Ranked Activation Shift for Post-Hoc Out-of-Distribution Detection

Gianluca Guglielmo and Marc Masana

Graz University of Technology
{guglielmo, mmasana}@tugraz.at

Abstract. State-of-the-art post-hoc out-of-distribution detection methods rely on intermediate layer activation editing. However, they exhibit inconsistent performance across datasets and models. We show that this instability is driven by differences in the activation distributions, and identify a failure mode of scaling-based methods that arises when penultimate layer activations are not rectified. Motivated by this analysis, we propose RAS, a hyperparameter-free post-hoc method that replaces sorted activation magnitudes with a fixed in-distribution reference profile. Our simple plug-and-play method shows strong and consistent performance across datasets and architectures without assumptions on the penultimate layer activation function, and without requiring any hyperparameter tuning, while empirically preserving in-distribution classification accuracy. We further analyze what drives the improvement, showing that both inhibiting and exciting activation shifts independently contribute to better out-of-distribution discrimination[1].

Keywords: Out-of-distribution detection · Hyperparameter-free methods · Representation Alignment

1 Introduction

Models deployed in real-world scenarios often face data that differs from what they were trained on. As a result, they must be able to identify such inputs and respond appropriately [1,23,41]. This problem, known as out-of-distribution (OoD) detection, is essential for building safe and reliable AI systems. Without these safety mechanisms, model predictions cannot be fully trusted, especially in high-stakes applications such as autonomous driving, medical imaging, or financial decision-making, where overconfident errors can have serious consequences [13, 15, 32, 41].

In recent years, several OoD detection methods have been proposed, covering a variety of methodological families [23, 41]. This paper focuses on *post-hoc* OoD detection methods, which operate without the need for retraining past the initial classification task. This provides a strong advantage since the models can be directly integrated into existing pipelines with very low computational

[1] Code is available at: https://github.com/gigug/RAS.

Table 1: *Technical Comparison of OoD Detection Methods*, extended from [20]. The table summarizes the requirements and capabilities of established OOD detection methods, and highlights the additional capabilities provided by RAS, namely requiring only unlabelled ID activations to compute $\boldsymbol{\mu}$, introducing no tunable threshold at inference, and operating on both features and logits. Notably, RAS makes no assumption about the sign of the penultimate layer activations, making it applicable to architectures such as ViTs [9] and ConvNeXt [22], unlike ASH-S and SCALE, which tend to fail when negative activations skew the Q/Q_p ratio.

	Method	Equation	Free of			Leverages		Applicable to Transformers
			ID access	ID labels	hyperparam.	features	logits	
Scoring	EBO [19]	$\log \sum_k e^{f_k(\mathbf{a})}$	✓	✓	✓		✓	✓
	GEN [20]	$G_\gamma(\mathbf{p}) = -\sum_{m=1}^{M} p_{im}^{\gamma}(1-p_{im})^{\gamma},\ p_{i1} \geq \cdots \geq p_{iC}$	✓	✓	✗ (γ)		✓	✓
	ViM [37]	$-\alpha\|\mathbf{a}^{P^\perp}\|_2 + \text{LogSumExp}\, f(\mathbf{a})$	✗ (α, P)	✓	✗ (α)	✓	✓	✓
	Maha [18]	$\max_c -(\mathbf{a}-\hat{\boldsymbol{\mu}}_c)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{a}-\hat{\boldsymbol{\mu}}_c)$	✗ ($\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}_c$)	✗	✓	✓		✓
Enhancing	ReAct [28]	$\tilde{\mathbf{a}} = \min(\mathbf{a}, b)$	✗ (b)	✓	✗ (b)	✓	✓	✓
	ASH-P [8]	$\tilde{a}_j = a_j \cdot \mathbf{1}[a_j > P_p(\mathbf{a})]$	✓	✓	✗ (p)	✓	✓	✓
	ASH-B [8]	$\tilde{a}_j = \bar{a}_p \cdot \mathbf{1}[a_j > P_p(\mathbf{a})],\ \bar{a}_p = \text{mean}_{a_j > P_p}(a_j)$	✓	✓	✗ (p)	✓	✓	✗
	ASH-S [8]	$\tilde{a}_j = a_j \exp(r) \cdot \mathbf{1}[a_j > P_p(\mathbf{a})],\ r = Q/Q_p$	✓	✓	✗ (p)	✓	✓	✗
	SCALE [40]	$\tilde{a}_j = a_j \exp(r)\ \forall j,\ r = \sum_j a_j / \sum_{a_j > P_p} a_j$	✓	✓	✗ (p)	✓	✓	✗
	DICE [29]	$\tilde{\mathbf{W}} = \mathbf{W} \odot \mathbf{M}_s$; top-$s$ weight-contribution mask	✗ (s)	✓	✗ (s)		✓	✓
	RAS (ours)	$\tilde{\mathbf{a}}_{\pi(j)} \leftarrow \mu_j;\ \boldsymbol{\mu} = \frac{1}{N}\sum_i \mathrm{r}(\mathbf{a}_i),\ \pi\colon a_{\pi(1)} \geq \cdots \geq a_{\pi(d)}$	✗ ($\boldsymbol{\mu}$)	✓	✓	✓	✓	✓

Notation: penultimate-layer activations $\mathbf{a}$, logits $f(\mathbf{a})$, prediction distribution $\mathbf{p} = \text{softmax}(f(\mathbf{a}))$, p-th percentile $P_p(\mathbf{a})$, null space of the top principal directions of ID activations $P^\perp$, $Q = \sum_j a_j$, $Q_p = \sum_{a_j > P_p} a_j$, per-class means and tied covariance of ID activations $\hat{\boldsymbol{\mu}}_c$ and $\hat{\boldsymbol{\Sigma}}$, mean ranked activations over ID training set $\boldsymbol{\mu}$, rank-order permutation of test activations π, and ranking function $\mathrm{r}(\cdot)$.

costs. Established post-hoc methods typically exploit either the penultimate layer activations or the model logits, which naturally exhibit useful information to detect OoD samples [14, 19, 41].

Score-enhancing post-hoc OoD detection methods improve the discriminative power of other logit-based OoD detection methods [20]. They establish a rule to adjust intermediate layer activations to suppress noise hindering OoD detection and amplify relevant signals. Among these, we find ReAct [28], DICE [29], ASH [8] and SCALE [40]. Although these methods perform well on a variety of settings, their performance tends to vary depending on the specific model architecture and datasets. Moreover, they require hyperparameter optimization, typically via access to an outlier dataset. This adds an additional layer of complexity to the task and ties the performance of the method to the correct choice of hyperparameters. We present a comparison of scoring and score-enhancing methods in Tab. 1.

