Revisit PCA-based technique for Out-of-Distribution Detection

Xiaoyuan Guan[1,3*] Zhouwu Liu[1,3*] Wei-Shi Zheng[1,3] Yuren Zhou[1†]
Ruixuan Wang[1,2,3†]
[1]School of Computer Science and Engineering, Sun Yat-sen Univerisity, Guangzhou, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China
{guanxy36,liuzhw33}@mail2.sysu.edu.cn, wszheng@ieee.org, {zhouyuren,wangruix5}@mail.sysu.edu.cn

Abstract

Out-of-distribution (OOD) detection is a desired ability to ensure the reliability and safety of intelligent systems. A scoring function is often designed to measure the degree of any new data being an OOD sample. While most designed scoring functions are based on a single source of information (e.g., the classifier's output, logits, or feature vector), recent studies demonstrate that fusion of multiple sources may help better detect OOD data. In this study, after detailed analysis of the issue in OOD detection by the conventional principal component analysis (PCA), we propose fusing a simple regularized PCA-based reconstruction error with other source of scoring function to further improve OOD detection performance. In particular, when combined with a strong energy score-based OOD method, the regularized reconstruction error helps achieve new state-of-the-art OOD detection results on multiple standard benchmarks. The code is available at https://github.com/SYSU-MIA-GROUP/pca-based-out-of-distribution-detection.

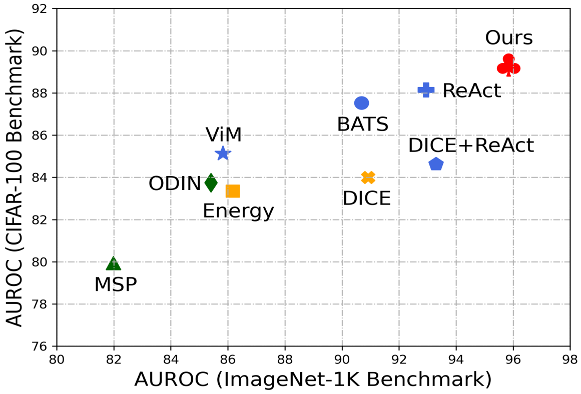

Figure 1. The OOD detection performance (AUROC) on two benchmarks from our method and representative strong baselines. The ImageNet-1K Benchmark includes the ID set ImageNet-1K and four OOD sets iNaturalist, SUN, Places, and Textures. The CIFAR-100 Benchmark includes the ID set CIFAR-100 and seven OOD sets SVHN, Tiny-ImageNet, iSUN, LSUN-Crop, LSUN-Resize, Textures and Places365.

1. Introduction

Out-of-distribution (OOD) detection refers to the ability of a model to correctly detect samples that are from a different distribution compared to that of the in-distribution (ID) training data [13, 35]. It is a crucial ability to ensure the reliability and safety of machine learning systems. For instance, in autonomous driving, the system should be able to detect previously unseen scenarios or objects and alert the human operator when it cannot make a safe decision. However, current intelligent systems are often limited in OOD detection.

Multiple approaches have been proposed to improve model's OOD detection ability [4, 9, 21, 26, 34]. The main objective is to design or derive a scoring function whose output is different between OOD input and ID input. For deep learning classifier models, the scoring function is often designed based on three sources of information, i.e., the output probability of the classifier [7, 8, 15, 38], the logit (i.e., input to the softmax function) at the last classifier layer [6, 10, 16, 31], and the feature vector output from the feature extractor part of the classifier [14, 18, 22, 29]. However, one source of information is often inherently limited. For example, softmax output and logit often discard some of the information in the feature vector even the discarded information might be helpful for OOD detection, while scoring function solely based the feature vector cannot utilize the class-associated weight parameters at the last classifier layer. Based on this observation, one recent study starts to explore combination of multiple sources for further improving OOD detection performance [29, 40]. Specifically,

[*] Authors contributed equally
[†] Corresponding author

certain discarded information from the feature vector is reused to generate an extra logit, which is then input to the softmax together with the original logits to obtain the final OOD score. Such multi-source fusion has achieved state-of-the-art performance on some benchmarks [29].

In this study, following the multi-source fusion idea, we propose a simple yet effective OOD score based on in-depth analysis of the effect of the conventional principal component analysis (PCA) on OOD detection. Starting from the observation that the PCA-based reconstruction error for feature vectors is often inseparable between ID and OOD data, we provide a detailed theoretical analysis and find that statistically the larger variance in components of feature vectors on the ID data is probably the main source of such inseparability. Based on this analysis, a regularized PCA-based reconstruction error is applied to improve the separability between ID and OOD data. Such regularized reconstruction error is post-hoc (i.e., working on any well-trained classifier) and can be flexibly combined with existing OOD detection methods such that multiple sources of information can be utilized together to better detect OOD data. When fusing this regularized error with the energy score-based OOD method, new state-of-the-art performance was obtained on multiple standard benchmarks.

2. Preliminaries

2.1. Out-of-distribution detection

For a supervised classification task, denote by D_{in} the training set of C classes which are randomly sampled from the in-distribution $\mathcal{D}_{\text{in}}$ of the C classes. The goal of OOD detection is to train a classifier G (e.g., a convolutional neural network) which can not only correctly classify any test data sampled from the in-distribution $\mathcal{D}_{\text{in}}$, but also detect whether a test data is from the in-distribution $\mathcal{D}_{\text{in}}$ or from a different and unknown distribution $\mathcal{D}_{\text{out}}$. A decision function F built on the classifier G needs to be designed to help the classifier detect any potential data $\mathbf{x}$ from the out-of-distribution $\mathcal{D}_{\text{out}}$, and ideally

$$F(\mathbf{x}; G) = \begin{cases} 1 & \text{if} \quad \mathbf{x} \sim \mathcal{D}_{\text{in}} \\ 0 & \text{if} \quad \mathbf{x} \sim \mathcal{D}_{\text{out}} \end{cases} . \tag{1}$$

