# CORES: Convolutional Response-based Score for Out-of-distribution Detection

Keke Tang[1*], Chao Hou[1*], Weilong Peng[1†], Runnan Chen[2],
Peican Zhu[3], Wenping Wang[4], Zhihong Tian[1]

[1]Guangzhou University [2]University of Hong Kong
[3]Northwestern Polytechnical University [4]Texas A&M University

tangbohutbh@gmail.com, houchaohk@gmail.com, wlpeng@gzhu.edu.cn, crnsmile@connect.hku.hk,
ericcan@nwpu.edu.cn, wenping@tamu.edu, tianzhihong@gzhu.edu.cn

## Abstract

*Deep neural networks (DNNs) often display overconfidence when encountering out-of-distribution (OOD) samples, posing significant challenges in real-world applications. Capitalizing on the observation that responses on convolutional kernels are generally more pronounced for in-distribution (ID) samples than for OOD ones, this paper proposes the COnvolutional REsponse-based Score (CORES) to exploit these discrepancies for OOD detection. Initially, CORES delves into the extremities of convolutional responses by considering both their magnitude and the frequency of significant values. Moreover, through backtracking from the most prominent predictions, CORES effectively pinpoints sample-relevant kernels across different layers. These kernels, which exhibit a strong correlation to input samples, are integral to CORES's OOD detection capability. Comprehensive experiments across various ID and OOD settings demonstrate CORES's effectiveness in OOD detection and its superiority to the state-of-the-art methods.*

## 1. Introduction

Deep neural networks (DNNs) [17, 36, 37] have achieved significant success in various tasks. Nevertheless, DNNs often suffer performance drop when exposed to novel, unforeseen categories of data, termed as out-of-distribution (OOD) samples, owing to their initial design for static, closed-world scenarios [1]. Indeed, many recent studies have demonstrated that DNN classifiers suffer from the issue of making overconfident predictions on OOD samples [6, 26, 35]. Hence, OOD detection is crucial for ensuring the safety and reliability of DNN deployments [10, 39].

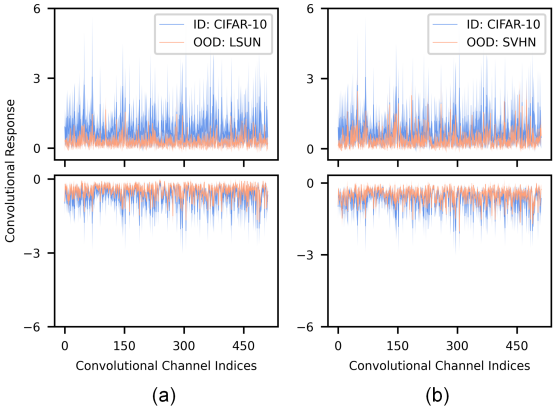


Figure 1. Response visualization of convolutional kernels in the penultimate layer of ResNet-18. Responses to in-distribution (ID) CIFAR-10 data are contrasted with OOD data from: (a) LSUN; and (b) SVHN. The $x$-axis enumerates the convolutional channel indices, and the $y$-axis shows the maximum positive and minimum negative values of the responses. It is evident that ID displays higher positive and lower negative responses than OOD.

A prevalent direction for OOD detection involves enhancing classifiers through exposure to auxiliary OOD samples [11, 18]. However, these approaches may lead to increased training costs and potentially compromise model performance. As a result, post-hoc methods that examine the inference process offer advantages in real-world applications. Existing post-hoc methods typically fall into two main categories: 1) those utilizing in-distribution (ID) samples, such as training data, and 2) those that do not rely on ID samples. Specifically, the first category aims to measure the difference between a test image and training samples [19, 27, 30, 34, 41]. In many real-world scenarios, access to training data is limited, and even when available, it may not comprehensively represent the entire ID domain, potentially causing inaccuracies in calculated discrepancies.

*Joint first authors. †Corresponding author.

The second ID-free category concentrates on the model's intrinsic characteristics rather than depending on training data, demonstrating potential for effective OOD detection.

In the realm of ID-free methodologies for OOD detection, researchers have investigated a plethora of signals as potential indicators. These include the maximum probabilities derived from softmax layers [10], its subsequent enhancements [21, 22], post-processed outputted features [32, 33], the norm of the gradient [14] and the norm of features [46]. However, these signals, which were primarily designed for ID categorization, may not be ideally suited for a totally different task, i.e., distinguishing between ID and OOD samples. Therefore, we aim to identify signals inherent within the DNNs' functioning that can assist in determining whether samples are ID or OOD, without leaning on category-specific data. Given that convolutional kernels in DNNs, generally exhibit more pronounced responses to ID samples, see Fig. 1, these responses are anticipated to be reliable signals for distinguishing OOD from ID.

In this paper, we present **CO**nvolutional **RE**sponse-based **S**core (**CORES**), a unique ID-free method that exploits responses of convolutional kernels for OOD detection. CORES harnesses the most pronounced responses of convolutional kernels in a layer, specifically the peak and trough values, to mitigate the impact of noise. We derive the OOD detection score from these extreme values, considering both their magnitude and frequency. Given that kernels yielding the strongest (either highest or lowest) confidence results are inherently relevant to the input samples, we trace back from these outputs to pinpoint a trajectory of sample-relevant kernels across layers. By collecting responses from these specific kernels, we establish the final CORES metric. We demonstrate the effectiveness of CORES across multiple DNN architectures, under diverse settings of ID and OOD datasets, encompassing both conventional OOD and near-OOD scenarios. Extensive experiments substantiate the robustness of CORES in OOD detection, and its superiority to the state-of-the-art methods.

Overall, our contribution is summarized as follows:

- We are the first to explore convolutional responses as ID-free signals for OOD detection.
- We develop a novel OOD score based on extreme response values on a trajectory of sample-relevant kernels.
- We show by experiments that CORES outperforms the state-of-the-art OOD detection methods.

## 2. Related Work

**OOD Detection.** The OOD overconfidence issue, where OOD samples are assigned excessively high confidence scores, poses a significant challenge for numerous DNNs [8, 9, 38, 44]. To mitigate it, a spectrum of strategies has been employed during DNN training, e.g., incorporating random noise and image shuffling [18], auxiliary datasets [9], and advanced generative models like Chamfer GAN [35] to simulate OOD samples resembling ID data closely. Alternatively, methods such as VOS [5] and NPOS [40] generate virtual features rather than actual OOD instances. These approaches, however, confront obstacles such as the need for extensive OOD data and the potential for intensive fine-tuning to detrimentally affect performance.