In this paper, we present a **hyperparameter-free, score-enhancing OoD detection method** which performs robustly and consistently on different datasets and architectures by shifting the penultimate layer activations to match the mean ranked activation intensities found in ID data.

Our contributions can be summarized as follows:

- we show where scaling-based enhancing methods underperform, including a failure mode arising from unrectified activations,
- we introduce RAS, a novel universal hyperparameter-free enhancing method that replaces ranked activations with a fixed ID reference profile, and

– we analyze what drives RAS's improvement, showing that, contrary to the prevailing assumption, both inhibitory and excitatory activation shifts independently contribute to OoD separation.

2 Preliminaries

Terminology. We consider a standard supervised classification setting with $\mathbf{x}_i \in \mathcal{X}$ representing the input space, and $y_i \in \mathcal{Y} = \{1, \dots, C\}$, $\mathcal{Y}$ representing the labels of C known classes. Model f_θ is trained on an ID dataset $\mathcal{D}_{\text{in}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, which is drawn from an unknown distribution $\mathcal{P}_{\text{in}} \in \mathcal{X} \times \mathcal{Y}$. During inference, the model processes samples drawn from a mixture of $\mathcal{P}_{\text{in}}$ and an external $\mathcal{P}_{\text{out}}$. The goal of OoD detection is to define a scoring function $S(\mathbf{x})$ that maps a given input sample to a scalar value representing the confidence that $\mathbf{x}$ is drawn from $\mathcal{P}_{\text{in}}$. As introduced by [14], a well constructed $S(\mathbf{x})$ should assign higher values to ID samples and lower values to OoD samples, such that a threshold τ can be chosen to effectively separate the two distributions. Consequently, the binary detector $g(\mathbf{x})$ is defined as:

$$g(\mathbf{x}) = \begin{cases} 0 & \text{if } S(\mathbf{x}) \geq \tau \\ 1 & \text{if } S(\mathbf{x}) < \tau\,, \end{cases} \tag{1}$$

where 0 indicates a sample being OoD, while 1 indicates ID. The threshold τ can be adjusted depending on the desired balance between sensitivity and specificity for OoD detection.

Post-hoc methods operate without the need for retraining past the classification task. Generally this allows for strong OoD detection capabilities without a trade-off with ID classification accuracy [4]. The de-facto OoD detection baseline, *Maximum Softmax Probability* (MSP) [14], relies on the maximum softmax output as the OoD score. EBO [19] improves this by defining an energy function that takes the whole logit vector as input. GEN [20] further pushes this by reweighting the softmax distribution, aggregating only the top predicted classes to produce a more discriminative score. Distance-based methods [30, 36] quantify how far a sample's intermediate-layer representation deviates from the ID representations, flagging it as OoD if the distance reaches a certain threshold. Some methods combine intermediate representations and logits information through a single unified score [41]. ViM [37] does it by defining an additional *virtual logit* equal to the sample distance from the ID-based principal subspace.

Score-enhancing methods act on intermediate activations to enhance logit-based methods and facilitate OoD detection at inference time. DICE [29] introduces a directed sparsification strategy for OoD detection, pruning weights in the penultimate layer based on their contribution to the output. ReAct [28] is based on the empirical observation that OoD samples often trigger abnormally

high activation values in the penultimate layer, leading to overconfident predictions. ReAct truncates these activations using a fixed threshold c, computed as a percentile of the activations from the ID training data. Conversely, ASH [8] simplifies the representation by removing a large portion of the penultimate layer activations. The authors introduce three versions: simply prune (ASH-P), prune and binarize (ASH-B) or prune and scale (ASH-S). SCALE [40] builds a theoretical framework to show why ASH [8] is effective and introduces a variation of it that scales activations without the need for pruning.

3 To prune or to scale?

ASH [8] and SCALE [40] rely heavily on auxiliary OoD data and achieve strong performance across many benchmarks [42]. Nonetheless, they have some unaddressed limitations, which we explore in this section. Specifically, we start by examining whether the choice between scaling and pruning is as straightforward as assumed, and we conclude by identifying a failure mode of scaling that occurs when the penultimate layer activations are not rectified.

Given an activation[2] vector $\mathbf{a}$, the following quantities are defined as in [40]:

$$r = \frac{Q}{Q_p}, \quad \text{where } Q = \sum_{j=1}^{D} a_j \text{ and } Q_p = \sum_{a_j > P_p(\mathbf{a})}^{D} a_j \,, \tag{2}$$

where $P_p(\mathbf{a})$ is the p-th percentile of $\mathbf{a}$. SCALE applies a scaling factor e^r to the whole vector $\mathbf{a}$, which is then passed to the final classifier:

$$\mathbf{z} = \mathbf{W}(e^r \mathbf{a}) + \mathbf{b} \,. \tag{3}$$

ASH-S [8] only scales the activations that are greater than $P_p(\mathbf{a})$ by e^r , pruning the remainder to zero. The rationale behind SCALE's choice stems from the following two assumptions: a) the mean of ID activations is higher than that of OoD activations, with the variances remaining equivalent, and b) the penultimate layer activations can be modeled by rectified Gaussian distributions (an assumption also found in the theoretical analysis of ReAct [28]).

With these assumptions, by defining ID activations as $\mathbf{a}_j^{(\text{ID})} \sim \mathcal{N}^R(\mu^{\text{ID}}, \sigma^{\text{ID}})$ and OoD ones as $\mathbf{a}_j^{(\text{OoD})} \sim \mathcal{N}^R(\mu^{\text{OoD}}, \sigma^{\text{OoD}})$, SCALE proves that if condition

$$\frac{\mu^{\text{ID}}}{\sigma^{\text{ID}}} > \frac{\mu^{\text{OoD}}}{\sigma^{\text{OoD}}} \tag{4}$$

is met, then there exists a percentile p such that:

$$\frac{Q_p^{\text{ID}}}{Q^{\text{ID}}} < \frac{Q_p^{\text{OoD}}}{Q^{\text{OoD}}} \,. \tag{5}$$

[2] We refer to the output of a layer as activations or feature vector, interchangeably.

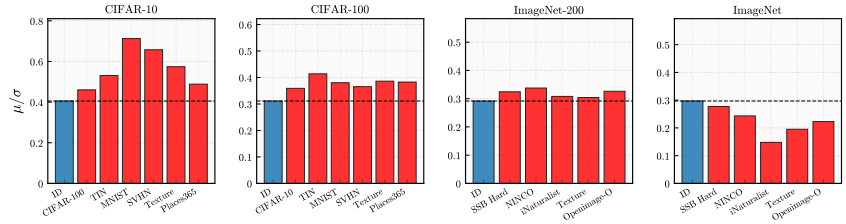
In turn, they prove that Condition 5 implies that *pruning* activations is detrimental to OoD detection, while *scaling* effectively helps it; when the condition is reversed, so are these effects too.

