \mathbf{W}^T \mathbf{F}^*\|_F \|\mathbf{W}^T \mathbf{R}\|_F = \\ &\stackrel{AM-GM}{\leq} \frac{1}{2} \|\mathbf{Y} - \mathbf{W}^T \mathbf{F}^*\|_F^2 + \frac{1}{2} \|\mathbf{R}\|_F^2 \end{aligned}$$

Substituting this for the inner product to establish a tighter inequality, we get $\|\mathbf{E}\|_F^2 + \mu \|\mathbf{R}\|_F^2 > \mu \|\mathbf{Y} - \mathbf{W}^T \mathbf{F}^*\|_F^2 + \mu \|\mathbf{R}\|_F^2$ reducing to $\|\mathbf{E}\|_F^2 > \mu \|\mathbf{Y} - \mathbf{W}^T \mathbf{F}^*\|_F^2$.

As the matrix of all zeros, i.e., $[0] \in \mathcal{S}$, inserting $[0]$ for $\mathbf{F}$ leads to a trivial upper bound for the minimum obtained over $\mathbf{F} \in \mathcal{S}$, i.e., $\|\mathbf{Y} - \mathbf{W}^T \mathbf{F}^*\|_F^2$ is upper bounded with $\|\mathbf{Y}\|_F^2$. Finding a condition for $\|\mathbf{E}\|_F^2 > \mu_{min} \|\mathbf{Y}\|_F^2$ guarantees the desired condition is satisfied. If $\|\mathbf{E}\|_F^2 > \mu_{min} \|\mathbf{Y}\|_F^2$ is met, the solution lies in $\mathcal{S}$ and RHS obtains the minimum, validating Lemma 1 for $\mathbf{F}^*$.

Otherwise, if the solution lies in $\mathcal{S}_2^c$ and is attained from the LHS such that it contravenes the dominance of the first principal component of $\mathbf{A}$, we will show by contradiction that the proper choice for μ avoids LHS to be less than the

$\mathcal{S}_2^c$ arbitrarily close to $\tilde{\mathbf{F}}$)

RHS in (5). To this end, we take a more profound look into $\|\mathbf{E}\|_F^2$. If $\mathbf{R}$ is to perturb the solution $\mathbf{F}^*$ such that the first principal component is not prominent, for $\mathbf{R} + \mathbf{F}^*$, we shall have $\sum_{i=2}^{N_l} \sigma_i^2 > \Delta + \alpha$ for some positive α violating the condition stated in the Theorem. This means there is at least one singular value of $\mathbf{F}^* + \mathbf{R}$, for which we have $\sigma_r > \sqrt{\frac{\Delta+\alpha}{N_l-1}} = \sqrt{\frac{\alpha}{N_l-1}} + \mathcal{O}(\sqrt[4]{\delta})$. As $\mathbf{F}^*$ inherits the square root of eigenvalues of $\mathbf{A}$, according to Lemma 1 and using Taylor series expansion, $\sigma_r(\mathbf{F}^*) = \mathcal{O}(\sqrt[4]{\delta})$. This yields $\sigma_r(R) > \sqrt{\frac{\alpha}{N_l-1}} + \mathcal{O}(\sqrt[4]{\delta})$. $\mathbf{E}$ is a symmetric matrix and therefore it has eigenvalue decomposition. $\|\mathbf{E}\|_F^2 \geq \lambda_r^2(\mathbf{E}) = \lambda_r^2(\mathbf{R}\mathbf{R}^T + \mathbf{R}\mathbf{F}^{*T} + \mathbf{F}^*\mathbf{R}^T) = \lambda_r^2(\mathbf{R}\mathbf{R}^T) + \mathcal{O}(\delta) > \frac{\alpha^2}{(N_l-1)^2} + \mathcal{O}(\delta)$. Knowing that $\|\mathbf{Y}\|_F^2 = N_l^2$, if $\mu < \frac{\alpha^2}{N_l^4}$, the condition for RHS<LHS is met. According to Lemma 1 and the previous bound found for μ_{min} , if $\mu_{min} < \min\{\frac{\alpha^2}{N_l^4}, \frac{m_1}{\mathcal{L}_{LL}(\mathbf{F})}\}$, the solution should be $\mathbf{F}^* = \mathbf{Q}\mathbf{A}^{\frac{1}{2}}\mathbf{V}^T$. Hence, for certain range of values for μ , the solution takes the form $\mathbf{Q}\mathbf{A}^{\frac{1}{2}}\mathbf{V}$ obeying the dominance of λ_1 in $\mathbf{A}$ and this concludes the proof.

