Out-of-Distribution Detection Using Union of 1-Dimensional Subspaces

Alireza Zaeemzadeh
University of Central Florida
zaeemzadeh@eecs.ucf.edu

Niccolò Bisagno
University of Trento
niccolo.bisagno@unitn.it

Zeno Sambugaro
University of Trento
zeno.sambugaro@unitn.it

Nicola Conci
University of Trento
nicola.conci@unitn.it

Nazanin Rahnavard
University of Central Florida
nazanin@eecs.ucf.edu

Mubarak Shah
University of Central Florida
shah@crcv.ucf.edu

Abstract

The goal of out-of-distribution (OOD) detection is to handle the situations where the test samples are drawn from a different distribution than the training data. In this paper, we argue that OOD samples can be detected more easily if the training data is embedded into a low-dimensional space, such that the embedded training samples lie on a union of 1*-dimensional subspaces. We show that such embedding of the in-distribution (ID) samples provides us with two main advantages. First, due to compact representation in the feature space, OOD samples are less likely to occupy the same region as the known classes. Second, the first singular vector of ID samples belonging to a* 1*-dimensional subspace can be used as their robust representative. Motivated by these observations, we train a deep neural network such that the ID samples are embedded onto a union of* 1*-dimensional subspaces. At the test time, employing sampling techniques used for approximate Bayesian inference in deep learning, input samples are detected as OOD if they occupy the region corresponding to the ID samples with probability* 0*. Spectral components of the ID samples are used as robust representative of this region. Our method does not have any hyperparameter to be tuned using extra information and it can be applied on different modalities with minimal change. The effectiveness of the proposed method is demonstrated on different benchmark datasets, both in the image and video classification domains.*

1. Introduction

Many classification methods are designed and deployed under the assumption that training data contains samples from all the possible classes that the classifier will encounter during testing. Of course, such assumption does not hold in many applications; as it may not be possible to cover every potential input class in the training set. Thus, it is desirable to detect out-of-distribution (OOD) samples; the input instances that do not belong to any of the training classes. In general, OOD detection techniques try to either use the class membership probabilities as a measure of uncertainty [12, 21, 36, 39, 14], or define a measure of similarity between the input samples and the training dataset in a feature space [2, 40, 20, 28]. As discussed in [20], the features extracted from a conventional softmax classifier follow a class-conditional Gaussian distribution. However, general class-conditional Gaussian embeddings are not particularly appropriate for outlier detection, as they are not easily distinguishable in the feature space.

In this work, we claim that we can improve the OOD detection performance by *constraining* the representation of in-distribution (ID) samples in the feature space. Particularly, if we embed the training samples such that the feature vectors belonging to each known class lie on a 1-dimensional subspace, OOD samples can be detected more robustly with higher probability, compared to a class-conditional non-degenerate Gaussian embeddings. Such a *union of* 1*-dimensional subspaces* representation provides us with two main advantages. First, due to compact representation in the feature space, OOD samples are less likely to occupy the same region as the known classes. In other words, a random vector in a high-dimensional space lies on a specific 1-dimensional line with probability 0. Second, we show that the first singular vector of a 1-dimensional subspace is a robust representative of its samples. We exploit these two desirable features and reject samples as OOD, if they occupy the region corresponding to the training samples with probability 0. This region is identified by the set of the first singular vectors of the training classes. To estimate the probability, we use Monte Carlo sampling techniques used in Bayesian deep learning such as [25, 8].

Our work is primarily motivated by the rich literature of spectral methods in signal processing and machine learning. Spectral techniques have been proven to be very effective

for different tasks such as robust estimation [6], learning mixture models [29], representative selection [43], and defense against backdoor attacks [35]. We are also inspired by the OOD detection method proposed in [20], in which authors use the ID feature vectors to estimate their distribution and to detect OOD samples. In contrast, we engineer the distribution of ID feature vectors to minimize the error probability, without knowing the distributions of OOD samples, and enforce our desired distribution on the feature vectors. Our proposed method does not need extra information or a subset of OOD examples for hyperparameter tuning or validation. This is in contrast to many existing methods that use some subset of the OOD samples, either during validation [21, 36, 20, 28], or even during training [13, 42]. Despite improving the results, the availability of such extra information is questionable in many real-world applications. Furthermore, our technique can be easily deployed on many existing frameworks and different modalities, e.g. images, videos, etc. In summary, this paper makes the following contributions:

- We demonstrate that if feature vectors lie on a union of 1-dimensional subspaces, the OOD samples can be robustly detected with high probability and we show how we can impose such constraint on the ID feature vectors (Section 3);

- We propose a new OOD detection test, which exploits the first singular vector of the feature vectors extracted from the training set, in conjunction with MC sampling (Section 4);

- Our framework does not have hyperparameters, does not need extra information, and can be easily applied to existing methods with minimal change. Furthermore, the proposed method can be applied to different domains. Here, we introduce a new baseline for OOD detection for human action classification in videos.

2. Related Work

The problem of detecting outliers and anomalies in the data has been extensively studied in machine learning and signal processing communities and is closely related to outlier detection, a topic that has been greatly studied both in the supervised [9] and unsupervised [38] settings. The literature in this area is sizable. Thus, we mainly focus on the recent deep learning approaches. These methods either estimate the distribution of ID samples [20, 28] or use a distance metric between the test samples and ID samples to detect OOD samples [21, 12].

