Shaping Parameter Contribution Patterns for Out-of-Distribution Detection

Haonan Xu[1], **Yang Yang**[1*]

[1]Nanjing University of Science and Technology
{xhnxhn, yyang}@njust.edu.cn

Abstract

Out-of-distribution (OOD) detection is a well-known challenge due to deep models often producing overconfident. In this paper, we reveal a key insight that trained classifiers tend to rely on sparse parameter contribution patterns, meaning that only a few dominant parameters drive predictions. This brittleness can be exploited by OOD inputs that anomalously trigger these parameters, resulting in overconfident predictions. To address this issue, we propose a simple yet effective method called Shaping Parameter Contribution Patterns (SPCP), which enhances OOD detection robustness by encouraging the classifier to learn boundary-oriented dense contribution patterns. Specifically, SPCP operates during training by rectifying excessively high parameter contributions based on a dynamically estimated threshold. This mechanism promotes the classifier to rely on a broader set of parameters for decision-making, thereby reducing the risk of overconfident predictions caused by anomalously triggered parameters, while preserving in-distribution (ID) performance. Extensive experiments under various OOD detection setups verify the effectiveness of SPCP.

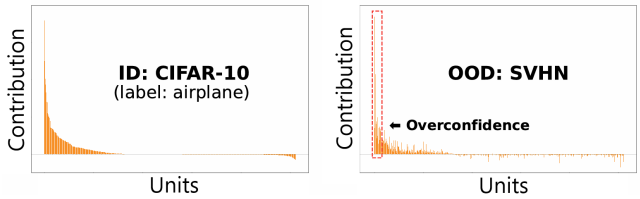


Figure 1: An illustrative example of parameter contribution patterns in a ResNet-18 classifier trained on CIFAR-10 using cross-entropy loss. Units are sorted in the same order. This example focuses on the 'airplane' class, showcasing the contributions of each unit in the classifier weights to the model outputs for both ID and OOD samples. The parameter contributions are defined with reference to Eq. (4). Since model outputs are directly determined by parameter contributions, OOD input that anomalously triggers dominant parameters can be confidently yet incorrectly classified as ID categories and hurt OOD detection.

1 Introduction

Deep neural networks have been widely applied in various fields (Radford et al. 2018; Dosovitskiy et al. 2021; Wan and Yang 2025) and have achieved remarkable success (Liu et al. 2021; Achiam et al. 2023; Yang et al. 2024). However, when deployed in real-world scenarios, deep models may fail by confidently yet erroneously classifying OOD data as one of the predefined training classes (Nguyen, Yosinski, and Clune 2015; Bendale and Boult 2016; Yang et al. 2021). The presence of such unreliable behavior can introduce considerable risks, particularly in safety-critical domains like autonomous driving (Geiger, Lenz, and Urtasun 2012) and medical diagnosis (Litjens et al. 2017). Therefore, equipping models with the capability to reliably identify and reject predictions for OOD inputs, a task known as OOD detection, is essential to ensure the trustworthiness of AI systems.

To date, extensive efforts have been dedicated to advancing reliable methods for OOD detection (Zhang et al. 2023b). One line of research focuses on designing suitable OOD scoring functions (Hendrycks and Gimpel 2017; Liu et al. 2020; Hendrycks et al. 2022) for a given well-trained model to estimate OOD uncertainty, or to further improve them through post-hoc network adjustments (Zhang et al. 2020; Zhu et al. 2022; Xu et al. 2024; Xu and Yang 2025). A complementary line of work aims to strengthen the model's OOD detection capabilities with training-time regularization. Among these, some methods utilize outlier exposure (Hendrycks, Mazeika, and Dietterich 2019; Zhang et al. 2023a; Wang et al. 2023) to guide the model in learning more robust decision boundaries. However, such outlier data may not always be readily available in practice. Alternatively, other methods tackle OOD detection by imposing favorable constraints during the training process (Wei et al. 2022; Regmi et al. 2024; Zhu et al. 2023; Zhang et al. 2024; Ghosal, Sun, and Li 2024), which do not require any additional data and offer a promising path toward improved performance, being the primary focus of this paper.

In this paper, we present a novel perspective on the causes of model overconfidence by examining how classifier parameters contribute to predictions. As illustrated in Figure 1, we reveal the empirical observation that a well-trained model's classifier tends to exhibit sparse contribution patterns in its predictions (see more examples in Appendix A),

*Corresponding author

a finding also supported by (Frankle and Carbin 2019; Sun and Li 2022). The sparsity of parameter contribution patterns indicates that a small subset of parameters dominates the classifier's predictions. This pattern makes OOD detection brittle because some of these dominant parameters can be anomalously triggered by OOD inputs (Nguyen, Yosinski, and Clune 2015; Sun, Guo, and Li 2021). Such parameters are exploited by OOD inputs, leading to overconfident predictions favoring ID categories (as shown in the right panel of Figure 1). Therefore, suppressing the dominance of parameters with excessive predictive power is crucial for robust OOD detection.

The above analysis naturally motivates the proposal of Shaping Parameter Contribution Patterns (SPCP), a simple yet effective method for OOD detection. The core idea behind SPCP is to induce contribution bounds for the classifier's parameters to enhance the robustness of OOD detection. To achieve this, SPCP truncates parameter contributions that exceed a dynamically estimated threshold during the training process. This strategy enforces bounded constraints on the dominant parameter contributions, compelling the classifier to rely on a broader subset of parameters during decision-making. As a result, SPCP reduces the risk of overconfident predictions potentially caused by anomalously triggered parameters, making the OOD scores derived from the calibrated model more reliable and improving the ID-OOD separation. Extensive experiments on the OpenOOD benchmark (Zhang et al. 2023b) verify the effectiveness of SPCP, demonstrating that our SPCP can enhance the model's OOD detection capability across various OOD scenarios, while also preserving ID task performance.

