ATS: Adaptive Temperature Scaling for Enhancing Out-of-Distribution Detection Methods

Gerhard Krumpl [†1,2] Henning Avenhaus[2] Horst Possegger[1] Horst Bischof[1]

[1]Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[2]KESTRELEYE GmbH, Austria

Abstract

Out-of-distribution (OOD) detection is essential to ensure the reliability and robustness of machine learning models in real-world applications. Post-hoc OOD detection methods have gained significant attention due to the fact that they offer the advantage of not requiring additional re-training, which could degrade model performance and increase training time. However, most existing post-hoc methods rely only on the encoder output (features), logits, or the softmax probability, meaning they have no access to information that might be lost in the feature extraction process. In this work, we address this limitation by introducing Adaptive Temperature Scaling (ATS), a novel approach that dynamically calculates a temperature value based on activations of the intermediate layers. Fusing this sample-specific adjustment with class-dependent logits, our ATS captures additional statistical information before they are lost in the feature extraction process, leading to a more robust and powerful OOD detection method. We conduct extensive experiments to demonstrate the efficacy of our approach. Notably, our method can be seamlessly combined with SOTA post-hoc OOD detection methods that rely on the logits, thereby enhancing their performance and improving their robustness.

1. Introduction

Applying deep learning-based neural networks in real-world settings, particularly in safety-critical domains, is often hindered by a significant challenge: the model must handle object classes or scenarios that were not seen during training. For example, an autonomous car could experience a fire brigade operation or a sorting system for food could be confronted with metal pieces that accidentally entered the sorting stream, but also effects such as dirty lenses, sensor failures, or adverse weather conditions can negatively impact the model performance. These samples, known as out-of-distribution (OOD) data, have the potential to cause problems due to misclassification and erode the trustworthiness of ML models. Consequently, the detection and handling of OOD samples are crucial in extending the applicability of ML models to real-world settings [13, 16, 25].

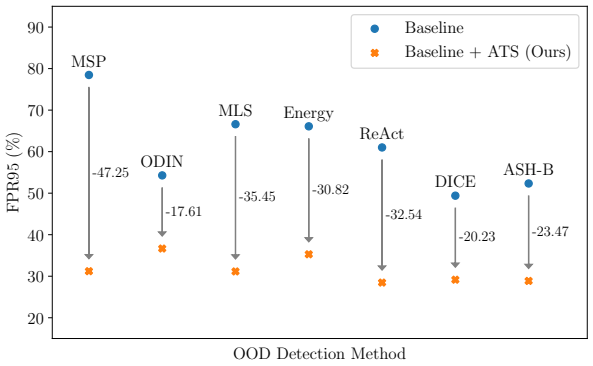

Figure 1. Improvement of the average FPR95 metric (false positive rate at 95% recall) for seven post-hoc OOD detection methods (MSP [14], ODIN [23], MLS [12], Energy [24], ReAct [32], DICE [34], and ASH-B [6]) with the proposed Adaptive Temperature Scaling (ATS) for ResNet18 trained on CIFAR-100.

A simple and desirable solution would be if trained neural networks (NNs) for classification tasks would exhibit high uncertainty (uniform activation over all classes, low confidence) when presented with samples that do not belong to the training distribution. Unfortunately, this is typically not the case, even for inputs that are completely unrecognizable [29] or irrelevant [26]. The inability of NNs to reliably detect and handle OOD samples in this way means that other methods for OOD detection must be developed, which is akin to adding an additional binary 'OOD' flag to the network output.

OOD detection methods should ideally fulfill two properties: i) They should be post-hoc, meaning they do not require re-training of the NN, which would incur additional training costs and could furthermore lead to performance degradation of the model for in-distribution (ID)

[†]Correspondence: gerhard.krumpl@icg.tugraz.at

samples [45], ii) They should be sample-free, *i.e*., not require OOD samples to be calibrated, as a typical feature of out-of-distribution samples is that they are not previously known (otherwise they could have been considered during the training process of the original network).

Current OOD detection methods fulfilling these two criteria typically utilize the softmax probability [14], logits [12, 24], or features from the penultimate layer [35] to compute a score that allows distinguishing between ID and OOD samples. However, these methods essentially assume that a single layer at the end of the neural network contains sufficient information to effectively detect OOD samples, despite their diverse range of patterns, characteristics, and properties. More recent works have experimented with using one or multiple intermediate layers to derive a score for OOD detection, *e.g*. [10, 31, 40, 48], omitting the final layer of the network. This comes with the risk of overlooking valuable cues, as only the logits of the network hold class-specific information.

We address these limitations with **A**daptive **T**emperature **S**caling (ATS), a method that leverages both class-agnostic and class-dependent information from intermediate layers and from the model output, respectively, to enhance OOD detection capabilities. To this end, ATS extracts a sample-specific temperature parameter based on the intermediate layer activation enabling temperature scaling and effectively combining information across the entire NN. ATS overcomes the limitation of relying on i) a fixed temperature parameter derived from OOD data [9, 23], ii) a single information source [12, 14, 24], or ii) intermediate-layer information without class-specific cues [10, 31, 40]. ATS can be applied to all SOTA post-hoc methods that derive their score for OOD detection from the model output to improve their performance and make the method more robust, as can be seen by the improved false positive rate in Fig. 1.

Our core contributions are threefold:

- We propose a novel OOD detection method utilizing intermediate feature maps and temperature scaling to enhance state-of-the-art OOD detectors and make them more robust.
- Our findings reveal the significance of features from all network layers, demonstrating that the sensitivity to OOD data varies across different layers and types of OOD data.
- We conduct extensive evaluations on widely used datasets (including 3 ID datasets and 11 OOD datasets), verifying the benefits of our ATS approach.