We empirically examine whether Condition 4 holds across standard benchmarks by defining the following quantity and studying its sign:

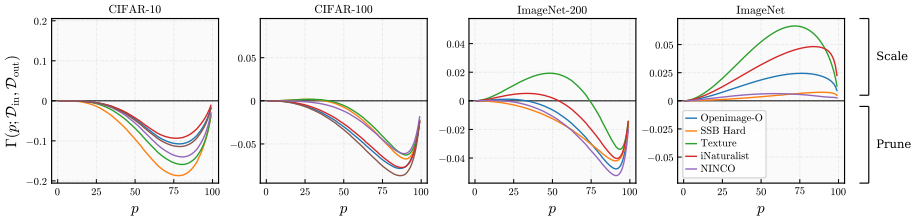
$$\Gamma(p; \mathcal{D}_{\text{in}}, \mathcal{D}_{\text{out}}) = \frac{Q_p^{\text{OoD}}}{Q^{\text{OoD}}} - \frac{Q_p^{\text{ID}}}{Q^{\text{ID}}} \,. \tag{6}$$

The four primary setups from OpenOOD [42] all employ CNN-based architectures with rectification at the penultimate layer (details in Sec. 5). We compute μ/σ for the ID and OoD datasets of each setup and show in Fig. 1a that Condition 4 is not met for three of the four setups.

We empirically evaluate the implication 4 $\implies$ 5 in Fig. 1b, which plots $\Gamma(p; \mathcal{D}_{\text{in}}, \mathcal{D}_{\text{out}})$ across percentiles p for the same datasets: when Condition 4 is false, $\Gamma(p; \mathcal{D}_{\text{in}}, \mathcal{D}_{\text{out}})$ is negative for most values of p. Consequently, for three of the four benchmarks, pruning activations is more discriminative than scaling them for a broad range of percentiles. The further this quantity deviates from zero, the stronger the advantage of either pruning or scaling, while for percentiles closer to zero the difference between the two becomes negligible. Since p is a hyperparameter, its selection is critical: the relative efficacy of pruning and scaling is highly sensitive to its choice.



(a) Pre-ReLU activations; ID is shown in blue; OoD are shown in red.



(b) $\Gamma(p; \mathcal{D}_{\text{in}}, \mathcal{D}_{\text{out}})$ for percentiles p. OoD datasets differentiated by colors.

Fig. 1: The condition required by SCALE's assumptions to hold is violated in three of the four primary OpenOOD setups. In these setups $\mu_{ID}/\sigma_{ID} < \mu_{OoD}/\sigma_{OoD}$, which leads to pruning being more discriminative than scaling for certain percentiles.

Table 2: Best performing ImageNet models [2] rarely have activations be all positive.

Model	Top-1 Acc.	Activation	$a_i \geq 0,\ \forall i$
ConvNeXt-XL [22]	87.76%	GELU	✗
DeiT-S [34]	87.76%	GELU	✗
ConvNeXt-L [22]	87.46%	GELU	✗
ConvNeXt-B [22]	86.82%	GELU	✗
DeiT-B [34]	85.67%	GELU	✗
EfficientNet-B6 [31]	83.93%	SiLU	✗
RegNet-Y-640 [26]	83.83%	ReLU	✓
EfficientNet-B5 [31]	83.36%	SiLU	✗
CvT-13 [39]	82.79%	GELU	✗
ResNet-152 [12]	82.50%	ReLU	✓

3.1 Absence of rectification

As shown in Tab. 2, many modern architectures do not apply any rectification to their penultimate layer activations, including both convolutional models such as ConvNeXt [22] and transformer-based models such as ViT [9]. This has direct consequences for the scaling factor r, used in SCALE and ASH-S. When activations are rectified, $Q_p \leq Q$ guarantees $r \geq 1$. However, in the absence of rectification, activations can take negative values, meaning Q may be small or even negative, and r is no longer constrained to be positive or greater than one. This breaks the monotonic relationship between r and the pruning threshold p that the theoretical analysis relies upon, and the interpretation of scaling as a strictly amplifying operation no longer holds.

4 Ranked Activation Shift

Current post-hoc methods [8,28,29,40] rely on threshold-based clipping or pruning, which does not leverage the global shape of the activation vector. We propose Ranked Activation Shift (RAS), which compares the sorted activations of new samples against an ID reference, as depicted in Fig. 2. During a one-time offline phase, we compute a reference profile vector $\boldsymbol{\mu} \in \mathbb{R}^d$, defined as the mean of the ranked[3] activation vectors over a subset of ID data:

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{r}(\mathbf{a}_i), \tag{7}$$

where a_i are the penultimate layer activations extracted from samples $\mathbf{x}_i \in \mathcal{D}_{\text{in}}$, while $\mathrm{r}(\cdot)$ is a function that arranges the elements of the vector in ascending order. As shown in Fig. 5, the reference profile can be successfully built using few

[3] We refer to the terms *ranking* and *sorting*, interchangeably.

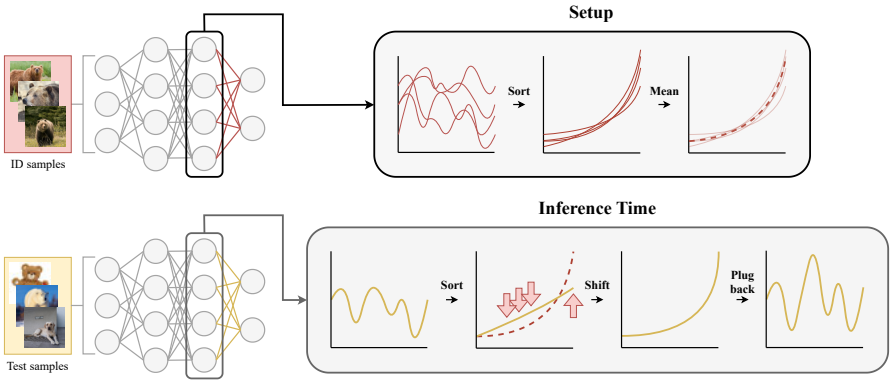


Fig. 2: Overview of RAS. At *setup* time, RAS computes a reference vector $\boldsymbol{\mu}$ by sorting each ID train sample by intensity and then averaging across all. At *inference* time, each sample's activations are shifted so that their ranked activation intensities match $\boldsymbol{\mu}$. Subsequently, they are passed to the classifier and the chosen OoD scorer is used.

samples, with the option of exploiting an held-out in-distribution dataset that was not seen during training.

