Discriminability-Driven Channel Selection for Out-of-Distribution Detection

Yue Yuan
Shandong University
yuanyueyy@mail.sdu.edu.cn

Rundong He*
Shandong University
rundong_he@mail.sdu.edu.cn

Yicong Dong
Shandong University
yicong_dong@mail.sdu.edu.cn

Zhongyi Han
King Abdullah University of Science and Technology
zhongyi.han@kaust.edu.sa

Yilong Yin*
Shandong University
ylyin@sdu.edu.cn

Abstract

Out-of-distribution (OOD) detection is essential for deploying machine learning models in open-world environments. Activation-based methods are a key approach in OOD detection, working to mitigate overconfident predictions of OOD data. These techniques rectifying anomalous activations, enhancing the distinguishability between in-distribution (ID) data and OOD data. However, they assume by default that every channel is necessary for OOD detection, and rectify anomalous activations in each channel. Empirical evidence has shown that there is a significant difference among various channels in OOD detection, and discarding some channels can greatly enhance the performance of OOD detection. Based on this insight, we propose Discriminability-Driven Channel Selection (DDCS), which leverages an adaptive channel selection by estimating the discriminative score of each channel to boost OOD detection. The discriminative score takes inter-class similarity and inter-class variance of training data into account. However, the estimation of discriminative score itself is susceptible to anomalous activations. To better estimate score, we pre-rectify anomalous activations for each channel mildly. The experimental results show that DDCS achieves state-of-the-art performance on CIFAR and ImageNet-1K benchmarks. Moreover, DDCS can generalize to different backbones and OOD scores.

1. Introduction

Deep neural networks are effective when the training and test sets share the same labels. However, in real-world applications, these models often encounter out-of-distribution (OOD) data, which includes labels not present in the training set's label space [11]. Existing models tend to misclassify OOD data as in-distribution (ID), leading to poor performance of OOD detection. To address this, researchers have introduced the OOD detection task, focusing on identifying whether a sample is ID or OOD. This task has gained significant attention in both academic and industrial sectors due to its critical role in various high-security applications, such as autonomous driving [2], medical diagnostics [29], and network intrusion detection [6].

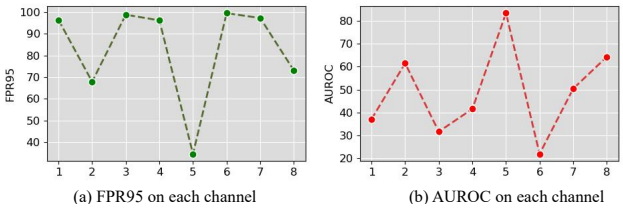


Figure 1. OOD detection performance across different channels. We randomly select 8 channels from the penultimate layer of Densenet-101 on the CIFAR-10 benchmark and evaluate the OOD detection performance using FPR95 and AUROC.

OOD detection methods are generally classified into two main types: density-based[19, 31, 43] and classification-based[22, 33]. Density-based methods, although comprehensive, are often complex and time-consuming, often resulting in performance that is inferior to classification-based methods[36]. This paper primarily examines classification-based methods, which are subdivided into training-time and test-time approaches. Training-time methods[8, 13, 41] necessitate model retraining, whereas test-time methods do not, thus eliminating training costs and preserving the model's multi-classification performance. Test-time methods are further categorized into six types: confidence-based [11, 25, 26], feature-based [34, 51], distance-based [22, 33], gradient-based [16], pruning-based methods [1, 37], and activation-based [1, 38, 44, 53]. The first five types do not account for abnormal activations in hidden layers, which can lead to overconfidence in OOD data pre-

*Corresponding authors.

dictions or underconfidence in ID data predictions, resulting in poor performance of OOD detection.

Activation-based methods [38, 44, 53] rectify anomalous activations, enhancing the distinguishability between ID data and OOD data. However, they assume by default that every channel is necessary for OOD detection, and rectify anomalous activations in each channel. To be noted, we adopt the term "channel" to denote the neurons in penultimate layer. Empirical evidence has shown that there is a significant difference among various channels in OOD detection, and discarding some channels can greatly enhance the performance of OOD detection. For example, in Figure 1, we randomly select 8 channels from the penultimate layer of DenseNet-101. Obviously, different channels exhibit varying levels of performance in OOD detection, and their contributions to this detection process also differ. Ahn et al. [1] also demonstrates that different channels react differently to ID and OOD data.

Based on this insight, we propose Discriminability-Driven Channel Selection (DDCS), which leverages an adaptive channel selection by estimating the discriminative score of each channel to boost OOD detection. We hope to select channels that greatly aid in OOD detection and discard those with minimal contribution to OOD detection. However, an important question is how to identify the contribution of diffident channels to OOD detection in the absence of OOD data. CIDER [27] shows that increasing ID discriminability can improve the separability of ID and OOD data, which demonstrates a positive correlation between ID discriminability and OOD detection. Inspired by this, we utilize ID discriminability to select channel for OOD detection. Specifically, we measure ID discriminability score through inter-class similarity and inter-class variance. Lower inter-class similarity and higher inter-class variance indicates higher ID discriminability [47, 48]. However, the estimation of discriminative score itself is susceptible to anomalous activations. To better estimate discriminability score, We pre-rectify anomalous activations in both local and global perspective.