Many of the existing approaches employ the OOD datasets during training [42, 13] or validation steps [21, 36, 20, 28, 19, 30]. For instance, in [42], the network is fine-tuned during the training to increase the distance between ID and OOD distributions. Other interesting methods, such as [21, 36, 20], apply a perturbation on each sample at test time to exploit the robustness of their network in detecting ID samples. However, they use part of the OOD samples to fine-tune the perturbation parameters. On the other hand, methods that rely on generative models or autoencoders, such as [28], also require hyperparameter tuning for loss terms, regularization terms, and/or latent space size. Authors in [32] propose to use extra supervision, in particular several word embeddings, to construct a better latent space and to detect OOD samples more accurately. A table summarizing the prior work and how they leverage extra information is provided in the supplementary material. Having access to extra information certainly helps with the performance. However, it can be argued that OOD detectors should be completely agnostic of unknown distributions, which is a more realistic scenario in the wild. On the other hand, only a few approaches, such as [12, 27, 40, 24, 14], do not require the OOD samples neither during training nor validation. For instance, Hendricks and Gimpel [12] show how the softmax layer can be used to detect OOD samples, when its prediction score is below a threshold. In [40], the authors rely on reconstructing the samples to produce a discriminative feature space. However, methods that rely on either reconstruction or generation [27, 40, 28] do not perform well in scenarios where sample generation or reconstruction is more difficult, such as large-scale datasets or video classification. While the problem of detecting OOD samples in image classification has been subject of many studies, in the human action classification domain the focus has been on zero-shot and few shot learning [26]. To the best of our knowledge, this work is the first one to benchmark OOD results on two different modalities, i.e. image classification and human action recognition in videos.

3. Union of 1-dimensional Subspaces for Out-of-Distribution Detection

Given a training dataset consisting of N sample-label pairs belonging to L known classes, our goal is to train a neural network such that at the test time it can be determined if an unlabeled sample is an out-of-distribution sample (not belonging to any of the L known classes) or not. We are particularly interested in the scenarios where OOD samples are not available. Thus, we do not use OOD samples during training or validation. We argue that OOD detection performance can be improved if the feature vectors from the known classes lie on a *union of* 1*-dimensional subspaces*. In short, such embedding has two main properties that we can take advantage for OOD detection: (i) Due to the compactness of ID samples in the feature space, OOD samples can be detected with higher probability, compared to conventional class-conditional non-degenerate Gaussian embeddings, and (ii) First singular vector of the samples in each class can be used as a robust representative of that class and can be ef-

fectively employed to distinguish between the ID and OOD samples. Below, we discuss each of these advantages in more details.

Distribution-agnostic minimization of error probability: Computing the error probability for OOD detection is a difficult task to carry out. This is due to the fact that, by definition, we do not have much information about the probability distribution of the OOD samples. However, it can be shown that the probability of error can be minimized by making the distribution of the known classes as compact as possible. Specifically, consider the binary classification problem of distinguishing between the OOD samples and samples from one of the known classes, following multivariate Gaussian distributions with different means and covariance matrices $\mathcal{N}(\boldsymbol{\mu}_o, \boldsymbol{\Sigma}_o)$ and $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, respectively. It has been shown [7] that the classification error probability p_e can be upper bounded by: $p_e \leq \sqrt{p_i p_o} e^{-B}$, where p_i and p_o are the probability of samples belonging to the known class and OOD samples, respectively. B is the Bhattacharyya distance defined as:

$$B = \frac{1}{8}\boldsymbol{\Delta}^T(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_o}{2})^{-1}\boldsymbol{\Delta} + \frac{1}{2}\ln(\frac{\det(\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_o}{2})}{\sqrt{\det(\boldsymbol{\Sigma}_i)\det(\boldsymbol{\Sigma}_o)}}),$$

where $\boldsymbol{\Delta} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_o$ is the distance between the means of the two distributions. The first term in B represents the Mahalanobis distance between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_o$, using $\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_o}{2}$ as the covariance matrix. The second term is a measure of compactness of the distributions. The larger the $\det(\boldsymbol{\Sigma}_i)$ is, the more its corresponding samples are spread out. Thus, even without any knowledge about $\boldsymbol{\mu}_o$, $\boldsymbol{\Sigma}_o$, p_i, and p_o, one can increase B by making $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ as compact as possible. In the extreme case, where the samples lie on a perfect 1-dimensional subspace, error probability will be 0, unless the OOD feature vectors have the exact same distribution as the known class. To demonstrate this in further details, consider the following toy examples:

Example 1: Let $\boldsymbol{\Sigma}_o = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ and $\boldsymbol{\Sigma}_i = \begin{bmatrix}1 & 0\\ 0 & \epsilon^2\end{bmatrix}, \epsilon \ll 1$, meaning that the ID samples occupy an almost 1-dimensional subspace of the 2-dimensional space. In this example, the second term in above equation becomes $\ln(\frac{1+\epsilon^2}{2\epsilon})$, which approaches infinity as $\epsilon \to 0$, making p_e very small. This is true even if $\boldsymbol{\mu}_i = \boldsymbol{\mu}_o$.

Example 2: Let $\boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}_i = \begin{bmatrix}1 & 0\\ 0 & \epsilon^2\end{bmatrix}, \epsilon \ll 1, \boldsymbol{\mu}_i = \begin{bmatrix}\mu_{i1}\\ \mu_{i2}\end{bmatrix}, \boldsymbol{\mu}_o = \begin{bmatrix}\mu_{o1}\\ \mu_{o2}\end{bmatrix}$, i.e., ID and OOD samples have the same degenerate covariance matrix. In this case, the second term becomes 0, but the first term, which is the Mahalanobis distance between the mean vectors, is $\frac{1}{8}[(\mu_{i1}-\mu_{o1})^2 + \frac{1}{\epsilon^2}(\mu_{i2}-\mu_{o2})^2]$. If $\epsilon \to 0$, p_e approaches 0, unless $(\mu_{i2}-\mu_{o2})^2 \to 0$ as well. This means that if the means of the distribution have some mismatch along the degenerate direction, even though very small, OOD samples can be detected with very small p_e.