At inference time, for a given input $\mathbf{x} \in \mathcal{D}_{\text{in}} \cup \mathcal{D}_{\text{out}}$, we extract its penultimate layer activations $\mathbf{a}$. Let π be the permutation of indices $\{1, \ldots, d\}$ that sorts $\mathbf{a}$ in ascending order, such that $s_j = a_{\pi(j)}$ corresponds to the j-th largest activation. Next, we construct the modified activation vector $\bar{\mathbf{a}}$ by assigning the reference value μ_j to the original position $\pi(j)$:

$$\bar{a}_{\pi(j)} = \mu_j \quad \text{for } j = 1, \ldots, d\,. \tag{8}$$

By strictly imposing the value distribution of $\boldsymbol{\mu}$ onto the input, we effectively perform a histogram matching step. This ensures that $\bar{\mathbf{a}}$ exhibits the average statistical profile of ID data, while preserving the spatial orientation of the original input $\mathbf{x}$. Consequently, if an OoD sample produces an anomalous activation distribution (e.g., peaked noise or heavy tails), this substitution forces it into an ID-like range before classification. The step-by-step process is shown in Alg. 1.

5 Experiments

We evaluate RAS using the OpenOOD benchmark [42], a standardized framework to compare OoD detection methods. It provides a unified pipeline with consistent data splits and pre-trained model backbones, ensuring a fair comparison between methods. We provide benchmark results on all available ID datasets and their corresponding models in OpenOOD: ImageNet [6] with ResNet50 [12], CIFAR-10, CIFAR-100 [17] and ImageNet-200 (a subset of ImageNet) with ResNet18 [12]. We further extend our evaluation to include EfficientNet [31], which does not use residual blocks, ConvNeXt [22], which aggregates different strategies to modernize

Table 3: OoD evaluation splits for each ID dataset.

	CIFAR-10 [17]	**CIFAR-100** [17]	**ImageNet-200** [6]	**ImageNet** [6]
Near-OoD	CIFAR-100 [17] TinyImageNet [6]	CIFAR-10 [17] TinyImageNet [6]	SSB-Hard [42] NINCO [3]	SSB-Hard [42] NINCO [3]
Far-OoD	Texture [5] MNIST [7] SVHN [25] Places365 [43]	Texture [5] MNIST [7] SVHN [25] Places365 [43]	Texture [5] iNaturalist [35] OpenImage-O [37]	Texture [5] iNaturalist [35] OpenImage-O [37]

convolutional neural networks, and vision-transformers, namely ViT-B/16 [9] and Swin-T [21]. For these models not available in the OpenOOD framework, we use checkpoints from Torchvision [33] and HuggingFace [38].

The evaluation protocol follows the splits summarized in Tab. 3. For each experiment, ID is represented by the corresponding test set of the pretrained backbone's dataset, containing unseen samples drawn from the same distribution. OoD datasets are divided into near- and far-OoD subsets, which indicates the semantic distance from the ID dataset [41, 42].

We employ the most commonly used OoD detection metrics [4, 14, 23, 41] for these scenarios: AUROC, the area under the receiver operating characteristic curve (the higher the better); AUPR, the area under the precision-recall curve (the higher the better); and FPR@TPR95, the false-positive rate when the true-positive rate is 95% (the lower the better).

Algorithm 1 RAS: Ranked Activation Shift

Require: ID training set $\mathcal{D}_{\text{in}}$, feature dimension d
Ensure: Modified activation vector $\bar{\mathbf{a}}$ at inference-time
Setup
Initialize reference vector $\boldsymbol{\mu} \leftarrow \mathbf{0} \in \mathbb{R}^d$
for each $\mathbf{x}_i \in \mathcal{D}_{\text{in}}$ **do**
  Extract activation vector $\mathbf{a}_i$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \mathrm{r}(\mathbf{a}_i)$
end for
$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}/|\mathcal{D}_{\text{in}}|$
Inference-Time Substitution
Extract activation vector[a] $\mathbf{a} \in \mathbb{R}^d$ for input $\mathbf{x}$
Compute permutation π such that $a_{\pi(1)} \geq \cdots \geq a_{\pi(d)}$
Initialize $\bar{\mathbf{a}} \leftarrow \mathbf{a}$
for $j = 1$ to d **do**
  $\bar{a}_{\pi(j)} \leftarrow \mu_j$
end for
Continue feed-forward, compute chosen $S(\mathbf{x})$

[a] In ViTs, $\mathbf{a}$ is the output representation of the class token.

Table 4: Across a comprehensive range of dataset–architecture combinations, RAS (hyperparameter-free) consistently exhibits best or competitive performance compared to state-of-the-art enhancing post-hoc OoD detection techniques on their optimized hyperparameter values. **Bold** indicates best result, underline second and third.

	CIFAR-10		**CIFAR-100**		**IN-200**		**ImageNet**									
	ResNet18		**ResNet18**		**ResNet18**		**ResNet50**		**EffNet-B0**		**ConvNx-T**		**ConvNx-B**		**ViT-B/16**	
	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓
EBO	90.00	48.24	80.15	56.26	87.51	45.01	84.03	50.46	76.76	72.29	69.36	83.03	59.36	93.29	72.35	88.48
ReAct	89.31	51.12	80.52	54.93	88.13	42.10	87.15	42.46	**84.27**	49.01	77.87	69.45	80.85	64.06	79.12	66.15
DICE	82.27	57.85	79.80	56.82	87.19	46.66	83.80	54.07	43.32	91.33	57.78	93.86	58.10	95.30	75.53	65.23
ASH-P	89.47	52.55	80.34	55.77	88.02	43.44	85.44	46.04	63.24	80.79	31.68	99.26	16.89	99.80	22.32	98.74
ASH-B	77.50	81.48	79.79	61.29	89.24	42.48	88.75	37.11	47.50	86.81	31.01	98.05	65.33	79.84	52.57	95.51
ASH-S	82.90	77.47	81.52	54.29	**90.51**	38.15	89.68	35.05	56.18	81.21	32.59	97.64	26.86	99.67	22.68	98.71
SCALE	85.18	71.77	81.29	54.60	90.29	38.89	**90.42**	**34.02**	59.05	79.70	60.10	84.32	65.22	76.53	68.64	89.52
RAS	**90.24**	**40.16**	**82.09**	**52.31**	89.51	**36.70**	86.55	40.92	83.78	**46.87**	**83.31**	**46.28**	**84.86**	**45.29**	**81.48**	**55.19**

5.1 Main benchmark results

We start by comparing RAS with state-of-the-art enhancing post-hoc OoD methods. We follow their original implementations by applying each method to the penultimate layer and using the EBO function to compute the final score $S(\mathbf{x})$. In the case of vision-transformers, RAS is applied to the class token, since the other penultimate layer tokens are discarded during classification.

