Catalyst: Out-of-Distribution Detection via Elastic Scaling

Supplementary Material

A. Description of Baseline Methods

In resonance with existing work [5, 36, 54, 55], for the reader's convenience, we summarize in detail a few common techniques for defining OOD scores that measure the degree of ID-ness on the given sample. All the methods derive the score post hoc on neural networks trained with in-distribution data only. By convention, a higher score is indicative of being in-distribution, and vice versa.

Softmax score One of the earliest works on OOD detection considered using the maximum softmax probability (MSP) to distinguish between $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$ [16]. In detail, suppose the label space is $\mathcal{Y} = \{1, 2, \cdots, C\}$. We assume the classifier f is defined in terms of a feature extractor $f : \mathcal{X} \to \mathbb{R}^m$ and a linear multinomial regressor with weight matrix $W \in \mathbb{R}^{C\times m}$ and bias vector $\mathbf{b} \in \mathbb{R}^C$. The prediction probability for each class is given by :

$$\mathbb{P}(y = c|\mathbf{x}) = \text{Softmax}(Wh(\mathbf{x}) + \mathbf{b})_c \tag{6}$$

The softmax score is defined as $S_{\text{MSP}}(\mathbf{x}; f) := \max_c \mathbb{P}(y = c|\mathbf{x})$.

ODIN [32] This method introduced temperature scaling and input perturbation to improve the separation of MSP for ID and OOD data. $\tilde{\mathbf{x}}$ denotes perturbed input.

$$\mathbb{P}(y = c|\tilde{\mathbf{x}}) = \text{Softmax}[(Wh(\tilde{\mathbf{x}}) + \mathbf{b})/T]_c \tag{7}$$

the ODIN score is defined as $S_{\text{ODIN}}(\mathbf{x}; f) := \max_c \mathbb{P}(y = c|\tilde{\mathbf{x}})$.

Energy score The energy function [36] maps the output logit to a scalar $S_{\text{Energy}}(\mathbf{x}; f) \in \mathbb{R}$, which is relatively lower for ID data:

$$S_{\text{Energy}}(\mathbf{x}; f) = -\text{Energy}(\mathbf{x}; f) = \log\left(\sum_{c=1}^{C} \exp(f_c(\mathbf{x}))\right) \tag{8}$$

They used the *negative energy score* for OOD detection, in order to align with the convention that $S(\mathbf{x}; f)$ is higher for ID data and vice versa.

ReAct They perform post hoc modification of penultimate layer of the neural network. It works by truncating the feature activations at a threshold c, i.e., replacing each activation with $\min(x, c)$. This limits the influence of abnormally large activations often caused by OOD inputs.The truncation threshold is set with the validation strategy in [55].Formally,

$$\begin{aligned} h^{\text{ReAct}}(\mathbf{x}) &= \text{ReAct}(h(\mathbf{x}); c) \\ &= \min(h(\mathbf{x}), c) \quad \text{(applied element-wise)} \end{aligned}$$

The final model output becomes:

$$f^{\text{ReAct}}(\mathbf{x}) = W^\top h^{\text{ReAct}}(\mathbf{x}) + \mathbf{b}$$

This method also uses energy score $S_{\text{Energy}}(\mathbf{x}; f^{\text{ReAct}}) \in \mathbb{R}$ for OOD detection.

DICE [54] It is a post hoc method to improve OOD detection by retaining only the most informative weights in the final layer of a pre-trained neural network. A *contribution matrix* $V \in \mathbb{R}^{m\times C}$ is computed, where each column is:

$$\mathbf{v}_c = \mathbb{E}_{\mathbf{x}\in\mathcal{D}}[\mathbf{w}_c \odot h(\mathbf{x})]$$

with $\odot$ denoting element-wise multiplication. Each entry in V quantifies the average contribution of a feature unit to class c. A binary *masking matrix* $M \in \mathbb{R}^{m\times C}$ selects the top-k highest-contributing weights, setting others to zero. The sparsified output is:

$$f^{\text{DICE}}(\mathbf{x}; \theta) = (M \odot W)^\top h(\mathbf{x}) + \mathbf{b}$$

This method also uses energy score $S_{\text{Energy}}(\mathbf{x}; f^{\text{DICE}}) \in \mathbb{R}$ for OOD detection.

ASH [5] It is also a post-hoc method that simplifies feature representations to improve OOD detection. They proposes three versions of ASH, we presented only the best performing version i.e, ASH-S. Given an input activation vector $h(\mathbf{x})$ and a pruning percentile p, ASH [5] proceeds as follows shaping the activation of penultimate layer $h(\mathbf{x})$ to get $h^{\text{ASH}}(\mathbf{x})$:

1. Compute the p-th percentile threshold t of $h(\mathbf{x})$.
2. Let $s_1 = \sum h(\mathbf{x})$, the sum of all activation values before pruning.
3. Set all values in $h(\mathbf{x})$ less than t to zero.
4. Let $s_2 = \sum h(\mathbf{x})$, the sum after pruning.
5. Scale all non-zero values in $h(\mathbf{x})$ by $\exp(s_1/s_2)$.

The final model output becomes, which is then used to compute energy score $S_{\text{Energy}}(\mathbf{x}; f^{\text{ASH}}) \in \mathbb{R}$ for OOD detection :

$$f^{\text{ASH}}(\mathbf{x}) = W^\top h^{\text{ASH}}(\mathbf{x}) + \mathbf{b}$$

SCALE [67] It is a post-hoc method designed to enhance out-of-distribution (OOD) detection by adaptively scaling the activation of the penultimate layer $h(\mathbf{x})$ before computing the final classifier output. Given an input activation vector $h(\mathbf{x})$ and a pruning percentile p, SCALE [67] proceeds as follows to obtain the scaled activation $h^{\text{SCALE}}(\mathbf{x})$:

1. Compute the p-th percentile threshold t of $h(\mathbf{x})$.
2. Let $s_1 = \sum h(\mathbf{x})$, the sum of all activation values before pruning.

3. Construct a binary mask $\mathbf{1}_{\{h(\mathbf{x})\geq t\}}$ that keeps only the top-p activations.
4. Let $s_2 = \sum h(\mathbf{x}) \cdot \mathbf{1}_{\{h(\mathbf{x})\geq t\}}$, the sum of the top-$p$ activations.
5. Compute the scaling ratio $r = \frac{s_1}{s_2}$.
6. Scale the original activations by $\exp(r)$:

$$h^{\text{SCALE}}(\mathbf{x};\theta) = \exp(r) \cdot h(\mathbf{x};\theta).$$

The final model output is then computed with the scaled activations, and the *energy score* is used for OOD detection:

$$f^{\text{SCALE}}(\mathbf{x}) = W^\top h^{\text{SCALE}}(\mathbf{x}) + \mathbf{b}, \quad S_{\text{Energy}}(\mathbf{x}; f^{\text{SCALE}}) \in \mathbb{R}.$$

KNN [56]. This post-hoc, feature-space method identifies OOD samples based on their distance from ID training manifold. Let $\mathcal{H}_{\text{train}} = \{h(\mathbf{x}_i) \in \mathbb{R}^d\}_{i=1}^N$ be the set of N penultimate-layer feature vectors stored from the ID training set. For a new test input $\mathbf{x}$ with feature $h(\mathbf{x})$, the kNN score is computed in three steps:

1. Compute Distances: The set of Euclidean distances $\{d_i\}$ between $h(\mathbf{x})$ and all stored ID features in $\mathcal{H}_{\text{train}}$ is computed:

$$d_i = \|h(\mathbf{x}) - h(\mathbf{x}_i)\|_2, \quad \forall h(\mathbf{x}_i) \in \mathcal{H}_{\text{train}}$$

.
2. Identify Neighbors: The k smallest distances are identified and sorted, $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(k)}$.
3. Calculate Score: The final kNN score is the average distance to these k nearest neighbors:

$$S_{\text{kNN}}(\mathbf{x}) = \frac{1}{k}\sum_{j=1}^{k} d_{(j)}$$

A large score $S_{\text{kNN}}(\mathbf{x})$ indicates that the sample lies far from the ID training manifold and is therefore flagged as out-of-distribution.

B. Statistical Analysis

In this section, we present a detailed statistical analysis of our method, `Catalyst`, exhibiting how it enhances the separation between in-distribution (ID) and out-of-distribution (OOD) samples. This increased separation leads to a sharper decision boundary between ID and OOD regions. Our analysis builds on key observations commonly made in prior work on OOD detection [5, 36, 54, 55, 67].

B.1. Framework and Objective

In this section, we provide a statistical analysis demonstrating that our method, `Catalyst`, improves OOD detection by increasing the distributional separation between the expected scores of in-distribution (ID) and out-of-distribution (OOD) data.

Let $S(\mathbf{x})$ be the baseline OOD score and $\gamma(\mathbf{x})$ be our input-dependent scaling factor. We analyze two fusion strategies multiplicative scaling (i.e, elastic scaling) and additive shift as described in Equation 5 of Section 3:

$$S^*(\mathbf{x}) = \gamma(\mathbf{x})S(\mathbf{x})$$
$$S^+(\mathbf{x}) = \gamma(\mathbf{x}) + S(\mathbf{x})$$

Our objective is to formally show that the separability of the re-calibrated scores (Δ_{scaled} and Δ_{shift}) is greater than or equal to the separability of the original score (Δ_{original}). The separation Δ is defined as the difference between the expected score for in-distribution and out-of-distribution data:

$$\Delta_{\text{shift}} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S^+(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S^+(\mathbf{x})]$$
$$\Delta_{\text{scaled}} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S^*(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S^*(\mathbf{x})]$$
$$\Delta_{\text{original}} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S(\mathbf{x})]$$

B.2. Rationale and Assumptions

The rationale for OOD scoring functions [16, 36] is to map inputs to a scalar $S(\mathbf{x})$ that separates ID from OOD data. For clarity, we will follow the convention where ID samples yield higher scores and OOD samples yield lower scores- The success of any post-hoc method relies on this baseline separation as a necessary condition.

Building on this, `Catalyst` introduces a complementary scaling factor, $\gamma(\mathbf{x})$. For the fusion to be effective, $\gamma(\mathbf{x})$ must also be larger for a typical ID samples than OOD samples. This property is a necessary condition for success of `Catalyst`.

Assumption 1. *The expected value of the scaling factor for in-distribution data is greater than or equal to its expected value for out-of-distribution data:*

$$\bar{\gamma}_{in} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[\gamma(\mathbf{x})] \geq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[\gamma(\mathbf{x})] = \bar{\gamma}_{out}$$

In other word, by fusing these two "higher-is-ID" signals multiplicatively ($S^*(\mathbf{x}) = S(\mathbf{x}) \times \gamma(\mathbf{x})$), `Catalyst` uses differential amplification to actively widen the ID-OOD gap:

- For a typical ID sample, the high baseline score $S(\mathbf{x})$ is amplified by the high $\gamma(\mathbf{x})$, pushing it further into the ID region.
- For a typical OOD sample, the low baseline score $S(\mathbf{x})$ is suppressed by the low $\gamma(\mathbf{x})$, pushing it further into the OOD region.

The additive fusion, $S^+(\mathbf{x}) = S(\mathbf{x}) + \gamma(\mathbf{x})$, achieves a similar separation by applying a differential shift.

Additionally, to simplify the theoretical analysis, we introduce the following sufficient condition, which is empirically supported by our observations (Figure 2, 4). We note that this condition is primarily for theoretical tractability; our method is empirically robust and does not strictly

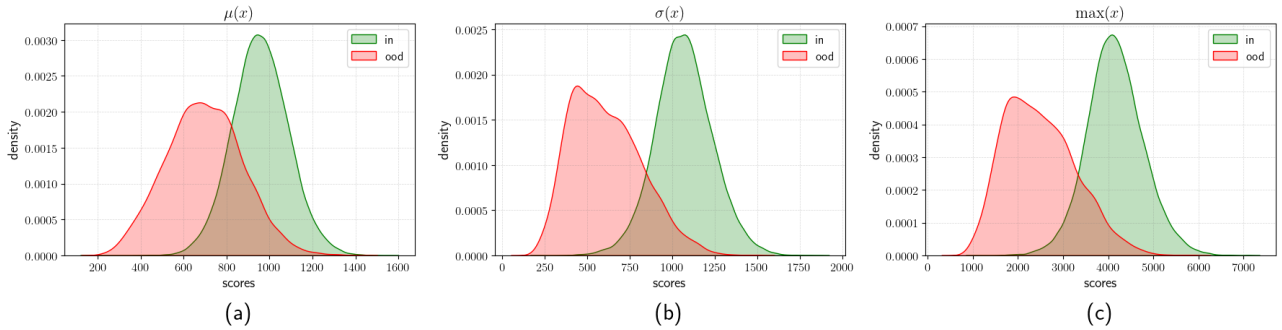

Figure 4. *Distribution of scaling factor γ from the penultimate layer of a ResNet-50 trained on ImageNet-1k, evaluated with Texture as the OOD dataset. The scales show clear separation between ID and OOD samples. Left to right: (a) $\mu(\mathbf{x})$: mean, (b) $\sigma(\mathbf{x})$: standard deviation, (c) $\max(\mathbf{x})$: max*

require this assumption to hold to achieve strong performance.

Assumption 2. *The mean scaling factor for ID data is larger than for OOD data, and both are bounded by one as illustrated in Figure 4. Formally, defining $\bar{\gamma}_{in} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{in}}[\gamma(\mathbf{x})]$ and $\bar{\gamma}_{out} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{out}}[\gamma(\mathbf{x})]$, we have*

$$\bar{\gamma}_{in} \geq \bar{\gamma}_{out} \geq 1 \tag{9}$$

Assumption 3. *The scaling factor $\gamma(\mathbf{x})$ and baseline score $S(\mathbf{x})$ are approximately uncorrelated. This is a simplifying assumption for the analysis that the covariance is negligible for both ID and OOD data.*

$$\mathrm{Cov}\big(\gamma(\mathbf{x}), S(\mathbf{x})\big) = 0 \tag{10}$$

B.3. `Catalyst`'s Improved Separation

In this section, we provide a formal characterization of how `Catalyst` widens the separability between the expected ID and OOD scores under both multiplicative (*) and additive ($^+$) fusion.

Theorem 1. *Under Assumptions 2 and 3, the distributional separation of the multiplicatively scaled score, $S^*(\mathbf{x})$, is at least as great as that of the original score, $S(\mathbf{x})$, i.e., $\Delta_{scaled} \geq \Delta_{original}$.*

Proof. By definition of Δ_{scaled}:

$$\begin{aligned}
\Delta_{\text{scaled}} &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[\gamma(\mathbf{x})S(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[\gamma(\mathbf{x})S(\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[\gamma(\mathbf{x})]\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[\gamma(\mathbf{x})]\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S(\mathbf{x})] \\
&= \bar{\gamma}_{\text{in}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S(\mathbf{x})] - \bar{\gamma}_{\text{out}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S(\mathbf{x})] \\
&\geq \bar{\gamma}_{\text{out}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S(\mathbf{x})] - \bar{\gamma}_{\text{out}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S(\mathbf{x})] \\
&= \bar{\gamma}_{\text{out}}\Big(\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S(\mathbf{x})]\Big) \\
&= \bar{\gamma}_{\text{out}}\Delta_{\text{original}}
\end{aligned}$$

$$\therefore \quad \Delta_{\text{scaled}} \geq \bar{\gamma}_{\text{out}}\Delta_{\text{original}}$$

$\because \bar{\gamma}_{\text{out}} \geq 1$, we conclude scaling increases the separation between typical ID and OOD samples. □

Theorem 2. *Under Assumption 1, the additive fusion score $S^+(\mathbf{x})$ increases or maintains the distributional separation compared to the baseline score, i.e., $\Delta_{shift} \geq \Delta_{original}$.*

Proof. By the definition of Δ_{shift} and the linearity of expectation:

$$\begin{aligned}
\Delta_{\text{shift}} &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S^+(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S^+(\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[\gamma(\mathbf{x}) + S(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[\gamma(\mathbf{x}) + S(\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[S(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[S(\mathbf{x})] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{in}}}[\gamma(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}[\gamma(\mathbf{x})] \\
&= \Delta_{\text{original}} + \Big(\bar{\gamma}_{\text{in}} - \bar{\gamma}_{\text{out}}\Big) \\
&\geq \Delta_{\text{original}} \qquad \Big(\because \bar{\gamma}_{\text{in}} - \bar{\gamma}_{\text{out}} \geq 0\Big)
\end{aligned}$$

$$\therefore \quad \Delta_{\text{shift}} \geq \Delta_{\text{original}}$$

We conclude shifting increases the separation between typical ID and OOD samples. □

C. Accuracy and Computational Overhead

Classification Accuracy. Our method, `Catalyst`, is designed to be post-hoc. The scaling factor γ is computed from penultimate pre-pooled activation map without altering the network's weights or its standard forward pass. Consequently, when used as a standalone method, `Catalyst` does not interfere with the model's inference process and maintains its original ID classification accuracy. We report the specific ID classification accuracy for all models used in our evaluation in Table 5.

Computational Overhead. The computational overhead introduced by `Catalyst` is negligible. The primary cost is computing one of channel-wise statistics (mean, std, max) from the $n\times k\times k$ pre-pooling map, followed by the clipping (Equation 3) and summation (Equation 4).

1. `Catalyst`(μ). This is the most efficient scenario. The mean statistic is simply the output of the GAP operation, which is already part of the standard forward pass. The only additional cost is the clipping

Dataset	Model	Accuracy
CIFAR-10	ResNet-18	93.89
	DenseNet-101	93.61
CIFAR-100	ResNet-18	75.20
	DenseNet-101	74.47
ImageNet	ResNet-34	73.31
	ResNet-50	76.13
	MobileNet-v2	71.88
	DenseNet-121	74.44

Table 5. *In-distribution classification accuracy (%) of the all the model used in evaluation of* `Catalyst`.

(Equation 3) and summation (Equation 4) of the resulting 2048-dimensional vector. This requires only 4,096 FLOP, an overhead of less than 0.0001% compared to the 5.42 GFLOPs of a ResNet-50.

2. `Catalyst`(σ) or `Catalyst`(m). These require computing a new statistic from the $n \times k \times k$ pre-pooling map. This is still negligible. For ResNet-50 (with a $2048 \times 7 \times 7$ map), computing the channel-wise maximum requires 0.1 MFLOPs, and the standard deviation requires 0.3 MFLOPs. In the worst-case scenario (standard deviation), the overhead is still less than 0.01% of the full forward pass. This confirms that `Catalyst` is lightweight and efficient post-hoc method.

D. Generalizability to Distance-Based Methods

A key question for our framework is its generalizability: is `Catalyst` merely an enhancement for logit-based methods, or is it a truly general-purpose framework? To answer this, we conducted a targeted study on its synergy with an entirely different family of OOD detectors: distance-based K-Nearest Neighbors (KNN) [56].

Setup. Our goal here is not to reproduce a specific, highly-optimized KNN baseline (which often rely on contrastive pre-training [11, 53, 56] to structure the feature space). Instead, our goal is to test a hypothesis: can `Catalyst` boost a generic KNN detector applied to a standard off-the-shelf pre-trained model?

To this end, we use the pre-trained models from our main experiments. Following the standard KNN OOD protocol [56], we build a Faiss [24] index of the (ID) training set's feature vectors. At inference, the baseline score $S_{\text{KNN}}(\mathbf{x})$ is the L_2 distance to the k-th nearest neighbor (we use $k = 50$, a standard value from prior work [11, 56]). A high distance indicates an OOD sample.