2 Preliminaries

Setup. In this paper, we focus on the setting of K-way image classification. Formally, let $\mathcal{X}$ denote the input space and let the ID label space be defined as $\mathcal{Y} = \{1, 2, ..., K\}$. The learner has access to a labeled training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$, where samples are drawn *i.i.d.* from a joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$. Let $f : \mathcal{X} \to \mathbb{R}^K$ denote the classification model. For typical architectures, f first extracts a D-dimensional penultimate feature representation $h(\mathbf{x}) \in \mathbb{R}^D$ from an input $\mathbf{x} \in \mathcal{X}$. The classifier layer, parameterized by a weight matrix $\mathbf{W} \in \mathbb{R}^{D\times K}$ and a bias vector $\mathbf{b} \in \mathbb{R}^K$, then maps $h(\mathbf{x})$ to the output vector $f(\mathbf{x}) \in \mathbb{R}^K$. Formally, the output of the model can be written as:

$$f(\mathbf{x}) = \mathbf{W}^\top h(\mathbf{x}) + \mathbf{b}. \tag{1}$$

Out-of-distribution Detection. The goal of OOD detection is to determine whether a given input $\mathbf{x}$ originates from an irrelevant distribution whose label set does not intersect with $\mathcal{Y}$ (Yang et al. 2021). In practice, this task is often formulated as a binary decision problem using level-set estimation:

$$g_\tau(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } S(\mathbf{x}; f) > \tau; \\ \text{OOD}, & \text{if } S(\mathbf{x}; f) \le \tau, \end{cases} \tag{2}$$

where τ is a threshold typically chosen such that the majority (*e.g.*, 95%) of ID data are correctly classified, and $S(\mathbf{x}; f) : \mathcal{X} \to \mathbb{R}$ is the scoring function that quantifies the model's predictive uncertainty as a scalar value. A widely adopted option for $S(\mathbf{x}; f)$ is the Energy score (Liu et al. 2020), which is defined as follows:

$$S_{\text{Energy}}(\mathbf{x}; f) = \log \sum_{k\in\mathcal{Y}} \exp(f_k(\mathbf{x})), \tag{3}$$

where $f_k(\cdot)$ denotes the k-th output logit. The underlying assumption is that OOD inputs are expected to receive lower scores than ID cases, as they lie outside the training label space. However, OOD detection remains a non-trivial task because deep models may behave overconfidently when presented with OOD data (Nguyen, Yosinski, and Clune 2015).

3 Proposed Method

3.1 Defining the Parameter Contribution

For a given input $\mathbf{x}$, the contribution $c_k(\mathbf{x}; \boldsymbol{\theta}_{ij})$ of a specific parameter $\boldsymbol{\theta}_{ij}$ to class k is defined as the change in the model's k-th output when $\boldsymbol{\theta}_{ij}$ is present versus when it is absent (*i.e.*, setting $\boldsymbol{\theta}_{ij}$ to 0) (Xu and Yang 2025), formally expressed as:

$$c_k(\mathbf{x}; \boldsymbol{\theta}_{ij}) = f_k(\mathbf{x}) - f_k(\mathbf{x}; \boldsymbol{\theta}_{ij} = 0). \tag{4}$$

Consistent with prior work (Sun and Li 2022; Chen et al. 2023; Xu and Yang 2025), we primarily focus on the contribution of the parameters that are directly responsible for making predictions, *i.e.*, the classifier weight matrix $\mathbf{W}$. This focus is motivated by two main considerations: (1) the adjustment of the deepest classifier layer is most effective for OOD detection, as the high-level semantics captured by the classifier layer are generally the most relevant and impactful for identifying OOD samples (Zhu et al. 2022; Xu and Yang 2025), and (2) the per-parameter contribution of the classifier layer is computationally efficient, since the contribution of the element $\mathbf{W}_{ij}$ to the k-th class can be expressed in a simplified form (see Appendix B for details):

$$c_k(\mathbf{x}; \mathbf{W}_{ij}) = \begin{cases} \mathbf{W}_{ij} \cdot h_i(\mathbf{x}), & \text{if } k = j, \\ 0, & \text{if } k \neq j, \end{cases} \tag{5}$$

where each $\mathbf{W}_{ij}$ is exclusively responsible for the corresponding class j. Accordingly, the model's k-th output $f_k(\mathbf{x})$ can be rewritten in terms of the classifier's weight parameter contributions as:

$$f_k(\mathbf{x}) = \sum_{d=1}^{D} c_k(\mathbf{x}; \mathbf{W}_{dk}) + \mathbf{b}_k. \tag{6}$$

3.2 Shaping Parameter Contribution Patterns

Motivation. As illustrated in Figure 1, the classifier layer typically relies on sparse contribution patterns when making predictions. That is, the model's decisions are driven by a small subset of parameters that exert disproportionately high influence. This brittleness can be exploited by OOD inputs that anomalously trigger these parameters (Nguyen, Yosinski, and Clune 2015; Sun, Guo, and Li 2021), resulting in overconfident predictions. This critical issue motivates us to propose SPCP, which encourages the model to learn a more robust contribution pattern for improved OOD detection.

Training Procedure. SPCP explicitly imposes an upper bound on the contributions during the training process, applied element-wise to the weight parameters $\mathbf{W}$ of the model's classifier layer:

$$\bar{c}_k^{\lambda}(\mathbf{x}; \mathbf{W}_{ij}) = \min(c_k(\mathbf{x}; \mathbf{W}_{ij}), \lambda), \tag{7}$$

where λ denotes the threshold. As $\lambda \to \infty$, Eq. (7) becomes equivalent to the original unbounded formulation. In effect, this operation truncates the contribution above λ to prevent the classifier's parameters from producing disproportionately large contributions. In this context, the model output for the k-th class after applying SPCP is given by:

$$f_k^{\text{SPCP}}(\mathbf{x}; \lambda) = \sum_{d=1}^{D} \bar{c}_k^{\lambda}(\mathbf{x}; \mathbf{W}_{dk}) + \mathbf{b}_k. \tag{8}$$