3. Experiments

In this section, we evaluate our proposed OOD detection method through extensive experimentation on different ID and OOD datasets with multiple architectures.

3.1. Datasets and Architecture

In our experiments, we used CIFAR-10 and CIFAR-100 [21] as ID datasets and 7 OOD datasets. OOD datasets utilized are TinyImageNet-crop (TINc), TinyImageNet-resize(TINr) [5], LSUN-resize (LSUN-r) [37], Places [41], Textures [4], SVHN [27] and iSUN [33]. For an architecture, we deployed WideResNet [40] with depth and width equal to 40 and 2, respectively, as an encoder in our experiments. However, the penultimate layer has been modified as compared to the baseline architecture as shown in Fig. 1.

3.2. Evaluation Metrics and Inference Criterion

As in [6, 31], the OOD detection performance of RODD is evaluated using the following metrics: (i) *FPR95* indicates the false positive rate (FPR) at 95% true positive rate (TPR) and (ii) *AUROC*, which is defined as the Area Under the Receiver Operating Characteristic curve. As RODD is a probabilistic approach, sampling is preformed on the ID and OOD data during the test time to ensure the probabilistic settings. We employ Monte Carlo sampling to estimate $p(\delta_t \leq \delta^{Th})$ for OOD detection, where δ^{Th} is the uncertainty score threshold calculated using training samples. During inference, 50 samples are drawn for a given sample, t . The evaluation metrics are then applied on ID test data and OOD data using the estimated δ^{Th} to calculate the difference in the feature space.

3.3. Results

We show the performance of RODD in Tables 1 and 2 for CIFAR-10 and CIFAR-100, respectively. Our method achieves an FPR95 improvement of 21.66%, compared to the most recently reported SOTA [6], on CIFAR-10. We obtain similar performance gains for CIFAR-100 dataset as well. For RODD, the model is first pre-trained using self-supervised adversarial contrastive learning [16]. We fine-tune the model following the training settings in [40].

4. Ablation Studies

In this section, we conduct extensive ablation studies to evaluate the robustness of RODD against corrupted ID and OOD test samples. Firstly, we apply the 14 corruptions in [9] on OOD data to generate *corrupted* OOD (OOD-C). Corruptions introduced can be benign or destructive based on their intensity which is defined by their severity level. To do comprehensive evaluations, 5 severity levels of the corruptions are infused. By introducing such corruptions in OOD datasets, the calculated mean detection error for both CIFAR-10 and CIFAR-100 is 0%, which highlights the inherit property of RODD that it shifts perturbed OOD features further away from the ID as shown in t-SNE plots in Fig. 2 which shows that perturbing OOD improves the RODD's performance. Secondly, we introduced corruptions [9] in the ID test data while keeping OOD data clean during testing. The performance of RODD on corrupted CIFAR-100 (CIFAR100-C) has been compared with VOS [6] in Table 3. Lastly, we compared the classification accuracy of our proposed method with the baseline WideResNet model [40] on clean and corrupted ID test samples in Table 4. RODD has improved accuracy on corrupted ID test data as compared to the baseline with a negligible drop on classification accuracy of clean ID test data.

5. Conclusion

In this work, we have proposed that in-distribution features can be aligned in a narrow region of the latent space using constrastive pre-training and uni-dimensional feature mapping. With such compact mapping, a representative first singular vector can be calculated from the features for each in-distribution class. The cosine similarity between these computed singular vectors and an extracted feature vector of the test sample is then estimated to perform OOD test. We have shown through extensive experimentation that our method achieves SOTA OOD detection results on CIFAR-10 and CIFAR-100 image classification benchmarks.

6. Acknowledgement

This research is based upon work supported by Leonardo DRS and partly by the National Science Foundation under Grant No. CCF-1718195 and ECCS-1810256.

Table 1. OOD detection results of RODD and comparison with competitive baselines trained on CIFAR-10 as ID dataset. All values are shown in percentages. $\uparrow$ indicates larger values are better and $\downarrow$ indicates smaller values are better.