We integrate `Catalyst` by fusing scaling factor γ to elastically scale this distance score. As γ is high for ID (low distance) and low for OOD (high distance) samples, the signals are anti-correlated. We therefore use the fusion as shown in Equation 11. This pushes ID scores even lower and OOD scores even higher, widening the separation.

$$S'_{\text{KNN}}(\mathbf{x}) = S_{\text{KNN}}(\mathbf{x})/\gamma(\mathbf{x}) \tag{11}$$

Results. As shown in Table 7, `Catalyst` provides an consistent improvement over the standard KNN baseline across CIFAR and ImageNet benchmark. For instance, elastically scaling using scaling factor derived from max statistics `Catalyst`(m) we observed:

- On CIFAR-10, `Catalyst` reduces the FPR95 by 49.64% for ResNet-18 (from 31.02% to 15.62%) and by 36.54% for DenseNet-101 (from 13.08% to 8.30%).
- On CIFAR-100, `Catalyst` reduces the FPR95 by 43.84% for ResNet-18 (from 66.81% to 37.52%) and by 23.61% for DenseNet-101 (from 41.97% to 32.06%).

Similarly, in Table 6, we can see an consistent improvement across the model over standard KNN baselines across all tested models on ImageNet-1k. For instance, we observe `Catalyst`(μ) reduces the FPR95 by 52.64%, 52.13%, 38.08% and 41.16% for ResNet-34, ResNet-50, MobileNet-v2, and DenseNet-121 respectively.

Discussion. These performance boost demonstrate that `Catalyst` is a general-purpose framework. It successfully modulates a distance-based score on a standard cross-entropy trained model, proving its utility without requiring specialized training. The discriminative signal from scaling factor γ provides additional complementary information captured by both logit-based and distance-based methods, making it a powerful, "plug-and-play" enhancer for diverse OOD detection paradigms.

While this principle could be extended to other families, such as gradient-based methods like GradOrth [2], we note that integrating with such methods requires substantial, non-trivial engineering to reproduce their codebases and is beyond our current scope. We therefore leave this as a promising direction for future work. Finally, we note that our evaluation omits a direct comparison to Mahalanobis [30]. This follows the precedent set by recent works [5, 54, 55], which has shown it to be computationally expensive while offering limiting performance on these benchmarks.

Model	Method	SUN		Places		Texture		iNaturalist		Average	
		FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
ResNet-34	KNN	88.42	91.38	88.21	90.55	31.08	98.76	85.31	93.21	73.26	93.47
	+ Catalyst(μ)	**40.16**	**97.55**	**53.81**	**96.38**	11.97	99.57	**32.82**	**98.45**	**34.69**	**97.99**
	+ Catalyst(σ)	56.43	96.20	70.31	94.44	**7.77**	**99.76**	39.11	98.34	43.40	97.18
	+ Catalyst(m)	54.22	96.30	68.95	94.48	8.30	99.72	41.16	98.18	43.16	97.17
ResNet-50	KNN	79.08	94.43	82.21	93.31	16.29	99.43	78.61	95.05	64.05	95.56
	+ Catalyst(μ)	**35.94**	**98.12**	**50.41**	**97.06**	10.53	99.74	**27.54**	**98.92**	**31.11**	**98.46**
	+ Catalyst(σ)	51.05	97.08	66.25	95.58	**6.88**	**99.83**	35.23	98.67	39.85	97.79
	+ Catalyst(m)	49.86	97.09	65.58	95.56	7.23	99.81	35.74	98.65	39.60	97.78
MobileNet-v2	KNN	94.24	88.46	94.29	87.77	20.48	99.31	93.16	91.46	75.54	91.75
	+ Catalyst(μ)	**53.00**	**96.68**	**69.59**	**94.97**	12.48	99.61	**52.00**	97.15	**46.77**	**97.10**
	+ Catalyst(σ)	62.90	95.84	77.39	93.64	**8.62**	**99.77**	54.50	**97.17**	50.85	96.61
	+ Catalyst(m)	61.41	95.88	76.40	93.66	8.88	99.76	55.39	97.14	50.52	96.61
DenseNet-121	KNN	91.80	89.06	91.66	88.94	21.86	99.22	90.70	91.23	74.01	92.11
	+ Catalyst(μ)	**54.87**	**96.78**	**65.96**	**95.28**	16.45	99.57	**36.91**	**98.43**	**43.55**	**97.52**
	+ Catalyst(σ)	61.96	95.85	73.53	93.92	**14.11**	**99.64**	43.79	97.97	48.35	96.85
	+ Catalyst(m)	61.73	95.76	73.68	93.80	14.97	99.62	45.20	97.83	48.89	96.75

Table 6. *Detailed KNN-based OOD detection results for ImageNet benchmarks, using ResNet-34, ResNet-50, MobileNet-v2, and DenseNet-121. All values are percentages and are averaged over four common OOD benchmark datasets: SUN [66], Places [73], Texture [3] and iNaturalist [60]. The symbol ↓ indicates lower values are better; ↑ indicates larger values are better.*

Dataset	Model	Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
			FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR-10	ResNet-18	KNN	13.72	97.96	49.67	89.77	38.52	93.94	29.42	96.82	16.25	97.61	38.55	93.92	31.02	95.00
		+ Catalyst(μ)	9.28	98.51	52.18	90.16	31.49	95.81	25.23	97.67	4.63	99.20	30.42	95.76	25.54	96.18
		+ Catalyst(σ)	4.83	99.16	41.62	91.80	19.54	97.34	14.45	98.58	**2.45**	99.55	18.32	97.26	16.87	97.28
		+ Catalyst(m)	**4.54**	**99.20**	**39.88**	**92.09**	**17.24**	**97.62**	**13.39**	**98.69**	2.48	**99.56**	**16.20**	**97.54**	**15.62**	**97.45**
	DenseNet-101	KNN	1.51	99.67	41.83	90.26	6.89	98.95	14.36	98.59	5.59	98.95	8.28	98.63	13.08	97.51
		+ Catalyst(μ)	0.96	99.74	41.61	90.67	2.31	99.55	8.81	99.09	**0.65**	**99.84**	2.57	99.44	9.49	98.05
		+ Catalyst(σ)	**0.82**	99.79	38.70	91.14	**1.92**	99.58	6.33	99.31	0.93	99.77	2.30	99.47	8.50	98.18
		+ Catalyst(m)	0.84	**99.80**	**37.51**	**91.42**	**1.92**	**99.60**	**6.15**	**99.35**	1.14	99.75	**2.21**	**99.49**	**8.30**	**98.23**
CIFAR-100	ResNet-18	KNN	60.35	92.40	**86.34**	**71.63**	69.50	82.53	40.78	94.58	76.15	77.29	67.73	81.96	66.81	83.40
		+ Catalyst(μ)	34.89	95.51	90.25	69.69	65.40	86.66	40.18	94.72	21.31	95.50	64.59	85.78	52.77	87.98
		+ Catalyst(σ)	8.33	98.56	89.69	70.39	46.61	91.39	22.22	97.19	14.48	96.92	48.38	**90.35**	38.28	90.80
		+ Catalyst(m)	**7.67**	**98.62**	88.64	71.25	**45.87**	**91.42**	**21.92**	**97.23**	**12.95**	**97.24**	**48.05**	90.20	**37.52**	**90.99**
	DenseNet-101	KNN	15.24	97.22	**88.30**	67.69	43.04	90.41	26.79	96.35	35.63	88.68	42.84	89.42	41.97	88.29
		+ Catalyst(μ)	11.07	98.20	89.38	69.47	44.03	93.20	23.65	96.70	**3.00**	**99.39**	47.41	92.12	36.42	91.51
		+ Catalyst(σ)	8.79	**98.53**	89.12	70.06	36.09	94.74	17.41	97.63	5.96	98.98	39.14	93.88	32.75	92.30
		+ Catalyst(m)	**8.45**	**98.53**	88.46	**70.71**	**34.32**	**95.01**	**16.26**	**97.75**	7.34	98.76	**37.51**	**94.15**	**32.06**	**92.48**

Table 7. *Detailed KNN-based OOD detection results for the CIFAR-10 and CIFAR-100 benchmarks, using ResNet-18 and DenseNet-101. Results are evaluated against six common OOD datasets: SVHN [46], Places365 [73], iSUN [68], Textures [3], LSUN-crop [70], and LSUN-resize [70]. ↓ indicates lower values are better and ↑ indicates larger values are better.*

E. Detailed OOD Detection Performance

E.1. Near-OOD Evaluation

We also evaluate Catalyst on the challenging near-OOD task of distinguishing CIFAR-10 from CIFAR-100, a commonly used setup used in prior work [11]. As shown in Table 8, while the separation is inherently more difficult for all methods, Catalyst(m) still provides a marginal performance improvement over the baselines when applied in tandem, demonstrating its robustness even in fine-grained detection scenarios. For instance, with ResNet-18, Catalyst(m) + ReAct reduces the FPR95 from 52.04% to 49.62%, an improvement of 4.65%. We attribute the limited improvement in near-OOD settings to the high similarity of the learned penultimate representations. A valuable direction for future research is to design a suitable scaling factor γ in near-ood evaluation settings.

Method	FPR95 ↓	AUROC ↑
MSP	65.85	88.17
+ Catalyst(μ)	68.10	71.94
+ Catalyst(σ)	60.88	86.55
+ Catalyst(m)	60.07	86.28
Energy	52.32	90.14
+ Catalyst(μ)	52.94	89.82
+ Catalyst(σ)	50.93	90.47
+ Catalyst(m)	50.98	90.38
ReAct	52.04	90.42
+ Catalyst(μ)	52.63	89.68
+ Catalyst(σ)	50.08	90.47
+ Catalyst(m)	**49.62**	**90.52**
DICE	56.56	89.12
+ Catalyst(μ)	65.00	86.87
+ Catalyst(σ)	57.03	88.69
+ Catalyst(m)	56.89	88.79
ReAct+DICE	55.94	89.50
+ Catalyst(μ)	69.11	84.33
+ Catalyst(σ)	59.34	87.99
+ Catalyst(m)	58.09	87.99
ASH	57.14	87.60
+ Catalyst(μ)	64.15	84.00
+ Catalyst(σ)	56.86	87.38
+ Catalyst(m)	56.33	87.40
SCALE	55.58	88.60
+ Catalyst(μ)	60.96	85.47
+ Catalyst(σ)	54.93	88.15
+ Catalyst(m)	54.34	88.17

Table 8. *Near-OOD detection evaluation using ResNet-18. CIFAR-10 is ID dataset and CIFAR-100 is OOD dataset. The symbol ↓ indicates lower values are better; ↑ indicates higher values are better.*

E.2. ImageNet Evaluation.

Evaluation. Table 9 showcases detailed evaluation on ImageNet benchmark, using broad pre-trained model , ResNet-34, ResNet-50, MobileNet-v2, and DenseNet-121 for which we re-evaluated all baselines to ensure a fair comparison. Since results for ResNet-34 and DenseNet-121 were not available in the original publications of foundational baselines (e.g., ReAct, DICE, ASH, SCALE), we rigorously re-evaluated these methods ourselves. To ensure a fair and direct comparison, we carefully followed the hyperparameter selection protocols described in their respective papers (Appendix G).

Discussion. In the Table 9, Catalyst shows limited performance improvement on the SUN and Places datasets, particularly when using the MobileNet-v2 backbone. We empirically observed that this is due to a high degree of overlap between the distribution of the scaling factor, γ, for these datasets and for the in-distribution ImageNet-1k data. This overlap can be attributed to the high scene similarity between these datasets, a challenge previously identified by ViM [62].

To demonstrate this, the Figure 5 presents the distributions of γ for the SUN, Places365, Texture, and iNaturalist datasets, generated using the pre-trained MobileNet-v2 model. A clear pattern emerges: the distributions for the scene-based datasets (SUN, Places365) exhibit a significantly greater overlap with the in-distribution data compared to the more distinct Texture and iNaturalist datasets. This effect is particularly prominent when using the standard deviation $\sigma(\mathbf{x})$ and maximum value $\max(\mathbf{x})$ as information cues.

For brevity, we omit the γ distribution plots for the ResNet-34 and ResNet-50 backbones, but we confirm they exhibit the same general pattern. However, we also find that the overlap is more prominent for MobileNet-V2 than for ResNet-34, and in turn, more prominent for ResNet-34 than for ResNet-50.

E.3. CIFAR Evaluation

Evaluation. We present detailed performance results across six OOD test datasets for models: ResNet-18, and DenseNet-101 trained on CIFAR-10 and CIFAR-100, in Table 10 and Table 11, respectively.

Discussion. As shown in Table 11, Catalyst yields limited improvement on the Places365 dataset for models trained on CIFAR-100. We empirically attribute this to a high degree of overlap between the scaling factor γ distributions of the in-distribution (CIFAR-100) and Places365 samples, as shown in Figures 6 and 7. This pattern is consistent with our analysis on the ImageNet benchmark, where similar overlaps led to reduced performance. This case illustrates a key requirement for our method: its success hinges on a significant distributional separation of γ,

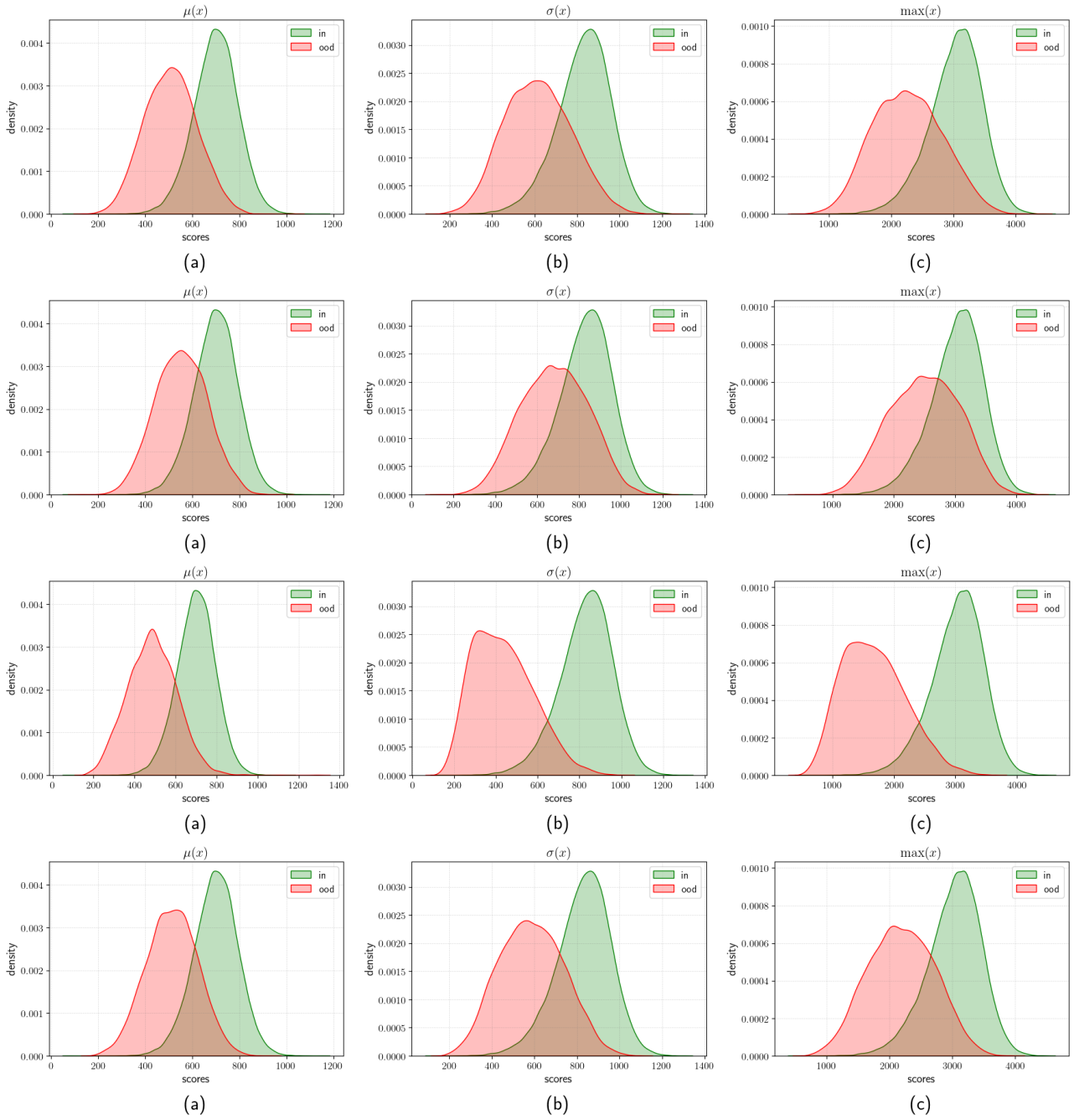

Figure 5. *Distributions of the scaling factor γ, derived from the penultimate layer of a MobileNet-V2 model trained on ImageNet-1k. The rows* (top to bottom) *correspond to the OOD datasets: SUN, Places365, Texture, and iNaturalist. The columns (left to right) correspond to the statistical cue used to compute γ: (a) mean: $\mu(\mathbf{x})$, (b) standard deviation: $\sigma(\mathbf{x})$, and (c) maximum value: $\max(\mathbf{x})$ (we used $\max(\mathbf{x})$ and $m(\mathbf{x})$ interchangeably). A clear pattern emerges: the distributions for the scene-based datasets (SUN, Places365) exhibit a significantly greater overlap with the in-distribution data compared to the more distinct Texture and iNaturalist datasets. This effect is particularly prominent when using the standard deviation $\sigma(\mathbf{x})$ and maximum value $\max(\mathbf{x})$ as information cues. We observe a similar pattern for ResNet-34 and ResNet-50 backbones. However, we also find that the overlap is more prominent for MobileNet-V2 than for ResNet-34, and in turn, more prominent for ResNet-34 than for ResNet-50.*

between ID/OOD data.

