Enhancing Out-of-Distribution Detection with Extended Logit Normalization

Yifan Ding[1] Xixi Liu[2] Jonas Unger[1]
Gabriel Eilertsen[1]
[1]Linköping University [2]Imperial College London
[1]{yifan.ding, jonus.unger, gabriel.eilertsen}@liu.se [2]x.liu2@imperial.ac.uk

Abstract

Out-of-distribution (OOD) detection is essential for the safe deployment of machine learning models. While extensive work has focused on designing effective scoring functions for OOD detection, relatively few studies explore training neural networks with calibration-oriented objectives, which often compromise predictive accuracy and restrict the choice of scoring functions. In this work, we first identify feature collapse in Logit Normalization (LogitNorm), and then propose a novel hyperparameter-free training formulation that significantly improves a wide range of post-hoc detection methods. Specifically, we introduce a feature distance-aware normalization objective, termed ***ELogitNorm****, which enhances both OOD detection performance and in-distribution (ID) confidence calibration. Extensive experiments on standard benchmarks demonstrate that our approach outperforms state-of-the-art training-time methods in OOD detection while preserving strong ID classification performance. Our code is available at: https://github.com/limchaos/ElogitNorm.*

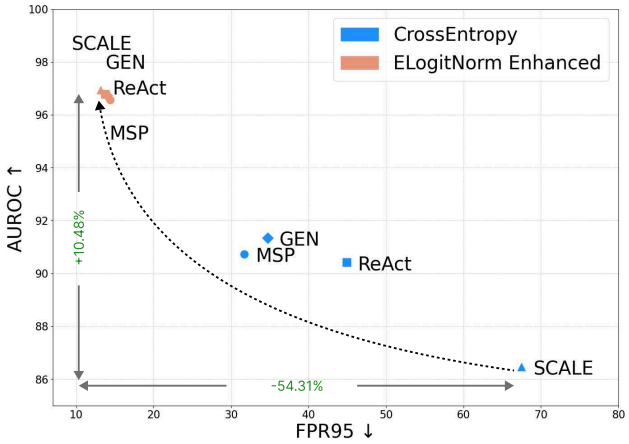


Figure 1. *Effectiveness of* ***ELogitNorm*** *applied with 4 Baseline OOD Methods.* The OOD performance, as measured by AUROC % (higher is better) and FPR95 % (lower is better). The in-distribution dataset is CIFAR-10 and the scores have been averaged over four far-OOD datasets in the OpenOOD benchmark [56].

1. Introduction

The reliability of learning-based systems is a cornerstone for their successful deployment in safety critical and socially impactful applications. Despite their remarkable performance, deep networks often assume that the training and test data share the same underlying distribution. In real world scenarios, however, this assumption rarely holds true. Models frequently encounter samples that differ from the training distribution, leading to unreliable predictions and degraded performance. Thus, a large body of research has been dedicated to developing methods for solving this problem by means of out-of-distribution (OOD) detection [7, 8, 15, 17, 18, 23, 27, 28, 30, 42, 49, 51]. Many previous works develop *post-hoc* OOD scoring methods using information from the feature space, logit space, or probability space of the trained model [15, 17, 27, 28]. Some approaches require access to training data statistics [17, 23, 48], while others focus on reshaping feature representations to enhance the separation between in-distribution (ID) and OOD features [7, 41, 42, 51]. Most of these methods assume that the classifier is trained solely with Cross-Entropy loss, yet their performance remains suboptimal. To address these limitations, several studies have proposed synthesizing outlier features and images from ID features [8, 9, 30, 45, 49], or reformulating classification as a deep metric learning task [30]. While these approaches improve OOD detection compared to post-hoc methods, they often rely on external generative models such as Stable Diffusion [39], require two-stage training [8, 9, 45], impose restrictions on OOD scoring methods, or suffer from a trade-off between classification accuracy and OOD detection performance.

LogitNorm [49] dives into the training dynamics of cross-entropy loss and enforces a normalization on the logit vec-

tor to mitigate the overconfidence of predictions. However, it improves OOD performance as a compromise of ID classification performance and limited post-hoc scoring functions. In this work, we diagnose the limitations of LogitNorm and reveal that the learned features tend to collapse toward the origin while being compressed into a few dominant directions. To address this issue, we introduce a novel training objective that accounts for the distances to class-specific decision boundaries within LogitNorm. We refer to this approach as *Extended Logit Normalization* (**ELogitNorm**) and demonstrate that it effectively prevents feature collapse while improving OOD detection performance. Moreover, models trained with our proposed loss can be seamlessly integrated with most post-hoc OOD scoring functions, whereas LogitNorm may exhibit degraded OOD performance under certain post-hoc scoring methods (see Fig. 3). In addition, our method achieves better confidence calibration (see Tab. 4). It has also been consistently observed that no single method dominates across all benchmarks [56]. For instance, approaches such as SCALE [51] perform strongly on ImageNet-1K but struggle on CIFAR-10. As shown in Fig. 1, SCALE [51] is outperformed by alternative methods. After applying calibrated training with our proposed **ELogitNorm**, all post-hoc OOD detection methods cluster around a similar level of OOD detection accuracy.

In summary, our approach offers several key contributions and advantages:

- We identify a feature collapse phenomenon in LogitNorm, which restricts its applicability to a wider range of post-hoc OOD scoring functions and leads to reduced classification accuracy.
- **ElogitNorm** improves OOD detection performance across diverse post-hoc OOD scoring methods. In contrast to existing training approaches, our method maintains classification accuracy while ensuring broader compatibility with various OOD scoring functions, cf. Fig. 5.
- The proposed method is simple, hyperparameter-free, and consistently achieves superior OOD performance on both near-OOD and far-OOD benchmarks. Moreover, it produces better-calibrated classifiers, as reflected by lower expected calibration error (ECE) values, cf. Table 4.

