ITP: Instance-Aware Test Pruning for Out-of-Distribution Detection

Haonan Xu[1], **Yang Yang**[1,2*]

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

Abstract

Out-of-distribution (OOD) detection is crucial for ensuring the reliable deployment of deep models in real-world scenarios. Recently, from the perspective of over-parameterization, a series of methods leveraging weight sparsification techniques have shown promising performance. These methods typically focus on selecting important parameters for in-distribution (ID) data to reduce the negative impact of redundant parameters on OOD detection. However, we empirically find that these selected parameters may behave overconfidently toward OOD data and hurt OOD detection. To address this issue, we propose a simple yet effective post-hoc method called Instance-aware Test Pruning (**ITP**), which performs OOD detection by considering both coarse-grained and fine-grained levels of parameter pruning. Specifically, ITP first estimates the class-specific parameter contribution distribution by exploring the ID data. By using the contribution distribution, ITP conducts coarse-grained pruning to eliminate redundant parameters. More importantly, ITP further adopts a fine-grained test pruning process based on the right-tailed Z-score test, which can adaptively remove instance-level overconfident parameters. Finally, ITP derives OOD scores from the pruned model to achieve more reliable predictions. Extensive experiments on widely adopted benchmarks verify the effectiveness of ITP, demonstrating its competitive performance.

1 Introduction

Deep neural networks (DNNs) have recently achieved remarkable success, driving significant progress in various fields, particularly in computer vision (Dosovitskiy et al. 2021; Yang et al. 2021, 2023a) and natural language processing (Radford et al. 2018; Achiam et al. 2023). However, when deployed in open-world environments, DNNs may fail by producing confident yet erroneous predictions for OOD data. Such unreliable behavior could lead to disastrous consequences, particularly in safety-critical fields like autonomous driving (Geiger, Lenz, and Urtasun 2012) and medical diagnosis (Litjens et al. 2017). Therefore, detecting and rejecting predictions on OOD inputs is crucial for ensuring the reliability of AI systems. This task, referred to as OOD detection, has gained widespread attention.

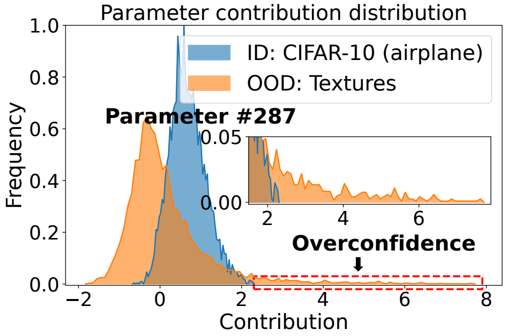


Figure 1: The distribution of parameter contributions to ID prediction for the CIFAR-10 class ('airplane') on both ID and OOD data. The parameter is selected from the subset of weight parameters that are important for ID prediction in the last layer of DenseNet-101, with pre-ReLU activations utilized for visualization. Since the model outputs are determined by the parameter contribution, overconfident behaviors in the parameters increase the risk of misclassifying OOD data as ID and hurt OOD detection.

Many techniques have been developed for OOD detection to enhance the discrimination between ID and OOD data. The training-based methods (Malinin and Gales 2018; Monteiro et al. 2023; Ghosal, Sun, and Li 2024) necessitate model training or fine-tuning, whereas post-hoc methods (Bendale and Boult 2016; Sun et al. 2022; Wang et al. 2022; Yang et al. 2023b, 2024) can be applied directly to pre-trained models off-the-shelf, eliminating the need for retraining process. This paper primarily focuses on post-hoc methods, which are easy to use, low-cost, and generally applicable. Early post-hoc methods, such as MSP (Hendrycks and Gimpel 2017), Energy (Liu et al. 2020), and GradNorm (Huang, Geng, and Li 2021), focus on devising a suitable OOD scoring function based on model outputs or gradients to indicate the likelihood that a sample originates from the OOD distribution. However, these methods overlook the fact that DNNs are typically over-parameterized to fit complex data distributions. This design makes the model more susceptible to noise from redundant parameters (Sun and Li 2022), leading to the brittleness of OOD detection.

To address this issue, a series of post-hoc methods that use

[*]Corresponding author

network adjustments to improve OOD scoring has emerged. Sparsification-based methods, represented by DICE (Sun and Li 2022) and LINe (Ahn, Park, and Kim 2023), provide effective solutions for model over-parameterization. Their key idea is to selectively use the weight parameters that are important for ID data to derive OOD scores, thereby reducing the noise interference caused by redundant parameters in OOD detection. However, as illustrated in Figure 1, our empirical findings reveal that these selected important parameters are not always beneficial for OOD detection. When processing OOD data, these parameters may contribute abnormally high to ID predictions, behaving overconfidently. Such overconfidence increases the risk of the model predicting OOD data as ID categories with high confidence. As a result, the derived OOD scores become unreliable, leading to confusion between ID and OOD data and hindering OOD detection. Therefore, addressing parameter-level overconfidence is crucial for better separating ID and OOD data.