2.2. OOD detection with energy score

Suppose the classifier G is a convolutional neural network which consists of multiple convolutional layers and a fully connected layer, with softmax output as the final classification prediction. Given a test sample $\mathbf{x}$, denote by $f_c(\mathbf{x})$ the c-th output of the final linear layer, i.e., the c-th logit corresponding to class c. One type of decision function F is based on the energy score E,

$$E(\mathbf{x}; G) = -\log \sum_{c=1}^{C} \exp(f_c(\mathbf{x})) , \tag{2}$$

In particular, Liu et al. [16] proposed the negative energy score $-E(\mathbf{x}; G)$ as the score measurement to detect OOD examples, with a higher score $-E(\mathbf{x}; G)$ suggesting that $\mathbf{x}$ is more likely from the in-distribution $\mathcal{D}_{\text{in}}$ and lower score suggesting that $\mathbf{x}$ is more likely from the out-of-distribution $\mathcal{D}_{\text{out}}$.

3. Method

In this section, we first analyze the reason why the reconstruction error based on conventional PCA is not helpful for OOD detection (Section 3.1), and then use a simple normalized PCA-based feature reconstruction error which can be combined with existing OOD scores (mainly energy score) to further improve the separability between ID and OOD data.

3.1. Issues in PCA-based OOD detection

Suppose a neural network classifier has been well trained with ID data. Intuitively, when applying PCA to the feature vector output from the well-trained feature extractor, one would expect that the reconstruction error is smaller for ID data while larger for OOD data, and thus the PCA-based reconstruction error would help separate ID from OOD data. However, it is observed that such expectation is not true statistically. In the following, we will reformulate the PCA-based reconstruction error and analyze why it is not working in separating ID from OOD data.

Denote by $h(\mathbf{x}) \in \mathbb{R}^K$ the feature vector from the penultimate layer of the well-trained neural network classifier for any input data $\mathbf{x}$. The covariance matrix $\mathbf{\Sigma}$ of feature vectors $h(\mathbf{x})$'s over all the ID training data can be obtained and then decomposed to $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$, where $\mathbf{U} \in \mathbb{R}^{K \times K}$ is a unitary matrix and constituted by the principal component directions, with each principal component corresponding to one eigenvector of the covariance matrix $\mathbf{\Sigma}$, and $\mathbf{\Lambda} \in \mathbb{R}^{K \times K}$ is the diagonal matrix with the diagonal entries corresponding to the eigenvalues of $\mathbf{\Sigma}$ in descending order. Since ID data often lie in a subspace of the original feature vector space, the projection of ID data into the subspace and then reconstruction to the original feature space would lose little information compared to the original feature vector for any ID data. Suppose the subspace is k-dimensional and its base is from the first k principal components which constitutes the first k columns of $\mathbf{U}$, denoted by $\mathbf{U}_k \in \mathbb{R}^{K \times k}$. k can be automatically determined as for traditional dimensionality reduction by PCA, e.g., choosing the k such that the first k eigenvalues can account for 95% of the variance in distribution of feature vectors for all the ID training data. Denote by $\boldsymbol{\mu}$ the mean feature vector over all ID training data in the original feature vector space, and $\mathbf{M} = \mathbf{U}_k\mathbf{U}_k^\intercal \in \mathbb{R}^{K \times K}$. Then any original feature vector $h(\mathbf{x})$ can be projected to the subspace and then projected

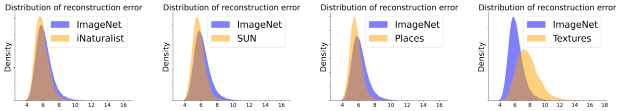

Figure 2. Distributions of the reconstruction error on the ID dataset (blue) and each of the four OOD datasets (orange). The dimension of the original feature space and the subspace is respectively 2048 and 256.

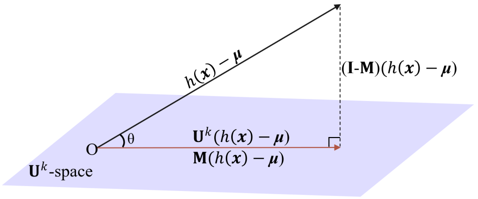
$h(\boldsymbol{x})-\boldsymbol{\mu}$
$(\mathbf{I}\text{-}\mathbf{M})(h(\boldsymbol{x})-\boldsymbol{\mu})$
O
θ
$\mathbf{U}^k(h(\boldsymbol{x})-\boldsymbol{\mu})$
$\mathbf{M}(h(\boldsymbol{x})-\boldsymbol{\mu})$
$\mathbf{U}^k$-space

Figure 3. The geometry relationship between the centralized feature vector $h(\boldsymbol{x})-\boldsymbol{\mu}$ and the reconstructed centralized feature vector $\mathbf{M}(h(\boldsymbol{x})-\boldsymbol{\mu})$. The reconstructed error is the magnitude of the difference between the two vectors.

back to the original feature space as below,

$$\hat{h}(\mathbf{x}) = \mathbf{M}(h(\mathbf{x})-\boldsymbol{\mu})+\boldsymbol{\mu}\,, \tag{3}$$

and the reconstruction error $e(\mathbf{x})$ can be directly obtained as

$$\begin{aligned} e(\mathbf{x}) &= \|h(\mathbf{x})-\hat{h}(\mathbf{x})\| && (4)\\ &= \|(\mathbf{I}-\mathbf{M})(h(\mathbf{x})-\boldsymbol{\mu})\|. && (5)\end{aligned}$$