2. Preliminaries

2.1. Out-of-distribution Detection

In image classification, ID samples are defined by a fixed set of semantic categories $\mathcal{Y}_{\text{ID}}$ with joint distribution $\mathcal{D}_{\text{ID}}$, where $\forall(\mathbf{x}, y) \sim \mathcal{D}_{\text{ID}},\ y \in \mathcal{Y}_{\text{ID}}$. In open-world scenarios, additional unseen categories naturally emerge, forming the OOD space $\mathcal{D}_{\text{OOD}}$. In this setting, OOD detection aims to achieve two primary objectives [56]. The first is to develop a discriminative model that accurately classifies ID samples drawn from $\mathcal{D}_{\text{ID}}$. The second objective is to construct a confidence score S that effectively distinguishes between ID and OOD samples. Formally, given an input sample x and a neural network parametrized by θ, the OOD detection is defined as follows:

$$x \in \begin{cases} \mathcal{X}_{\text{ID}}, & \text{if } S_\theta(x) \geq \lambda \\ \mathcal{X}_{\text{OOD}}, & \text{if } S_\theta(x) < \lambda, \end{cases} \tag{1}$$

where λ is a predefined threshold, and $S_\theta(\cdot)$ is an OOD scoring function such as [15, 17, 27, 28].

2.2. Logit Normalization

Logit Normalization (LogitNorm) [49] is a technique used to stabilize the output logits of a neural network by constraining their norm. In classification models, raw logits are typically unbounded, which can lead to overconfident predictions and suboptimal generalization, particularly in the presence of OOD samples. To address this, LogitNorm ensures that the logits maintain a controlled magnitude, improving robustness and calibration. Mathematically, let $\mathbf{f} = f(\mathbf{x}; \theta)$ be the logit vector corresponding to an input $\mathbf{x}$ and neural network f parametrized by θ. LogitNorm normalizes the logits to produce a unit logit vector:

$$\tilde{\mathbf{f}} = \frac{\mathbf{f}}{||\mathbf{f}||}, \tag{2}$$

where $||\mathbf{f}||$ denotes the ℓ_2-norm of the logit vector. This operation projects the logits onto a unit sphere, preventing excessive scaling and enforcing a uniform logit distribution. The corresponding loss function for LogitNorm is given by:

$$\mathcal{L}_{\text{LogitNorm}}(f(\mathbf{x}; \theta), y) = -\log \frac{e^{f_y/\tau||\mathbf{f}||}}{\sum_{i=1}^{c} e^{f_i/\tau||\mathbf{f}||}}, \tag{3}$$

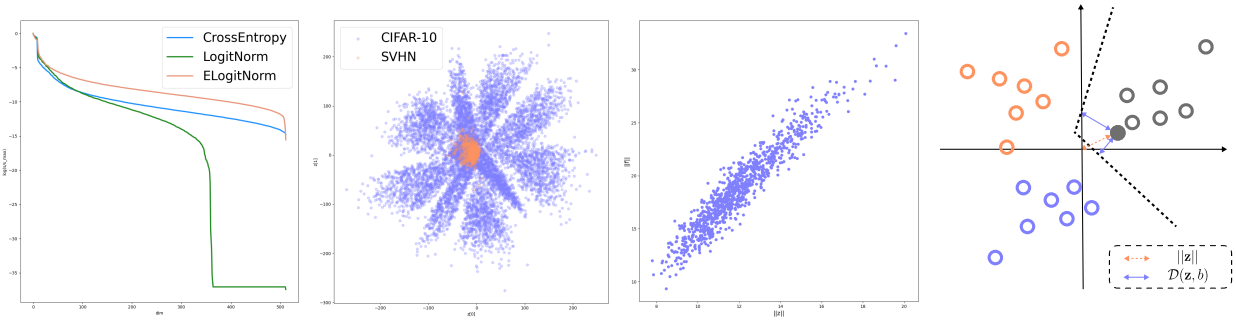
where c is the number of classes, f_y represents the logit corresponding to the correct class, and τ is a temperature hyperparameter. Replacing the standard cross-entropy objective, the LogitNorm training objective is optimizing:

$$\min_{\theta} \mathbb{E}_{\mathcal{P}_{\mathcal{X}\mathcal{Y}}} \mathcal{L}_{\text{LogitNorm}}(f(\mathbf{x}; \theta), y). \tag{4}$$

3. Method: Extended Logit Normalization

In the following, we first identify two types of feature collapse in LogitNorm, and then demonstrate that they can be overcome by our proposed method.

Notation. To analyze the limitations of LogitNorm, we consider the feature space of the classifier. To this end, let $\mathbf{f} = \mathbf{W}^\top \mathbf{z} + \mathbf{b}$ be the logits of a neural network classifier, where $\mathbf{W} \in \mathbb{R}^{m \times c}$ is the weight matrix, $\mathbf{z} \in \mathbb{R}^m$ is the feature vector before the final layer, and $\mathbf{b} \in \mathbb{R}^c$ is the bias term.



(a) Singular value spectrum $\Sigma_{\text{cov}}^{\text{ID}}$ (b) CIFAR-10 v.s. SVHN (c) $||\mathbf{f}||$ v.s. $||\mathbf{z}||$ (d) $\mathcal{D}(\mathbf{z}, b)$ v.s. $||\mathbf{z}||$

Figure 2. (a): The singular value spectrum (log scale for better visualization) of the covariance of training ID embedding $\mathbf{z}$ under different training losses including Cross-Entropy, LogitNorm [49], and **ELogitNorm** (Ours). The ID data is CIFAR-10 trained on ResNet-18. (b): We train a ResNet18 on CIFAR-10 and set feature $\mathbf{z} \in \mathbb{R}^2$ before the last penultimate layer for visualization. (c) $||\mathbf{z}||$ and $||\mathbf{f}||$ denote the feature norm and logit norm, respectively. (d): Distance to origin $||\mathbf{z}||$ and distance to decision boundary $\mathcal{D}(\mathbf{z}, b)$, where b is the corresponding decision boundary.