2. Related Work

The field of out-of-distribution (OOD) detection has seen a surge in research interest in recent years, leading to the development of various methodologies, *e.g*., classification-based [12,14,24,38], distance-based [22,30,35,36], density-based [1, 22, 27, 51], and reconstruction-based methods [5, 46, 50]. For a detailed survey, we refer the interested reader to [45]. In the following, we emphasize the major concepts and notable contributions of post-hoc and sample-free OOD detection methods.

Hendrycks *et al*. [14] introduced the maximum softmax probability (MSP) as an initial baseline for OOD detection, where the maximum softmax probability is used as the detection score. Building upon this, ODIN [23] improved the MSP score by incorporating fixed temperature scaling and input perturbation. Furthermore, other approaches have leveraged the logits for OOD detection. Hendrycks *et al*. [12] introduced the maximum logit score (MLS) and KL-Matching, utilizing the minimum KL-divergence between the softmax output and the mean class-conditional distribution as a score to distinguish between ID and OOD samples. On the other hand, Liu *et al*. [24] introduced the energy score, calculated based on the logits, to effectively identify OOD samples. ReAct [32] showed that the activation pattern of ID and OOD samples in the penultimate layer differs and improved the energy score by activation clipping. DICE [34] and ASH [6] further improved the energy score by weight and activation sparsification, respectively. KNN [35] analyses the nearest-neighbor distance in the embedding space (penultimate layer) for OOD detection. ViM [39] combines information from the feature space and the logits to preserve class-agnostic and class-dependent features, respectively. All these OOD detection methods utilize either the penultimate layer, the model output, or both to calculate their OOD score and thus neglect information from shallow layers.

In order to improve the robustness of OOD detection to the diverse patterns, features, and properties of OOD samples, recent research has explored the utilization of intermediate layers. Two approaches have emerged: (i) selecting the optimal layer [48] and (ii) aggregating information from multiple layers [7, 22, 31, 40].

Yu *et al*. [48] proposed a method for selecting the optimal layer for OOD detection based on the ID dataset and deriving the OOD score from that layer. The OOD score is based on the average L2-norm of the layer activation, and the optimal layer is selected based on the ratio of ID and synthetic OOD data. Methods such as [7] use feature maps from intermediate layers as input for an auxiliary OOD classifier. The usage of an auxiliary model implies that training on OOD or synthetic OOD data is required. Methods that utilize multiple intermediate layers typically calculate scores for each layer and aggregate the layer-specific scores to distinguish between ID and OOD samples [10, 22, 31]. Methods such as GRAM [31] and MDS [22] utilize the Gram matrix and the Mahalanobis distance from class cen-

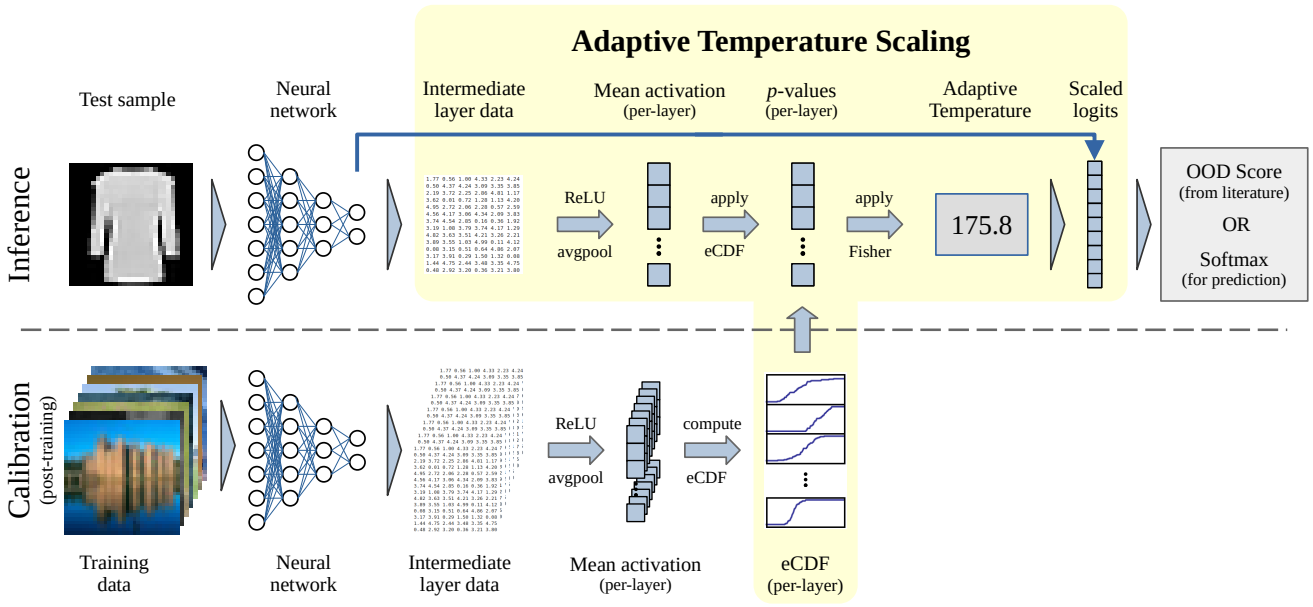

Figure 2. Overview of our Adaptive Temperature Scaling (ATS) method: At test-time (inference), ATS utilizes a per-sample specific temperature derived from the intermediate layer activation to scale the logits. The per-sample temperature uses the empirical cumulative distribution function (eCDF), which is precomputed on the training set in the calibration phase. ATS can be seamlessly combined with various OOD detection methods that leverage their OOD score from the logits, effectively enhancing the distinguishability between in-distribution (ID) and out-of-distribution (OOD) samples.

troids to obtain a per-layer score, respectively, to handle OOD samples. Haroush [10] formulated the OOD detection as statistical hypothesis testing considering intermediate layers. One different approach is HDFF [40], which uses techniques from hyperdimensional computing to derive a score for OOD detection. All these approaches derive their scores from intermediate layer activations only, not considering the class-dependent logits.