Table 4 summarizes the results for various ID datasets and architectures. RAS consistently outperforms every compared methods on all but one scenario (ImageNet with ResNet50), where it still achieves a strong improvement over the standard EBO performance. As expected, ASH and SCALE fail on architectures whose penultimate layer contains negative-valued activations (ConvNeXt, ViT-B/16). It is crucial to note that RAS achieves these results without requiring any hyperparameter optimization. In contrast, the other enhancing methods require a hyperparameter (ID-OoD threshold) tuning on a held-out OoD dataset, with the reported values representing their optimal configurations. This underlines the robustness of RAS, as it improves over or remains competitive against methods displaying their peak potential performance. Extended results in Supplementary.

Preserving ID accuracy is critical for OoD detection methods, especially for enhancing post-hoc ones, as these can directly interfere with the forward pass. As shown in Tab. 5, RAS keeps ID accuracy virtually unchanged.

5.2 RAS with different scoring strategies

A key advantage of enhancing post-hoc methods is their modularity, which enables integration with a variety of scoring methods. To evaluate RAS's flexibility, we pair it with EBO [19], ViM [37] and GEN [20], and compare the results to their corresponding standard performance. Table 6 presents the results over the four main OpenOOD dataset scenarios. Applying RAS guarantees a consistent performance improvement, regardless of the scoring method used.

Table 5: RAS empirically preserves ID accuracy.

	C-10	C-100	IN-200	ImageNet				
	RN18	RN18	RN18	RN50	ENB0	CNx-T	CNx-B	ViT-B/16
Orig.	95.06	77.25	86.37	76.17	77.80	82.11	83.74	81.14
RAS	95.03	77.26	86.16	76.04	77.73	82.15	83.71	81.14

Table 6: RAS performance with different scoring strategies. ViM* is initialized with evaluation data given the lack of access to some of the training data. **Bold** indicates best performance.

	CIFAR-10			CIFAR-100			ImageNet-200			ImageNet		
Method	AUC↑	FPR↓	AUPR↑	AUC↑	FPR↓	AUPR↑	AUC↑	FPR↓	AUPR↑	AUC↑	FPR↓	AUPR↑
EBO	90.00	48.24	91.88	80.15	56.26	81.85	87.51	45.01	88.11	84.03	50.46	58.54
EBO + **RAS**	**90.24**	**40.16**	**92.03**	**82.09**	**52.31**	**83.98**	**89.51**	**36.70**	**90.06**	**86.55**	**40.92**	**64.03**
ViM*	90.61	34.31	91.94	75.26	62.69	**78.57**	82.86	**48.39**	83.59	**84.29**	**43.04**	**61.44**
ViM* + **RAS**	**91.12**	**33.54**	**92.62**	**75.28**	**61.41**	78.48	**83.47**	48.51	**85.18**	82.47	45.13	56.95
GEN	90.30	41.04	91.76	80.23	55.95	81.94	88.29	41.34	88.58	84.60	47.49	57.60
GEN + **RAS**	**90.61**	**37.43**	**92.45**	**82.47**	**51.88**	**84.65**	**89.03**	**37.91**	**89.69**	**85.89**	**42.80**	**60.74**

5.3 Results beyond selecting the penultimate layer representations

Comparison with ℓ_2-normalization. MDS [18] scores samples by their Mahalanobis distance to the nearest class mean in feature space, while RMDS [27] refines this by computing distances relative to a class-agnostic background distribution, reducing the influence of uninformative feature dimensions. Recent methods, such as MDS++ and RMDS++ [24], have demonstrated that normalizing activations to match the ID ℓ_2-norm can significantly bolster OoD detection performance compared to the original unnormalized versions. RAS takes this further by aligning the full activation distribution to the ID reference via $\boldsymbol{\mu}$. This transformation implicitly enforces ℓ_2-norm matching, since every activation vector is replaced by the same reference vector, but produces a qualitatively different vector geometry from simple radial scaling, as illustrated in Fig. 3. We show that RAS provides a more robust enhancement to MDS and RMDS than the standard ℓ_2-normalized variants. Empirical results in Tab. 7 confirm that this activation shifting strategy consistently outperforms standard ℓ_2-normalization across both baseline frameworks.

Single layer selection. Our proposed RAS method is applicable to any intermediate model representation, offering flexibility on the layer selection. A natural choice is the penultimate layer of the model, which has been shown to be the most effective for other enhancing post-hoc methods, such as ReAct [28] and ASH [8]. Nevertheless, non-penultimate intermediate layers have been proven to be effective for OoD detection [10, 11, 16]. We present results in Fig. 4 for the CIFAR-10, CIFAR-100 and ImageNet-200 scenarios, where RAS has been

	C-10	C-100	IN		
	RN18	**RN18**	**RN50**	**CN-B**	**SWIN**
MDS	69.5	79.5	49.5	33.6	52.4
MDS++	46.2	72.9	52.0	24.3	38.9
MDS + **RAS**	**41.0**	**50.0**	**48.9**	**22.1**	**37.6**
RMDS	52.5	76.1	62.5	31.7	48.7
RMDS++	54.1	77.5	70.4	29.5	39.3
RMDS + **RAS**	**47.1**	**52.1**	**49.8**	**23.3**	**38.8**

Table 7: RAS vs ℓ_2-normalization. Average FPR on OpenOOD datasets using the checkpoints used in the MDS++ paper [24], rather than the OpenOOD defaults. **Bold** indicates best.

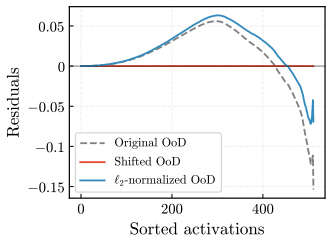


Fig. 3: RAS (shifted) vs ℓ_2-normalization, shown as residuals with respect to the average ranked ID activations.

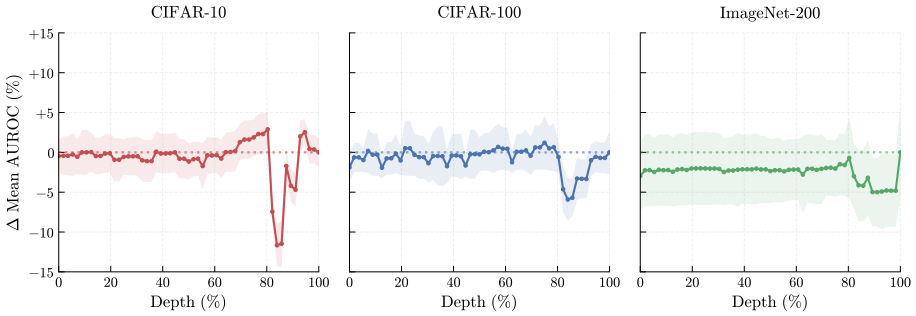
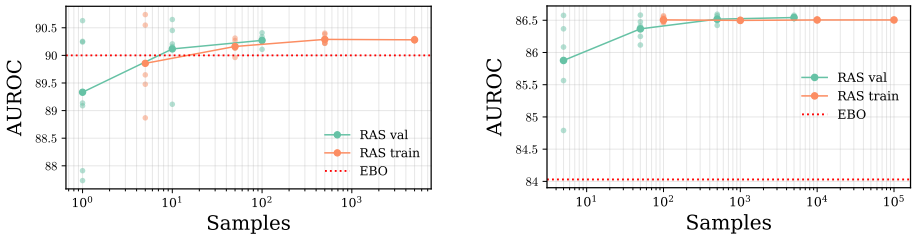