3.1. Motivating Observation

Although LogitNorm aims to mitigate overconfidence by enforcing normalization on logits $\mathbf{f}$ during training, we find that it inadvertently induces feature dimension collapse. In addition, even when trained solely with cross-entropy loss, OOD feature representations tend to cluster near the origin rather than being distributed across other low-likelihood regions (*e.g.* decision boundaries). We summarize these findings through two key observations:

(a) Dimensional collapse. As illustrated in Fig. 2a, the singular value spectrum of the learned features under LogitNorm exhibits several small singular values approaching zero, indicating a collapse of feature variance along many dimensions. This suggests that LogitNorm may cause a potential loss of representational information, as features become concentrated in a few dominant directions. Such reduction in the effective feature dimensionality not only limits the expressiveness of the learned representation, but may also hinder downstream tasks that rely on rich feature diversity [19]. In extreme cases, the network may overfit to a low-dimensional subspace, making it less robust to variations in the input and less capable of separating complex decision boundaries.

(b) Origin collapse. We further observe that OOD data tend to reside closer to the origin in the feature space compared to ID data, implying that neural networks inherently produce lower-norm embeddings for OOD samples. To confirm this, we train a classifier with a feature dimensionality of $m = 2$, *i.e.* $\mathbf{z} \in \mathbb{R}^2$, to allow direct visualization. We prefer this approach over non-linear dimensionality reduction methods such as t-SNE [47] or UMAP [29], which are highly sensitive to hyperparameters and often difficult to interpret. As shown in Fig. 2b, when CIFAR-10 is used as the ID dataset and SVHN as the OOD dataset, the OOD samples are observed to collapse near the origin. We demonstrate how LogitNorm encourage this types of collapse in Section 3.2.

3.2. LogitNorm Further Encourages Collapse

Building on the observations from the previous section, we aim to analyze the collapsing behavior of LogitNorm through its normalization factor, $\tau||\mathbf{f}||$, in Eq. (3). To this end, we examine the phenomenon in the feature space, $\mathbf{z} \in \mathbb{R}^m$, rather than directly normalizing the logits $\mathbf{f}$. This is justified by the proportional relationship between $||\mathbf{f}||$ and $||\mathbf{z}||$, indicating that normalization in logit space corresponds to an equivalent operation in the feature space.

Proposition 1. *The norm of the logits $||\mathbf{f}||$ is approximately proportional to the feature norm $||\mathbf{z}||$, up to an additive noise term, such that $||\mathbf{f}|| \approx \bar{\sigma}||\mathbf{z}|| + \eta$, where $\bar{\sigma}$ is the weighted mean of the singular values. More formally, the following bound holds:*

$$\sigma_{\min}||\mathbf{z}|| - ||\mathbf{b}|| \leq ||\mathbf{f}|| \leq \sigma_{\max}||\mathbf{z}|| + ||\mathbf{b}||,$$

where $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest singular values of the weight matrix $\mathbf{W}$, respectively.

The proof of Proposition 1 is provided in Appendix A. It shows that the LogitNorm objective can be approximated as,

$$\mathcal{L}_{\text{LogitNorm}}(f(\mathbf{x};\theta), y) \approx -\log \frac{e^{f_y/\hat{\tau}||\mathbf{z}||}}{\sum_{i=1}^{c} e^{f_i/\hat{\tau}||\mathbf{z}||}}, \quad (5)$$

where $\hat{\tau}$ is a new hyperparameter. We evaluate the proportionality between $||\mathbf{f}||$ and $||\mathbf{z}||$ through experiments on

ImageNet-1K, ImageNet-200, CIFAR-100, and CIFAR-10 using ResNet18 and ResNet50. Empirical results show that the norm $||\mathbf{b}||$ lies within the range $[0.05, 0.35]$, and the relative magnitude $\frac{||\mathbf{b}||}{\sigma_{\max}}$ remains around 2%–4%. This indicates that the additive term $||\mathbf{b}||$ introduces only a marginal shift in $||\mathbf{f}||$, with the observed proportionality primarily governed by singular values, confirming the tightness of the bound. As shown in Fig. 2c, there exists a proportional relationship between $||\mathbf{f}||$ and $||\mathbf{z}||$. $||\mathbf{z}||$ represents the distance to the origin, which implies that LogitNorm implicitly enforces the network based on feature distance from the origin, thereby encouraging feature collapse toward the origin.

3.3. Extended Logit Normalization

We have empirically and theoretically shown that LogitNorm implicitly constrains feature representations based on their distance from the origin $||\mathbf{z}||$, which can lead to feature collapse. Alternatively, this notion of distance can be generalized from the origin to the decision boundaries. Intuitively, samples closer to a boundary exhibit higher uncertainty, while those farther away are classified with greater confidence. Therefore, a well-calibrated feature representation should capture each sample's proximity to the decision boundaries rather than collapsing toward the origin.

To this end, we replace the scaling factor based on the distance to a singular point (the origin) with one based on the distance to the decision boundaries, introducing a new scaling mechanism. This extends LogitNorm into a more general formulation, which we refer to as **ELogitNorm**. Let $f_{\max}$ denote the index of the maximum entry in the logit vector $\mathbf{W}^\top \mathbf{z} + \mathbf{b}$, defined as:

$$f_{\max} = \arg\max_{i \in \{1,\dots,c\}} \left(\mathbf{W}^\top \mathbf{z} + \mathbf{b}\right)_i . \tag{6}$$