Consequently, the focus has shifted to post-hoc OOD scoring techniques that differentiate model responses to OOD and ID data through analysis of feature discrepancies [24], Mahalanobis distance computation [19], and detection of anomalies in patterns and predictions [30]. Despite Sun et al.'s development of a purportedly assumption-free method, the necessity for some ID samples persists, presenting challenges when only pre-trained models are available without corresponding training data or when such data inadequately represents the ID space [34]. Our research, therefore, prioritizes advancing post-hoc OOD detection methods that operate independently of ID data access.

**Post-hoc ID-free OOD Detection.** Hendrycks et al. [10] pioneered the use of maximum softmax probability, observing that classifiers typically exhibit lower confidence for OOD samples than for ID ones. Expanding on this, ODIN [21] integrates input preprocessing and temperature scaling to sharpen the distinction between ID and OOD samples. Energy Score [22] employs an energy-based model to reduce prediction bias. Further advancements include ReAct [33], LHAct [47], and DICE [32], which refine outputted features through activation rectification or selective sparsification for improved discrimination. Recognizing the notable gradient intensity in ID data, GradientNorm [14] utilizes the vector norm of gradients for OOD detection. Similarly, FeatureNorm [46] leverages the norm of feature maps from a strategically chosen block for identifying OOD samples.

However, these signals, primarily tailored for ID categorization, may falter when redirected to other tasks, such as differentiating between ID and OOD samples. Differently, our chosen signal, specifically the convolutional response, is not crafted for categorization purposes. It focuses exclusively on expressing the inherent characteristics of the samples, resulting in enhanced performance in OOD detection.

**Convolutional Response.** Convolutional responses, the output feature maps from each convolutional kernel on input data such as images, illuminate key attributes including edges, textures, and colors. Analyzing these responses deepens understanding of the features learned by deep convolutional networks [23, 49] and supports model refinement through the identification and pruning of ineffective kernels, thus conserving computational resources [20]. Our research presents a novel OOD detection approach based on the observation that convolutional kernels exhibit more pronounced responses to ID data compared to OOD data.

## 3. Problem Formulation

**Preliminaries.** This work considers the setting in multi-category classification problems. Given an input image $x$, the DNN classifier $f$, which consists of $L$ convolutional layers, processes it as follows:

$$f : x \to \mathcal{R}^{(1)} \to \mathcal{R}^{(2)} \to \ldots \to \mathcal{R}^{(L)} \to \mathcal{O}$$

Here, $\mathcal{R}^{(i)}$ represents the convolutional responses at the $i$-th layer, and $\mathcal{O}$ denotes the final output.

By identifying which category in $\mathcal{O}$ has the highest prediction confidence,

$$c_{\max} = \arg\max_i \mathcal{O}_i$$

$f$ predicts $x$ to be in the $c_{\max}$-th category. However, in some open-world scenarios, when the input $x$ is out-of-distribution (OOD), $f$ still classifies it into the predefined categories. As a result, it is essential to conduct OOD detection in practical applications.

A prevalent strategy for OOD detection entails the construction of a score, i.e., $\text{Score}_{OOD}$, subsequently utilizing thresholding to discriminate between in-distribution (ID) and OOD samples [10],

$$g_\gamma(\mathbf{x}) = \begin{cases} \text{ID} & \text{Score}_{OOD} \geq \gamma \\ \text{OOD} & \text{Score}_{OOD} < \gamma \end{cases} \tag{1}$$

where $\gamma$ is a threshold.

**Discussion.** Existing methodologies for $\text{Score}_{OOD}$ largely hinge on the derived probability or logit outputs, such as maximum softmax probabilities (MSP) [10], or the Energy Score [22], as well as on particular feature characteristics, as seen in FeatureNorm [46]. However, these indicators, crafted for categorizing ID classes, may falter when applied to a fundamentally different task, namely, distinguishing between ID and OOD samples. Consequently, our goal is to derive signals from the DNNs' inference phase, that are not principally used for categorization but rather dedicated to capturing the inherent properties of the samples, irrespective of their classification as ID or OOD.

**Response-based OOD Detection.** We observe that the convolutional kernels in a trained deep neural network are inherently attuned to extracting the fundamental attributes of samples. These kernels exhibit a strong response to patterns they recognize — those consistent with ID inputs. In contrast, their response diminishes with patterns they do not recognize, characteristic of OOD inputs. Thus, we suggest that the significant difference in the responses, represented as $\{\mathcal{R}^{(i)}\}_{i=1:L}$, is a strong indicator for determining if inputs are ID or OOD,

$$\{\mathcal{R}^{(i)}\}_{i=1:L} \to \text{Score}_{OOD} \tag{2}$$

## 4. Convolutional Response-based Score

In this section, we first detail the convolutional response-based score (CORES), derived from the responses of a single, specified layer. We then elaborate on the method for pinpointing a trajectory of sample-relevant kernels in different layers. Finally, we combine these elements to unveil the final version of CORES.

### 4.1. Single-layer CORES

Given $c$ channels of convolutional kernels in the $l$-th layer of a trained deep neural network, we calculate $\text{Score}_{OOD}$ utilizing the responses from these kernels, i.e., $\mathcal{R}^{(l)} = \{R_1^{(l)}, R_2^{(l)}, \ldots, R_c^{(l)}\}$. For clarity, we will omit the superscript $(l)$ unless specific distinction is required.

Since DNNs are trained on ID data, their responses $\mathcal{R}^{(l)}$ align better with ID than OOD inputs. This difference in response strength helps identify ID and OOD inputs. Recognizing the intrinsic noise in the responses, we emphasize that extreme values in a kernel often convey more pertinent information. Thus, we exclusively concentrate on the highest positive value and the lowest negative value for each kernel. Specifically, we design $\text{Score}_{OOD}$ to capture both the *magnitude* of these extreme values and their *frequency* within the entire response spectrum.

**Response Magnitude.** Convolutional kernels are generally more responsive to ID inputs than to OOD ones, producing stronger responses. To capture this, we introduce a metric focusing on the maximum positive and minimum negative values of the response:

$$\begin{aligned} RM^+(\mathcal{R}) &= \frac{1}{|\mathcal{R}|}\sum_{i=1}^{|\mathcal{R}|}\max(\max(R_i) - \tau_+, 0) \\ RM^-(\mathcal{R}) &= \frac{1}{|\mathcal{R}|}\sum_{i=1}^{|\mathcal{R}|}\max(\tau_- - \min(R_i), 0) \end{aligned} \tag{3}$$

Here, $\tau_+$ and $\tau_-$ are thresholds that help decide if the response strength is significant.