Our contributions can be summarized as follows:

- Each channel contributes differently to OOD detection. Moreover, improving the class distinguishability of ID data can enhance OOD detection. Inspired by these two points, we propose DDCS to select channels with strong class distinguishability for OOD detection.
- To achieve DDCS, we design a discriminative score, which utilizes the inter-class difference of ID class prototypes to evaluate the ID discriminability of each channel.
- We find that the estimation of discriminative score itself is susceptible to anomalous activations. To address this problem, we implement pre-rectification for anomalous activations in both local and global perspective.
- We conduct extensive experiments on the CIFAR and ImageNet-1K benchmarks, and the results show that our proposed DDCS is state-of-the-art and can be generalized to other backbone and OOD scores.

2. Related Work

2.1. Out-of-Distribution Detection

Determining whether inputs are OOD is an essential problem for the deployment of multimedia applications. The methods of OOD detection can be categorized into two main branches: classification-based methods [25, 26], density-based methods [32, 43]. Classification-based methods in OOD detection aim to model conditional distribution and then design scoring function to measure uncertainty of test data. Density-based methods in OOD detection explicitly model the in-distribution with some probabilistic models and consider test data in low-density regions as OOD data. Density-based methods are challenging to train and optimize, and the performance often lags behind the classification-based methods [46].

The research of classification-based methods can be categorized into two main branches: test-time OOD detection methods [12, 16, 22, 25, 26, 28, 34, 38] and training-time OOD detection methods [4, 13, 17, 18, 26]. Training-time OOD detection methods aim to calibrate the model by using auxiliary OOD datasets [4, 7, 9, 13, 23, 26]. The training-time OOD detection methods use a set of extra collected large-scale auxiliary OOD data during training to help the model learn ID/OOD discrepancy. These collected OOD data help the model output lower confidence on OOD data. Lee et al. [21] and Hendrycks et al. [13] force the predictive distribution of auxiliary OOD data to uniform distribution. Chen et al. [3] present informative OOD data mining to select valuable OOD data for improving the performance of OOD detection, and enhancing the method's robustness. Liu et al. [26] proposes an energy-bounded learning objective, where the neural network is fine-tuned to explicitly create an energy gap by assigning lower energy to the ID data and higher energy to the OOD data. Test-time OOD detection methods have the advantage of being easy to use without modifying the training procedure and objective [46]. This paper focuses on test-time OOD detection methods.

Test-Time Out-of-Distribution Detection. Test-time OOD detection methods are an important branch of OOD detection. The test-time approaches do not require retraining the model, performs well, and is easy to implement in the real world. In addition, the test-time approaches are naturally suitable for privacy protection tasks where it is impossible to fine-tune the model using private data. Test-time OOD detection methods can be categorized into confidence-based, feature-based, distance-based, gradient-

based, pruning-based, and activation-based methods. **Confidence-based methods** use the confidence score of a pre-trained classifier to detect OOD data. The underlying assumption is that the ID data should receive a high confidence score, while the OOD data should receive a low confidence score. MSP [11] directly uses the maximum SoftMax score to determine whether the test sample is an ID or OOD. ODIN [24] improves the SoftMax score by perturbing the input and applying temperature scaling to the logits. Energy [26] demonstrates that the Energy score (i.e., logsumexp of logits) outperforms the SoftMax score in distinguishing between ID and OOD data. **Feature-based methods** include GRAM [34] and SHE [51]. GRAM computes the gram matrix within the hidden layers. SHE uses the energy function defined in modern Hopfield networks. **Distance-based methods** consider OOD data to be farther away from the training set than ID data. Mahalanobis [22] calculates the minimum Mahalanobis distance between the test data and the class centroids of the training set as an OOD score. **Gradient-based approaches** [16] uses gradient statistics to calculate OOD score. **Pruning-based methods** prunes the weights of model to address overconfident prediction of OOD data. DICE [37] prunes the weights of the classification layer to address overconfidence in the model's prediction of OOD data. These five OOD detection methods fail to detect abnormal activations in the neural network. However, abnormal activations can cause the model to be overconfident in predicting OOD data or underconfident in predicting ID data, which can impact the performance of OOD detection.

Activation-based methods attempt to maximize the gap between ID and OOD data by truncating abnormally low or high activations. ReAct [38] observes that OOD inputs trigger abnormally high activations, which causes the model to assign higher confidence to OOD inputs. Therefore, ReAct truncates abnormally high activations using a precomputed threshold. LHAct [50] finds that abnormally low activations cause the model to be underconfident in predicting the ID data. Therefore, LHAct truncates abnormally low activations. Additionally, LHAct designs a constrained Butterworth filter to truncate abnormally high activations. VRA [44] zeroes out abnormally low activations and truncates abnormally high activations through a variational method. BATS [53] exhibits efficacy by truncating both abnormally low and abnormally high activations of each channel. However, they assume by default that every channel is necessary for OOD detection, and rectify anomalous activations in each channel. Empirical evidence has shown that there is a significant difference among various channels in OOD detection, and discarding some channels can greatly enhance the performance of OOD detection. LINe [1] is a hybrid pruning-based and activation-based approach. In detail, LINe prunes activations and weights by measuring category contributions using Shapley values and rectifies extremely high activations. Ahn et al. [1] also demonstrates that different channels respond differently to ID and OOD data. Therefore, selecting effective channels is important for OOD detection.