Although it is expected that $e(\mathbf{x})$ would be smaller for ID data and larger for OOD data, surprisingly we observed that the distribution of the reconstruction error on the ID dataset is often not separable from the distribution of the reconstruction error on the OOD dataset (Figure 2). To explore the underlying cause to such inseparability between ID and OOD data, we reformulate the reconstruction error formula (Eq. 5) based on the geometry relationship between the centralized feature vector $h(\boldsymbol{x})-\boldsymbol{\mu}$ and the reconstructed centralized feature vector $\mathbf{M}(h(\boldsymbol{x})-\boldsymbol{\mu})$. As Figure 3 illustrates, the reconstructed centralized feature vector $\mathbf{M}(h(\boldsymbol{x})-\boldsymbol{\mu})$ coincide with the projection of the centralized feature vector $h(\boldsymbol{x})-\boldsymbol{\mu}$ to the subspace ($\mathbf{U}^k$-space) spanned by the first k principal components (see proof in Supplementary Section B), and the reconstruction error $e(\mathbf{x})$ is actually the L_2 norm of the difference between $h(\boldsymbol{x})-\boldsymbol{\mu}$ and $\mathbf{M}(h(\boldsymbol{x})-\boldsymbol{\mu})$. Denote by $\theta(\mathbf{x})$ the angle between the two vectors $h(\boldsymbol{x})-\boldsymbol{\mu}$ and $\mathbf{M}(h(\boldsymbol{x})-\boldsymbol{\mu})$, then the reconstruction error can be reformulated as

$$\begin{aligned} e(\mathbf{x}) &= \|h(\boldsymbol{x})-\boldsymbol{\mu}\|\cdot\frac{\|(\mathbf{I}-\mathbf{M})(h(\mathbf{x})-\boldsymbol{\mu})\|}{\|h(\boldsymbol{x})-\boldsymbol{\mu}\|} && (6)\\ &= \|h(\boldsymbol{x})-\boldsymbol{\mu}\|\cdot\sin\theta\,. && (7)\end{aligned}$$

Since relatively large variations in ID signals are already in the subspace, the variations in ID signals along any direction orthogonal to the subspace would be relatively small. In other words, the angle θ is expected to be relatively small for ID data. In contrast, OOD signals are not utilized to determine the first k principal components and therefore some of large variations in OOD signals may not be in the subspace, and the angle θ may be relatively large at least for some OOD data. This has been confirmed on all the four OOD datasets as shown in Figure 4.

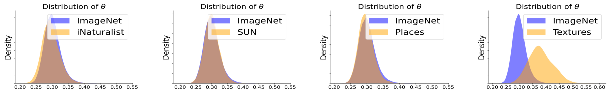

Figure 4. Distributions of the angle θ between the centralized feature vector $h(\boldsymbol{x})-\boldsymbol{\mu}$ and its reconstruction $\mathbf{M}(h(\boldsymbol{x})-\boldsymbol{\mu})$ on the ID dataset ImageNet-1K (blue) and on the OOD dataset iNaturalist (first subfigure, orange), SUN (second, orange), Places (third, orange), and Textures (fourth, orange), respectively. Model with backbone ResNet50 is trained on the ID dataset ImageNet-1K here and below by default.

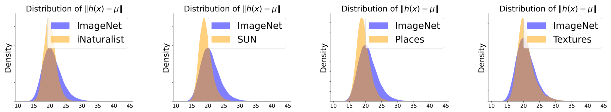

Figure 5. Distributions of the magnitude $\|h(\boldsymbol{x})-\boldsymbol{\mu}\|$ of the centralized feature vector on the ID dataset (blue) and each of the four OOD datasets (orange).

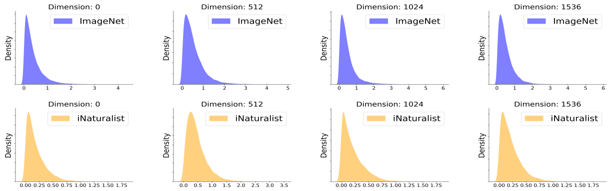

Figure 6. Positively skewed distribution of each z_i on both ID dataset ImageNet-1K (first row) and an OOD dataset iNaturalist (second row). Four z_i's are uniformly sampled for demonstration. Model with backbone ResNet50 is trained on the ID dataset ImageNet-1K.

From Eq. (7) and the observations in Figures 2 and 4, it can be inferred that the magnitude $\|h(\boldsymbol{x})-\boldsymbol{\mu}\|$ of the centralized feature vector $h(\boldsymbol{x})-\boldsymbol{\mu}$ should be statistically larger for ID data while smaller for OOD data. Again, this is confirmed by the empirical observation on the four OOD datasets as demonstrated in Figure 5. In the following, further exploration is performed to search for the source of such difference in magnitude $\|h(\boldsymbol{x})-\boldsymbol{\mu}\|$ between ID and OOD data.

3.1.1 Variance analysis

For ease of analysis, let $z_i(\mathbf{x})$ denote the i-th component of $h(\mathbf{x})$, $\mu_{\text{in},i}$ denote the i-th component of $\boldsymbol{\mu}$, and define $v(\mathbf{x}) = \|h(\boldsymbol{x})-\boldsymbol{\mu}\|^2$. Then the expectation of $v(\mathbf{x})$ on the ID dataset $\mathcal{D}_{\text{in}}$ and an OOD dataset $\mathcal{D}_{\text{out}}$ can be respectively calculated by

$$\mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{D}_{\text{in}}}[v(\mathbf{x})] = \sum_{i=1}^{K} \mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{D}_{\text{in}}}[(z_i(\mathbf{x}) - \mu_{\text{in},i})^2] \tag{8}$$

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{D}_{\text{out}}}[v(\mathbf{x})] &= \sum_{i=1}^{K} \mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{D}_{\text{out}}}[(z_i(\mathbf{x}) - \mu_{\text{in},i})^2] \qquad (9)\\ &= \sum_{i=1}^{K}\Big\{ \mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{D}_{\text{out}}}[(z_i(\mathbf{x}) - \mu_{\text{out},i})^2] + \Delta\mu_i \Big\}\,,\end{aligned}$$

where $\mu_{\text{out},i}$ is the average of the i-th component of feature vector output over all the OOD data in $\mathcal{D}_{\text{out}}$, and $\Delta\mu_i = (\mu_{\text{out},i} - \mu_{\text{in},i})^2$.