**Frequency of Significant Responses.** Convolutional kernels typically produce more distinct responses to ID inputs than to OOD ones. To capture this, we introduce a metric that calculates the frequency of these significant responses:

$$\begin{aligned} RF^+(\mathcal{R}) &= \frac{1}{|\mathcal{R}|}\mathbb{1}(\max(R_i) > \tau_+) \\ RF^-(\mathcal{R}) &= \frac{1}{|\mathcal{R}|}\mathbb{1}(\min(R_i) < \tau_-) \end{aligned} \tag{4}$$

**Definition of Single-layer CORES.** We integrate the above metrics to define a straightforward OOD score:

$$S(\mathcal{R}) = RM^+(\mathcal{R})^{\lambda_1} \times RM^-(\mathcal{R})^{\lambda_1} \times RF^+(\mathcal{R})^{\lambda_2} \times RF^-(\mathcal{R})^{\lambda_2} \tag{5}$$

where $\lambda_1$ and $\lambda_2$ are weighting hyperparameters.

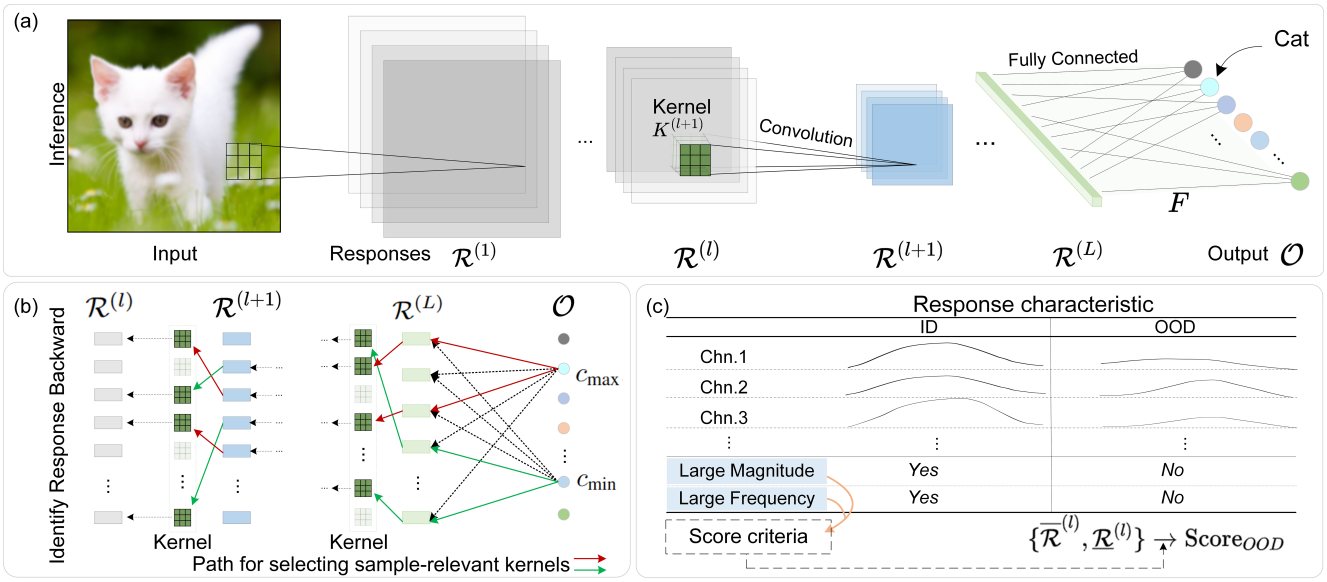


Figure 2. Illustration of CORES for OOD detection. (a) Inference pathway in neural networks, with an input progressing through convolutional layers to the final classification. (b) Backward retracing from extreme predictions to identify sample-relevant kernels. (c) Strategy for deriving the OOD score based on response magnitude and frequency, highlighting differences between ID and OOD characteristics.

### 4.2. Selection of Sample-relevant Kernels

In DNNs, there are many convolutional kernels across different layers. However, only a few of these kernels play a key role in making a prediction for a given input. Specifically, kernels leading to extreme predictions such as $c_{\max}$ and $c_{\min}$ tend to have a stronger connection or mismatch with the input. Signals on these sample-relevant kernels tend to be more reliable for OOD detection. Hence, we focus on these kernels by retracing from the extreme predictions through the network's layers.

**Selection for the Last Convolution Layer.** We define the weight matrix of the fully-connected (FC) layer that connects the last convolutional layer's responses to the output $\mathcal{O}$ as $F \in \mathbb{R}^{o \times c}$, where $o$ is the category number and $c$ is the kernel number. To identify kernels in the last convolution layer—preceding the FC layer—that are relevant to the sample, we focus on the weights in $F$ corresponding to the highest and lowest confidence categories, $c_{\max}$ and $c_{\min}$. The selection process is as follows:

$$\begin{aligned} \overline{I}^{(L)} &= \text{TopK}(\{F_{c_{\max},j}\}_{j=1:c}) \\ \underline{I}^{(L)} &= \text{TopK}(\{F_{c_{\min},j}\}_{j=1:c}) \end{aligned} \tag{6}$$

Indices $\overline{I}$ and $\underline{I}$, derived from $F$, indicate the kernel indices from the last convolutional layer that are associated with $c_{\max}$ and $c_{\min}$, respectively. The operation TopK selects a subset of entries with the largest values.

We assume that the responses from the last convolutional layer have a $1 \times 1$ dimension before connecting to the fully-connected layer. If the response is of a larger dimension, say $t \times t$ where $t > 1$, flattening is required. Before conducting the TopK operation, we average each $t \times t$ block of weights.

In summary, the responses from the kernels deemed relevant to the sample in the last convolutional layer are identified as follows:

$$\begin{aligned} \overline{\mathcal{R}}^{(L)} &= \{R_i^{(L)} | i \in \overline{I}^{(L)}\} \\ \underline{\mathcal{R}}^{(L)} &= \{R_i^{(L)} | i \in \underline{I}^{(L)}\} \end{aligned} \tag{7}$$

**Selection for Intermediate Convolutional Layers.** To trace the indices of sample-relevant kernels from the $l+1$-th to the $l$-th layer, we employ the weight matrix $K \in \mathbb{R}^{c^{(l+1)} \times c^{(l)} \times w \times h}$, which represents the inter-layer connections. Here, $c^{(l)}$ and $c^{(l+1)}$ denote the number of kernels in the $l$-th and $l+1$-th layers, respectively, while $w \times h$ indicates the kernel size. We select kernels in the $l$-th layer based on their influence on the subsequent layer's kernels. The selection process is as follows:

$$\begin{aligned} \overline{I}^{(l)} &= \text{TopK}\left(\left\{\sum_{i \in \overline{I}^{(l+1)}} \max(K_{i,j,:,:})\right\}_{j=1:c^{(l)}}\right) \\ \underline{I}^{(l)} &= \text{BotK}\left(\left\{\sum_{i \in \underline{I}^{(l+1)}} \min(K_{i,j,:,:})\right\}_{j=1:c^{(l)}}\right) \end{aligned} \tag{8}$$