3. Preliminaries

3.1. Overview

In this section, we first describe the general process of OOD detection. Next, we mainly discuss typical activation-based and pruning-based methods, which are relevant to DDCS.

3.2. Learning Set-Up

OOD detection aims to detect test samples that are drawn from a distribution that differs from the training distribution. In supervised multi-classification tasks, this means that OOD samples should not have overlapping labels with the training data. The training set $\mathcal{D}_{in}^{train} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ is drawn i.i.d from the joint data distribution $\mathcal{P}(\mathcal{X}, \mathcal{Y})$. Here, $\mathcal{X}$ represents the input space, $\mathcal{Y} = \{1, 2, \ldots, C\}$ represents the label space, n is the number of instances in $\mathcal{D}_{in}^{train}$, and $N = \{n^1, n^2, \ldots, n^C\}$ represents the number of instances per ID class. Let $\mathcal{P}(\mathcal{X})$ represent the marginal distribution of $\mathcal{X}$. Let $F : \mathcal{X} \to \mathbb{R}^M$ be a feature extractor pre-trained by $\mathcal{D}_{in}^{train}$, where M represents the number of channels. Let $G : \mathbb{R}^M \to \mathbb{R}^C$ be a classifier pre-trained by $\mathcal{D}_{in}^{train}$ to predict labels for input samples.

During the test, we design an OOD detector, denoted as G_λ, to determine whether a given sample x^{test} is ID or not. The OOD detector can make a binary decision based on the OOD score S:

$$G_\lambda(x^{test}) = \begin{cases} ID, & \text{if } S(x^{test}) \geq \lambda; \\ OOD, & \text{if } S(x^{test}) < \lambda. \end{cases} \tag{1}$$

According to Eq. (1) (where λ represents the threshold), test samples with higher OOD scores are classified as ID, while those with lower scores are classified as OOD. The pre-trained model refuses to make predictions for test data that are identified as OOD by the OOD detector. Next, we detail several typical activation-based OOD detection methods that are highly related to our DDCS.

3.3. Activation-based OOD detection

ReAct [38]: truncating extremely high activations. We consider $z = F(x^{test})$ as the activations triggered by a given test sample x^{test}, where F denotes the feature extractor. To address the extremely high activations, ReAct proposes to truncate them with a pre-defined threshold t, which is set based on the percentile of ID activation distribution of the training set $\mathcal{D}_{in}^{train}$.

BATS [53]: truncating activations into typical set. BATS finds that the activation distribution for each channel approximate a different Gaussian distribution and treats the high-density region of each channel as a typical set. BATS precomputes the mean and standard deviation of the activations on each channel using the training data. Here, $z = \{z_1, z_2, \cdots, z_M\}$ represents the activations on each channel, while $\mu = \{\mu_1, \mu_2, \cdots, \mu_M\}$ and $\sigma = \{\sigma_1, \sigma_2, \cdots, \sigma_M\}$ represent the mean and standard deviation of activation distributions on each channel, respectively. For the k-th channel, BATS considers the high-probability region (e.g. typical set) as $[\mu_k - \lambda\sigma_k, \mu_k + \lambda\sigma_k]$.

LINe [1]: class-wise activation pruning. LINe prunes activations based on Shapley values[35]. We consider x^c as an training data from class c, the Shapley value of the i-th channel in class c is calculated as: $\gamma_i^c = |F(x^c) - F(x^c; z_i \leftarrow 0)|$, where $z_i \leftarrow 0$ represents setting activations of the i-th channel to zero. In other words, γ_i^c denotes the contribution of the i-th channel to the class c. For each class, LINe selects top-k channels based on the k-largest Shapley values. Moreover, LINe defines an activation mask matrix $A \in \mathbb{R}^{M \times C}$ (i.e. pruning matrix), where M and C denotes the number of channels and classes, respectively, and set 1 for the k-largest elements from every column, otherwise 0. The activation pruning operation for a given class c can be defined as $F(x^c) \odot A^c$, where $\odot$ denotes the element-wise multiplication, and A^c represents the c-th column of the activation pruning matrix $A \in \mathbb{R}^{M \times C}$.

However, LINE has three drawbacks. First, calculating the Shapley value requires significant computational overhead. Secondly, in the inference phase, LINe is not suitable for selecting channels based on the ID category with the highest prediction probability because the OOD test data does not belong to any ID category. Finally, the performance of LINe is weaker than that of the DDCS proposed in this paper. This is because the DDCS effectively utilizes the knowledge that ID discriminability and OOD detection are positively correlated. DDCS selects channels that are favorable for OOD detection based on ID discriminability.

4. Method

Current activation-based methods typically assume that every channel is essential for Out-of-Distribution (OOD) detection. However, empirical studies have demonstrated significant variation in the contributions of different channels to OOD detection. Our hypothesis is that selectively removing certain channels can markedly enhance OOD detection performance. Additionally, there is a positive link between In-Distribution (ID) discriminability and OOD detection, leading us to focus on channels with high ID discriminability for effective OOD detection.

This section elaborates on our method, termed Discriminability-Driven Channel Selection (DDCS). First, to minimize the impact of abnormal activations on assessing ID discriminability, we introduce the Channel-level Anomalous activations Pre-rectifying (CAP) module in Section 4.1. This involves pre-rectifying anomalous activations in both local and global perspective. Next, in Section 4.2, the Channel Selection (CS) module is proposed for evaluating the ID discriminability of each channel through inter-class similarity and variance. Subsequently, we utilize only the activations from channels exhibiting high ID discriminability to calculate the OOD scores (e.g., Energy) for OOD detection, as detailed in Section 4.3.