Thus, by enforcing the ID feature vectors to lie on 1-dimensional subspaces, we can detect slight mismatches between the distribution of the OOD samples in feature space and the distribution of ID samples, which leads to better OOD detection.

First singular vector as a robust representative: In the context of robust statistics, the first singular vector has been shown to be a great tool to define robust mean and covariance estimators [6]. In addition, the first singular vector has been used to select the representatives of the class[43]. It can be shown that the first singular vector is robust to perturbations and noise. Let $\boldsymbol{X}_l$ denote an $M \times N$ matrix containing N M-dimensional *feature vectors* belonging to class l. Furthermore, consider the autocorrelation matrix of the class l defined as $\boldsymbol{C}_l = \boldsymbol{X}_l\boldsymbol{X}_l^T$. Eigenvectors and eigenvalues of $\boldsymbol{C}_l$ are the left singular vectors and the square of singular values of $\boldsymbol{X}_l$, respectively. Adding noise or adding a new noisy column in $\boldsymbol{X}_l$ perturbs $\boldsymbol{C}_l$, without changing its dimensions. To quantify the sensitivity of eigenvectors of $\boldsymbol{C}_l$ against perturbations, we use the following Lemma.

Lemma 1 *(from [43]) Assume square matrix $\boldsymbol{C}$ and its spectrum $[\lambda_i, \boldsymbol{v}_i]$. Then, $\|\partial \boldsymbol{v}_i\|_2 \leq \sqrt{\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}}\|\partial \boldsymbol{C}\|_F$, where $\|.\|_F$ denotes Frobenius norm and the partial derivative is taken with respect to any scalar variable.*

If we take the partial derivative with respect to an entry in $\boldsymbol{C}$, we can see that the sensitivity of the i^{th} spectral component, $\boldsymbol{v}_i$, to perturbations in $\boldsymbol{C}$, is inversely related to the gap between its corresponding eigenvalue λ_i and other eigenvalues $\lambda_j, j \neq i$. Therefore, we can define the sensitivity coefficient of the i^{th} eigenvector of a square matrix as $s_i \triangleq \sqrt{\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}}$. In general, the first singular component $\boldsymbol{v}_1$ is the least sensitive direction to the perturbations. This is because, in many scenarios, the gap between consecutive eigenvalues is decreasing (see [3] and references therein), which leads to $s_1 < s_i, \ \ \forall i \geq 2$. However, we can further increase the robustness, by embedding the ID feature vectors onto a union of 1-dimensional subspaces. Since the singular values represent the amount of energy concentrated along their corresponding singular vector, if almost all of the energy of the data points in each class is concentrated along its corresponding first singular vector, we will have large λ_1 and small $\lambda_i, i \geq 2$ for all the classes. Therefore, if the feature vectors belonging to the same class lie on a 1-dimensional subspace, we can use the first singular vector of $\boldsymbol{X}_l$ as a robust representative of the class subspace in the feature space and to reject outliers, as shown in Section 4.

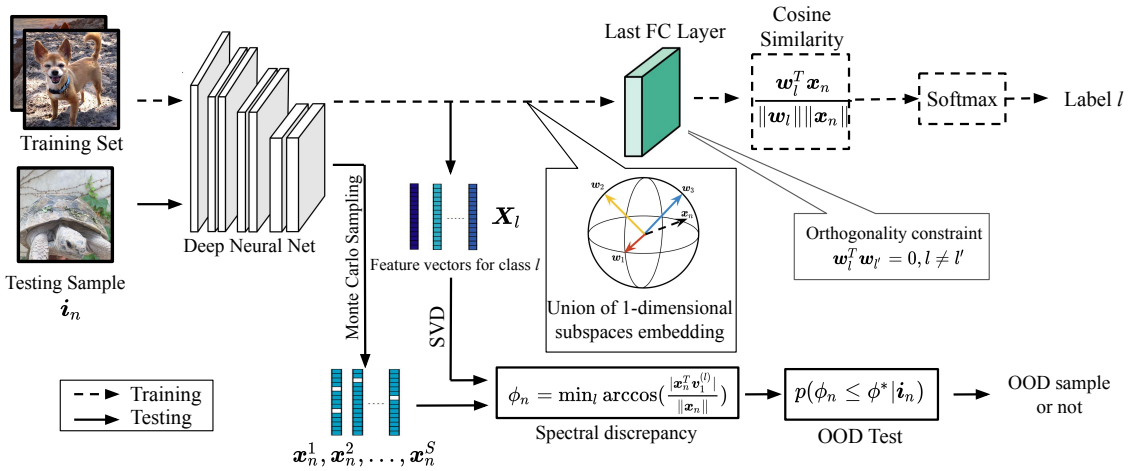


Figure 1. Overall architecture of the proposed framework. A neural network (e.g., WideResnet28) maps the input onto a feature space. Then, the cosine similarities between the extracted feature $\boldsymbol{x}_n$ and the class vectors $\boldsymbol{w}_l$ are used to compute the class membership probabilities. $\boldsymbol{w}_l$s are set to predefined orthonormal vectors and are not updated during training. This leads to the desired embedding, union of uncorrelated 1-dimensional subspaces. At test time, the cosine similarity between the test samples and the first singular vector corresponding to each class is used to distinguish between the ID and OOD samples.