The function BotK is defined as the operation selecting a subset of entries with the lowest values. Fig. 3 illustrates this selection process. Utilizing these indices, we can deter-

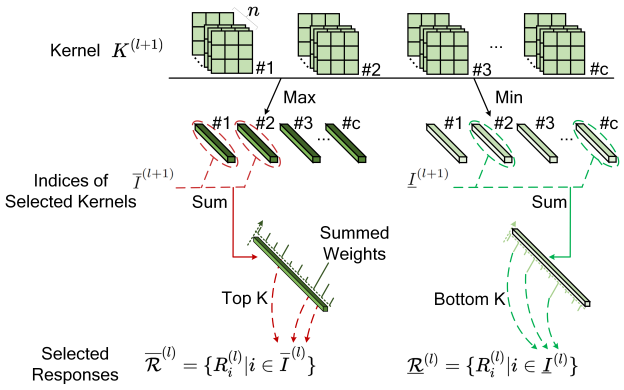


Figure 3. Demonstration of identifying sample-relevant kernels in the $l$-th layer, leveraging those kernels in the $l+1$-th layer.

mine the responses of the sample-relevant kernels at the $l$-th layer, as detailed in Eq. (7).

### 4.3. Final CORES

Leveraging responses from sample-relevant kernels, we refine Eq. (5) to establish a revised OOD score for the pertinent layer:

$$S(\overline{\mathcal{R}} \cup \underline{\mathcal{R}}) = RM^{+}(\overline{\mathcal{R}})^{\lambda_1} \cdot RM^{-}(\underline{\mathcal{R}})^{\lambda_1} \cdot RF^{+}(\overline{\mathcal{R}})^{\lambda_2} \cdot RF^{-}(\underline{\mathcal{R}})^{\lambda_2} \tag{9}$$

Practically, we aggregate responses from the sample-relevant kernels at multiple layers. The formula is applied to each respective layer to calculate individual scores, which are then averaged to yield the final OOD score.

## 5. Experimental Results

### 5.1. Experimental Setup

**Implementation.** The parameters $\lambda_1$ and $\lambda_2$ are set to 10 and 1. In the process of selecting sample-relevant kernels, the TopK and BotK operations are configured to select entries corresponding to the top and bottom $20\%$ of kernels, respectively. For CORES construction, 4 convolutional layers are utilized for models on CIFAR-10, while 5 layers are used for models on CIFAR-100 and ImageNet. We tune the thresholds $\tau_+$ and $\tau_-$ by calibrating them to minimize the false positive rate of Gaussian noise and uniform noise. All experiments are implemented in PyTorch [28] and conducted on a workstation equipped with dual 2.40 GHz CPU, 128 GB of RAM, and four NVIDIA RTX 3090Ti GPUs.

**ID/OOD Settings.** We investigate four ID/OOD scenarios: (1) *Small-scale Common OOD Setting*: CIFAR-10 and CIFAR-100 [16] serve as ID datasets. SVHN [25], Textures [2], LSUN-R [45], iSUN [43], and TinyImageNet-R are employed as OOD datasets, with "-R" denoting the resized variants. All images are uniformly resized to a $32 \times 32$ resolution. (2) *Large-scale Common OOD Setting*: We designate ImageNet-1k [4] as the ID dataset and select OpenImage-O [41], SUN [42], Places [50], and Textures [2] as OOD datasets. (3) *Small-scale Near-OOD Setting*: CIFAR-10 is utilized as the ID dataset, with CIFAR-100 acting as the OOD dataset, reflecting the semantic closeness of the categories. (4) *Large-scale Near-OOD Setting*: ImageNet-1k [4] is the ID dataset, while ImageNet-O [12] is adopted as the OOD dataset. In particular, the term "near-OOD" is applied to samples that bear semantic resemblance to ID samples [29].

**Setup.** For CIFAR benchmarks [16], we employ ResNet-18 [7], DenseNet-101 [13], and WideResNet-28-10 (WRN-28) [48]. For ImageNet-1k benchmark [4], we utilize VGG-16 [31] and SqueezeNet [15]. The overall settings are consistent with [32, 33, 46].

**Baselines.** We select seven post-hoc OOD detection methods as baselines, including maximum softmax probability (MSP) [10], ODIN [21], Energy [22], ReAct [33], DICE [32], TFEM [51] and FeatureNorm [46].

**Evaluation Metrics.** We evaluate OOD detection using the area under the receiver operating characteristic curve (AUROC) [3] and the false positive rate at a true positive rate of $95\%$ (FPR95) [21], both expressed as percentages.

### 5.2. Experimental Results

**Performance on OOD Detection.** Tab. 1 presents the OOD detection results for ResNet-18, DenseNet-101, and WRN-28 on small-scale datasets. It is evident that our CORES outperforms all competing approaches across different cases. Notably, on the CIFAR-10 dataset, CORES achieves a reduction in average FPR95 of $19.97\%$, $10.01\%$, and $5.37\%$ over the next best method, while for CIFAR-100, the reductions are $27.99\%$, $18.67\%$, and $13.91\%$, respectively. These findings underscore CORES's effectiveness and its pronounced advantage in OOD detection.

Furthermore, Tab. 2 reports the OOD detection performance of VGG-16 and SqueezeNet on large-scale datasets, which pose a greater challenge. Despite this, our CORES still attains the lowest FPR95 and highest AUROC in most cases, underscoring its strong generalization capabilities on large-scale datasets.

**Performance on Near-OOD Detection.** We evaluate our method's robustness in near-OOD scenarios. Tab. 3 presents the performance with CIFAR-10 as ID and CIFAR-100 as OOD, using ResNet-18 and WRN-28 as the DNN classifiers. In this challenging setting, our method consistently outperforms others, despite a general decline in metrics.