In contrast, our approach distinguishes itself by leveraging both class-agnostic features from multiple intermediate layers (shallow to deep) and class-dependent information from the model output.

3. ATS: Adaptive Temperature Scaling

Our approach for OOD detection, Adaptive Temperature Scaling (ATS), is outlined in Figure 2. The core idea is to perform temperature scaling of the logits (Sec. 3.1) using a sample-specific temperature derived from the intermediate layer activations (Sec. 3.2). The temperature extraction step is designed to return low temperatures for in-distribution samples while returning high temperatures for out-of-distribution samples. After that, any scoring function that is based on the model output can be used to distinguish between in-distribution (ID) and out-of-distribution (OOD) samples (Sec. 3.3).

Definitions & Notations. We consider OOD detection for an image classification model. Let $\mathcal{X}$ be the input space (typically, $\mathcal{X} = \mathbb{R}^{C\times H\times W}$) and $\mathcal{Y} = \{\text{class}_1, ..., \text{class}_K\}$ be the output space of a supervised classification problem and let furthermore $\mathcal{X}_{ID} \subset \mathcal{X}$ be the space of in-distribution samples (the space of OOD samples is the complement of the ID subspace, $\mathcal{X}_{OOD} = \mathcal{X}'_{ID}$). We have a deep neural network, denoted as $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$, which has been trained in a supervised manner on the training dataset $\mathcal{D}_{\text{in}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. The samples in $\mathcal{D}_{\text{in}}$ are drawn from the joint data space $\mathcal{X}_{ID} \times \mathcal{Y}$.

Problem Statement. Ideally, a deep NN should know what it does not know when it is deployed in the real world, such that it is able to warn if a given sample is outside the distribution it was trained on, on top of correctly classifying ID samples. Consequently, when deployed, the model's objective is twofold: i) accurately classifying ID samples and ii) correctly identifying OOD samples. The OOD detection is a binary classification problem with a scoring function that describes how likely a given sample $\mathbf{x}$ is from the ID space, $\mathbf{x} \in \mathcal{X}_{ID}$. The main objective in OOD detection research is to develop a scoring function that effectively distinguishes between ID and OOD samples.

3.1. Temperature Scaling

Typically, confidence values for the different classes are calculated using the softmax probability, where the softmax $S_i(\mathbf{x})$ is defined as

$$S_i(\mathbf{x}) = \frac{\exp(f_i(\mathbf{x}))}{\sum_{k=1}^{K} \exp(f_k(\mathbf{x}))}. \quad (1)$$

Temperature scaling scales the logits before applying the softmax by $1/T$, such that

$$S_i(\mathbf{x}, T) = \frac{\exp(f_i(\mathbf{x})/T)}{\sum_{k=1}^{K} \exp(f_k(\mathbf{x})/T)}, \quad (2)$$

where T is the temperature parameter. Prior works have established temperature scaling for knowledge distillation [17], confidence calibration [9], and also OOD detection [23]. ODIN [23] utilized a fixed temperature derived from an OOD validation dataset in combination with input perturbation to improve OOD detection. However, this fixed scaling does not generalize well across different OOD data sources, considering their diverse characteristics.

3.2. Adaptive Temperature Scaling

Our key concept is to calculate a sample-specific temperature. A temperature that adapts to the input sample by utilizing information from multiple intermediate layers improves the robustness of a neural network. This concept is based on the idea that intermediate layer activations can be a good indicator for OOD detection and should thus be able to outperform a pre-computed, fixed temperature. It should also generalize well because it is able to dynamically adapt to new (previously unseen) types or examples of OOD data. We denote the activation response of the l-th layer of the classification model f for a given input $\mathbf{x}$ by $\mathbf{z}_l \in \mathbb{R}^{\mathrm{C}_l \times \mathrm{H}_l \times \mathrm{W}_l}$, where C_l, H_l, and W_l refer to the number of channels, height, and width, respectively. The mean layer activation of a given feature map $\mathbf{z}_l$ is calculated as

$$\mu_l(\mathbf{x}) = \frac{1}{C_l H_l W_l} \sum_{c}^{C_l} \sum_{h}^{H_l} \sum_{w}^{W_l} \max(\mathbf{z}_l(c, h, w), 0), \quad (3)$$

where $\mathbf{z}_l(\mathrm{c}, \mathrm{h}, \mathrm{w})$ is the c-th, h-th, w-th element of the given feature map. In other words, μ_l is the mean over all values, ignoring negative values, and represents the mean activation of the l-th layer after ReLU activation.

We then estimate how likely it is that a mean layer activation $\mu_l(\mathbf{x})$ is drawn from the same distribution as the corresponding mean activation from the training dataset. To do so, we pre-calculate the empirical cumulative distribution function (eCDF) from the in-distribution training dataset $\mathcal{D}_{\text{in}}$ as

$$\hat{F}_{l,\mathcal{D}_{\text{in}}}(\tau) = \frac{1}{N} \sum_{\mathbf{x}_i \in \mathcal{D}_{\text{in}}} \mathbb{1}_{\mu_l(\mathbf{x}_i) \leq \tau}, \quad (4)$$

where $\mathbb{1}$ is the indicator function. Pre-computing the eCDF is done in an initial, post-training calibration phase, as illustrated in Figure 2.

