Multi-Method Ensemble for Out-of-Distribution Detection

Lucas Rakotoarivony
lucas.rakotoarivony@thalesgroup.com

Thales, cortAIx Labs
Palaiseau, 91120, France

Abstract

Detecting out-of-distribution (OOD) samples is essential for neural networks operating in open-world settings, particularly in safety-critical applications. Existing methods have improved OOD detection by leveraging two main techniques: feature truncation, which increases the separation between in-distribution (ID) and OOD samples, and scoring functions, which assign scores to distinguish between ID and OOD data. However, most approaches either focus on a single family of techniques or evaluate their effectiveness on a specific type of OOD dataset, overlooking the potential of combining multiple existing solutions. Motivated by this observation, we theoretically and empirically demonstrate that state-of-the-art feature truncation and scoring functions can be effectively combined. Moreover, we show that aggregating multiple scoring functions enhances robustness against various types of OOD samples. Based on these insights, we propose the Multi-Method Ensemble (MME) score, which unifies state-of-the-art OOD detectors into a single, more effective scoring function. Extensive experiments on both large-scale and small-scale benchmarks, covering near-OOD and far-OOD scenarios, show that MME significantly outperforms recent state-of-the-art methods across all benchmarks. Notably, using the BiT model, our method achieves an average FPR95 of 27.57% on the challenging ImageNet-1K benchmark, improving performance by 6% over the best existing baseline.

1 Introduction

With the growing deployment of neural networks in real-world applications, they may encounter samples outside their trained class set. The unpredictable nature of such out-of-distribution (OOD) instances can prevent the model from making confident and reliable predictions [36]. Therefore, OOD detection is essential for ensuring the safety of machine learning systems, particularly in high-stakes applications such as autonomous driving [2, 30], medical imaging [18, 23] and biosynthesis [27].

The standard approach to OOD detection involves deriving a score function from a trained neural network, where in-distribution (ID) samples receive relatively higher scores than OOD samples. Researchers have developed scoring functions using either feature representations [13, 26] or logit values [9, 16], aiming to identify properties that naturally distinguish ID and OOD samples. More recent work [6, 20, 31] has even combined both types of information sources to better handle the wide variety of OOD instances. However, while combining multiple inputs has been shown to be beneficial, these methods some-

© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

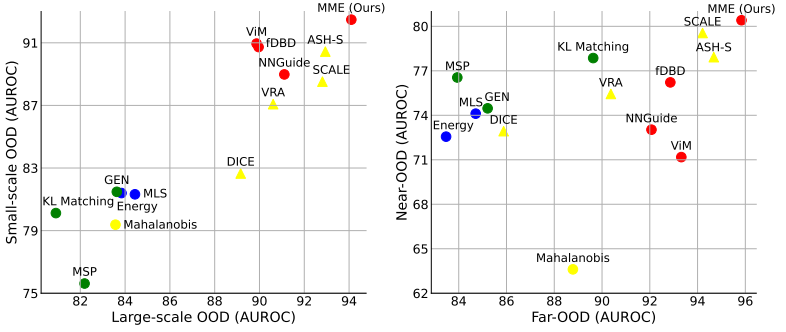

Figure 1: AUROC performance trade-offs for 14 algorithms, comparing small- vs. large-scale OOD detection (left) and near-OOD vs. far-OOD detection (right). Results are averaged across multiple datasets. For the small-scale setup, a DenseNet-100 model trained on CIFAR-100 is used, while the large-scale and near/far-OOD setups use a BiT model trained on ImageNet-1K. △ denotes feature truncation methods combined with the Energy score, while ◯ represents scoring functions. Green symbols indicate probability-based methods, blue symbols correspond to logit-based methods, yellow symbols represent feature-based methods, and red symbols denote approaches that leverage multiple sources of information.

times achieve suboptimal performance in specific applications [38], even compared to simpler scoring functions. In contrast, feature truncation methods [24, 25] aim to increase the separation between ID and OOD samples by modifying the network's propagated signals or weights. Despite their simplicity, feature truncation techniques have been shown to effectively reduce the overlap between ID and OOD samples at both the feature and logit levels. However, these methods cannot be used independently and are typically combined with simple scoring functions, such as Energy [16] or MLS [9, 29], for OOD detection. Therefore, even state-of-the-art feature truncation methods [33, 34], when paired with such scoring functions, often fail to outperform more advanced scoring methods due to the inherent simplicity of their approach.

