Exploring Channel-Aware Typical Features for Out-of-Distribution Detection

Rundong He[1*], Yue Yuan[1*], Zhongyi Han[2†], Fan Wang[1], Wan Su[1], Yilong Yin[1†], Tongliang Liu[2, 3], Yongshun Gong[1]

[1]Shandong University
[2] Mohamed bin Zayed University of Artificial Intelligence
[3] The University of Sydney
{rundong_he, yuanyueyy, suwan}@mail.sdu.edu.cn, {hanzhongyicn, fanwangsail}@gmail.com, ylyin@sdu.edu.cn, tongliang.liu@sydney.edu.au, yongshun2512@hotmail.com

Abstract

Detecting out-of-distribution (OOD) data is essential to ensure the reliability of machine learning models when deployed in real-world scenarios. Different from most previous test-time OOD detection methods that focus on designing OOD scores, we delve into the challenges in OOD detection from the perspective of typicality and regard the feature's high-probability region as the feature's typical set. However, the existing typical-feature-based OOD detection method implies an assumption: the proportion of typical feature sets for each channel is fixed. According to our experimental analysis, each channel contributes differently to OOD detection. Adopting a fixed proportion for all channels results in several channels losing too many typical features or incorporating too many abnormal features, resulting in low performance. Therefore, exploring the channel-aware typical features is crucial to better-separating ID and OOD data. Driven by this insight, we propose expLoring channel-Aware tyPical featureS (LAPS). Firstly, LAPS obtains the channel-aware typical set by calibrating the channel-level typical set with the global typical set from the mean and standard deviation. Then, LAPS rectifies the features into channel-aware typical sets to obtain channel-aware typical features. Finally, LAPS leverages the channel-aware typical features to calculate the energy score for OOD detection. Theoretical and visual analyses verify that LAPS achieves a better bias-variance trade-off. Experiments verify the effectiveness and generalization of LAPS under different architectures and OOD scores.

Introduction

Deep neural networks exhibit remarkable effectiveness when applied to scenarios where the training and test datasets share the same label space (Sun et al. 2023, 2022; Huang et al. 2023b,a). However, in open and dynamic environments, data outside the training label space poses a significant challenge for deep neural networks (Yang et al. 2023b,a). For example, misclassifying novel diseases as known could lead to significant medical errors (Han et al. 2022). To mitigate such scenarios and guarantee the security of AI applications, substantial research endeavors have recently been dedicated to OOD detection.

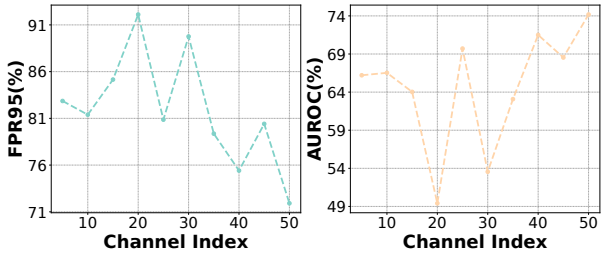


Figure 1: Different channels contribute differently to OOD detection. FPR95 and AUROC evaluate performance.

OOD detection methods can be classified into two main categories: density-based and classification-based methods. Since the performance of density-based methods (Zhou and Levine 2021; Jiang, Sun, and Yu 2021) often lags behind that of classification-based methods, and the training and optimization processes are more complex (Song, Sebe, and Wang 2022), this paper focuses on classification-based methods. The classification-based methods can be further dissected into training-time and test-time methods. The training-time methods (Du et al. 2022; He et al. 2022) require model training or fine-tuning, while the test-time methods do not require any retraining, making them convenient. The latter can be divided into four types: output-based (Hendrycks and Gimpel 2016), distance-based (Lee et al. 2018), gradient-based (Huang, Geng, and Li 2021), and rectified-activation-based methods. The first three types overlook exploring abnormal activations within the neural network's hidden layers. These abnormal activations cause overconfidence in predicting OOD data and underconfidence in predicting ID data, enlarging the overlap of ID and OOD distribution and weakening the performance of OOD detection (Sun, Guo, and Li 2021; Zhu et al. 2022).

Rectified-activation-based methods aim to address abnormal activations (features), which can be classified into two categories: clip-based and typical-set-based methods. Clip-based methods (Sun, Guo, and Li 2021; Kong and Li 2023) employ a pre-defined threshold to clip extremely high activations or employ a low-pass filter to exclude activations. Different from clip-based methods, the typical-set-based methods consider typical features. Zhu et al. (2022) regards

*These authors contributed equally.
†Corresponding authors.

the feature's high-probability region as the feature's typical set and exhibits higher efficacy by leveraging the typical features. However, existing typical-set-based methods imply an assumption: the proportion of typical feature sets for each channel is fixed. Through Fig. 1, we discover a phenomenon: different channels contribute differently to OOD detection. Several studies have observed analogous phenomena to ours, like Sun and Li (2022); Zhang and Xiang (2023). Adopting a fixed proportion for all channels results in several channels losing too many typical features or incorporating too many abnormal features, resulting in low performance. Therefore, **exploring channel-aware typical features is crucial to better-separating ID and OOD data.**

Based on the insight, we propose a novel typical-set-based method called expLoring channel-Aware tyPical featureS (LAPS). Firstly, LAPS obtains the channel-level typical set based on BATS (Zhu et al. 2022) and the global typical set by averaging the mean and standard deviation of the channel-level typical set. Then, LAPS estimates the channel-aware typical set by calibrating the channel-level typical set with the global typical set from two perspectives of mean and standard deviation. After that, LAPS rectifies the features into channel-aware typical sets to obtain channel-aware typical features. Finally, LAPS leverages the channel-aware typical features to calculate the energy score for OOD detection. Theoretical and visual analyses verify that LAPS achieves better bias-variance trade-off, thus facilitating the distinction between ID and OOD data.