During testing, we apply this eCDF to calculate the p-value p_l from the mean layer activation $\mu_l(\mathbf{x})$ of the given sample $\mathbf{x}$ by:

$$p_l(\mathbf{x}) = 2 \min\Big(\hat{F}_{l,\mathcal{D}_{\text{in}}}(\mu_l(\mathbf{x})),\, 1 - \hat{F}_{l,\mathcal{D}_{\text{in}}}(\mu_l(\mathbf{x}))\Big), \quad (5)$$

where we perform a two-sided test (giving rise to the leading constant of 2) since outliers can be on both sides of the mean layer activation. Using the eCDF to transform the mean layer activation μ_l is important since all layers have different ranges, and otherwise, some layers would dominate the temperature calculation.

The sample-specific temperature $\hat{T}(\mathbf{x})$ is determined by aggregating all layers via Fisher's method [8]:

$$\hat{T}(\mathbf{x}) = -2 \sum_{l}^{L} \log(p_l(\mathbf{x})). \quad (6)$$

This results in low temperature values for ID and high temperature values for OOD samples.

3.3. Out-of-Distribution Detector with ATS

After the logits have been scaled with the sample-specific temperature $\hat{T}(\mathbf{x})$, we can apply any scoring function $G(f(\mathbf{x}))$ that relies on the logits $f(\mathbf{x})$ (or, by extension, the softmax probabilities) in order to perform the OOD detection. The OOD detector H is then defined as

$$H(\mathbf{x}, \lambda) = \begin{cases} \text{ID} & \text{if } G(f(\mathbf{x}) \,/\, \hat{T}(\mathbf{x})) \geq \lambda \\ \text{OOD} & \text{otherwise,} \end{cases} \quad (7)$$

where λ is a threshold parameter to distinguish between ID and OOD samples.

The threshold is typically chosen such that a high fraction of ID data (*e.g.*, 95%) is correctly classified. Given that the classification of a sample is calculated using the *argmax* function and the *softmax* function does not change the relative ordering of values in the logits, the network's classification performance is unaffected. ATS can be combined with all state-of-the-art methods that rely on the final output of the model (logits or softmax probabilities), such as MLS [12], ReAct [32], DICE [34], and ASH [6]. We do not require re-training the model (post-hoc), prior knowledge of OOD data (sample-free), or any hyperparameter tuning. Furthermore, temperature scaling does not impact the classification performance of ID samples. These characteristics make ATS highly applicable for real-world applications.

4. Experiments

Comprehensive experiments have been conducted to evaluate our method. Detailed information regarding the

experimental setup can be found in Sec. 4.1. In Sec. 4.2, we use CIFAR [20] and also the large-scale dataset based on the ImageNet benchmark, commonly used in the literature [15, 24, 33] and various OOD datasets. Sec. 4.3 provides an in-depth analysis of why our method proves to be effective and valuable in the context of OOD detection.

4.1. Experimental Setup

Datasets. The conventional approach for constructing OOD detection benchmarks involves designating a complete dataset as the in-distribution (ID) dataset and gathering several datasets that are unrelated to any ID categories as OOD datasets [44]. We conducted our studies on three ID datasets (CIFAR-10/100 [20] and ImageNet [4]) and eleven OOD datasets (SVHN [28], Textures [3], iSUN [43], LSUN [47], LSUN-Crop [47], Places365 [49], MNIST [21], Fashion-MNIST [41], iNaturalist [37], SUN [42], and NINCO [2]).

Setup for CIFAR. We evaluate CIFAR-10 and CIFAR-100 [20] as ID datasets using the default split with 50000 training and 10000 test images, respectively. We consider eight commonly used benchmark datasets: SVHN [28], Textures [3], iSUN [43], LSUN [47], LSUN-Crop [47], Places365 [49], MNIST [21], and Fashion-MNIST [41]. We test on two standard architectures: i) ResNet18 [11] and ii) DenseNet-101 [18]. Both models are trained with a batch size of 64 for 100 epochs. The initial learning rate is 0.1 and decreases by a factor of 1×10^{-1} at 50, 75, and 90 training epochs. We use the SGD optimizer with a momentum of 0.9 and a decay rate of 1×10^{-4}. Parameters for ReAct [32], DICE [34], and ATS (our approach) are pre-computed from the entire training set.

Setup for ImageNet. We also evaluate our method on the large-scale ImageNet [4] dataset. The OOD datasets used are based on [19], who provided a subset of the following four datasets where all the overlapping categories with ImageNet-1k are removed. The four subsets are: iNaturalist [37], SUN [42], Places365 [49], and Textures [3]. We additionally added the NINCO [2] dataset, which contains 5879 manually selected images with no overlap with the ImageNet-1k dataset, as well as Fashion-MNIST [41] as an extreme OOD dataset. We use a ResNet-50 [11], trained on ImageNet-1k. The trained weights are provided by Pytorch. To pre-compute training set statistics, 200000 images are randomly sampled from the training set.

Intermediate layer selection. We uniformly select intermediate layers across the entire network. This eliminates the need for tuning the layer selection for either the ID dataset, the OOD data, or the network architecture, ensuring a simple and consistent process that does not require manual adjustments of parameters relating to layer selection.

Evaluation metric. Following the common protocol, *e.g.*, [6, 12, 14, 32, 34], we report the false positive rate at recall 95% (FPR95) and the area under the receiver operating characteristic curve (AUROC).

Test time and evaluation. At test time, all images are resized to 32×32 for CIFAR [20]. For ImageNet, all images are resized to 256×256 and center cropped to 224×224 For evaluation, we apply our approach to recent post-hoc OOD detection methods: MSP [14], ODIN [23], MLS [12], Energy [24], ReAct [32], DICE [34], and ASH-B [6] and compare the performance with and without ATS.