Then, the average distance to decision boundaries for a given feature vector $\mathbf{z}$, whose predicted class corresponds to $f_{\max}$, is defined as a point to a plane equation:

$$\mathcal{D}(\mathbf{z}) := \frac{1}{c-1} \sum_{i \neq f_{\max}}^{c} \frac{\left|(\mathbf{w}_{f_{\max}} - \mathbf{w}_i)^T \mathbf{z} + (b_{f_{\max}} - b_i)\right|}{\|\mathbf{w}_{f_{\max}} - \mathbf{w}_i\|_2}, \tag{7}$$

where $f_{\max}$ is the predicted class index, and the summation excludes the term corresponding to $f_{\max}$. We replace the scaling factor in Eq. (3) with $s = \mathcal{D}(\mathbf{z})$, leading to our **training objective**:

$$\mathcal{L}_{\text{ELogitNorm}}(f(\mathbf{x};\theta), y) = -\log \frac{e^{f_y/\mathcal{D}(\mathbf{z})}}{\sum_{i=1}^{k} e^{f_i/\mathcal{D}(\mathbf{z})}}. \tag{8}$$

LogitNorm implicitly uses $||\mathbf{z}||$ as a proxy for margin, we replace it with the actual multi-class margin, computed via average point-to-plane distances to all competing classes.

Proposition 2. *(Dimension of Minimum Scaling Factor Space) If $m \geq c-1$, the minimum distance to decision boundaries $\mathcal{D}_{\min}(\mathbf{z}) = 0$ is attained when $\mathbf{z} \in \bigcap_{i \neq f_{\max}} H_{i\ f_{\max}}$, where $H_{i\ f_{\max}}$ denotes the decision boundary between the predicted class $f_{\max}$ and class i. This intersection forms an affine subspace of dimension $m - c + 1$.*

From Proposition 2 (see proof in Appendix B), the minimum scaling factor space in **ELogitNorm** exhibits a substantially higher dimensionality compared to that of LogitNorm. For instance, in a ResNet-18 model trained on CIFAR-10, where the feature dimension is $m = 512$ and the number of classes is $c = 10$, the resulting affine subspace has a dimension of 503. In contrast, LogitNorm enforces a minimum norm constraint $||\mathbf{z}||_{\min} = 0$, which corresponds to a singular point at the origin. Since the minimum scaling factor space represents the optimal region for neural network optimization, our method effectively prevents collapse to a singular point. As illustrated in Fig. 2a, the singular value spectrum in **ELogitNorm** is more evenly distributed, avoiding the dominance of a few singular vectors. Furthermore, as demonstrated in Section 4 and Fig. 3, our proposed hyperparameter-free **ELogitNorm** consistently improves multiple post-hoc OOD scoring methods, whereas LogitNorm often underperforms or fails to generalize effectively.

4. Experiments

In this section, we describe the experimental setup in detail and evaluate our method on several standard benchmarks, including small-scale OOD benchmarks (*i.e.* CIFAR-10 and CIFAR-100), and a large-scale OOD benchmark (*i.e.* ImageNet-200 and ImageNet-1k). We closely follow the OpenOOD evaluation[1] [56]. All experiments are repeated three times with different random seeds, and we report the average performance. We run our experiments on an NVIDIA A100.

4.1. Experiment settings

Benchmark. We evaluate *ELogitNorm* using OpenOOD [56], which includes four vision benchmarks: *CIFAR-10* [22], *CIFAR-100* [22], *ImageNet-200* [40], and *ImageNet-1k* [40] for training-based OOD detection methods. Each benchmark consists of an in-distribution dataset $\mathcal{D}_{\text{ID}}$ and multiple out-of-distribution (OOD) datasets $\mathcal{D}_{\text{OOD}}$, further categorized into *Near-OOD* and *Far-OOD* datasets. The distinction between "near" and "far" is based on the similarity of OOD samples to ID samples, with Near-OOD samples posing a greater challenge for separation. OpenOOD also provides pre-trained model checkpoints for CIFAR-10,

[1]The codebase is: https://github.com/Jingkang50/OpenOOD

Table 1. *Per-Dataset Performance of OOD Detection Methods and Their ELogitNorm-Enhanced Variants (denoted with $*$).* The image encoder is ResNet18. The ID dataset is **CIFAR-10**. Blue indicates improvement, and orange indicates degradation.

Method	CIFAR-100		TIN		Near-OOD		MNIST		SVHN		Textures		Places365		Far-OOD	
	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR
MSP [15]	87.19	53.08	88.87	43.27	88.03	48.17	92.63	23.64	91.46	25.82	89.89	34.96	88.92	42.47	90.73	31.72
MSP*	91.05	33.27	93.77	23.99	92.89 (↑4.86)	26.49(↓21.68)	99.12	3.96	98.31	7.66	94.61	23.21	94.22	22.81	96.68 (↑5.95)	13.73 (↓17.99)
ReAct [42]	85.93	67.40	88.29	59.71	87.11	63.56	92.81	33.77	89.12	50.23	89.38	51.42	90.35	44.20	90.42	44.90
ReAct*	90.79	33.97	93.83	23.74	92.31 (↑5.20)	28.86 (↓34.70)	99.44	2.53	98.45	7.13	94.79	22.77	94.43	22.54	96.78 (↑6.36)	13.74 (↓31.16)
KNN [43]	89.73	37.64	91.56	30.37	90.64	34.01	94.26	20.05	92.67	22.60	93.16	24.06	91.77	30.38	92.96	24.27
KNN*	90.96	33.23	93.29	25.88	92.56 (↑1.92)	28.06 (↓5.95)	98.53	7.20	97.81	11.44	95.74	19.82	93.96	24.07	96.60 (↑3.64)	15.19 (↓9.08)
GEN [28]	87.21	58.75	89.20	48.59	88.20	53.67	93.83	23.00	91.97	28.14	90.14	40.74	89.46	47.03	91.35	34.73
GEN*	90.96	33.58	93.88	23.83	92.42 (↑4.22)	28.70 (↓24.97)	99.41	2.74	98.41	7.29	94.67	23.48	94.44	22.50	96.73 (↑5.38)	14.00 (↓20.73)
fDBD [26]	89.56	39.61	91.65	30.57	90.61	35.09	94.71	19.33	92.93	22.50	93.13	24.35	92.01	29.15	93.19	23.84
fDBD*	90.31	36.28	93.44	24.72	91.87 (↑1.26)	30.50 (↓4.59)	99.28	3.31	98.67	6.05	95.78	17.93	93.94	23.23	96.92 (↑3.73)	12.63 (↓11.21)
SCALE [51]	81.27	81.78	83.98	78.87	82.62	80.32	90.58	48.69	84.91	70.17	83.93	80.54	86.41	70.57	86.46	67.49
SCALE*	90.86	34.17	93.86	23.75	92.36 (↑9.74)	28.96 (↓51.36)	99.54	2.09	98.78	5.79	95.23	21.63	94.19	23.22	96.94 (↑10.48)	13.18 (↓54.31)