Targeting this important problem, we propose Instance-aware Test Pruning (ITP), a simple yet effective method for OOD detection that considers parameter pruning from both coarse-grained and fine-grained perspectives. Concretely, ITP first estimates the class-specific parameter contribution distribution by exploring the ID data. Then, by leveraging the contribution distribution, ITP considers the following two key points to improve OOD detection performance: (1) ITP conducts coarse-grained pruning to remove noise interference caused by redundant parameters based on DICE (Sun and Li 2022). (2) ITP adopts a fine-grained test pruning process based on the right-tailed Z-score test, which adaptively removes instance-level overconfident parameters to reduce the risk of the model making confident yet erroneous predictions. We further provide insightful justification of the working mechanism of ITP from the perspective of the OOD score distribution. As a result of ITP, we show that the OOD scores derived from the model are more reliable and become more separable between ID and OOD data. Moreover, by examining parameter behavior in the weight space, ITP operates orthogonally to activation-based OOD detection methods (e.g., ReAct (Sun, Guo, and Li 2021)), facilitating their integration to push ITP's performance further.

2 Related Work

OOD detection aims to enable models to identify and reject predictions for OOD inputs, thereby ensuring the reliability of AI systems. We highlight three major lines of work.

OOD Scoring Methods are designed to provide appropriate criteria for indicating the likelihood that an input sample is OOD. Distance-based (Lee et al. 2018; Huang et al. 2021; Sun et al. 2022) methods identify OOD data as being farther from the training set compared to ID data. Gradient-based methods (Huang, Geng, and Li 2021; Behpour et al. 2023) detect OOD inputs by utilizing information extracted from the gradient space. Output-based methods rely on model output logits to identify OOD data. MSP (Hendrycks and Gimpel 2017) directly uses the maximum SoftMax score to classify a test sample as either ID or OOD. ODIN (Liang, Li, and Srikant 2018) improves the MSP score by perturbing the input and applying temperature scaling to the logits. Energy score (Liu et al. 2020) uses the logsumexp of the output logits, which is consistent with input density and less susceptible to overconfidence problems. However, output-based methods are often disrupted by redundant or overconfident parameters, which can negatively impact OOD detection.

Sparsification-Based Methods perform OOD detection by pruning the weights of the model. DICE (Sun and Li 2022) proposes selectively using the most salient weights to derive the output for OOD detection. LINe (Ahn, Park, and Kim 2023) adopts the Shapley value (Shapley et al. 1953) for more precise pruning of redundant parameters and neurons, and it further considers the number of activated features by clipping activations. OPNP (Chen et al. 2023) prunes the parameters and neurons with exceptionally large or nearly zero sensitivities to mitigate over-fitting. These methods typically focus on selecting parameters that are important for ID prediction before testing for OOD detection. However, at test time, these selected parameters may exhibit overconfidence, which can impact the performance of OOD detection.

Activation-Based Methods attempt to rectify activations to widen the gap between ID and OOD data. ReAct (Sun, Guo, and Li 2021) truncates activations above a pre-computed threshold to treat all activated features equally, thereby incorporating the number of activated features into consideration for OOD detection. VRA (Xu et al. 2023) zeros out anomalously low activations and truncates anomalously high activations. BATS (Zhu et al. 2022) proposes rectifying activations towards their typical set, while LAPS (He et al. 2024) improves BATS by considering channel-aware typical sets. These methods only examine anomalies at the activation level, whereas managing overconfidence anomalies at a more granular parameter level is important for more effective OOD detection.

3 Proposed Method

3.1 Preliminaries

Setup. In this paper, we follow previous work (Yang et al. 2022) and focus on the setting of K-way image classification. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, 2, ..., K\}$ be the ID label space. Suppose that the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is drawn *i.i.d* from a joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ defined over $\mathcal{X} \times \mathcal{Y}$. We denote $\mathcal{P}_{in}$ as the marginal distribution of $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ on $\mathcal{X}$, representing the ID distribution.

Let f be a model pre-trained from $\mathcal{D}$. For typical image classification 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 last fully connected (FC) 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 $f(\mathbf{x}) \in \mathbb{R}^K$. Mathematically, the model output can be expressed as:

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

Out-of-distribution Detection. The goal of OOD detection is to determine whether a test input $\mathbf{x}$ is from $\mathcal{P}_{in}$ (ID) or not (OOD). In practice, the OOD detection task is often formu-