In this paper, we empirically and theoretically demonstrate that combining multiple scoring functions with feature truncation methods improves OOD detection by better handling the diverse range of OOD examples encountered in real-world scenarios. Building on this ensemble approach, we propose a Multi-Method Ensemble (MME) score, which integrates recent scoring functions with feature truncation techniques. Specifically, the scoring process begins by extracting features, which are then refined using SCALE [33] and VRA [34] feature truncation to obtain sparse and clipped representations. These processed features are subsequently used by state-of-the-art scoring functions, including fDBD [15], PCA [6], ViM [31], and our proposed NME$^+$ method, to generate multiple scalar scores. Finally, these individual scores are combined to produce a final OOD score. The design of MME intuitively ensures that even if one method fails to detect an OOD sample, another method is likely to do so, thereby improving overall detection performance, as illustrated in Figure 1.

To verify MME effectiveness, we perform experiments on several benchmark datasets and models, including CIFAR-10, CIFAR-100 [12], and the more challenging ImageNet-1K [22]. The results show that our method outperforms current post-hoc strategies, achieving state-of-the-art performance. The key contributions of this paper are summarized as follows:

- **Theory:** We theoretically demonstrate that combining multiple scoring functions improves OOD detection. Additionally, we empirically show that scoring functions generally perform better when fused with feature truncation methods.

- **Methodology:** We propose an effective post-hoc strategy called MME, which combines multiple state-of-the-art OOD detection methods to better handle the wide variety of OOD samples.

- **Performance:** We evaluate MME on popular benchmarks, including both large-scale and small-scale scenarios, as well as near-OOD and far-OOD cases. Our results show that MME consistently outperforms existing post-hoc methods for OOD detection.

2 Related Work

OOD detection research focuses on two approaches: feature truncation [25] and creating a scalar score function to distinguish between OOD and ID samples [8].

Feature truncation methods [24, 25] aim to better separate ID and OOD samples by altering network signals or weights. ReAct [25] clips high activation values to disrupt OOD signals while preserving ID ones. VRA [34] goes further by also suppressing low and amplifying mid-range activations. ASH [4] prunes and scales activations, while SCALE [33] avoids pruning to reduce ID-OOD overlap. However, feature truncation methods cannot be applied by themselves for detecting OOD and must be combined with a score function.

Scalar score functions for OOD detection are generally divided into classifier-based and distance-based approaches. Classifier-based methods, or confidence scores, use neural network outputs to derive scores. MSP [8] uses maximum softmax probability for ID detection, while MLS [9, 29] uses logits to preserve more information. Energy score [16] quantifies prediction uncertainty by mapping model outputs to an energy value. Distance-based methods determine OOD samples by measuring distances in feature space. The Mahalanobis detector [13] uses class-wise mean distances and shared covariances. In contrast, fDBD [15] focuses on distances from decision boundaries, showing ID features lie further away.

Classifier-based detectors effectively identify near-OOD instances near decision boundaries but exhibit overconfidence in far-OOD regions [7]. In contrast, distance-based methods perform well in far-OOD detection but struggle with near-OOD samples close to ID classes. Recent methods [6, 20, 31] address these limitations by combining classifier- and distance-based cues. ViM [31] combines virtual logits from feature residuals with an energy score, while NNGuide [20] improves logits-based confidence using KNN guidance in feature space.