Method	SSB-hard		NINCO		Near-OOD		iNaturalist		Textures		OpenImage-O		Far-OOD	
	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR
Cross-Entropy	**72.09**	**74.49**	79.95	56.88	76.02	65.68	88.41	43.34	82.43	60.87	84.86	50.13	85.23	51.45
LogitNorm [49]	67.50	82.08	81.73	55.04	74.62	68.56	94.57	20.75	89.30	40.82	90.75	32.38	91.54	31.32
ELogitNorm	70.52	76.85	**83.24**	**52.22**	**76.88**	**64.54**	**96.15**	**16.56**	**91.46**	**36.58**	**91.97**	**30.07**	**93.19**	**27.74**

Table 2. *Per-Dataset Performance of OOD Detection Training Methods including Cross-Entropy, LogitNorm [49], and **ELogitNorm** (Ours).* The image encoders is ResNet50. The ID dataset is **ImageNet1K**. MSP [15] is used for OOD score.

CIFAR-100, ImageNet-200, and ImageNet-1k trained using standard Cross-Entropy loss. We primarily evaluate the improvements achieved by our training method over Cross-Entropy, while also comparing against alternative training methods under the same fixed post-hoc detection techniques.

Metric. We employ two conventional metrics to evaluate OOD detection performance. The first is a threshold-independent metric: Area Under the Receiver Operating Characteristic Curve (AUROC), where higher percentages indicate better performance. The second metric is the False Positive Rate at 95% True Positive Rate (FPR95), where lower percentages reflect better performance.

Implementation Details. We train ResNet-18 on CIFAR-10/100 for 100 epochs and on ImageNet-200 for 90 epochs using SGD with momentum 0.9, weight decay 5×10^{-4}, batch size 128, and an initial learning rate of 0.1 with standard scheduling. For ImageNet-1K, we finetune a ResNet-50 for 30 epochs with a learning rate of 0.001. Our method is hyperparameter-free and adds only minimal computational overhead, with an efficient implementation that scales well even for large numbers of classes $c = 1000$ (see Appendix).

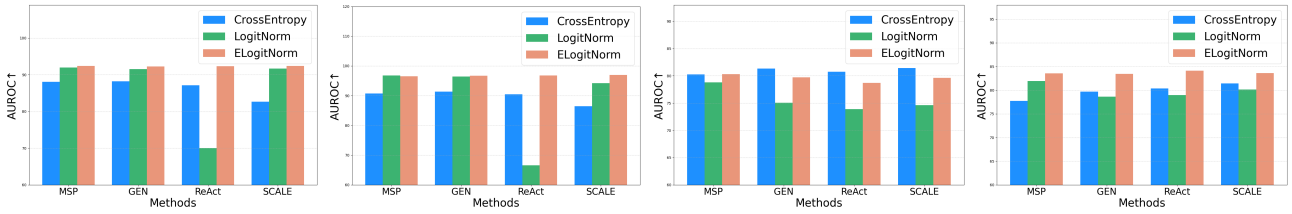
4.2. Main results

In this section, we demonstrate the effectiveness of our method from three perspectives: (a) *OOD detection*, where we evaluate its ability to enhance post-hoc methods and compare it with other training-based approaches, (b) *model calibration*, measured by the expected calibration error and (c) in-distribution data *classification accuracy*.

ELogitNorm enhances post-hoc OOD detection. The baseline post-hoc methods include OOD scoring techniques such as MSP [15], GEN [28], KNN [43], ReAct [42] and SCALE [51]. Table 1 demonstrates that **ELogitNorm** significantly improves OOD detection performance in both AUROC and FPR95 across all methods. A notable observation is that **ELogitNorm** leads to substantial improvements in far-OOD detection compared to near-OOD detection. For instance, in SCALE [51], the AUROC for far-OOD datasets improves by +10.48%, while the FPR95 is reduced by 54.31%. This pattern is consistent across other methods, where enhancements are more pronounced in far-OOD scenarios. Results for CIFAR-100 and ImageNet-200 show that the improvements are particularly noticeable in far-OOD datasets and MSP [15] shows minor performance trade-offs on some near-OOD datasets, details can be found in the supplementary material.

ELogitNorm enables better enhancement than LogitNorm and alternatives. As shown in Table 2, **ELogitNorm** consistently improves OOD detection on ImageNet-1K, with especially strong gains in far-OOD (FPR95 reduced from 51.45% to 27.74%), exceeding the improvement of LogitNorm. For near-OOD, **ELogitNorm** remains stable even when LogitNorm degrades. Similar trends appear on CIFAR-10/100, where both methods offer improve-

Figure 3. *Average far-OOD and near-OOD performance with 4 post-hoc methods MSP [15], ReAct [42], GEN [28] and SCALE [51].* ResNet18 are trained by Cross-Entropy, LogitNorm [49], and **ELogitNorm** (Ours), respectively. ID data are CIFAR-10 and CIFAR-100.