Our contributions can be summarized as follows:

- A fixed proportion of typical feature sets on different channels hinders the detection of OOD data. Based on this finding, we propose an insight: exploring channel-aware typical features to enhance OOD detection.
- We propose a new typical-set-based OOD detection method called LAPS, which rectifies the features into channel-aware typical sets. Both theoretical and visual analyses prove that LAPS can achieve a better bias-variance trade-off.
- We perform extensive experimental evaluations on ImageNet-1K, CIFAR benchmarks, showing that our method outperforms the existing methods and can generalize to other architectures and OOD scores.

Related Work

Test-Time Out-of-Distribution Detection. Test-time OOD detection methods save computing resources and are naturally suitable for privacy protection tasks, as they do not require retraining the model. These methods can be categorized into output-based, distance-based, gradient-based, and rectified-activation-based methods. (1) **Output-based methods** harness the output of a pre-trained classifier to devise OOD scores. MSP (Hendrycks and Gimpel 2016) directly leverages the highest SoftMax score for OOD detection. Meanwhile, ODIN (Liang, Li, and Srikant 2017) incorporates temperature scaling alongside gradient-based input perturbations. Furthermore, Energy (Liu et al. 2020a) demonstrates that energy scores offer superior discrimination between ID and OOD data compared to softmax scores. MaxLogit (Hendrycks et al. 2019) uses maximum logit instead. (2) **Distance-based methods** design OOD scores based on the distance between test data and ID data. Mahalanobis (Lee et al. 2018) measures the minimum Mahalanobis distance between test data and training class centroids. RMD (Ren et al. 2021) proposes relative Mahalanobis distance, a simple fix to Mahalanobis. (3) **Gradient-based methods** derive OOD scores from the gradient space. GradNorm (Huang, Geng, and Li 2021) employs the vector norm of gradients to enhance OOD detection. (4) **Rectified-activation-based methods** aim to enhance the separability of ID and OOD data by rectifying activations, which can be classified into two categories: clip-based methods and typical-set-based methods. Clip-based methods (Sun, Guo, and Li 2021; Kong and Li 2023) employ a pre-defined threshold to clip extremely high activations or employ a low-pass filter to exclude activations. The typical-set-based method (Zhu et al. 2022) exhibits higher efficacy by leveraging the typical features of each channel for OOD detection. However, the rectified-activation-based methods ignore the channel-level typical set or the differences between channels, causing suboptimal rectification of activation.

Preliminaries

OOD Detection

Following Zhu et al. (2022), we provide a summary of the out-of-distribution detection from the perspective of hypothesis testing (Ahmadian, Lindsten, and Zhou 2021; Haroush et al. 2022; Zhang, Goldstein, and Ranganath 2021; Bergamin et al. 2022). We consider a K-way classification problem with a training set $\mathcal{D}_{in}^{train} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ drawn i.i.d. from a joint distribution $P(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X}$ is the input space, $\mathcal{Y}$ is the label space, $\mathcal{Y} = \{1, 2, \ldots, K\}$ is the set of ID classes, and n is the number of instances in $\mathcal{D}_{in}^{train}$. We denote the marginal distribution of $P(\mathcal{X}, \mathcal{Y})$ for the input variable X by P_0. Given a test input $\hat{\boldsymbol{x}}$ from a test set $\mathcal{D}^{test}$, the problem of out-of-distribution detection can be formulated as a single-sample hypothesis testing task:

$$\mathcal{H}_0 : \hat{\boldsymbol{x}} \in P_0, \quad \textbf{vs.} \quad \mathcal{H}_1 : \hat{\boldsymbol{x}} \notin P_0. \tag{1}$$

According to Eq. (1), the null hypothesis $\mathcal{H}_0$ assumes that the test input $\hat{\boldsymbol{x}}$ is an in-distribution sample. In out-of-distribution (OOD) detection tasks, the goal is to define criteria based on $\mathcal{D}_{in}^{train}$ to determine whether the null hypothesis $\mathcal{H}_0$ should be rejected. This involves establishing a reject region $\mathcal{R}$, such that for any test input $\hat{\boldsymbol{x}} \in \mathcal{D}^{test}$, the null hypothesis is rejected if $\hat{\boldsymbol{x}} \in \mathcal{R}$. Typically, a test statistic and a threshold define the reject region $\mathcal{R}$. Let $G : \mathcal{X} \mapsto \mathbb{R}^M$ be a feature extractor pre-trained from $\mathcal{D}_{in}^{train}$, where M denote the dimension of features. Let $F : \mathbb{R}^M \mapsto \mathbb{R}^K$ be a classifier pre-trained from $\mathcal{D}_{in}^{train}$. The test-time OOD detection methods use G and F to construct a test statistic $T(\hat{\boldsymbol{x}}; F \circ G)$. Then the reject region can be written as $\mathcal{R} = \{\hat{\boldsymbol{x}} : T(\hat{\boldsymbol{x}}; F \circ G) \leq \gamma\}$, where γ is the threshold.