Model	Method	SUN		Places		Texture		iNaturalist		Average	
		FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
ResNet-34	MSP	72.39	79.81	73.76	79.20	69.98	79.12	59.24	86.61	68.84	81.19
	ODIN	59.34	86.13	64.62	84.14	51.95	87.45	47.69	90.91	55.90	87.16
	Energy	57.39	86.59	62.61	84.59	54.95	86.45	53.86	89.73	57.20	86.84
	ReAct	25.03	94.28	34.32	91.67	46.21	90.63	23.40	95.75	32.24	93.08
	DICE	38.03	90.20	48.40	87.18	34.72	90.24	35.34	92.22	39.12	89.96
	ReAct+DICE	22.33	94.79	32.85	91.84	30.39	93.21	19.42	96.13	26.25	93.99
	ASH	36.22	91.72	47.53	88.58	14.18	97.12	19.34	96.44	29.32	93.46
	SCALE	32.00	92.91	42.51	90.50	16.97	96.24	16.59	96.93	27.02	94.14
	Catalyst(μ)	33.46	91.85	43.78	89.39	24.86	93.39	25.60	94.99	31.92	92.41
	Catalyst(σ)	37.78	90.74	48.87	87.82	16.08	95.80	24.90	95.10	31.91	92.36
	Catalyst(m)	36.90	90.88	48.37	87.87	16.83	95.58	25.22	95.00	31.83	92.34
	Catalyst(μ) + **ReAct**	**21.44**	**95.18**	**31.74**	**92.56**	13.39	97.02	**12.81**	**97.47**	**19.84**	**95.56**
	Catalyst(σ) + **ReAct**	21.80	95.06	32.11	92.41	12.73	97.11	13.01	97.41	19.91	95.50
	Catalyst(m) + **ReAct**	22.03	94.99	32.58	92.31	**12.55**	**97.15**	13.47	97.33	20.16	95.44
ResNet-50	MSP	68.58	81.75	71.57	80.63	66.13	80.46	52.77	88.42	64.76	82.82
	ODIN	60.15	84.59	67.89	81.78	50.23	85.62	47.66	89.66	56.48	85.41
	Energy	58.28	86.73	65.40	84.13	52.29	86.73	53.95	90.59	57.48	87.05
	ReAct	23.68	94.44	33.33	91.96	46.33	90.30	19.73	96.37	30.77	93.27
	DICE	36.11	91.01	47.62	87.76	32.38	90.48	26.48	94.53	35.65	90.94
	ReAct+DICE	24.05	94.31	34.28	91.71	28.40	93.33	14.90	97.06	25.41	94.10
	ASH	28.01	94.02	39.84	90.98	11.95	97.60	11.52	97.87	22.83	95.12
	SCALE	25.78	94.54	36.86	91.96	14.56	96.75	10.37	98.02	21.89	95.32
	Catalyst(μ)	30.79	92.67	42.59	89.78	22.29	94.01	18.02	96.46	28.42	93.23
	Catalyst(σ)	35.73	91.47	48.35	88.04	15.85	95.94	19.05	96.21	29.75	92.92
	Catalyst(m)	35.79	91.40	48.68	87.82	16.08	95.88	19.00	96.18	29.89	92.82
	Catalyst(μ) + **ReAct**	**18.46**	**95.82**	**28.98**	**93.31**	12.11	97.38	**8.54**	**98.19**	**17.02**	**96.18**
	Catalyst(σ) + **ReAct**	19.13	95.61	29.58	93.04	**12.04**	**97.38**	9.10	98.06	17.46	96.02
	Catalyst(m) + **ReAct**	19.02	95.52	29.77	92.92	12.06	97.31	9.71	97.97	17.64	95.93
MobileNet-v2	MSP	74.20	78.88	76.89	78.14	70.99	78.95	59.86	86.72	70.49	80.67
	ODIN	54.07	85.88	57.36	84.71	49.96	85.03	55.39	87.62	54.20	85.81
	Energy	59.36	86.24	66.27	83.21	54.54	86.58	55.31	90.34	58.87	86.59
	ReAct	52.46	87.26	59.89	84.07	40.25	90.96	43.05	92.72	48.91	88.75
	DICE	37.84	90.81	52.35	86.17	32.57	91.46	41.53	91.30	41.07	89.94
	ReAct+DICE	30.60	92.98	45.93	88.29	16.03	96.33	31.68	93.76	31.06	92.84
	ASH	43.63	90.02	58.85	84.73	13.12	97.10	39.13	91.94	38.68	90.95
	SCALE	38.74	91.64	53.49	87.34	14.79	96.65	30.09	94.46	34.28	92.52
	Catalyst(μ)	37.74	91.43	52.21	87.33	23.42	94.17	33.47	93.84	36.71	91.69
	Catalyst(σ)	38.20	91.26	53.04	86.84	14.02	96.37	29.25	94.63	33.63	92.27
	Catalyst(m)	37.41	91.37	52.24	86.89	14.18	96.35	28.78	94.70	33.15	92.33
	Catalyst(μ) + **ReAct**	**32.82**	**92.93**	**48.62**	**88.59**	13.60	96.83	28.19	94.89	30.81	93.31
	Catalyst(σ) + **ReAct**	37.53	91.22	51.32	87.19	10.18	97.31	27.21	95.12	31.56	92.71
	Catalyst(m) + **ReAct**	34.77	92.26	49.77	88.06	**8.69**	**97.76**	**24.08**	**95.66**	**29.33**	**93.43**
DenseNet-121	MSP	67.49	81.41	69.53	80.95	67.23	79.18	49.58	89.05	63.46	82.65
	ODIN	54.13	86.33	60.39	84.14	50.82	85.81	32.47	93.66	49.45	87.48
	Energy	52.51	87.27	58.24	85.05	52.22	85.42	39.75	92.66	50.68	87.60
	ReAct	41.06	91.23	48.48	88.17	33.46	93.65	20.98	96.04	35.99	92.27
	DICE	38.75	89.91	49.29	86.24	40.85	88.09	25.78	94.37	38.67	89.65
	ReAct+DICE	31.36	92.99	43.91	89.11	24.38	95.14	17.68	96.44	29.33	93.42
	ASH	37.20	91.51	46.54	88.79	21.76	95.04	15.50	97.03	30.25	93.09
	SCALE	33.85	92.16	42.92	89.62	22.27	94.63	**13.21**	**97.40**	28.06	93.45
	Catalyst(μ)	33.24	91.84	42.94	89.01	25.59	93.29	16.41	96.69	29.54	92.71
	Catalyst(σ)	34.12	91.57	44.34	88.53	21.06	94.52	16.95	96.57	29.12	92.80
	Catalyst(m)	34.29	91.47	44.74	88.35	21.26	94.43	17.50	96.46	29.45	92.68
	Catalyst(μ) + **ReAct**	31.58	93.41	42.77	90.30	12.71	97.44	14.66	97.11	25.43	94.56
	Catalyst(σ) + **ReAct**	**30.04**	**93.44**	**41.50**	**90.24**	**11.37**	**97.61**	14.12	97.16	**24.26**	**94.61**
	Catalyst(m) + **ReAct**	30.25	93.37	41.61	90.13	11.74	97.52	14.48	97.09	24.52	94.53

Table 9. *Detailed OOD detection results on ImageNet benchmarks. All values are percentages and are averaged over four common OOD benchmark datasets: SUN [66], Places [73], Texture [3] and iNaturalist [60]. The symbol ↓ indicates lower values are better; ↑ indicates larger values are better.*

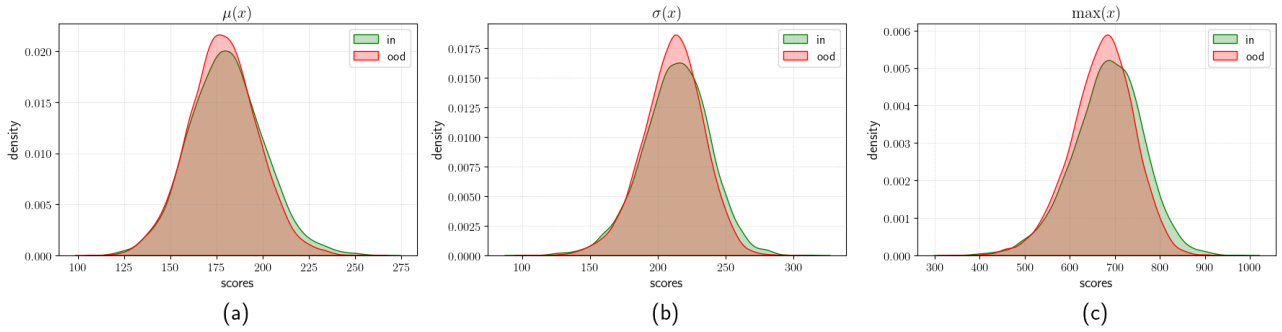

Figure 6. *Distribution of scaling factor γ from the penultimate layer of a ResNet-18 trained on CIFAR-100, evaluated with Places365 as the OOD dataset. The scales shows high overlap between ID and OOD samples.* Left to right*: (a) $\mu(\mathbf{x})$: mean, (b) $\sigma(\mathbf{x})$: standard deviation, (c) $\max(\mathbf{x})$: max*

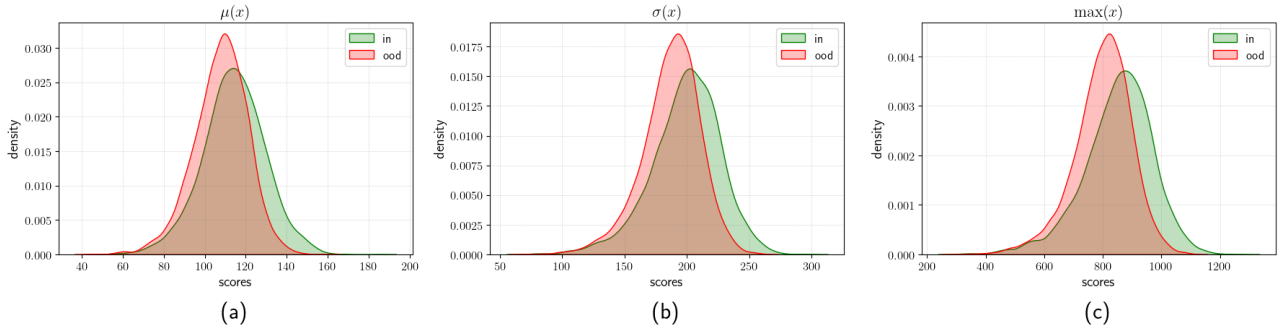

Figure 7. *Distribution of scaling factor γ from the penultimate layer of a DenseNet-101 trained on CIFAR-100, evaluated with Places365 as the OOD dataset. The scales shows high overlap between ID and OOD samples.* Left to right*: (a) $\mu(\mathbf{x})$: mean, (b) $\sigma(\mathbf{x})$: standard deviation, (c) $\max(\mathbf{x})$: max*

Model	Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet-18	MSP	60.39	92.40	63.69	88.37	56.74	91.32	62.66	90.10	51.87	93.64	54.63	91.87	58.33	91.28
	ODIN	35.96	94.70	41.11	92.06	23.36	96.56	46.74	91.97	6.66	98.71	20.04	96.93	28.98	95.16
	Energy	44.32	94.04	41.43	91.72	35.22	94.70	50.30	91.11	9.77	98.19	31.97	95.26	35.50	94.17
	ReAct	42.31	94.12	40.70	92.25	23.07	96.37	40.44	93.69	12.27	97.90	19.78	96.80	29.76	95.19
	DICE	17.60	97.09	46.14	90.66	39.08	94.32	44.65	91.80	1.90	99.57	36.52	94.70	30.98	94.69
	ReAct+DICE	11.05	98.07	47.53	91.14	17.19	97.04	24.33	95.91	1.56	99.66	16.24	97.19	19.65	96.50
	ASH	6.24	98.80	53.83	88.05	21.61	96.44	21.81	96.41	1.94	99.52	20.31	96.49	20.96	95.95
	SCALE	7.73	98.54	50.51	89.81	21.43	96.62	22.29	96.27	4.18	99.18	20.17	96.75	21.05	96.19
	Catalyst(μ)	15.73	97.32	43.65	91.25	26.26	96.08	35.98	94.25	3.20	99.26	24.26	96.30	24.85	95.74
	Catalyst(σ)	10.33	98.13	37.74	92.68	16.90	97.32	24.31	96.16	1.38	99.63	15.64	97.41	17.72	96.89
	Catalyst(m)	9.93	98.24	36.59	92.97	14.89	97.61	23.01	96.43	1.31	99.65	13.80	97.68	16.59	97.10
	Catalyst(μ) + **ReAct**	14.37	97.48	43.18	91.46	16.71	97.22	25.04	95.82	4.26	99.13	15.70	97.36	19.88	96.41
	Catalyst(σ) + **ReAct**	9.38	98.29	36.51	93.10	10.82	98.10	16.86	97.27	1.57	99.61	10.38	98.14	14.25	97.42
	Catalyst(m) + **ReAct**	8.86	98.39	35.04	93.38	9.08	98.32	15.64	97.48	1.52	99.63	9.00	98.35	13.19	97.59
DenseNet-101	MSP	64.76	88.33	60.30	88.55	33.57	95.41	56.67	90.17	23.41	96.75	33.87	95.37	45.43	92.43
	ODIN	33.09	94.41	36.68	92.34	3.22	99.20	38.49	91.61	1.84	99.53	2.89	99.28	19.37	96.06
	Energy	37.91	93.59	36.42	92.38	7.33	98.27	43.87	90.48	1.95	99.47	6.97	98.38	22.41	95.43
	ReAct	23.18	96.28	33.96	92.97	5.56	98.49	32.23	93.98	2.47	99.33	5.37	98.59	17.13	96.61
	DICE	16.66	96.98	37.59	92.04	2.31	99.42	27.98	92.71	0.15	99.94	2.44	99.36	14.52	96.74
	ReAct+DICE	4.60	99.02	35.94	92.91	1.78	99.51	17.07	96.78	0.12	99.95	2.02	99.47	10.26	97.94
	ASH	5.18	98.90	42.80	90.42	2.97	99.27	15.80	97.04	0.45	99.80	3.06	99.25	11.71	97.44
	SCALE	29.23	95.23	37.86	92.14	6.71	98.46	36.99	92.28	1.71	99.50	6.80	98.48	19.88	96.01
	Catalyst(μ)	15.12	97.51	33.93	92.75	3.32	98.98	25.71	94.88	0.61	99.79	3.70	98.92	13.73	97.14
	Catalyst(σ)	10.95	98.11	32.51	92.99	1.73	99.47	18.26	96.34	0.26	99.90	1.88	99.43	10.93	97.71
	Catalyst(m)	10.86	98.13	31.69	93.16	1.64	99.48	17.93	96.51	0.30	99.89	1.83	99.43	10.71	97.77
	Catalyst(μ) + **ReAct**	5.82	98.76	31.59	93.50	2.87	99.15	16.91	96.83	0.91	99.75	3.32	99.09	10.24	97.85
	Catalyst(σ) + **ReAct**	5.82	98.83	30.35	93.71	1.49	99.54	11.26	97.78	0.34	99.88	1.69	99.51	8.49	98.21
	Catalyst(m) + **ReAct**	5.86	98.86	29.97	93.89	1.49	99.55	11.06	97.88	0.48	99.87	1.68	99.51	8.42	98.26

Table 10. *Detailed results on six common OOD benchmark datasets: SVHN [46], Places365 [73], iSUN [68], Textures [3], LSUN-crop [70], LSUN-resize [70]. We used the same ResNet-18 and DenseNet-101 pre-trained on CIFAR-10.* ↓ *indicates lower values are better and* ↑ *indicates larger values are better.*

Model	Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet-18	MSP	74.26	83.20	82.37	75.31	84.13	71.57	85.04	74.02	70.79	82.78	82.96	73.10	79.92	76.66
	ODIN	70.30	88.06	80.14	77.02	60.26	86.98	81.56	76.56	47.73	91.84	56.35	88.23	66.06	84.78
	Energy	66.64	89.53	81.39	76.83	71.46	83.02	85.18	75.68	48.01	91.63	68.57	84.53	70.21	83.54
	ReAct	56.62	91.69	80.38	77.28	53.40	89.25	57.27	88.63	49.29	90.69	49.59	90.27	57.76	87.97
	DICE	40.89	92.97	81.33	76.23	62.61	85.83	75.28	76.29	12.44	97.65	61.39	86.84	55.66	85.97
	ReAct+DICE	34.16	94.18	83.57	74.79	54.50	89.85	52.96	87.36	10.40	97.95	53.78	90.22	48.23	89.06
	ASH	22.00	96.16	86.10	69.25	64.55	84.17	37.87	91.77	23.39	95.57	63.19	84.25	49.52	86.86
	SCALE	22.12	96.38	81.96	74.95	61.62	86.65	44.50	90.72	18.62	96.78	59.76	86.74	48.10	88.70
	Catalyst(μ)	31.13	95.02	81.53	76.00	64.83	85.24	62.06	85.32	16.45	97.17	61.59	86.00	52.93	87.46
	Catalyst(σ)	20.60	96.54	82.09	75.57	55.69	88.63	54.61	87.27	10.36	98.19	54.42	88.87	46.29	89.18
	Catalyst(m)	19.94	96.66	81.83	76.16	55.48	88.74	54.66	87.38	9.47	98.37	54.35	88.89	45.96	89.37
	Catalyst(μ) + **ReAct**	19.43	96.65	85.03	73.97	50.51	89.41	31.76	93.30	16.52	96.91	48.33	89.70	41.93	89.99
	Catalyst(σ) + **ReAct**	12.01	97.78	84.81	73.90	38.70	92.53	28.69	93.87	8.56	98.30	38.15	92.51	35.15	91.48
	Catalyst(m) + **ReAct**	11.47	97.85	83.96	74.67	38.04	92.72	28.58	93.96	7.65	98.46	38.23	92.55	34.66	91.70
DenseNet-101	MSP	81.38	75.71	82.68	74.06	82.52	70.50	87.11	68.39	51.82	87.93	79.31	72.21	77.47	74.80
	ODIN	85.94	80.35	75.59	77.62	48.03	89.12	83.37	67.83	12.78	97.70	40.28	91.35	57.67	84.00
	Energy	70.99	86.66	77.28	76.94	59.39	85.68	83.49	67.47	11.45	97.89	50.90	88.57	58.92	83.87
	ReAct	69.82	86.30	79.23	74.09	41.50	92.40	72.09	80.38	18.14	96.26	36.53	93.64	52.89	87.18
	DICE	32.93	94.09	79.90	75.43	35.50	92.50	64.84	71.95	1.93	99.57	30.81	93.96	40.98	87.92
	ReAct+DICE	25.10	95.70	84.17	73.56	27.98	95.06	41.79	87.82	1.06	99.70	27.76	95.16	34.64	91.17
	ASH	10.32	97.99	85.80	71.97	37.68	92.45	35.48	91.77	5.43	98.98	40.35	91.96	35.84	90.85
	SCALE	16.26	97.05	78.54	76.97	43.56	91.21	45.60	87.23	3.23	99.30	42.69	91.02	38.31	90.46
	Catalyst(μ)	22.45	96.11	78.72	77.16	48.77	89.75	52.09	83.58	1.54	99.68	44.92	90.44	41.42	89.45
	Catalyst(σ)	21.13	96.30	78.19	77.16	42.78	91.48	44.34	86.19	1.18	99.72	40.25	92.02	37.98	90.48
	Catalyst(m)	19.90	96.45	77.30	77.67	41.02	91.81	42.48	87.12	1.28	99.70	38.78	92.25	36.79	90.83
	Catalyst(μ) + **ReAct**	11.73	97.67	83.17	74.06	25.66	95.16	26.12	93.93	1.69	99.52	27.77	94.98	29.36	92.56
	Catalyst(σ) + **ReAct**	14.13	97.32	83.87	74.36	25.15	95.51	23.21	94.55	1.55	99.55	26.41	95.37	29.05	92.78
	Catalyst(m) + **ReAct**	13.70	97.36	83.00	75.14	23.26	95.83	21.68	94.95	2.07	99.44	24.63	95.64	28.06	93.06

Table 11. *Detailed results on six common OOD benchmark datasets: SVHN [46], Places365 [73], iSUN [68], Textures [3], LSUN-crop [70], LSUN-resize [70]. We used the same ResNet-18 and DenseNet-101 pre-trained on CIFAR-100.* ↓ *indicates lower values are better and* ↑ *indicates larger values are better.*

F. Comparison with Other Baselines

While in the main paper we restrict our comparison to foundational representative techniques (i.e., MSP, ODIN, Energy, ReAct, DICE, ASH and SCALE), we provide a comparison of our method, `Catalyst`, with additional baselines fDBD [34] and NCI [35] in this section. A comprehensive re-evaluation of fDBD and NCI across all architectures used in our study was determined to be beyond the scope of this work due to a fundamental difference in their design philosophy.

Methods like ReAct, DICE, ASH, SCALE and our own `Catalyst` are modular, post-hoc techniques that primarily modify the penultimate feature vector itself. In contrast, fDBD and NCI introduce entirely new scoring functions derived from the geometric relationship between features and the classifier's decision boundaries (fDBD) or class weight vectors (NCI). Integrating `Catalyst` into these structurally different frameworks would require significant, non-trivial engineering effort. Therefore, for these two methods, we present a comparison limited to the overlapping architectures and datasets from their original publications.

Additionally, we conduct a large-scale benchmark comparison on ImageNet-1k against plethora of existing literature using both ResNet-50 and MobileNet-v2. As shown in Table 12, we compare our method against 19 existing baselines for ResNet-50 [2, 5, 16, 18, 21, 30, 32, 34, 36, 49, 54–56, 62, 67, 74] and 15 baselines for MobileNet-v2 [2, 5, 16, 21, 30, 32, 36, 49, 54, 55, 62, 67, 74], with all competitors' results taken directly from their original publications. This comprehensive evaluation demonstrates that `Catalyst` achieves competitive and consistent performance compared to all prior post-hoc methods on this challenging benchmark.