3.1. Enforcing the Structural Constraints

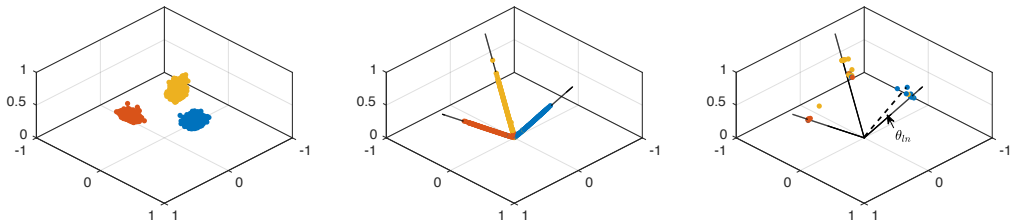
Intraclass Constraint: We can make the feature vectors for each known class to lie on a 1-dimensional subspace by employing cosine similarity. This can be achieved by modifying the softmax function to predict the membership probability using $p_{ln} = \frac{e^{|\cos(\theta_{ln})|}}{\sum_l e^{|\cos(\theta_{ln})|}}$, where p_{ln} is the probability of membership of feature vector n in class l and $\cos(\theta_{ln}) = \frac{\boldsymbol{w}_l^T \boldsymbol{x}_n}{\|\boldsymbol{w}_l\|\|\boldsymbol{x}_n\|}$ is the cosine similarity between the learned feature vector $\boldsymbol{x}_n$ and the weights of the last fully connected layer corresponding to class l, i.e.,$\boldsymbol{w}_l$. Note that, unlike other methods which employ angular margin [37, 23], we use the absolute value of the cosine similarity to compute the class memberships. This is due to the fact that the subspace membership, and therefore the class membership, does not change if a vector is multiplied by -1. By employing such activation function, the feature vectors of each class are aligned to its corresponding weight vector $\boldsymbol{w}_l$. In other words, class l forms a 1-dimensional subspace along the direction of $\boldsymbol{w}_l$ in the feature space. Therefore the final loss function to be minimized is defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} -\log\left(\frac{e^{|\cos(\theta_n^*)|}}{\sum_l e^{|\cos(\theta_{ln})|}}\right), \tag{1}$$

where θ_n^* is angle between the n^{th} feature vector and the weight vector corresponding to its true label.

Interclass Constraint: By using the absolute cosine similarity as the classification criteria, we can ensure the feature vectors are angularly distributed in the space and form a union of 1-dimensional subspaces. To boost the interclass separation of the known classes, we need to decrease the interclass similarity, in terms of cosine similarity. Minimum interclass cosine similarity can be enforced by ensuring that $\boldsymbol{w}_l$ are orthogonal to each other. We achieve this by simply initializing the weight matrix with orthonormal vectors, as described in [31], and freezing them during the training. Orthogonal initialization requires that $M > L$, which is often the case in practice (feature space dimension is larger than number of classes). In other words, the feature extractor, i.e., the deep neural network, is trained such that it can map each input sample in class l onto a predefined 1-dimensional subspace represented by the direction of $\boldsymbol{w}_l$.

Figure 1 shows the overall architecture of the proposed framework. The neural network maps the input sample onto a low-dimensional space, where the known classes are represented by a set of orthonormal vectors. The cosine similarity between the extracted feature from the n^{th} input sample, $\boldsymbol{x}_n$, and the vector corresponding to the class subspace, $\boldsymbol{w}_l$, is used to determine the class membership probability and therefore the label. Figure 2 demonstrates the effectiveness of the proposed framework in enforcing the desired embedding. It shows a 3-dimensional embedding, obtained by PCA, of the feature vectors belonging to the first 3 classes of CIFAR10. The neural network, WideResnet28, is trained on all the classes of CIFAR10 with and without enforcing the proposed structural constraints. Figure 2(a) shows that the feature vectors belonging to each class extracted from a plain WideResnet have a fairly isometric Gaussian structure, meaning that they are spread out in different direction uniformly. On the other hand, as shown in Figure 2(b), the feature vectors extracted from the same network trained us-



(a) (b) (c)

Figure 2. 3-dimensional representation of the features belonging to the first 3 classes of CIFAR10 training set, extracted from WideResNet with and without the proposed embedding: (a) features extracted from a plain WideResnet, (b) features extracted after enforcing the proposed embedding, and (c) same as (b) after ℓ_2-*normalizing* the feature vectors. The solid lines represent the direction of the first singular vector corresponding to each class. All the figures contain $3,000$ feature vectors.

ing our proposed technique lie on a union of 1-dimensional subspaces. We also show the ℓ_2-normalized feature vectors in Figure 2(c) to remove the scale of the feature vectors and emphasize the angle between each vector and the singular vector corresponding to its class.

4. Out-of-distribution Detection Test

If the feature vectors belonging to the known classes lie on a union of 1-dimensional subspaces, their corresponding region in the feature space has no volume. Thus, the probability of OOD samples being in the region corresponding to any of the known classes, which is the probability of false negative p_{fn}, is zero. This can be seen using the Bhattacharyya bound, discussed in Section 3, $p_e = p_o p_{fn} + p_i p_{fp} \leq \sqrt{p_i p_o} e^{-B}$. Therefore, if we make the known classes occupy a tiny region with no volume in the space, we will have $B \to \infty$ and $p_{fn} \to 0$. We use this property to classify samples as OOD if they lie inside the region corresponding to any of the known classes with probability 0. More specifically, given an input instance $\boldsymbol{i}_n$ and corresponding feature vector $\boldsymbol{x}_n$, this probability can be estimated using the singular vectors of each class as $p(\phi_n \leq \phi^* | \boldsymbol{i}_n)$, where ϕ_n is defined as:

$$\phi_n = \min_l \arccos\left(\frac{|\boldsymbol{x}_n^T \boldsymbol{v}_1^{(l)}|}{\|\boldsymbol{x}_n\|}\right), \tag{2}$$

which is the minimum angular distance of the test feature vector $\boldsymbol{x}_n$, from the first singular vector of any of the classes. We name this measure as *spectral discrepancy*. ϕ^* is a critical spectral discrepancy and defines the region belonging to the known classes. Smaller values of ϕ^* corresponds to more compact regions. In the extreme case of $\phi^* = 0$, the input instance $\boldsymbol{i}_n$ is detected as OOD, if it does not have the exact same direction as one of the singular vectors. It is worthwhile to mention that in the ideal case, the first singular vector of class l, $\boldsymbol{v}_1^{(l)}$, would be the same as $\boldsymbol{w}_l$. However, in practice, the first singular vector is a better representative of the subspace after training, as training feature vectors may not perfectly align with $\boldsymbol{w}_l$. $\boldsymbol{v}_1^{(l)}$ can be computed using the extracted features from training ID samples of class l. Time complexity order of computing the first singular vector is linear w.r.t both the number and the dimensions of the feature vectors [4, 1]. To estimate $p(\phi_n \leq \phi^* | \boldsymbol{i}_n)$, we employ Monte Carlo sampling. Specifically:

$$p(\phi_n \leq \phi^* | \boldsymbol{i}_n) = \int_0^{\phi^*} p(\phi_n | \boldsymbol{i}_n) d\phi_n \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}(\phi_n^s < \phi^*), \tag{3}$$

where S is the number of the Monte Carlo samples and ϕ_n^s is the spectral discrepancy of the s^{th} Monte Carlo sample, given input instance $\boldsymbol{i}_n$. Furthermore, $\mathbb{I}(.)$ is the indicator function that takes value 1 if $\phi_n^s < \phi^*$ and 0 otherwise. To obtain the samples, we can use the methods proposed for approximate Bayesian inference in [25, 8]. ϕ^* is the decision parameter, which can be set to achieve a problem-specific precision and/or recall requirements using different methods such as [22] or by using the training set (as will be discussed in Section 5).

Figure 3 demonstrates the effectiveness of employing spectral discrepancy in distinguishing between ID and OOD samples. Similar to Figure 2, this figure shows a 3-dimensional representation of the features that are close to the first 3 classes of the CIFAR10, meaning that the classifier would classify them as one of these classes. The first two subfigures show the features extracted from a plain WideResNet. Comparing ID samples (Figure 3(a)) with OOD samples (Figure 3(b)), it is clear that both ID and OOD samples follow a very similar structure, which makes OOD detection more difficult. On the other hand, the last two subfigures illustrate the ℓ_2-normalized features extracted from the WideResNet trained using our proposed embedding. Comparing the ID (Figure 3(c)) and OOD (Figure 3(d)) samples, most of the OOD samples have larger angular distance to their closest singular vector, compared to the ID samples, which can be exploited to detect them more accurately. A quantitative evaluation of this example, including the histogram of spectral discrepancies for ID and OOD samples, is provided in

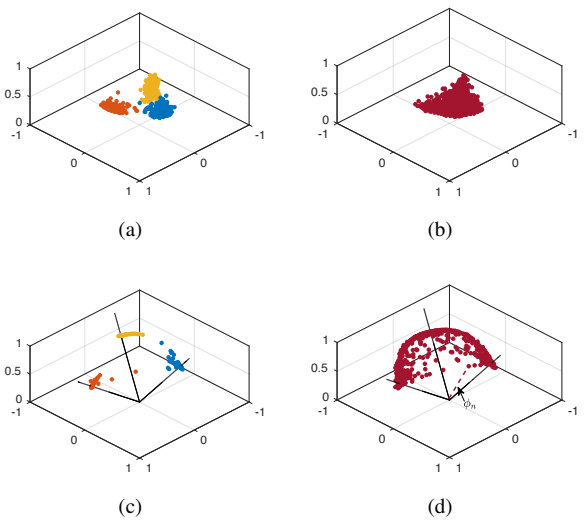


Figure 3. 3-dimensional representation of the features extracted from a plain WideResNet and the same network with our proposed embedding. (a) ID features extracted from plain network, (b) OOD features extracted from plain network, (c) ID features extracted using our embedding, and (d) OOD features extracted using our embedding. The solid lines represent the direction of the first singular vector corresponding to each class. OOD samples, extracted using our embedding, have larger angular distance to their closest singular vector. All the figures contain 3000 samples.

Section 5 (e.g., Figure 4). Furthermore, an algorithmic description of the training and testing phases of our proposed method is provided in the supplementary material.

5. Experiments

Datasets: For the *image classification* task, we train the WideResNet model on CIFAR-10 and CIFAR-100 [16] datasets, which consist of 50,000 images for training and 10,000 images for testing, with an image size of 32×32. The testing set is used as the ID testing samples. Similarly to prior work [20, 21, 24], for the OOD testing samples, we use the following datasets: (i) **TinyImagenet:** The Tiny ImageNet dataset consists of 10,000 test images of size 36×36 belonging to 200 different classes, which are sampled from the original 1,000 classes of ImageNet [5]. As in [21, 36] we construct two datasets from TinyImagenet: TinyImagenet-crop (TINc) and TinyImagenet-resize (TINr), by either randomly cropping or downsampling each image to a size of 32×32. (ii) **LSUN:** LSUN [41] consists of 10,000 test images from 10 different scene categories. Like before, we randomly crop and downsample the LSUN test set to construct two datasets LSUN-crop (LSUNc) and LSUN-resize (LSUNr).