OOD Detection with Rectified Activations

Rectified-activation-based OOD detection methods aim to enhance the separability of ID and OOD data by rectify-

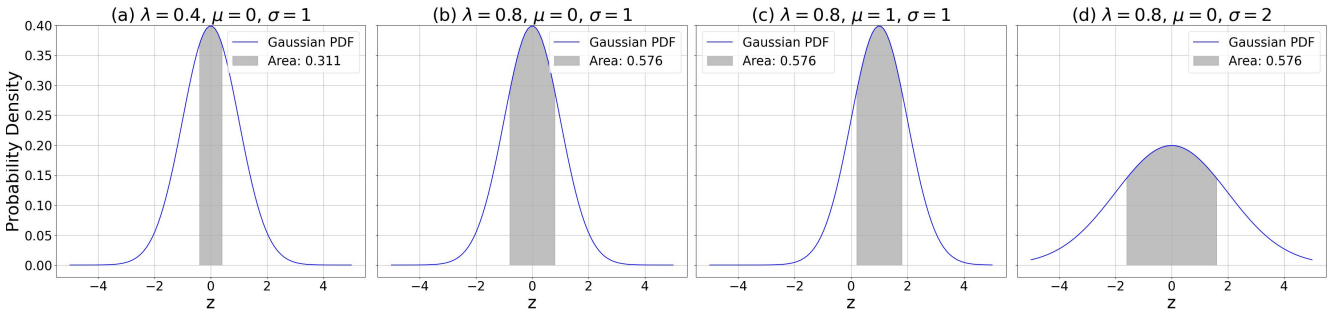


Figure 2: The visualization to verify that the proportion of high-density region $[\mu - \lambda\sigma, \mu + \lambda\sigma]$ is only determined by λ.

ing activations. According to Zhu et al. (2022), the representative rectified-activation-based methods contain two types: (1) clip-based methods, including ReAct (Kong and Li 2023) and BFAct (Kong and Li 2023); (2) typical-set-based method, including BATS (Zhu et al. 2022).

ReAct ReAct considers extremely high activations as abnormal activations because extremely high activations cause overconfidence in predicting OOD data. To address the extremely high activations, ReAct proposes to clip them with a pre-defined activation threshold c:

$$\text{ReAct}\left(\boldsymbol{z}\right) = \min(\boldsymbol{z}, c)\,, \tag{2}$$

where activation $\boldsymbol{z}$ denotes the output of feature extractor G, c is set based on the percentile of ID activation distribution of $\mathcal{D}_{in}^{train}$. Then the reject region can be rewritten as $\mathcal{R} = \{\hat{\boldsymbol{x}} : T(\hat{\boldsymbol{x}}; F \circ \text{ReAct} \circ G) \leq \gamma\}$, where γ is the threshold.

BFAct Since the extremely high activations are likely to belong to the OOD data, clipping it to the upper bound c still outputs large activations. To further eliminate the side effect of extremely high activations, BFAct addresses the extremely high activations by a low-pass filter:

$$\text{BFAct}\left(\boldsymbol{z}\right) = \frac{\boldsymbol{z}}{\sqrt{1 + \left(\frac{\boldsymbol{z}}{\lambda_2}\right)^{2N}}}\,, \tag{3}$$

where λ_2 is equivalent to the threshold c used in ReAct and N represents the order of the Butterworth filter. Then the reject region can be rewritten as $\mathcal{R} = \{\hat{\boldsymbol{x}} : T(\hat{\boldsymbol{x}}; F \circ \text{BFAct} \circ G) \leq \gamma\}$, where γ is the threshold.

BATS BATS rectifies the features into the feature's typical set and then uses these typical features to calculate the OOD score. The feature's typical set is set:

$$\text{BATS}\left(\boldsymbol{z}\right) = \begin{cases} \mu + \lambda\sigma, & \text{if} \quad \boldsymbol{z} \geq \mu + \lambda\sigma\,; \\ \boldsymbol{z}, & \text{if} \quad \mu - \lambda\sigma < \boldsymbol{z} \leq \mu + \lambda\sigma\,; \\ \mu - \lambda\sigma, & \text{if} \quad \boldsymbol{z} < \mu - \lambda\sigma\,, \end{cases} \tag{4}$$

where λ is a tuning parameter. μ and σ denote the mean and standard deviation of the channel-level feature distribution of the training dataset. Then the reject region can be rewritten as $\mathcal{R} = \{\hat{\boldsymbol{x}} : T(\hat{\boldsymbol{x}}; F \circ \text{BATS} \circ G) \leq \gamma\}$.

Although these rectified-activation-based OOD detection methods enhance the separability of ID and OOD data by rectifying activation, they ignore the channel-level typical set or the differences between channels, causing sup-optimal rectification of activation. Through Fig. 1, we discover that different channels contribute differently to OOD detection. Therefore, exploring the channel-aware typical features is crucial to better-separating ID and OOD data.

Exploring Channel-Aware Typical Features

Motivation

According to Zhu et al. (2022), exploring feature's typical set to enhance OOD detection is an effective practice. However, the existing typical-feature-based OOD detection method implies an assumption: the proportion of typical feature sets for each channel is fixed. Through Fig. 1, we discover a phenomenon: different channels contribute differently to out-of-distribution (OOD) detection. Several studies have observed analogous phenomena to ours, like Sun and Li (2022); Zhang and Xiang (2023). Adopting a fixed proportion for all channels results in several channels losing too many typical features or incorporating too many abnormal features, resulting in low performance. Therefore, each channel's fixed proportion of typical feature sets is unreasonable. Based on these findings, we propose exp**L**oring channel-**A**ware ty**P**ical feature**S** (**LAPS**).

Identifying Typical Features

According to Zhu et al. (2022), the distribution of the deep features is consistent with the Gaussian distribution. There are high-probability regions and low-probability regions in deep features. We define the features that fall in high-probability regions as typical features, and the corresponding regions are called feature's typical sets. In contrast, we define the features that fall in low-probability regions as abnormal features, including extremely high features and extremely low features. According to Kong and Li (2023), extremely high features cause overconfidence in predicting OOD data. According to Zhu et al. (2022), extremely low features cause underconfidence in predicting ID data. The overconfidence in OOD data and underconfidence in ID data enlarges the overlap of ID and OOD data, weakenings the performance of OOD detection.