F.1. Neural Collapse Inspired (NCI) OOD Detector

As shown in Table 13 for CIFAR-10 and Table 14 for ImageNet, in a direct comparison against NCI's [35] reported results, our method `Catalyst` demonstrates a clear and significant advantage. On CIFAR-10 with a ResNet-18 backbone, $\texttt{Catalyst}(m) + \texttt{ReAct}$ decisively outperforms NCI, reducing the average FPR95 by 33.43%. This strong performance is maintained on the large-scale ImageNet benchmark, where our method reduces the FPR95 by 42.83% on ResNet-50.

F.2. Fast Decision Boundary based OOD Detector

As shown in Tables 15 and 16, our method, `Catalyst`, demonstrates a decisive and substantial performance advantage over the fDBD's reported results in all comparable, overlapping settings. The strength of `Catalyst` is most apparent when it is composed with existing techniques, creating a powerful synergistic effect that dramatically improves OOD detection. On CIFAR-10, this combination is particularly effective. Using a ResNet-18, $\texttt{Catalyst}(m) + \texttt{ReAct}$ slashes the average FPR95 by 44.80% relative to fDBD (from 31.09% down to 17.16%). The gains are even more pronounced on a DenseNet-101, where $\texttt{Catalyst}(m) + \texttt{ReAct}$ achieves an FPR95 reduction of 34.67%.

As demonstrated in Table 16, this commanding performance extends to the large-scale ImageNet benchmark. While fDBD struggles with a high average FPR95 of 51.19%, our $\texttt{Catalyst}(m) + \texttt{ReAct}$ achieves an FPR95 of just 17.64% – a massive 65.54% relative reduction. These results validate `Catalyst` performed significantly better than recent baseliens like NCI and fDBD.

F.3. AdaSCALE OOD Detection

AdaSCALE is a post-hoc OOD detection method that replaces fixed activation-scaling strategies with an adaptive, sample-dependent mechanism. Existing approaches (ASH, SCALE, LTS) prune activations using a static percentile threshold, which cannot reliably distinguish ID from OOD data. AdaSCALE leverages the observation that OOD samples experience larger shifts in their top activated neurons under small pixel perturbations, while ID activations remain stable. It measures this activation shift (Q), adjusts it with a correction term (Co), and maps the resulting OODness score through a CDF to produce a dynamic pruning percentile. This causes ID samples to receive stronger scaling and OOD samples weaker scaling, yielding more separated energy scores and improved detection performance.

We compare `Catalyst` against AdaScale in Tables 17 and 18, strictly following the restricted dataset protocol from the original AdaScale paper for a fair comparison. On the CIFAR (DenseNet-101) benchmark, `Catalyst` consistently outperforms AdaScale. For instane, on CIFAR-10, the best baseline AdaScale-L achieves an average FPR95 of 40.03%. Our standalone $\texttt{Catalyst}(m)$ is already significantly better at 20.16%, and our combined $\texttt{Catalyst}(m) + \texttt{ReAct}$ further extends this lead to 15.63%. On CIFAR-100, our $\texttt{Catalyst}(m) + \texttt{ReAct}$ (FPR95 39.46%) likewise outperforms the best AdaScale-A baseline (58.42%). This trend holds on the ImageNet (ResNet-50) benchmark. As shown in Table 18, our $\texttt{Catalyst}(\mu) + \texttt{ReAct}$ (FPR95 16.54%) achieves similar level of performance compared to best AdaScale-L (16.92%).

G. Reproducibility Statement

We are committed to ensuring the reproducibility of our research. To this end, we provide detailed information regarding our code, experimental setup, hyperparameter selection, and computational environment.

Model	Method	SUN		Places		Texture		iNaturalist		Average	
		FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
ResNet-50	MSP* [16]	68.58	81.75	71.57	80.63	66.13	80.46	52.77	88.42	64.76	82.82
	ODIN* [32]	60.15	84.59	67.89	81.78	50.23	85.62	47.66	89.66	56.48	85.41
	GODIN [18]	60.83	85.60	63.70	83.81	77.85	73.27	61.91	85.40	66.07	82.02
	Mahalanobis [30]	68.36	84.35	73.32	81.46	16.05	94.96	39.90	93.76	49.41	88.63
	KNN ($\alpha = 100\%$) [56]	68.82	80.72	76.28	75.76	11.77	97.07	59.00	86.47	53.97	85.01
	KNN ($\alpha = 1\%$) [56]	69.53	80.10	77.09	74.87	11.56	97.18	59.08	86.20	54.32	84.59
	GradOrth [2]	19.61	95.76	33.67	91.78	**11.19**	**98.06**	11.04	98.00	18.57	96.31
	GradNorm [21]	42.81	87.26	55.62	81.85	38.15	87.73	23.73	93.97	40.08	87.70
	NN-Guide [49]	31.62	91.66	38.88	90.12	24.93	91.52	12.02	97.47	26.86	92.69
	ViM [62]	43.10	89.39	52.86	86.61	17.18	93.58	20.34	96.24	33.37	91.45
	fDBD [34]	60.60	86.97	66.40	84.27	37.50	92.12	40.24	93.67	51.19	89.26
	BATS [74]	22.62	95.33	34.34	91.83	38.90	92.27	12.57	97.67	27.11	94.20
	LAPS [14]	**15.81**	**96.18**	**24.71**	**93.64**	41.49	91.81	12.72	97.50	23.68	94.78
	Energy* [36]	58.28	86.73	65.40	84.13	52.29	86.73	53.95	90.59	57.48	87.05
	ReAct* [55]	23.68	94.44	33.33	91.96	46.33	90.30	19.73	96.37	30.77	93.27
	DICE* [54]	36.11	91.01	47.62	87.76	32.38	90.48	26.48	94.53	35.65	90.94
	ReAct+DICE* [54, 55]	24.05	94.31	34.28	91.71	28.40	93.33	14.90	97.06	25.41	94.10
	ASH* [5]	28.01	94.02	39.84	90.98	11.95	97.60	11.52	97.87	22.83	95.12
	SCALE* [67]	25.78	94.54	36.86	91.96	14.56	96.75	10.37	98.02	21.89	95.32
	Catalyst(μ)	30.79	92.67	42.59	89.78	22.29	94.01	18.02	96.46	28.42	93.23
	Catalyst(σ)	35.73	91.47	48.35	88.04	15.85	95.94	19.05	96.21	29.75	92.92
	Catalyst(m)	35.79	91.40	48.68	87.82	16.08	95.88	19.00	96.18	29.89	92.82
	Catalyst(μ) + **ReAct**	18.46	95.82	28.98	93.31	12.11	97.38	**8.54**	**98.19**	**17.02**	**96.18**
	Catalyst(σ) + **ReAct**	19.13	95.61	29.58	93.04	12.04	97.38	9.10	98.06	17.46	96.02
	Catalyst(m) + **ReAct**	19.02	95.52	29.77	92.92	12.06	97.31	9.71	97.97	17.64	95.93
MobileNet-v2	MSP* [16]	74.20	78.88	76.89	78.14	70.99	78.95	59.86	86.72	70.49	80.67
	ODIN* [32]	54.07	85.88	57.36	84.71	49.96	85.03	55.39	87.62	54.20	85.81
	Mahalanobis [30]	54.79	86.33	53.77	83.69	88.72	37.28	62.04	82.37	64.83	72.40
	GradOrth [2]	30.82	**93.18**	40.27	89.12	12.69	97.52	26.81	93.17	27.65	93.25
	GradNorm [21]	42.15	89.65	56.56	83.93	34.95	90.99	33.70	92.46	41.84	89.20
	NN-Guide [49]	79.57	76.10	81.87	74.23	38.78	89.32	68.24	82.07	67.12	80.43
	ViM [62]	88.67	66.37	92.16	62.43	40.71	89.59	86.86	69.57	77.10	71.99
	BATS [74]	41.68	90.21	52.43	86.26	38.69	90.76	31.56	94.33	41.09	90.39
	LAPS [14]	**30.07**	92.98	**39.70**	**90.10**	51.37	88.29	**18.82**	**96.76**	34.99	92.03
	Energy* [36]	59.36	86.24	66.27	83.21	54.54	86.58	55.31	90.34	58.87	86.59
	ReAct* [55]	52.46	87.26	59.89	84.07	40.25	90.96	43.05	92.72	48.91	88.75
	DICE* [54]	37.84	90.81	52.35	86.17	32.57	91.46	41.53	91.30	41.07	89.94
	ReAct+DICE* [54, 55]	30.60	92.98	45.93	88.29	16.03	96.33	31.68	93.76	31.06	92.84
	ASH* [5]	43.63	90.02	58.85	84.73	13.12	97.10	39.13	91.94	38.68	90.95
	SCALE* [67]	38.74	91.64	53.49	87.34	14.79	96.65	30.09	94.46	34.28	92.52
	Catalyst(μ)	37.74	91.43	52.21	87.33	23.42	94.17	33.47	93.84	36.71	91.69
	Catalyst(σ)	38.20	91.26	53.04	86.84	14.02	96.37	29.25	94.63	33.63	92.27
	Catalyst(m)	37.41	91.37	52.24	86.89	14.18	96.35	28.78	94.70	33.15	92.33
	Catalyst(μ) + **ReAct**	32.82	92.93	48.62	88.59	13.60	96.83	28.19	94.89	30.81	93.31
	Catalyst(σ) + **ReAct**	37.53	91.22	51.32	87.19	10.18	97.31	27.21	95.12	31.56	92.71
	Catalyst(m) + **ReAct**	34.77	92.26	49.77	88.06	**8.69**	**97.76**	24.08	95.66	**29.33**	**93.43**

Table 12. *Detailed Comparison with existing OOD detection methods on the ImageNet-1k benchmark, using ResNet-50 and MobileNet-v2. Methods marked with * were reproduced by us; results for all other methods are taken from their original publications. The symbol ↓ indicates lower values are better; ↑ indicates larger values are better.*

G.1. Code and Data Availability

The complete source code used in `Catalyst`, along with the scripts used to run all experiments and generate figures, is publicly available on GitHub.[2] We will also provide the model weights for our trained CIFAR models. All datasets used in this work (CIFAR-10, CIFAR-100, ImageNet-1k, and all OOD benchmarks) are publicly available and were used without modification, following the standard preprocessing steps described in their original publications and common benchmarks.

Experimental Setup.

- **CIFAR Benchmarks:** Our primary models include ResNet-18 and DenseNet-101. Following established protocols [5, 43, 54, 55, 67], all models were trained from scratch for 100 epochs using SGD with a momentum of 0.9, a weight decay of 0.0001, and a batch size of 64. The learning rate was initialized at 0.1 and decayed by a factor of 10 at epochs 50, 75, and 90.
- **ImageNet Benchmark:** For our large-scale experiments, we used the official pre-trained models provided by PyTorch for ResNet-34, ResNet-50, MobileNet-v2, and DenseNet-121. No fine-tuning was performed.

[2] https://github.com/bingabid/Catalyst

Method	SVHN		Places365		Texture		Average	
	FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
`NCI`	28.92	90.81	34.01	90.74	26.53	92.18	29.82	91.24
`Catalyst`(μ)	15.73	97.32	43.65	91.25	35.98	94.25	31.79	94.27
`Catalyst`(σ)	10.33	98.13	37.74	92.68	24.31	96.16	24.13	95.66
`Catalyst`(m)	9.93	98.24	36.59	92.97	23.01	96.43	23.18	95.88
`Catalyst`(μ) + `ReAct`	14.37	97.48	43.18	91.46	25.04	95.82	27.53	94.92
`Catalyst`(σ) + `ReAct`	9.38	98.29	36.51	93.10	16.86	97.27	20.92	96.22
`Catalyst`(m) + `ReAct`	**8.86**	**98.39**	**35.04**	**93.38**	**15.64**	**97.48**	**19.85**	**96.42**

Table 13. *A direct comparison of* `Catalyst` *against the NCI baseline, using their originally reported results for CIFAR-10 with a ResNet-18 backbone. The evaluation is restricted to the SVHN, Texture, and Places365 OOD datasets to ensure a fair comparison that matches the protocol from the original NCI paper.*

Method	Texture		iNaturalist		Average	
	FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
`NCI`	23.79	96.63	14.31	96.95	19.05	96.79
`Catalyst`(μ)	22.29	94.01	18.02	96.46	20.16	95.24
`Catalyst`(σ)	15.85	95.94	19.05	96.21	17.45	96.08
`Catalyst`(m)	16.08	95.88	19.00	96.18	17.54	96.03
`Catalyst`(μ) + `ReAct`	12.11	**97.38**	**8.54**	**98.19**	**10.33**	**97.79**
`Catalyst`(σ) + `ReAct`	**12.04**	**97.38**	9.10	98.06	10.57	97.72
`Catalyst`(m) + `ReAct`	12.06	97.31	9.71	97.97	10.89	97.64

Table 14. *A direct comparison of* `Catalyst` *against the NCI baseline, using their originally reported results for ImageNet with a ResNet-50 backbone. The evaluation is restricted to the iNaturalist and Texture OOD datasets to ensure a fair comparison that matches the protocol from the original NCI paper.*

G.2. 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 do not tune $\boldsymbol{c}$ directly; instead, we control it by setting it to the p-th percentile of the ID activation distribution (e.g, when $p = 95$, it indicates that 95% of the ID activations are less than the threshold $\boldsymbol{c}$). The choice of this percentile p is the key hyperparameter to be tuned. To select the optimal p, we follow established protocols from prior work [54, 55] and create a proxy OOD validation set, which is generated by adding pixel-wise Gaussian noise $\mathcal{N}(0, 0.2)$ 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 – using a percentile for the mechanism and a proxy set for tuning – is a robust tuning strategy grounded in prior work. The selected p values are:

- **CIFAR:** We found it optimal to tune the percentile for each statistic individually. These values are fixed for all CIFAR models (ResNet-18, DenseNet-101) and across all baselines (e.g., Energy, ReAct). The selected percentiles are $p_{\text{mean}} = 60$, $p_{\text{std}} = 95$, and $p_{\text{max}} = 95$.
- **ImageNet:** For our default method (`Catalyst` + Energy), a single percentile of $p = 75$ is used for all ImageNet models. When combining with baselines like ReAct (`Catalyst` + `ReAct`), we found it optimal to use a single, shared percentile p across all three statistics (mean, std, and max). The optimal shared percentile p varies by model: $p = 15$ for ResNet-34 and ResNet-50, $p = 35$ for MobileNet-v2, and $p = 52$ for DenseNet-121.

As our hyperparameter search for ImageNet demonstrated, the optimal shared percentile p varies across different architectures (e.g., $p = 15$ for ResNet-50 vs. $p = 52$ for DenseNet-121). This empirical finding is highly intuitive and aligns with our method's design. The optimal p (which sets the clipping threshold c) is naturally coupled with the model's architecture, particularly the dimension (n) of the penultimate layer. This is because our scaling factor γ (Equation 4) is an aggregation (a sum) over all n channels. A model with a larger channel dimension (e.g., ResNet-50, $n = 2048$) will produce a sum of a very different magnitude than a model with a smaller dimension (e.g., DenseNet-121, $n = 1024$). Therefore, a different clipping percentile p is required for each architecture to produce the most discriminative γ signal.

G.3. Baseline Hyperparameter Tuning

A core principle of our evaluation is to ensure a fair and rigorous comparison against all baselines. For all methods (ODIN, ReAct, DICE, ASH, SCALE, KNN), we strictly

Model	Method	SVHN		Places365		iSUN		Texture		Average	
		FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
ResNet-18	`fDBD`	22.58	96.07	46.59	90.40	23.96	95.85	31.24	94.48	31.09	94.20
	`Catalyst`(μ)	15.73	97.32	43.65	91.25	26.26	96.08	35.98	94.25	30.41	94.73
	`Catalyst`(σ)	10.33	98.13	37.74	92.68	16.90	97.32	24.31	96.16	22.32	96.07
	`Catalyst`(m)	9.93	98.24	36.59	92.97	14.89	97.61	23.01	96.43	21.11	96.31
	`Catalyst`(μ) + `ReAct`	14.37	97.48	43.18	91.46	16.71	97.22	25.04	95.82	24.83	95.99
	`Catalyst`(σ) + `ReAct`	9.38	98.29	36.51	93.10	10.82	98.10	16.86	97.27	18.89	96.69
	`Catalyst`(m) + `ReAct`	**8.86**	**98.39**	**35.04**	**93.38**	**9.08**	**98.32**	**15.64**	**97.48**	**17.16**	**96.89**
DenseNet-101	`fDBD`	5.89	98.67	39.52	91.53	5.90	98.75	22.75	95.81	18.52	96.19
	`Catalyst`(μ)	15.12	97.51	33.93	92.75	3.32	98.98	25.71	94.88	19.52	96.53
	`Catalyst`(σ)	10.95	98.11	32.51	92.99	1.73	99.47	18.26	96.34	15.86	96.73
	`Catalyst`(m)	10.86	98.13	31.69	93.16	1.64	99.48	17.93	96.51	15.53	96.82
	`Catalyst`(μ) + `ReAct`	5.82	98.76	31.59	93.50	2.87	99.15	16.91	96.83	14.30	**97.56**
	`Catalyst`(σ) + `ReAct`	**5.82**	98.83	30.35	93.71	**1.49**	99.54	11.26	97.78	12.23	97.47
	`Catalyst`(m) + `ReAct`	5.86	**98.86**	**29.97**	**93.89**	**1.49**	**99.55**	**11.06**	**97.88**	**12.10**	97.55

Table 15. *Direct comparison of* `Catalyst` *against the fDBD baseline on CIFAR-10. To ensure a fair comparison, the evaluation is restricted to the four OOD datasets reported in the original fDBD paper: SVHN, Places365, iSUN, and Texture.*

Method	SUN		Places365		Texture		iNaturalist		Average	
	FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
`fDBD`	60.60	86.97	66.40	84.27	37.50	92.12	40.24	93.67	51.19	89.26
`Catalyst`(μ)	30.79	92.67	42.59	89.78	22.29	94.01	18.02	96.46	28.42	93.23
`Catalyst`(σ)	35.73	91.47	48.35	88.04	15.85	95.94	19.05	96.21	29.75	92.92
`Catalyst`(m)	35.79	91.40	48.68	87.82	16.08	95.88	19.00	96.18	29.89	92.82
`Catalyst`(μ) + `ReAct`	**18.46**	**95.82**	**28.98**	**93.31**	12.11	**97.38**	**8.54**	**98.19**	**17.02**	**96.18**
`Catalyst`(σ) + `ReAct`	19.13	95.61	29.58	93.04	**12.04**	**97.38**	9.10	98.06	17.46	96.02
`Catalyst`(m) + `ReAct`	19.02	95.52	29.77	92.92	12.06	97.31	9.71	97.97	17.64	95.93

Table 16. *This table presents a direct comparison of* `Catalyst` *against the fDBD baseline on ImageNet using a ResNet-50 backbone. To ensure a fair comparison, the evaluation is restricted to the iNaturalist and Texture OOD datasets, matching the protocol in the original fDBD paper.*

followed the hyperparameter selection protocols described in their respective papers.