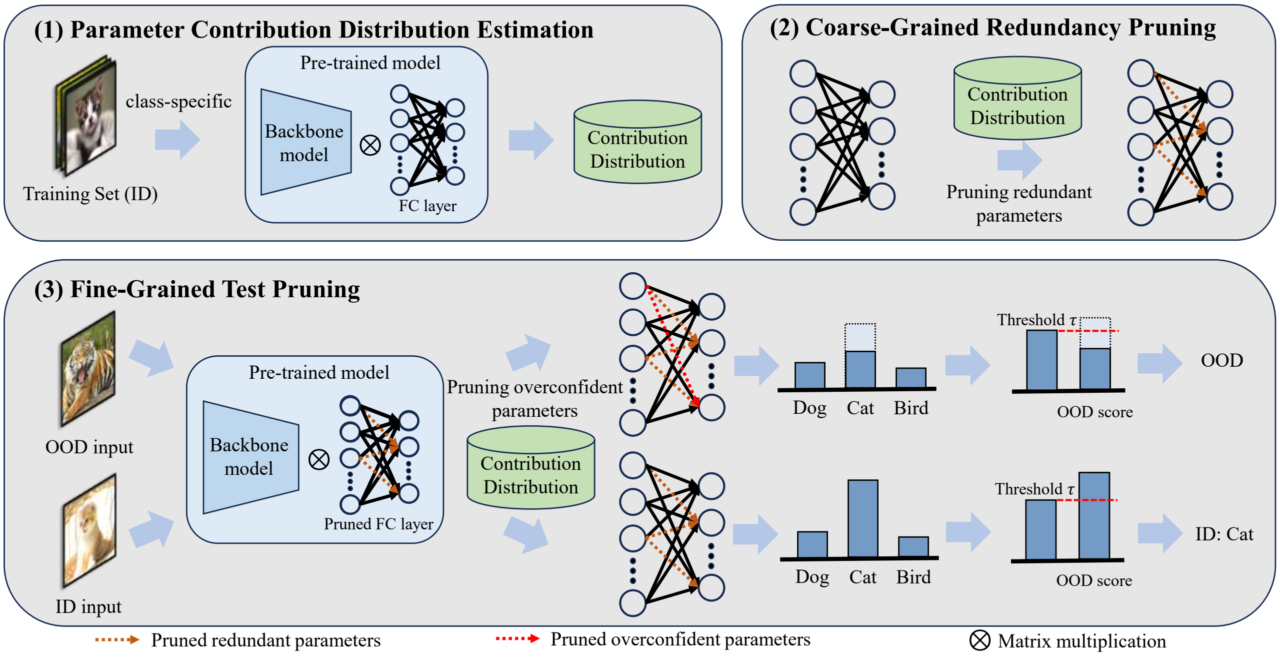

(1) Parameter Contribution Distribution Estimation
class-specific
Training Set (ID)
Pre-trained model
Backbone model
FC layer
Contribution Distribution
(2) Coarse-Grained Redundancy Pruning
Contribution Distribution
Pruning redundant parameters
(3) Fine-Grained Test Pruning
OOD input
ID input
Pre-trained model
Backbone model
Pruned FC layer
Pruning overconfident parameters
Contribution Distribution
Dog
Cat
Bird
Threshold τ
OOD score
OOD
Dog
Cat
Bird
Threshold τ
OOD score
ID: Cat
Pruned redundant parameters
Pruned overconfident parameters
⊗ Matrix multiplication

Figure 2: Illustration of OOD detection using ITP. The overall procedure involves three main steps. (1) Training data are used to estimate the class-specific parameter contribution distribution for a pre-trained model. (2) Coarse-grained redundancy pruning applies a fixed pruning pattern to the model's last layer to remove redundant parameters. (3) Fine-grained test pruning applies a customized pruning pattern to remove overconfident parameters for each test sample at test time. After applying ITP, the OOD scores derived from the model are better able to distinguish between ID and OOD data.

lated as the following binary decision problem:

$$G(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } S(\mathbf{x}) > \tau, \\ \text{OOD}, & \text{if } S(\mathbf{x}) \leq \tau, \end{cases} \tag{2}$$

where $S(\cdot)$ represents the OOD scoring function, and τ is a chosen threshold to ensure that the majority of ID data are correctly classified (*e.g.*, 95%). By convention, samples with higher OOD scores are heuristically classified as ID and vice versa. Given that the energy score (Liu et al. 2020) has been proven to be consistent with the input density and performs well, we mainly adopt the negative energy score as the OOD score, expressed as:

$$S(\mathbf{x}) = -E\left(\mathbf{x}\right) = \log \sum_{k=1}^{K} \exp(f_k(\mathbf{x})), \tag{3}$$

where $E(\mathbf{x})$ denotes the energy of $\mathbf{x}$, and $f_k(\mathbf{x})$ represents the k-th output of the model.

3.2 Parameter Contribution Distribution Estimation

Figure 2 illustrates the overall procedure of our proposal. In this section, we first provide a detailed description of how to estimate the parameter contribution distribution, which will guide subsequent parameter pruning.

Defining the Parameter Contribution. For a given input $\mathbf{x}$, the contribution of a specific parameter $\boldsymbol{\theta}_{ij}$ to the category k is defined as the change in the k-th output of the model by the presence or absence (setting $\boldsymbol{\theta}_{ij}$ to 0) of the parameter $\boldsymbol{\theta}_{ij}$, *i.e.*,

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

Previous studies (Zhu et al. 2022) highlight that the features extracted by early layers show similarities between ID and OOD data. In contrast, the later layers, particularly the penultimate layer, can extract more separable features. In this paper, we primarily focus on the last layer's model parameters, which significantly impact OOD detection by processing separable features and directly influencing particular class outputs. Especially, the contributions of the last layer's parameters $\mathbf{W}_{ij}$ can be expressed more simply using Equation 4 as follows (see Appendix for details):