(a) CIFAR-10, near-OOD (b) CIFAR-10, far-OOD (c) CIFAR-100, near-OOD (d) CIFAR-100, far-OOD

Method	CIFAR10				CIFAR100				ImageNet-200			
	Near-OOD		Far-OOD		Near-OOD		Far-OOD		Near-OOD		Far-OOD	
	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR
MSP [15]												
Cross-Entropy	88.03	48.17	90.73	31.72	**80.27**	**54.80**	77.76	58.70	**83.34**	**54.82**	90.13	35.43
LogitNorm [49]	92.33	29.34	**96.74**	13.81	78.47	62.89	81.53	53.61	82.66	56.46	93.04	26.11
ELogitNorm (Our)	**92.89**	**26.49**	96.68	**13.73**	79.68	59.48	**84.51**	**46.86**	83.06	54.99	**93.58**	**25.08**
KNN [43]												
Cross-Entropy	90.64	34.01	92.96	24.27	**80.18**	61.22	82.40	53.65	81.57	60.18	93.16	27.27
CIDER [31]	90.71	32.11	94.71	20.72	73.10	72.02	80.49	54.22	80.58	60.10	90.66	30.17
NPOS [45]	89.78	32.64	94.07	20.59	78.35	63.35	82.29	**51.13**	**84.37**	62.09	94.83	21.76
ELogitNorm (Our)	**92.56**	**28.06**	**96.60**	**15.19**	80.12	**58.10**	**82.84**	54.23	82.73	**56.65**	**96.08**	**18.04**

Table 3. *Far-OOD and near-OOD performance from different training methods with fixed OOD scoring function.* The image encoders is ResNet18. The ID datasets are **CIFAR-10**, **CIFAR-100**, and **ImageNet-200**. KNN [43] and MSP [15] used as post-hoc methods.

ments, but **ELogitNorm** delivers stronger and more reliable performance, avoiding the severe degradation LogitNorm encounters when combined with ReAct [42] (Section 3). Results on ImageNet-200 further confirm this robustness. As summarized in Table 2 and Fig. 3, **ELogitNorm** provides broader compatibility and stronger overall enhancement across post-hoc methods. When comparing with other training-based approaches under fixed post-hoc scoring (MSP [15], KNN [43]; Table 3), **ELogitNorm** again achieves superior results. Despite CIDER [31] and NPOS [45] being designed for KNN [43], our method attains higher performance—for example, an AUROC of 96.08 on ImageNet-200, surpassing CIDER [31] (90.66) and NPOS [45] (94.83). Although minor drops on near-OOD are observed, such degradation is common for all training-time methods [56]. Notably, **ELogitNorm** maintains more stable near-OOD performance than alternatives, representing a modest but meaningful step toward narrowing this long-standing performance gap.

Calibration performance. Expected calibration error (ECE) is a standard and widely adopted metric for quantifying confidence calibration in multi-class classifiers. Table 4 reports ECE under various training objectives and logit-scaling strategies, allowing a controlled and comprehensive comparison. We consider three forms of scaled logits: the raw prediction vector $\mathbf{f}$, the normalized logits $\frac{\mathbf{f}}{\tau\|\mathbf{f}\|}$ derived from LogitNorm with a learned temperature τ, and the boundary-aware scaling $\frac{\mathbf{f}}{\mathcal{D}(\mathbf{z})}$, where $\mathcal{D}(\mathbf{z})$ measures the average distance from input $\mathbf{z}$ to nearby decision boundaries. This setup highlights how calibration quality depends jointly on the training loss and the normalization applied at inference. For instance, cross-entropy achieves its best calibration performance when evaluated with unscaled logits, while LogitNorm attains its lowest ECE when logits are normalized by the product of the learned temperature and their magnitude. Across all settings, however, our method demonstrates consistently superior calibration, achieving the lowest ECE regardless of the loss function or scaling strategy, thereby indicating improved robustness and stability in calibration performance.

Method	$\mathbf{f}$	$\frac{\mathbf{f}}{\tau\|\mathbf{f}\|}$	$\frac{\mathbf{f}}{\mathcal{D}(\mathbf{z})}$
Cross-Entropy	**3.3**	4.8	23.3
LogitNorm [49]	58.7	**4.1**	52.3
ELogitNorm (Ours)	26.7	4.7	**1.8**

Table 4. *Expected Calibration Error (ECE) (%).* The binning size is set to be 15. The model is ResNet18 trained on CIFAR-10.

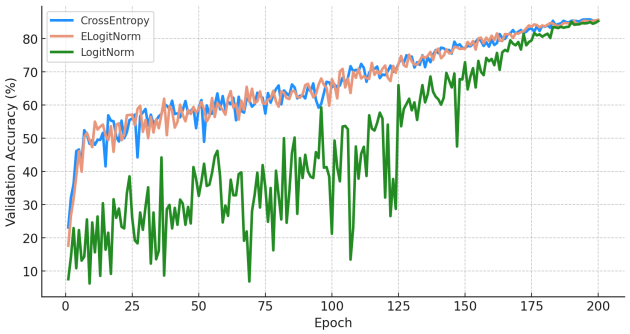


Figure 4. *Validation accuracy during training.* ResNet18 trained on ImageNet-200 over 200 training epochs. The proposed **ELogitNorm** (Ours) achieves a stable training curve compared to LogitNorm [49], while maintaining competitive final accuracy as CrossEntropy.

Method	CIFAR-10	CIFAR-100	ImageNet200
100 Epochs / 90 Epoch for ImageNet200			
Cross-Entropy	**95.06**	**77.25**	86.37
LogitNorm [49]	94.30	76.34	86.04
ELogitNorm (Ours)	94.47	76.67	**86.39**
200 Epochs			
Cross-Entropy	95.10	**77.47**	86.58
LogitNorm [49]	94.83	76.06	86.41
ELogitNorm (Ours)	**95.11**	77.37	**87.12**

Table 5. *ID Accuracy.* Comparison of Cross-Entropy, LogitNorm [49] and **ELogitNorm** (Ours) on **CIFAR-10**, **CIFAR-100** and **ImageNet-200** datasets.