3 Methods

3.1 Preliminaries

To ensure consistency with previous work [20, 25, 34], we define the OOD detection task as follows: let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be a label space containing k distinct categories. Consider a neural network $f : \mathcal{X} \to \mathbb{R}^k$ trained on samples drawn from the in-distribution data. The goal of OOD detection is to design a scoring function, S, that determines whether a test input $\mathbf{x} \in \mathcal{X}$ belongs to the ID or OOD space based on its score $S(\mathbf{x})$ and a threshold τ.

$$\mathbf{x} \in \begin{cases} \text{ID} & \text{if} \quad S(\mathbf{x}) \geq \tau \\ \text{OOD} & \text{if} \quad S(\mathbf{x}) < \tau \end{cases} \tag{1}$$

By convention, scoring functions are designed to produce high values for in-distribution data and low values for out-of-distribution data. Therefore, an effective scoring function should satisfy the following

$$\mathbb{E}_{\text{in}}[S(\mathbf{x})] \geq \mathbb{E}_{\text{out}}[S(\mathbf{x})] \tag{2}$$

Let define $\mathbb{V}(S) = \mathbb{E}_{\text{in}}[S(\mathbf{x})] - \mathbb{E}_{\text{out}}[S(\mathbf{x})]$, by definition, OOD detection objective is to find S that maximizes $\mathbb{V}(S)$.

Feature truncation aims to increase the separation between ID and OOD samples by adjusting the network's propagated signals or weights. Let g be a feature truncation method and z a feature vector, where $z = f(\mathbf{x})$. We adopt the terminology introduced by [25, 34].

$$\mathbb{E}_{\text{in}}[g(z)] - \mathbb{E}_{\text{out}}[g(z)] \geq \mathbb{E}_{\text{in}}[z] - \mathbb{E}_{\text{out}}[z] \tag{3}$$

For convenience, since scoring functions typically use feature vectors or logits to compute a confidence score, we adopt the notation $S(\mathbf{x})$ or $S(z)$, depending on the context.

In the Section 3.3, we will demonstrate that combining two scoring functions can improve OOD detection performance. To achieve this, we make the following assumption:

Assumption 1. *Let S_1 and S_2 be two scoring functions and $Cov(S_1, S_2)$ the covariance of S_1 and S_2. We assume $Cov_{in}(S_1, S_2) \geq Cov_{out}(S_1, S_2)$*

By definition, covariance is a measure of the degree to which two variables change together. Assumption 1 means that two confidence scores will more likely produce correlated values when encountering ID data rather than OOD samples. To validate Assumption 1, we will use two classifiers that can also be employed for OOD detection: the well-known MLS [8] classifier (classifier-based) and the Nearest-Mean-of-Exemplars (NME) classifier (distance-based) proposed by [21]. To ensure consistency with Equation (1), we will adapt the NME classifier to NME^+, defined as follows $\text{NME}^+(\mathbf{x}) = \max(\frac{1}{\text{softmax}(\frac{\text{NME}(\mathbf{x})}{T})})$, where T is a temperature parameter that controls softmax smoothness. For more details about NME^+, please refer to Appendix. Since NME^+ and MLS are two scoring functions that can be easily derived from classification tasks, we introduce a piecewise function called COnsistency (CO), along with its corresponding scoring function, CO^+, as follows:

$$\text{CO}(\mathbf{x}) = \begin{cases} 1, & \text{if } c_{\text{NME+}}(\mathbf{x}) = c_{\text{MLS}}(\mathbf{x}), \\ 0, & \text{otherwise.} \end{cases} . \tag{4}$$

$$\text{CO}^+(\mathbf{x}) = \begin{cases} \lambda, & \text{if } c_{\text{NME+}}(\mathbf{x}) = c_{\text{MLS}}(\mathbf{x}), \\ 1, & \text{otherwise.} \end{cases} . \tag{5}$$

Intuitively, CO can be interpreted as a proxy for the covariance between NME^+ and MLS, as it measures agreement between their class predictions (denoted by c), while CO^+ is a scoring function derived from CO. In CO^+, the parameter $\lambda \geq 1$ controls the influence of the prediction consistency between logits and features on the final score. Table 1 presents the consistency ratio for ID and OOD datasets across various models. From these results, we observe that NME^+ and MLS are more likely to make similar predictions when encountering ID data, compared to OOD data, which empirically validates Assumption 1. For simplicity, we use NME^+ and MLS as confidence scores in the definition of CO. However, as demonstrated in Appendix, the same reasoning applies equally well to other confidence scores.