Similarly to previous methods (Sun, Guo, and Li 2021; Chen et al. 2023), we obtain the threshold λ by using a percentile ρ. Let $C(\mathbf{x})$ denote the contribution matrix *w.r.t.* the classifier weights $\mathbf{W}$ for a given sample $\mathbf{x}$. The threshold λ is set to the values corresponding to the top ρ-th percentile of $C(\mathbf{x})$, averaged over the entire training set $\mathcal{D}_{\text{train}}$. To adapt λ to the dynamic behavior of training while adhering to mini-batch processing, we apply an Exponential Moving Average (EMA) for estimation at each iteration t:

$$\lambda_{t+1} = \beta \cdot \lambda_t + (1 - \beta) \cdot \frac{1}{|\mathcal{B}_t|} \sum_{\mathbf{x}_i \in \mathcal{B}_t} \text{Top}(\rho, C(\mathbf{x}_i)), \tag{9}$$

where $\beta \in [0, 1]$ is the smoothing factor controlling the update rate, $\mathcal{B}_t$ denotes the mini-batch of data at the t-th iteration, $|\mathcal{B}_t|$ denotes the cardinality of the set $\mathcal{B}_t$, and $\text{Top}(\rho, C(\mathbf{x}))$ refers to the value corresponding to the top ρ-th percentile of the contribution matrix $C(\mathbf{x})$ for the given sample $\mathbf{x}$. To stabilize early training, λ_0 is intentionally initialized with a relatively large value for warm-up. The learning objective, based on the cross-entropy loss ℓ_{CE}, is to minimize the following expected risk:

$$\mathcal{R}(f^{\text{SPCP}}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{\text{train}}} \ell_{\text{CE}}(f^{\text{SPCP}}(\mathbf{x}; \lambda), \mathbf{y}). \tag{10}$$

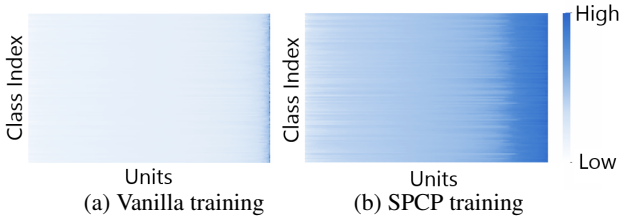
Inferring Procedure. SPCP adopts the Energy Score (Liu et al. 2020) as the default OOD scoring function, as it is provably aligned with the input density and generally performs well. Given a test input $\mathbf{x}$, the predictive uncertainty of the model after applying SPCP is quantified as:

$$S_{\text{SPCP}}(\mathbf{x}; f^{\text{SPCP}}) = \log \sum_{k\in\mathcal{Y}} \exp(f_k^{\text{SPCP}}(\mathbf{x}; \lambda)). \tag{11}$$

To ensure consistency, λ is set to the value estimated from Eq. (9) at the point of training completion. The pseudo code of SPCP is available in Appendix K.

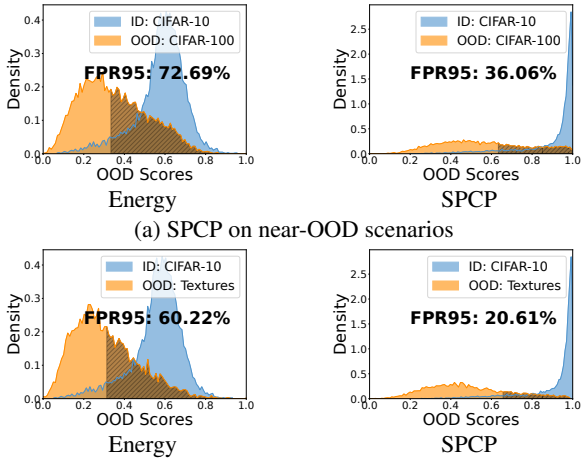
3.3 Insight Justification

Remark 1. SPCP effectively prevents predictions from being dominated by a small subset of classifier's parameters. In modern deep learning models, many design choices drive sparse contribution patterns, including over-parameterization (Sun and Li 2022), regularization techniques (*e.g.*, L_1 regularization (Tibshirani 1996)), and sparsity-inducing activations (*e.g.*, ReLU (Agarap 2018)). Such sparsity causes a small subset of parameters to dominate the prediction disproportionately. In response, our proposed SPCP explicitly enforces an upper bound λ on the contributions of the classifier's parameters. This restriction limits their excessive influence on predictions during training and encourages the adoption of relatively denser contribution patterns. As shown in Figure 2, the model's decisions are based on a broader set of parameters after SPCP training, effectively mitigating the prediction from being dominated by only a few parameters.



(a) Vanilla training (b) SPCP training

Figure 2: Comparison of the classifier's parameter contribution patterns before and after applying SPCP, with the average contribution matrix on the ID test set sorted for clarity.



(a) SPCP on near-OOD scenarios

(b) SPCP on far-OOD scenarios

Figure 3: Comparison of normalized OOD score distributions before and after applying SPCP.

Remark 2. SPCP mitigates overconfidence by shaping boundary-oriented dense contribution patterns. Due to deep models that may activate neurons even for unfamiliar inputs (Nguyen, Yosinski, and Clune 2015; Sun, Guo, and Li 2021), the behavior of dominant parameters may pose significant risks under sparse contribution patterns. Specifically, anomalous triggering of dominant parameters can drive the prediction toward incorrect ID categories with high confidence, resulting in overconfident predictions. To address this issue, our proposed SPCP guides the model to learn boundary-oriented dense contribution patterns, thereby enhancing the model's resilience to anomalous parameter contributions induced by OOD inputs. As shown in Figure 3,