If the distribution of each z_i (i.e., $z_i(\mathbf{x})$, omitting $\mathbf{x}$ for simplicity below) is normal, the variance of z_i and consequently the expectation of $v(\mathbf{x})$ on either ID or OOD dataset can be unbiasedly estimated based on the collection of ID or OOD data. However, largely due to the ReLU activation operator (only keeping non-negative signals) which is adopted in most CNN models, the distribution of each z_i is often positively skewed on both ID and OOD dataset, as demonstrated in Figure 6. In this case, the initially estimated variance of each z_i needs to be modified by considering the non-symmetrical property of each z_i's distribution. Following the previous studies [17, 22], the epsilon-skew-norm (ESN) distribution is used to approximate the positively skewed distribution of each z_i. Formally, suppose $\sigma_{\text{out},i}$ is the initially estimated standard deviation of z_i on the collection of OOD data, and $m_{\text{out},i}$ is the estimated mode of z_i over all the OOD data, then the positively skewed distribution of the z_i on the OOD dataset can be modelled by a ESN distribution as below,

$$q(z_i) = \begin{cases} \frac{1}{\sigma_{\text{out, i}}}\phi(\frac{z_i - m_{\text{out},i}}{\sigma_{\text{out},i}(1+\epsilon_{\text{out},i})}) & \text{if } z_i < m_{\text{out},i} \\ \frac{1}{\sigma_{\text{out, i}}}\phi(\frac{z_i - m_{\text{out},i}}{\sigma_{\text{out},i}(1-\epsilon_{\text{out},i})}) & \text{if } z_i \geq m_{\text{out},i} \end{cases}, \tag{10}$$

where $\phi(\cdot)$ represents the p.d.f of standard normal distribution, and the hyper-parameter $\epsilon_{\text{out},i} \in (-1, 1)$ controls the skewness. $q(z_i)$ will become the well-known half-normal distributions when $\epsilon_{\text{out},i} \to \pm 1$, and reduce to the normal distribution when $\epsilon_{\text{out},i} = 0$. In particular, $\epsilon_{\text{out},i} < 0$ for positively-skewed ESN distribution. In this case, the relationship between initial estimated variance of z_i and the variance can be formulated as [17]

$$\text{Var}(z_{\text{out,i}}) = \frac{\sigma^2_{\text{out,i}}}{\pi}[(3\pi - 8)\epsilon^2_{\text{out,i}} + \pi]. \tag{11}$$

The upper bound of the variance for each z_i can be obtained by setting $\epsilon_{\text{out,i}} = -1$. Similarly on the ID dataset, the variance of z_i can be estimated as well, with the parameters in Eqs. (10) and (11) estimated based on the ID dataset, and the lower bound of the variance for each z_i can be obtained by setting $\epsilon_{\text{in,i}} = 0$. After replacing $\mathbb{E}_{\mathbf{x}\in\mathcal{D}_{\text{out}}}[(z_i - \mu_{\text{out},i})^2]$ (Eq. 9) by the upper bound of $\text{Var}(z_i)$ on the OOD dataset, and replacing $\mathbb{E}_{\mathbf{x}\in\mathcal{D}_{\text{in}}}[(z_i(\mathbf{x}) - \mu_{\text{in},i})^2]$ (Eq. 8) by the lower bound of $\text{Var}(z_i)$ on the ID dataset, the upper bound of $\mathbb{E}_{\mathbf{x}\in\mathcal{D}_{\text{out}}}[v(\mathbf{x})]$ (Eq. 9) and the the lower bound of $\mathbb{E}_{\mathbf{x}\in\mathcal{D}_{\text{in}}}[v(\mathbf{x})]$ (Eq. 8) can be directly obtained. Table 1 (last row) shows that even the lower bound of $\mathbb{E}_{\mathbf{x}\in\mathcal{D}_{\text{in}}}[v(\mathbf{x})]$ on the ID dataset (first column) is clearly larger than the upper bound of $\mathbb{E}_{\mathbf{x}\in\mathcal{D}_{\text{out}}}[v(\mathbf{x})]$ on each of the four OOD datasets (last four columns).

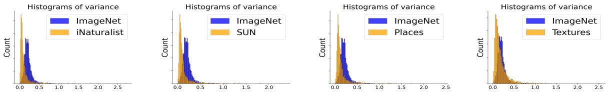

Figure 7. Histograms of the initial variance estimates of feature vector components respectively on the ID dataset ImageNet-1k (blue) and on each of the four OOD dataset iNaturalist, SUN, Places, and Textures (orange).

This supports the above finding that the magnitude $\|h(\boldsymbol{x}) - \boldsymbol{\mu}\|$ of the centralized feature vector $h(\boldsymbol{x}) - \boldsymbol{\mu}$ is statistically larger for ID data than for OOD data. Together with Eq. (11), it also indicates that the initial estimates $\{\sigma^2_{\text{in},i}\}_{i=1}^{K}$ on the ID dataset should be statistically larger than the initial estimates $\{\sigma^2_{\text{out},i}\}_{i=1}^{K}$ on the OOD dataset. This is confirmed by the histogram of the initial variance estimates respectively on the ID dataset and on each OOD dataset (Figure 7).

In summary, we find that statistically the larger variance in components (z_i's) of the feature vector $h(\mathbf{x})$ on the ID data, together with the smaller angle between the original feature vector of ID data and its projection in the subspace, is probably the source of the inseparability in PCA-based reconstruction error between ID and OOD data. One may further wonder why feature vector components have relatively smaller variances on OOD data, which is beyond the scope of this study and can be investigated as a future work.

