# Catalyst: Out-of-Distribution Detection via Elastic Scaling

Abid Hassan, Tuan Ngo, Saad Shafiq, Nenad Medvidovic
University Southern California, Los Angeles
{mdskabid, tkngo, sshafiq, neno}@usc.edu

## Abstract

*Out-of-distribution (OOD) detection is critical for the safe deployment of deep neural networks. State-of-the-art post-hoc methods typically derive OOD scores from the output logits or penultimate feature vector obtained via global average pooling (GAP). We contend that this exclusive reliance on the logit or feature vector discards a rich, complementary signal: the raw channel-wise statistics of the pre-pooling feature map lost in GAP. In this paper, we introduce* `Catalyst`*, a post-hoc framework that exploits these under-explored signals.* `Catalyst` *computes an input-dependent scaling factor ($\gamma$) on-the-fly from these raw statistics (e.g., mean, standard deviation, and maximum activation). This $\gamma$ is then fused with the existing baseline score, multiplicatively modulating it – an "elastic scaling" – to push the ID and OOD distributions further apart. We demonstrate* `Catalyst` *is a generalizable framework: it seamlessly integrates with logit-based methods (e.g., Energy, ReAct, SCALE) and also provides a significant boost to distance-based detectors like KNN. As a result,* `Catalyst` *achieves substantial and consistent performance gains, reducing the average False Positive Rate by 32.87% on CIFAR-10 (ResNet-18), 27.94% on CIFAR-100 (ResNet-18), and 22.25% on ImageNet (ResNet-50). Our results highlight the untapped potential of pre-pooling statistics and demonstrate that* `Catalyst` *is complementary to existing OOD detection approaches. Our code is available here: https://github.com/bingabid/Catalyst*

## 1. Introduction

A deep neural network deployed in real-world environments will inevitably encounter out-of-distribution (OOD) samples drawn from novel contexts whose class labels are disjoint from the training distribution, referred as in-distribution (ID) data. Unlike ID samples that the model was trained on, these OOD instances should not be confidently classified but be detected and flagged for human review. Robust OOD detection is particularly crucial for safety-critical applications where erroneous predictions can have severe consequences, e.g., in medical diagnosis [52, 63] or autonomous driving [9] systems.

Early methods to OOD detection primarily focused on designing scoring functions to distinguish ID from OOD samples. The seminal work [16] proposed using the maximum softmax probability (MSP) as a confidence measure, based on observation that OOD samples yield lower softmax scores. However, subsequent studies [15, 48] revealed a critical flaw: neural networks often produce overconfident softmax predictions even for far-OOD inputs, rendering MSP unreliable. To address this, Energy [36] introduced the energy-based score, which maps inputs to a scalar value such that ID samples yield lower energy than OOD samples. This score provided a more robust uncertainty measure, inspiring a series of improvements aimed at enhancing ID-OOD separability. Recent advances have focused on post-hoc activation manipulation to amplify this separation. Notable methods include ReAct [55], DICE [54], ASH [5], SCALE [67] achieving state-of-the-art performance.

These methods share a common paradigm: they derive their scores using the penultimate feature vector (generally obtained via GAP) as their foundational input. These techniques process this feature vector to derive energy-based scores [36, 55, 67] or distance-based scores [11, 56]. We contend that exclusive reliance on the feature vector creates an information bottleneck, as it discards complementary signals, namely the raw channel-wise statistics of the pre-pooling feature map, which could otherwise be used in tandem with existing methods for improved OOD detection.

Figure 1 illustrates the distribution of these untapped information cues, extracted from the penultimate layer's pre-pooling activation map in an ImageNet-trained ResNet-50, using Textures as the OOD dataset. In exemplary visualization, we observed that pre-pooled activation map encode important channel-specific characteristics that exhibit discriminative attributes between ID (blue) and OOD (orange) samples. Each point on the x-axis corresponds to a single channel, while the y-axis represents the strength of four statistical cues: (a) mean, (b) standard deviation, (c) maximum activation, and (d) entropy values.

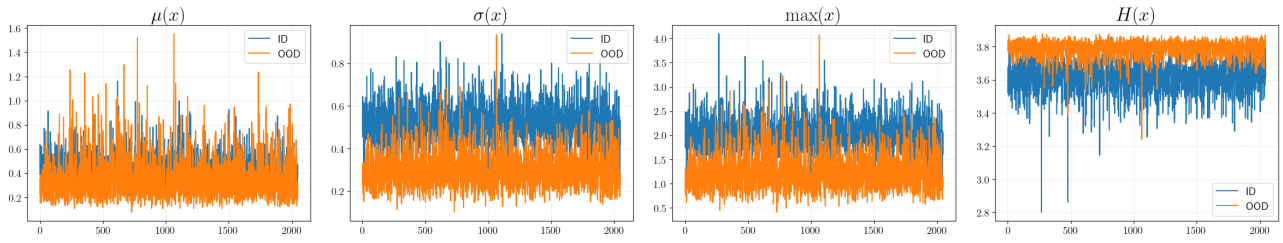


Figure 1. *Information cues from each channel before the penultimate layer of a ResNet-50 trained on ImageNet-1k, evaluated with Texture as the OOD dataset. The x-axis shows channel indices; the y-axis shows cue strength.* **Left to right***: (a) $\mu(\mathbf{x})$: mean activation, (b) $\sigma(\mathbf{x})$: standard deviation, (c) $\max(\mathbf{x})$: dominant activation, and (d) $H(\mathbf{x})$: entropy per channel.*

The existing methods have under-explored these distinctive statistical information. The approach exclusively relies on a score derived from the output logits [5, 16, 32, 36, 55, 67]: discards potent raw cues (e.g., standard deviation, maximum) and fails to leverage independent discriminative power of raw mean statistics. To address this critical limitation, we propose `Catalyst`, a simple yet powerful framework that computes an input-dependent *scaling factor* ($\gamma$) designed to be fused in tandem with an existing scoring function. This scaling factor is computed on-the-fly, leveraging these distribution-sensitive cues embedded in the pre-pooled activation maps. `Catalyst` is designed to integrate seamlessly with established approaches while significantly improving their ability to distinguish between ID and OOD data. Our key contributions are:

1. `Catalyst`, a complementary post-hoc OOD detection framework that leverages pre-pooling channel-wise statistics to augment existing methods, generalizing across architectures like ResNet, DenseNet, and MobileNet.
2. An extensive evaluation showing `Catalyst` complements and substantially improves established competitive baselines. Specifically, on the ImageNet benchmark, `Catalyst` reduces average FPR95 by 22.25% using ResNet-50. On CIFAR benchmarks, it reduces FPR95 by 32.87% on CIFAR-10 and 27.94% on CIFAR-100 using ResNet-18.
3. Statistical analysis (Appendix B) and extensive ablation studies (Section 5) validate our design choices.