4.1. Channel-level Anomalous Activations Pre-rectifying

According to [38], rectifying abnormal activations can improve the ID discriminability to a certain extent. Therefore, we introduce the Channel-level Anomalous activations Pre-rectifying (CAP) module to pre-rectify abnormal activations in both local and global perspective.

Specially, we rectify channel-level abnormal activations into the high-density area of the activation distributions by

$$\text{Local}\,(\boldsymbol{z}) = \begin{cases} \mu + \lambda\sigma, & \text{if} \quad \boldsymbol{z} \geq \mu + \lambda\sigma\,; \\ \boldsymbol{z}, & \text{if} \quad \mu - \lambda\sigma < \boldsymbol{z} \leq \mu + \lambda\sigma\,; \\ \mu - \lambda\sigma, & \text{if} \quad \boldsymbol{z} < \mu - \lambda\sigma\,, \end{cases} \tag{2}$$

where λ is a hyper-parameter, $\boldsymbol{z}$ denotes activations, μ and σ denote the mean and standard deviation of the channel-level activation distribution of the training dataset.

Next, we pre-rectify the extremely high activations using a global threshold, t. The formula for this pre-rectification is defined as follows:

$$\text{Global}\,(\boldsymbol{z}) = \min(\text{Local}\,(\boldsymbol{z}), t)\ . \tag{3}$$

4.2. Channel Selection

In this section, we propose the Channel Selection (CS) module, which evaluates the ID discriminability of each channel based on inter-class similarity and variance. First, for each class from the training set, we calculate the average of the features to obtain a class prototype. Then, the inter-class similarity and inter-class variance between all class prototypes are calculated to determine the discriminant score for each channel. Finally, we select channels with the highest discriminability.

4.2.1 ID class prototypes estimating

The inter-class difference among ID prototypes is used as an evaluation criterion for the channel's class discriminability. Thus, we precompute the class prototypes $H = \{h^1, h^2, \ldots, h^C\}$ for the training set $\mathcal{D}_{in}^{train}$. For a given ID

class c, we consider x_i as the i-th training data from class c, and n^c as the number of samples in class c. We then calculate the mean of features for class c as the prototype, which is defined by

$$h^c = \frac{1}{n^c} \sum_{i=1}^{n^c} \text{CAP} \circ F(x_i), \tag{4}$$

where F denotes the feature extractor and CAP denotes the channel abnormality pre-processing.

4.2.2 ID discriminative scoring

For the k-th channel, where $k \in \{1, 2, \cdots, M\}$, the inter-class difference consists of two components: inter-class similarity and inter-class variance. We consider channels with low inter-class similarity and high inter-class variance as discriminative channels. These channels are discriminative for multi-classification. First, we calculate the average similarity S_k between ID prototypes as follows,

$$S_k = \frac{1}{C(C-1)} \sum_{i=1}^{C} \sum_{j=1, j\neq i}^{C} \delta(h_k^i, h_k^j), \tag{5}$$

where $\delta(\cdot, \cdot)$ denotes cosine similarity, and $i, j \in \{1, 2, \cdots, C\}$ represent two different classes.

In addition to measuring inter-class similarity, we also introduce inter-class variance to evaluate the channel's sensitivity to inter-class differences, which is defined by

$$V_k = \frac{1}{C} \sum_{i=1}^{C} (h_k^i - \tilde{h_k})^2, \tag{6}$$

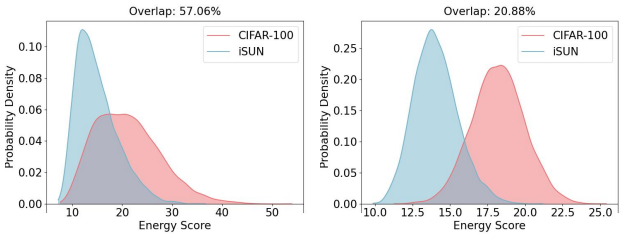
where $\tilde{h_k} = \frac{1}{C} \sum_{i=1}^{C} h_k^i$ represents the mean of ID prototypes for the k-th channel. A greater inter-class variance signiffes that the classes are more distinctly separated, thereby enhancing class discriminability.

Finally, we combine the inter-class similarity and variance with the balance factor a to calculate the discriminative score. For the k-th channel, we formulate discriminative score as

$$J_k = aS_k - (1-a)V_k. \tag{7}$$

The K channels with the lowest discriminative score are selected as class-discriminative channels, indicating that these channels have the greatest inter-class differences. We define a pruning matrix $B \in \mathbb{R}^{M \times C}$ that sets the elements in the K channels (i.e., rows) to 1 and the elements in the remaining channels to 0. Finally, we summarize DDCS in a complete formula by

$$DDCS(x_i) = (\text{CAP} \circ F(x_i)) \odot B. \tag{8}$$



(a) Without DDCS (b) With DDCS

Figure 2. DDCS promotes the separation of ID and OOD data. (a) shows the Energy score distribution for ID (CIFAR-100) and OOD (iSUN) data when DDCS is not utilized. (b) shows the Energy score distribution when utilizing DDCS. With the application of DDCS, the overlap between the Energy score distributions of ID and OOD data is significantly reduced.