When re-evaluating these baselines on new architectures not present in their original work, we performed a new hyperparameter search using the same validation procedures and search spaces they described. This ensures that every baseline is as strong as possible for each specific model. Key hyperparameters for these methods are summarized below:

- **ODIN:** We adopted the optimal hyperparameter values reported in the original publication. Accordingly, we set the temperature to $T = 1000$, with a noise magnitude ϵ of 0.004 for CIFAR and 0.0015 for ImageNet.
- **ReAct:** The clipping percentile p was selected from $\{85, 90, 95\}$. While we found $p = 90$ to be optimal for the standalone ReAct baseline, consistent with the original paper, the optimal value shifted to $p = 95$ when ReAct was combined with our `Catalyst`.
- **DICE:** We selected the sparsity ratio p from $\{70, 75, 80, 85, 90, 95\}$. Our validation process consistently identified $p = 70\%$ as the optimal value.
- **ASH:** The pruning percentile p was selected from $\{80, 85, 90\}$. The optimal value was found to be dependent on the dataset and architecture. We report the specific optimal value for each major setting to ensure the strongest and fairest possible comparison.
  - For **ImageNet**, the optimal value was consistently $p = 90$ for most architectures, with the exception of EfficientNet-b0, which required a less aggressive pruning of $p = 50$.
  - For **CIFAR-10**, the optimal values were $p = 80$ for both ResNet models, $p = 90$ for DenseNet, and $p = 70$ for MobileNet-v2. These values held for both the standalone baseline and when combined with `Catalyst`.
  - For **CIFAR-100**, the optimal value for the ResNet models was consistently $p = 80$. For other architectures, we observed an interaction effect: the optimal percentile for DenseNet shifted from $p = 90$ (baseline) to $p = 80$ (with `Catalyst`), and for MobileNet-v2, it shifted from $p = 90$ to $p = 85$.
- **SCALE:** For the SCALE baseline, the pruning percentile p was set to a fixed value of $p = 85$ across all experiments. We adopted this value directly from the original SCALE paper [67] to ensure our re-implementation was consistent with the authors' reported optimal setting, providing a fair comparison.

Dataset	Method	SVHN		Places365		Texture		Average	
		FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
CIFAR-10	`AdaScale-L`	25.04	94.05	36.77	91.20	58.28	87.35	40.03	90.87
	`AdaScale-A`	26.43	93.87	37.03	91.25	58.59	87.19	40.68	90.77
	`Catalyst`(μ)	15.12	97.51	33.93	92.75	25.71	94.88	24.92	95.05
	`Catalyst`(σ)	10.95	98.11	32.51	92.99	18.26	96.34	20.57	95.15
	`Catalyst`(m)	10.86	98.13	31.69	93.16	17.93	96.51	20.16	95.27
	`Catalyst`(μ) + `ReAct`	**5.82**	98.76	31.59	93.50	16.91	96.83	18.77	96.36
	`Catalyst`(σ) + `ReAct`	**5.82**	98.83	30.35	93.71	11.26	97.78	15.81	96.77
	`Catalyst`(m) + `ReAct`	5.86	**98.86**	**29.97**	**93.89**	**11.06**	**97.88**	**15.63**	**96.88**
CIFAR-100	`AdaScale-L`	46.29	84.31	**61.70**	**78.86**	71.40	76.59	59.13	79.25
	`AdaScale-A`	43.97	85.30	61.97	78.69	69.31	77.71	58.42	80.57
	`Catalyst`(μ)	22.45	96.11	78.72	77.16	52.09	83.58	51.09	85.62
	`Catalyst`(σ)	21.13	96.30	78.19	77.16	44.34	86.19	47.89	86.55
	`Catalyst`(m)	19.90	96.45	77.30	77.67	42.48	87.12	46.56	87.08
	`Catalyst`(μ) + `ReAct`	**11.73**	**97.67**	83.17	74.06	26.12	93.93	40.34	88.55
	`Catalyst`(σ) + `ReAct`	14.13	97.32	83.87	74.36	23.21	94.55	40.40	88.74
	`Catalyst`(m) + `ReAct`	13.70	97.36	83.00	75.14	**21.68**	**94.95**	**39.46**	**89.15**

Table 17. *A direct comparison of* `Catalyst` *against the AdaScale baseline, using their originally reported results for CIFAR with a DenseNet-101 backbone. The evaluation is restricted to the SVHN, Texture, and Places365 OOD datasets to ensure a fair comparison that matches the protocol from the original AdaScale paper [50].*

Method	Places		Texture		iNaturalist		Average	
	FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
`AdaScale-L`	32.60	92.74	10.57	97.88	7.61	98.31	16.92	96.98
`AdaScale-A`	32.97	92.63	**10.33**	**97.92**	**7.78**	**98.29**	17.03	96.95
`Catalyst`(μ)	42.59	89.78	22.29	94.01	18.02	96.46	27.63	93.42
`Catalyst`(σ)	48.35	88.04	15.85	95.94	19.05	96.21	27.08	93.40
`Catalyst`(m)	48.68	87.82	16.08	95.88	19.00	96.18	27.25	93.29
`Catalyst`(μ) + `ReAct`	**28.98**	**93.31**	12.11	97.38	8.54	98.19	**16.54**	**96.96**
`Catalyst`(σ) + `ReAct`	29.58	93.04	12.04	97.38	9.10	98.06	16.91	96.83
`Catalyst`(m) + `ReAct`	29.77	92.92	12.06	97.31	9.71	97.97	17.18	96.73

Table 18. *A direct comparison of* `Catalyst` *against the* `AdaScale` *baseline, using their originally reported results for ImageNet with a ResNet-50 backbone. The evaluation is restricted to the Places, Texture, and iNaturalist OOD datasets to ensure a fair comparison that matches the protocol from the original AdaScale paper [50].*

G.4. Computational Environment

All CIFAR model training and OOD detection experiments were conducted on an Apple M2 Max system with 96 GB of RAM. The experiments were implemented in Python using PyTorch (v2.1) and the Torchvision library.

H. Choice of Layer for Computing γ

A necessary condition for Catalyst to be effective is that its scaling factor (γ) must be inherently distinguishable between ID and OOD samples. Consequently, a core methodological decision is to identify which network stage produces the most discriminative information cues. To justify our focus on the penultimate layer, we conducted an analysis to locate the most potent source of information cues within the network. We used a ResNet-50 model trained on ImageNet (ID) and computed γ using the channel-wise average activation from each of its four main residual stages (Layer 1-4). We then compared the ID γ distribution against multiple OOD datasets, including Places, SUN, Texture, and iNaturalist.

A clear and consistent trend emerged, as illustrated representatively in Figure 8 for the Texture dataset. The analysis reveals that the γ distributions from the early-to-mid stages (Layer 1-3) are not sufficiently discriminative. As shown in the Figure 8, they exhibit high overlap overlap between ID (ImageNet) and OOD (Texture) samples, rendering them ineffective for our scaling purposes. In sharp contrast, the signal from the final residual stage (Layer 4) demonstrates a sufficiently clear between the two distributions. This finding was consistent across all tested OOD datasets.

This analysis empirically validates our methodological focus. The penultimate layer's pre-pooling feature map is not just a layer of convenience; it is the most reliable source of a potent signal for constructing γ. While we only illustrate this with ResNet-50 for brevity, we observed this same trend with the final feature map consistently providing the most signal separation across all tested architectures. This confirms our choice is a generalizable one, not specific to a

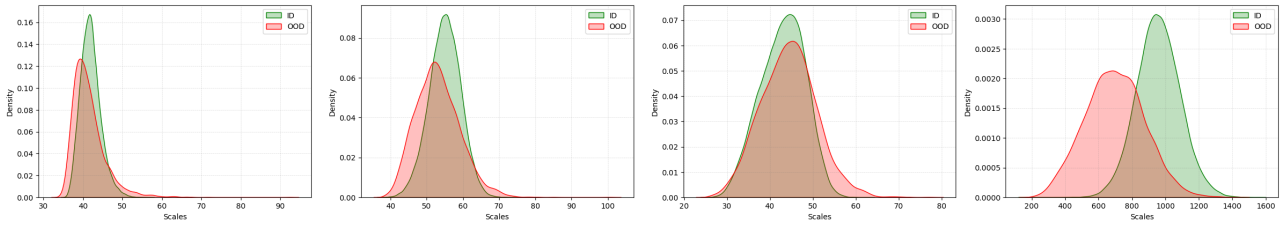

Figure 8. *Distributions of the scaling factor (γ) are computed from the four residual stages. The model was trained on ImageNet-1K (ID) and evaluated against Texture (OOD). (a-c) The γ distributions from the early-to-mid stages (Layer 1 to Layer 3) show significant overlap between ID and OOD samples, rendering them ineffective as a discriminative signal. (d) In sharp contrast, the distribution from the final residual stage (Layer 4) provides a clear and distinct separation. This result, consistent across OOD datasets, validates our methodological focus on the penultimate layer's pre-pooling feature map as the most potent and reliable signal source.*

single model.

I. Analysis of Fusion Strategy

In Section 3, we introduced two potential fusion strategies for integrating our scaling factor $\gamma(\mathbf{x})$ with a baseline score $S(\mathbf{x})$: multiplicative (Eq. 12a) and additive (Eq. 12b).

$$S^*(\mathbf{x};\theta,\gamma) = \gamma(\mathbf{x};f) \times S(\mathbf{x};\theta) \tag{12a}$$

$$S^+(\mathbf{x};\theta,\gamma) = \gamma(\mathbf{x};f) + S(\mathbf{x};\theta) \tag{12b}$$

As shown in Table 19, a comparative evaluation on the ImageNet benchmark reveals that both strategies can achieve a similar level of performance. This confirms that the core discriminative power originates from the γ signal itself, not the specific mathematical operator.

However, a critical distinction emerged when analyzing their hyperparameter sensitivity and robustness. The scaling factors γ^* (multiplicative) and γ^+ (additive) are both derived from Eq. 4, but they require different optimal settings for the clipping threshold, c^* and c^+ respectively, as detailed in Equation 13.

$$\gamma^*(\mathbf{x};f) = \sum_{i=1}^{n} \min(f_i(\mathbf{x}), c^*) \tag{13a}$$

$$\gamma^+(\mathbf{x};f) = \sum_{i=1}^{n} \min(f_i(\mathbf{x}), c^+) \tag{13b}$$

For proposed multiplicative strategy, the optimal threshold c^* is set by selecting a moderate percentile as detailed in reproducibility section in Appendix G of the ID training data's $f_i(\mathbf{x})$ values. This procedure is robust, stable, and aligns with foundational post-hoc methods like ReAct and SCALE (details in Appendix G). For the additive strategy, we empirically found that the optimal threshold c^+ was consistently an order of magnitude smaller, (e.g., $c^+ \approx 0.1 \times c^*$). On ImageNet, this optimal c^+ value corresponds to an extremely low percentile of the ID data (e.g., below the 1st percentile for ResNet-50).

This low-percentile tuning makes the additive strategy operationally fragile. Tuning at the extreme low activation threshold could be fragile and sensitive to small shifts in data or model, making it a poor choice for a general-purpose method. Therefore, while both methods achieve similar performance, we chose multiplicative fusion as our primary strategy. It provides not only competitive performance but also the practical robustness and hyperparameter stability required of a *plug-and-play* framework. This choice aligns conceptually with our *elastic scaling* narrative, where γ acts as an input-dependent modulator of the baseline score.

J. Alternate Statistics: Median and Entropy

To justify our final methodological choice of using mean, standard deviation, and maximum statistics, we conducted a rigorous analysis of two common alternatives: median and Shannon entropy. For a statistic to be viable for our framework, it must produce a scaling factor γ that has a distinct and reliable signature for ID versus OOD samples.

Setup. As we discussed in Section 3 of main paper, we compute the scaling factor γ using a trained deep neural network $\theta : \mathbb{R}^d \to \mathbb{R}^C$ 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 output classes. The network's penultimate layer produces a feature vector $h(\mathbf{x}) \in \mathbb{R}^n$ by applying global average pooling 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.

Median. We begin by extracting the *median* from each activation map of $g(\mathbf{x}) \in \mathbb{R}^{n\times k\times k}$, transforming it into an n-dimensional feature vector $h(\mathbf{x}) \in \mathbb{R}^n$ using global median pooling, as defined in Equation 14:

Model	Fusion	Method	SUN		Places365		Texture		iNaturalist		Average	
			FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
ResNet-34	Catalyst(+)	Energy	57.39	86.59	62.61	84.59	54.95	86.45	53.86	89.73	57.20	86.84
		Catalyst(μ)	35.50	91.69	46.05	89.07	12.66	96.66	20.03	96.20	28.56	93.41
		Catalyst(σ)	39.93	90.29	49.61	87.65	14.45	95.83	25.53	95.03	32.38	92.20
		Catalyst(m)	45.07	89.94	57.08	86.80	10.46	97.61	26.55	95.17	34.79	92.38
		Catalyst(μ) + **ReAct**	22.23	94.80	32.07	92.09	18.19	95.91	15.96	96.89	22.11	94.92
		Catalyst(σ) + **ReAct**	23.10	94.66	33.18	91.92	17.43	96.07	17.18	96.75	22.72	94.85
		Catalyst(m) + **ReAct**	37.74	92.96	49.69	89.86	11.21	97.60	23.36	96.01	30.50	94.11
	Catalyst(*)	Energy	57.39	86.59	62.61	84.59	54.95	86.45	53.86	89.73	57.20	86.84
		Catalyst(μ)	33.46	91.85	43.78	89.39	24.86	93.39	25.60	94.99	31.92	92.41
		Catalyst(σ)	37.78	90.74	48.87	87.82	16.08	95.80	24.90	95.10	31.91	92.36
		Catalyst(m)	36.90	90.88	48.37	87.87	16.83	95.58	25.22	95.00	31.83	92.34
		Catalyst(μ) + **ReAct**	21.44	95.18	31.74	92.56	13.39	97.02	12.81	97.47	19.84	95.56
		Catalyst(σ) + **ReAct**	21.80	95.06	32.11	92.41	12.73	97.11	13.01	97.41	19.91	95.50
		Catalyst(m) + **ReAct**	22.03	94.99	32.58	92.31	12.55	97.15	13.47	97.33	20.16	95.44
ResNet-50	Catalyst(+)	Energy	58.28	86.73	65.40	84.13	52.29	86.73	53.95	90.59	57.48	87.05
		Catalyst(μ)	30.99	93.13	43.36	89.97	9.72	97.71	12.94	97.45	24.25	94.57
		Catalyst(σ)	34.07	91.87	45.43	88.64	10.44	97.19	15.25	96.88	26.30	93.64
		Catalyst(m)	42.08	91.70	55.43	88.17	9.47	98.10	20.81	96.44	31.95	93.60
		Catalyst(μ) + **ReAct**	19.73	95.28	29.47	92.73	12.68	97.10	9.99	97.89	17.97	95.75
		Catalyst(σ) + **ReAct**	20.64	95.12	30.94	92.51	10.53	97.56	10.25	97.79	18.09	95.74
		Catalyst(m) + **ReAct**	37.59	93.66	50.32	90.69	9.88	98.07	18.78	96.93	29.14	94.84
	Catalyst(*)	Energy	58.28	86.73	65.40	84.13	52.29	86.73	53.95	90.59	57.48	87.05
		Catalyst(μ)	30.79	92.67	42.59	89.78	22.29	94.01	18.02	96.46	28.42	93.23
		Catalyst(σ)	35.73	91.47	48.35	88.04	15.85	95.94	19.05	96.21	29.75	92.92
		Catalyst(m)	35.79	91.40	48.68	87.82	16.08	95.88	19.00	96.18	29.89	92.82
		Catalyst(μ) + **ReAct**	18.46	95.82	28.98	93.31	12.11	97.38	8.54	98.19	17.02	96.18
		Catalyst(σ) + **ReAct**	19.13	95.61	29.58	93.04	12.04	97.38	9.10	98.06	17.46	96.02
		Catalyst(m) + **ReAct**	19.02	95.52	29.77	92.92	12.06	97.31	9.71	97.97	17.64	95.93
MobileNet-v2	Catalyst(+)	Energy	59.36	86.24	66.27	83.21	54.54	86.58	55.31	90.34	58.87	86.59
		Catalyst(μ)	39.84	90.64	54.28	86.57	12.20	96.47	32.62	94.00	34.73	91.92
		Catalyst(σ)	42.03	90.52	56.96	86.35	9.73	97.27	34.94	93.66	35.92	91.95
		Catalyst(m)	43.30	89.94	57.90	85.84	10.07	97.07	36.45	93.34	36.93	91.55
		Catalyst(μ) + **ReAct**	40.64	90.34	53.27	86.40	11.05	96.97	29.73	94.68	33.67	92.10
		Catalyst(σ) + **ReAct**	41.41	90.61	55.85	86.46	8.17	97.78	31.46	94.37	34.22	92.31
		Catalyst(m) + **ReAct**	41.98	90.31	55.76	86.19	8.19	97.71	31.75	94.27	34.42	92.12
	Catalyst(*)	Energy	59.36	86.24	66.27	83.21	54.54	86.58	55.31	90.34	58.87	86.59
		Catalyst(μ)	37.74	91.43	52.21	87.33	23.42	94.17	33.47	93.84	36.71	91.69
		Catalyst(σ)	38.20	91.26	53.04	86.84	14.02	96.37	29.25	94.63	33.63	92.27
		Catalyst(m)	37.41	91.37	52.24	86.89	14.18	96.35	28.78	94.70	33.15	92.33
		Catalyst(μ) + **ReAct**	32.82	92.93	48.62	88.59	13.60	96.83	28.19	94.89	30.81	93.31
		Catalyst(σ) + **ReAct**	37.53	91.22	51.32	87.19	10.18	97.31	27.21	95.12	31.56	92.71
		Catalyst(m) + **ReAct**	34.77	92.26	49.77	88.06	8.69	97.76	24.08	95.66	29.33	93.43
DenseNet-121	Catalyst(+)	Energy	52.51	87.27	58.24	85.05	52.22	85.42	39.75	92.66	50.68	87.60
		Catalyst(μ)	36.68	90.59	45.60	87.91	20.62	94.00	19.43	96.07	30.58	92.14
		Catalyst(σ)	35.10	90.97	44.54	88.22	16.90	95.11	18.40	96.24	28.73	92.64
		Catalyst(m)	38.70	91.31	50.39	88.13	11.86	97.36	21.68	95.89	30.66	93.17
		Catalyst(μ) + **ReAct**	38.24	91.84	47.00	88.61	14.34	97.03	17.80	96.50	29.35	93.49
		Catalyst(σ) + **ReAct**	32.72	92.71	43.39	89.39	10.69	97.76	15.21	96.95	25.50	94.20
		Catalyst(m) + **ReAct**	37.89	92.34	51.97	88.63	7.32	98.41	20.80	96.28	29.50	93.92
	Catalyst(*)	Energy	52.51	87.27	58.24	85.05	52.22	85.42	39.75	92.66	50.68	87.60
		Catalyst(μ)	33.24	91.84	42.94	89.01	25.59	93.29	16.41	96.69	29.54	92.71
		Catalyst(σ)	34.12	91.57	44.34	88.53	21.06	94.52	16.95	96.57	29.12	92.80
		Catalyst(m)	34.29	91.47	44.74	88.35	21.26	94.43	17.50	96.46	29.45	92.68
		Catalyst(μ) + **ReAct**	31.58	93.41	42.77	90.30	12.71	97.44	14.66	97.11	25.43	94.56
		Catalyst(σ) + **ReAct**	30.04	93.44	41.50	90.24	11.37	97.61	14.12	97.16	24.26	94.61
		Catalyst(m) + **ReAct**	30.25	93.37	41.61	90.13	11.74	97.52	14.48	97.09	24.52	94.53

Table 19. *Analysis of fusion strategies on detection performance on ImageNet benchmarks. All values are percentages and are averaged over four common OOD benchmark datasets.* `Catalyst`$(+)$ *represents additive strategy and* `Catalyst`$(*)$ *represents multiplicative strategy. The symbol* ↓ *indicates lower values are better;* ↑ *indicates larger values are better.*

$$h(\mathbf{x}) = \texttt{median}\left(g(\mathbf{x})\right) \tag{14}$$

Here, `median` denotes a global median pooling operation applied independently to each of the n activation maps in $g(\mathbf{x})$.