For large-scale datasets, Tab. 4 confirms a similar trend, with our method exhibiting considerable improvements over existing techniques. The MSP baseline shows over a $95\%$ FPR95 for VGG-16 and SqueezeNet, indicating a near-complete breakdown in detection. In contrast, our method achieves at least a $2\%$ better FPR95 and a $4\%$ higher AUROC than any alternative. These findings underscore

Table 1. Comparison on the OOD detection performance of different methods under the small-scale setting.

| ID | Model | Method | OOD | | | | | | | | | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | SVHN | | Textures | | LSUN-R | | iSUN | | TinyImageNet-R | | | |
| | | | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ |
| CIFAR-10 | ResNet-18 | MSP | 52.12 | 92.20 | 59.47 | 89.56 | 48.35 | 93.07 | 50.30 | 92.58 | 59.70 | 90.48 | 53.99 | 91.58 |
| | | ODIN | 33.83 | 93.03 | 45.49 | 90.01 | 20.05 | 96.56 | 23.09 | 96.01 | 19.82 | 96.63 | 28.46 | 94.45 |
| | | Energy | 30.47 | 94.05 | 45.83 | 90.37 | 23.62 | 95.93 | 27.14 | 95.34 | 38.80 | 93.50 | 33.17 | 93.84 |
| | | ReAct | 40.54 | 90.54 | 48.61 | 88.44 | 27.01 | 94.74 | 30.57 | 93.95 | 32.87 | 94.56 | 35.92 | 92.45 |
| | | DICE | 25.95 | 94.66 | 47.22 | 89.82 | 27.70 | 95.01 | 31.07 | 94.42 | 40.86 | 93.17 | 34.56 | 93.42 |
| | | TFEM | 38.42 | 93.53 | 43.81 | 92.32 | 19.85 | 96.59 | 23.43 | 96.00 | 33.78 | 94.05 | 31.86 | 94.50 |
| | | FeatureNorm | 7.13 | 98.65 | 31.18 | 92.31 | 27.08 | 95.25 | 26.02 | 95.38 | 33.91 | 94.31 | 25.06 | 95.18 |
| | | Ours | **3.92** | **99.24** | **10.92** | **97.85** | **2.42** | **99.31** | **2.13** | **99.37** | **6.08** | **98.76** | **5.09** | **98.91** |
| | DenseNet-101 | MSP | 47.24 | 93.48 | 64.15 | 88.15 | 42.10 | 94.51 | 42.31 | 94.52 | 50.34 | 93.14 | 49.23 | 92.76 |
| | | ODIN | 25.29 | 94.57 | 57.50 | 82.38 | 3.09 | 99.02 | 3.98 | 98.90 | 5.97 | 98.73 | 19.17 | 94.72 |
| | | Energy | 40.61 | 93.99 | 56.12 | 86.43 | 9.28 | 98.12 | 10.07 | 98.07 | 18.25 | 96.99 | 26.87 | 94.72 |
| | | ReAct | 41.64 | 93.87 | 43.58 | 92.47 | 11.46 | 97.87 | 12.72 | 97.72 | 18.25 | 96.99 | 25.53 | 95.78 |
| | | DICE | 25.99 | 95.90 | 41.90 | 88.18 | 3.91 | 99.20 | 4.36 | 99.14 | 9.08 | 98.33 | 17.05 | 96.15 |
| | | TFEM | 30.78 | 95.27 | 50.89 | 90.09 | 8.15 | 98.24 | 8.81 | 98.21 | 15.89 | 97.21 | 22.90 | 95.80 |
| | | FeatureNorm | 16.53 | 95.81 | 30.43 | 90.76 | 8.66 | 98.20 | 10.36 | 97.94 | 14.23 | 97.09 | 16.04 | 95.96 |
| | | Ours | **6.07** | **98.75** | **13.10** | **96.92** | **2.30** | **99.39** | **2.79** | **99.30** | **5.90** | **98.77** | **6.03** | **98.63** |
| | WRN-28 | MSP | 42.10 | 91.85 | 53.30 | 87.45 | 37.81 | 93.71 | 40.11 | 93.05 | 42.51 | 92.12 | 43.17 | 91.64 |
| | | ODIN | 37.08 | 88.36 | 47.58 | 82.85 | 20.51 | 95.04 | 22.95 | 94.22 | 28.60 | 90.24 | 31.34 | 90.14 |
| | | Energy | 33.11 | 90.54 | 46.06 | 85.09 | 22.68 | 94.90 | 25.12 | 94.17 | 32.29 | 92.88 | 31.85 | 91.52 |
| | | ReAct | 15.92 | 97.06 | 44.38 | 91.97 | 27.98 | 94.75 | 31.52 | 94.17 | 35.64 | 93.76 | 31.09 | 94.34 |
| | | DICE | 37.84 | 86.99 | 50.77 | 79.70 | 26.30 | 92.89 | 28.30 | 92.14 | 31.70 | 91.67 | 34.98 | 88.68 |
| | | TFEM | 50.60 | 89.50 | 45.30 | 91.18 | 11.98 | 97.70 | 26.14 | 94.70 | 32.29 | 92.88 | 33.26 | 93.19 |
| | | FeatureNorm | 3.83 | 99.18 | 14.23 | 97.06 | 8.13 | 98.32 | 5.98 | 98.71 | 10.91 | 97.86 | 8.62 | 98.23 |
| | | Ours | **2.80** | **99.50** | **6.03** | **98.79** | **1.66** | **99.52** | **1.27** | **99.59** | **4.50** | **99.06** | **3.25** | **99.29** |
| CIFAR-100 | ResNet-18 | MSP | 81.32 | 77.74 | 85.11 | 73.36 | 82.46 | 75.73 | 82.26 | 76.16 | 80.07 | 77.89 | 82.24 | 76.18 |
| | | ODIN | 40.94 | 93.29 | 83.63 | 72.37 | 79.61 | 82.13 | 76.66 | 83.51 | 75.35 | 84.40 | 71.24 | 83.14 |
| | | Energy | 81.74 | 84.56 | 85.87 | 74.94 | 73.57 | 82.99 | 73.36 | 83.80 | 71.68 | 84.76 | 77.24 | 82.21 |
| | | ReAct | 70.81 | 88.24 | 59.15 | 87.96 | 54.47 | 89.56 | 51.89 | 90.12 | 54.10 | 89.44 | 58.08 | 89.06 |
| | | DICE | 50.84 | 91.15 | 72.68 | 76.71 | 75.29 | 81.20 | 74.06 | 82.14 | 73.30 | 83.11 | 69.23 | 82.86 |
| | | TFEM | 61.48 | 90.63 | 52.36 | 89.72 | 52.68 | 90.04 | 48.48 | 90.94 | 51.97 | 89.73 | 53.39 | 90.21 |
| | | FeatureNorm | 22.82 | 95.33 | 48.51 | 83.92 | 96.73 | 62.34 | 94.32 | 65.30 | 90.78 | 67.01 | 70.63 | 74.78 |
| | | Ours | **6.78** | **98.47** | **24.38** | **94.61** | **34.27** | **93.57** | **30.01** | **94.36** | **31.58** | **93.60** | **25.40** | **94.92** |
| | DenseNet-101 | MSP | 81.70 | 75.40 | 84.79 | 71.48 | 85.24 | 69.18 | 85.99 | 70.17 | 82.67 | 73.97 | 84.08 | 72.04 |
| | | ODIN | 41.35 | 92.65 | 82.34 | 71.48 | 65.22 | 84.22 | 67.05 | 83.84 | 54.50 | 87.55 | 62.09 | 83.95 |
| | | Energy | 87.46 | 81.85 | 84.15 | 71.03 | 70.65 | 80.14 | 74.54 | 78.95 | 68.45 | 80.84 | 77.05 | 78.56 |
| | | ReAct | 83.81 | 81.41 | 77.78 | 78.95 | 60.08 | 87.88 | 65.27 | 86.55 | 64.84 | 84.88 | 70.36 | 83.93 |
| | | DICE | 54.65 | 88.84 | 65.04 | 76.42 | 49.40 | 91.04 | 48.72 | 90.08 | 43.58 | 90.96 | 52.28 | 87.47 |
| | | TFEM | 66.99 | 88.24 | 59.49 | 85.95 | 48.02 | 91.40 | 51.10 | 91.10 | 48.52 | 91.33 | 54.82 | 89.61 |
| | | FeatureNorm | 48.39 | 90.98 | 48.60 | 84.89 | 79.34 | 83.27 | 69.93 | 85.99 | 63.05 | 87.38 | 61.86 | 86.50 |
| | | Ours | **25.57** | **94.50** | **30.07** | **92.48** | **40.06** | **91.94** | **38.39** | **92.09** | **30.07** | **92.48** | **33.61** | **92.68** |
| | WRN-28 | MSP | 75.31 | 81.68 | 81.13 | 78.25 | 79.85 | 76.69 | 80.02 | 75.54 | 80.47 | 73.65 | 79.36 | 77.16 |
| | | ODIN | 27.39 | 94.39 | 74.29 | 76.66 | 65.74 | 81.62 | 66.36 | 81.26 | 67.08 | 79.47 | 60.17 | 82.68 |
| | | Energy | 70.93 | 82.94 | 80.11 | 79.04 | 74.42 | 79.35 | 75.56 | 78.02 | 76.82 | 75.72 | 75.57 | 79.01 |
| | | ReAct | 66.05 | 86.61 | 68.51 | 83.00 | 73.27 | 74.45 | 72.92 | 73.77 | 75.94 | 70.91 | 71.34 | 77.75 |
| | | DICE | 68.37 | 83.14 | 79.27 | 78.90 | 74.17 | 79.45 | 75.14 | 78.13 | 77.75 | 75.49 | 74.94 | 79.02 |
| | | TFEM | 71.01 | 87.50 | 80.46 | 78.42 | 57.01 | 88.04 | 72.78 | 81.48 | 76.03 | 78.14 | 71.46 | 82.72 |
| | | FeatureNorm | 19.90 | 96.40 | 57.96 | 79.15 | 81.92 | 80.02 | 81.87 | 79.11 | 80.54 | 78.09 | 64.44 | 82.55 |
| | | Ours | **9.47** | **98.17** | **31.86** | **92.88** | **65.46** | **86.63** | **61.25** | **87.22** | **63.27** | **86.32** | **46.26** | **90.24** |