To identify typical features, we should better define high-probability and low-probability regions in each channel. We

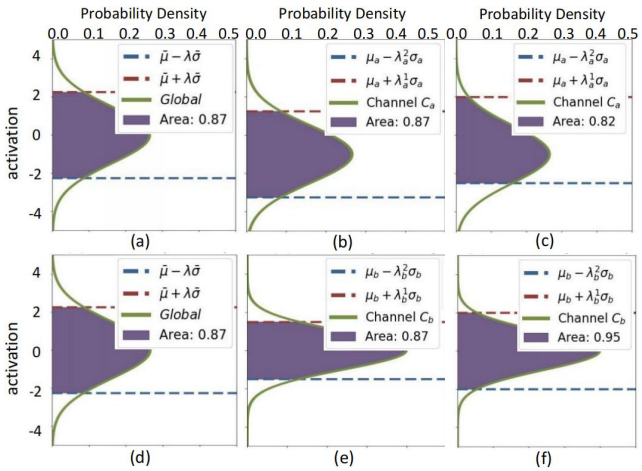


Figure 3: Estimating channel-aware typical sets. The shadowed denotes typical sets. The red and the blue denote the boundary line. (a) and (d) denote global typical sets. (b) and (e) denote typical sets estimated by BATS. (c) and (f) denote typical sets estimated by LAPS.

denote the proportion of typical feature sets in i-th channel C_i by r_i. According to Fig. 2, r_i is only related to λ_i, a hyperparameter in the high-probability interval. The mean and standard deviation of i-th channel's feature distribution of the training dataset is μ_i and σ_i. Therefore, the high-probability interval is $[\mu_i - \lambda_i\sigma_i, \mu_i + \lambda_i\sigma_i]$. The features that fall into $[\mu_i - \lambda_i\sigma_i, \mu_i + \lambda_i\sigma_i]$ can be regarded as typical features, and the others as abnormal features. Since these abnormal features hinder OOD detection, we need to rectify them into the feature's high-probability interval by

$$\text{LAPS}\left(\boldsymbol{z}_{\text{i}}\right) = \begin{cases} \mu_i + \lambda_i\sigma_i, & \text{if} \quad \boldsymbol{z}_{\text{i}} \geq \mu_i + \lambda_i\sigma_i\,; \\ \mu_i - \lambda_i\sigma_i, & \text{if} \quad \boldsymbol{z}_{\text{i}} \leq \mu_i - \lambda_i\sigma_i\,; \\ \boldsymbol{z}_{\text{i}}, & \text{others}\,, \end{cases} \tag{5}$$

where $\boldsymbol{z}_{\text{i}}$ denote the feature from i-th channel. According to Eq. (5), LAPS rectifies the channel-level abnormal features by performing channel-level truncating. Then, we perform LAPS in Eq. (5) for all channels to obtain typical features.

Estimating Channel-Aware Typical Sets

The typical-feature-based OOD detection method implies an assumption: the proportion of typical feature sets for each channel is fixed. Through Fig. 1, we discover that different channels contribute differently to out-of-distribution (OOD) detection. Adopting a fixed proportion for all channels results in several channels losing too many typical features or incorporating too many abnormal features, resulting in low performance. Therefore, each channel's fixed r_i is unreasonable. Based on these findings, we propose an insight: exploring channel-aware typical features is crucial to OOD detection. According to Fig. 2, the proportion r_i of typical feature sets is only related to hyperparameter λ_i. To obtain channel-aware typical features, we estimate channel-aware hyperparameter λ_i by calibrating the channel-level typical set with the global typical set from the mean and standard deviation.

Firstly, we obtain the global typical set $[\bar{\mu} - \lambda\bar{\sigma}, \bar{\mu} + \lambda\bar{\sigma}]$ by global feature distribution $\mathcal{D}_g(\bar{\mu}, \bar{\sigma})$, where $\bar{\mu}$ and $\bar{\sigma}$ are defined by

$$\bar{\mu} = \frac{1}{M}\sum_{i=1}^{M}\mu_i, \quad \bar{\sigma} = \frac{1}{M}\sum_{i=1}^{M}\sigma_i\,, \tag{6}$$

where M denotes the number of channels, μ_i and σ_i denote the mean and standard deviation of i-th channel's feature distribution. Next, we introduce the calibrating process from mean and standard deviation.

Calibrating from Mean Fig. 3(a) shows the global feature distribution. The interval between the red and blue boundary lines denotes the global typical set. Fig. 3(b) shows the feature distribution of channel C_a. Compared with the global typical set, C_a's typical set is biased towards the below, and partial C_a's extremely high features are not abnormal. Therefore, C_a's red boundary line should be moved up. Similarly, compared with the global typical set, partial C_a's typical features are below the global blue boundary line, which should be considered abnormal. Therefore, C_a's blue boundary line should also be moved up. To achieve the movement of C_a's red and blue boundary lines, we update λ by

$$\lambda_a^1 = \lambda + m(\bar{\mu} - \mu_a), \quad \lambda_a^2 = \lambda - m(\bar{\mu} - \mu_a)\,, \tag{7}$$

where λ_a^1 and λ_a^2 correspond to C_a's red and blue boundary lines, and m is a coefficient to control the item of mean discrepancy. The above assumes that the feature distribution of channel C_a is below the global feature distribution. When the feature distribution of channel C_a is above the global feature distribution, updating λ by Eq. (7) still holds.