Shannon Entropy. In addition to the median, we compute the *Shannon entropy* for each activation map. For the i-th channel activation $g_i(\mathbf{x}) \in \mathbb{R}^{k\times k}$, the entropy is computed as shown in Equation 16. To do so, we first flatten $g_i(\mathbf{x})$ into a vector of length k^2, and normalize it to define a discrete

probability distribution p_{ij}, as described in Equation 15. By collecting the entropy values across all channels, we obtain the final feature representation $h(\mathbf{x}) \in \mathbb{R}^n$, as defined in Equation 17 .

$$p_{ij} = \frac{g_i(\mathbf{x})_j}{\sum_{l=1}^{k^2} g_i(\mathbf{x})_l}, \quad j = 1, \ldots, k^2 \tag{15}$$

$$\texttt{entropy}_i(\mathbf{x}) = -\sum_{j=1}^{k^2} p_{ij} \log p_{ij} \tag{16}$$

$$\begin{aligned} h(\mathbf{x}) &= \texttt{entropy}\left(g(\mathbf{x})\right) \\ &= \left[\texttt{entropy}_1(\mathbf{x}), \ldots, \texttt{entropy}_n(\mathbf{x})\right]^\top \end{aligned} \tag{17}$$

Computing the Scaling Factor γ. The $\gamma(\mathbf{x})$ computation using the median follows the same principle described in Section 3: a higher median activation is assumed to indicate an ID sample. The Shannon entropy, however, exhibits the opposite behavior. As alluded to in our motivation (Section 1) and empirically demonstrated (Figure 1), ID samples typically have lower entropy (i.e., less uncertainty) than OOD samples. So, to maintain the convention that an ID sample exhibits higher score than OOD samples, the entropy-based scaling factor must be inverted (i.e, $\frac{1}{\gamma(\mathbf{x})}$) before it is applied to the baseline score.

Evaluation. Using Shannon entropy as the information cue for our scaling factor γ yields a notable performance improvement, particularly when combined with strong baselines like ReAct+DICE. This enhancement is especially pronounced for the MobileNet-V2 architecture. As shown in Table 23, our entropy-based scaling improves upon the vanilla ReAct+DICE baseline by 14.65%, achieving superior performance among all foundational methods compared. A slight improvement of 5.90% is also observed for the ResNet-50 backbone. For brevity, the table presents results for these two representative architectures.

We present the performance of our method, `Catalyst`, across both ImageNet and CIFAR benchmarks when the scaling factor (γ) is computed using the median and entropy statistics. For the ImageNet evaluation, detailed results for the ResNet-50 and MobileNet-V2 architectures are shown in Table 22 and Table 23, respectively. For the CIFAR benchmarks, we present detailed results for two architectures. For ResNet-18, the performance on CIFAR-10 and CIFAR-100 is shown in Table 24 and Table 25, respectively. Similarly, for DenseNet-101, the results for CIFAR-10 and CIFAR-100 are in Table 26 and Table 27.

Discussion. Across all evaluated benchmarks (Tables 22–27), the results for the median statistic are conclusive. Across all evaluated benchmarks, using the median to compute γ consistently degrades performance. This degradation occurs because the median violates our method's core assumption: its statistical signature fails to separate ID and OOD samples, resulting in a γ with high overlap. Figure 9 provides a representative example of this distribution collapse. Given its consistent failure, the median was conclusively rejected as a viable statistic.

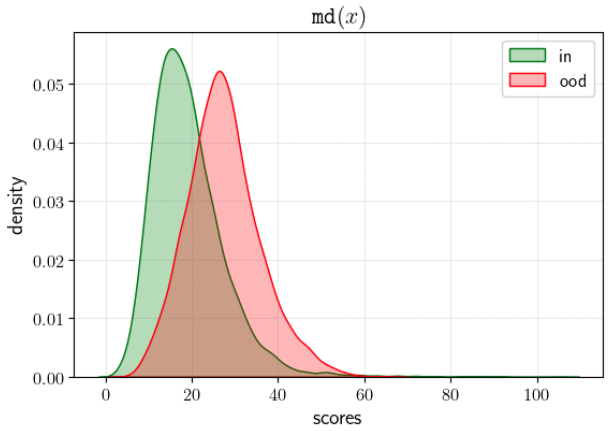

Figure 9. *Distribution of the scaling factor, γ, computed using the median statistic. The model is a DenseNet-101 trained on CIFAR-100 (ID), evaluated against the SVHN dataset (OOD). The plot reveals that the OOD distribution is shifted to the right of the ID distribution, indicating that OOD samples produce a higher γ value than ID samples. This contradicts the core assumption of our method, leading to degraded OOD detection performance.*

The analysis of Shannon entropy is more nuanced and reveals a critical insight. Entropy is not consistently ineffective; rather, it is inconsistent. In specific cases, entropy can be very effective. For example, on the ImageNet benchmark (Table 23), the entropy-based γ improves the `ReAct+DICE` baseline by 14.65% on the MobileNet-V2 architecture, achieving the best performance for that specific model. However, this strong performance is not generalizable. On the same dataset but with a ResNet-50 backbone (Table 22), the improvement is minimal (5.90%) and lags behind our proposed mean/std/max combination. This inconsistency can also be seen in ablation study of using scaling factor standalone as scoring metric in Appendix K, where we show that γ_{entropy} as a standalone score was dominant on CIFAR but failed to generalize to ImageNet.

This rigor confirms that our chosen statistics (mean, standard deviation, maximum) are the most effective choice, providing a robust and consistently high-performing signal across all models and datasets.

K. Analysis of γ as a Standalone OOD Score

The core hypothesis of our work is that the scaling factor $\gamma(\mathbf{x})$, contains significant discriminative information. While our primary method uses γ as a modulator for existing OOD scores (e.g., Energy, MSP, ODIN), an important question is whether γ is powerful enough to serve as a standalone OOD scoring function. Furthermore, this analysis

allows us to identify which of its component statistics are the most robust and generalizable.

To investigate this, we conducted a standalone analysis of γ, comparing it directly to the strong Energy score. We computed γ individually from four distinct channel-wise statistics: mean, standard deviation, maximum, and entropy. (We omit median as an initial analysis, detailed in Appendix J, showed insufficient discriminative power). For entropy, we found its reciprocal ($1/\gamma_{\text{entropy}}$) without thresholding was its most potent configuration, and we use that for this analysis.

Analysis on ImageNet. To test the generalizability of these findings, we repeated the analysis on the large-scale ImageNet benchmark (Table 20). Here, the trend dramatically reversed. The scores derived from γ_{std}, γ_{max}, and even γ_{mean} remained robust and generalizable, consistently outperforming the Energy baseline by a significant margin (e.g., 19.32% for γ_{std} on ResNet-50).

In sharp contrast, the γ_{entropy} score, which was dominant on CIFAR, failed to generalize. It not only performed worse than the other γ statistics but also failed to consistently beat the Energy baseline. For instance, on DenseNet-121, it scored a poor 53.09% FPR95 compared to Energy's 50.68% (a 2.4-point gap), and on ResNet-50, it merely matched the Energy score (56.73% vs. 57.48%). This demonstrates that entropy, while powerful on simpler datasets, is not a reliable or generalizable statistic for OOD detection on more complex, large-scale tasks.

This analysis provides a critical insight and directly justifies our final methodological design (Equation 3 and 4). Our Catalyst framework is constructed by combining the statistics that proved to be consistently robust across all benchmarks (mean, standard deviation, maximum), while entropy is deliberately excluded due to its clear lack of generalizability.

Analysis on CIFAR. As shown in Table 21, the standalone performance of γ on CIFAR is remarkably strong. The scores from γ_{std} and γ_{max} are highly competitive, matching or exceeding the Energy score baseline. The γ_{mean} score is less effective, which aligns with the poor signal separation observed in our motivational analysis (Figure 1).

Notably, the γ_{entropy} score is effective on these benchmarks. It dramatically outperforms the Energy baseline by 69.03% (CIFAR-10) and 31.62% (CIFAR-100) on ResNet-18, and by 33.69% (CIFAR-10) and 15.29% (CIFAR-100) on DenseNet-101. Based on this initial finding, entropy would appear to be the most powerful standalone signal. As a representative, Figure 10 visually demonstrates the superior distribution separation achieved by the standalone γ_{entropy} score compared to the Energy baseline on the CIFAR-100 benchmark.

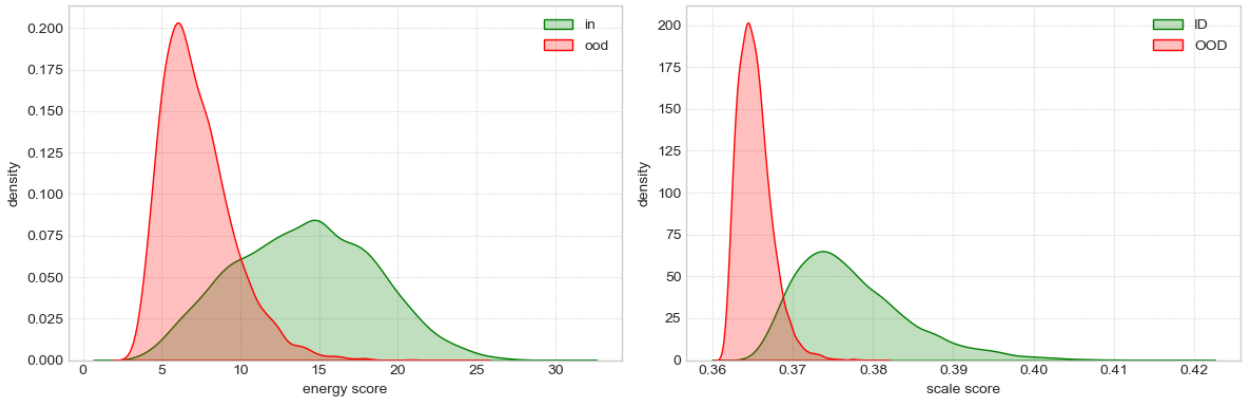

Figure 10. *Superior OOD separation of $\gamma_{entropy}$ as a standalone score on CIFAR-100. The model is a ResNet-18 trained on CIFAR-100 (ID), evaluated against the Texture dataset (OOD). (Left) Significant distribution overlap between ID and OOD using the baseline Energy score. (Right) Dramatically improved separation using the standalone $\gamma_{entropy}$ score.This visualization confirms the finding from Table 21 that entropy is an exceptionally powerful standalone signal on the CIFAR benchmarks.*

Model	**Method**	SUN		Places		Texture		iNaturalist		Average	
		FPR95 ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
ResNet-34	Energy	57.39	86.59	62.61	84.59	54.95	86.45	53.86	89.73	57.20	86.84
	Mean	44.83	96.01	56.88	94.29	27.99	98.88	35.96	97.69	41.42	96.72
	Std	56.73	94.09	69.36	91.74	18.81	99.22	42.05	97.14	46.74	95.55
	Max	54.94	94.12	68.05	91.63	19.11	99.19	42.87	97.01	46.24	95.49
	Entropy	63.75	91.93	75.47	88.55	20.64	99.03	52.63	95.25	53.12	93.69
ResNet-50	Energy	58.28	86.73	65.40	84.13	52.29	86.73	53.95	90.59	57.48	87.05
	Mean	43.71	96.52	56.03	94.69	27.23	99.00	30.67	98.22	39.41	97.11
	Std	56.57	94.83	69.33	92.40	20.36	99.20	39.24	97.61	46.37	96.01
	Max	56.12	94.63	69.41	92.03	20.28	99.19	39.67	97.53	46.37	95.84
	Entropy	71.43	90.98	81.43	87.52	21.84	98.90	52.21	95.38	56.73	93.20
MobileNet-v2	Energy	59.36	86.24	66.27	83.21	54.54	86.58	55.31	90.34	58.87	86.59
	Mean	44.23	96.84	60.08	94.60	21.14	99.40	45.29	96.98	42.68	96.96
	Std	50.97	96.02	67.22	93.28	14.54	99.60	46.03	97.19	44.69	96.52
	Max	50.33	95.96	66.70	93.10	14.56	99.60	46.58	97.22	44.54	96.47
	Entropy	56.30	94.59	71.04	90.89	15.43	99.52	50.40	96.28	48.29	95.32
DenseNet-121	Energy	52.51	87.27	58.24	85.05	52.22	85.42	39.75	92.66	50.68	87.60
	Mean	56.90	95.02	68.16	93.11	36.99	98.58	40.51	97.42	50.64	96.03
	Std	58.44	94.80	69.69	92.64	29.63	98.89	41.84	97.36	49.90	95.92
	Max	58.35	94.65	70.09	92.37	29.68	98.89	43.02	97.21	50.29	95.78
	Entropy	61.31	92.74	73.05	89.28	29.86	98.56	48.13	95.38	53.09	93.99

Table 20. *Detailed OOD detection results on ImageNet benchmarks using scaling factor γ as a standalone scoring metric. ↓ indicates lower values are better and ↑ indicates larger values are better.*

Dataset	Model	Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
			FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR-10	ResNet-18	Energy	44.32	94.04	41.43	91.72	35.22	94.70	50.30	91.11	9.77	98.19	31.97	95.26	35.50	94.17
		Mean	29.84	85.75	97.48	38.81	84.09	57.59	58.63	79.92	30.37	91.33	87.08	53.23	64.58	67.77
		Std	5.56	98.97	56.34	88.34	17.21	97.48	12.43	98.56	0.35	99.87	20.93	96.74	18.80	96.66
		Max	5.39	99.00	54.87	88.51	14.49	97.81	11.42	98.62	0.57	99.84	17.56	97.16	17.38	96.82
		Entropy	6.14	98.81	36.31	93.29	7.28	98.83	8.09	99.07	0.73	99.81	8.65	98.55	11.20	98.06
	DenseNet-101	Energy	37.91	93.59	36.42	92.38	7.33	98.27	43.87	90.48	1.95	99.47	6.97	98.38	22.41	95.43
		Mean	7.66	98.49	87.02	71.43	50.85	93.13	28.39	94.51	9.86	98.39	59.84	91.56	40.60	91.25
		Std	10.35	97.69	77.55	74.92	15.17	97.59	18.28	96.94	1.54	99.62	16.99	97.06	23.31	93.97
		Max	10.47	97.62	77.68	74.83	16.35	97.47	17.52	97.11	2.73	99.42	18.48	96.87	23.87	93.89
		Entropy	8.80	97.67	53.28	86.95	7.13	98.79	9.79	98.56	2.68	99.43	7.54	98.59	14.86	96.67
CIFAR-100	ResNet-18	Energy	66.64	89.53	81.39	76.83	71.46	83.02	85.18	75.68	48.01	91.63	68.57	84.53	70.21	83.54
		Mean	88.94	80.05	98.54	42.35	96.46	58.22	63.79	79.95	57.49	79.04	97.24	54.51	83.74	65.68
		Std	36.12	93.51	96.38	49.35	76.44	82.19	46.88	88.67	26.59	95.08	81.57	79.82	60.66	81.43
		Max	35.18	93.69	95.98	50.69	77.05	83.24	46.92	88.95	25.07	95.52	82.80	80.45	60.50	82.09
		Entropy	22.71	95.73	93.66	62.23	56.89	90.97	36.17	93.60	16.57	97.24	62.06	89.39	48.01	88.19
	DenseNet-101	Energy	70.99	86.66	77.28	76.94	59.39	85.68	83.49	67.47	11.45	97.89	50.90	88.57	58.92	83.87
		Mean	31.77	94.25	95.11	55.02	78.44	81.61	40.41	92.03	18.11	97.11	87.01	77.82	58.47	82.97
		Std	30.65	94.73	93.83	60.93	65.01	87.40	34.13	94.21	11.03	98.26	73.04	85.11	51.28	86.77
		Max	30.81	94.70	93.85	62.60	64.63	88.43	33.24	94.68	14.43	97.74	73.70	86.22	51.77	87.39
		Entropy	34.67	93.78	93.64	65.11	58.01	89.88	30.53	95.32	16.15	97.34	66.51	88.03	49.91	88.24

Table 21. *Detailed OOD detection results on CIFAR benchmarks using scaling factor γ as a standalone scoring metric.↓ indicates lower values are better and ↑ indicates larger values are better.*

L. Societal Impact

The reliable detection of out-of-distribution (OOD) inputs is a fundamental requirement for safe and trustworthy deployment of machine learning systems. This capability is critical in high-stakes domains such as autonomous transportation, where an unexpected object on the road must be identified as anomalous, and in medical diagnostics, where a model must recognize that a scan presents features of an unseen disease. By improving the separation between in-distribution (ID) and OOD data, our work directly contributes to building more robust and dependable AI. The primary benefit of our approach, `Catalyst`, is its potential to reduce critical failure rates, which is crucial for ensuring user safety and earning public trust in automated systems.

The broader impact of this research lies in enhancing the safety and reliability of AI. Our work adheres to ethical research standards, does not involve human subjects, and uses publicly available datasets. While any powerful technology can have unforeseen applications, our work is fundamentally aimed at mitigating the harm that arises from brittle AI models that fail silently or unpredictably when faced with novel inputs. By releasing our code to the public, we hope to foster further research, encourage reproducibility, and accelerate the development of more robust AI systems that can be deployed responsibly in society.

M. Synergy with Existing Methods

`Catalyst` is designed to be fully compatible with existing post-hoc OOD detection techniques, enabling seamless integration with widely used methods such as `MSP` [16], `ODIN` [32], `Energy` [36], `ReAct` [55], `DICE` [54], `ASH` [5] and `KNN` [56]. Rather than replacing these techniques, `Catalyst` acts as a complementary module. It enhances their ability to separate in-distribution and out-of-distribution samples through an elastic scaling mechanism, introducing an additional degree of freedom that works in tandem with them. In our evaluation, we omit ODIN [32] from our analysis due to its high computational cost and limited performance on large-scale datasets like ImageNet. The method's expense stems from requiring an FGSM-based perturbation for every input sample.

To demonstrate the effectiveness of this synergy on the ImageNet benchmark, we present detailed experimental results in Table 22 and Table 23, showing the performance of each baseline method when combined with `Catalyst`. The results consistently indicate performance improvements, validating the benefit of integrating `Catalyst` with established methods. For brevity, the table presents results for these two representative architectures, ResNet-50 and MobileNet-V2.

To demonstrate the effectiveness of this synergy on the CIFAR benchmarks, we present detailed experimental results in Table 24, 26 (CIFAR-10) and Table 25, 27 (CIFAR-100), showing the performance of each baseline method when combined with `Catalyst`. The results demonstrate consistent performance improvements, validating the benefit of integrating `Catalyst` with established baselines.