Table 1. The lower bound of Eq. 8 on ID data (1st column) and the upper bound of Eq. 9 on OOD data (last 4 columns).

Dataset	ImageNet1K	iNaturalist	SUN	Places	Textures
ϵ	0	-1	-1	-1	-1
$\mathbb{E}[v(\boldsymbol{x})]$	393.22	370.03	335.06	333.76	388.19

3.2. Regularized reconstruction error

While it is still unclear why feature vector components have relatively larger variance on the ID data, it is well-known that a variable whose values are at a larger scale would more likely have a relatively larger variance. If that is the case, we may use the relatively larger norm of feature vector to alleviate the negative effect of the larger variance from ID data. Fortunately, previous studies [28] have shown that statistically the norm of feature vector from ID data is larger than that from OOD data (also see the re-implemented result in Figure 8). With all the above con-

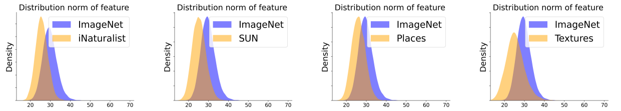

Figure 8. Distributions of the norm of feature vector on the ID dataset ImageNet-1K (blue) and each of the four OOD datasets iNaturalist, Plcaces, SUN, and Textures (orange).

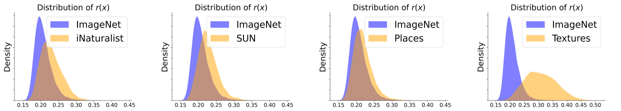

Figure 9. Distributions of the regularized reconstruction error on the ID dataset (blue) and each of the four OOD datasets (orange).

sideration, we choose to use a simple regularized version of the original reconstruction error $e(\mathbf{x})$ (Eq. 5) as below

$$r(\mathbf{x}) = \|h(\mathbf{x}) - \hat{h}(\mathbf{x})\| / \|h(\mathbf{x})\|. \tag{12}$$

With the regularized reconstruction error $r(\mathbf{x})$, statistically smaller reconstruction error is observed on the ID dataset compared to that on each OOD dataset (Figure 9). Although the separability in the regularized reconstruction error between ID and OOD data is still weak, the regularized reconstruction error $r(\mathbf{x})$ can be combined with existing OOD scores to potentially further improve the OOD detection performance. For example, the fusion of $r(\mathbf{x})$ with the energy score results in a new OOD score as below

$$D(\mathbf{x}) = (1 - r(\mathbf{x})) \log \sum_{c=1}^{C} \exp(f_c(\mathbf{x})) \,. \tag{13}$$

The multiplication rather than the addition operator is used for the fusion of two OOD scores to improve the influence of the regularized reconstruction error $r(\mathbf{x})$ and meanwhile avoid using the extra trade-off coefficient between the two scores. It is expected that ID data would result in a higher score $D(\mathbf{x})$ and OOD data lower score. Also, it is worth noting that the proposed $r(\mathbf{x})$ is compatible with multiple existing OOD methods, i.e., it can be fused with not only energy-based OOD scores but also others like MSP [7] for OOD detection (see Table 3 below).

4. Experiments

4.1. Experimental settings

Our method was extensively evaluated on four OOD detection benchmarks. Each benchmark contains a training ID set, a test ID set, and multiple test OOD sets. The ImageNet-1K and ImageNet-100 [25] Benchmarks respectively use the ImageNet-1K and ImageNet-100 as ID sets, and both use four OOD sets, including iNaturalist [27], SUN [32], Places [39], and Textures [2]. The CIFAR-10 and CIFAR-100 Benchmarks respectively use the CIFAR-10 and CIFAR-100 as ID sets, and both use seven OOD sets, including SVHN [19], Tiny-ImageNet [1], iSUN [33], LSUN-Crop [36], LSUN-Resize [36], Textures, and Places365 [39]. There are no overlapped classes between ID set and each OOD set. Please refer to Supplementary Section A for more details of datasets.

On the ImageNet-1K Benchmark, a publicly released pre-trained CNN classifier with certain backbone (ResNet50 [5] or MobileNet-v2 [20]) was directly used for evaluation of our method and each baseline method. On the other three benchmarks, a CNN classifier with certain backbone was trained from scratch with the associated training ID set. The stochastic gradient descent optimizer with momentum (0.9) and weight decay (0.0005) was used to train each classifier up to 100 epochs on the ImageNet-100 training set or up to 200 epochs on the CIFAR [12] training datasets. The batch size was set to 128. The initial learning was set to 0.1, and decayed by a factor of 10 at the 50-th, 75-th, and 90-th epoch on ImageNet-100, or at the 100-th and 150-th epoch on CIFAR. On the ImageNet-100 training set, each image was resized to 256 × 256 pixels and then randomly cropped to 224 × 224 pixels. On the CIFAR-10 or CIFAR-100 training set, each training image was padded from 32 × 32 pixels to 36 × 36 pixels and then randomly cropped to 32 × 32 pixels. Random horizontal flipping was adopted together with random cropping on each training image. During test, only center cropping together resizing was used on the ImageNet [3] dataset.

Following previous studies [22], ResNet50 and the light-weight MobileNet-v2 were used as the classifier backbone on the ImageNet Benchmarks, ResNet34 and WideResNet28-10 [31, 40] were used on CIFAR Benchmarks. The feature dimension of the penultimate layer is respectively 2048, 1280, 512, and 640 for ResNet50, MobileNet-v2, ResNet34, and WideResNet28-10. In all experiments, the reduced feature dimension was empirically set to 256, 128, 9 (on CIFAR10) or 49 (on CIFAR100), and 16 (on CIFAR10) or 64 (on CIFAR100) respectively for the four backbones, with the constraint that around 90% of the original feature signal was preserved after dimension reduction on the ID training sets.