Fig. 4: Mean AUROC gain of RAS (solid line) when applied to each ResNet layer, relative to the commonly selected penultimate layer (dashed line).

independently applied to each intermediate layer. Across all results, there is only a noticeable improvement over the penultimate layer for the CIFAR-10 scenario, which occurs in a batch normalization layer inside the 4th residual block. In the other cases, the improvement is either not significant or not observed at all, leading to the conclusion that applying RAS to the penultimate layer remains the most straightforward single-layer choice. We explore multiple layer selection in the Supplementary Material.

5.4 How much ID data is needed?

Earlier, we computed the reference profile using the entire training set. Here, we evaluate RAS with varying numbers of samples, using both training data and unlabeled held-out validation data, showing that retaining the original training set is also unnecessary. Figure 5 (log scale) confirms nearly identical results across both splits. Only a few samples are enough to improve over the EBO baseline (red dashed line), and held-out validation data not used for training can also be used.



(a) CIFAR-10 **(b)** ImageNet

Fig. 5: AUROC on CIFAR-10 (left) and ImageNet (right) as ID.

6 Discussion

Methods such as ReAct operate under the assumption that OoD samples are characterized by abnormally high activations, and improve detection by suppressing them. It is natural to ask whether the opposite direction of correction is also effective: can pushing below-mean activations upward toward $\boldsymbol{\mu}$ also improve OoD detection? Is the benefit of RAS tied to a single direction of activation modification?

To study this, we define RAS-inhibit and RAS-excite, two extensions of RAS which only shift the ranked activations that are *above* and *below* the ranked ID mean, respectively. We report this ablation in Tab. 8 with the energy score statistics for each variant relative to the unmodified EBO baseline. Each entry shows both the baseline and the post-transform energy score distribution $\mu \pm \sigma$, along with the change in mean $\Delta\mu$ and the change in AUROC, allowing us to disentangle the respective contributions of RAS-inhibit and RAS-excite to the overall score separation. Table 8 shows that, crucially, *both* directions improve over EBO in the majority of settings, demonstrating that the benefit of RAS is not tied to a single direction of activation modification. Full RAS compounds both effects simultaneously, achieving the largest average AUROC gain in almost every scenario, up to +2.45 and +2.86 for ImageNet-200 and ImageNet respectively. Consistently across all benchmarks and variants, RAS also reduces the standard deviation of energy scores for both ID and OoD samples, with the improvement driven primarily by this variance compression rather than any shift in means.

Therefore, the benefit of activation editing is not tied to any particular direction of correction: both upward and downward shifts independently improve OoD separation, and RAS captures the full gain by applying both.

6.1 Activation study

A further question is whether the improvement holds regardless of the relative behavior of OoD and ID activations across rank positions, or whether it is specific to certain configurations. To study this, we compare the mean ranked activation profiles of ID and OoD datasets. Let $\bar{a}_{\pi_i}$ denote the mean activation at rank i

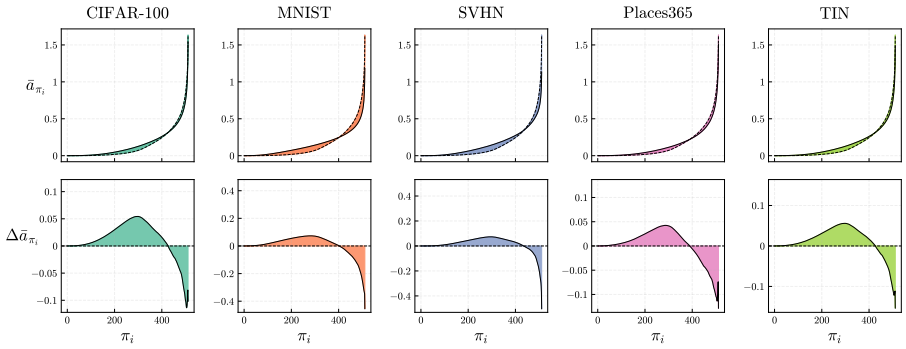
Table 8: Ablation of RAS into RAS-inhibit (activations above $\boldsymbol{\mu}$ shifted down) and RAS-excite (activations below $\boldsymbol{\mu}$ shifted up). EBO ($\mu\pm\sigma$) is the baseline score distribution for each OoD dataset. $\Delta\mu$ and ΔAUROC are relative to EBO scores. ID is shown in bold in each header.