	Dataset	ViT	BiT
ID	ImageNet	0.90	0.74
OOD	iNaturalist	0.43	0.32
	Places	0.66	0.52
	SUN	0.65	0.50
	Textures	0.43	0.38

Table 1: Consistency ratio evaluated on the ImageNet-1K dataset (ID) and four OOD datasets. Results are reported for both ViT and BiT architectures.

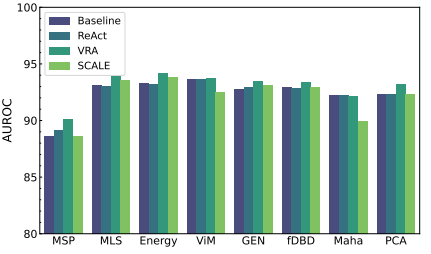

Figure 2: AUROC OOD detection performance on the ImageNet-1K benchmark for eight scoring functions combined with different feature truncation methods. Results are averaged across four OOD datasets.

3.2 Motivation

Current post-hoc OOD detection methods can be divided into two main categories: feature truncation [24, 25] and scoring functions [8, 16]. Feature truncation methods aim to increase the gap between ID and OOD samples through scaling, pruning or clipping operations on features. However, feature truncation alone cannot perform OOD detection and is generally paired with simple scoring functions such as MLS [8] or Energy [16]. Scoring functions use various types of inputs, such as logits or features, to produce a score that can distinguish between ID and OOD samples. Recent advancements [6, 20, 31] suggest using multiple input sources to better handle the diversity of OOD samples. These observations lead us to the following questions:

1. Does combining multiple scoring functions improve OOD detection?
2. Can feature truncation methods improve the performance of recent scoring functions?

In this paper, we answer these questions through both theoretical and empirical analysis.

3.3 Theoretical analysis

Combining multiple scoring function. Recent state-of-the-art OOD detection methods have explored the benefits of combining feature and logits information [6, 31] to better handle the diversity of OOD samples. To effectively merge these two sources of input, these methods typically combine the Energy score (logits-based information) with a custom scoring function based on features to achieve state-of-the-art performance. Building on these insights, we make the following proposition:

Proposition 1. *Let $n \in \mathbb{N}$ with $n \geq 2$, and let $S_1, \ldots, S_n$ be distinct scoring functions. Under Assumption 1 and using Equation (2), The inequality below is satisfied:*

$$\mathbb{V}(S_1 \cdot \ldots \cdot S_n) \geq \max\left(\mathbb{V}(S_1), \ldots, \mathbb{V}(S_n)\right).$$

Simplified proof. The full proof is in Appendix. For simplicity, we consider the case where $n = 2$, as the case $n > 2$ can be trivially derived by recurrence. Using the mathematical expression of the covariance and the scalability properties of a scoring function, we obtain:

$$\begin{aligned}\mathbb{V}(S_1 \cdot S_2) &\geq \mathbb{E}_{\text{in}}(S_1(\mathbf{x}))\mathbb{E}_{\text{in}}(S_2(\mathbf{x})) - \mathbb{E}_{\text{out}}(S_1(\mathbf{x}))\mathbb{E}_{\text{out}}(S_2(\mathbf{x})) \\ &\geq \mathbb{V}(S_1)\end{aligned} \tag{6}$$

By symmetry, we have $\mathbb{V}(S_1 \cdot S_2) \geq \mathbb{V}(S_2)$, which leads to $\mathbb{V}(S_1 \cdot S_2) \geq \max(\mathbb{V}(S_1), \mathbb{V}(S_2))$

Proposition 1 states that combining multiple scoring functions results in a scoring function that separates ID and OOD data at least as effectively as the best individual scoring function, thus answering our first question.