The main difference between LINe and DDCS in terms of activation pruning is how the pruning matrix is defined. For each column (i.e., class), LINe keeps different channels. However, it ignores the fact that certain channels with small differences between classes can cause trouble in separating ID and OOD data. Therefore, DDCS removes these channels on each class. It is worth mentioning that these two activation pruning strategies are orthogonal and can be used simultaneously.

4.3. OOD detection with selected channels

DDCS is compatible with any downstream OOD score. We use the Energy score by default, and DDCS is also applied to other scores in the generalization analysis (Sec. 5.5.2).

Given an test input x^{test}, we first perform abnormality pre-processing for each channel. Then, we select channels based on the discriminative score. Finally, we calculate the Energy score:

$$S_{energy}(x^{test}) = -\log \sum_{c=1}^{C} \exp\left(G \circ \text{DDCS}(x^{test})\right)_c. \tag{9}$$

Following Equation 1, we can redefine the binary decision function:

$$G_\lambda(x^{test}) = \begin{cases} \text{ID} & \text{if } S_{energy}(x^{test}) \geq \lambda, \\ \text{OOD} & \text{if } S_{energy}(x^{test}) < \lambda. \end{cases} \tag{10}$$

The rule for selecting λ is allowing Eq. 10 to correctly predict the majority of the ID data (e.g. 95%). Figure. 2 shows that the gap between ID and OOD data expands largely after applying DDCS. It proves that selecting the channels with high class discriminability helps to widen the gap between ID and OOD data, thereby enhancing OOD detection.

Table 1. OOD detection performance on CIFAR-10. We use DenseNet-101 as the backbone. All values are percentages. ↑ indicates larger values are better, and ↓ indicates smaller values are better. The bold are superior results.

	OOD Datasets													
	SVHN		Textures		iSUN		LSUN-resize		LSUN-crop		Places365		**Avg**	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP [11]	47.24	93.48	64.15	88.15	42.31	94.52	42.10	94.51	33.57	95.54	63.02	88.57	48.73	92.46
ODIN [24]	25.29	94.57	57.50	82.38	**3.98**	98.90	**3.09**	99.02	4.70	98.86	52.85	88.55	24.57	93.71
Mahalanobis [22]	6.42	98.31	21.51	92.15	9.78	97.25	9.14	97.09	56.55	86.96	85.14	63.15	31.42	89.15
Energy [26]	40.61	93.99	56.12	86.43	10.07	98.07	9.28	98.12	3.81	99.15	39.40	91.64	26.55	94.57
ReAct [38]	41.64	93.87	43.58	92.47	12.72	97.72	11.46	97.87	5.96	98.84	43.31	91.03	26.45	94.67
BATS [53]	25.86	95.91	41.61	92.42	8.19	98.26	7.60	98.34	4.22	99.09	**39.31**	**92.05**	21.13	96.01
DICE [37]	25.99	95.90	41.90	88.18	4.36	**99.14**	3.91	99.20	**0.26**	**99.92**	48.59	89.13	20.83	95.24
SHE [51]	28.12	94.72	51.98	83.07	10.99	97.95	9.73	98.15	0.76	99.84	59.35	84.16	26.82	92.98
VRA [44]	18.75	96.68	34.89	93.42	5.70	98.69	5.80	98.69	1.32	99.63	39.98	91.69	17.74	96.47
LINe[1]	11.38	97.75	23.44	95.12	4.90	99.01	4.19	99.09	0.61	99.83	43.96	91.17	14.75	96.99
DDCS (ours)	**9.90**	**97.95**	**20.16**	**95.96**	4.45	99.11	3.31	**99.29**	0.70	99.85	42.90	91.19	**13.57**	**97.22**

Table 2. OOD detection performance on CIFAR-100 with DenseNet-101 as the backbone.

	OOD Datasets													
	SVHN		Textures		iSUN		LSUN-resize		LSUN-crop		Places365		**Avg**	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP [11]	81.70	75.40	84.79	71.48	85.99	70.17	85.24	69.18	60.49	85.60	82.55	74.31	80.13	74.36
ODIN [24]	41.35	92.65	82.34	71.48	67.05	83.84	65.22	84.22	10.54	97.93	82.32	76.84	58.14	84.49
Mahalanobis [22]	**22.44**	**95.67**	62.39	79.39	31.38	93.21	23.07	94.20	68.90	86.30	92.66	61.39	55.37	82.73
Energy [26]	87.46	81.85	84.15	71.03	74.54	78.95	70.65	80.14	14.72	97.43	79.20	77.72	68.45	81.19
ReAct [38]	83.81	81.41	77.78	78.95	65.27	86.55	60.08	87.88	25.55	94.92	82.65	74.04	62.27	84.47
BATS [53]	67.61	87.85	58.17	86.19	51.34	90.94	48.40	91.20	22.32	95.59	77.95	77.30	54.30	88.18
DICE [37]	54.65	88.84	65.04	76.42	48.72	90.08	49.40	91.04	**0.93**	**99.74**	79.58	77.26	49.72	87.23
SHE [51]	41.89	90.61	61.49	76.57	72.73	76.14	78.18	73.97	1.06	99.68	85.33	70.53	56.78	81.25
VRA [44]	70.91	87.46	47.64	90.17	38.53	93.42	38.52	93.49	10.73	98.04	**76.39**	**78.66**	47.12	90.21
LINe[1]	31.10	91.90	39.29	87.84	24.12	94.76	25.37	94.54	5.75	98.85	88.41	64.18	35.67	88.68
DDCS (ours)	31.34	92.58	**35.30**	**90.29**	**18.46**	**96.17**	**20.90**	**95.71**	3.84	99.21	87.11	67.91	**32.83**	**90.31**