4.2. Results

CIFAR evaluation. In Table 1, we report the performance of OOD detection methods over eight OOD datasets with and without ATS for a ResNet18 trained on CIFAR-100 [20]. As shown in the table, our method reduces the FPR95 of the baseline methods on average by 29.62%. We also see that ATS improves the performance of all baseline methods on seven out of eight OOD datasets. For Places365 [49], we are not able to extract useful information from the intermediate layers, see additional investigations in Sec. 4.3. The best method with ATS (ReAcT+) provides, on average, an FPR95 that is 20.93% lower than the best non-ATS baseline method (DICE [34]). When enhanced with ATS, the baseline methods exhibit a noteworthy improvement in performance, particularly when detecting far-OOD data such as MNIST [21] or SVHN [28].

Table 2 presents the average performance across various OOD datasets, including SVHN [28], Textures [3], iSUN [43], LSUN [47], LSUN-Crop [47], Places365 [49], MNIST [21]. The evaluation is based on CIFAR-10 and CIFAR-100 [20] as the ID dataset, using ResNet18 and DenseNet as the backbone architectures. The performance increase on CIFAR [20] is consistent across multiple existing methods and network architectures we considered.

ImageNet evaluation. Table 3 presents the performance of the baseline methods, both with and without ATS, on the ImageNet benchmark. Remarkably, ATS reduces the FPR95 of the baseline methods on average by 7.91%. Specifically, methods such as MSP [14], ODIN [23], and MLS [12] benefit significantly from ATS, as they exhibit relatively lower baseline performance. Even for methods that already achieve a very strong baseline performance, *i.e.*, ReAct [32], DICE [34], and ASH [6], our ATS can still improve the results. We attribute the slightly less significant improvement on high baselines to two key factors:

Method	OOD-Dataset																Average	
	SVHN		Textures		iSUN		LSUN		LSUN-Crop		Places365		MNIST		fMNIST			
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MSP	73.74	84.24	73.57	85.34	74.15	83.00	74.32	81.88	83.11	68.81	75.33	82.28	75.02	81.56	87.30	60.59	77.07	78.46
MSP+	94.55	28.33	87.76	56.01	93.18	34.12	93.96	32.60	99.47	2.21	71.05	86.98	98.83	6.23	99.23	3.22	92.25	31.21
ODIN	88.99	58.27	73.32	85.53	83.90	74.44	83.22	75.83	94.52	30.04	73.33	86.13	98.24	6.82	97.07	17.21	86.57	54.28
ODIN+	94.66	28.50	83.68	65.83	91.35	45.97	91.87	46.60	98.60	8.34	64.01	91.85	**99.62**	**0.17**	98.77	6.09	90.32	36.67
MLS	80.77	82.17	74.10	87.23	82.66	73.10	82.90	71.94	93.49	39.36	78.26	80.15	86.34	65.86	95.02	33.04	84.19	66.61
MLS+	94.55	28.22	87.75	55.96	93.18	34.04	93.97	32.51	**99.47**	**2.20**	70.95	87.01	98.84	6.15	**99.24**	**3.15**	92.24	31.15
Energy	80.90	83.02	73.89	88.26	82.99	72.94	83.24	71.45	94.05	36.18	78.21	80.99	86.89	64.74	95.38	31.20	84.45	66.10
Energy+	91.53	35.64	81.62	65.51	90.73	39.17	92.02	37.24	99.33	3.17	56.77	93.93	99.31	1.44	98.64	6.14	88.74	35.28
ReAct	83.98	80.35	87.82	62.13	87.96	60.96	87.17	62.36	93.32	37.00	78.43	80.55	86.75	67.96	94.42	36.76	87.48	61.01
ReAct+	95.13	25.05	**91.74**	**41.42**	93.46	31.65	94.15	29.73	99.29	3.18	67.20	89.29	99.11	3.24	98.99	4.21	92.39	**28.47**
DICE	82.17	71.04	74.40	81.91	85.69	65.96	86.09	65.57	98.39	8.59	**79.26**	**76.22**	97.26	14.12	97.93	11.75	87.65	49.40
DICE+	94.20	30.41	87.33	56.26	**94.43**	**27.78**	**95.36**	**24.59**	99.45	2.47	73.01	84.58	99.24	3.54	99.18	3.68	**92.78**	29.16
ASH-B	87.64	60.00	86.64	62.50	85.48	66.30	85.20	66.31	96.70	19.90	75.89	82.54	93.94	36.59	96.10	24.55	88.45	52.34
ASH-B+	**95.65**	**22.72**	91.17	43.63	92.77	34.34	93.64	32.73	99.44	2.50	66.86	89.75	99.44	1.38	99.07	3.90	92.25	28.87

Table 1. Performance of OOD detection methods for a ResNet18 backbone trained on CIFAR-100. *Method*+ denotes that our ATS is applied on top of the method (*i.e.*, rows with gray background). ↑/↓ indicate that larger/smaller values are better. The **best** and second-best results for each OOD dataset (*i.e.*, each column) are shown in bold or underlined, respectively. All values are reported as percentages.

First, the relevant information from the images (both ID and OOD) is only becoming available (through extraction) in the later layers of the network, making the information in the intermediate layers less relevant for OOD detection (see Sec. 4.3). Second, the standard benchmark datasets such as iNaturalist [37], Places [49], and Textures [3] overlap with ImageNet [4], as pointed out by the authors of NINCO [2]. Our findings demonstrate that utilizing information from the intermediate layer can also enhance performance on large-scale datasets like ImageNet. Notably, the improved robustness against far-OOD samples is particularly pronounced, showcasing the efficacy of incorporating information from several layers in OOD detection methods.