Calibrating from Standard Deviation Assuming that the mean is fixed, we only consider the standard deviation in this part. Fig. 3(d) shows the global feature distribution and Fig. 3(e) shows the feature distribution of channel C_b. C_b's typical set is contained within the global typical set, and partial C_b's extremely high and low features are not abnormal. Therefore, C_b's red boundary line should be moved up, and C_b's blue boundary line should also be moved down. To achieve the movement of C_b's red and blue boundary lines, we update λ by

$$\lambda_b^1 = \lambda + n(\bar{\sigma} - \sigma_b), \quad \lambda_b^2 = \lambda + n(\bar{\sigma} - \sigma_b)\,, \tag{8}$$

where λ_b^1 and λ_b^2 correspond to C_b's red and blue boundary lines, and n is a coefficient to control the item of standard deviation discrepancy. The above assumes that the standard deviation of channel C_b's feature distribution is smaller than the global feature distribution. When the standard deviation of channel C_b's feature distribution is larger than the global feature distribution, updating λ by Eq. (8) still holds.

We summarize the above updating of λ into a complete formula by

$$\begin{aligned} \lambda_i^1 &= \lambda + m(\bar{\mu} - \mu_i) + n(\bar{\sigma} - \sigma_i) \\ \lambda_i^2 &= \lambda - m(\bar{\mu} - \mu_i) + n(\bar{\sigma} - \sigma_i)\,. \end{aligned} \tag{9}$$

We can effectively estimate channel-aware typical sets $[\mu_i - \lambda_i^2\sigma_i, \mu_i + \lambda_i^1\sigma_i]$ with Eq. (9).

	OOD Datasets									
	Textures		SVHN		LSUN		iSUN		Avg	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	83.67	73.41	84.22	71.50	66.58	83.73	82.71	75.51	79.28	76.04
ODIN	78.48	76.04	89.28	70.01	40.53	92.33	72.11	81.51	70.10	79.97
Energy	79.83	76.19	85.85	73.79	35.89	93.52	81.12	**78.90**	70.67	80.60
ReAct	68.38	83.37	77.59	87.59	33.66	92.99	79.91	74.66	64.89	84.65
BFAct	71.09	83.01	82.58	85.06	34.27	93.09	80.75	74.85	67.17	84.00
BATS	69.12	84.11	82.65	84.48	33.10	**93.63**	78.51	77.80	65.85	85.01
LAPS (ours)	**63.59**	85.06	**70.71**	**89.18**	**30.80**	93.46	**78.03**	74.92	**60.78**	**85.66**

Table 1: Comparison of OOD detection performance between LAPS and other baselines with CIFAR-100 as ID dataset.

	OOD Datasets									
	iNaturalist		SUN		Places		Textures		Avg	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	64.29	85.32	77.02	77.10	79.23	76.27	73.51	77.30	73.51	79.00
ODIN	55.39	87.62	54.07	85.88	57.36	84.71	49.96	85.03	54.20	85.81
Energy	59.50	88.91	62.65	84.50	69.37	81.19	58.05	85.03	62.39	84.91
ReAct	45.27	92.40	53.29	87.58	61.04	84.39	41.13	90.85	50.18	88.80
BFAct	40.24	92.99	51.05	87.35	58.01	83.99	**37.50**	**91.72**	46.70	89.01
BATS	50.63	91.26	57.36	86.30	64.46	83.06	40.00	91.14	53.11	87.94
LAPS (ours)	**18.82**	**96.76**	**30.07**	**92.98**	**39.70**	**90.10**	51.37	88.29	**34.99**	**92.03**

Table 2: Comparison of performance between LAPS and other baselines under MobileNet with ImageNet-1K as the ID dataset.

Exploiting Typical Features for OOD Detection

LAPS is similar to ReAct (Sun, Guo, and Li 2021) and BATS (Zhu et al. 2022) in that it is compatible with any downstream OOD scores. We default to using the Energy (Liu et al. 2020b) score as the OOD score but extend to other scores on generalization analysis.

Given a test input $\hat{\boldsymbol{x}}$ from a test set $\mathcal{D}^{test}$, we first obtain the channel-aware typical features by applying the LAPS function in Eq. (5) for each channel. Based on channel-aware typical features, we calculate the energy score:

$$S_{energy}(\hat{\boldsymbol{x}}) = -\log \sum_{k=1}^{K} \exp\left(F \circ \text{LAPS} \circ G(\hat{\boldsymbol{x}})\right)_k . \quad (10)$$

Then, the reject region can be rewritten as $\mathcal{R} = \{\hat{\boldsymbol{x}} : -S_{energy}(\hat{\boldsymbol{x}}) \leq \gamma\}$, where γ is the threshold. We usually set the threshold γ to accurately classify a significant portion of the ID data (e.g., 95%).

Theoretical and Visualization Analysis

Understanding from Variance Reduction

Following Zhu et al. (2022), we analyze the benefits of LAPS from the perspective of variance reduction. Assuming the original feature distribution of channel C_i is $N(\mu, \sigma^2)$, where μ represents the mean and σ^2 represents the variance. After rectifying the original feature distribution with the **LAPS** function, the variance becomes:

$$\begin{aligned}\tilde{\sigma}^2 &= \sigma^2 \times C(\lambda_i^1, \lambda_i^2) \\ &= \sigma^2 \times \left(1 - \frac{\lambda_i^1 \phi(\lambda_i^1) + \lambda_i^2 \phi(-\lambda_i^2)}{\Phi(\lambda_i^1) - \Phi(-\lambda_i^2)} - \left(\frac{\phi(\lambda_i^1) - \phi(-\lambda_i^2)}{\Phi(\lambda_i^1) - \Phi(-\lambda_i^2)}\right)^2\right),\end{aligned} \quad (11)$$

where λ_i^1 and λ_i^2 denote the updated λ in Eq. (9), $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and cumulative distribution function of the standard normal distribution, respectively. According to Burkardt (2014), we can know that $C(0,0) = 0$ and $C(+\infty, +\infty) = 1$. The smaller λ_i^1 and λ_i^2, the smaller the variance. Zhu et al. (2022) points out that the extreme features increase the uncertainty and lead to overconfidence in predicting OOD data and underconfidence in predicting ID data. Our LAPS solves the above problem by reducing the variance of the feature distribution, which improves the estimation accuracy of the reject region.