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

Estimating the Distribution of Parameter Contribution. The distribution estimation relies on the assumption that the contribution of the last layer's parameters approximately follows Gaussian distributions parameterized by (μ, σ), as observed empirically (see Appendix). According to the Equation 5, the parameter $\mathbf{W}_{ij}$ is specifically associated with class j. To minimize potential bias from including data from other classes, we estimate the contribution distribution of parameter $\mathbf{W}_{ij}$ using only the training data for class j. Let $\mathcal{D}_j$ denote the set of data points belonging to class j. The mean μ_{ij} and standard deviation σ_{ij} of the contribution distribution for the parameter $\mathbf{W}_{ij}$ are estimated using the following class-specific formulas:

$$\mu_{ij} = \frac{1}{|\mathcal{D}_j|} \sum_{x \in \mathcal{D}_j} c_j(x; \mathbf{W}_{ij}),$$
$$\sigma_{ij} = \left(\frac{1}{|\mathcal{D}_j| - \delta} \sum_{x \in \mathcal{D}_j} \left(c_j(x; \mathbf{W}_{ij}) - \mu_{ij} \right)^2 \right)^{\frac{1}{2}}, \tag{6}$$

where $|\mathcal{D}_j|$ denotes the cardinality of the set $\mathcal{D}_j$, and δ represents the correction factor. To correct the bias in the estimation of the population standard deviation, we adopt Bessel's correction by setting δ to 1.

3.3 Instance-Aware Test Pruning (ITP)

In this section, we introduce two parameter pruning strategies with different levels of granularity used in ITP for post-hoc enhancement in OOD detection: coarse-grained redundancy pruning (Figure 2(2)) and fine-grained test pruning (Figure 2(3)). Detailed descriptions of each strategy are provided below.

Coarse-Grained Redundancy Pruning (CRP) remove redundant parameters in the over-parameterized weight space of the model. CRP operates at a coarse-grained level by applying a uniform pruning pattern across all test samples. Specifically, CRP measures each parameter's redundancy based on its average contribution to ID prediction. Parameters that fall within the lowest $p\%$ of average contributions are deemed redundant and pruned. To implement this, we define a mask matrix $\mathbf{M}^{\text{CRP}}$ for CRP, where the average contribution of a parameter is directly obtained from the mean μ of its contribution distribution. The (i, j)-th entry of the mask matrix $\mathbf{M}_{ij}^{\text{CRP}} \in \mathbf{M}^{\text{CRP}}$ is defined as follows:

$$\mathbf{M}_{ij}^{\text{CRP}} = \begin{cases} 1, & \text{if} \quad \mu_{ij} > \Omega_p, \\ 0, & \text{if} \quad \mu_{ij} \leq \Omega_p, \end{cases} \tag{7}$$

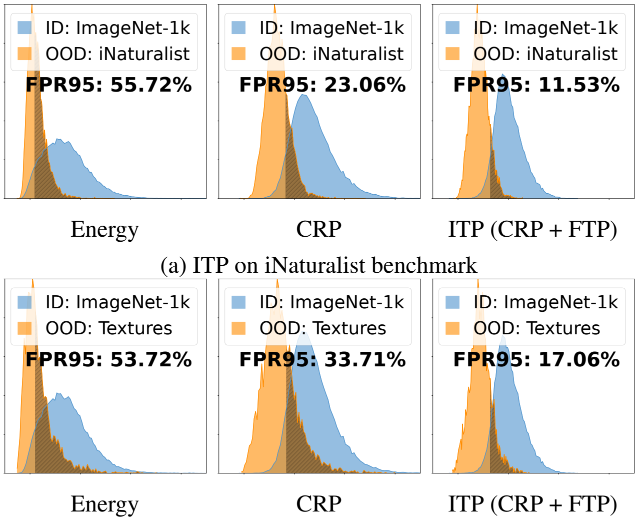
where Ω_p represents the average contribution threshold at the lowest p percentile. The model output after applying CRP can be expressed as follows:

$$f^{\text{CRP}}(\mathbf{x}) = \left(\mathbf{W} \odot \mathbf{M}^{\text{CRP}} \right)^{\top} h(\mathbf{x}) + \mathbf{b}, \tag{8}$$

where $\odot$ denotes the element-wise multiplication. Through CRP, we remove noise interference from redundant parameters at a coarse-grained level in OOD detection by retaining only the parameters important for ID data, thereby enhancing the distinction between ID and OOD data.

Fine-Grained Test Pruning (FTP) prunes overconfident parameters with anomalously high contributions to ID prediction. FTP operates at a fine-grained level by customizing parameter pruning patterns for each test sample at test time. Specifically, FTP determines whether the parameter is overconfident by performing a right-tail test based on the Z-score. The Z-score quantifies the deviation of a data point from the mean of the distribution, expressed as $(X - \mu)/\sigma$. In this context, X represents the contribution being evaluated, while μ and σ denote the mean and standard deviation of the contribution distribution, respectively. FTP can be framed as a single-sample hypothesis testing task:

$$\mathcal{H}_0 : \frac{X - \mu}{\sigma} \leq \lambda, \quad \text{vs.} \quad \mathcal{H}_1 : \frac{X - \mu}{\sigma} > \lambda, \tag{9}$$



(a) ITP on iNaturalist benchmark

(b) ITP on Textures benchmark

Figure 3: Changes in Energy score distribution using ITP on (a) iNaturalist benchmark and (b) Textures benchmark.