4.3. Ablation Study

Intermediate layer selection. Figure 3 shows the importance of intermediate layers in the temperature calculation of a ResNet18 model trained on CIFAR-100 [20]. We evaluate the performance using the first l layers (left plot of Fig. 3) and the last l layers (right plot of Fig. 3) for the sample-specific temperature calculation. Our findings demonstrate that the optimal layer configuration varies depending on the specific OOD dataset. Detecting outliers from SVHN [28] or iSUN [43] is best done using shallow-layer activations. On the other hand, for outlier detection from LSUN-Crop [47], utilizing activations from deeper layers yields better results. The performance decline of ATS on Places365 [49] can be attributed to the semantic and intrinsic similarities and overlaps between samples in Places365 [49] and the CIFAR [20] in-distribution classes. For a more in-depth analysis, we refer readers to the supplementary material.

Figure 4 shows the importance of intermediate layers in the temperature calculation of a ResNet50 trained on ImageNet-1k [4]. For the large-scale ImageNet benchmark, we see that for all OOD datasets (iNaturalist [37], SUN [42], Places [49], Textures [3], NINCO [2]) except for Fashion-MNIST [41], deeper layers are more important for temperature calculation. We posit that the difference in the impact of intermediate layers on OOD detection performance between ImageNet [4] and CIFAR [20] can be attributed to two factors. First, ImageNet [4] is a more complex dataset than CIFAR [20], which implies that shallow layers in the network extract more general features that are less useful for distinguishing near-OOD samples. Second, the OOD datasets used in our evaluation exhibit greater statistical and semantic similarity to ImageNet [4], making shallow layers less important. As a result, the improvement achieved by our ATS on strong baselines, such as ReAct [32], DICE [34], and ASH [6], is not as substantial for ImageNet [4] as it is for CIFAR [20] performance. However, it is also evident that the first layers are important for far-OOD samples such as those from Fashion-MNIST [41], making existing methods more robust against extreme OOD outliers while only moderately impacting OOD detection performance for near-OOD samples.

Three points are clearly observable: i) that the best layer depends on the ID vs. OOD dataset (in accordance with existing OOD detection methods that also leverage intermediate layers [7, 31, 40]), ii) combing intermediate layer information improves the performance, and iii) selecting intermediate layers across the model depth improves robustness against the diverse characteristics and properties present in both ID and OOD samples. These findings not

	Method	CIFAR-10 AUROC ↑	CIFAR-10 FPR95 ↓	CIFAR-100 AUROC ↑	CIFAR-100 FPR95 ↓
ResNet18	AT-only	89.16	36.73	86.67	40.53
	MSP	91.69	55.49	77.07	78.46
	MSP+	**96.93**	**14.21**	92.25	31.21
	ODIN	94.64	24.35	86.57	54.28
	ODIN+	96.14	16.60	90.32	36.67
	MLS	94.01	33.50	84.19	66.61
	MLS+	96.91	14.24	92.24	31.15
	Energy	94.07	32.96	84.45	66.10
	Energy+	94.54	23.00	88.74	35.28
	ReAct	70.61	77.12	87.48	61.01
	ReAct+	86.59	45.84	92.39	**28.47**
	DICE	75.51	62.22	87.65	49.40
	DICE+	89.06	38.90	**92.78**	29.16
	ASH-B	68.12	75.14	88.45	52.34
	ASH-B+	84.84	48.67	92.25	28.87
DenseNet	AT-only	91.00	29.33	87.30	41.30
	MSP	92.59	49.71	77.66	78.37
	MSP+	98.04	9.62	93.96	25.74
	ODIN	96.17	18.61	84.96	58.70
	ODIN+	97.92	10.08	**94.12**	**24.17**
	MLS	95.66	21.54	83.87	64.47
	MLS+	98.04	9.63	93.96	25.70
	Energy	95.72	20.88	83.93	64.12
	Energy+	97.03	12.65	90.90	32.75
	ReAct	96.36	19.64	80.83	70.18
	ReAct+	**98.10**	**9.31**	92.96	24.65
	DICE	96.74	14.92	84.48	51.88
	DICE+	97.93	10.16	93.56	26.85
	ASH-B	97.34	12.86	89.90	40.72
	ASH-B+	97.67	10.35	93.04	25.85

Table 2. Average performance over the OOD datasets SVHN [28], Textures [3], iSUN [43], LSUN [47], LSUN-Crop [47], Places365 [49], MNIST [21], and Fashion-MNIST for CIFAR-10 and CIFAR-100 as ID data, considering two architectures. *AT-only* denotes the results when the per-sample temperature is used directly as the OOD detection score. *Method+* (highlighted rows) denotes that our ATS is applied on top of the method. Detailed results can be found in the supplemental material.

only validate our proposed methodology, particularly the hyperparameter-free layer selection, but also provide opportunities for future research in also selecting the best N layers for the temperature calculation dynamically based on the input sample.

Adaptive Temperature only. We evaluate the performance of using the sample-specific temperature directly (considering only information of intermediate layers) as the OOD detection score compared to ATS (combining information from intermediate features and class-dependent logits). In Table 2 the method "AT-only" shows the average performance over SVHN [28], Textures [3], iSUN [43], LSUN [47], LSUN-Crop [47], Places365 [49], and MNIST [21] for CIFAR [20] as ID data, utilizing the per-sample temperature directly to distinguish ID and OOD samples. Our findings indicate that the performance of the per-sample temperature on its own is not particularly strong. However, when combined with the class-dependent logits through ATS, we observe a significant improvement in performance across all settings. In summary, leveraging information from all layers of the network, rather than relying on parts, leads to a better and significantly more robust performance of the OOD detection method.