## 2. Preliminaries

**Setup.** This paper focuses on the post-hoc analysis of multiclass classification in supervised settings. Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, 2, \cdots, C\}$ the output label space. A neural network $\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ is trained on a dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^{N}$ drawn *i.i.d.* from an unknown joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ over $\mathcal{X} \times \mathcal{Y}$. The network outputs a logit vector, which is used to predict the label of an input sample. $\mathcal{D}_{\text{in}}$ represents the marginal distribution of $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ over $\mathcal{X}$, corresponding to the ID data.

**Scoring Function.** As introduced in Section 1, the core challenge in OOD detection lies in designing effective scoring functions that reliably distinguish between ID and OOD samples. The evolution of scoring functions began with the MSP [16] approach and progressed to more robust energy-based scores [36]. While other scoring functions exist (e.g., ODIN [32], Mahalanobis [30], KNN [56]), we focus on the energy-based score $S_{\texttt{energy}}(\mathbf{x}; \theta)$ due to its prevalence, superior performance and simplicity [5, 36, 54, 55, 67]. Without loss of generality, all subsequent mentions of "score" refer to $S_{\texttt{energy}}(\mathbf{x}; \theta)$ unless specified otherwise. We adopt the negative free energy formulation from [36]. Formally, given a logit vector $f(\mathbf{x}) \in \mathbb{R}^C$ produced by the model $\theta$, the scoring functions is defined as:

$$S_{\texttt{energy}}(\mathbf{x}; \theta) = \log \left( \sum_{j=1}^{C} e^{f_j(\mathbf{x})} \right) \tag{1}$$

**Out-of-distribution Detection.** At inference time, the model $\theta$ operating in real-world will inevitably encounter OOD samples $\mathcal{D}_{\text{out}}$ whose label sets are disjoint from $\mathcal{Y}$. These samples should not be confidently predicted by $\theta$ as one of the known classes, instead necessitating robust OOD detection. Formally, we frame OOD detection as learning a decision boundary $G_\lambda(\mathbf{x}; \theta)$ that classifies a test sample $\mathbf{x} \in \mathcal{X}$ as either ID or OOD:

$$G_\lambda(\mathbf{x}; \theta) = \begin{cases} \text{ID} & \text{if } \mathbf{x} \sim \mathcal{D}_{\text{in}} \\ \text{OOD} & \text{if } \mathbf{x} \sim \mathcal{D}_{\text{out}} \end{cases} = \begin{cases} \text{ID} & \text{if } S(\mathbf{x}; \theta) \geq \lambda \\ \text{OOD} & \text{if } S(\mathbf{x}; \theta) < \lambda \end{cases} \tag{2}$$

where $S(\mathbf{x}; \theta)$ represents a downstream OOD scoring function, and by convention [36] $\lambda$ is a threshold calibrated such that 95% of ID data ($\mathcal{D}_{\text{in}}$) is correctly classified.

**Evaluation metrics.** In line with standard evaluation protocol in OOD detection [36], we evaluate the performance of `Catalyst` using two key metrics: FPR95 and AUROC:

1. *FPR95* measures the False Positive Rate when 95% of in-distribution (ID) samples are correctly classified. A lower FPR95 (↓) indicates better OOD detection performance.
2. *AUROC* is a threshold-free metric that computes the area under the receiver operating characteristic curve. Higher AUROC (↑) signifies superior discriminative capability.

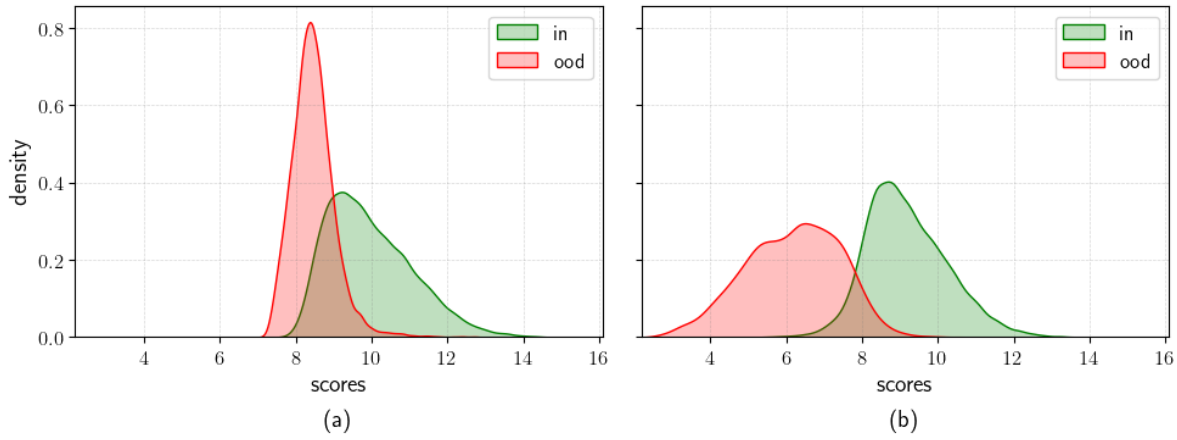


Figure 2. *Illustration of* `Catalyst`*'s effectiveness. The model is ResNet-50 trained on ImageNet-1k, evaluated on Texture (OOD). Here, we apply $\gamma$ computed from the channel-maximum statistic (m) multiplicatively to the baseline ReAct. (a) The unscaled score distribution shows more significant overlap than (b) the* `Catalyst`*-scaled score distribution.*

## 3. Methodology

The key contribution of this paper is `Catalyst`, a novel elastic scaling mechanism for enhanced OOD detection. We propose an input-dependent scaling factor ($\gamma$), derived from the overlooked channel-wise statistics of the penultimate layer's pre-pooling activation map. When this factor is fused with a baseline score, it significantly enhances the separability between ID and OOD samples. Figure 2 illustrates this effect with a trend representative of what we observe across the diverse models and OOD datasets in our evaluation. For instance, in the specific case depicted using a ResNet-50 trained on ImageNet, we see that while baseline score distributions for ID and Texture (OOD) data exhibit significant distributional overlap (Figure 2a), multiplicatively fusing $\gamma$ markedly reduces this overlap, enabling a much clearer separation (Figure 2b).