5. Experiment

5.1. Set Up

To maintain consistency with previous research, we utilize CIFAR [20] as the ID data for the small-scale OOD detection benchmark. To ensure that there is no overlap in categories between ID and OOD test data, we use SVHN [30], LSUN-crop [49], LSUN-resize [49], iSUN [45], Textures [5], and Places365 [52] as OOD data. For the large-scale OOD detection benchmark, we choose ImageNet-1K [15] as our ID dataset. For the OOD test datasets, we choose iNaturalist [40], SUN [42], Places [52], and Textures [5]. Compared to the CIFAR benchmark, the ImageNet benchmark has a much larger label space and higher resolution. This also means that the ImageNet-1K benchmark is more realistic and challenging.

In line with previous research, we use FPR95 and AUROC as evaluation metrics for OOD detection. FPR95 measures the false positive rate of OOD data when the true positive rate of ID data is 95%; AUROC measures the area under the receiver operating characteristic curve. The lower the FPR95 or the higher the AUROC, the better the performance of OOD detection. We choose MSP [11], ODIN [24], Mahalanobis [22], Energy [26], ReAct [38], BATS [53], DICE [37], SHE[51], VRA [44], and LINe [1] as baselines. For the CIFAR benchmark, we use DenseNet-101 [14] as the backbone, which has been pre-trained on CIFAR. In the ImageNet-1K benchmark, we utilize ResNet50 [10] pre-trained on ImageNet-1K.

5.2. Main Results

Table 1 and 2 display the performance of OOD detection on the CIFAR-10 and CIFAR-100. The results show that common OOD detection methods, like MSP, ODIN, Mahalanobis, Energy, DICE, and SHE, perform poorly in detecting OOD data. This is because they do not eliminate any abnormal activations, causing the model to be overconfident in predicting OOD data. Activation-based methods, such as ReAct, BATS, VRA, and LINe, increase the separability between ID and OOD data by truncating abnormal activations, resulting in improved performance. Unlike aforementioned activation-based methods, our proposed DDCS considers the differences between different channels for OOD detection. The results show that DDCS performs best on average FPR95 and AUROC. This demonstrates that it is important to select channels with strong ID discrimination.

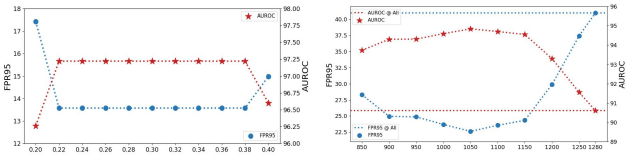
Table 3 display the OOD detection performance on the more challenging ImageNet-1K benchmark. LINe is the most advanced among the previous activation-based OOD detection methods. Surprisingly, DDCS outperforms LINe on almost all OOD datasets, particularly on iNaturalist and Texture. Compared with LINe, DDCS reduces the FPR95 by 10.89% on iNaturalist and 10.14% on Textures. With Fig. 3, we can see that DDCS widens the gap between ID and OOD data more dramatically than LINe. This proves

Table 3. OOD detection performance on ImageNet-1K. We use ResNet50 as the backbone.

	OOD Datasets									
	iNaturalist		SUN		Places		Textures		**Avg**	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP [11]	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
ODIN [24]	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
Mahalanobis [22]	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47
Energy [26]	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
ReAct [38]	20.38	96.22	24.20	94.20	33.85	91.58	47.30	89.80	31.43	92.95
BATS [53]	12.57	97.67	22.62	95.33	34.34	91.83	38.90	92.27	27.11	94.28
DICE [37]	25.63	94.49	35.15	90.83	46.49	87.48	31.72	90.30	34.75	90.77
SHE [51]	34.22	90.18	54.19	84.69	45.35	90.15	45.09	87.93	44.71	88.24
VRA [44]	15.70	97.12	26.94	94.25	37.85	91.27	21.47	95.62	25.49	94.57
LINe[1]	22.52	94.44	19.48	95.26	**12.24**	**97.56**	28.54	92.84	20.69	95.03
DDCS (ours)	**11.63**	**97.85**	**18.63**	**95.68**	28.78	92.89	**18.40**	**95.77**	**19.36**	**95.55**

Table 4. Ablation study. CAP denotes the Channel-level Anomalous activations Pre-rectifying module, and CS denotes the Channel Selection module. DDCS w/o all is equivalent to Energy[26].

Method	FPR95 ↓	AUROC ↑
DDCS w/o all	26.55	94.57
DDCS w/o CAP	23.65	95.32
DDCS w/o CS	14.70	97.01
DDCS (ours)	**13.57**	**97.22**



(a) Sensitivity of a (b) Sensitivity of K

Figure 3. Parameter sensitivity analysis. For each ID dataset, we report the averaged results across multiple OOD datasets. In (a), for the weighting factor a of the CSC score, we use CIFAR-10 as the ID dataset and DenseNet-101 as the backbone. In (b), for the number of selected channels K, we use ImageNet as the ID dataset and MobileNetV2 as the backbone.

that using ID discrimination to select channels is effective.