Pairing feature truncation with scoring function. The second question is partially addressed in the literature. For instance, VRA [34] demonstrates that feature truncation increases the gap between ID and OOD distributions, as shown in Equation (3). Additionally, ReAct [25] shows that clipping high activation values reduces both feature and logits activations more for OOD samples than for ID samples, leading to a greater separation between ID and OOD. Feature truncation methods [24, 25, 33, 34] indicate that this increased separation also transfers to the logits space and energy-based scores when combined with simple scoring functions, such as the Energy score [16] or MLS [8]. While recent scoring functions are more complex than Energy or MLS, they all rely on either features or logits as inputs. Based on these observations, we make the following hypothesis:

Hypothesis 1. *Let g be a feature truncation method such that $\mathbb{E}_{in}[g(z)] - \mathbb{E}_{out}[g(z)] \geq \mathbb{E}_{in}[z] - \mathbb{E}_{out}[z]$ and let S be a scoring function. The following inequality holds:*

$$\mathbb{E}_{in}[S(g(z))] - \mathbb{E}_{out}[S(g(z))] \geq \mathbb{E}_{in}[S(z)] - \mathbb{E}_{out}[S(z)]$$

While Hypothesis 1 is central to our method, a general proof is elusive due to the nature of the scoring function S, which typically relies on the output of a neural network. Consequently, S often does not satisfy properties such as monotonicity, Lipschitz continuity, or convexity. These conditions are commonly required to apply standard analytical tools like Jensen's inequality or continuity bounds. As a result, we treat this statement as a working hypothesis. We support it empirically in Figure 2, which shows that recent state-of-the-art scoring functions consistently benefit from feature truncation methods. It is important to note that Mahalanobis [13] is the only tested scoring function that does not benefit from feature truncation. This can be explained by the fact that, as a distance-based method, Mahalanobis relies on accurate feature distance values, which may be degraded by feature truncation.

3.4 MME implementation

Based on Proposition 1 and Hypothesis 1, we have shown that various feature truncation methods and scoring functions can be combined to create a single scoring function maximizing the gap between ID and OOD. However, although it is theoretically possible to combine an infinite number of scoring functions, this approach is impractical due to computational overhead and does not necessarily guarantee better performance. Therefore, we focus on recent state-of-the-art feature truncation methods, namely SCALE [33] and VRA [34], along with scoring functions such as fDBD [15], PCA [6] and ViM [31], as well as our proposed NME^+ and CO^+ functions, to construct MME scoring function. The MME function is defined as follows:

$$\text{MME}(\mathbf{x}) = \exp(\text{SCALE}(\mathbf{x}) - \text{ViM}(\text{VRA}(\mathbf{x}))) \cdot \text{fDBD}(\text{VRA}(\mathbf{x})) \cdot \text{PCA}(\text{VRA}(\mathbf{x})) \cdot \text{CO}^+(\mathbf{x}) \cdot \text{NME}^+(\mathbf{x}) \tag{7}$$

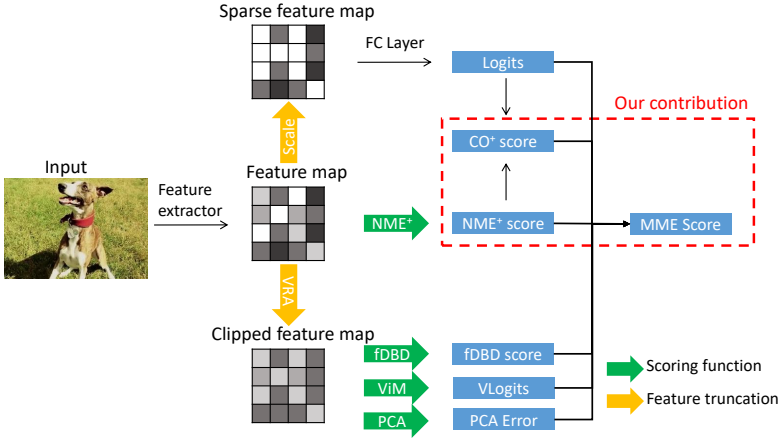

Figure 3: Overview of the MME pipeline. First, features are extracted from the input and processed using the VRA and SCALE feature truncation methods. Next, multiple scoring functions are applied to the feature maps to generate scalar scores. Finally, these scores are combined to produce MME score, which can be used with a threshold τ for OOD detection.

To ensure positivity of MME, we apply the exponential function to $\text{SCALE}(\mathbf{x}) - \text{ViM}(\text{VRA}(\mathbf{x}))$. Note that NME$^+$ is the only scoring function that does not use feature truncation, as it is a distance-based method, as suggested by Figure 2. Further details about the choice of combinations will be provided in Appendix. A complete overview of MME pipeline is provided by Figure 3.