In this section, we describe the method to compute input-dependent $\gamma$ and how it is fused with score. Finally, we discuss the compatibility of `Catalyst` with other scoring functions and its integration with existing baselines.

### 3.1. Computing the Scaling Factor $\gamma$

To compute scaling factor $\gamma$, we consider a trained DNN $\theta : \mathbb{R}^d \to \mathbb{R}^C$ that maps an input $\mathbf{x} \in \mathbb{R}^d$ to a logit vector $f(\mathbf{x}) \in \mathbb{R}^C$, where $C = |\mathcal{Y}|$ denotes the number of classes. The network's penultimate layer produces a feature vector $h(\mathbf{x}) \in \mathbb{R}^n$ by applying GAP operation to the activation map $g(\mathbf{x}) \in \mathbb{R}^{n \times k \times k}$. Here, $n$ is the number of channels, and each channel has spatial resolution $k \times k$. A weight matrix $\mathbf{W} \in \mathbb{R}^{n \times C}$ projects $h(\mathbf{x})$ to the final logit vector.

In this work, we deliberately focus on this activation map $g(\mathbf{x})$ as our source of statistics. As we will empirically demonstrate in our ablation study (Section 5.1, Appendix H), this specific layer provides the most potent and reliable discriminative information cues for $\gamma$. The *earlier layers* provide less informative signal with high ID/OOD overlap.

`Catalyst` is built upon the core insight, illustrated in Figure 1, that the existing baselines' exclusive reliance on the feature vector fails to leverage valuable, channel-specific statistical information. Building upon this, we identify and extract three key statistical cues from $g(\mathbf{x})$:

- *Channel Mean* [$\mu(\mathbf{x}) \in \mathbb{R}^n$] is equivalent to the penultimate feature vector $h(\mathbf{x})$ obtained via GAP.[1]
- *Channel Standard Deviation* [$\sigma(\mathbf{x}) \in \mathbb{R}^n$] measures the spatial variability of activations within each channel.
- *Channel Maximum* [$m(\mathbf{x}) \in \mathbb{R}^n$] captures the peak activation response in each channel.

The information cues $\mu(\mathbf{x}), \sigma(\mathbf{x})$, and $m(\mathbf{x})$ for OOD samples may exhibit extreme unit activations. Prior work [55] presented a similar phenomenon of abnormally high unit activations that result in overconfident predictions for OOD samples, subsequently distorting the energy score. Extreme values in $\mu(\mathbf{x}), \sigma(\mathbf{x})$, and $m(\mathbf{x})$ can similarly distort scaling factor $\gamma$ for OOD samples. To mitigate this effect, we introduce a clipping mechanism that bounds each statistic by a threshold $c > 0$. Specifically, for each input, we compute rectified features via element-wise clipping:

$$\bar{f}(\mathbf{x}) = \min(f(\mathbf{x}), c) \tag{3}$$

where $f(\mathbf{x}) \in \{\mu(\mathbf{x}), \sigma(\mathbf{x}), m(\mathbf{x})\}$. This operation ensures that activation values are capped at $c$, preventing them from disproportionately influencing $\gamma$. The rectified vectors are the basis for $\gamma$'s calculation:

$$\gamma(\mathbf{x}; f) = \sum_{i=1}^{n} \bar{f}_i(\mathbf{x}) \tag{4}$$

where the subscript $i$ denotes the $i$-th channel. The selection of this clipping threshold $c$ is discussed in Section 4.5 and detailed in Appendix G.

[1] We use $\mu(\mathbf{x})$ and $h(\mathbf{x})$ interchangeably.

While we primarily focus on $\mu(\mathbf{x}), \sigma(\mathbf{x})$ and $m(\mathbf{x})$, our framework readily accommodates other channel-wise statistics derived from $g(\mathbf{x})$, such as entropy and median. We provide a detailed comparative analysis of these cues in our ablation study (Section 5.3; Appendix J), which justifies our design and validates our focus on $\mu(\mathbf{x}), \sigma(\mathbf{x})$, and $m(\mathbf{x})$ as robust and generalizable set of statistics for computing $\gamma$.

### 3.2. Elastic Scaling of the Score

To create a more discriminative score, we dynamically recalibrate the baseline score $S(\mathbf{x};\theta)$ using scaling factor $\gamma$. We explore the two fusion strategies: *multiplicative* and *additive*. We term the multiplicative strategy *"Elastic Scaling"* because it truly scales (i.e., multiplies) the baseline score, elastically stretching or shrinking it based on the $\gamma$. The additive approach, in contrast, is a simple offset or shift, not a scaling. These are defined in Equation 5:

$$S^{*}_{\text{mul}}(\mathbf{x};\theta,\gamma) = \gamma(\mathbf{x};f) \times S(\mathbf{x};\theta) \tag{5a}$$

$$S^{+}_{\text{add}}(\mathbf{x};\theta,\gamma) = \gamma(\mathbf{x};f) + S(\mathbf{x};\theta) \tag{5b}$$

where $\gamma(\mathbf{x};f)$ is the scaling factor computed from an information cue $f(\mathbf{x}) \in \{\mu(\mathbf{x}), \sigma(\mathbf{x}), m(\mathbf{x})\}$.

While our analysis in Section 5.2 shows that both strategies can achieve similar peak performance, we adopt multiplicative fusion (Eq. 5a) as our primary framework. This choice is not arbitrary, as we demonstrate that the additive approach, while effective, is operationally fragile due to its hyperparameter sensitivity. The multiplicative fusion provides not only competitive performance but also the practical robustness and stability required of a general-purpose usage. Therefore, in the remainder of this paper, we will refer multiplicative fusion as *Elastic Scaling*. This final recalibrated score is subsequently used in the decision rule defined in Equation 2 to classify as ID or OOD.

### 3.3. Generalizability of `Catalyst`

While our analysis primarily focuses on energy-based scoring (given its primary role for competitive methods like ReAct and SCALE), `Catalyst` is a general framework. It can be seamlessly integrated with other scoring functions such as MSP [16], ODIN [32], and KNN [11, 56] – by replacing the baseline score $S(\mathbf{x};\theta)$ in equation (Eq. 5a) with the alternate score.