Model	Combined Method	SUN		Place365		Textures		iNaturalist		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet-50	MSP	68.58	81.75	71.57	80.63	66.13	80.46	52.77	88.42	64.76	82.82
	+ Catalyst(μ)	56.58	87.33	62.93	85.10	52.48	87.81	39.20	92.41	52.80	88.16
	+ Catalyst(σ)	59.13	85.63	65.72	82.89	42.66	91.76	39.26	92.31	51.69	88.15
	+ Catalyst(m)	58.92	85.51	65.71	82.53	42.82	91.67	39.02	92.29	51.62	88.00
	+ Catalyst(md)	53.93	88.34	59.01	86.99	63.49	79.42	42.50	91.27	54.73	86.51
	+ Catalyst(e)	67.59	82.41	70.90	80.79	63.87	83.21	51.05	89.05	63.35	83.86
	Energy	58.28	86.73	65.40	84.13	52.29	86.73	53.95	90.59	57.48	87.05
	+ Catalyst(μ)	30.79	92.67	42.59	89.78	22.29	94.01	18.02	96.46	28.42	93.23
	+ Catalyst(σ)	35.73	91.47	48.35	88.04	15.85	95.94	19.05	96.21	29.75	92.92
	+ Catalyst(m)	35.79	91.40	48.68	87.82	16.08	95.88	19.00	96.18	29.89	92.82
	+ Catalyst(md)	30.23	93.24	38.47	91.31	47.70	87.77	25.46	95.29	35.46	91.90
	+ Catalyst(e)	52.40	87.75	62.06	84.76	41.35	89.27	44.22	92.20	50.01	88.50
	ReAct	23.68	94.44	33.33	91.96	46.33	90.30	19.73	96.37	30.77	93.27
	+ Catalyst(μ)	18.46	95.82	28.98	93.31	12.11	97.38	8.54	98.19	17.02	96.18
	+ Catalyst(σ)	19.13	95.61	29.58	93.04	12.04	97.38	9.10	98.06	17.46	96.02
	+ Catalyst(m)	19.02	95.52	29.77	92.92	12.06	97.31	9.71	97.97	17.64	95.93
	+ Catalyst(md)	24.41	95.55	31.75	93.63	57.62	87.82	22.06	96.18	33.96	93.30
	+ Catalyst(e)	22.64	94.55	34.09	91.58	22.27	94.91	13.94	97.25	23.23	94.57
	DICE	36.11	91.01	47.62	87.76	32.38	90.48	26.48	94.53	35.65	90.94
	+ Catalyst(μ)	31.28	92.16	43.09	89.11	19.91	94.46	16.59	96.58	27.72	93.08
	+ Catalyst(σ)	32.86	91.85	45.67	88.45	20.05	94.45	20.18	95.82	29.69	92.65
	+ Catalyst(m)	33.56	91.77	47.04	88.18	18.62	95.06	20.63	95.82	29.96	92.71
	+ Catalyst(md)	30.26	93.53	38.80	91.09	46.90	88.00	25.14	95.17	35.27	91.95
	+ Catalyst(e)	36.11	90.98	47.72	87.72	32.38	90.48	26.60	94.50	35.70	90.92
	ReAct+DICE	24.05	94.31	34.28	91.71	28.40	93.33	14.90	97.06	25.41	94.10
	+ Catalyst(μ)	23.47	94.82	34.08	92.18	14.45	96.91	10.86	97.86	20.72	95.44
	+ Catalyst(σ)	23.76	94.65	34.58	92.06	15.07	96.70	11.40	97.75	21.20	95.29
	+ Catalyst(m)	25.33	94.46	36.92	91.67	13.81	97.02	12.35	97.61	22.10	95.19
	+ Catalyst(md)	24.25	95.25	32.08	93.19	46.68	90.50	18.80	96.63	30.45	93.89
	+ Catalyst(e)	22.07	94.78	31.61	92.32	28.00	93.54	13.98	97.25	23.91	94.47
	ASH	28.01	94.02	39.84	90.98	11.95	97.60	11.52	97.87	22.83	95.12
	+ Catalyst(μ)	28.97	93.75	41.04	90.53	11.47	97.79	12.08	97.74	23.39	94.95
	+ Catalyst(σ)	29.76	93.73	41.75	90.77	11.56	97.57	12.24	97.75	23.83	94.95
	+ Catalyst(m)	28.23	93.97	40.20	90.90	11.49	97.73	11.60	97.85	22.88	95.11
	+ Catalyst(md)	26.32	94.43	36.36	91.67	19.52	96.17	13.95	97.35	24.04	94.91
	+ Catalyst(e)	27.96	94.01	39.81	90.97	11.93	97.60	11.50	97.87	22.80	95.11
	SCALE	25.78	94.54	36.86	91.96	14.56	96.75	10.37	98.02	21.89	95.32
	+ Catalyst(μ)	25.38	94.57	36.55	91.83	11.90	97.45	10.11	98.06	20.98	95.48
	+ Catalyst(σ)	25.58	94.51	36.99	91.77	11.83	97.48	10.31	98.03	21.18	95.45
	+ Catalyst(m)	25.60	94.51	37.09	91.76	11.79	97.48	10.32	98.02	21.20	95.44
	+ Catalyst(md)	23.67	95.12	33.19	92.83	23.37	95.15	12.91	97.56	23.28	95.17
	+ Catalyst(e)	25.77	94.47	37.04	91.81	14.01	96.84	10.34	98.02	21.79	95.29

Table 22. *Detailed results of post-hoc methods combined with* `Catalyst` *on four OOD benchmarks: SUN, Places365, Textures, and iNaturalist using ResNet-50 trained on ImageNet-1K.* ↑ *indicates higher is better;* ↓ *indicates lower is better. The symbols denote the statistic used:* μ *(mean),* σ *(std. deviation),* m *(maximum),* md *(median), and* e *(Shannon entropy).*

Model	Combined Method	SUN		Place365		Textures		iNaturalist		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MobileNet-V2	MSP	74.20	78.88	76.89	78.14	70.99	78.95	59.86	86.72	70.49	80.67
	+ Catalyst(μ)	63.48	85.01	69.78	82.57	57.46	86.55	48.69	90.06	59.85	86.05
	+ Catalyst(σ)	63.13	84.81	69.71	81.73	43.67	91.44	45.33	90.88	55.46	87.22
	+ Catalyst(m)	63.19	84.78	69.63	81.84	46.12	90.69	46.05	90.73	56.25	87.01
	+ Catalyst(md)	66.47	83.52	70.73	82.29	73.14	76.11	57.43	87.31	66.94	82.31
	+ Catalyst(e)	72.66	80.27	75.86	78.90	67.94	81.88	57.41	87.59	68.47	82.16
	Energy	59.36	86.24	66.27	83.21	54.54	86.58	55.31	90.34	58.87	86.59
	+ Catalyst(μ)	37.74	91.43	52.21	87.33	23.42	94.17	33.47	93.84	36.71	91.69
	+ Catalyst(σ)	38.20	91.26	53.04	86.84	14.02	96.37	29.25	94.63	33.63	92.27
	+ Catalyst(m)	37.41	91.37	52.24	86.89	14.18	96.35	28.78	94.70	33.15	92.33
	+ Catalyst(md)	52.89	88.64	62.06	85.82	67.66	82.51	62.71	87.50	61.33	86.12
	+ Catalyst(e)	52.16	87.95	61.69	84.32	41.17	89.95	45.70	92.08	50.18	88.58
	ReAct	52.46	87.26	59.89	84.07	40.25	90.96	43.05	92.72	48.91	88.75
	+ Catalyst(μ)	32.82	92.93	48.62	88.59	13.60	96.83	28.19	94.89	30.81	93.31
	+ Catalyst(σ)	37.53	91.22	51.32	87.19	10.18	97.31	27.21	95.12	31.56	92.71
	+ Catalyst(m)	34.77	92.26	49.77	88.06	8.69	97.76	24.08	95.66	29.33	93.43
	+ Catalyst(md)	50.96	89.62	59.73	86.80	71.45	82.57	63.14	86.49	61.32	86.37
	+ Catalyst(e)	36.32	91.14	50.91	86.16	13.71	96.34	23.20	95.70	31.03	92.33
	DICE	37.84	90.81	52.35	86.17	32.57	91.46	41.53	91.30	41.07	89.94
	+ Catalyst(μ)	36.22	91.56	51.48	87.17	16.45	95.78	34.79	93.38	34.74	91.97
	+ Catalyst(σ)	36.31	91.29	51.20	87.03	17.15	95.40	35.07	93.25	34.93	91.74
	+ Catalyst(m)	34.90	91.80	50.45	87.32	14.86	96.22	32.41	93.85	33.15	92.30
	+ Catalyst(md)	47.96	89.30	58.91	85.37	61.90	83.26	61.27	84.74	57.51	85.67
	+ Catalyst(e)	37.96	90.76	52.11	86.28	28.88	92.59	36.19	93.07	38.79	90.68
	ReAct+DICE	30.60	92.98	45.93	88.29	16.03	96.33	31.68	93.76	31.06	92.84
	+ Catalyst(μ)	33.90	92.65	49.33	88.22	9.52	97.87	31.04	94.44	30.95	93.30
	+ Catalyst(σ)	32.67	92.57	47.68	88.29	9.88	97.61	29.20	94.66	29.86	93.28
	+ Catalyst(m)	32.60	92.58	47.77	88.30	9.79	97.63	29.29	94.66	29.86	93.29
	+ Catalyst(md)	48.95	89.51	59.14	85.64	65.62	83.49	63.50	83.56	59.30	85.55
	+ Catalyst(e)	26.33	93.86	40.71	89.65	13.87	96.60	25.12	95.22	26.51	93.83
	ASH	43.63	90.02	58.85	84.73	13.12	97.10	39.13	91.94	38.68	90.95
	+ Catalyst(μ)	40.05	90.86	55.47	85.83	14.49	96.77	37.05	92.59	36.76	91.51
	+ Catalyst(σ)	41.76	90.45	57.32	85.12	10.92	97.64	34.17	93.32	36.04	91.63
	+ Catalyst(m)	42.01	90.41	57.41	85.02	11.12	97.62	36.11	92.94	36.66	91.50
	+ Catalyst(md)	43.12	89.49	57.70	83.57	20.02	95.86	45.78	88.40	41.65	89.33
	+ Catalyst(e)	43.33	89.97	58.82	84.47	12.71	97.24	38.66	91.97	38.38	90.91
	SCALE	38.74	91.64	53.49	87.34	14.79	96.65	30.09	94.46	34.28	92.52
	+ Catalyst(μ)	37.12	91.82	52.31	87.00	13.97	97.01	31.85	93.82	33.81	92.41
	+ Catalyst(σ)	39.23	91.57	54.34	87.00	11.81	97.43	31.22	94.20	34.15	92.55
	+ Catalyst(m)	38.70	91.66	53.53	86.95	11.33	97.56	30.97	94.27	33.63	92.61
	+ Catalyst(md)	40.23	91.51	53.09	87.71	28.53	93.86	40.72	91.62	40.64	91.18
	+ Catalyst(e)	38.73	91.62	53.46	87.20	14.47	96.74	29.90	94.46	34.14	92.51

Table 23. *Detailed results of post-hoc methods combined with* Catalyst *on four OOD benchmarks: SUN, Places365, Textures, and iNaturalist using MobileNet-V2 trained on ImageNet-1K.* ↑ *indicates higher is better;* ↓ *indicates lower is better. The symbols denote the statistic used:* μ *(mean),* σ *(std. deviation),* m *(maximum),* md *(median), and* e *(Shannon entropy).*

Model	Combined Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet-18	MSP	60.39	92.40	63.69	88.37	56.74	91.32	62.66	90.10	51.87	93.64	54.63	91.87	58.33	91.28
	+ Catalyst(μ)	28.18	95.48	62.71	85.40	48.66	93.02	49.93	92.15	14.77	97.72	46.89	92.97	41.86	92.79
	+ Catalyst(σ)	11.66	97.82	54.35	90.08	30.90	95.83	25.07	96.22	4.01	98.93	31.90	95.67	26.32	95.76
	+ Catalyst(m)	10.60	97.97	51.94	90.72	25.71	96.30	22.29	96.56	3.61	99.00	27.37	96.14	23.59	96.11
	+ Catalyst(md)	97.77	44.48	86.80	56.39	87.60	62.30	95.18	42.70	90.74	51.01	85.24	65.28	90.55	53.70
	+ Catalyst(e)	58.23	93.11	62.81	89.84	55.29	92.70	61.06	91.96	48.72	94.37	53.22	93.00	56.56	92.50
	Energy	44.32	94.04	41.43	91.72	35.22	94.70	50.30	91.11	9.77	98.19	31.97	95.26	35.50	94.17
	+ Catalyst(μ)	15.73	97.32	43.65	91.25	26.26	96.08	35.98	94.25	3.20	99.26	24.26	96.30	24.85	95.74
	+ Catalyst(σ)	10.33	98.13	37.74	92.68	16.90	97.32	24.31	96.16	1.38	99.63	15.64	97.41	17.72	96.89
	+ Catalyst(m)	9.93	98.24	36.59	92.97	14.89	97.61	23.01	96.43	1.31	99.65	13.80	97.68	16.59	97.10
	+ Catalyst(md)	98.21	73.52	79.74	78.51	80.55	83.81	94.36	68.85	82.37	84.90	77.77	85.43	85.50	79.17
	+ Catalyst(e)	41.92	94.36	41.20	91.90	34.44	94.94	49.01	91.58	9.16	98.30	31.11	95.46	34.47	94.42
	ReAct	42.31	94.12	40.70	92.25	23.07	96.37	40.44	93.69	12.27	97.90	19.78	96.80	29.76	95.19
	+ Catalyst(μ)	14.37	97.48	43.18	91.46	16.71	97.22	25.04	95.82	4.26	99.13	15.70	97.36	19.88	96.41
	+ Catalyst(σ)	9.38	98.29	36.51	93.10	10.82	98.10	16.86	97.27	1.57	99.61	10.38	98.14	14.25	97.42
	+ Catalyst(m)	8.86	98.39	35.04	93.38	9.08	98.32	15.64	97.48	1.52	99.63	9.00	98.35	13.19	97.59
	+ Catalyst(md)	98.91	65.96	84.00	74.67	84.16	83.07	95.57	65.25	90.41	76.61	80.67	84.90	88.95	75.08
	+ Catalyst(e)	40.32	94.25	39.76	92.54	22.24	96.49	37.87	93.99	13.64	97.70	18.89	96.92	28.79	95.31
	DICE	17.60	97.09	46.14	90.66	39.08	94.32	44.65	91.80	1.90	99.57	36.52	94.70	30.98	94.69
	+ Catalyst(μ)	8.39	98.56	58.07	88.42	29.21	95.58	30.30	95.00	0.61	99.81	30.51	95.48	26.18	95.47
	+ Catalyst(σ)	6.39	98.88	44.70	91.48	15.07	97.40	19.04	96.86	0.41	99.89	16.09	97.28	16.95	96.96
	+ Catalyst(m)	5.98	98.96	42.31	92.01	12.18	97.82	17.02	97.17	0.36	99.90	12.91	97.69	15.13	97.26
	+ Catalyst(md)	99.11	68.64	91.50	68.78	94.43	74.02	98.37	58.86	86.39	80.37	93.52	76.03	93.89	71.12
	+ Catalyst(e)	15.85	97.30	44.64	90.97	36.25	94.74	41.95	92.43	1.60	99.61	33.86	95.08	29.02	95.02
	ReAct+DICE	11.05	98.07	47.53	91.14	17.19	97.04	24.33	95.91	1.56	99.66	16.24	97.19	19.65	96.50
	+ Catalyst(μ)	7.81	98.50	64.61	86.47	22.48	96.21	21.29	96.29	0.80	99.72	23.08	96.04	23.35	95.54
	+ Catalyst(σ)	5.41	98.98	47.06	91.11	11.54	97.85	12.55	97.83	0.38	99.88	12.91	97.64	14.98	97.21
	+ Catalyst(m)	5.13	99.03	45.17	91.44	9.74	98.14	11.33	98.00	0.39	99.88	10.58	97.94	13.72	97.41
	+ Catalyst(md)	99.48	56.14	93.84	58.15	96.45	66.78	98.76	49.11	94.49	67.60	95.53	69.35	96.42	61.19
	+ Catalyst(e)	10.23	98.21	46.14	91.57	15.65	97.29	22.27	96.32	1.30	99.69	15.00	97.41	18.43	96.75
	ASH	6.24	98.80	53.83	88.05	21.61	96.44	21.81	96.41	1.94	99.52	20.31	96.49	20.96	95.95
	+ Catalyst(μ)	5.68	98.90	62.59	83.99	24.06	95.79	20.51	96.34	2.09	99.53	23.79	95.66	23.12	95.03
	+ Catalyst(σ)	4.13	99.19	49.53	89.46	13.09	97.61	13.01	97.73	0.62	99.81	13.41	97.50	15.63	96.88
	+ Catalyst(m)	3.85	99.24	47.16	90.05	11.05	97.89	11.56	97.91	0.53	99.82	11.46	97.78	14.27	97.11
	+ Catalyst(md)	94.21	79.42	83.91	72.01	83.37	79.32	92.84	67.47	40.13	92.93	81.73	80.74	79.36	78.65
	+ Catalyst(e)	5.86	98.84	52.81	88.49	20.22	96.63	20.46	96.60	1.77	99.55	19.18	96.67	20.05	96.13
	SCALE	7.73	98.54	50.51	89.81	21.43	96.62	22.29	96.27	4.18	99.18	20.17	96.75	21.05	96.19
	+ Catalyst(μ)	6.80	98.69	58.87	86.60	22.20	96.17	21.03	96.22	3.28	99.30	21.93	96.09	22.35	95.51
	+ Catalyst(σ)	4.80	99.06	46.79	90.50	12.12	97.71	13.44	97.62	1.04	99.70	12.83	97.61	15.17	97.03
	+ Catalyst(m)	4.56	99.10	45.10	90.80	10.30	97.93	12.34	97.78	1.02	99.71	11.33	97.82	14.11	97.19
	+ Catalyst(md)	80.58	86.08	84.81	72.82	76.89	84.83	84.50	77.21	52.92	89.48	73.54	85.92	75.54	82.72
	+ Catalyst(e)	7.39	98.59	49.40	90.11	20.22	96.78	20.98	96.45	3.82	99.23	19.24	96.89	20.17	96.34

Table 24. *Detailed results of post-hoc methods combined with* `Catalyst` *on six OOD benchmarks: SVHN, Places365, iSUN, Textures, LSUN-c, and LSUN-r using ResNet-18 trained on CIFAR-10.* ↑ *indicates higher is better;* ↓ *indicates lower is better. The symbols denote the statistic used:* μ *(mean),* σ *(std. deviation),* m *(maximum),* md *(median), and* e *(Shannon entropy).*