Multiple types of competitive OOD detection methods were adoped as baselines for comprehensive evaluation, including the Maximum Softmax Probability (MSP) [7], ODIN [15], Energy [16], Mahalanobis [14], ViM [29], DICE [23], ReAct [22] and BATS [40]. All the baselines are post-hoc and can obtain the OOD score based on a pre-trained CNN classifier. In addition, LogitNorm [31] was used on Benchmark III because it achieves the state-of-the-art performance on the benchmark. For all experiments, FPR95 (i.e., the false positive rate of OOD examples when the true positive rate of ID examples is 95%) and AUROC

Table 2. Comparison between different methods in OOD detection on the ImageNet-1K Benchmark with two different model backbones. ↑ indicates that larger values are better and ↓ indicates that smaller values are better. All values are percentages.

ID Dataset Model	Method	OOD Datasets									
		iNaturalist		SUN		Places		Textures		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ImageNet-1K ResNet50	MSP	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
	ODIN	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
	Mahalanobis	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47
	Energy	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
	ViM	68.86	87.13	79.62	81.67	83.81	77.80	**14.95**	**96.74**	61.81	85.83
	BATS	42.26	92.75	44.70	90.22	55.85	86.48	33.24	93.33	44.01	90.69
	DICE	26.66	94.49	36.08	90.98	47.63	87.73	32.46	90.46	35.71	90.92
	ReAct	20.38	96.22	24.20	94.20	33.85	91.58	47.30	89.80	31.43	92.95
	DICE+ReAct	20.08	96.11	26.50	93.83	38.34	90.61	29.36	92.65	28.57	93.30
	Ours	**10.17**	**97.97**	**18.50**	**95.80**	**27.31**	**93.39**	18.67	95.95	**18.66**	**95.78**
ImageNet-1K MobileNet	MSP	64.29	85.32	77.02	77.10	79.23	76.27	73.51	77.30	73.51	79.00
	ODIN	58.54	87.51	57.00	85.83	59.87	84.77	52.07	85.04	56.87	85.79
	Mahalanobis	62.11	81.00	47.82	83.66	**52.09**	83.63	92.38	33.06	63.60	71.01
	Energy	59.50	88.91	62.65	84.50	69.37	81.19	58.05	85.03	62.39	84.91
	ViM	91.83	77.47	94.34	70.24	93.97	68.26	37.62	92.65	79.44	77.15
	BATS	49.57	91.50	57.81	85.96	64.48	82.83	39.77	91.17	52.91	87.87
	DICE	43.28	90.79	**38.86**	90.41	53.48	85.67	33.14	91.26	42.19	89.53
	ReAct	43.07	92.72	52.47	87.26	59.91	84.07	40.20	90.96	48.91	88.75
	DICE+ReAct	41.75	89.84	39.07	90.39	54.41	84.03	19.98	**95.86**	38.80	90.03
	Ours	**35.84**	**93.66**	40.35	**90.77**	52.38	**86.76**	**18.44**	95.39	**36.75**	**91.65**

(the area under the receiver operating characteristic curve) in OOD detection were used as metric, with lower FPR95 values and higher AUROC values indicating better OOD detection performance.

4.2. Quantitative evaluations

On the ImageNet-1K Benchmark, Table 2 summarizes the OOD detection performance of each method on each OOD set (together with the ID test set) and the average performance over the four OOD sets. By default, our method is built on the state-of-the-art baseline ReAct which uses the energy score for OOD detection. With the ResNet-50 backbone, our method achieves state-of-the-art performance on three of the four OOD sets, and in average outperforms the best baseline DICE+ReAct by a large margin (95.78% vs. 93.30% on AUROC, and 18.66% vs. 28.57% on FPR95). Similar finding can be obtained with the MobileNet-v2 backbone (Table 2, lower half), achieving the best AUROC performance on three OOD sets and the state-of-the-art average performance over the four OOD sets.

Our method can be flexibly combined with existing baselines to further improve their OOD detection performance. As Table 3 shows, when fusing the proposed regularized reconstruction error term $(1-r)$ into the OOD scores in existing methods MSP, Energy, BATS, and ReAct, the OOD detection performance is boosted on each OOD set. With the ResNet50 backbone, the average AUROC is improved by $1.87\% - 2.83\%$, and the average FPR95 is successfully reduced by $4.24\% - 12.77\%$. Similar result can be observed with the MobileNet-v2 backbone (Table 3, lower half).

Consistently on the ImageNet-100 Benchmark, our method achieves the best performance in average (Table 4, last two columns). On individual OOD sets, our method performs either best or similar to the the best baseline. Note that for each method, the OOD performance on the ImageNet-100 Benchmark is slightly worse than that on ImageNet-1K Benchmark probably because the classifier trained on ImageNet-100 is more confident about its predictions, which causes more confident false positive predictions (i.e., OOD samples considered as ID classes) and correspondingly lower AUROC values.

On the CIFAR Benchmarks, the regularized reconstruction error also helps achieve state-of-the-art performance in average on the seven OOD sets with both CNN backbones, as shown in Table 5. Note that on the CIFAR-10 Benchmark, the regularized reconstruction error is combined with the state-of-the-art method LogitNorm by default. In addition, Table 6 consistently confirms that the regularized reconstruction error can be flexibly combined with multiple OOD methods to further improve their original detection performance.

4.3. Sensitivity study

Here we evaluate the sensitivity of our method to the dimension k of the subspace which is crucial to compute the regularized reconstruction error. Smaller k statistically corresponds to less percent of the original feature signal preserved after dimension reduction on the ID training set. As shown in Figure 10, our method works quite stable when varying the value of k in a large range on both benchmarks. Although the OOD detection performance slowly decreases when k gradually becomes smaller, our method still works

Table 3. Fusion of the regularized reconstruction error with various OOD methods on the ImageNet-1K Benchmark. For each paired values by '/': the left one is from the original baseline and the right one is from the fusion one.