Additionally, this elastic scaling retains all advantages of post-hoc methods while transforming scores into a more discriminative metric. `Catalyst` is designed to complement existing techniques, including Energy, ReAct, DICE, ASH, SCALE, and KNN. In Appendix B, we provide a formal characterization of why `Catalyst` enhances ID-OOD separability, offering deeper insight.

## 4. Experiments

In this section, we evaluate the efficacy of `Catalyst` across a diverse set of OOD datasets. We begin with an in-depth empirical analysis on standard CIFAR benchmarks. We then extend our evaluation to a large-scale OOD detection setting using ImageNet, demonstrating the versatility and robustness of `Catalyst`. Our evaluation does not assume the availability of an OOD validation set and incorporates a wide range of OOD datasets to provide a realistic assessment of `Catalyst`.

We use the Energy score [36] as our default baseline. For brevity, when `Catalyst` is applied to Energy, we simply denote it as `Catalyst`. When applying `Catalyst` to other baselines or scoring functions, we state it explicitly (e.g., `Catalyst` + `ReAct`, `Catalyst` + `KNN`).

We ensure a fair and direct comparison against prior work. As the architectures used in our evaluation (e.g., ResNet-18 for the CIFAR benchmarks; ResNet-34 and DenseNet-121 for ImageNet) were not included in the primary baselines like ReAct, DICE, ASH, and SCALE, we undertook a rigorous re-evaluation of these methods. We carefully followed the official hyperparameter selection protocols and open-sourced implementations from their respective papers to ensure the integrity of our comparisons.

### 4.1. CIFAR Evaluation

| Model | Method | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|---|
| | | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ |
| ResNet-18 | MSP | 58.33 | 91.28 | 79.92 | 76.66 |
| | ODIN | 28.98 | 95.16 | 66.06 | 84.78 |
| | Energy | 35.50 | 94.17 | 70.21 | 83.54 |
| | ReAct | 29.76 | 95.19 | 57.76 | 87.97 |
| | DICE | 30.98 | 94.69 | 55.66 | 85.97 |
| | ReAct+DICE | 19.65 | 96.50 | 48.23 | 89.06 |
| | ASH | 20.96 | 95.95 | 49.52 | 86.86 |
| | SCALE | 21.05 | 96.19 | 48.10 | 88.70 |
| | **`Catalyst`**$(\mu)$ | 24.85 | 95.74 | 52.93 | 87.46 |
| | **`Catalyst`**$(\sigma)$ | 17.72 | 96.89 | 46.29 | 89.18 |
| | **`Catalyst`**$(m)$ | 16.59 | 97.10 | 45.96 | 89.37 |
| | **`Catalyst`**$(\mu)$ + **`ReAct`** | 19.88 | 96.41 | 41.93 | 89.99 |
| | **`Catalyst`**$(\sigma)$ + **`ReAct`** | 14.25 | 97.42 | 35.15 | 91.48 |
| | **`Catalyst`**$(m)$ + **`ReAct`** | **13.19** | **97.59** | **34.66** | **91.70** |
| DenseNet-101 | MSP | 45.43 | 92.43 | 77.47 | 74.80 |
| | ODIN | 19.37 | 96.06 | 57.67 | 84.00 |
| | Energy | 22.41 | 95.43 | 58.92 | 83.87 |
| | ReAct | 17.13 | 96.61 | 52.89 | 87.18 |
| | DICE | 14.52 | 96.74 | 40.98 | 87.92 |
| | ReAct+DICE | 10.26 | 97.94 | 34.64 | 91.17 |
| | ASH | 11.71 | 97.44 | 35.84 | 90.85 |
| | SCALE | 19.88 | 96.01 | 38.31 | 90.46 |
| | **`Catalyst`**$(\mu)$ | 13.73 | 97.14 | 41.42 | 89.45 |
| | **`Catalyst`**$(\sigma)$ | 10.93 | 97.71 | 37.98 | 90.48 |
| | **`Catalyst`**$(m)$ | 10.71 | 97.77 | 36.79 | 90.83 |
| | **`Catalyst`**$(\mu)$ + **`ReAct`** | 10.24 | 97.85 | 29.36 | 92.56 |
| | **`Catalyst`**$(\sigma)$ + **`ReAct`** | 8.49 | 98.21 | 29.05 | 92.78 |
| | **`Catalyst`**$(m)$ + **`ReAct`** | **8.42** | **98.26** | **28.06** | **93.06** |

Table 1. *OOD detection results on CIFAR benchmarks. All values are percentages, averaged across six OOD test datasets. Full results for each dataset are available in Appendix E.3.* ↓ / ↑ *indicates lower / higher values are better.*

**Experimental Setup.** We evaluate on the CIFAR datasets [27]. Following standard protocols [5, 36, 55], we use six common OOD datasets for evaluation: Textures [3], SVHN [46], Places365 [73], LSUN-Crop [70],

| Method | ResNet-34 | | ResNet-50 | | MobileNet-v2 | | DenseNet-121 | |
|---|---|---|---|---|---|---|---|---|
| | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ |
| MSP | 68.84 | 81.19 | 64.76 | 82.82 | 70.49 | 80.67 | 63.46 | 82.65 |
| ODIN | 55.90 | 87.16 | 56.48 | 85.41 | 54.20 | 85.81 | 49.45 | 87.48 |
| Energy | 57.20 | 86.84 | 57.48 | 87.05 | 58.87 | 86.59 | 50.68 | 87.60 |
| ReAct | 32.24 | 93.08 | 30.77 | 93.27 | 48.91 | 88.75 | 35.99 | 92.27 |
| DICE | 39.12 | 89.96 | 35.65 | 90.94 | 41.07 | 89.94 | 38.67 | 89.65 |
| ReAct+DICE | 26.25 | 93.99 | 25.41 | 94.10 | 31.06 | 92.84 | 29.33 | 93.42 |
| ASH | 29.32 | 93.46 | 22.83 | 95.12 | 38.68 | 90.95 | 30.25 | 93.09 |
| SCALE | 27.02 | 94.14 | 21.89 | 95.32 | 34.28 | 92.52 | 28.06 | 93.45 |
| Catalyst($\mu$) | 31.92 | 92.41 | 28.42 | 93.23 | 36.71 | 91.69 | 29.54 | 92.71 |
| Catalyst($\sigma$) | 31.91 | 92.36 | 29.75 | 92.92 | 33.63 | 92.27 | 29.12 | 92.80 |
| Catalyst($m$) | 31.83 | 92.34 | 29.89 | 92.82 | 33.15 | 92.33 | 29.45 | 92.68 |
| Catalyst($\mu$) + ReAct | **19.84** | **95.56** | **17.02** | **96.18** | 30.81 | 93.31 | 25.43 | 94.56 |
| Catalyst($\sigma$) + ReAct | 19.91 | 95.50 | 17.46 | 96.02 | 31.56 | 92.71 | **24.26** | **94.61** |
| Catalyst($m$) + ReAct | 20.16 | 95.44 | 17.64 | 95.93 | **29.33** | **93.43** | 24.52 | 94.53 |