For the *action classification* task, we train a 3DResNet model on **UCF101** [33] and **HMDB51** [17] datasets, which consist of 13320 videos with 101 classes and 6766 videos with 51 classes, respectively. As in previous works in zero-shot learning domain [26], we perform a random split the datasets between OOD classes and ID classes. UCF101 is divided in 50/51 ID/OOD classes, while UCF101 is divided in 25/26 ID/OOD classes.

Evaluation Metrics: We evaluate the OOD detection performance using the following metrics: **FPR at 95% TPR** indicates the false positive rate (FPR) at 95% true positive rate (TPR). **Detection Error** indicates the minimum misclassification probability. It is computed by the minimum misclassification rate over all possible values of ϕ^*. **AUROC**, defined as the Area Under the Receiver Operating Characteristic curve, is computed as the area under the FPR against TPR curve. **AUPR In** is computed as the area under the precision-recall curve. For AUPR In, ID images are treated as positive. **AUPR Out** is similar to the metric AUPR-In. Opposite to AUPR In, OOD images are treated as positive. **F1 Score** is the maximum average F1 score over all possible critical spectral discrepancy values ϕ^*.

We deploy WideResNet with depth 28 and width 10 as the neural network architecture for the image classification task and a 3DResNet [11] with 32 residual layers as the neural network for the action classification task. As in [26], our 3DResNet is initialized with weights pretrained on the Kinetics dataset [15]. Both network parameters are set as the original implementations in [44, 11], except the last layer, which is modified as discussed in Section 3. At the test time, unless otherwise stated, we draw 50 Monte Carlo samples to estimate $p(\phi_n \leq \phi^*)$ and to detect the OOD samples. To draw MC samples for the image classification task, we employ the SWAG-Diag method proposed in [25]. However, the storage and computation requirements of [25] makes it less practical for larger networks. Thus, for the video classification setup we employ the method in [8] to draw samples. Other uncertainty estimation methods such as [18, 34, 10] can also be used to estimate the uncertainty in conjunction with our proposed method. Additional training details are provided in the supplementary material[1].

Table 4 compares our results with recent OOD detection techniques in terms of F1-score. As denoted in the table, we use the code provided by the authors from most of the baselines to generate the results under a fair setting, i.e., same architecture, same datasets, and same metrics. For [27, 40], we provide the results reported by the authors, as these methods rely on reconstruction and/or generation of samples and the same architecture cannot be used. In addition, since these methods only report their performance using F1-score, we also use this metric for all the methods. Our proposed method is able to consistently outperform the

[1]Code for the image classification task is available at https://github.com/zaeemzadeh/OOD and the code for the action recognition task is available at https://github.com/mmlab-cv/OOD_video

Table 1. A comparison of OOD detection results, in terms of F1-score, for different ID and OOD datasets. † represents the results achieved by our re-run of the publicly available codes. The bottom section summarizes the performance of the methods that use a subset of OOD samples for hyperparameter tuning, such as finding the best perturbation magnitude. Our method does not have any parameters to be tuned.

ID dataset	CIFAR10				CIFAR100			
OOD dataset	TINc	TINr	LSUNc	LSUNr	TINc	TINr	LSUNc	LSUNr
SoftMax Pred. [12]†	0.803	0.807	0.794	0.815	0.683	0.683	0.664	0.693
Counterfactual [27]	0.636	0.635	0.650	0.648	-	-	-	-
CROSR [40]	0.733	0.763	0.714	0.731	-	-	-	-
OLTR [24]†	0.860	0.852	0.877	0.877	0.746	0.721	0.753	0.747
Ours	**0.930**	**0.936**	**0.962**	**0.961**	**0.810**	**0.860**	**0.769**	**0.886**
Methods that use OOD samples for validation and hyperparameter tuning.								
ODIN [21]†	0.902	0.926	0.894	0.937	0.834	0.863	0.828	0.875
Mahalanobis [20]†	0.985	0.969	0.985	0.975	0.974	0.944	0.963	0.952

Table 2. Performance of the proposed framework for distinguishing ID and OOD test set data for the image classification task, using a WideResnet with depth 28 and width 10. ↑ indicates larger value is better and ↓ indicates lower value is better. All the methods use the same network architecture.

Training dataset	**OOD dataset**	**FPR at 95% TPR** ↓	**Detection Error** ↓	**AUROC** ↑	**AUPR In** ↑	**AUPR Out** ↑
		Softmax. Pred. [12]/OLTR [24]/ Ours				
CIFAR10	TINc	38.9/25.6/**9.0**	21.9/14.8/**6.8**	92.9/91.3/**98.1**	92.5/93.2/**98.2**	91.9/88.3/**98.1**
	TINr	45.6/28.8/**7.6**	25.3/15.8/**6.2**	91.0/90.3/**98.5**	89.7/92.3/**98.6**	89.9/87.1/**98.4**
	LSUNc	35.0/21.3/**2.8**	20.0/13.0/**3.7**	94.5/92.9/**99.4**	95.1/94.4/**99.4**	93.1/90.8/**99.4**
	LSUNr	35.0/21.7/**3.4**	20.0/13.2/**3.8**	93.9/92.6/**99.3**	93.8/94.4/**99.4**	92.8/90.0/**99.3**
CIFAR100	TINc	66.6/63.8/**41.7**	35.8/29.0/**18.9**	82.0/77.4/**88.6**	83.3/78.7/**89.1**	80.2/74.4/**87.0**
	TINr	79.2/72.9/**29.42**	42.1/32.1/**14.2**	72.2/73.1/**93.7**	70.4/73.8/**94.0**	70.8/69.8/**93.8**
	LSUNc	74.0/59.2/**38,8**	39.5/29.1/**13.9**	80.3/76.9/**93.8**	83.4/80.0/**93.6**	77.0/72.9/**93.1**
	LSUNr	82.2/61.9/**20.3**	43.6/29.2/**11.3**	73.9/77.0/**95.7**	75.7/79.2/**96.0**	70.1/73.3/**95.7**