4 Experiments

4.1 Experimental settings

Datasets MME was evaluated on three OOD detection benchmarks, each with training ID, test ID, and multiple OOD sets that don't class overlap with the ID sets. For small-scale benchmarks, CIFAR-10 and CIFAR-100 [12] serve as ID sets, with six OOD sets: LSUN-Crop [37], LSUN-Resize [37], iSUN [35], SVHN [19], Places365 [40], and Textures [3]. For large-scale OOD detection, ImageNet-1K [22] is the ID set, with iNaturalist [28], Places [40], SUN [32], and Textures [3] as OOD sets. Following recent studies [33, 38], ImageNet-1K [22] is used for near-OOD (NINCO [1], SSB-hard [29]) and far-OOD (iNaturalist [28], Textures [3], OpenImage-O [31]) detection. Refer to Appendix for dataset details.

Models Following previous studies [14, 17, 31], we used ViT-B/16 [8] (Transformer-based) and BiT-S-R101x1 [11] (CNN-based) as the classifier backbones for the ImageNet-1K benchmarks, while DenseNet-BC-100 [10] was used for the CIFAR benchmarks.

Methods To ensure a fair comparison, we compare MME with several state-of-the-art post-hoc OOD detectors. We consider 5 feature truncation methods combined with the Energy score: ReAct [25], DICE [24], ASH-S [4], SCALE [33] and VRA [34]. We also evaluate 11 scoring functions: MSP [8], MLS [8, 29], Energy [16], Mahalanobis [13] (Maha), KL Matching [9] (KL), ViM [31], SHE [39], GEN [17], PCA [6], NNGuide [20], and fDBD [15]. We use the same hyperparameters as in the original papers and reproduce results using

their code. Experiments were conducted on an NVIDIA GeForce RTX 3090, with CUDA 11.0 and PyTorch 1.7. In the following experiments, $T = 0.5$ for the ImageNet-1k benchmark and $T = 0.1$ for the CIFAR benchmark, while λ is set to 2 for both benchmarks. MME is robust to T and λ, requiring minimal hyperparameter tuning, as shown in Appendix.

Evaluation metrics We report two metrics: AUROC, the area under the ROC curve (higher is better), and FPR95 (false positive rate at 95% true positive rate, lower is better). Both metrics are presented as percentages, with the best results in bold.

4.2 Evaluation on ImageNet-1k

On the ImageNet-1K large-scale benchmark, Table 3 summarizes the OOD detection performance of each method across different OOD datasets, as well as the average performance over all four datasets. With the BiT backbone, MME outperforms all methods on three out of the four datasets in terms of FPR95, achieving an average FPR95 of 27.57%, an improvement of 6% over the best existing baseline. Furthermore, with the ViT model, MME achieves the highest performance on the Textures dataset and the best average performance overall. On the other three datasets, MME consistently delivers competitive results, generally ranking among the top methods (second or third), further highlighting its robustness.

Across all experiments, at least nine different methods achieve a top-three performance on a specific model and dataset, highlighting the sensitivity of existing OOD detectors to both the model architecture and dataset characteristics. In contrast, MME consistently ranks among the top three, demonstrating that our ensemble strategy enhances robustness and improves the detection of various types of OOD samples.

4.3 Evaluation on CIFAR benchmarks

On small-scale CIFAR benchmarks, the ensembling strategy also contributes to achieving state-of-the-art performance with greater consistency. Table 2 presents the average performance across six OOD datasets for both CIFAR-10 and CIFAR-100 as ID sets, demonstrating that MME outperforms existing post-hoc detectors. Notably, while ViM achieves the second-best performance on both CIFAR-10 and CIFAR-100 benchmarks and even comes within 1% of MME in terms of FPR95 on CIFAR-10, it falls significantly short on CIFAR-100. In this case, MME achieves an FPR95 of 28.96%, offering a substantial improvement of nearly 10% over ViM, which records an FPR95 of 38.36%.