5.3. Ablation Study

As shown in Table 4, we conduct ablation experiments to verify the effectiveness of two key modules of DDCS: Channel-level Anomalous activations Pre-rectifying (CAP) and Channel Selection (CS). We conduct ablation experiments on these two modules using the CIFAR-10 benchmark to demonstrate their impact on the performance of DDCS. From Table 4, we find that the performance significantly degrades when either module is missing.

5.4. Sensitivity Analysis

The trade-off factor a. In Figure 3 (a), a presents trade-off factor for inter-class similarity and inter-class variance. We use DenseNet as the backbone pretrained on CIFAR-10. OOD detection performs best when a is in the range of [0.22, 0.38]. This suggests that both inter-class similarity and inter-class variance are useful for assessing the level of class discriminability of each channel. At the same time, the trade-off factor a for inter-class similarity and inter-class variance should not be too small or large.

The number of selected channels K. In Figure 3 (b), we use MobileNetV2 as the backbone pretrained on ImageNet-1K, and the number of channels in the penultimate layer is 1280. We investigate the sensitivity of the number of selected channels, K, which produces the best performance at 1050. DDCS fails when K is either too small or too large. When too few channels are selected, a significant number of features are lost. When the number of selected channels is too large, the noise from redundant channels can impede the model's ability to detect OOD data. The most extreme case is shown as the dotted line: OOD detection performance is poor when all channels are selected. This proves that our proposed channel selection is necessary.

5.5. Generalization Analysis

5.5.1 Generalizing to different backbones

As shown in Table 5, we use MobileNet as the backbone on the ImageNet benchmark to evaluate the generalization of DDCS across various backbones. In this case, DDCS uses LHAct as the CAP module. We find that the average FPR95 of DDCS is reduced by 18.30% and the average AUROC is improved by 4.23% compared to LHAct. This shows that DDCS can be applied to various backbones. It also proves that the CAP module is not limited to specific abnormal activation processing methods, such as ReAct, BATS, etc.

5.5.2 Generalizing to different OOD scores

Table 6 shows the generalization of DDCS to different OOD scores such as MSP and Energy. Compared to existing

Table 5. Generalizing to different backbones. For MobileNetV2 pretrained on ImageNet-1K, DDCS consistently performs better than baselines across all the OOD datasets.

	OOD Datasets									
	iNaturalist		SUN		Places		Textures		**Avg**	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP [11]	64.29	85.32	77.02	77.10	79.23	76.27	73.51	77.30	73.51	79.00
ODIN [24]	55.39	87.62	54.07	85.88	57.36	84.71	49.96	85.03	54.20	85.81
Mahalanobis [22]	62.11	81.00	47.82	86.33	52.09	83.63	92.38	33.06	63.60	71.01
Energy [26]	59.50	88.91	62.65	84.50	69.37	81.19	58.05	85.03	62.39	84.91
ReAct [38]	42.40	91.53	47.69	88.16	51.56	86.64	38.42	91.53	45.02	89.47
BATS [53]	50.63	91.26	57.36	86.30	64.46	83.06	40.00	91.14	53.11	87.94
DICE [37]	43.09	90.83	38.69	90.46	53.11	85.81	32.80	91.30	41.92	89.60
LHAct [50]	34.49	94.07	46.34	89.00	55.26	85.34	27.55	94.02	40.91	90.61
DDCS (ours)	**17.44**	**96.87**	**17.42**	**95.83**	**30.49**	**91.80**	**25.11**	**94.86**	**22.61**	**94.84**

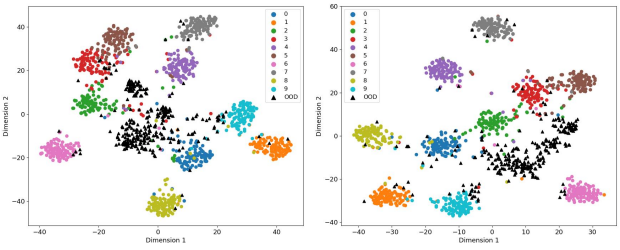
Table 6. Generalizing to different OOD scores. We use ImageNet-1K as the ID dataset and ResNet50 as the pre-trained model. The results are averaged over four OOD test datasets.

Method	FPR95 ↓	AUROC ↑
MSP	66.95	81.99
MSP + ReAct	55.68	87.28
MSP + DICE	67.41	82.24
MSP + BATS	53.89	88.23
MSP + VRA	47.09	89.62
MSP + DDCS	**43.20**	**89.95**
Energy	58.41	86.17
Energy + ReAct	32.68	93.08
Energy + DICE	34.75	90.77
Energy + BATS	30.16	93.59
Energy + VRA	25.49	94.57
Energy + DDCS	**19.36**	**95.55**

activation-based methods such as ReAct, BATS and VRA, and the pruning-based method DICE, the experimental results show that DDCS maximally corrects the model's OOD scores (e.g.MSP and Energy) for the input, and is able to reduce the model's confidence in the OOD data and increase the confidence in the ID data.