The Bias Introduced by LAPS

While LAPS decreases the variance of the feature distribution, it can also introduce a bias term that captures the shift in the feature distribution. A large bias can damage the performance (Zhu et al. 2022), which is defined by

$$\mathbb{E}[\text{LAPS}(z)] - \mathbb{E}[z] = -\sigma \frac{\phi(\lambda_i^1) - \phi(-\lambda_i^2)}{\Phi(\lambda_i^1) - \Phi(-\lambda_i^2)} . \quad (12)$$

According to Eq. (12), the bias term $\mathbb{E}[\text{LAPS}(z)] - \mathbb{E}[z]$ converges to zero as λ_i^1 and λ_i^2 approaches positive infinity. If λ_i^1 and λ_i^2 are large enough, the bias can be very small. Thus, there exists a bias-variance trade-off in LAPS.

A Better Bias-variance Trade-off of LAPS

Fig. 4(a), (b), and (c) show that a proper selection of λ can improve the detection performance by significantly reducing variance and slightly changing the distribution of the features (small bias). For example, we can get a good bias-variance trade-off when $\lambda = 1.5$. Through Fig. 4(d), (e), and (f), channel-aware λ_i^1 and λ_i^2 can achieve better bias-variance trade-off by a proper selection of m and n.

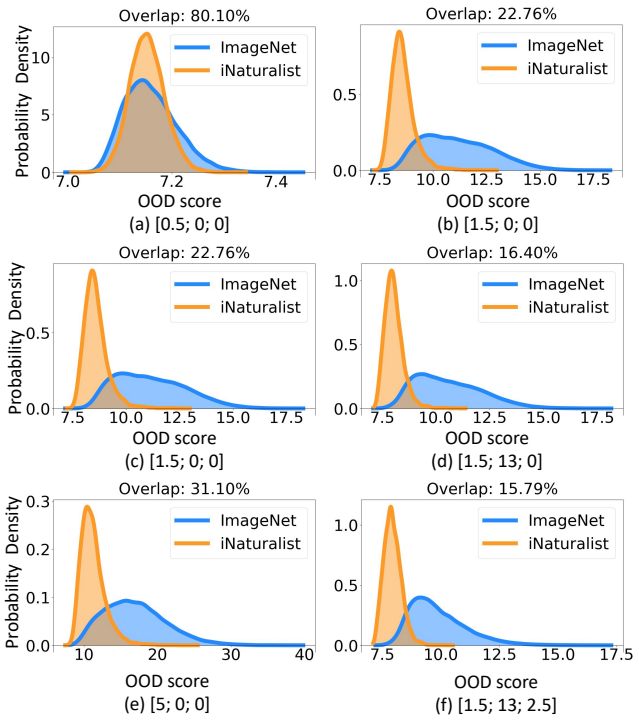
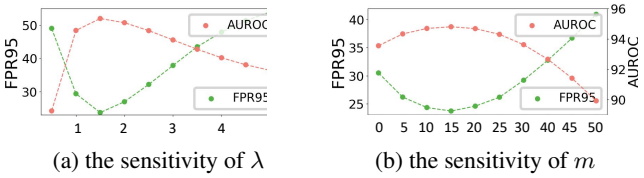


Figure 4: Bias-variance trade-off under $[\lambda; m; n]$ with ID (ImageNet) and OOD (iNaturalist) data.



(a) the sensitivity of λ (b) the sensitivity of m

Figure 5: The sensitivity analysis of λ and m.

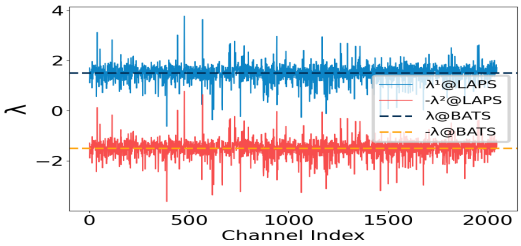


Figure 6: The visualization of fixed $\{-\lambda, \lambda\}$ by BATS and channel-aware $\{-\lambda_i^2, \lambda_i^1\}$ by LAPS across each channel. The FPR95 for BATS and LAPS are 30.16% and 23.68%.

Experiments

Set Up

We use CIFAR-100 (Krizhevsky and Hinton 2009) as ID data for small-scale OOD detection benchmark. For the OOD test data, we utilize Textures (Cimpoi et al. 2014), SVHN (Netzer et al. 2011), LSUN (Yu et al. 2015) and iSUN (Xu et al. 2015). For large-scale OOD detection benchmark, we select ImageNet-1K (Huang and Li 2021) as our ID dataset. For the OOD test datasets, following Zhu et al. (2022), we choose iNaturalist (Horn et al. 2017), SUN (Xiao et al. 2010), Places (Zhou et al. 2018), and Textures (Cimpoi et al. 2014).

Following Zhu et al. (2022), our experiments use AUROC and FPR95 as test metrics for OOD detection. We choose **MSP** (Hendrycks and Gimpel 2016), **ODIN** (Liang, Li, and Srikant 2017), **Energy** (Liu et al. 2020a), **ReAct** (Sun, Guo, and Li 2021), **BFAct** (Kong and Li 2023), and **BATS** (Zhu et al. 2022) as baselines. Code is available at: https://github.com/rm1972/LAPS.git.