ID datasets		CIFAR-10		CIFAR-100	
OOD datasets		FPR95↓	AUROC↑	FPR95↓	AUROC↑
		Vanilla training / **SPCP training (Ours)**			
Near-OOD	CIFAR-10	–	–	**59.21**±0.75 / 60.10±0.95	79.05±0.11 / **79.09**±0.22
	CIFAR-100	66.60±4.46 / **35.74**±0.51	86.36±0.58 / **90.42**±0.17	–	–
	TIN	56.08±4.83 / **27.59**±0.48	88.80±0.36 / **92.72**±0.07	52.03±0.50 / **50.33**±0.85	82.76±0.08 / **83.40**±0.28
	Average	61.34±4.63 / **31.67**±0.25	87.58±0.46 / **91.57**±0.11	55.62±0.61 / **55.21**±0.85	80.91±0.08 / **81.25**±0.19
Far-OOD	MNIST	24.99±12.93 / **19.09**±1.14	94.32±2.53 / **94.76**±0.69	52.62±3.83 / **49.33**±3.55	79.18±1.37 / **80.98**±2.70
	SVHN	35.12±6.11 / **16.52**±2.74	91.79±0.98 / **95.35**±1.33	53.62±3.14 / **43.00**±6.73	82.03±1.74 / **86.02**±3.60
	Textures	51.82±6.11 / **20.09**±0.57	89.47±0.70 / **95.00**±0.03	62.35±2.06 / **54.87**±1.21	78.35±0.83 / **81.32**±0.28
	Places365	54.85±6.52 / **26.47**±0.54	89.25±0.78 / **93.09**±0.12	57.75±0.86 / **55.16**±0.61	79.52±0.23 / **80.37**±0.38
	Average	41.69±5.32 / **20.54**±0.67	91.21±0.92 / **94.55**±0.42	56.59±1.38 / **50.59**±1.83	79.77±0.61 / **82.17**±1.32

Table 1: OOD detection performance on CIFAR benchmarks. All values are percentages, and the best results are in bold. ↑ indicates that larger values are better, ↓ indicates that smaller values are better.

ID datasets		ImageNet-200	
OOD datasets		FPR95↓	AUROC↑
		Vanilla training / **SPCP training (Ours)**	
Near-OOD	SSB-hard	**69.77**±0.32 / 70.23±0.21	**79.83**±0.02 / 79.80±0.05
	NINCO	50.70±0.89 / **49.32**±0.40	85.17±0.11 / **85.24**±0.06
	Average	60.24±0.57 / **59.77**±0.25	82.50±0.05 / **82.52**±0.03
Far-OOD	iNaturalist	26.41±2.29 / **23.86**±0.94	92.55±0.50 / **92.91**±0.08
	Textures	41.43±1.85 / **33.83**±0.69	90.79±0.16 / **91.73**±0.21
	OpenImage-O	36.74±1.14 / **33.59**±0.34	89.23±0.26 / **89.75**±0.16
	Average	34.86±1.30 / **30.43**±0.45	90.86±0.21 / **91.46**±0.14

Table 2: OOD detection performance on large-scale ImageNet benchmark.

SPCP reduces the overlap in the right tails of the OOD score distributions between OOD and ID samples, achieving improved ID-OOD separation in both near-OOD and far-OOD scenarios. These results demonstrate that SPCP effectively mitigates the model's overconfidence on OOD inputs.

4 Experiments

4.1 Experimental Setup

Our evaluation is based on the standard practice of the OpenOOD v1.5 benchmark (Zhang et al. 2023b), which includes both small-scale CIFAR benchmarks and large-scale ImageNet benchmark, spanning near-OOD scenarios with semantic shifts and far-OOD scenarios with further obvious covariance shifts [1].

Datasets. The setup for the small-scale experiment uses CIFAR-10/100 (Krizhevsky 2009) as the ID dataset. Evaluations cover near-OOD group, including CIFAR-100/10 and TinyImageNet (TIN) (Le and Yang 2015), as well as far-OOD group, including MNIST (Deng 2012), SVHN (Netzer et al. 2011), Textures (Cimpoi et al. 2014), and Places365 (Zhou et al. 2018).

For the large-scale experimental setup, a 200-class subset of ImageNet-1K (Deng et al. 2009), referred to as ImageNet-200, is adopted as the ID dataset. The near-OOD group includes SSB-hard (Vaze et al. 2022) and NINCO (Bitterwolf, Müller, and Hein 2023), while the far-OOD group contains iNaturalist (Horn et al. 2018), Textures (Cimpoi et al. 2014), and OpenImage-O (Wang et al. 2022).

Evaluation Metrics. We report the following three widely adopted metrics for comparison: (1) FPR95, the false positive rate of OOD data at a 95% true positive rate of ID data; (2) AUROC, the area under the receiver operating characteristic curve; and (3) ID ACC, the classification accuracy on the ID test set. The evaluation adopts the same number of independent runs and random seed settings as in OpenOOD (Zhang et al. 2023b), with results reported as the mean and standard deviation.

Baselines. We compare SPCP with a broad range of competitive baselines: (1) post-hoc methods, including OOD scoring methods: MSP (Hendrycks and Gimpel 2017), Energy (Liu et al. 2020); and network adjustment methods: DICE (Sun and Li 2022), ReAct (Sun, Guo, and Li 2021), ASH (Djurisic et al. 2023), SCALE (Xu et al. 2024); and (2) training-time regularization methods: LogitNorm (Wei et al. 2022), CIDER (Ming et al. 2023), UMAP (Zhu et al. 2023), SNN (Ghosal, Sun, and Li 2024) and T2FNorm (Regmi et al. 2024). Baseline results are sourced from the authoritative OpenOOD leaderboard (Zhang et al. 2023b).