Table 2. *OOD detection results on ImageNet benchmarks. All values are percentages and are averaged over four common OOD benchmark datasets. Complete results for each individual dataset are available in Appendix E.2.* ↓ / ↑ *indicates lower / higher values are better.*

LSUN-Resize [70], and iSUN [68]. To ensure fair comparison with prior work, we use a DenseNet-101 backbone [19]. To demonstrate architectural generality, we extend our evaluation to ResNet-18 [13]. Training details are detailed in Appendix G.

**Results.** Table 1 summarizes our results on the CIFAR benchmarks. The table clearly shows the two key benefits(1) Catalyst (e.g., Catalyst($m$)) significantly outperforms the standard energy score baseline, proving the inherent value of scaling factor. (2) When composed with ReAct, Catalyst establishes a new benchmark. For instance, on CIFAR-10, Catalyst($m$) + ReAct reduces FPR95 by 32.87%, 28.10% with ResNet-18 and DenseNet-101 respectively. On CIFAR-100, Catalyst($m$) + ReAct reduces FPR95 by 27.94% and 18.99% with ResNet-18 and DenseNet-101 respectively. The detailed per-dataset results are provided in Appendix E.3.

**Near-OOD Evaluation.** We evaluate Catalyst on the challenging near-OOD task of distinguishing CIFAR-10 from CIFAR-100 [11]. While it yields marginal gains over SCALE, the improvements are less pronounced than in far-OOD settings, likely due to the high similarity of learned penultimate representations. Designing a more effective $\gamma$ for near-OOD settings remains an important direction for future work. Detailed results are provided in Appendix E.1.

## 4.2. ImageNet Evaluation

**Experimental Setup.** To assess scalability in a more realistic setting, we evaluate on the ImageNet-1k benchmark. We use four OOD datasets: iNaturalist [60], SUN [66], Places365 [73], and Textures [3]. These datasets are carefully curated to avoid class overlap with ImageNet, while spanning distinct semantic domains to rigorously assess generalization performance [36, 55].

Our evaluation showcases broad architectural robustness by using pre-trained ResNet-34, ResNet-50, DenseNet-121, and MobileNet-v2. Since primary baselines (e.g., ReAct, SCALE) did not originally report results on all of these architectures (such as ResNet-34 and DenseNet-121), we undertook a rigorous re-evaluation of all methods.

**Results**. Table 2 shows that Catalyst yields consistent improvements at ImageNet scale. Compared to energy score, Catalyst($m$) reduces FPR95 by 44.35%, 47.99%, 43.69%, and 21.23% using ResNet-34, ResNet-50, MobileNet-v2, and DenseNet-121 architectures respectively. The most significant gains are achieved when composing Catalyst with existing primary methods like ReAct. Specifically, Catalyst($m$)+ReAct improves FPR95 by 25.39%, 19.41%, 5.57% and 12.62% compared to previous best results using ResNet-34, ResNet-50, MobileNet-v2, and DenseNet-121 respectively. These results validate that the principles of Catalyst are effective in complex, large-scale datasets and across diverse architectural families. The performance boost confirms that the scaling factor $\gamma$ provides significant discriminative information that is complementary to existing competitive techniques. The detailed per-dataset results are in Appendix E.2.

**Discussion.** While we acknowledge standardized benchmarks like OpenOOD [72], we adopted a more challenging and principled evaluation for two key reasons: *(a)* OpenOOD's dataset selection excludes several difficult, widely-used testbeds like SUN [66], Places [73], and four complex categories (bubbly, honeycombed, cobwebbed, and spiralled) from Texture [3, 62]. *(b)* OpenOOD's setup uses held-out OOD validation set for hyperparameter tuning. Our evaluation is conducted without assuming the availability of an OOD validation set and incorporates these difficult datasets to provide a more rigorous and realistic assessment of Catalyst.

We also explored combining statistical cues (e.g., mean + std) to compute $\gamma$. Our empirical analysis showed these multivariate combinations did not yield significant per-

formance gains over the best-performing single statistic. This finding reinforces our framework's simplicity and efficiency, as a single, well-chosen statistic is sufficient to provide a robust performance boost.

### 4.3. Synergy with Existing Baselines

We evaluated the performance of existing baselines when applied in tandem with `Catalyst` to demonstrate its complementary effect. The results show that `Catalyst` provides consistent relative performance boost across baselines on CIFAR and ImageNet. For example, on ImageNet `Catalyst`$(\mu)$ + `DICE` improves relative FPR95 by 22.24% (ResNet-50) and 15.41% (MobileNet-v2) respectively. A detailed breakdown is provided in Appendix M.

### 4.4. Generalizability to Distance-Based Methods

To validate `Catalyst` as a general-purpose framework, we test its synergy with a distance-based K-Nearest Neighbors (KNN) [11, 56] OOD detector. The results in Tables 3 and 4 confirm our hypothesis, showing that `Catalyst` provides significant improvement over the KNN baseline across all benchmarks. For instance, on CIFAR-100 (ResNet-18), `Catalyst`$(m)$ achieves a 43.84% reduction. Similarly on, large-scale ImageNet benchmark, where `Catalyst`$(\mu)$ on a ResNet-50 results in a 52.13% reduction in average FPR95. These performance boost highlight that `Catalyst` is a general-purpose modulator, providing complementary information for both logit and distance based methods, making it a true *plug-and-play* framework. The full experimental setup and detailed per-dataset results are in Appendix D.