Methods	OOD Datasets													
	SVHN		iSUN		LSUNr		TINc		TINr		Places		Textures	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$
MSP [10]	48.49	91.89	56.03	89.83	52.15	91.37	53.15	87.33	54.24	79.35	59.48	88.20	59.28	88.50
ODIN [24]	33.55	91.96	32.05	93.50	26.52	94.57	36.75	89.20	49.15	81.64	57.40	84.49	49.12	84.97
Mahalanobis [23]	12.89	97.62	44.18	92.66	42.62	93.23	42.75	88.85	52.25	80.33	92.38	33.06	15.00	97.33
Energy [25]	35.59	90.96	33.68	92.62	27.58	94.24	35.69	89.05	50.45	81.33	40.14	89.89	52.79	85.22
OE [11]	4.36	98.63	6.32	98.85	5.59	98.94	13.45	96.44	15.67	96.78	19.07	96.16	12.94	97.73
VOS [6]	8.65	98.51	7.56	98.71	14.62	97.18	11.76	97.58	28.08	94.26	37.61	90.42	47.09	86.64
FS [39]	24.71	95.31	17.41	96.61	4.84	96.28	12.45	97.83	9.65	97.95	11.56	96.42	5.55	98.64
RODD (Ours)	1.82	99.63	4.07	99.32	4.49	99.25	10.29	98.10	6.30	99.0	9.59	98.47	3.87	99.43

Table 2. OOD detection results of RODD and comparison with competitive baselines trained on CIFAR-100 as ID dataset. All values are shown in percentages. $\uparrow$ indicates larger values are better and $\downarrow$ indicates smaller values are better.

Methods	OOD Datasets													
	SVHN		iSUN		LSUNr		TINc		TINr		Places		Textures	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$
MSP [10]	84.59	71.44	82.80	75.46	82.42	75.38	69.82	79.77	79.95	72.36	82.84	73.78	83.29	73.34
ODIN [24]	84.66	67.26	68.51	82.69	71.96	81.82	45.55	87.77	57.34	80.88	87.88	71.63	49.12	84.97
Mahalanobis [23]	57.52	86.01	26.10	94.58	21.23	96.00	43.45	86.65	44.45	85.68	88.83	67.87	39.39	90.57
Energy [25]	85.52	73.99	81.04	78.91	79.47	79.23	68.85	78.85	77.65	74.56	40.14	89.89	52.79	85.22
OE [11]	65.91	86.66	72.39	78.61	69.36	79.71	46.75	85.45	78.76	75.89	57.92	85.78	61.11	84.56
VOS [6]	65.56	87.86	74.65	82.12	70.58	83.76	47.16	90.98	73.78	81.58	84.45	72.20	82.43	76.95
FS [39]	22.75	94.33	45.45	85.61	40.52	87.21	11.76	97.58	44.08	86.26	47.61	88.42	47.09	86.64
RODD (Ours)	19.89	95.76	39.79	88.40	36.61	89.73	44.42	85.95	42.56	87.67	41.72	89.10	24.64	94.14

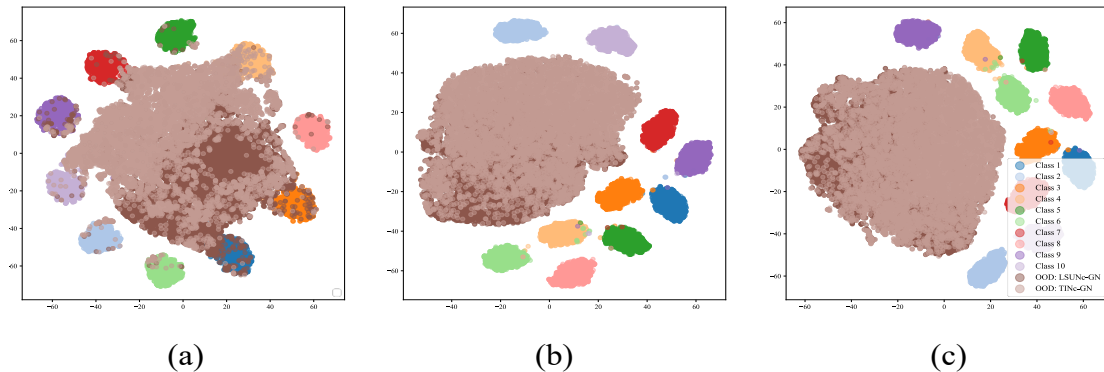


Figure 2. t-SNE representation of features extracted by introducing Gaussian noise on OOD dataset. 10,000 samples each of TINc and LSUNc while 1,000 sample of each class from ID CIFAR-10 test set are used to generate 2D t-SNE plot. (a) Features extracted from the baseline model with severity level 1. (b) Features extracted using RODD with corruption severity level 1. (c) Features extracted from the RODD with corruption severity level 5.