Implementation Details. Our implementation strictly fol-

[1]Code is available at https://github.com/njustkmg/AAAI2026-SPCP

Method	CIFAR-10					CIFAR-100				
	Near-OOD		Far-OOD		ID ACC↑	Near-OOD		Far-OOD		ID ACC↑
	FPR95↓	AUROC↑	FPR95↓	AUROC↑		FPR95↓	AUROC↑	FPR95↓	AUROC↑	
Post-hoc methods (vanilla training with cross-entropy)										
MSP (Hendrycks and Gimpel 2017)	48.17±3.92	88.03±0.25	31.72±1.84	90.73±0.43	95.06±0.30	**54.80±0.33**	80.27±0.11	58.70±1.06	77.76±0.44	77.25±0.10
Energy (Liu et al. 2020)	61.34±4.63	87.58±0.46	41.69±5.32	91.21±0.92	95.06±0.30	55.62±0.61	80.91±0.08	56.59±1.38	79.77±0.61	77.25±0.10
DICE (Sun and Li 2022)	70.04±7.64	78.34±0.79	51.76±4.42	84.23±1.89	95.06±0.30	57.95±0.53	79.38±0.23	56.25±0.60	80.01±0.18	77.25±0.10
ReAct (Sun, Guo, and Li 2021)	63.56±7.33	87.11±0.61	44.90±8.37	90.42±1.41	95.06±0.30	56.39±0.34	80.77±0.05	54.20±1.56	80.39±0.49	77.25±0.10
ASH (Djurisic et al. 2023)	86.78±1.82	75.27±1.04	79.03±4.22	78.49±2.58	95.06±0.30	65.71±0.24	78.20±0.15	59.20±2.46	80.58±0.66	77.25±0.10
SCALE (Xu et al. 2024)	80.45±4.02	82.55±0.36	67.53±7.50	86.39±1.86	95.06±0.30	55.68±0.69	80.99±0.12	54.09±1.07	81.42±0.43	77.25±0.10
Training-time regularization methods										
LogitNorm (Wei et al. 2022)	29.34±0.81	92.33±0.08	13.81±0.20	96.74±0.06	94.30±0.25	62.89±0.57	78.47±0.31	53.61±3.45	81.53±1.26	76.34±0.17
CIDER (Ming et al. 2023)	32.11±0.94	90.71±0.16	20.72±0.85	94.71±0.36	–	72.02±0.31	73.10±0.39	54.22±1.24	80.49±0.68	–
UMAP (Zhu et al. 2023)	33.12±0.06	91.00±0.07	21.70±1.57	94.20±0.36	95.06±0.30	59.71±0.65	79.49±0.23	52.11±2.36	81.62±1.37	77.25±0.10
SNN (Ghosal, Sun, and Li 2024)	37.21±0.70	90.25±0.09	26.05±2.34	92.49±0.78	**95.11±0.13**	60.32±1.44	80.33±0.22	53.52±1.77	82.17±0.69	77.56±0.27
T2FNorm (Regmi et al. 2024)	**26.47±0.35**	**92.79±0.13**	**12.75±0.73**	**96.98±0.23**	94.69±0.07	58.47±1.35	79.84±0.40	51.25±2.52	**82.73±1.01**	76.43±0.13
SPCP	31.67±0.25	91.57±0.11	20.54±0.67	94.55±0.42	94.91±0.27	55.21±0.85	**81.25±0.19**	**50.59±1.83**	82.17±1.32	**77.70±0.20**

Table 3: Comparison on CIFAR benchmarks. All values are percentages, and OOD detection results are averaged over multiple OOD datasets. Detailed results for each OOD dataset are provided in Appendix D. The best results are in **bold**, with the second and third best results underlined.

Method	Near-OOD		Far-OOD		ID ACC↑
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
Post-hoc methods (vanilla training with cross-entropy)					
MSP (Hendrycks and Gimpel 2017)	**54.82±0.35**	83.34±0.06	35.43±0.38	90.13±0.09	86.37±0.08
Energy (Liu et al. 2020)	60.24±0.57	82.50±0.05	34.86±1.30	90.86±0.21	86.37±0.08
DICE (Sun and Li 2022)	61.88±0.67	81.78±0.14	36.51±1.18	90.80±0.31	86.37±0.08
ReAct (Sun, Guo, and Li 2021)	62.49±2.19	81.87±0.98	28.50±0.95	92.31±0.56	86.37±0.08
ASH (Djurisic et al. 2023)	64.89±0.90	82.38±0.19	27.29±1.12	93.90±0.27	86.37±0.08
SCALE (Xu et al. 2024)	57.29±0.90	**84.84±0.28**	26.46±0.81	93.98±0.25	86.37±0.08
Training-time regularization methods					
LogitNorm (Wei et al. 2022)	56.46±0.37	82.66±0.15	26.11±0.52	93.04±0.21	86.04±0.15
CIDER (Ming et al. 2023)	60.10±0.73	80.58±1.75	30.17±2.75	90.66±1.68	–
UMAP (Zhu et al. 2023)	60.81±0.84	81.08±0.39	32.47±0.67	91.62±0.29	86.37±0.08
SNN (Ghosal, Sun, and Li 2024)	59.85±0.46	81.33±0.19	28.04±0.64	92.28±0.21	86.56±0.03
T2FNorm (Regmi et al. 2024)	55.01±0.36	83.00±0.07	25.37±0.55	93.55±0.17	**86.87±0.19**
SPCP	59.77±0.25	82.52±0.03	30.43±0.45	91.46±0.14	86.59±0.10
LogitNorm+SPCP	55.33±0.45	83.20±0.07	**21.95±0.73**	**94.11±0.28**	86.37±0.09

Table 4: Comparison on large-scale ImageNet benchmark.

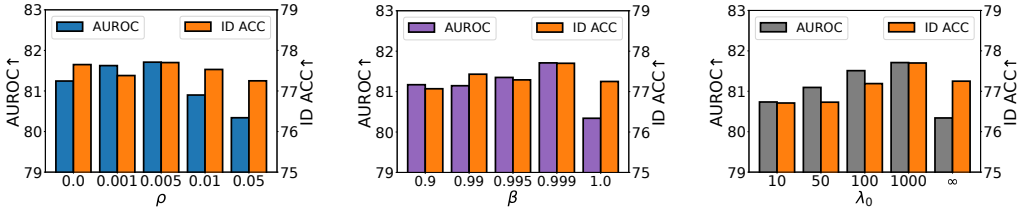
lows the OpenOOD benchmark (Zhang et al. 2023b). Specifically, ResNet-18 (He et al. 2016) is employed as the backbone architecture. The models are trained for 100 epochs using stochastic gradient descent (SGD) with a learning rate of 0.1, following a cosine annealing decay schedule (Loshchilov and Hutter 2017), a momentum of 0.9, and a weight decay of 5×10^{-4}. The batch size is set to 128 for the CIFAR experiment and 256 for the ImageNet experiment. More details are available in Appendix C.