Extending our framework to gradient-based methods [2, 21] remains future work due to engineering challenges. Furthermore, we omit Mahalanobis [30] as a baseline, following recent precedents [5, 54, 55], owing to its high computational cost and limiting performance.

| Model | Method | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|---|
| | | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ |
| ResNet-18 | KNN | 31.02 | 95.00 | 66.81 | 83.40 |
| | + `Catalyst`$(\mu)$ | 25.54 | 96.18 | 52.77 | 87.98 |
| | + `Catalyst`$(\sigma)$ | 16.87 | 97.28 | 38.28 | 90.80 |
| | + `Catalyst`$(m)$ | **15.62** | **97.45** | **37.52** | **90.99** |
| DenseNet-101 | KNN | 13.08 | 97.51 | 41.97 | 88.29 |
| | + `Catalyst`$(\mu)$ | 9.49 | 98.05 | 36.42 | 91.51 |
| | + `Catalyst`$(\sigma)$ | 8.50 | 98.18 | 32.75 | 92.30 |
| | + `Catalyst`$(m)$ | **8.30** | **98.23** | **32.06** | **92.48** |

Table 3. *Generalizability of* `Catalyst` *to KNN-based OOD detection on the CIFAR benchmarks. All values are averaged across six OOD test datasets. ↓ / ↑ indicates lower / higher values are better. Full per-dataset results are in Appendix D.*

### 4.5. Hyperparameter Selection

The clipping threshold $\boldsymbol{c}$ (Eq. 3) is crucial for enhanced performance, as it must be set to optimally distinguish ID from OOD data. Analogous to ReAct [55], we set $\boldsymbol{c}$ to the $p$-th percentile of the ID activation distribution. The choice of this percentile $p$ is the key hyperparameter to be tuned. To demonstrate its sensitivity, we summarize the OOD detection performance of `Catalyst`$(m)$ in Figure 3, varying $p$ from 10 to 100 at 5-point intervals. To this end, we follow established protocols [54, 55] and create a proxy OOD validation set, generated by adding pixel-wise Gaussian noise to images from the ID validation set. We then select the percentile $p$ that yields the best OOD separation on this proxy task. This two-step procedure, which uses a percentile for the mechanism and a proxy set for tuning, is a robust tuning strategy grounded in prior work. The specific details and the selected $p$ values are provided in Appendix G.

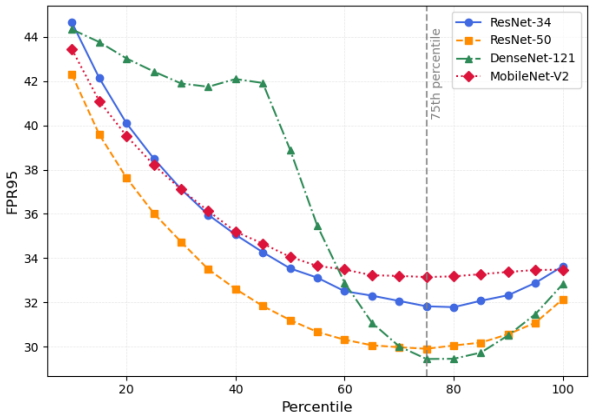


Figure 3. *Sensitivity analysis of the clipping percentile (p) on* `Catalyst`$(m)$ *performance. All values averaged over 4 OOD test datasets for a ResNet-50 (ImageNet).*

### 4.6. Comparison with Other Baselines

Comparing `Catalyst` against three contemporary methods, AdaScale [50], NCI [35] and fDBD [34], confirms its superiority, particularly when used as a complementary module. Aginst, AdaScale's reported reslults using DenseNet-101 on CIFAR-100, `Catalyst`$(m)$ + `ReAct` outperforms AdaScale yielding a 32.45% gain over the best AdaScale variant. Against NCI's reported results (which were obtained on the OpenOOD settings), `Catalyst`$(m)$ + `ReAct` achieves the average FPR95 by a significant 33.43% on CIFAR-10 (ResNet-18). This advantage is even more pronounced against fDBD, where our method achieves a substantial FPR95 reduction of 65.54% on ImageNet (ResNet-50). We also provide a detailed comparison with 19 existing OOD detection methods in literature in Appendix F.

### 4.7. Accuracy and Computational Overhead

Our post-hoc method, `Catalyst`, maintains the original ID classification accuracy of the base model, as it does not alter its inference path. Furthermore, its computational

| Method | ResNet-34 | | ResNet-50 | | MobileNet-v2 | | DenseNet-121 | |
|---|---|---|---|---|---|---|---|---|
| | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ |
| KNN | 73.26 | 93.47 | 64.05 | 95.56 | 75.54 | 91.75 | 74.01 | 92.11 |
| + Catalyst($\mu$) | **34.69** | **97.99** | **31.11** | **98.46** | **46.77** | **97.10** | **43.55** | **97.52** |
| + Catalyst($\sigma$) | 43.40 | 97.18 | 39.85 | 97.79 | 50.85 | 96.61 | 48.35 | 96.85 |
| + Catalyst($m$) | 43.16 | 97.17 | 39.60 | 97.78 | 50.52 | 96.61 | 48.89 | 96.75 |

Table 4. *Generalizability of* Catalyst *to KNN-based OOD detection on the ImageNet benchmarks. All values are averaged across six OOD test datasets. ↓ / ↑ indicates lower / higher values are better. Full per-dataset results are in Appendix D.*

overhead is negligible. The cost depends on the statistic used. Catalyst($\mu$) is the most efficient, as the mean is already computed by the standard GAP. The additional cost is less than 0.0001% of a ResNet-50's forward pass. Whereas, Catalyst($\sigma$), our most complex statistic, still adds less than 0.01% overhead. This confirms Catalyst is lightweight and efficient framework. A detailed breakdown of accuracy and FLOPs is provided in Appendix C.

## 5. Ablation Study

### 5.1. Choice of Layer for Computing $\gamma$

A core methodological decision is which network layer provides the most discriminative signal for $\gamma$. We conducted an analysis to locate this optimal signal source and discovered a critical and consistent trend. Using a pre-trained ResNet-50 on ImageNet-1k as a representative example, we found that the $\gamma$ distributions from the early-to-mid residual stages (Layers 1-3) are not sufficiently discriminative, exhibiting high overlap between ID and OOD data and rendering them ineffective (As shown in Figure 8 of Appendix H).