ID Dataset	Metrics	Methods								
		MSP	MLS	ViM	ASH-S	GEN	SCALE	VRA	fDBD	MME
CIFAR-10	AUROC↑	92.17	93.48	96.58	94.93	94.88	93.90	94.92	96.14	**96.92**
	FPR95↓	47.11	29.63	16.48	23.36	30.01	28.59	27.14	20.52	**15.60**
CIFAR-100	AUROC↑	75.62	81.32	90.95	90.44	81.48	88.52	87.08	90.73	**92.48**
	FPR95↓	77.52	68.64	38.36	40.95	68.54	47.59	61.63	44.61	**28.96**

Table 2: Comparison between different methods in OOD detection on small-scale CIFAR benchmarks using a DenseNet model. Values are average percentages over six OOD datasets.

Method	iNaturalist		Places		SUN		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Backbone : BiT										
MSP (ICLR'17)	87.90	64.58	79.27	79.00	81.82	71.82	79.76	77.17	82.19	73.14
MLS (ICML'22)	86.78	70.51	82.58	74.36	86.74	63.49	81.65	73.60	84.44	70.49
Energy (NeurIPS'20)	84.52	74.96	82.56	73.21	87.17	60.81	81.10	73.91	83.84	70.72
ReAct (NeurIPS'21)	91.50	48.65	87.96	54.67	92.33	38.99	90.65	50.14	90.61	48.11
Maha (NeurIPS'18)	85.80	64.69	73.06	82.35	78.10	73.38	97.33	13.95	83.57	58.59
KL (ICML'22)	90.84	38.19	73.17	79.10	76.43	75.88	83.22	55.16	80.91	62.08
ViM (CVPR'22)	89.38	55.09	83.33	67.94	87.85	57.79	**98.92**	**4.63**	89.87	46.36
DICE (ECCV'22)	91.44	43.64	88.99	49.20	92.52	35.54	83.67	55.10	89.16	45.87
ASH-S (ICLR'23)	93.87	36.85	88.49	49.17	91.77	34.69	97.62	12.03	92.94	33.19
SHE (ICLR'23)	77.65	72.25	80.31	71.63	84.65	56.68	84.58	48.84	81.80	62.35
GEN (CVPR'23)	88.67	68.32	80.19	79.66	84.17	72.23	81.48	77.91	83.63	74.53
PCA (ICCV'23)	50.37	99.77	42.19	99.80	50.15	99.25	69.69	85.35	53.1	96.04
NNGuide (ICCV'23)	88.48	52.83	88.17	50.76	**92.92**	33.58	94.94	21.51	91.11	40.60
SCALE (ICLR'24)	93.77	39.37	88.96	50.10	92.39	35.28	96.11	18.99	92.81	35.94
VRA (NeurIPS'24)	91.50	48.65	87.96	54.67	92.33	38.99	90.65	50.14	90.61	48.11
fDBD (ICML'24)	91.71	51.42	84.20	67.45	88.47	56.03	95.45	24.38	89.96	49.82
MME (Ours)	**95.53**	**26.91**	**89.73**	**45.03**	92.70	**30.64**	98.45	7.71	**94.10**	**27.57**
Backbone : ViT										
MSP (ICLR'17)	96.13	19.15	85.09	59.97	86.13	57.00	87.13	48.45	88.62	46.14
MLS (ICML'22)	98.57	6.53	89.24	46.65	91.50	38.97	93.05	30.27	93.09	30.61
Energy (NeurIPS'20)	98.66	6.04	89.30	45.27	91.75	37.09	93.42	28.12	93.28	29.13
ReAct (NeurIPS'21)	99.01	4.19	89.11	47.37	91.54	39.04	93.38	28.37	93.26	29.74
Maha (NeurIPS'18)	**99.63**	2.15	86.29	60.04	89.11	51.29	94.21	25.27	92.31	34.69
KL (ICML'22)	96.62	16.45	84.34	63.50	85.91	59.41	88.26	45.99	88.78	46.34
ViM (CVPR'22)	99.41	2.56	88.47	50.44	91.53	39.48	95.31	20.14	93.68	28.15
DICE (ECCV'22)	65.49	91.69	61.56	93.88	64.75	88.51	75.49	74.26	66.82	87.09
ASH-S (ICLR'23)	61.96	95.34	61.23	95.68	65.52	92.90	56.79	91.12	61.38	93.76
SHE (ICLR'23)	90.69	43.50	85.63	59.25	87.97	51.93	92.18	32.31	89.12	46.75
GEN (CVPR'23)	98.63	5.83	89.22	49.61	91.00	43.36	92.35	34.05	92.80	33.21
PCA (ICCV'23)	95.36	27.37	76.36	79.58	80.92	72.92	90.20	43.59	85.71	55.86
NNGuide (ICCV'23)	87.25	52.54	80.85	62.61	86.04	51.16	91.97	32.66	88.03	46.05
SCALE (ICLR'24)	98.88	5.00	90.14	**42.89**	92.57	**34.28**	93.39	28.51	93.75	27.67
VRA (NeurIPS'24)	99.52	**1.89**	**90.75**	42.98	**92.99**	34.49	93.37	29.19	94.16	27.14
fDBD (ICML'24)	98.39	7.15	89.07	48.68	91.13	41.35	93.10	29.81	92.92	31.75
MME (Ours)	99.48	2.16	89.66	46.71	92.49	35.92	**95.48**	**19.77**	**94.28**	**26.14**