our approach's effectiveness and confirm its superiority in near-OOD detection.

### 5.3. Ablation Studies and More Analysis

**Effect of Four Metrics in CORES.** To assess the influence of the four metrics in CORES, we use CIFAR-10 as the ID dataset and contrasting it with a suite of OOD datasets including SVHN, Textures, LSUN-R, iSUN, and TinyImageNet-R, with WRN-28 serving as the classifier. We consider five configurations of CORES: without the frequency of significant responses (w/o RF), without the magnitude of responses (w/o RM), ignoring negative convolutional responses, disregarding positive convolutional responses, and the complete CORES framework.

Table 2. Comparison on the OOD detection performance of different methods under the large-scale setting.

| ID | Model | Method | OOD | | | | | | | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | OpenImage-O | | SUN | | Places | | Textures | | | |
| | | | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ |
| ImageNet-1k | VGG-16 | MSP | 74.29 | 81.24 | 75.66 | 78.31 | 77.89 | 77.60 | 64.84 | 81.66 | 73.17 | 79.70 |
| | | ODIN | 68.96 | 82.88 | 61.31 | 86.51 | 67.33 | 83.87 | 44.57 | 89.82 | 60.54 | 85.77 |
| | | Energy | 70.06 | 84.74 | 59.34 | 86.82 | 66.27 | 83.95 | 43.90 | 89.94 | 59.89 | 86.36 |
| | | ReAct | 97.83 | 47.58 | 99.87 | 35.01 | 99.25 | 37.54 | 96.45 | 49.12 | 98.35 | 42.31 |
| | | DICE | 71.81 | 74.02 | 58.42 | 86.71 | 68.97 | 83.04 | 38.95 | 90.66 | 59.54 | 83.61 |
| | | TFEM | 65.64 | 76.29 | 44.03 | 91.50 | 48.67 | 88.84 | 37.94 | 88.99 | 49.07 | 86.41 |
| | | FeatureNorm | 57.11 | 82.91 | 28.09 | 94.37 | 41.78 | 90.21 | 23.53 | 95.05 | 37.63 | 90.64 |
| | | Ours | **51.58** | **87.24** | **23.36** | **95.02** | **38.49** | **90.61** | **14.11** | **97.26** | **31.89** | **92.53** |
| | SqueezeNet | MSP | 86.12 | 71.15 | 83.03 | 72.25 | 87.27 | 67.00 | 94.61 | 41.84 | 87.76 | 63.06 |
| | | ODIN | 80.27 | 73.21 | 78.32 | 78.37 | 83.23 | 73.31 | 92.25 | 43.43 | 83.52 | 67.08 |
| | | Energy | 74.15 | 74.28 | 56.41 | 87.88 | 67.74 | 82.73 | 67.16 | 64.51 | 66.37 | 77.35 |
| | | ReAct | 75.75 | 68.66 | 87.57 | 66.37 | 88.80 | 66.20 | 51.05 | 76.57 | 75.79 | 69.45 |
| | | DICE | 71.81 | 74.02 | 50.55 | 87.46 | 60.14 | 83.19 | 39.89 | 87.53 | 55.60 | 83.05 |
| | | TFEM | 67.46 | 79.18 | 44.61 | 91.40 | 56.64 | **86.38** | **19.63** | **96.18** | 47.08 | 88.29 |
| | | FeatureNorm | 68.72 | 77.99 | 43.91 | 90.20 | 56.95 | 84.83 | 24.15 | 94.47 | 48.43 | 86.87 |
| | | Ours | **67.35** | **80.99** | **40.23** | **91.70** | **54.63** | 85.78 | 20.55 | 95.99 | **45.69** | **88.62** |

Table 3. Performance comparison of near-OOD detection using CIFAR-10 as ID and CIFAR-100 as OOD.