where the alternative hypothesis $\mathcal{H}_1$ implies that the parameter behaves overconfidently, and λ is the threshold, with $\lambda > 0$. In practice, we define $\mathbf{M}^{\text{FTP}}(\mathbf{x})$ as the mask matrix customized for $\mathbf{x}$ to perform FTP. The (i, j)-th entry of the mask matrix $\mathbf{M}_{ij}^{\text{FTP}}(\mathbf{x}) \in \mathbf{M}^{\text{FTP}}(\mathbf{x})$ is defined as follows:

$$\mathbf{M}_{ij}^{\text{FTP}}(\mathbf{x}) = \begin{cases} 1, & \text{if} \quad \dfrac{c_j(\mathbf{x}; \mathbf{W}_{ij}) - \mu_{ij}}{\sigma_{ij}} \leq \lambda, \\ 0, & \text{if} \quad \dfrac{c_j(\mathbf{x}; \mathbf{W}_{ij}) - \mu_{ij}}{\sigma_{ij}} > \lambda. \end{cases} \tag{10}$$

The model output after applying FTP can be expressed as:

$$f^{\text{FTP}}(\mathbf{x}) = \left(\mathbf{W} \odot \mathbf{M}^{\text{FTP}}(\mathbf{x}) \right)^{\top} h(\mathbf{x}) + \mathbf{b}. \tag{11}$$

Through FTP, we can adaptively prevent the abnormal increase in ID confidence caused by anomalously high contributions from overconfident parameters. This effectively reduces the risk of the model making confident yet erroneous predictions, thereby making ID data and OOD data more distinguishable.

Overall Methods. Both CRP and FTP pruning strategies are designed to remove parameters that negatively impact OOD detection. CRP utilizes a fixed, coarse-grained pruning pattern across all test samples to reduce interference from noisy signals. FTP applies a customized, fine-grained pruning pattern to overconfident parameters for each test sample, mitigating the risk of overconfident predictions. ITP achieves both ways to improve OOD detection. As a result, the model outputs using ITP can be expressed as follows:

$$f^{\text{ITP}}(\mathbf{x}) = \left(\mathbf{W} \odot \mathbf{M}^{\text{CRP}} \odot \mathbf{M}^{\text{FTP}}(\mathbf{x}) \right)^{\top} h(\mathbf{x}) + \mathbf{b}. \tag{12}$$

Similar to previous works (Ahn, Park, and Kim 2023), we can always use the original FC layer for prediction to preserve ID accuracy with negligible additional overhead.

3.4 Insight Justification

The following remarks are provided to further explain how ITP widens the gap between ID and OOD data.

Remark 1. CRP enhances the disparity between the left tail of the OOD score distributions for ID and OOD data. After employing CRP to prune redundant parameters, the logits reduction for the k-th class is given by:

$$\Delta f_j(\mathbf{x}) = \sum_{d=1}^{D} (1 - \mathbf{M}^{\text{CRP}}_{dj}) \cdot c_j(\mathbf{x}; \mathbf{W}_{dj}). \tag{13}$$

CRP eliminates parameters with the least average contribution to the ID distribution. Hence, redundant parameters generally have higher contributions (manifesting as noise) to ID prediction for OOD data compared to ID data, *i.e.*,

$$\sum_{\mathbf{M}^{\text{CRP}}_{dj}=0} c_j(\mathbf{x}^{\text{OOD}}; \mathbf{W}_{dj}) > \sum_{\mathbf{M}^{\text{CRP}}_{dj}=0} c_j(\mathbf{x}^{\text{ID}}; \mathbf{W}_{dj}). \tag{14}$$

Therefore, the reduction in logits for OOD data is greater than that for ID data $\Delta f_j(\mathbf{x}^{\text{OOD}}) > \Delta f_j(\mathbf{x}^{\text{ID}})$. This improves the differentiation between the left tail of the energy score distributions for ID and OOD data, due to the positive correlation between energy scores and logits. This effect is empirically validated in Figure 3.

Remark 2. FTP alleviates the overconfidence of OOD data at the right tail of the OOD score distribution. FTP removes the abnormally high contributions caused by overconfident parameters, thereby increasing the left skewness in the parameter contribution distribution, *i.e.*,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\left[\left(\frac{c_j(\mathbf{x}; \mathbf{W}_{ij}) - \mu_{ij}}{\sigma_{ij}}\right)^3 \cdot \mathbf{M}^{\text{FTP}}_{ij}(\mathbf{x})\right] \\ < \mathbb{E}_{\mathbf{x}}\left[\left(\frac{c_j(\mathbf{x}; \mathbf{W}_{ij}) - \mu_{ij}}{\sigma_{ij}}\right)^3\right].\end{aligned} \tag{15}$$

Since parameter contributions directly determine the energy score, the left skewness of the energy score distribution will also increase accordingly. This helps reduce the risk of parameters being overly confident when handling OOD data. As a result, the overlap between the right tail of the OOD energy score distribution and the ID energy score distribution is diminished, as illustrated in Figure 3.

4 Experiments

In this section, we first describe our experimental setup, then present the main results on multiple OOD detection benchmarks, followed by ablation studies and further analysis.

4.1 Experimental Setup

In line with other OOD literature (Sun and Li 2022), we evaluate our methods both on the small-scale CIFAR benchmarks and the large-scale ImageNet benchmark[1]. Moreover, we provide a further evaluation of our proposal on the OpenOOD v1.5 benchmark (Zhang et al. 2023) in the appendix. We default to using the entire training set for estimating the parameter contribution distribution.