Model	Method	ImageNet AUROC ↑	ImageNet FPR95 ↓
ResNet50	MSP	82.95	66.52
	MSP+	89.57	42.73
	ODIN	87.46	49.65
	ODIN+	88.90	40.77
	MLS	85.86	62.86
	MLS+	89.56	42.75
	Energy	85.53	63.52
	Energy+	79.94	63.52
	ReAct	90.33	43.75
	ReAct+	90.62	38.13
	DICE	87.55	46.52
	DICE+	88.38	42.69
	ReAct+DICE	88.75	42.43
	ReAct+DICE+	88.28	41.96
	ASH-B	**92.69**	34.05
	ASH-B+	91.61	**33.43**

Table 3. Average performance over the OOD datasets iNaturalist [37], SUN [42], Places [49], Textures [3], NINCO [2] and Fashion-MNIST [41], where the ID data comes from ImageNet. *Method+* (highlighted rows) denotes that ATS is applied on top of the method. See supplemental material for detailed results.

Computational Efficiency. Our approach results in a modest computational overhead: 6.18% for ResNet18, 11.94% for ResNet50, and 4.71% for DenseNet100. This minor increase in computational cost is justified when weighed against the notable performance gains and enhanced robustness. A comprehensive breakdown of this runtime analysis is available in the supplementary material.

4.4. Limitations

As a post-hoc method, ATS's effectiveness relies on the quality of the intermediate features and the method used for their extraction. ATS exhibits less performance improve-

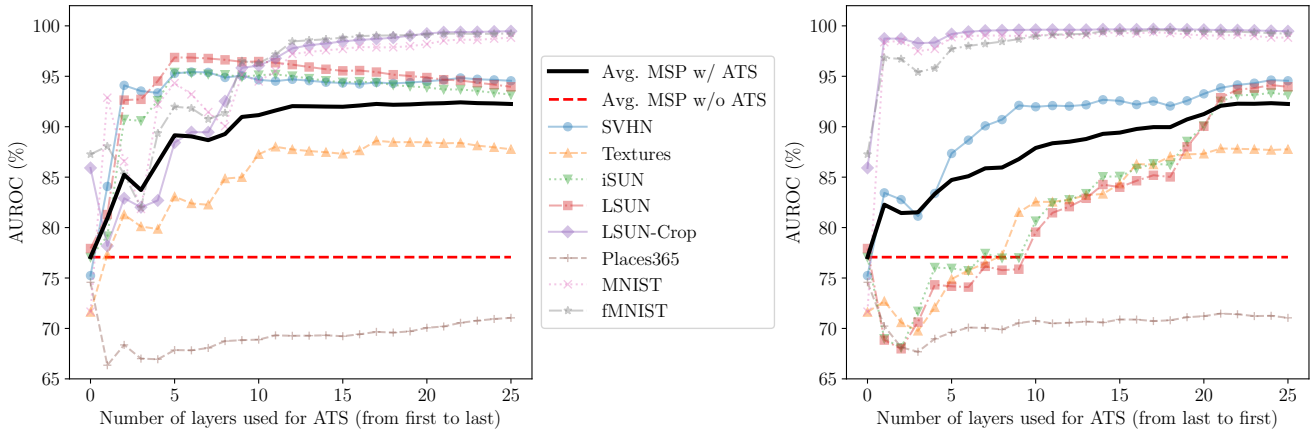

Figure 3. Evaluation of intermediate layer importance for temperature calculation on ResNet18 using CIFAR-100 as the in-distribution dataset and various OOD datasets: SVHN [28], Textures [3], iSUN [43], LSUN [47], LSUN-Crop [47], Places365 [49], MNIST [21], and Fashion-MNIST [41]. **Left figure:** The first l layers are used for the adaptive temperature calculation and we plot the AUROC for different l. **Right figure:** The last l layers are used for the adaptive temperature calculation and we plot the AUROC for different l.

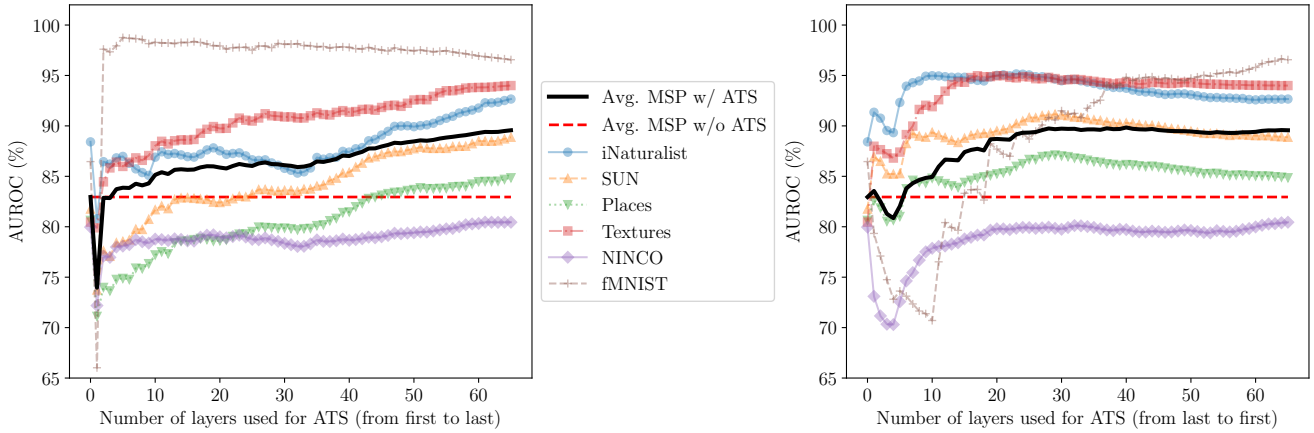

Figure 4. Evaluation of intermediate layer importance for temperature calculation on ResNet50 using ImageNet as the in-distribution dataset and various OOD datasets: iNaturalist [37], SUN [42], Places [49], Textures [3], NINCO [2] and Fashion-MNIST [41]. **Left figure:** The first l layers are used for the adaptive temperature calculation and we plot the AUROC for different l. **Right figure:** The last l layers are used for the adaptive temperature calculation and we plot the AUROC for different l.

ment on OOD datasets where the relevant information is predominantly present in deeper layers, as observed in the case of SUN on the ImageNet benchmark (see Sec. 4.3). Still, ATS enhances the robustness of the considered methods, especially against far-OOD samples, which are crucial to detect due to their potential impact and the erosion of trust in ML systems when not recognized. Consequently, ATS contributes to the overall improvement and practical applicability of OOD detection methods.