Table 3. Evaluation using corrupted ID test samples for CIFAR-100. All values are in % and averaged over 7 OOD datasets discussed in Section 3.1 whereas corruption severity is varied from 1-5 as in [9]. $\uparrow$ indicates larger values are better and $\downarrow$ indicates smaller values are better.

Dataset	Method	Clean	Noise			Blur			Weather				Digital			
			Gauss	Shot	Impulse	Defocus	Motion	Zoom	Snow	Frost	Fog	Bright	Cont.	Elastic	Pixel	JPEG
$\downarrow$ FPR95	VOS	66.79	72.55	76.95	90.36	84.50	83.62	84.56	87.0	83.34	83.84	86.11	86.67	85.81	89.58	89.25
	RODD	39.76	67.91	65.42	65.53	49.51	71.81	55.87	53.92	59.84	52.23	48.39	52.98	57.31	55.42	66.47
$\uparrow$ AUROC	VOS	81.9	74.26	72.90	60.00	68.35	69.83	68.55	65.31	68.14	68.50	66.54	66.82	66.98	61.18	62.38
	RODD	88.1	77.18	78.40	78.41	84.70	74.64	82.42	83.50	80.60	83.85	85.54	83.44	81.91	83.11	78.19

Table 4. Clean and corruption accuracy (%) of RODD and Baseline on CIFAR10-C and CIFAR100-C.

Dataset	Method	Clean	Noise			Blur			Weather				Digital			
			Gauss	Shot	Impulse	Defocus	Motion	Zoom	Snow	Frost	Fog	Bright	Cont.	Elastic	Pixel	JPEG
CIFAR10-C	Baseline	94.52	46.54	57.72	56.45	69.15	62.98	58.85	74.88	72.18	84.26	92.19	75.14	74.31	68.27	77.34
	RODD	94.45	49.63	59.89	55.62	69.77	64.81	61.79	78.59	74.48	86.56	93.08	73.37	75.49	70.79	80.12
CIFAR100-C	Baseline	72.35	18.80	26.56	25.56	49.80	40.45	39.37	45.38	42.62	56.40	69.14	52.87	48.32	40.70	46.11
	RODD	72.20	18.40	27.13	26.25	50.32	41.82	40.40	46.25	43.46	57.13	70.0	51.81	49.05	40.86	47.62

References

- [1] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15701–15710, 2021.
- [2] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–445. Springer, 2021.
- [3] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020.
- [4] Mircea Cimpoi, Subhansu 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] 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.
- [6] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don’t know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.
- [7] Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2020.
- [8] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.
- [9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.
- [10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.
- [11] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.
- [12] Chih-Hui Ho and Nuno Vnasconcelos. Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33:17081–17093, 2020.
- [13] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.
- [14] 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.
- [15] Taewon Jeong and Heeyoung Kim. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. *Advances in Neural Information Processing Systems*, 33, 2020.
- [16] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33:16199–16210, 2020.
- [17] Mohsen Joneidi, Saeed Vahidian, Ashkan Esmaeili, Weijia Wang, Nazanin Rahnavard, Bill Lin, and Mubarak Shah. Select to better learn: Fast and accurate deep learning using data selection from nonlinear manifolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7819–7829, 2020.
- [18] Nazmul Karim, Umar Khalid, Nick Meeker, and Sarinda Samarasinghe. Adversarial training for face recognition systems using contrastive adversarial learning and triplet loss fine-tuning, 2021.
- [19] Umar Khalid, Nazmul Karim, and Nazanin Rahnavard. Rf signal transformation and classification using deep neural networks. *arXiv preprint arXiv:2204.03564*, 2022.
- [20] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.
- [21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [22] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.
- [23] 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.
- [24] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- [25] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [26] Siyu Luan, Zonghua Gu, Leonid B Freidovich, Lili Jiang, and Qingling Zhao. Out-of-distribution detection for deep neural networks with isolation forest and local outlier factor. *IEEE Access*, 9:132980–132989, 2021.
- [27] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.
- [28] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31, 2018.
- [29] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

- [30] Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. *Advances in Neural Information Processing Systems*, 31, 2018.
- [31] Yiyu Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34, 2021.
- [32] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 550–564, 2018.
- [33] 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.
- [34] Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise, 2022.
- [35] Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. Improved ood generalization via adversarial training and pre-training, 2021.
- [36] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019.
- [37] 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.
- [38] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9518–9526, 2019.
- [39] Alireza Zaeemzadeh, Niccolò Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9452–9461, 2021.
- [40] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [41] 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.