Method	CIFAR-10		CIFAR-100	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	48.73	92.46	80.13	74.36
Energy	26.55	94.57	68.45	81.19
ODIN	24.57	93.71	58.14	84.49
ReAct	26.45	94.67	62.27	84.47
DICE	20.83	95.24	49.72	87.23
OPNP	22.07	95.14	51.79	87.20
LAPS	19.40	96.10	50.50	88.07
ITP (Ours)	**16.72**	**96.64**	**35.03**	**91.39**
DICE + ReAct	16.48	96.64	49.57	85.07
OPNP + ReAct	18.46	96.35	42.98	88.55
LINe (w/ ReAct)	14.72	96.99	35.67	88.67
ITP + ReAct (Ours)	**14.50**	**97.13**	**30.13**	**91.91**

Table 1: OOD detection performance on CIFAR benchmarks with DenseNet-101 as the backbone. All values in the table are averaged over six OOD test datasets and are percentages. The best results are in bold. ↑ indicates that larger values are better, while ↓ indicates that smaller values are better. Detailed results for each OOD dataset are provided in the Appendix.

CIFAR. We use CIFAR-10 and CIFAR-100 (Krizhevsky 2009) as ID datasets and consider six OOD datasets: SVHN (Netzer et al. 2011), Textures (Cimpoi et al. 2014), iSUN (Xu et al. 2015), LSUN-Resize (Yu et al. 2015), LSUN-Crop (Yu et al. 2015), and Places365 (Zhou et al. 2018). For consistency with previous work (Sun and Li 2022), we use the same model architecture and pre-trained weights, namely DenseNet-101 (Huang et al. 2017).

ImageNet. For the large-scale ImageNet experiments, we use the ImageNet-1k as the ID dataset and consider (subsets of) iNaturalist (Horn et al. 2018), Places (Zhou et al. 2018), SUN (Xiao et al. 2010), and Textures (Cimpoi et al. 2014) with non-overlapping categories from ImageNet-1k as OOD datasets. We adopt the widely used ResNet-50 (He et al. 2016) model architectures, and we obtain the pre-trained weights from the torchvision library.

Baselines. We compare ITP with the most competitive OOD detection methods: MSP (Hendrycks and Gimpel 2017), Energy (Liu et al. 2020), ODIN (Liang, Li, and Srikant 2018), ReAct (Sun, Guo, and Li 2021), DICE (Sun and Li 2022), LINe (Ahn, Park, and Kim 2023), OPNP (Chen et al. 2023), and LAPS (He et al. 2024). Moreover, to align with standard sparsification practices, we also report the results of a comparison with ReAct (e.g., ITP + ReAct). In particular, LINe integrates ReAct within its framework, which we refer to as "LINE (w/ ReAct)" in the table. All methods are post-hoc and can be directly applied to pre-trained models.

Evaluation Metric. We adopt two threshold-free metrics for evaluation. FPR95: the false positive rate of OOD data at 95% true positive rate of ID data. AUROC: the area under the receiver operating characteristic curve.

[1] Code is available at https://github.com/njustkmg/AAAI25-ITP

Method	**OOD Datasets**								**Average**	
	iNaturalist		**SUN**		**Places**		**Textures**			
	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
Energy	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
ODIN	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
ReAct	20.38	96.22	24.20	94.20	33.85	91.58	47.30	89.80	31.43	92.95
DICE	25.63	94.49	35.15	90.83	46.49	87.48	31.72	90.30	34.75	90.77
OPNP	18.89	96.03	18.50	95.62	30.14	93.46	36.17	91.70	25.93	94.20
LAPS	12.72	97.50	**15.81**	**96.18**	**24.71**	**93.64**	41.49	91.81	23.68	**94.78**
ITP (Ours)	**11.53**	**97.83**	25.82	93.58	35.63	90.75	**17.06**	**96.03**	**22.51**	94.55
DICE + ReAct	18.64	96.24	25.45	93.94	36.86	90.67	28.07	92.74	27.25	93.40
OPNP + ReAct	14.72	96.78	19.73	**95.65**	30.23	**93.34**	27.78	94.13	23.12	94.98
LINe (w/ ReAct)	12.26	97.56	**19.48**	95.26	**28.52**	92.85	22.54	94.44	20.70	95.03
ITP + ReAct (Ours)	**9.78**	**98.02**	22.82	94.47	30.87	92.03	**18.09**	**95.98**	**20.39**	**95.13**

Table 2: OOD detection performance on ImageNet with ResNet-50 as the backbone.

Dataset	**CRP**	**FTP**	**FPR95** ↓	**AUROC** ↑
CIFAR-10 DenseNet-101	×	×	26.55	94.57
	✓	×	21.29	95.09
	×	✓	19.16	96.36
	✓	✓	**16.96**	**96.59**
CIFAR-100 DenseNet-101	×	×	68.45	81.19
	✓	×	53.60	85.33
	×	✓	53.73	87.56
	✓	✓	**35.03**	**91.39**
ImageNet-1k ResNet-50	×	×	58.41	86.17
	✓	×	33.96	91.26
	×	✓	51.38	88.68
	✓	✓	**22.51**	**94.55**

Table 3: Ablation study for our proposed method. All values are percentages and averaged over multiple OOD datasets.