This finding is intuitively aligned with the principles of hierarchical feature learning [47, 71]. These initial layers learn general, low-level features such as edges, textures, and color blobs, which are fundamental properties shared by all natural images. Since both ID and OOD samples contain these common features, their activation statistics in these early layers are highly similar, resulting in the non-discriminative, overlapping $\gamma$ distributions we observed. In sharp contrast, the distribution from the final residual stage (Layer 4), immediately preceding GAP, provides a better separation, because it is trained to recognize the complex, high-level concepts and structures specific to the ID classes, which OOD samples lack. This analysis, which held true across all tested OOD datasets and architectures, empirically validates our focus: the penultimate layer's pre-pooling feature map is not a layer of convenience but the most reliable source of a potent signal for Catalyst. The complete details are presented in Appendix H.

### 5.2. Analysis of Fusion Strategy

In Section 3, we alluded to two fusion strategies: multiplicative(*) and additive(+). We investigated both to validate our design choice. Our analysis on the ImageNet (Table 19 in Appendix I) reveals that both strategies can achieve a similar high level of performance, confirming the discriminative power of the scaling factor $\gamma$ itself.

However, we found a critical difference in their hyperparameter robustness. The optimal additive method required tuning its clipping threshold $c^+$ at an extremely low percentile (e.g., $\leq$ 1st percentile for ResNet-50), making it operationally fragile and highly sensitive to data shifts. In sharp contrast, our proposed multiplicative method tunes its threshold $c^*$ at a stable, moderate percentile, aligning with robust foundational methods like ReAct and SCALE.

Given its superior robustness and practical stability, we selected multiplicative fusion as our primary strategy. A detailed analysis of this comparison is provided in Appendix I.

### 5.3. Alternate Statistics: Median and Entropy

To validate our choice of statistics (mean, std, max), we performed a rigorous analysis of two alternatives: median and Shannon entropy. This study found that median is not a viable statistic. It consistently degrades performance across all benchmarks, as its statistical signature fails to produce a discriminative $\gamma$ (see Fig. 9 in Appendix J). The study of Shannon entropy revealed it to be inconsistent. While it provided a strong 14.65% improvement in a specific case (MobileNet-V2 on ImageNet), this performance was not generalizable, with minimal gains on other architectures like ResNet-50.

This confirms our design choice: median was skipped for being ineffective, and entropy was rejected for being unreliable. Our proposed combination of mean, std, and max provides the most robust and consistently high-performing signal. Our full analysis is presented in Appendix J.

### 5.4. Scaling Factor $\gamma$ as a Scoring Metric

We conducted an analysis to determine if $\gamma$ is powerful enough to serve as a standalone OOD score, similar to MSP [16] or Energy [36]. Our findings show that $\gamma$ computed from std ($\gamma_{\text{std}}$) and max ($\gamma_{\text{max}}$) are consistently robust signals. On both the CIFAR benchmarks (Table 21) and the large-scale ImageNet benchmark (Table 20), these two statistics are consistently better than Energy baseline, proving they are viable and generalizable standalone scores. In contrast, the entropy provides a critical insight. While

$\gamma_{\text{entropy}}$ appears to be the distinguishable signal on CIFAR, this trend is inconsistent on ImageNet. On this more complex benchmark, $\gamma_{\text{entropy}}$ fails to generalize, suffering a performance collapse and lagging behind Energy. This analysis confirms that entropy, while potent in some cases is not a reliable or generalizable statistic for a robust, all-purpose method. The complete details are provided in Appendix K.

## 6. Scope and Future Work

We evaluated `Catalyst` using three specific statistics: mean, standard deviation, and max. As our ablations (Appendix J) demonstrated, this choice was deliberate, as other statistics like median were ineffective and entropy was not generalizable. While other aggregate functions could be explored, our focus remained on this robust set.

Additionally, we provide a comprehensive analysis focused on CNN-based architectures, with `Catalyst` applied within this setting. This focus is motivated by two factors: *(a)* Competitive baselines in the literature [5, 33, 36, 44, 54, 55, 61, 67] extensively use CNN-based architectures. For fair comparison, we adopt similar architectures to evaluate `Catalyst`. *(b)* CNN-based architectures continue to be widely used in both the research community and real-world applications. A comprehensive benchmark study carried out in prior work [12] has shown convolutional networks such as ResNet [13] and ConvNeXt [37, 65] remain the default choice in real-world vision systems (including object detection, segmentation, retrieval, and classification) due to their strong inductive bias (translation invariance), computational efficiency, strong performance on moderate-scale data, and extensive ecosystem of pretrained models.

The core principle of our method, leveraging statistical cues from penultimate pre-pooled activation map is a general strategy that can be extended beyond CNNs to architectures like Vision Transformers (ViTs) [6]. However, adapting `Catalyst` to derive an effective scaling factor $\gamma$ from the intermediate blocks of a transformer requires substantial research and engineering. We are actively exploring the extension of our framework to transformer-based models.

## 7. Related Work

**Scoring-based OOD Detection.** Post-hoc OOD detection is dominated by the design of scoring functions. Early work on MSP [16] and its variants [18, 20, 32] was shown to be vulnerable to model overconfidence [15, 48]. This led to the development of energy-based scores [36], which have become the foundation for most logit-based OOD detection methods [5, 54, 55, 67] due to their superior performance. Other families of scores exist, including distance-based (e.g., Mahalanobis [30], KNN [11, 49, 53, 56], fDBD [34], NCI [35]), gradient-based (e.g., GradNorm [21], GradOrth [2]), Virtual-logit [62] and Bayesian approaches [10, 28, 38–40], etc. `Catalyst` is designed to complement and enhance these scoring paradigms.

**Post-hoc Pruning based OOD Detection.** Recent approaches like ReAct [55], DICE [54], ASH [5], and SCALE [67] operate post-hoc by pruning [1, 31] or modifying feature representations, often using simple heuristics over penultimate activations. Our method, `Catalyst`, reveals that activation channels of layers prior to penultimate layer possess rich statistical information cues that, when exploited, can substantially improve OOD detection performance when combined with these approaches. It complements existing sparse representation techniques and is easy to integrate into standard pipelines.