ID Dataset Model	Method	OOD Datasets									
		iNaturalist		SUN		Places		Textures		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ImageNet-1K ResNet50	MSP	54.99/**51.47**	87.74/**88.95**	70.83/**67.64**	80.86/**82.71**	73.99/**71.20**	79.76/ **80.87**	68.00/**60.53**	79.61/**85.86**	66.95/**62.71**	81.99/**84.60**
	Energy	55.72/**50.36**	89.95/**91.09**	59.26/**54.19**	85.89/**87.55**	64.92/**64.13**	82.86/**84.00**	53.72/**29.33**	85.99/**92.59**	58.41/**49.50**	86.17/**88.81**
	BATS	42.26/**29.66**	92.75/**94.49**	44.70/**38.11**	**91.40**/90.03	55.85/**51.70**	86.48/**87.25**	33.24/**13.46**	93.33/**97.09**	44.01/**33.23**	90.69/**92.56**
	ReAct	20.38/**10.17**	96.22/**97.97**	24.20/**18.50**	94.20/**95.80**	33.85/**27.31**	91.58/**93.39**	47.30/**18.67**	89.80/**95.95**	31.43/**18.66**	92.95/**95.78**
ImageNet-1K MobileNet	MSP	64.29/**59.49**	85.32/**86.87**	77.02/**73.75**	77.10/**79.41**	79.23/**76.79**	76.27/ **77.94**	73.51/**65.71**	77.30/**83.46**	73.51/**68.93**	79.00/**81.92**
	Energy	59.50/**56.92**	88.91/**89.62**	62.65/**60.07**	84.50/**85.80**	69.37/**69.23**	81.19/**81.72**	58.05/**34.22**	85.03/**91.66**	62.39/**55.11**	84.91/**87.20**
	BATS	**49.57**/50.51	**91.50**/90.86	57.81/**55.41**	85.96/**87.00**	64.48/**66.43**	**82.83**/82.60	39.77/**23.26**	91.17/**94.70**	52.91/**48.90**	87.87/**88.79**
	ReAct	43.07/**35.84**	92.72/**93.66**	52.47/**40.35**	87.26/**90.77**	59.91/**52.38**	84.07/**86.76**	40.20/**18.44**	90.96/**95.39**	48.91/**36.75**	88.75/**91.65**

Table 4. Comparison between different methods in OOD detection on the ImageNet-100 Benchmark.

ID Dataset Model	Method	OOD Datasets									
		iNaturalist		SUN		Places		Textures		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ImageNet-100 ResNet50	MSP	69.28	85.84	70.14	84.20	69.43	84.29	64.27	84.09	68.28	84.60
	ODIN	44.22	92.42	54.71	88.94	57.52	88.01	42.87	89.76	49.83	89.78
	Mahalanobis	98.09	35.06	98.75	34.70	98.66	34.95	44.08	83.07	84.98	44.08
	Energy	64.60	89.08	62.70	88.03	60.70	87.78	51.38	87.89	59.85	88.20
	ViM	84.92	81.92	83.18	81.47	81.45	81.55	**20.00**	**96.07**	67.39	85.25
	BATS	43.05	92.65	58.75	87.83	57.72	87.43	38.88	91.97	49.60	89.97
	DICE	35.08	93.29	**36.89**	**92.53**	43.71	90.66	31.84	92.08	36.46	92.11
	ReAct	30.60	94.40	47.55	89.99	47.21	89.53	50.89	87.57	44.06	90.45
	DICE+ReAct	29.20	93.83	47.01	89.28	**38.40**	**92.01**	31.84	92.08	36.61	91.80
	Ours	**26.60**	**94.83**	41.55	90.96	44.50	90.22	22.18	95.07	**33.71**	**92.77**

well even just 70% of the original signals is preserved in the feature subspace.

5. Related work

OOD score design. One line of studies perform OOD detection by designing scoring functions. Based on the information sources used in designing scoring functions, existing methods can be divided into three main categories, i.e., (1) probability space based, such as maximum softmax probability (MSP) [7] and ODIN [15]; (2) logit space based, such as MaxLogit [6], energy-scores [16, 30]; and (3) feature space based, such as Mahalanobis distance [14]. Very recently, ReAct [22] was proposed to improve the effect of the energy score by feature pruning, and similarly BATS [40] boosts the energy score by limiting features to a so-called typical space. Both of them utilize the feature space information to enhance the performance of logit space-based scoring function for OOD detection, yet neither of them adopt multiple sources information. ViM [29] is the most relevant work to our study. It maps features to the orthogonal complement space of the principal space and computes a virtual logit based on the projection of this space. ViM uses fusion of information from both features and logits and achieves state-of-the-art OOD detection performance on some benchmark datasets. In constrast, our method fuses a regularized PCA-based reconstruction error with other source of scoring functions to improve OOD detection performance. Our method leverages the differences in the distribution of information in the feature space between ID and OOD data, as well as differences in feature norms, and achieves stable improvement on OOD detection performance when combined with different scoring functions across multiple sources of information.

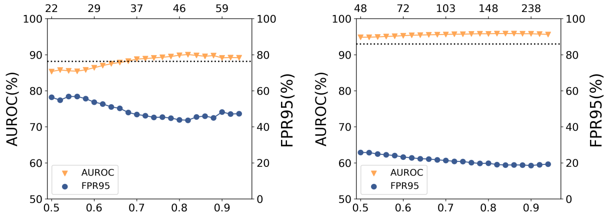

Figure 10. Effect of the dimension k of the subspace on OOD detection performance. Left: on the CIFAR-100 Benchmark with the ResNet34 backbone. Right: on the ImageNet-1K Benchamrk with the ResNet50 backbone. For each subfigure, the bottom horizontal axis represents the percent of the original feature signal preserved after dimension reduction, while the top horizontal axis represents the corresponding dimension k of the subspace. The horizontal dotted line in each subfigure represents the AUROC performance of best baseline on each Benchmark.