4.2 Main Results

In this section, we report the performance of our ITP on commonly used CIFAR benchmarks and the more realistic and challenging ImageNet benchmark. Baseline results are sourced from (Ahn, Park, and Kim 2023; Chen et al. 2023; He et al. 2024), with additional baselines (*e.g.*, LAPS, OPNP, and OPNP + ReAct on CIFAR) reproduced by us.

For the CIFAR benchmarks, Table 1 lays out the performance of OOD detection on the CIFAR-10 and CIFAR-100, respectively. As we can see, the proposed ITP outperforms all baselines considered and achieves state-of-the-art performance. In CIFAR-100, ITP reduces FPR95 by 33.42% compared to the energy baseline, showing the effectiveness of our proposal with the same OOD scoring function. Remarkably, ITP + ReAct outperforms the most competitive method LINe by 5.54% in FPR95 and 3.24% in AUROC, highlighting the importance of further fine-grained pruning of overconfident parameters.

	$p = 10$	$p = 30$	$p = 50$	$p = 70$	$p = 90$
$\lambda = 0.5$	73.19	34.19	33.70	33.75	35.96
$\lambda = 1.0$	33.18	25.44	26.83	27.26	30.96
$\lambda = 1.5$	26.30	**22.51**	24.18	24.67	29.50
$\lambda = 2.0$	27.58	24.69	25.95	26.45	31.59
$\lambda = 3.0$	31.95	29.14	30.37	30.73	36.53
$\lambda = 5.0$	35.90	33.00	34.14	34.52	40.23

Table 4: Impact of varying hyperparameters on FPR95. We use ImageNet-1k as the ID dataset and ResNet-50 as the pretrained model. All values are percentages and averaged over four OOD datasets.

For the large-scale ImageNet benchmark, Table 2 reports detailed performances for each OOD dataset and the average over the four datasets. Our proposed method, ITP, outperforms recent approaches such as DICE, OPNP, and LAPS, achieving the best performance among the baseline methods with an FPR95 of 22.51%. Moreover, the combination of ITP and ReAct outperforms recent approaches DICE + ReAct, OPNP + ReAct, and LINe. The experimental results demonstrate that our ITP is state-of-the-art and effective for OOD detection on large-scale real-world datasets.

4.3 Ablation Study

Ablation on proposed pruning strategies. To fully demonstrate the impact of different granularity pruning strategies in ITP, we conduct a comprehensive empirical analysis on CIFAR-10, CIFAR-100, and ImageNet-1k, and report the results in Table 3. As shown in the table, both FTP and CRP improve performance, and ITP further markedly boosts OOD detection performance by integrating these two coarse-grained and fine-grained pruning strategies. However, the improvement of FTP on ResNet-50 is less pronounced, likely due to its larger feature space (2048 dimensions) compared to DenseNet-101 (342 dimensions). This larger space allows noise to dominate and interfere with FTP.

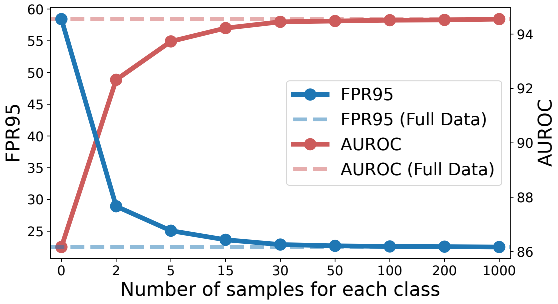


Figure 4: The FFPR95 and AUROC with different number of training samples on ImageNet benchmark. The results are averaged over five independent runs.

Method	Preprocessing Time (hours)	Additional Backpropagation	Batch Support
OPNP	8.7940	✓	×
LINe	7.3096	✓	×
ITP	0.2411	×	✓
ITP (30)	**0.0073**	×	✓

Table 5: Comparison of preprocessing overhead. We assess the preprocessing overhead by averaging the preprocessing times measured across three runs on ImageNet-1k. "ITP (30)" denotes that only 30 images per class are used for ITP while maintaining the original performance (see Figure 4).

The significant improvement observed when FTP is applied after noise removal with CRP supports this explanation. The ablation study verifies the effectiveness of the two strategies and demonstrates that they mutually enhance and complement each other.

Effect of the Hyperparameter. Table 4 shows the results of varying the percentile p used for pruning redundant parameters and the threshold λ for identifying overconfident parameters. The optimal performance is observed at $p = 30$ and $\lambda = 1.5$, achieving an FPR95 of 22.51%. Conversely, we notice that selecting excessively large values for p and overly small values for λ can lead to the erroneous removal of critical parameters, thereby impairing OOD detection.

Effect of the Amount of Training Samples. In Figure 4, we show the effect of utilizing different numbers of training samples to estimate the parameter contribution distribution. Remarkably, even with just two samples per class, ITP can significantly improve OOD detection performance, resulting in a drastic 29.48% reduction in FPR95 compared to the energy baseline (without pruning). Furthermore, empirical evidence suggests that using only 30 samples per class can yield performance nearly equivalent to that achieved with the full dataset. Therefore, to reduce computational overhead, it is feasible to use a suitably sized subset of the training set (*e.g.*, 30 samples per class) for distribution estimation, while still achieving comparable performance.