competing methods over different datasets, and is the closest competitor to the techniques that use OOD sample for validation. Table 2 compares the performance of our proposed solution with two of the more competitive baselines over different metrics, using the same network architecture for all the methods. Our results are consistent over different OOD datasets and different metrics, meaning that our method can perform well for different types of OOD samples, without any hyperparameter tuning for each OOD dataset.

In the ablation study, Table 4 investigates the impact of enforcing structure on the OOD detection using spectral discrepancy. AUROC is computed by using spectral discrepancy for the different variants. This table shows that, while enforcing the proposed embedding slightly hurts the ID classification accuracy and does not improve the representation ability of the network, it is an effective technique to distinguish between ID and OOD samples. This table also shows the effect of MC samples, which are used to compute the probabilities. As expected, introducing MC sampling improves the OOD detection performance, regardless of the feature space structure. However, the improvement is more significant for networks on which our proposed structure is enforced. Further, MC sampling alone or enforcing 1D subspace alone does not make a significant difference. But the combination of 1D subspaces and MC samples improves the results significantly. This is mainly because our method is a probabilistic approach and only works in a probabilistic setting.

In Table 3, we show our result for the action classification task. To best of our knowledge, we are the first to tackle the task of Out-Of-Distribution detection in the action recognition domain. To establish a baseline, we apply the Softmax threshold method as in [12] on the output of our network. We are able to consistently outperform the baseline, even if enforcing the structure hurts the results when is not combined with our OOD detector, which is consistent with the ablation study shown in Table 4. This illustrates the fact that our method can be easily applied to different network architectures and even different modalities, by only replacing the last fully connected layer of the network.

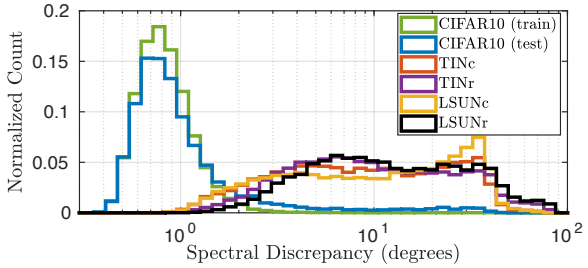
As a guideline to set the value of the critical spectral discrepancy ϕ^*, Figure 4(a) shows the histogram of the spectral discrepancy for samples belonging to CIFAR10, as the ID dataset, and different real OOD datasets. It is evident that samples from both the testing and training set of the ID dataset follow a very similar behaviour. Thus, the training set can be used to estimate the possible interval of spectral discrepancies for the ID samples. For instance, about 98% of the samples in CIFAR10 have a spectral discrepancy of less than 2 degrees. On the other hand, Figure 4(b) demonstrates the detection error for different values of the critical spectral discrepancy ϕ^*. This figure shows that best detection error is achieved by setting ϕ^* to a value in range $[1.3, 2]$ degrees, regardless of the OOD dataset. Hence, this figure shows that

Table 3. Performance of the proposed framework for distinguishing ID and OOD test set data for the action recognition task, using a 3DResNet [11] with 32 residual layers. ↑ indicates larger value is better and ↓ indicates lower value is better. All the methods use the same network architecture. As in [26], we use 50/51 splits of the UCF101 dataset and 25/26 splits of the HMDB51 dataset.

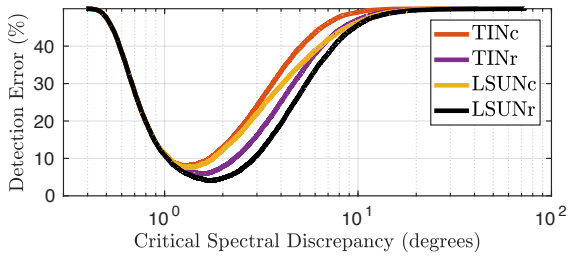
Training dataset	**OOD dataset**	**FPR at 95% TPR** ↓	**Detection Error** ↓	**AUROC** ↑	**AUPR In** ↑	**AUPR Out** ↑
		SoftMax. Pred. (Baseline) [12]/ SoftMax. Pred. (Orthogonal Subs.) [12]/ Ours				
UCF50	UCF51	86.3/82.44/**71.6**	36.8/36.1/**30.0**	66.0/68.3/**75.7**	89.8/**90.1**/74.3	25.6/27.8/**72.5**
HMDB25	HMDB26	**82.0**/85.2/84.5	41.8/44.5/**40.8**	59.7/56.4/**61.9**	**88.9**/87.6/65.4	20.4/19.7/**56.6**

Table 4. Ablation study of the proposed framework using CIFAR10 (ID) and TINr (OOD). While enforcing the structure hurts the ID accuracy slightly, it improves the OOD detection performance significantly. The remaining two combinations, (No, Yes, No) and (No, Yes, No), are not meaningful.