Training regularization. Another line of studies perform OOD detection by adding certain regularization term into the loss function for model training [24, 11, 37]. For example, G-ODIN [35] uses a dividend/divisor structure to measure ID-ness of inputs. VOS [4] produces better predictions by utilizing synthetic virtual outliers. LogitNorm [31] normalizes logit's before the logits are sent to the CE loss function, which results an angular dispersed feature representation and achieve state-of-the-art performance on CIFAR-10

Table 5. Comparison between different methods in OOD detection on the CIFAR benchmarks. Values are average percentages over seven OOD datasets. See more detailed results on each OOD dataset in the Supplementary Section C.

ID Dataset	Metrics	Methods										
Model		MSP	Mahalanobis	ODIN	DICE	VIM	Energy	BATS	ReAct	DICE+ReAct	LogitNorm	Ours
CIFAR-10	FPR95↓	39.69	60.09	30.37	33.51	40.62	24.71	28.62	25.90	33.74	15.03	**13.06**
ResNet34	AUROC↑	92.54	91.38	90.79	91.65	93.56	93.85	93.98	93.90	91.65	97.25	**97.59**
CIFAR-10	FPR95↓	37.45	88.37	28.98	32.56	21.51	25.75	28.66	26.26	32.82	15.36	**13.32**
WideResNet	AUROC↑	92.64	62.93	90.23	89.16	95.16	93.17	94.30	93.56	90.01	96.63	**97.14**
CIFAR-100	FPR95↓	77.26	94.43	63.91	67.32	63.03	68.13	56.88	51.11	64.81	63.41	**43.64**
ResNet34	AUROC↑	79.93	54.43	83.74	83.99	85.12	83.34	87.52	88.14	84.62	81.90	**90.04**
CIFAR-100	FPR95↓	74.42	66.99	64.29	60.48	58.43	68.55	67.28	65.02	67.89	51.23	**51.79**
WideResNet	AUROC↑	80.50	68.31	83.80	83.90	87.37	82.88	85.27	86.05	74.13	87.22	**88.94**

Table 6. Fusion of the regularized reconstruction error with various OOD methods on the CIFAR Benchmark

Method	ResNet34				WideResNet			
	CIFAR10		CIFAR100		CIFAR10		CIFAR100	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP	39.69/**36.08**	92.54/**92.95**	77.26/**74.48**	**79.93**/79.71	37.45/**32.08**	92.64/**94.12**	74.42/**72.89**	80.50/**81.62**
Energy	**24.71**/25.19	93.85/**94.88**	68.13/**57.21**	83.34/**86.30**	25.75/**24.82**	93.17/**94.40**	68.55/**63.00**	82.88/**85.00**
BATS	**28.62**/29.45	93.98/**94.23**	56.88/**45.56**	87.52/**89.48**	28.66/**28.42**	94.30/**94.69**	67.28/**52.11**	85.27/**88.78**
ReAct	**25.90**/27.49	93.90/**94.60**	54.11/**43.64**	88.14/**90.04**	26.26/**25.29**	93.56/**94.49**	65.02/**51.79**	86.05/**88.94**

benchmarks. As a post-hoc strategy, our method does not require model retraining, and the derived regularized reconstruction error can serve as a plugin and be conveniently combined with existing OOD methods with different model architectures.

6. Conclusion

In this study, we analyze in detail why the conventional PCA-based reconstruction error is not working well for OOD detection in deep learning models, and propose fusing a regularized PCA-based reconstruction error with existing scoring functions to further improve OOD detection performance. A simple combination of the regularized reconstruction error with the energy-based scoring function achieves state-of-the-art performance on multiple benchmark datasets. This study can be considered as one more evidence for the fusion of multiple information sources in the scoring function to better detect OOD data.

Acknowledgement: This work is supported in part by the Major Key Project of PCL (grant No. PCL2023AS7-1), the National Natural Science Foundation of China (grant No. 62071502), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

References

[1] P. Chrabaszcz, I. Loshchilov, and F. Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. In *CVPR*, 2009.

[4] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: learning what you don't know by virtual outlier synthesis. In *ICLR*, 2022.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[8] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

[9] 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 *CVPR*, 2020.

[10] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021.

[11] Rui Huang and Yixuan Li. Mos: towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021.

[12] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

[13] Johnson Kuan and Jonas Mueller. Back to the basics: revisiting out-of-distribution detection baselines. In *ICML*, 2022.

[14] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

[15] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[16] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

[17] Govind S Mudholkar and Alan D Hutson. The epsilon–skew–normal distribution for analyzing near-normal data. *Journal of statistical planning and inference*, 83(2), 2000.

[18] Ibrahima Ndiour, Nilesh Ahuja, and Omesh Tickoo. Out-of-distribution detection with subspace techniques and probabilistic modeling of features. *arXiv preprint arXiv:2012.04250*, 2020.

[19] Y. Netzer, T. Wang, A. Coates, A. Bissacco, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS*, 2011.

[20] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[21] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: rank-1 feature removal for out-of-distribution detection. In *NeurIPS*, 2022.

[22] Y. Sun, C. Guo, and Y. Li. React: out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

[23] Yiyou Sun and Yixuan Li. Dice: leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022.

[24] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020.

[25] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *ICCV*, 2023.

[26] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

[27] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

[28] S Vaze, K Han, A Vedaldi, and A Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *ICLR*, 2022.

[29] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: out-of-distribution with virtual-logit matching. In *CVPR*, 2022.

[30] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In *NeurIPS*, 2021.

[31] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022.

[32] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[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] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *CVPR*, 2021.

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

[36] 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 e-prints*, 2015.

[37] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *ICCV*, 2019.

[38] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: regularization strategy to train strong classifiers with localizable features. In *CVPR*, 2019.

[39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 40(6):1452–1464, 2017.

[40] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, Yaowu Chen, et al. Boosting out-of-distribution detection with typical features. In *NeurIPS*, 2022.