| Method | Model | | | |
|---|---|---|---|---|
| | ResNet-18 | | WRN-28 | |
| | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ |
| MSP | 64.99 | 88.35 | 52.06 | 88.31 |
| ODIN | 52.83 | 88.57 | 60.08 | 73.79 |
| Energy | 50.32 | 89.77 | 44.92 | 88.36 |
| ReAct | 50.60 | 90.03 | 48.03 | 88.66 |
| DICE | 66.95 | 84.40 | 44.90 | 86.48 |
| TFEM | 50.04 | 89.73 | 44.92 | 88.35 |
| FeatureNorm | 76.17 | 79.96 | 58.33 | 86.64 |
| Ours | **47.39** | **90.25** | **43.98** | **91.64** |

Table 4. Performance comparison of near-OOD detection using ImageNet-1k as ID and ImageNet-O as OOD.

| Method | Model | | | |
|---|---|---|---|---|
| | VGG-16 | | SqueezeNet | |
| | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ |
| MSP | 96.90 | 52.29 | 97.35 | 52.22 |
| ODIN | 87.50 | 67.94 | 94.30 | 60.21 |
| Energy | 95.25 | 64.08 | 89.65 | 62.07 |
| ReAct | 98.80 | 39.53 | 82.60 | 61.12 |
| DICE | 95.25 | 64.08 | 85.60 | 65.46 |
| TFEM | 90.85 | 56.77 | 80.55 | 71.81 |
| FeatureNorm | 80.20 | 74.81 | 81.40 | 72.65 |
| Ours | **72.20** | **79.10** | **77.70** | **78.10** |

Tab. 5 reveals that each metric is essential to CORES's performance, with their removal impairing the model's OOD detection capabilities. The elimination of response magnitude metrics, $RM^+$ and $RM^-$, notably affects performance more than that of frequency metrics, $RF^+$ and $RF^-$, highlighting the value of response magnitudes. The absence of metrics associated with positive responses is particularly detrimental, leading to an approximate $3\%$ increase in the FPR95 metric. This analysis underscores the integral role each metric plays within CORES, where they contribute collectively to its success.

Table 5. Ablation study on the four CORES metrics, using the WRN-28 model trained on CIFAR-10 as the DNN classifier. The values are averaged over multiple OOD datasets.

| $RM^+$ | $RM^-$ | $RF^+$ | $RF^-$ | FPR95↓ | AUROC↑ |
|---|---|---|---|---|---|
| ✓ | ✓ | | | 3.81 | 99.18 |
| | | ✓ | ✓ | 6.33 | 98.72 |
| ✓ | | ✓ | | 5.04 | 99.02 |
| | ✓ | | ✓ | 5.46 | 98.79 |
| ✓ | ✓ | ✓ | ✓ | **3.25** | **99.29** |

**Effect of Sample-relevant Kernel Selection.** We evaluate the influence of kernel selection on OOD detection efficacy by analyzing the CORES OOD score distributions. Using CIFAR-10 and CIFAR-100 as ID datasets, we implement models WRN-28, DenseNet-101, and ResNet-18, with LSUN-R as the OOD dataset. For ImageNet-1k, VGG-16 is utilized with Textures as the OOD dataset. Fig. 4 demonstrates that applying kernel selection within CORES significantly diminishes the overlap between ID and OOD score distributions, leading to a notable decrease in FPR95. This reduction is especially marked for CIFAR-100, where FPR95 decreases by approximately $55\%$, highlighting the benefits of kernel selection in improving model discernment between ID and OOD samples.

**Analysis of CORES Metric Values.** To illustrate the efficacy of the four CORES metrics, we present a visualization of $RM^+$, $RM^-$, $RF^+$, and $RF^-$ across ID and OOD images. Utilizing WRN-28, we select an ID image from each of the CIFAR-10 categories and two OOD images from each of SVHN, LSUN-R, iSUN, TinyImageNet-R, and Textures. As depicted in Fig. 5, ID images consistently show higher values for $RM^+$, $RM^-$, $RF^+$, and $RF^-$ compared

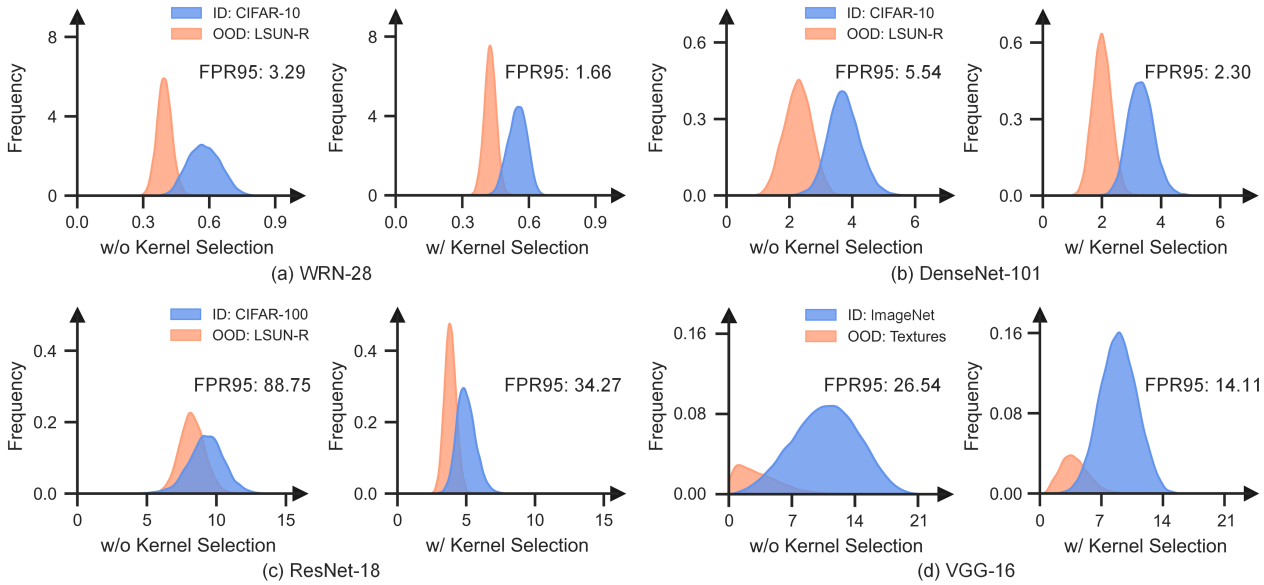


Figure 4. Comparison on the OOD score distributions and FPR95 values of CORES with and without applying sample-relevant kernel selection using different DNN models: (a) WRN-28, (b) DenseNet-101, (c) ResNet-18, and (d) VGG-16.