	EBO	**EBO + RAS-excite**			**EBO + RAS-inhibit**			**EBO + RAS**		
	$\mu\pm\sigma$	$\mu\pm\sigma$	$\Delta\mu$	ΔAUC	$\mu\pm\sigma$	$\Delta\mu$	ΔAUC	$\mu\pm\sigma$	$\Delta\mu$	ΔAUC
CIFAR-10	9.09 ± 1.51	9.32±1.24	+0.23	–	8.47±1.05	−0.63	–	8.67±0.84	−0.42	–
CIFAR-100	6.37±1.83	7.04±1.56	+0.67	+0.53	6.25±1.50	−0.12	+1.82	6.89±1.32	+0.52	+0.99
TIN	6.07±1.72	6.75±1.53	+0.68	+0.25	5.95±1.45	−0.12	+1.53	6.72±1.29	+0.64	+0.00
MNIST	5.37±1.41	6.34±1.32	+0.97	−0.64	5.41±1.32	+0.04	+0.03	6.39±1.26	+1.02	−1.95
SVHN	5.81±1.54	6.80±1.36	+0.99	−1.52	5.83±1.37	+0.02	+0.19	6.81±1.22	+1.00	−2.97
Texture	6.15±1.87	6.80±1.57	+0.64	+1.18	5.90±1.40	−0.26	+3.60	6.63±1.26	+0.47	+2.73
Places365	6.01±1.77	6.68±1.52	+0.68	+0.98	5.90±1.46	−0.10	+2.03	6.55±1.26	+0.54	+2.42
Average	5.96±1.69	6.74±1.47	+0.77	+0.13	5.87±1.42	−0.09	+1.53	6.66±1.27	+0.70	+0.20
CIFAR-100	11.14 ± 3.75	11.71±3.13	+0.57	–	10.01±2.35	−1.13	–	10.65±1.84	−0.49	–
CIFAR-10	7.74±2.07	8.95±1.80	+1.21	−0.83	7.59±1.69	−0.14	+0.26	8.84±1.51	+1.11	−1.52
TIN	7.35±1.79	8.46±1.64	+1.11	+0.52	7.18±1.47	−0.17	+1.30	8.32±1.35	+0.96	+1.26
MNIST	7.68±1.67	8.89±1.50	+1.21	−0.26	7.56±1.49	−0.12	+0.48	8.73±1.37	+1.05	−0.07
SVHN	7.39±1.79	8.43±1.56	+1.04	+1.31	7.20±1.45	−0.20	+1.56	8.27±1.29	+0.88	+2.27
Texture	7.80±2.14	8.57±1.82	+0.77	+3.47	7.41±1.54	−0.39	+2.93	8.27±1.30	+0.47	+6.10
Places365	7.64±1.92	8.69±1.65	+1.04	+1.08	7.51±1.57	−0.13	+0.42	8.57±1.40	+0.93	+1.17
Average	7.60±1.90	8.66±1.66	+1.06	+0.88	7.41±1.54	−0.19	+1.16	8.50±1.37	+0.90	+1.53
ImageNet-200	13.99 ± 4.57	14.96±3.81	+0.97	–	12.37±2.62	−1.62	–	13.34±1.98	−0.65	–
SSB-Hard	9.67±3.14	11.28±2.64	+1.60	+0.77	9.33±2.35	−0.34	+0.55	10.94±2.02	+1.26	+0.50
NINCO	8.88±2.34	10.42±2.05	+1.54	+1.85	8.77±1.98	−0.11	+0.13	10.22±1.71	+1.34	+2.29
Texture	7.92±2.33	9.39±2.10	+1.47	+1.65	7.60±1.69	−0.32	+1.94	9.13±1.55	+1.20	+3.04
iNaturalist	7.80±1.49	8.94±1.37	+1.14	+3.19	7.65±1.20	−0.15	+1.30	8.80±1.18	+1.00	+3.75
Openimage-O	8.25±1.85	9.80±1.69	+1.55	+1.84	8.10±1.55	−0.15	+1.12	9.59±1.46	+1.33	+2.66
Average	8.51±2.23	9.97±1.97	+1.46	+1.86	8.29±1.75	−0.22	+1.01	9.73±1.58	+1.23	+2.45
ImageNet	17.16 ± 4.38	17.94±4.03	+0.78	–	16.16±3.29	−0.99	–	16.94±3.12	−0.21	–
SSB-Hard	14.02±3.25	15.07±2.98	+1.05	−0.38	13.56±2.74	−0.46	+0.64	14.61±2.71	+0.60	−0.80
NINCO	13.14±2.55	13.70±2.46	+0.56	+3.37	12.69±2.24	−0.45	+1.61	13.32±2.17	+0.18	+3.45
iNaturalist	11.39±1.77	11.76±1.51	+0.37	+3.55	10.81±1.30	−0.58	+3.12	11.24±1.28	−0.15	+4.79
Texture	11.47±2.75	12.21±2.21	+0.74	+2.28	11.01±1.84	−0.46	+2.86	11.75±1.69	+0.28	+3.76
Openimage-O	11.55±2.08	12.27±1.85	+0.73	+2.02	11.16±1.68	−0.38	+1.84	11.83±1.51	+0.28	+3.12
Average	12.31±2.48	13.00±2.20	+0.69	+2.17	11.85±1.96	−0.47	+2.01	12.55±1.87	+0.24	+2.86

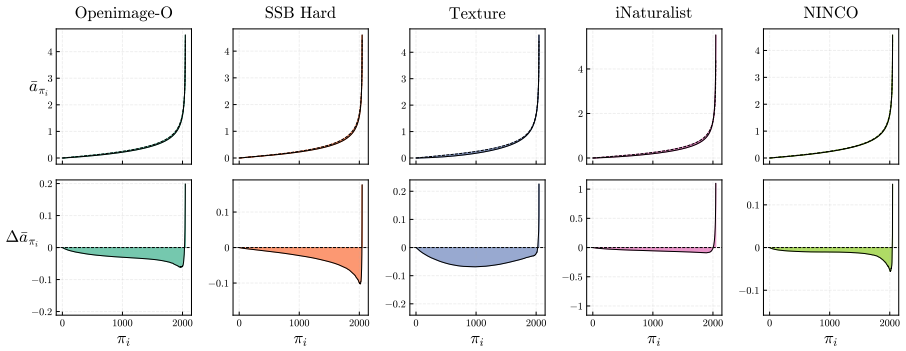
across a selected OoD dataset, where π is the rank-order permutation. We define the signed deviation between the OoD dataset profile and the ID reference at each rank position i as the *residual*:

$$\Delta\bar{a}_{\pi_i} = \bar{a}_{\pi_i} - \mu_i \,, \tag{9}$$

Figure 6 shows two qualitatively distinct behaviors when observing the mean ranked profiles and their corresponding residuals for CIFAR-10 and ImageNet as ID datasets. For CIFAR-10, OoD samples exhibit a characteristic profile where high-rank activations fall below the ID reference, while low-rank activations exceed it; while for ImageNet the pattern is reversed. RAS-excite and RAS-inhibit are both effective across all cases, confirming that RAS does not rely



(a) ID dataset: CIFAR-10. Model: ResNet-18.



(b) ID dataset: ImageNet. Model: ResNet-50.

Fig. 6: Comparison of activation profiles for ID (dashed line) and OoD datasets (solid line). a_{π_i} shows the average ranked activations, while Δa_{π_i} indicates the residual difference between the sorted ID and OoD activations. Two distinct behaviors emerge across benchmarks: in the CIFAR-10 case, OoD samples consistently exhibit weaker high-rank activations than ID; in the ImageNet case, the opposite holds across all OoD datasets.

on any particular configuration of the OoD activation profile relative to the ID reference. This suggests that the benefit of activation shifting is a general property of the shifting to the ID reference profile, rather than a consequence of any specific OoD behavior. Additional profiles are provided in the Supplementary Material.

7 Conclusions

Our analysis reveals why competing methods such as ASH and SCALE exhibit inconsistent behavior across benchmarks: their underlying assumptions on the μ/σ ratio of the penultimate layer activations are often violated, and their reliance on

a ratio of activation sums breaks down entirely in the absence of rectification. RAS sidesteps these failure modes by operating on rank order rather than activation magnitude, and by anchoring each sample to the same ID reference regardless of the direction or magnitude of its deviation.