Method	FPR95 ↓	AUROC ↑
MSP	66.95	81.99
MSP + **ITP**	**62.44**	**82.99**
ODIN	56.48	85.41
ODIN + **ITP**	**42.32**	**90.10**
GradNorm	36.49	90.18
GradNorm + **ITP**	**29.93**	**92.13**
MLS	58.05	87.00
MLS + **ITP**	**26.43**	**93.45**
Energy	58.41	86.17
Energy + **ITP**	**22.51**	**94.55**

Table 6: ITP on other OOD scores. We use ResNet-50 as the pre-trained model and ImageNet-1k as the ID dataset. The results are averaged over four OOD datasets.

4.4 Further Analysis

Analysis of Preprocessing Overhead. Table 5 compares the preprocessing overhead of ITP with the most competitive weight sparsification methods: LINe and OPNP. In contrast to ITP, both LINe (Ahn, Park, and Kim 2023) and OPNP (Chen et al. 2023) require additional backpropagation to compute gradient information and lack support for batch processing. The results indicate that our proposal demonstrates a significant advantage in terms of preprocessing overhead. Notably, with only 30 samples per class, we can further substantially reduce the overhead while maintaining comparable performance (see Figure 4). Therefore, ITP is efficient and well-suited for real-world applications.

Compatibility with Other OOD Scores. Table 6 presents the OOD detection performance of ITP using various OOD scoring methods, including MSP (Hendrycks and Gimpel 2017), ODIN (Liang, Li, and Srikant 2018), GradNorm (Huang, Geng, and Li 2021), MLS (Hendrycks et al. 2022), and Energy (Liu et al. 2020). Our ITP consistently improves FPR95 and AUROC across different OOD scores. In particular, ITP can effectively complement gradient-based methods such as GradNorm. These results indicate that the parameters our ITP selectively used are also applicable to other OOD scores and show strong compatibility.

5 Conclusion

In this paper, we reveal that parameters important for ID data prediction are not always beneficial for OOD detection. To address this issue, we propose a parameter pruning method called ITP, which utilizes class-specific parameter contribution distributions for post-hoc OOD detection. ITP is based on two powerful pruning strategies: CRP performs coarse-grained pruning to remove redundant parameters, while FTP executes fine-grained pruning to eliminate overconfident parameters. Experimental results show that our ITP method significantly improves OOD detection performance and can be integrated with a wide range of other OOD scoring methods. We hope our work can raise more attention to the importance of test parameter pruning for OOD detection.

Acknowledgements

National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081), the Fundamental Research Funds for the Central Universities (No.30922010317, No.30923011007), Key Laboratory of Target Cognition and Application Technology (2023-CXPT-LC-005).

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.

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

Behpour, S.; Doan, T. L.; Li, X.; He, W.; Gou, L.; and Ren, L. 2023. GradOrth: A Simple yet Efficient Out-of-Distribution Detection with Orthogonal Projection of Gradients. In *NeurIPS*.

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

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*, pages 3606–3613.

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*.

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

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

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

He, R.; Yuan, Y.; Han, Z.; Wang, F.; Su, W.; Yin, Y.; Liu, T.; and Gong, Y. 2024. Exploring Channel-Aware Typical Features for Out-of-Distribution Detection. In *AAAI*, pages 12402–12410.

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*, volume 162, pages 8759–8773.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. 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*, pages 8769–8778.

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

Huang, H.; Li, Z.; Wang, L.; Chen, S.; Zhou, X.; and Dong, B. 2021. Feature Space Singularity for Out-of-Distribution Detection. In *AAAI*, volume 2808.

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

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

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*, pages 7167–7177.

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: pages 60–88.

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

Malinin, A.; and Gales, M. J. F. 2018. Predictive Uncertainty Estimation via Prior Networks. In *NeurIPS*, pages 7047–7058.

Monteiro, J.; Rodríguez, P.; Noël, P.; Laradji, I. H.; and Vázquez, D. 2023. Constraining Representations Yields Models That Know What They Don't Know. 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.

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

Shapley, L. S.; et al. 1953. A value for n-person games. *Classics in game theory*.

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

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

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. In *ICML*, volume 162, pages 20827–20840.

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

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492.

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

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, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Hendrycks, D.; Li, Y.; and Liu, Z. 2022. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. In *NeurIPS*.

Yang, Y.; Huang, Y.; Guo, W.; Xu, B.; and Xia, D. 2023a. Towards Global Video Scene Segmentation with Context-Aware Transformer. In *AAAI*, 3206–3213.

Yang, Y.; Jiang, N.; Xu, Y.; and Zhan, D. 2024. Robust Semi-Supervised Learning by Wisely Leveraging Open-Set Data. *TPAMI*, 46(12): 8334–8347.

Yang, Y.; Zhang, C.; Xu, Y.; Yu, D.; Zhan, D.; and Yang, J. 2021. Rethinking Label-Wise Cross-Modal Retrieval from A Semantic Sharing Perspective. In *IJCAI*, 3300–3306.

Yang, Y.; Zhang, Y.; Song, X.; and Xu, Y. 2023b. Not All Out-of-Distribution Data Are Harmful to Open-Set Active Learning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *NeurIPS*.

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, 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. 2023. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *CoRR*, abs/2306.09301.

Zhou, B.; Lapedriza, À.; Khosla, A.; Oliva, A.; and Torralba, A. 2018. Places: A 10 Million Image Database for Scene Recognition. *TPAMI*.

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*.