Main Results

Table 1 reports the results on ID dataset CIFAR-100 with four OOD datasets. Our algorithm significantly outperforms all the comparison methods by a large margin on average FPR95 and AUROC. It is evident that the energy score obtained by applying the ReAct, BFAct, and BATS to the activations is superior to the vanilla energy score. Our LAPS further enhances the performance of the energy score.

We tackle the more challenging task of using ImageNet-1K as the ID data and used MobileNet-V2 as the backbone. Table 2 shows that our method outperforms all baselines on average performance. Compared to BATS, LAPS shows a 18.12% decrease in average FPR95 and a 4.09% increase in average AUROC. These results verify that using channel-aware typical features to calculate the energy score can improve the separability of ID and OOD data.

Sensitivity Analysis

LAPS has three important hyper-parameters (including λ, m, and n). Here, we empirically show the influence of the hyperparameter λ in Fig. 5(a). As λ tends to infinity, BATS approaches the Energy Score (the horizontal lines). As λ tends to zero, a large bias damages the performance. The bias-variance trade-off shows that selecting proper λ is important. Fig. 5(b) and Table. 5 show the influence of the hyperparameter m and n, which verifies that selecting a proper m and n help achieve better bias-variance trade-off.

Fig. 6 visualizes the fixed $\{-\lambda, \lambda\}$ by BATS and channel-aware $\{-\lambda^2, \lambda^1\}$ by LAPS across each channel. The channel-aware $\{-\lambda^2, \lambda^1\}$ contributes to OOD detection, which verifies our insight that exploring channel-aware typical features is crucial to better-separating ID and OOD data.

Ablation Study

Table 6 presents the results of the ablation study. We adopt energy as a vanilla OOD score. **BATS**, **BATS+variance**, **BATS+mean**, and **BATS+mean+variance** (i.e., LAPS) obtain feature's typical set by fixed λ, Eq. (7), Eq. (8), and Eq. (9), respectively. According to Table 6, calibrating from the mean and standard deviation contributes to obtaining channel-aware typical features to enhance OOD detection.

	OOD Datasets									
	iNaturalist		SUN		Places		Textures		Avg	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
ODIN	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
Energy	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
ReAct	19.99	96.31	29.60	93.42	39.70	90.95	41.42	91.62	32.68	93.08
BFAct	20.69	96.20	21.02	95.24	30.33	92.62	54.20	87.78	31.56	92.96
BATS	24.98	95.51	25.68	94.27	37.34	91.11	**32.62**	**93.47**	30.16	93.59
LAPS (ours)	**12.72**	**97.50**	**15.81**	**96.18**	**24.71**	**93.64**	41.49	91.81	**23.68**	**94.78**

Table 3: Comparison of performance between LAPS and other baselines under ResNet50 with ImageNet-1K as ID dataset.

	OOD Datasets									
	iNaturalist		SUN		Places		Textures		Avg	
Method	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	58.21	87.06	73.45	78.90	75.90	78.06	71.22	78.43	69.70	80.61
ODIN	46.13	91.66	59.83	85.03	66.44	82.64	50.53	86.65	55.73	86.50
Energy	56.47	89.75	59.73	86.04	64.97	83.56	53.19	86.25	58.59	86.40
ReAct	35.20	93.97	40.58	91.64	48.57	88.83	40.35	91.16	41.18	91.40
BFAct	27.31	94.95	29.88	93.76	38.25	90.52	40.23	91.77	33.92	92.75
BATS	39.71	92.68	40.83	91.45	49.46	88.15	**37.61**	**91.96**	41.90	91.06
LAPS (ours)	**15.18**	**97.20**	**15.69**	**96.39**	**25.99**	**93.31**	47.59	89.47	**26.11**	**94.09**

Table 4: Comparison of performance between LAPS and other baselines under ResNet18 with ImageNet-1K as ID dataset.

n	0.5	0.6	0.7	0.8	0.9	1.0
FPR95 ↓	29.62	29.49	**29.33**	29.43	29.43	29.46

Table 5: Effect of n for OOD detection.

Method	FPR95 ↓	AUROC ↑
Energy	58.41	86.17
BATS	30.16	93.59
BATS + variance	29.33	93.66
BATS + mean	26.73	94.21
BATS + mean + variance (LAPS)	**23.68**	**94.78**

Table 6: Ablation studies on ImageNet-1K with ResNet50.

Generalization Analysis

Generalizing to Other Architectures. We validate the generalization of LAPS with different model architectures: ResNet50 and ResNet18. Table 3 and 4 show that our approach outperforms competitive baselines, achieving the best performance across different network architectures. When utilizing ResNet50 as the backbone, LAPS achieved a 6.48% reduction in average FPR95 compared to BATS, alongside a 1.19% increase in average AUROC. These outcomes underscore LAPS's remarkable capacity to generalize and maintain robustness across diverse architectures.

Generalizing to Other OOD Scores. Table 7 demonstrates the generalizability of LAPS with different OOD scores. Compared with MSP, ODIN, Energy score using BATS, after replacing BATS with LAPS, the average FPR95 decreased by 9.16%, 4.14%, 6.48% respectively. Experimental findings consistently show that our method improves the performance of all scoring functions.

Method	MSP	ODIN	Energy
base	66.95	56.48	58.41
+ ReAct	55.68	37.27	32.68
+ BFAct	56.02	35.08	31.56
+ BATS	53.89	38.54	30.16
+ LAPS (ours)	**44.73**	**34.40**	**23.68**

Table 7: FPR95 on ImageNet-1K with ResNet50.