5.5.3 DDCS widens the gap between ID and OOD data

To visually demonstrate the contribution of DDCS to ID-OOD separation. We use the CIFAR-10 benchmark as an example to demonstrate the data distribution before and after discriminability-driven channel selection (DDCS). Figure 4 shows the t-SNE[39] visualization results of the (a) original deep feature space and (b) the feature space after DDCS. Dots indicate samples from the ID dataset, while triangles indicate samples from the OOD dataset. For ID data, different categories are represented by different colors. OOD data is all colored in black. It is clear that DDCS can widen the gap between various categories. Therefore, DDCS reduces the difficulty to detect OOD data.



(a) Without DDCS (b) With DDCS

Figure 4. DDCS favors widening interclass differences. (a) and (b) are t-SNE visualizations before and after applying DDCS on features. We visualize the features of the penultimate layer of DenseNet-101, with colored dots indicating the ten categories of CIFAR-10 (ID) and black triangles indicating the OOD data. Interclass differences increase significantly when applying DDCS.

6. Conclusion

In this paper, we found that each channel contributes differently to OOD detection. In addition, empirical evidence demonstrated that increasing the ID discriminability enhances OOD detection. Therefore, we propose DDCS to select channels with strong class distinguishability for OOD detection. Firstly, we design a discriminant score that utilizes the inter-class differences to assess the discrimination of each channel. Secondly, to address this issue that the estimation of the discriminant score itself is susceptible to anomalous activation, we designed a CAP module. Experiments results showed that DDCS performs best. Also, DDCS is plug-and-play, making it applicable to various scenarios. Furthermore, DDCS can be generalized to other backbones and OOD scores. In the future, we will integrate DDCS into other dynamic open-world tasks.

7. Acknowledgment

This work was supported by the National Natural Science Foundation of China (62176139). This paper was also supported by the Major Basic Research Project of Natural Science Foundation of Shandong Province (ZR2021ZD15).

References

[1] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19852–19862, 2023. 1, 2, 3, 4, 6, 7

[2] Alexander Amini, Ava Soleimany, Sertac Karaman, and Daniela Rus. Spatial uncertainty sampling for end-to-end control. *arXiv preprint arXiv:1805.04829*, 2018. 1

[3] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Informative outlier matters: Robustifying out-of-distribution detection using outlier mining. *arXiv preprint arXiv:2006.15207*, 2020. 2

[4] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–445. Springer, 2021. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 6

[6] Andrea Corsini and Shanchieh Jay Yang. Are existing out-of-distribution techniques suitable for network intrusion detection? *arXiv preprint arXiv:2308.14376*, 2023. 1

[7] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[8] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 1

[9] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 6

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 3, 6, 7, 8

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

[13] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 1, 2

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[15] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 6

[16] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *arXiv preprint arXiv:2110.00218*, 2021. 1, 2, 3

[17] Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. In *ICLR*, 2023. 2

[18] Zhuo Huang, Miaoxi Zhu, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, Bo Du, and Tongliang Liu. Robust generalization against photon-limited corruptions via worst-case sharpness minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16175–16185, 2023. 2

[19] Dihong Jiang, Sun Sun, and Yaoliang Yu. Revisiting flow generative models for out-of-distribution detection. In *International Conference on Learning Representations*, 2021. 1

[20] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 6

[21] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 2

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1, 2, 3, 6, 7, 8

[23] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13218–13227, 2020. 2

[24] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 3, 6, 7, 8

[25] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 1, 2

[26] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*, 2020. 1, 2, 3, 6, 7, 8

[27] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022. 2

[28] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. *arXiv preprint arXiv:2112.00787*, 2021. 2

[29] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020. 1

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS*, 2011. 6

[31] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019. 1

[32] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019. 2

[33] 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. *arXiv preprint arXiv:2106.09022*, 2021. 1

[34] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020. 1, 2, 3

[35] Lloyd S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952. 4

[36] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. *arXiv preprint arXiv:2209.08590*, 2022. 1

[37] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022. 1, 3, 6, 7, 8

[38] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 3, 4, 6, 7, 8

[39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 8

[40] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 6

[41] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022. 1

[42] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 6

[43] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *arXiv preprint arXiv:2003.02977*, 2020. 1, 2

[44] Mingyu Xu, Zheng Lian, Bin Liu, and Jianhua Tao. Vra: Variational rectifed activation for out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3, 6, 7

[45] 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. 6

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

[47] Yang Yang, Hongchen Wei, Zhen-Qiang Sun, Guangyu Li, Yuanchun Zhou, Hui Xiong, and Jian Yang. S2OSC: A holistic semi-supervised approach for open set classification. *ACM Trans. Knowl. Discov. Data*, 16(2):34:1–34:27, 2022. 2

[48] Yang Yang, Zhen-Qiang Sun, Hengshu Zhu, Yanjie Fu, Yuanchun Zhou, Hui Xiong, and Jian Yang. Learning adaptive embedding considering incremental class. *IEEE Trans. Knowl. Data Eng.*, 35(3):2736–2749, 2023. 2

[49] 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. 6

[50] Yue Yuan, Rundong He, Zhongyi Han, and Yilong Yin. Lhact: Rectifying extremely low and high activations for out-of-distribution detection. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 3, 8

[51] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *International Conference on Learning Representations (ICLR'23)*, 2023. 1, 3, 6, 7

[52] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 6

[53] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, bolun zheng, and Yaowu Chen. Boosting out-of-distribution detection with typical features, 2022. 1, 2, 3, 4, 6, 7, 8