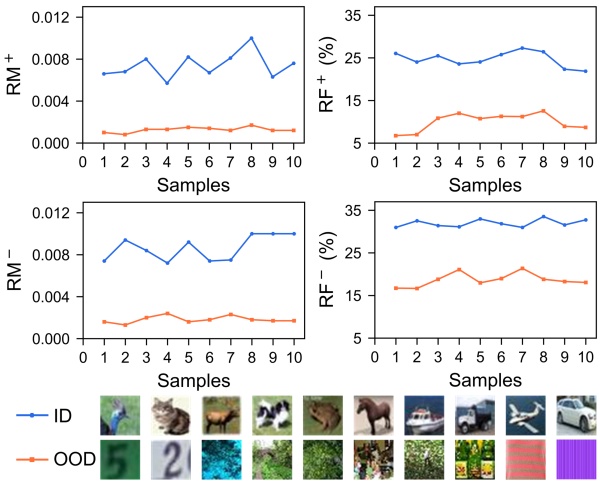


Figure 5. Comparative analysis of four CORES metrics across ID (CIFAR-10) and OOD (SVHN, LSUN-R, iSUN, TinyImageNet-R, Textures) images using WRN-28 as the DNN classifier.

to OOD images, confirming that the CORES metrics proficiently differentiate between ID and OOD samples.

**Effect of Convolutional Layer Count.** To assess how the convolutional layer count influences the performance of CORES in OOD detection, we experiment with the DenseNet-101 network, comparing CIFAR-10 and CIFAR-100 as ID datasets with various OOD datasets. Fig. 6 shows that CORES achieves the lowest performance with a single convolutional layer. Performance improves as more layers are included, with the FPR95 value decreasing and stabilizing at four layers for CIFAR-10 and five layers for CIFAR-100. These findings confirm that CORES is more effective when leveraged across multiple convolutional layers.

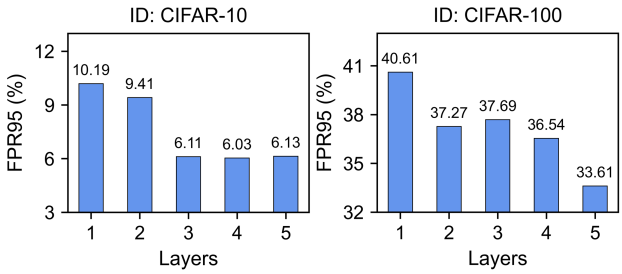


Figure 6. Comparison of CORES's OOD detection performance using different numbers of convolutional layers in DenseNet-101. FPR95 values are averaged over multiple OOD datasets.

## 6. Conclusion

In this paper, we have proposed a novel COnvolutional REsponse-based Score (CORES) for OOD detection. The rationale is that convolutional kernels generally produce more pronounced responses for ID samples compared to OOD ones, as the kernels are trained for fitting ID samples. Extensive experiments validate the effectiveness of CORES in OOD detection, and its superiority to the state-of-the-art methods. Future plans include applying convolutional responses to more security issues, such as adversarial attacks.

**Acknowledgements.** This work was supported by NSFC (62102105, U20B2046, 62073263, 61902082), Guangdong Basic and Applied Basic Research Foundation (2022A1515011501, 2022A1515010138), S&T Program of Guangzhou (202201020229), and GZHU Program (YJ2023048).

## References


[1] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, pages 1893–1902, 2015. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 5

[3] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, pages 233–240, 2006. 5

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[5] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *ICLR*, 2021. 2

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[8] Xu He, Keke Tang, Yawen Shi, Yin Li, Weilong Peng, and Peican Zhu. Are deep point cloud classifiers suffer from out-of-distribution overconfidence issue? In *SMC*, pages 2620–2627. IEEE, 2023. 2

[9] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, pages 41–50, 2019. 2

[10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2016. 1, 2, 3, 5

[11] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2018. 1

[12] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 5

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 5

[14] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021. 2

[15] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 2016. 5

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

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[18] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018. 1, 2

[19] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177, 2018. 1, 2

[20] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *ICLR*, 2017. 2

[21] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 2, 5

[22] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, pages 21464–21475, 2020. 2, 3, 5

[23] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, pages 5188–5196, 2015. 2

[24] 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. 2

[25] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 5

[26] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015. 1

[27] Bartłomiej Olber, Krystian Radlak, Adam Popowicz, Michal Szczepankiewicz, and Krystian Chachuła. Detection of out-of-distribution samples using binary neuron activation patterns. In *CVPR*, pages 3378–3387, 2023. 1

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 5

[29] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. In *ICML workshop on Uncertainty and Robustness in Deep Learning*, 2021. 5

[30] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *ICML*, pages 8491–8501, 2020. 1, 2

[31] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*. Computational and Biological Learning Society, 2015. 5

[32] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, pages 691–708, 2022. 2, 5

[33] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 2, 5

[34] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022. 1, 2

[35] Keke Tang, Dingruibo Miao, Weilong Peng, Jianpeng Wu, Yawen Shi, Zhaoquan Gu, Zhihong Tian, and Wenping Wang. Codes: Chamfer out-of-distribution examples against overconfidence issue. In *ICCV*, pages 1153–1162, 2021. 1, 2

[36] Keke Tang, Yuhong Chen, Weilong Peng, Yanling Zhang, Meie Fang, Zheng Wang, and Peng Song. Reppvconv: attentively fusing reparameterized voxel features for efficient 3d point cloud perception. *The Visual Computer*, pages 1–12, 2022. 1

[37] Keke Tang, Yuexin Ma, Dingruibo Miao, Peng Song, Zhaoquan Gu, Zhihong Tian, and Wenping Wang. Decision fusion networks for image classification. *IEEE TNNLS*, pages 1–14, 2022. 1

[38] Keke Tang, Xujian Cai, Weilong Peng, Shudong Li, and Wenping Wang. Ood attack: Generating overconfident out-of-distribution examples to fool deep neural classifiers. In *ICIP*, pages 1260–1264, 2023. 2

[39] Keke Tang, Xujian Cai, Weilong Peng, Daizong Liu, Peican Zhu, Pan Zhou, Zhihong Tian, and Wenping Wang. Matching words for out-of-distribution detection. In *ICDM*, pages 578–587. IEEE, 2023. 1

[40] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *ICLR*, 2022. 2

[41] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, pages 4921–4930, 2022. 1, 5

[42] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. 5

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

[44] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

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

[46] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *CVPR*, pages 15701–15711, 2023. 2, 3, 5

[47] Yue Yuan, Rundong He, Zhongyi Han, and Yilong Yin. Lhact: Rectifying extremely low and high activations for out-of-distribution detection. In *ACM Multimedia*, pages 8105–8113, 2023. 2

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 5

[49] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014. 2

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

[51] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Xiang Tian, Yaowu Chen, et al. Boosting out-of-distribution detection with typical features. In *NeurIPS*, 2022. 5