Conclusion

In this paper, we found that a fixed proportion of typical feature sets across various channels hampers OOD detection. Based on this finding, we proposed an insight: exploring channel-aware typical features to enhance ID-OOD separability. We designed a new typical-set-based OOD detection method called LAPS, which rectifies the features into channel-aware typical sets. Both theoretical and visual analyses proved that LAPS could boost the trade-off between bias and variance. LAPS effectively reduced the variance in the OOD score while maintaining bias control. Additionally, extensive experiments demonstrated that LAPS outperforms existing methods on CIFAR and ImageNet-1K.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62176139, 62202270), the Major Basic Research Project of Natural Science Foundation of Shandong Province (ZR2021ZD15), Taishan Scholar Project of Shandong Province (tsqn202306066).

References

Ahmadian, A.; Lindsten, F.; and Zhou, Z.-H. 2021. Likelihood-free Out-of-Distribution Detection with Invertible Generative Models. In *IJCAI*, 2119–2125.

Bergamin, F.; Mattei, P.-A.; Havtorn, J. D.; Senetaire, H.; Schmutz, H.; Maaløe, L.; Hauberg, S.; and Frellsen, J. 2022. Model-agnostic out-of-distribution detection using combined statistical tests. In *International Conference on Artificial Intelligence and Statistics*, 10753–10776. PMLR.

Burkardt, J. 2014. The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, 1: 35.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 3606–3613.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. *arXiv preprint arXiv:2202.01197*.

Han, Z.; Gui, X.-J.; Sun, H.; Yin, Y.; and Li, S. 2022. Towards accurate and robust domain adaptation under multiple noisy environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 6460–6479.

Haroush, M.; Frostig, T.; Heller, R.; and Soudry, D. 2022. A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

He, R.; Han, Z.; Lu, X.; and Yin, Y. 2022. Safe-Student for Safe Deep Semi-Supervised Learning With Unseen-Class Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14585–14594.

Hendrycks, D.; Basart, S.; Mazeika, M.; Zou, A.; Kwon, J.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2019. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Horn, G. V.; Aodha, O. M.; Song, Y.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. J. 2017. The iNaturalist Challenge 2017 Dataset. *CoRR*, abs/1707.06642.

Huang, R.; Geng, A.; and Li, Y. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34: 677–689.

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-distribution Detection for Large Semantic Space. *CoRR*, abs/2105.01879.

Huang, Z.; Xia, X.; Shen, L.; Han, B.; Gong, M.; Gong, C.; and Liu, T. 2023a. Harnessing out-of-distribution examples via augmenting content and style. In *ICLR*.

Huang, Z.; Zhu, M.; Xia, X.; Shen, L.; Yu, J.; Gong, C.; Han, B.; Du, B.; and Liu, T. 2023b. Robust Generalization against Photon-Limited Corruptions via Worst-Case Sharpness Minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16175–16185.

Jiang, D.; Sun, S.; and Yu, Y. 2021. Revisiting flow generative models for Out-of-distribution detection. In *International Conference on Learning Representations*.

Kong, H.; and Li, H. 2023. BFAct: Out-of-Distribution Detection with Butterworth Filter Rectified Activations. In Sun, F.; Cangelosi, A.; Zhang, J.; Yu, Y.; Liu, H.; and Fang, B., eds., *Cognitive Systems and Information Processing*, 115–129. Singapore: Springer Nature Singapore. ISBN 978-981-99-0617-8.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020a. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33: 21464–21475.

Liu, W.; Wang, X.; Owens, J. D.; and Li, Y. 2020b. Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. *NIPS*.

Ren, J.; Fort, S.; Liu, J.; Roy, A. G.; Padhy, S.; and Lakshminarayanan, B. 2021. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*.

Song, Y.; Sebe, N.; and Wang, W. 2022. RankFeat: Rank-1 Feature Removal for Out-of-distribution Detection. *arXiv preprint arXiv:2209.08590*.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34.

Sun, Y.; and Li, Y. 2022. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 691–708. Springer.

Sun, Y.; Peng, D.; Huang, H.; and Ren, Z. 2022. Feature and semantic views consensus hashing for image set classification. In *Proceedings of the 30th ACM International conference on multimedia*, 2097–2105.

Sun, Y.; Wang, X.; Peng, D.; Ren, Z.; and Shen, X. 2023. Hierarchical hashing learning for image set classification. *IEEE Transactions on Image Processing*, 32: 1732–1744.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *CoRR*, abs/1504.06755.

Yang, Y.; Sun, Z.; Zhu, H.; Fu, Y.; Zhou, Y.; Xiong, H.; and Yang, J. 2023a. Learning Adaptive Embedding Considering Incremental Class. *IEEE Trans. Knowl. Data Eng.*, 35(3): 2736–2749.

Yang, Y.; Zhou, D.; Zhan, D.; Xiong, H.; Jiang, Y.; and Yang, J. 2023b. Cost-Effective Incremental Deep Model: Matching Model Capacity With the Least Sampling. *IEEE Trans. Knowl. Data Eng.*, 35(4): 3575–3588.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *CoRR*, abs/1506.03365.

Zhang, L.; Goldstein, M.; and Ranganath, R. 2021. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, 12427–12436. PMLR.

Zhang, Z.; and Xiang, X. 2023. Decoupling MaxLogit for Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3388–3397.

Zhou, A.; and Levine, S. 2021. Amortized Conditional Normalized Maximum Likelihood: Reliable Out of Distribution Uncertainty Estimation. In *International Conference on Machine Learning*, 12803–12812. PMLR.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464.

Zhu, Y.; Chen, Y.; Xie, C.; Li, X.; Zhang, R.; Xue, H.; Tian, X.; bolun zheng; and Chen, Y. 2022. Boosting Out-of-distribution Detection with Typical Features. arXiv:2210.04200.