4.2 Main Results

In this section, we report the performance of SPCP on the CIFAR benchmarks as well as on the more realistic and challenging ImageNet benchmark. Specifically, Tables 1 and 2 provide a fine-grained comparison of our SPCP with vanilla training, while Tables 3 and 4 show comparisons with other competitive methods. The results reveal that: (1) SPCP boosts OOD detection in nearly all cases while preserving ID performance. On the CIFAR-10 benchmark, SPCP outperforms the baseline by a large margin, reducing the average FPR95 by 29.67% in near-OOD settings and 21.25% in far-OOD settings. (2) In comparison to post-hoc methods, SPCP offers notable improvements by establishing more robust contribution patterns for OOD detection. Moreover, SPCP is complementary to various existing post-hoc methods to push their performance further, as will be discussed in Section 4.4. (3) Compared to a suite of training-time regularization methods, SPCP delivers competitive OOD detection performance and achieves promising results. It is noteworthy that no single method emerges as the definitive winner on the authoritative OpenOOD benchmark (Zhang et al. 2023b). Our SPCP achieves top or near-top performance in most OOD settings, indicating the effectiveness of our proposed shaping contribution pattern strategy.

Stage		Near-OOD		Far-OOD		ID ACC↑
Training	Inferring	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
×	×	55.62±0.61	80.91±0.08	56.59±1.38	79.77±0.61	77.25±0.10
✓	×	55.49±0.77	81.09±0.21	51.96±2.25	81.71±1.45	**77.70±0.24**
×	✓	60.28±0.46	79.50±0.08	53.66±3.13	80.59±1.10	76.94±0.12
✓	✓	**55.21±0.85**	**81.25±0.19**	**50.59±1.83**	**82.17±1.32**	**77.70±0.20**

Table 5: Ablation study of contribution truncation in different stages on the CIFAR-100 benchmark.



(a) Diff. Percentile for ρ (b) Diff. EMA Factor β (c) Diff. Initial Threshold λ_0

Figure 4: Ablation studies on the hyperparameters: (a) effect of varying the percentile ρ used for threshold estimation; (b) impact of the EMA smoothing factor β; (c) influence of initial threshold λ_0. The OOD results are averaged over both near- and far-OOD groups on the CIFAR-100 benchmark.

4.3 Ablation Study

Effects of Truncating Contributions at Different Stages. Table 5 presents an ablation study on the impact of truncating contributions during training and inference. The results indicate that: (1) Imposing contribution truncation during training is more effective than doing so only during inference, highlighting the critical role of shaping contribution patterns in the training process. (2) Applying contribution truncation exclusively during inference helps mitigate overconfidence and leads to modest improvements in far-OOD scenarios. However, the absence of corresponding regularization during training can hurt performance in near-OOD settings and cause a slight degradation in ID task performance. (3) SPCP achieves contribution truncation in both the training and inference processes, which preserves pattern consistency and further drives performance improvements.

Effect of the Percentile ρ. In Figure 4a, we investigate the impact of varying the percentile ρ, which is used to determine the parameter contribution threshold λ in Eq. (9). Compared to the baseline (*i.e.*, $\rho = 0$), properly constraining the upper bound of parameter contributions with a suitable ρ improves OOD detection while maintaining generalization on ID tasks. However, an inappropriately small ρ may cause the model to rely excessively on numerous and overlapping parameter decisions, leading to inter-class conflicts and degraded performance.

Effect of the EMA Factor β. Figure 4b presents an analysis of the effect of the EMA smoothing factor as defined in Eq. (9). As shown, performance degrades when the dynamic updating strategy is disabled (*i.e.*, $\beta = 1.0$), highlighting the critical role of dynamically adjusting the threshold during training. However, a smaller β may lead to suboptimal performance due to unstable threshold estimation from rapid updates. Therefore, striking a moderate balance is more beneficial for performance improvement.

Effect of the Initial Threshold λ_0. Figure 4c explores the impact of different initial threshold values set for λ_0 in Eq. (9). The results indicate that a smaller initial value of λ_0 leads to excessive intervention before sufficient learning has occurred, thereby disrupting the early learning process. In contrast, a moderately large λ_0 serves as a form of warm-up and leads to performance improvements. Consequently, selecting a suitably large λ_0 is essential for facilitating effective early-stage learning and improving final performance.

4.4 Further Analysis

Generalizing to Different Backbones. In Table 6, we evaluate the generalization of SPCP to different backbones, including the widely used and lightweight ResNet-18 (He et al. 2016), the high-capacity WideResNet-28-10 (Zagoruyko and Komodakis 2016), and the densely connected DenseNet-101 (Huang et al. 2017). As shown in the table, our SPCP consistently improves FPR95 and AUROC across both near- and far-OOD scenarios on various backbones, while also preserving ID generalization. These results indicate that SPCP generalizes well across different backbones and effectively enhances OOD detection robustness.

Compatibility with Other OOD Detection Methods. To validate the compatibility of our SPCP, we integrate representative OOD detection methods spanning diverse information sources: probability space (MSP (Hendrycks and Gimpel 2017)), logits space (Energy (Liu et al. 2020)), gradient space (GradNorm (Huang, Geng, and Li 2021)), feature space (KNN (Sun et al. 2022)), penultimate activation manipulations (ASH (Djurisic et al. 2023)), and training-time regularization (LogitNorm (Wei et al. 2022)). The results in Table 7 show that SPCP consistently improves OOD detection performance across a range of methods, indicating that the contribution patterns shaped by SPCP are broadly applicable and compatible with various algorithmic paradigms.