Model	Combined Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet-18	MSP	74.26	83.20	82.37	75.31	84.13	71.57	85.04	74.02	70.79	82.78	82.96	73.10	79.92	76.66
	+ Catalyst(μ)	67.71	85.20	83.18	68.78	83.14	71.63	80.32	77.35	54.85	90.25	81.69	72.16	75.15	77.56
	+ Catalyst(σ)	56.50	90.40	82.87	70.28	80.36	79.05	71.63	84.67	48.21	92.39	78.98	78.88	69.76	82.61
	+ Catalyst(m)	55.56	90.74	82.24	71.81	79.93	79.93	71.83	84.85	46.07	92.94	78.50	79.67	69.02	83.32
	+ Catalyst(md)	81.67	67.33	83.60	66.01	87.54	52.52	87.06	62.24	67.72	80.46	85.95	54.73	82.26	63.88
	+ Catalyst(e)	72.43	84.84	82.29	74.78	83.52	75.54	83.99	78.22	68.74	85.00	82.42	76.35	78.90	79.12
	Energy	66.64	89.53	81.39	76.83	71.46	83.02	85.18	75.68	48.01	91.63	68.57	84.53	70.21	83.54
	+ Catalyst(μ)	31.13	95.02	81.53	76.00	64.83	85.24	62.06	85.32	16.45	97.17	61.59	86.00	52.93	87.46
	+ Catalyst(σ)	20.60	96.54	82.09	75.57	55.69	88.63	54.61	87.27	10.36	98.19	54.42	88.87	46.29	89.18
	+ Catalyst(m)	19.94	96.66	81.83	76.16	55.48	88.74	54.66	87.38	9.47	98.37	54.35	88.89	45.96	89.37
	+ Catalyst(md)	86.35	81.98	82.54	74.78	81.34	76.38	87.39	71.44	35.48	93.53	78.69	78.37	75.30	79.41
	+ Catalyst(e)	59.49	90.88	81.10	77.00	68.86	84.21	82.27	77.55	41.10	92.92	65.83	85.56	66.44	84.69
	ReAct	56.62	91.69	80.38	77.28	53.40	89.25	57.27	88.63	49.29	90.69	49.59	90.27	57.76	87.97
	+ Catalyst(μ)	19.43	96.65	85.03	73.97	50.51	89.41	31.76	93.30	16.52	96.91	48.33	89.70	41.93	89.99
	+ Catalyst(σ)	12.01	97.78	84.81	73.90	38.70	92.53	28.69	93.87	8.56	98.30	38.15	92.51	35.15	91.48
	+ Catalyst(m)	11.47	97.85	83.96	74.67	38.04	92.72	28.58	93.96	7.65	98.46	38.23	92.55	34.66	91.70
	+ Catalyst(md)	94.03	75.11	86.93	70.50	84.67	75.27	76.97	78.69	48.19	89.96	82.43	76.92	78.87	77.74
	+ Catalyst(e)	41.81	93.69	79.65	77.65	51.03	89.82	52.52	89.67	37.41	93.16	47.55	90.74	51.66	89.12
	DICE	40.89	92.97	81.33	76.23	62.61	85.83	75.28	76.29	12.44	97.65	61.39	86.84	55.66	85.97
	+ Catalyst(μ)	18.07	96.70	85.71	73.54	63.43	87.39	50.48	87.32	7.72	98.53	63.17	87.40	48.10	88.48
	+ Catalyst(σ)	17.98	96.65	87.65	70.66	52.29	90.21	49.82	87.43	6.52	98.77	54.88	89.98	44.86	88.95
	+ Catalyst(m)	17.13	96.73	86.35	72.10	50.76	90.77	50.04	87.40	5.63	98.91	53.29	90.42	43.87	89.39
	+ Catalyst(md)	91.36	73.12	88.97	66.71	91.18	68.13	86.83	64.19	31.66	93.47	90.88	69.56	80.15	72.53
	+ Catalyst(e)	32.64	94.27	81.06	76.22	58.24	87.49	70.66	78.98	9.96	98.11	57.36	88.25	51.65	87.22
	ReAct+DICE	34.16	94.18	83.57	74.79	54.50	89.85	52.96	87.36	10.40	97.95	53.78	90.22	48.23	89.06
	+ Catalyst(μ)	24.94	95.56	89.97	68.73	66.15	86.65	37.75	89.95	13.29	97.43	68.14	86.26	50.04	87.43
	+ Catalyst(σ)	21.45	95.77	90.46	65.11	55.33	88.99	39.57	89.35	9.84	98.00	59.02	88.56	45.95	87.63
	+ Catalyst(m)	20.28	96.00	89.36	67.31	53.37	89.95	39.68	89.69	8.55	98.26	57.50	89.37	44.79	88.43
	+ Catalyst(md)	96.81	61.79	93.08	58.01	95.51	60.58	84.40	65.93	45.31	88.46	95.54	61.27	85.11	66.00
	+ Catalyst(e)	27.53	95.18	83.83	74.35	49.97	91.02	47.41	88.90	8.75	98.25	50.10	91.19	44.60	89.82
	ASH	22.00	96.16	86.10	69.25	64.55	84.17	37.87	91.77	23.39	95.57	63.19	84.25	49.52	86.86
	+ Catalyst(μ)	17.35	96.98	83.85	72.96	64.02	85.53	40.44	91.61	18.42	96.74	61.84	85.85	47.65	88.28
	+ Catalyst(σ)	12.61	97.77	84.80	72.26	53.65	89.25	37.20	92.12	9.87	98.23	53.33	89.29	41.91	89.82
	+ Catalyst(m)	11.99	97.84	83.71	73.24	52.11	89.63	37.25	92.09	8.90	98.39	51.84	89.57	40.97	90.13
	+ Catalyst(md)	62.56	88.69	85.74	69.88	80.30	76.19	66.83	82.06	28.96	94.35	77.19	77.42	66.93	81.43
	+ Catalyst(e)	24.94	96.03	82.40	75.43	63.05	86.60	52.87	88.78	23.38	96.15	60.69	87.30	51.22	88.38
	SCALE	22.12	96.38	81.96	74.95	61.62	86.65	44.50	90.72	18.62	96.78	59.76	86.74	48.10	88.70
	+ Catalyst(μ)	18.05	96.77	86.73	69.73	62.42	85.29	36.51	91.75	15.39	97.04	62.41	84.95	46.92	87.59
	+ Catalyst(σ)	12.08	97.67	85.87	70.50	51.94	88.95	32.80	92.64	9.75	98.11	53.54	88.47	41.00	89.39
	+ Catalyst(m)	11.65	97.72	85.21	71.04	51.65	89.03	32.66	92.71	9.09	98.21	53.29	88.45	40.59	89.53
	+ Catalyst(md)	61.84	88.20	86.79	67.62	79.72	75.40	60.34	83.57	27.00	94.26	77.43	75.95	65.52	80.83
	+ Catalyst(e)	19.13	96.79	81.84	74.91	58.96	87.54	41.77	91.37	16.32	97.16	57.53	87.53	45.93	89.22

Table 25. *Detailed results of post-hoc methods combined with* `Catalyst` *on six OOD benchmarks: SVHN, Places365, iSUN, Textures, LSUN-c, and LSUN-r using ResNet-18 trained on* $\gamma(\mathbf{x})$. ↑ *indicates higher is better;* ↓ *indicates lower is better. The symbols denote the statistic used:* μ *(mean),* σ *(std. deviation),* m *(maximum),* md *(median), and* e *(Shannon entropy).*

Model	Combined Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
DenseNet-101	MSP	64.76	88.33	60.30	88.55	33.57	95.41	56.67	90.17	23.41	96.75	33.87	95.37	45.43	92.43
	+ Catalyst(μ)	29.11	95.90	54.33	86.64	10.82	98.13	30.73	93.72	1.15	99.59	12.19	97.99	23.06	95.33
	+ Catalyst(σ)	13.81	97.59	50.90	89.94	9.03	98.47	18.00	97.08	2.09	99.48	10.46	98.32	17.38	96.81
	+ Catalyst(m)	13.24	97.59	50.14	90.02	9.17	98.42	16.91	97.15	2.59	99.40	10.49	98.26	17.09	96.81
	+ Catalyst(md)	88.99	59.01	73.67	72.19	54.88	88.98	89.11	55.92	8.64	97.86	51.84	89.42	61.19	77.23
	+ Catalyst(e)	63.89	90.76	59.99	89.29	32.43	95.72	55.53	92.04	22.36	96.98	32.83	95.68	44.50	93.41
	Energy	37.91	93.59	36.42	92.38	7.33	98.27	43.87	90.48	1.95	99.47	6.97	98.38	22.41	95.43
	+ Catalyst(μ)	15.12	97.51	33.93	92.75	3.32	98.98	25.71	94.88	0.61	99.79	3.70	98.92	13.73	97.14
	+ Catalyst(σ)	10.95	98.11	32.51	92.99	1.73	99.47	18.26	96.34	0.26	99.90	1.88	99.43	10.93	97.71
	+ Catalyst(m)	10.86	98.13	31.69	93.16	1.64	99.48	17.93	96.51	0.30	99.89	1.83	99.43	10.71	97.77
	+ Catalyst(md)	82.84	79.69	63.50	84.13	41.24	94.40	85.67	72.20	3.12	99.23	38.13	94.80	52.42	87.41
	+ Catalyst(e)	36.00	93.90	35.84	92.49	6.36	98.37	42.18	90.89	1.81	99.50	6.30	98.46	21.42	95.60
	ReAct	23.18	96.28	33.96	92.97	5.56	98.49	32.23	93.98	2.47	99.33	5.37	98.59	17.13	96.61
	+ Catalyst(μ)	5.82	98.76	31.59	93.50	2.87	99.15	16.91	96.83	0.91	99.75	3.32	99.09	10.24	97.85
	+ Catalyst(σ)	5.82	98.83	30.35	93.71	1.49	99.54	11.26	97.78	0.34	99.88	1.69	99.51	8.49	98.21
	+ Catalyst(m)	5.86	98.86	29.97	93.89	1.49	99.55	11.06	97.88	0.48	99.87	1.68	99.51	8.42	98.26
	+ Catalyst(md)	85.79	78.26	65.55	82.53	49.13	92.97	88.55	69.17	3.63	99.09	45.31	93.49	56.33	85.92
	+ Catalyst(e)	21.19	96.49	33.40	93.11	5.06	98.59	30.64	94.32	2.30	99.38	4.99	98.68	16.26	96.76
	DICE	16.66	96.98	37.59	92.04	2.31	99.42	27.98	92.71	0.15	99.94	2.44	99.36	14.52	96.74
	+ Catalyst(μ)	6.23	98.80	44.96	91.11	2.35	99.30	20.18	95.57	0.07	99.94	2.67	99.22	12.74	97.32
	+ Catalyst(σ)	5.33	98.95	41.71	91.84	1.88	99.48	15.28	96.95	0.09	99.95	2.06	99.43	11.06	97.77
	+ Catalyst(m)	4.95	99.01	39.57	92.20	1.58	99.53	14.08	97.18	0.11	99.95	1.88	99.48	10.36	97.89
	+ Catalyst(md)	83.46	75.99	74.72	78.40	53.58	92.11	86.81	67.01	2.24	99.42	53.74	92.19	59.09	84.19
	+ Catalyst(e)	15.57	97.17	37.28	92.17	2.15	99.44	26.81	93.08	0.14	99.95	2.27	99.39	14.04	96.86
	ReAct+DICE	4.60	99.02	35.94	92.91	1.78	99.51	17.07	96.78	0.12	99.95	2.02	99.47	10.26	97.94
	+ Catalyst(μ)	2.68	99.30	48.65	90.81	2.76	99.20	14.27	97.29	0.09	99.93	3.19	99.13	11.94	97.61
	+ Catalyst(σ)	3.78	99.21	43.99	91.70	2.02	99.44	10.80	98.01	0.11	99.93	2.38	99.39	10.51	97.95
	+ Catalyst(m)	3.59	99.24	42.01	92.03	1.83	99.49	9.72	98.17	0.15	99.93	2.24	99.45	9.92	98.05
	+ Catalyst(md)	85.94	75.87	78.12	75.30	57.04	91.52	88.16	65.34	2.45	99.38	56.79	91.52	61.42	83.16
	+ Catalyst(e)	4.49	99.06	35.95	93.02	1.73	99.52	16.38	96.99	0.13	99.95	1.99	99.49	10.11	98.01
	ASH	5.18	98.90	42.80	90.42	2.97	99.27	15.80	97.04	0.45	99.80	3.06	99.25	11.71	97.44
	+ Catalyst(μ)	7.02	98.61	38.76	91.80	2.72	99.22	17.98	96.62	0.29	99.85	3.18	99.15	11.66	97.54
	+ Catalyst(σ)	5.31	98.91	35.92	92.56	1.70	99.48	12.85	97.61	0.26	99.89	1.92	99.44	9.66	97.98
	+ Catalyst(m)	5.23	98.95	35.13	92.77	1.56	99.50	12.32	97.73	0.32	99.89	1.89	99.46	9.41	98.05
	+ Catalyst(md)	68.46	84.08	63.48	83.16	31.93	95.40	76.86	75.56	1.83	99.56	30.24	95.54	45.47	88.88
	+ Catalyst(e)	4.85	98.93	42.53	90.54	2.86	99.29	15.21	97.15	0.44	99.80	3.01	99.27	11.48	97.50
	SCALE	29.23	95.23	37.86	92.14	6.71	98.46	36.99	92.28	1.71	99.50	6.80	98.48	19.88	96.01
	+ Catalyst(μ)	11.85	97.95	35.98	92.36	3.68	98.94	22.91	95.56	0.65	99.75	3.94	98.86	13.17	97.23
	+ Catalyst(σ)	8.99	98.36	34.54	92.59	1.88	99.38	16.74	96.74	0.32	99.87	2.17	99.32	10.77	97.71
	+ Catalyst(m)	8.96	98.38	34.02	92.76	1.87	99.38	16.47	96.89	0.38	99.86	2.23	99.32	10.66	97.76
	+ Catalyst(md)	69.80	84.20	54.23	87.24	23.01	96.29	75.48	76.37	1.76	99.51	18.90	96.73	40.53	90.06
	+ Catalyst(e)	28.81	95.31	37.44	92.21	6.51	98.48	36.60	92.41	1.63	99.51	6.61	98.50	19.60	96.07

Table 26. *Detailed results of post-hoc methods combined with* `Catalyst` *on six OOD benchmarks: SVHN, Places365, iSUN, Textures, LSUN-c, and LSUN-r using DenseNet-101 trained on CIFAR-10.* ↑ *indicates higher is better;* ↓ *indicates lower is better. The symbols denote the statistic used:* μ *(mean),* σ *(std. deviation),* m *(maximum),* `md` *(median), and* e *(Shannon entropy).*

Model	Combined Method	SVHN		Place365		iSUN		Textures		LSUN-c		LSUN-r		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
DenseNet-101	MSP	81.38	75.71	82.68	74.06	82.52	70.50	87.11	68.39	51.82	87.93	79.31	72.21	77.47	74.80
	+ Catalyst(μ)	67.10	85.98	82.63	72.87	76.61	77.17	73.87	80.77	22.08	96.58	74.63	77.39	66.15	81.79
	Catalyst(σ)	61.17	88.41	82.09	73.45	73.98	80.04	64.43	86.18	22.22	96.42	72.44	79.91	62.72	84.07
	+ Catalyst(m)	60.15	88.82	81.77	74.22	73.23	80.90	62.82	86.96	22.51	96.35	71.72	80.72	62.03	84.66
	+ Catalyst(md)	94.34	47.23	86.86	66.68	93.52	45.23	97.41	30.76	40.44	90.25	90.66	50.25	83.87	55.07
	+ Catalyst(e)	78.98	79.30	82.44	74.49	81.08	74.98	84.65	75.59	46.99	90.11	77.79	75.99	75.32	78.41
	Energy	70.99	86.66	77.28	76.94	59.39	85.68	83.49	67.47	11.45	97.89	50.90	88.57	58.92	83.87
	+ Catalyst(μ)	22.45	96.11	78.72	77.16	48.77	89.75	52.09	83.58	1.54	99.68	44.92	90.44	41.42	89.45
	+ Catalyst(σ)	21.13	96.30	78.19	77.16	42.78	91.48	44.34	86.19	1.18	99.72	40.25	92.02	37.98	90.48
	+ Catalyst(m)	19.90	96.45	77.30	77.67	41.02	91.81	42.48	87.12	1.28	99.70	38.78	92.25	36.79	90.83
	+ Catalyst(md)	98.09	53.27	88.48	68.06	96.62	55.42	99.40	30.52	20.27	95.70	93.82	61.85	82.78	60.81
	+ Catalyst(e)	58.57	89.86	76.92	77.59	53.03	88.03	76.86	72.19	7.68	98.66	45.28	90.32	53.06	86.11
	ReAct	69.82	86.30	79.23	74.09	41.50	92.40	72.09	80.38	18.14	96.26	36.53	93.64	52.89	87.18
	+ Catalyst(μ)	11.73	97.67	83.17	74.06	25.66	95.16	26.12	93.93	1.69	99.52	27.77	94.98	29.36	92.56
	+ Catalyst(σ)	14.13	97.32	83.87	74.36	25.15	95.51	23.21	94.55	1.55	99.55	26.41	95.37	29.05	92.78
	+ Catalyst(m)	13.70	97.36	83.00	75.14	23.26	95.83	21.68	94.95	2.07	99.44	24.63	95.64	28.06	93.06
	+ Catalyst(md)	99.29	36.35	92.19	60.93	98.68	41.22	99.70	19.61	41.86	91.88	97.16	47.82	88.15	49.63
	+ Catalyst(e)	46.52	91.94	78.18	75.18	25.51	95.11	49.66	87.42	10.23	97.95	23.32	95.79	38.90	90.56
	DICE	32.93	94.09	79.90	75.43	35.50	92.50	64.84	71.95	1.93	99.57	30.81	93.96	40.98	87.92
	+ Catalyst(μ)	16.64	96.95	84.89	74.34	33.11	94.02	46.44	84.49	1.14	99.65	32.25	94.27	35.74	90.62
	Catalyst(σ)	17.27	96.77	83.48	74.87	32.32	93.97	48.19	83.15	1.15	99.67	30.98	94.38	35.57	90.47
	+ Catalyst(m)	17.95	96.62	82.44	75.04	32.04	93.84	48.97	82.26	1.11	99.67	30.23	94.34	35.46	90.30
	+ Catalyst(md)	97.75	56.78	90.67	65.03	97.62	54.76	99.27	29.22	9.86	97.71	95.27	60.53	81.74	60.67
	+ Catalyst(e)	25.98	95.41	79.38	77.04	37.18	92.38	56.77	77.17	0.80	99.75	33.23	93.48	38.89	89.21
	ReAct+DICE	25.10	95.70	84.17	73.56	27.98	95.06	41.79	87.82	1.06	99.70	27.76	95.16	34.64	91.17
	+ Catalyst(μ)	16.40	96.93	86.87	72.82	31.87	94.55	31.65	92.18	1.41	99.56	33.77	94.34	33.66	91.73
	Catalyst(σ)	17.69	96.80	84.49	74.41	30.06	94.83	33.94	91.27	0.97	99.69	30.73	94.84	32.98	91.97
	+ Catalyst(m)	17.50	96.83	84.73	74.34	30.34	94.81	33.76	91.33	0.99	99.68	31.03	94.80	33.06	91.97
	+ Catalyst(md)	98.37	49.23	92.03	60.76	98.36	45.84	99.40	25.45	11.36	97.48	96.42	51.22	82.66	55.00
	+ Catalyst(e)	23.39	96.01	83.50	74.62	28.59	95.03	40.64	89.05	1.06	99.73	28.83	95.13	34.34	91.60
	ASH	10.32	97.99	85.80	71.97	37.68	92.45	35.48	91.77	5.43	98.98	40.35	91.96	35.84	90.85
	+ Catalyst(μ)	9.00	98.12	87.63	70.70	38.40	92.42	27.70	93.91	5.38	98.94	42.25	91.72	35.06	90.97
	Catalyst(σ)	8.82	98.16	86.33	71.57	37.41	92.38	29.27	93.53	5.43	98.95	40.76	91.78	34.67	91.06
	+ Catalyst(m)	10.63	97.85	87.59	70.89	38.11	92.67	26.45	94.16	4.40	99.09	42.01	92.02	34.87	91.11
	+ Catalyst(md)	67.65	85.31	90.45	67.10	89.48	71.59	90.98	59.95	4.17	99.08	86.53	73.09	71.54	76.02
	+ Catalyst(e)	9.50	98.08	85.59	71.97	35.82	92.84	32.34	92.69	4.89	99.07	38.53	92.33	34.45	91.16
	SCALE	16.26	97.05	78.54	76.97	43.56	91.21	45.60	87.23	3.23	99.30	42.69	91.02	38.31	90.46
	+ Catalyst(μ)	10.37	97.96	83.56	74.76	42.16	91.85	33.35	92.48	1.84	99.52	44.44	91.07	35.95	91.27
	+ Catalyst(σ)	11.08	97.85	83.49	74.93	38.60	92.77	30.53	93.09	1.58	99.57	41.10	92.17	34.40	91.73
	+ Catalyst(m)	10.79	97.90	83.16	75.44	37.69	93.03	29.08	93.56	1.89	99.52	40.53	92.39	33.86	91.97
	+ Catalyst(md)	82.79	78.26	87.59	70.09	92.87	64.90	95.57	50.40	5.80	98.79	89.44	68.02	75.68	71.75
	+ Catalyst(e)	14.85	97.27	78.26	77.29	41.02	91.93	42.84	88.33	2.86	99.37	40.71	91.70	36.76	90.98

Table 27. *Detailed results of post-hoc methods combined with* `Catalyst` *on six OOD benchmarks: SVHN, Places365, iSUN, Textures, LSUN-c, and LSUN-r using DenseNet-101 trained on CIFAR-100.* ↑ *indicates higher is better;* ↓ *indicates lower is better. The symbols denote the statistic used:* μ *(mean),* σ *(std. deviation),* m *(maximum),* `md` *(median), and* e *(Shannon entropy).*