Classification accuracy. A reliable classifier should not only detect OOD samples but also maintain strong ID performance. Table 5 reports accuracy on CIFAR-10, CIFAR-100, and ImageNet-200. The model trained with LogitNorm shows a clear degradation in accuracy compared to the standard cross-entropy baseline across all datasets. In contrast, **ELogitNorm** attains the highest accuracy on CIFAR-10 and ImageNet-200. As illustrated in Fig. 4, LogitNorm also displays unstable training behavior and inferior classification performance relative to the other two methods. Additional results for the remaining datasets are provided in the Appendix.

5. Discussion

Distance Awareness. Distance awareness has been extensively studied in OOD detection. KNN [43] estimates OOD likelihood based on the distance to the nearest neighbor, while fDBD [26], closely related to our approach, assumes that OOD samples lie near decision boundaries and designs its scoring function accordingly. CIDER [31] also incorporates distance awareness by optimizing feature representations to maximize inter-class separation and enforce intra-class compactness. Inspired by these approaches, we extend the formulation of LogitNorm by incorporating distances to decision boundaries, leading to our proposed method, **ELogitNorm**. By amplifying distance awareness during training, **ELogitNorm** significantly enhances distance-based OOD scoring techniques such as KNN [43] and fDBD [26]. This perspective opens a broader direction for refining feature representations through alternative distance metrics to further improve OOD detection. By examining the relationship between ID and OOD data in the classifier's representation space, we advocate for a deeper understanding of representation geometry and distance-based separation within learned feature spaces.

Adaptive temperature scaling Temperature scaling [13] applies a fixed scaling factor to all logits, whereas both LogitNorm and ELogitNorm adopt an *adaptive temperature scaling* mechanism, where the scaling factor varies for each sample. We define this adaptive scaling in a general form:

$$\mathcal{L}_s(f(\mathbf{x};\theta),y) = -\log \frac{e^{f_y/s}}{\sum_{i=1}^{c} e^{f_i/s}}, \tag{9}$$

where $s = \tau||\mathbf{f}||$ in Eq. (3) and $s = \mathcal{D}(\mathbf{z})$ in Eq. (8). The training of both objectives can thus be interpreted as a form of dynamic calibration governed by the sample-dependent scaling factor s. Notably, in LogitNorm, $s = \tau||\mathbf{f}||$ depends solely on the logits $\mathbf{f}$, whereas in ELogitNorm, $s = \mathcal{D}(\mathbf{z})$ incorporates information from the feature space. This allows ELogitNorm to achieve better calibration by explicitly accounting for each sample's relative position within the feature space. By aligning scaling with decision boundaries, **ELogitNorm** improves OOD separation while preserving strong ID discrimination. This is closely related to uncertainty estimation, where confidence scores should reflect epistemic uncertainty. Moreover, the definition of *adaptive temperature scaling* s suggests alternative strategies for further improving OOD detection and confidence calibration, potentially through uncertainty-aware representations [33].

6. Related work

OOD Detection OOD detection has been extensively studied from three perspectives: 1) *designing OOD scoring functions* using information available at different stages of a pre-trained classifier. Maximum softmax probability (MSP) [15] and GEN [28] operate in the probability space (*i.e.*, after the Softmax layer). Max-Logit [17], energy score [27], and NN-guide [35] utilize information from logits. Other methods operate in the feature

space and commonly require access to training data statistics [1, 2, 6, 17, 21, 26], which restricts their applicability. ViM [48] combines information from both feature and logit spaces. 2) *reshaping the extracted features* to enhance OOD performance. These methods typically work in a post-hoc manner and do not interfere with training [7, 25, 41, 42, 51]. In particular, ReAct [42] observes that OOD features tend to have unusually large activations in the feature space (*i.e.*, features extracted before the penultimate layer). To bring activations into a "normal" range, ReAct clips features whose magnitude exceeds the p-th percentile estimated on ID data. ASH [7] and SCALE [51] apply similar rescaling strategies to penultimate-layer features, and both require training data statistics to set the rescaling factor. ASH only rescales a specified portion of the features, whereas SCALE does not. SCALE can also be incorporated during training, but shows limited performance on small-scale OOD benchmarks, as in Fig. 1. Our method not only improves baseline OOD scoring methods, but also further boosts enhancement methods. 3) *new training objectives*. Such methods either add a regularization loss to the cross-entropy loss [8, 9, 16, 44] or reformulate the problem as a deep metric learning problem [30, 45]. The regularization loss often relies on real OOD samples [16] or synthesized OOD samples/features [8, 9]. However, these methods commonly involve multiple training stages or additional hyperparameters [37, 38], which makes them less favorable in practice. Our method falls into this category and is completely hyperparameter-free. Compared to LogitNorm [49], our method does not require a held-out dataset to select a proper temperature value. Moreover, it works well with a wide range of OOD scoring methods, including MSP [15], KNN [43], GEN [28], and fDBD [26], as well as enhancement methods such as ReAct [42] and SCALE [51]. Feature normalization has also been explored as a post-hoc strategy in [34, 52]. Beyond these approaches, several works address broader problems such as OOD generalization and detection, including Scone [4], AHA [5], and InfoBound [57].

Confidence Calibration Modern neural networks trained with trained with cross-entropy loss are prone to be over-confident [12]. Later work [50] empirically shows that models tend to be under-confident for lower proximity samples and over-confident for higher proximity samples. To mitigate the issue of miscalibration, there are roughly two types of methods consisting of 1) post hoc methods and 2) training loss modification. Post hoc methods includes parameterized temperature scaling [46], Mix-n-Match [55] and binning-based methods such as classic histogram binning [53], mutual information maximization-based binning [36], and isotonic regression [54]. Another line of work requires training with additional loss such as [37] and logit-norm [49].