Model	Method	Near-OOD		Far-OOD		ID ACC↑
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	
ResNet-18	Vanilla	55.62±0.61	80.91±0.08	56.59±1.38	79.77±0.61	77.25±0.10
	SPCP	**55.21**±0.85	**81.25**±0.19	**50.59**±1.83	**82.17**±1.32	**77.70**±0.20
WideResNet-28-10	Vanilla	54.48±1.30	82.31±0.41	55.34±1.67	80.87±0.80	80.53±0.13
	SPCP	**53.48**±0.52	**82.63**±0.24	**52.84**±0.93	**81.69**±0.31	**80.56**±0.22
DenseNet-101	Vanilla	61.57±1.36	79.27±0.36	65.49±2.45	76.16±0.81	76.54±0.42
	SPCP	**58.93**±1.14	**80.28**±0.48	**56.75**±2.62	**78.68**±1.62	**76.86**±0.29

Table 6: Generalization to different backbones on the CIFAR-100 benchmark.

Method	Near-OOD		Far-OOD		ID ACC↑
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
MSP (Hendrycks and Gimpel 2017)	**54.80**±0.33	80.27±0.11	58.70±1.06	77.76±0.44	77.25±0.10
MSP+SPCP	55.04±0.68	**80.41**±0.28	**54.31**±0.95	**80.22**±0.93	**77.70**±0.20
Energy (Liu et al. 2020)	55.62±0.61	80.91±0.08	56.59±1.38	79.77±0.61	77.25±0.10
Energy+SPCP	**55.21**±0.85	**81.25**±0.19	**50.59**±1.83	**82.17**±1.32	**77.70**±0.20
GradNorm (Huang, Geng, and Li 2021)	85.58±0.46	70.13±0.47	83.68±1.92	69.14±1.05	77.25±0.10
GradNorm+SPCP	**80.33**±1.28	**73.71**±0.71	**79.47**±2.60	**73.70**±1.69	**77.70**±0.20
KNN (Sun et al. 2022)	61.22±0.14	80.18±0.15	53.65±0.28	82.40±0.17	77.25±0.10
KNN+SPCP	**59.13**±0.46	**80.55**±0.07	**51.97**±0.97	**82.52**±0.79	**77.70**±0.20
ASH (Djurisic et al. 2023)	65.71±0.24	78.20±0.15	59.20±2.46	80.58±0.66	77.25±0.10
ASH+SPCP	**61.15**±0.18	**79.75**±0.13	**56.06**±1.33	**82.07**±0.29	**77.70**±0.20
LogitNorm (Wei et al. 2022)	62.89±0.57	78.47±0.31	53.61±3.45	81.53±1.26	**76.34**±0.17
LogitNorm+SPCP	**60.70**±0.77	**79.20**±0.31	**47.81**±1.16	**83.99**±0.66	75.94±0.20

Table 7: Compatibility with other OOD detection methods on the CIFAR-100 benchmark.

5 Related Work

Post-hoc Methods aim to provide suitable measures to indicate the likelihood that a given sample is OOD. Early methods for detecting OOD samples work by scoring the network outputs (Liang, Li, and Srikant 2018; Liu et al. 2020; Huang, Geng, and Li 2021; Hendrycks et al. 2022). For example, MSP (Hendrycks and Gimpel 2017) uses the maximum SoftMax score as an indicator. Based on the OOD score, various post-hoc network adjustment methods (Xu et al. 2023; Ahn, Park, and Kim 2023; Xu and Yang 2025) are proposed to further enhance OOD score reliability. For example, ReAct (Sun, Guo, and Li 2021) and SCALE (Xu et al. 2024) improve ID-OOD separability by rectifying and scaling penultimate-layer activations, respectively. In this work, we reveal that shaping the parameter contribution patterns can help reduce model overconfidence and boost the performance of most existing post-hoc methods.

Training-Time Regularization Methods aim to provide better discriminative representations for OOD detection by calibrating the model. One category of methods leverages outlier exposure (Hendrycks, Mazeika, and Dietterich 2019; Zhang et al. 2023a; Wang et al. 2023; Jiang et al. 2024) to help the model learn more robust decision boundaries. As a representative method, Hendrycks et al. (Hendrycks, Mazeika, and Dietterich 2019) propose enforcing a uniform predictive distribution for outlier data. Although effective, such methods often hurt ID task performance, and access to outlier data may not always be available. Another category tackles OOD detection by imposing favorable constraints during the training process (DeVries and Taylor 2018; Huang and Li 2021; Du et al. 2022; Yang and Xu 2025). For instance, SNN (Ghosal, Sun, and Li 2024) mitigates the curse-of-dimensionality issue by learning the most relevant subspace. LogitNorm (Wei et al. 2022) and T2FNorm (Regmi et al. 2024) offer a simple fix to cross-entropy loss by decoupling the impact of logits and feature norms, respectively. However, these methods overlook the classifier's tendency to develop a sparse contribution pattern, which can easily push the model to overconfidence when dominant parameters are spuriously triggered. Our methods alleviate this issue by shaping a more robust contribution pattern for OOD detection while preserving ID performance.

6 Conclusion

This paper investigates the underlying factors contributing to the brittleness of OOD detection from the perspective of parameter contribution patterns. We identify that sparse contribution patterns can increase the risk of overconfident predictions and hurt OOD detection. To mitigate this issue, we propose SPCP, a method that constrains the upper bound of parameter contributions during training, thereby promoting the development of dense and bounded contribution patterns. As a result, SPCP enhances the model's resilience to anomalous parameter contributions induced by OOD inputs. Extensive experiments across various OOD scenarios confirm the effectiveness and broad applicability of SPCP.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (2022YFF0712100), NSFC (62276131), the Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081), and the Research on the Teaching Reform of Artificial Intelligence General Education Courses in Jiangsu Undergraduate Universities.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *CoRR*, abs/2303.08774.