Union of 1D Subspaces	Orthogonal Subspaces	MC Samples	In Disribution Accuracy (%)	OOD AUROC
No	No	No	96.0	95.2
No	No	Yes	96.0	96.3
Yes	No	No	95.4	95.6
Yes	No	Yes	95.4	96.8
Yes	Yes	No	95.4	95.9
Yes	Yes	Yes	95.4	98.5



(a)



(b)

Figure 4. (a) Empirical probability distribution of the spectral discrepancy of samples belonging to CIFAR10 (ID) and different OOD datasets. (b) Detection error for different values of critical spectral discrepancy ϕ^*. Both the spectral discrepancy histogram and the best ϕ^* do not change significantly for different datasets.

ϕ^* is not sensitive to the OOD dataset and can be set using only the training set. However, it should be mentioned that in general the best value for ϕ^* depends on the task at hand and the precision and/or recall requirements. As mentioned earlier, ϕ^* can also be set by many of the threshold estimation techniques such as [22]. More experimental results such as quantifying the impact of the number MC samples, robustness of the first singular vector to perturbations, and ROC curves are provided in the supplementary material.

6. Conclusion

We show that the distribution of the ID samples in the feature space plays an important role in the OOD detection. Particularly, we propose to embed the ID samples into a low-dimensional feature space such that each known class lies on a 1-dimensional subspace. Such embedding gives us two main advantages in the OOD detection task: (i) ID samples occupy a tiny region in the space and (ii) ID samples have robust representatives. By exploiting these desirable features, our proposed method is able to outperform state-of-the-art methods in several performance metrics and different domains. We also establish a new baseline for OOD detection in the action classification in videos.

7. Acknowledgements

This research is based upon work supported in parts by the National Science Foundation (NSF) under Grants No. CCF-1718195 and No. ECCS-181256 and the Intelligence Advanced Research Projects Activity (IARPA), via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, IARPA, DOI/IBC, or the U.S. Government.

References

[1] James Baglama and Lothar Reichel. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, 2005. 5

[2] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. 1

[3] Daguang Chen, Tao Zheng, and Hongcang Yang. Estimates of the gaps between consecutive eigenvalues of Laplacian. *Pacific Journal of Mathematics*, 2016. 3

[4] P Comon and G H Golub. Tracking a few extreme singular values and vectors in signal processing. *Proceedings of the IEEE*, 78(8):1327–1343, 1990. 5

[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, 6 2009. 6

[6] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *34th International Conference on Machine Learning, ICML 2017*, 2017. 2, 3

[7] Moataz M.H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Toward a tight upper bound for the error probability of the binary Gaussian classification problem. *Pattern Recognition*, 2008. 3

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *PMLR*, 2016. 1, 5, 6

[9] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, 2018. 2

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *34th International Conference on Machine Learning, ICML 2017*, 2017. 6

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-Temporal features with 3D residual networks for action recognition. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2017. 6, 8

[12] 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. 1, 2, 7, 8

[13] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. 2

[14] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data. 2020. 1, 2

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. 5 2017. 6

[16] Alex Krizhevsky and G Hinton. Learning multiple layers of features from tiny images.(2009). Technical report, 2009. 6

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011. 6

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. 6

[19] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018. 2

[20] Kimin Kibok Lee, Kimin Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018. 1, 2, 6, 7

[21] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*, 2018. 1, 2, 6, 7

[22] Si Liu, Risheek Garrepalli, Thomas G. Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with PAC guarantees. In *35th International Conference on Machine Learning, ICML 2018*, 2018. 5, 8

[23] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. 4

[24] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-Scale Long-Tailed Recognition in an Open World. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2019. 2, 6, 7

[25] Wesley J. Maddox, Timur Garipov, Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, 2019. 1, 5, 6

[26] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. 2, 6, 8

[27] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11210 LNCS, pages 620–635, 2018. 2, 6, 7

[28] Stanislav Pidhorskyi, Ranya Almohsen, Donald A. Adjeroh, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in Neural Information Processing Systems*, 2018. 1, 2

[29] Kannan Ravindran, Salmasian Hadi, and Vempala Santosh. The spectral method for general mixture models. *SIAM Journal on Computing*, 2008. 2

[30] 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. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neu-*

ral Information Processing Systems 32, pages 14680–14691. Curran Associates, Inc., 2019. 2

[31] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 12 2014. 4

[32] Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. In *Advances in Neural Information Processing Systems*, 2018. 2

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. 12 2012. 6

[34] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019. 6

[35] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, 2018. 2

[36] 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 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018. 1, 2, 6

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. 4

[38] Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft. Effective End-to-end Unsupervised Outlier Detection via Inlier Priority of Discriminative Network. *Nips*, 2019. 2

[39] Raanan Yehezkel Rohekar, Yaniv Gurwicz, Shami Nisimov, and Gal Novik. Modeling Uncertainty by Learning a Hierarchy of Deep Neural Connections. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4246–4256. Curran Associates, Inc., 2019. 1

[40] Ryota Yoshihashi, Shaodi You, Wen Shao, Makoto Iida, Rei Kawakami, and Takeshi Naemura. Classification-Reconstruction Learning for Open-Set Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 6, 7

[41] 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, 2015. 6

[42] Qing Yu and Kiyoharu Aizawa. Unsupervised Out-of-Distribution Detection by Maximum Classifier Discrepancy. In *The IEEE International Conference on Computer Vision (ICCV)*, 10 2019. 2

[43] Alireza Zaeemzadeh, Mohsen Joneidi, Nazanin Rahnavard, and Mubarak Shah. Iterative Projection and Matching: Finding Structure-preserving Representatives and Its Application to Computer Vision. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[44] Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *British Machine Vision Conference 2016, BMVC 2016*, 2016. 6