Feature collapse. Feature collapse occurs when learned representations degenerate into a low-diversity, low-rank subspace that fails to distinguish samples from different classes. In self-supervised learning, [19] showed that whitening batch normalization can mitigate collapse. [20] provided a theoretical analysis of dimensional collapse in contrastive methods, and several subsequent works linked this phenomenon to performance degradation [10, 11, 14, 24]. Related issues also arise in generative modeling: GANs are well known to exhibit mode collapse [3, 32], and vector-quantized models can similarly suffer from feature collapse that reduces generative quality [58]. Overall, preventing collapse requires balancing representation compactness and expressiveness so that embeddings retain meaningful semantic structure rather than converging to trivial solutions. This balance is also critical for OOD detection, where overly uniform or over-regularized features reduce inter-class separation and degrade sensitivity.

7. Conclusion

In this paper, we introduced **ELogitNorm**, a generalized and hyperparameter-free extension of LogitNorm that leverages distances to decision boundaries in the feature space of neural network classifiers. Our approach effectively mitigates the feature collapse issue inherent in LogitNorm, leading to improved OOD detection, better confidence calibration, and preserved classification accuracy. Extensive experiments demonstrate that **ELogitNorm** consistently enhances the OOD detection capability of post-hoc methods and achieves superior performance compared to prior training-based approaches, particularly on far-OOD benchmarks. While the improvements on near-OOD settings are more modest, the results remain robust and stable across architectures and datasets. Beyond performance gains, this work underscores the importance of understanding the geometric structure of feature representations in deep networks. We hope that **ELogitNorm** will serve as a foundation for future research exploring boundary-aware calibration, adaptive scaling mechanisms, and more principled ways to align feature geometry for reliable open-world recognition.

Acknowledgement

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the WASP NEST _main_, the strategic research environment ELLIIT, and the Zenith career development program at Linköping University. The computational resources were provided by the National Academic Infrastructure for Super computing and the National Supercomputer Centre in Sweden. We thank Jens Sjölund for interesting discussions.

References

[1] Amirhossein Ahmadian, Yifan Ding, Gabriel Eilertsen, and Fredrik Lindsten. Unsupervised novelty detection in pretrained representation space with locally adapted likelihood ratio. In *AISTATS*, 2024. 8

[2] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. NECO: NEural collapse based out-of-distribution detection. In *ICLR*, 2024. 8

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 8

[4] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D. Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *ICML*, 2023. 8

[5] Haoyue Bai, Jifan Zhang, and Robert Nowak. Aha: Adaptive human-assisted out-of-distribution generalization and detection. In *NeurIPS*, 2024. 8

[6] Yifan Ding, Arturas Aleksandraus, Amirhossein Ahmadian, Jonas Unger, Fredrik Lindsten, and Gabriel Eilertsen. Revisiting likelihood-based out-of-distribution detection by modeling representations. In *SCIA*. Springer, 2025. 8

[7] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection, 2023. 1, 8

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

[9] Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *NeurIPS*, 2023. 1, 8

[10] Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *ICML*, 2023. 8

[11] Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022. 8

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 8

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 7

[14] Bobby He and Mete Ozay. Exploring the gap between collapsed & whitened features in self-supervised learning. In *ICML*, 2022. 8

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 2, 5, 6, 7, 8

[16] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 8

[17] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022. 1, 2, 7, 8

[18] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, 2020. 1

[19] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, 2021. 3, 8

[20] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021. 8

[21] Ryo Kamoi and Kei Kobayashi. Why is the mahalanobis distance effective for anomaly detection? *arXiv preprint arXiv:2003.00402*, 2020. 8

[22] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 4

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

[24] Alexander C Li, Alexei A Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning. In *ECCV*, 2022. 8

[25] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 8

[26] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. *arXiv preprint arXiv:2312.11536*, 2023. 5, 7, 8

[27] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 1, 2, 7

[28] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *CVPR*, 2023. 1, 2, 5, 6, 7, 8

[29] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3

[30] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv:2203.04450*, 2022. 1, 8

[31] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv preprint arXiv:2203.04450*, 2022. 6, 7

[32] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 8

[33] Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *NeurIPS*, 2024. 7

[34] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *CVPR*, 2023. 8

[35] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *ICCV*, 2023. 7

[36] Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. In *ICLR*, 2021. 8

[37] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR*, 2017. 8

[38] Sudarshan Regmi, Bibek Panthi, Sakar Dotel, Prashnna K Gyawali, Danail Stoyanov, and Binod Bhattarai. T2FNorm: Train-time Feature Normalization for OOD Detection in Image Classification. In *CVPRW*, 2024. 8

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 4

[41] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *NeurIPS*, 2022. 1, 8

[42] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 1, 5, 6, 8

[43] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022. 5, 6, 7, 8

[44] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *NeurIPS*, 2020. 8

[45] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *ICLR*, 2023. 1, 6, 8

[46] Christian Tomani, Daniel Cremers, and Florian Buettner. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *ECCV*, 2022. 8

[47] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 3

[48] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022. 1, 8

[49] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022. 1, 2, 3, 5, 6, 7, 8

[50] Miao Xiong, Ailin Deng, Pang Wei Koh, Jiaying Wu, Shen Li, Jianqing Xu, and Bryan Hooi. Proximity-informed calibration for deep neural networks. In *NeurIPS*, 2023. 8

[51] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *ICLR*, 2024. 1, 2, 5, 6, 8

[52] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *CVPR*, 2023. 8

[53] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, 2001. 8

[54] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD*, 2002. 8

[55] Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, 2020. 8

[56] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. In *DMLR*, 2024. 1, 2, 4, 6

[57] Lin Zhu, Yifeng Yang, Zichao Nie, Yuan Gao, Jiarui Li, Qinying Gu, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. Infobound: A provable information-bounds inspired framework for both ood generalization and ood detection. *TPAMI*, 2025. 8

[58] Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *ICCV*, 2025. 8