Table 3: OOD detection for MME and state-of-the-art methods. The ID dataset is ImageNet-1K and the OOD datasets are iNaturalist, Places, SUN and Textures. Results are reported using AUROC and FPR95 metrics (in percentage). A pre-trained BiT and a ViT model are tested. The best method is highlighted in bold, and the 2nd and 3rd best are underlined.

4.4 Evaluation on near- and far-OOD

Figure 1 illustrates the trade-off between near-OOD and far-OOD detection performance across various state-of-the-art methods. As highlighted by [20], simple probability-based methods such as MLS [9] and KL Matching [9] achieve competitive results on near-OOD but perform poorly on far-OOD. In contrast, distance-based scoring functions such as Mahalanobis [13] struggle with near-OOD detection due to the difficulty in defining clear decision boundaries. Recent methods like fDBD [15] and ViM [31], which integrate multiple input sources, offer a better balance between near-OOD and far-OOD performance. Additionally, feature truncation techniques such as SCALE [33] and ASH-S [4] achieve a strong trade-off due to their effective scaling properties. Ultimately, MME outperforms all other approaches on both near-OOD and far-OOD tasks, further emphasizing the advantage of combining feature truncation with multiple scoring functions to detect a wide variety of OOD samples.

5 Conclusion

In this study, we analyzed existing feature truncation and scoring function methods and demonstrated both theoretically and empirically that combining these methods improves OOD detection performance. Based on this insight, we proposed a MME strategy that fuses state-of-the-art OOD detectors to better handle the diversity of OOD samples. MME has been thoroughly tested on both large and small-scale benchmarks, considering both near- and far-OOD scenarios. MME has been shown to achieve state-of-the-art performance across various models and datasets, demonstrating its effectiveness and robustness in detecting a wide variety of OOD samples.

References

[1] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

[2] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15793–15803, 2021.

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[4] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Xiaoyuan Guan, Zhouwu Liu, Wei-Shi Zheng, Yuren Zhou, and Ruixuan Wang. Revisit pca-based technique for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19431–19439, 2023.

[7] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 41–50, 2019.

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

[9] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

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

[11] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

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

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

[16] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[17] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23946–23955, 2023.

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

[19] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[20] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1695, 2023.

[21] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

[23] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

[24] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European conference on computer vision*, pages 691–708. Springer, 2022.

[25] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021.

[26] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[27] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

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

[29] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? 2021.

[30] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15651–15660, 2021.

[31] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[32] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[33] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. *arXiv preprint arXiv:2310.00227*, 2023.

[34] Mingyu Xu, Zheng Lian, Bin Liu, and Jianhua Tao. Vra: variational rectified activation for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 36:28941–28959, 2023.

[35] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[36] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

[37] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[38] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

[39] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2022.

[40] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.