We present a hyperparameter-free post-hoc method for out-of-distribution detection that shifts penultimate layer activations to match the average ranked activation profile of in-distribution data. Unlike existing enhancing methods, RAS requires no threshold tuning, no access to OoD data, and no assumptions about the sign or distribution of the penultimate layer activations, making it directly applicable to a broad range of architectures, including ReLU-based CNNs, GELU-based ConvNeXts, and vision transformers.

We validate RAS across four standard, well-established benchmarks, multiple architectures, and three scoring functions, consistently matching or outperforming state-of-the-art enhancing methods that operate at their optimized hyperparameter values. Crucially, as shown in the Tab. 5, RAS leaves ID classification accuracy virtually unchanged. We further show that the improvement is not tied to a single direction of activation modification: both inhibiting and excitatory shifts toward $\boldsymbol{\mu}$ independently improve OoD separation, with RAS compounding both effects through variance compression.

Acknowledgments

Gianluca Guglielmo acknowledges the support of KAI GmbH and Infineon Technologies Austria. This research was partially funded by the Austrian Science Fund (FWF) 10.55776/COE12. The authors would like to thank Christoph Griesbacher for the fruitful discussions during the development of this research. We would also like to thank Thomas Pock for his academic supervision at Graz University of Technology.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv e-prints pp. arXiv–1606 (2016) 1
2. Bekhouche, M.: ImageNet-1k Leaderboard. https://huggingface.co/spaces/Bekhouche/ImageNet-1k_leaderboard (2025), accessed: 2026-03-02 6
3. Bitterwolf, J., Müller, M., Hein, M.: In or out? fixing ImageNet out-of-distribution detection evaluation. In: International Conference on Machine Learning. pp. 2471–2506 (2023) 8
4. Borlino, F.C., Lu, L., Tommasi, T.: Foundation models and fine-tuning: A benchmark for out-of-distribution detection. IEEE Access **12**, 79401–79414 (2024) 3, 8
5. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Computer Vision and Pattern Recognition Conference. pp. 3606–3613 (2014) 8
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition Conference. pp. 248–255. IEEE (2009) 7, 8

7. Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine **29**(6), 141–142 (2012) 8
8. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. In: International Conference on Learning Representations (2023) 2, 4, 6, 10
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021) 2, 6, 8
10. Guglielmo, G., Masana, M.: Leveraging intermediate representations for better out-of-distribution detection. In: Computer Vision Winter Workshop (2025) 10
11. Harun, M.Y., Lee, K., Gallardo, J., Krishnan, G., Kanan, C.: What variables affect out-of-distribution generalization in pretrained models? Advances in Neural Information Processing Systems (2024) 10
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition Conference. pp. 770–778 (2016) 6, 7
13. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: International Conference on Machine Learning. pp. 8759–8773. PMLR (2022) 1
14. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017) 2, 3, 8
15. Hong, Z., Yue, Y., Chen, Y., Cong, L., Lin, H., Luo, Y., Wang, M.H., Wang, W., Xu, J., Yang, X., et al.: Out-of-distribution detection in medical image analysis: A survey. arXiv e-prints pp. arXiv–2404 (2024) 1
16. Meza De la Jara, I., Rodriguez-Opazo, C., Teney, D., Ranasinghe, D., Abbasnejad, E.: Mysteries of the deep: Role of intermediate representations in out-of-distribution detection. In: Advances in Neural Information Processing Systems (2025) 10
17. Krizhevsky, A., Hinton, G., et al.: Learning Multiple Layers of Features from Tiny Images. Master's thesis, Department of Computer Science, University of Toronto (2009) 7, 8
18. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems **31** (2018) 2, 10
19. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems **33**, 21464–21475 (2020) 2, 3, 9
20. Liu, X., Lochman, Y., Zach, C.: GEN: Pushing the limits of softmax-based out-of-distribution detection. In: Computer Vision and Pattern Recognition Conference. pp. 23946–23955 (2023) 2, 3, 9
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision. pp. 10012–10022 (2021) 8
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Computer Vision and Pattern Recognition Conference. pp. 11976–11986 (2022) 2, 6, 7
23. Miyai, A., Yang, J., Zhang, J., Ming, Y., Lin, Y., Yu, Q., Irie, G., Joty, S., Li, Y., Li, H.H., et al.: Generalized out-of-distribution detection and beyond in vision language model era: A survey. Transactions on Machine Learning Research (2024) 1, 8
24. Müller, M., Hein, M.: Mahalanobis++: Improving ood detection via feature normalization. In: International Conference on Machine Learning (2025) 10, 11

25. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., et al.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada (2011) 8
26. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: Computer Vision and Pattern Recognition Conference. pp. 10428–10436 (2020) 6
27. Ren, J., Fort, S., Liu, J., Guha Roy, A., Padhy, S., Lakshminarayanan, B.: A simple fix to mahalanobis distance for improving near-ood detection. arXiv e-prints pp. arXiv–2106 (2021) 10
28. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. Advances in Neural Information Processing Systems **34**, 144–157 (2021) 2, 3, 4, 6, 10
29. Sun, Y., Li, Y.: DICE: Leveraging sparsification for out-of-distribution detection. In: European Conference on Computer Vision. pp. 691–708. Springer (2022) 2, 3, 6
30. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning. pp. 20827–20840. PMLR (2022) 3
31. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019) 6, 7
32. Thimonier, H., Popineau, F., Rimmel, A., Doan, B.L., Daniel, F.: Comparative evaluation of anomaly detection methods for fraud detection in online credit card payments. In: International Congress on Information and Communication Technology. pp. 37–50. Springer (2024) 1
33. TorchVision maintainers and contributors: TorchVision: PyTorch's computer vision library. https://github.com/pytorch/vision (2016) 8
34. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021) 6
35. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Computer Vision and Pattern Recognition Conference. pp. 8769–8778 (2018) 8
36. Venkataramanan, A., Benbihi, A., Laviale, M., Pradalier, C.: Gaussian latent representations for uncertainty estimation using mahalanobis distance in deep classifiers. In: International Conference on Computer Vision. pp. 4488–4497 (2023) 3
37. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Computer Vision and Pattern Recognition Conference. pp. 4921–4930 (2022) 2, 3, 8, 9
38. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Conference on Empirical Methods in Natural Language Processing. pp. 38–45. Association for Computational Linguistics (2020) 8
39. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: CvT: Introducing convolutions to vision transformers. In: International Conference on Computer Vision. pp. 22–31 (2021) 6
40. Xu, K., Chen, R., Franchi, G., Yao, A.: Scaling for training time and post-hoc out-of-distribution detection enhancement. In: International Conference on Learning Representations (2024) 2, 4, 6

41. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024) 1, 2, 3, 8
42. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., et al.: Openood v1.5: Enhanced benchmark for out-of-distribution detection. Journal of Data-centric Machine Learning Research (2024) 4, 5, 7, 8
43. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (2017) 8