**Generative OOD Detection.** Generative models identify OOD samples by estimating data density [4, 22, 26, 51, 57, 59], but recent work [45] has shown they may assign high likelihoods to OOD inputs. Moreover, these models are often harder to train and less reliable than discriminative approaches [5, 16, 21, 32, 36, 54, 55, 67]. Thus, we primarily focused on such discriminative approaches, while showcasing generality using KNN [56]. However, if a generative method relies on a scalar-based scoring function, then `Catalyst` can also be extended to such generative methods.

**Training-Time OOD Detection Methods.** A distinct line of work involves modifying the model's training objective with regularization techniques to improve OOD separation [15, 17, 23, 25, 29, 36, 39, 41, 42, 58, 64, 69]. These methods often encourage uniform predictions for outliers [17, 29] or explicitly penalize low energy scores for out-of-distribution samples during training [7, 8, 25, 36, 42]. In contrast, our work is entirely post-hoc, requiring no changes to the training process. This is highly practical and broadly applicable, particularly in scenarios involving large models where retraining is costly or infeasible.

## 8. Conclusion

`Catalyst` is a simple yet powerful post-hoc framework for OOD detection that challenges the conventional paradigm of using only the pooled feature vector from the penultimate layer. We demonstrated that rich, discriminative information cues were being discarded, namely, the channel-wise statistics embedded in penultimate layer's pre-pooled feature map. `Catalyst` effectively harnesses this under-explored information by computing an input-dependent scaling factor ($\gamma$) that modulates existing baseline scores, significantly enhancing the separation between ID and OOD distributions. Extensive experiments across diverse models and datasets demonstrate that `Catalyst` consistently outperforms recent competitive baselines OOD detection methods. These empirical findings are further supported by ablation studies and statistical analysis.

## References


[1] Mohammad Babaeizadeh, Paris Smaragdis, and Roy H. Campbell. Noiseout: A simple way to prune neural networks. *CoRR*, abs/1611.06211, 2016. 8

[2] Sima Behpour, Thang Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 6, 8, 4, 13, 14

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, page 3606–3613, 2014. 4, 5, 6, 9, 11, 12

[4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. 8

[5] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 4, 6, 8, 13, 14, 24

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 8

[7] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[8] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, 2022. 8

[9] Angelos Filos, Panagiotis Tigas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning (ICML)*, 2020. 1

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059, 2016. 8

[11] Soumya Suvra Ghosal, Yiyou Sun, and Yixuan Li. How to overcome curse-of-dimensionality for out-of-distribution detection? In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, 2024. 1, 4, 5, 6, 8, 7

[12] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Uday Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 8

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 8

[14] Rundong He, Yue Yuan, Zhongyi Han, Fan Wang, Wan Su, Yilong Yin, Tongliang Liu, and Yongshun Gong. Exploring channel-aware typical features for out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:12402–12410, 2024. 14

[15] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 1, 8

[16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 1, 2, 4, 7, 8, 13, 14, 24

[17] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10948–10957, 2020. 8, 13, 14

[19] 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 (CVPR)*, 2017. 5

[20] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8710–8719, 2021. 8

[21] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, pages 677–689. Curran Associates, Inc., 2021. 6, 8, 13, 14

[22] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[23] Taewon Jeong and Heeyoung Kim. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. In *Advances in Neural Information Processing Systems*, pages 3907–3916, 2020. 8

[24] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 4

[25] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training OOD detectors in their natural habitats.

In *Proceedings of the 39th International Conference on Machine Learning*, pages 10848–10865, 2022. 8

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 8

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. 8

[29] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. 8

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018. 2, 6, 8, 4, 13, 14

[31] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017. 8

[32] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 2, 4, 8, 1, 13, 14, 24

[33] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15308–15318, 2021. 8

[34] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. *ICML Workshop or arXiv preprint*, 2023. 6, 8, 13, 14

[35] Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6, 8, 13

[36] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475. Curran Associates, Inc., 2020. 1, 2, 4, 5, 7, 8, 13, 14, 24

[37] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 8

[38] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, 2019. 8

[39] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018. 8

[40] Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019. 8

[41] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations*, 2020. 8

[42] Yifei Ming, Ying Fan, and Yixuan Li. POEM: Out-of-distribution detection with posterior sampling. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15650–15665, 2022. 8

[43] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2023. 14

[44] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI conference on Aritificial Intelligence*, 2021. 8

[45] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. 8

[46] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 4, 6, 11, 12

[47] Tuan Ngo, Abid Hassan, Saad Shafiq, and Nenad Medvidovic. Dnn modularization via activation-driven training, 2025. 7

[48] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015. 1, 8

[49] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection, 2023. 8, 13, 14

[50] Sudarshan Regmi. Adascale: Adaptive scaling for ood detection, 2025. 6, 17

[51] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models, 2014. 8

[52] A.G. Roy, J Ren, S Azizi, A Loh, V Natarajan, B Mustafa, N Pawlowski, J Freyberg, Y Liu, and Z Beaver. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *CoRR*, arXiv:2104.03829, 2021. 1

[53] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *Proceedings of the International Conference on Learning Representations*, 2021. 8, 4

[54] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Computer Vision – ECCV 2022*, pages 691–708. Springer Nature Switzerland, 2022. 1, 2, 6, 8, 4, 13, 14, 15, 24

[55] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, pages 144–157. Curran Associates, Inc., 2021. 1, 2, 3, 4, 5, 6, 8, 13, 14, 15, 24

[56] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20827–20840, 2022. 1, 2, 4, 6, 8, 13, 14, 24

[57] E. Tabak and Turner Cristina. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66:145–164, 2013. 8

[58] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9690–9700, 2020. 8

[59] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 2016. 8

[60] 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 (CVPR)*, 2018. 5, 9

[61] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In *Advances in Neural Information Processing Systems*, 2021. 8

[62] 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*, 2022. 5, 8, 7, 13, 14

[63] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[64] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. 8

[65] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, 2023. 8

[66] 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, 2010. 5, 9

[67] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 8, 13, 14, 16

[68] Pingmei Xu, Krista A. Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *CoRR*, 1504.06755, 2015. 5, 6, 11, 12

[69] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically Coherent Out-of-Distribution Detection . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8

[70] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, 1506.03365, 2015. 4, 5, 6, 11, 12

[71] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013. 7

[72] 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 Li Hai. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 5

[73] 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. 4, 5, 6, 9, 11, 12

[74] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, bolun zheng, and Yaowu Chen. Boosting out-of-distribution detection with typical features, 2022. 13, 14