5. Conclusion

In this paper, we present a simple yet highly effective extension to existing logit-based post-hoc OOD detection methods that works by adaptively scaling the logits with a per-sample temperature calculated from intermediate layer activations. ATS, our proposed method, can be seamlessly applied to all OOD detection methods that utilize the model output for the OOD score calculation. We conduct extensive experiments and evaluations on widely used OOD benchmarks, demonstrating the favorable performance of ATS. The results highlight the simplicity and effectiveness of ATS in enhancing OOD detection capabilities across different methods and datasets.

Acknowledgements We gratefully acknowledge the financial support from KESTRELEYE GmbH and its steadfast commitment to advancing this research.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent Space Autoregression for Novelty Detection. In *Proc. CVPR*, 2019.

[2] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In *Proc. ICLR Workshops*, 2023.

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proc. CVPR*, 2014.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[5] Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance. *arXiv preprint arXiv:1812.02765*, 2018.

[6] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *Proc. ICLR*, 2023.

[7] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and H.T. Kung. Neural Mean Discrepancy for Efficient Out-of-Distribution Detection. In *Proc. CVPR*, pages 19217–19227, June 2022.

[8] Ronald Aylmer Fisher. *Statistical methods for research workers*. Springer, 1992.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proc. ICML*, pages 1321–1330, 2017.

[10] Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks. In *Proc. ICLR*, 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. CVPR*, 2016.

[12] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proc. ICML*, 2022.

[13] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved Problems in ML Safety. *arXiv preprint arXiv:2109.13916*, 2021.

[14] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proc. ICLR*, 2017.

[15] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proc. ICLR*, 2017.

[16] Dan Hendrycks and Mantas Mazeika. X-Risk Analysis for AI Research. *arXiv preprint arXiv:2206.05862*, 2022.

[17] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proc. CVPR*, 2017.

[19] Rui Huang and Yixuan Li. MOS: Towards Scaling Out-of-distribution Detection for Large Semantic Space. In *Proc. CVPR*, 2021.

[20] Alex Krizhevsky and Geoffrey E. Hinton. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[21] Yann Lecun, Lé'on Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based Learning Applied to Document Recognition. *IEEE*, 86(11):2278–2324, 1998.

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*, 2018.

[23] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proc. ICLR*, 2018.

[24] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *NeurIPS*, 2020.

[25] Sina Mohseni, Haotao Wang, Zhiding Yu, Chaowei Xiao, Zhangyang Wang, and Jay Yadawa. Practical Machine Learning Safety: A Survey and Primer. *arXiv preprint arXiv:2106.04823*, 2021.

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proc. CVPR*, 2017.

[27] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? . In *Proc. ICLR*, 2019.

[28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. *NeurIPS Workshops*, 2011.

[29] Anh M Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *Proc. CVPR*, 2015.

[30] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection. *arXiv preprint arXiv:2106.09022*, 2021.

[31] Chandramouli Shama Sastry and Sageev Oore. Detecting Out-of-Distribution Examples with Gram Matrices. In *Proc. ICML*, 2020.

[32] Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution Detection With Rectified Activations. In *NeurIPS*, 2021.

[33] Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution Detection With Rectified Activations. *NeurIPS*, 2021.

[34] Yiyou Sun and Yixuan Li. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *Proc. ECCV*, 2022.

[35] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution Detection with Deep Nearest Neighbors. In *Proc. ICML*, 2022.

[36] Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. Hyperparameter-Free Out-of-Distribution Detection Using Cosine Similarity. In *Proc. ACCV*, 2020.

[37] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The INaturalist Species Classification and Detection Dataset. In *Proc. CVPR*, 2018.

[38] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore Willke. Out-of-Distribution Detection Using an Ensemble of Self Supervised Leave-Out Classifiers. In *Proc. ECCV*, 2018.

[39] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-Of-Distribution with Virtual-logit Matching. In *Proc. CVPR*, 2022.

[40] Samuel Wilson, Tobias Fischer, Niko Sünderhauf, and Feras Dayoub. Hyperdimensional Feature Fusion for Out-Of-Distribution Detection. In *Proc. WACV*, 2023.

[41] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[42] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, 2010.

[43] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[44] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. In *NeurIPS Datasets and Benchmarks Track*, 2022.

[45] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection: A Survey. *arXiv preprint arXiv:2110.11334*, 2021.

[46] Yijun Yang, Ruiyuan Gao, and Qiang Xu. Out-of-Distribution Detection with Semantic Mismatch under Masking. In *Proc. ECCV*, 2022.

[47] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365*, 2016.

[48] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block Selection Method for Using Feature Norm in Out-of-Distribution Detection. In *Proc. CVPR*, 2023.

[49] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million Image Database for Scene Recognition. *TPAMI*, 2017.

[50] Yibo Zhou. Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection. In *Proc. CVPR*, 2022.

[51] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *Proc. ICLR*, 2018.