Agarap, A. F. 2018. Deep Learning using Rectified Linear Units (ReLU). *CoRR*, abs/1803.08375.

Ahn, Y. H.; Park, G.; and Kim, S. T. 2023. LINe: Out-of-Distribution Detection by Leveraging Important Neurons. In *CVPR*, 19852–19862.

Bendale, A.; and Boult, T. E. 2016. Towards Open Set Deep Networks. In *CVPR*, 1563–1572.

Bitterwolf, J.; Müller, M.; and Hein, M. 2023. In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In *ICML*, 2471–2506.

Chen, C.; Fu, Z.; Liu, K.; Chen, Z.; Tao, M.; and Ye, J. 2023. Optimal Parameter and Neuron Pruning for Out-of-Distribution Detection. In *NeurIPS*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Deng, L. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

DeVries, T.; and Taylor, G. W. 2018. Learning Confidence for Out-of-Distribution Detection in Neural Networks. *CoRR*, abs/1802.04865.

Djurisic, A.; Bozanic, N.; Ashok, A.; and Liu, R. 2023. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *ICLR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

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

Frankle, J.; and Carbin, M. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *ICLR*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 3354–3361.

Ghosal, S. S.; Sun, Y.; and Li, Y. 2024. How to Overcome Curse-of-Dimensionality for Out-of-Distribution Detection? In *AAAI*, 19849–19857.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Hendrycks, D.; Basart, S.; Mazeika, M.; Zou, A.; Kwon, J.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. In *ICML*, 8759–8773.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. G. 2019. Deep Anomaly Detection with Outlier Exposure. In *ICLR*.

Horn, G. V.; Aodha, O. M.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. J. 2018. The INaturalist Species Classification and Detection Dataset. In *CVPR*, 8769–8778.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269.

Huang, R.; Geng, A.; and Li, Y. 2021. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In *NeurIPS*, 677–689.

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space. In *CVPR*, 8710–8719.

Jiang, W.; Cheng, H.; Chen, M.; Wang, C.; and Wei, H. 2024. DOS: Diverse Outlier Sampling for Out-of-Distribution Detection. In *ICLR*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *ICLR*.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88.

Liu, W.; Wang, X.; Owens, J. D.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *NeurIPS*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*, 9992–10002.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*.

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *ICLR*.

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

Nguyen, A. M.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 427–436.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Regmi, S.; Panthi, B.; Dotel, S.; Gyawali, P. K.; Stoyanov, D.; and Bhattarai, B. 2024. T2FNorm: Train-time Feature Normalization for OOD Detection in Image Classification. In *CVPRW*, 153–162.

Sun, Y.; Guo, C.; and Li, Y. 2021. ReAct: Out-of-distribution Detection With Rectified Activations. In *NeurIPS*, 144–157.

Sun, Y.; and Li, Y. 2022. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *ECCV*, volume 13684, 691–708.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. In *ICML*, 20827–20840.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1): 267–288.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In *ICLR*.

Wan, F.; and Yang, Y. 2025. Probabilistic Group Mask Guided Discrete Optimization for Incremental Learning. In *ICML*.

Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022. ViM: Out-Of-Distribution with Virtual-logit Matching. In *CVPR*, 4911–4920.

Wang, Q.; Ye, J.; Liu, F.; Dai, Q.; Kalander, M.; Liu, T.; Hao, J.; and Han, B. 2023. Out-of-distribution Detection with Implicit Outlier Transformation. In *ICLR*.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating Neural Network Overconfidence with Logit Normalization. In *ICML*, 23631–23644.

Xu, H.; and Yang, Y. 2025. ITP: Instance-Aware Test Pruning for Out-of-Distribution Detection. In *AAAI*, 21743–21751.

Xu, K.; Chen, R.; Franchi, G.; and Yao, A. 2024. Scaling for Training Time and Post-hoc Out-of-distribution Detection Enhancement. In *ICLR*.

Xu, M.; Lian, Z.; Liu, B.; and Tao, J. 2023. VRA: Variational Rectified Activation for Out-of-distribution Detection. In *NeurIPS*.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2021. Generalized Out-of-Distribution Detection: A Survey. *CoRR*, abs/2110.11334.

Yang, Y.; Wan, F.; Jiang, Q.; and Xu, Y. 2024. Facilitating Multimodal Classification via Dynamically Learning Modality Gap. In *NeurIPS*.

Yang, Y.; and Xu, H. 2025. Strengthen Out-of-Distribution Detection Capability with Progressive Self-Knowledge Distillation. In *ICML*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *BMVC*.

Zhang, J.; Inkawhich, N.; Linderman, R.; Chen, Y.; and Li, H. 2023a. Mixture Outlier Exposure: Towards Out-of-Distribution Detection in Fine-grained Environments. In *WACV*, 5520–5529.

Zhang, J.; Yang, J.; Wang, P.; Wang, H.; Lin, Y.; Zhang, H.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Li, Y.; Liu, Z.; Chen, Y.; and Li, H. 2023b. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *CoRR*, abs/2306.09301.

Zhang, Y.; Lan, Z.; Dai, Y.; Zeng, F.; Bai, Y.; Chang, J.; and Wei, Y. 2020. Prime-Aware Adaptive Distillation. In *ECCV*, volume 12364, 658–674.

Zhang, Y.; Lu, J.; Peng, B.; Fang, Z.; and Cheung, Y. 2024. Learning to Shape In-distribution Feature Space for Out-of-distribution Detection. In *NeurIPS*.

Zhou, B.; Lapedriza, À.; 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*.

Zhu, J.; Li, H.; Yao, J.; Liu, T.; Xu, J.; and Han, B. 2023. Unleashing Mask: Explore the Intrinsic Out-of-Distribution Detection Capability. In *ICML*, 43068–43104.

Zhu, Y.; Chen, Y.; Xie, C.; Li, X.; Zhang, R.; Xue, H.; Tian, X.; Zheng, B.; and Chen, Y. 2022. Boosting Out-of-distribution Detection with Typical Features. In *NeurIPS*.