IMPACT OF THE LAST FULLY CONNECTED LAYER ON OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection, a task that aims to detect OOD data during deployment, has received lots of research attention recently, due to its importance for the safe deployment of deep models. In this task, a major problem is how to handle the overconfidence problem in OOD data. While this problem has been explored from several perspectives in previous works, such as the measure of OOD uncertainty and the activation function, the connection between the last fully connected (FC) layer and this overconfidence problem is still less explored. In this paper, we find that the weight of the last FC layer of the model trained on in-distribution (ID) data can be an important source of the overconfidence problem, and we propose a simple yet effective OOD detection method to assign the weight of the last FC layer with small values instead of using the original weight trained on ID data. We analyze in Sec. 5 that our proposed method can make the OOD data and the ID data to be more separable, and thus alleviate the overconfidence problem. Moreover, our proposed method can be flexibly applied on various off-the-shelf OOD detection methods. We show the effectiveness of our proposed method through extensive experiments on the ImageNet dataset, the CIFAR-10 dataset, and the CIFAR-100 dataset.

1 INTRODUCTION

Recently, deep models have achieved good performance in various computer vision tasks, but with a severe reliance on the assumption that the testing data comes from the same distribution as the training set (i.e., *in-distribution* (ID) test data) (Ben-David et al., 2010; Vapnik, 1991). This assumption, however, can be violated in the open world where *out-of-distribution* (OOD) data can be often encountered, and these OOD data as inputs can lead models to produce unrelated predictions and result in severe consequences, especially in many safety-critical applications, such as autonomous driving (Filos et al., 2020) and medical diagnosis (Zadorozhny et al., 2021). Due to the severe implications of OOD data in these applications, the task of OOD detection, which aims to detect OOD data during deployment, is important and has received lots of research attention recently (Liang et al., 2017; Hendrycks & Gimpel, 2016; Hendrycks et al., 2019; Liu et al., 2020; Sun et al., 2021; Huang & Li, 2021; Huang et al., 2021; Lee et al., 2018).

To detect OOD data, a naive idea is to classify the OOD data and the ID data based on the confidence of the model in the data input. However, as deep models can be overconfident in the OOD data inputs (Nguyen et al., 2015), it can be non-trivial to separate the OOD data and the ID data based on such a naive idea. To better cope with the overconfidence problem and make the OOD data and the ID data more separable, previous works have proposed methods from several perspectives, such as redefining the measure of OOD uncertainty (Liu et al., 2020; Wang et al., 2022; Hendrycks et al., 2019) and rectifying the activation function (Sun et al., 2021). However, the connection between the last fully connected (FC) layer and the overconfidence problem is still less explored.

In this work, we argue that the weight of the last FC layer of the model trained on ID data can be an important source of the overconfidence problem. To justify our aforementioned argumentation, as a preliminary of our method, in Fig. 1, we use a ResNet-50 (He et al., 2016) model trained on ImageNet and conduct OOD detection experiments on various other datasets including iNaturalist, SUN, Places365, and Textures. Specifically, we compare the baseline that uses the original weight of the last FC layer (**original weight**) with a variant that assigns the weight of the last FC layer sim-

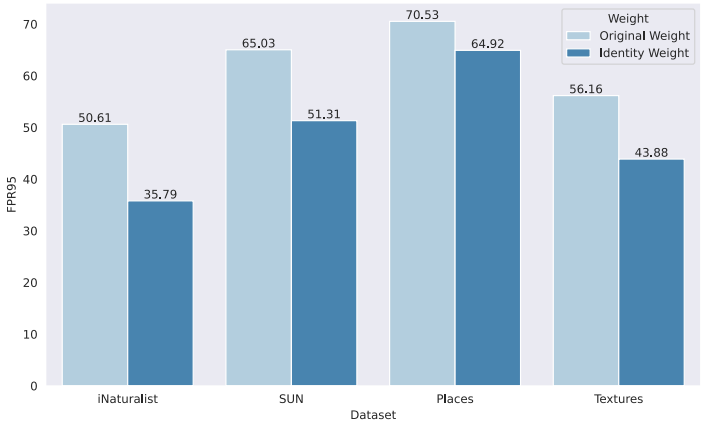


Figure 1: Comparison between the baseline method that uses the original weight of the last FC layer and the variant that assigns the weight of the last FC layer simply with ones. Note that in both the baseline method and the variant, following previous works (Liu et al., 2020; Sun et al., 2021; Wang et al., 2022), we consistently use the energy score (Liu et al., 2020) as the measure of OOD uncertainty.

ply with ones (**identity weight**). As illustrated, compared to the baseline, this variant consistently reduces the false positive rate (FPR95) over various datasets. This demonstrates that the weight of the last FC layer of the model trained on ID data is not the optimal weight for OOD detection, and there can exist a weight that is more suitable.

Inspired by the above argumentation, in this work, to better cope with the overconfidence problem, we aim to assign the last FC layer of the model with a new weight so that the OOD data and the ID data can be made more separable. We find that this can be achieved via simply assigning small values (e.g. 0.01) to the weight of the last FC layer of the model. To theoretically show the effectiveness of our method, in Sec. 5, we first analyze why assigning constant values (e.g., ones) to the weight of the last FC layer can separate OOD data and ID data; we then explain why assign the weight of the last FC layer with small value can even make OOD data and ID data to be more separable. We also want to point out that, as the original weight of the last FC layer can still be used for the original task, via using our method, the classification accuracy on the original task is completely preserved.

Also, note that, as we just need to assign the last FC layer of the model with small values, our method is simple yet effective and needs neither a retraining process of the model nor additional OOD data. Besides, with only the weight of the last FC layer modified, our method can also be flexibly applied to various off-the-shelf OOD detection methods. We experiment our method with various OOD detection methods and achieve consistent improvement in OOD detection performance.

The contributions of our work are summarized as follows.

- From the novel perspective of the last FC layer, we propose a simple and effective OOD detection method to detect OOD data by simply assigning the weight of the last FC layer with small values.
- We perform theoretical analysis (in Sec. 5) on why assigning constant values (e.g., ones) to the weight of the last FC layer can separate OOD data and ID data. Moreover, we also analyze why a small value can even make the OOD data and ID data to be more separable. Our method thus can improve the OOD detection performance.
- Our method achieves significant OOD detection performance improvement when applied to various OOD detection methods on various evaluation benchmarks (Deng et al., 2009; Krizhevsky et al., 2009).

The rest of the paper is organized as follows. In Sec. 2, we discuss the related works of our paper. In Sec. 3, we provide the background of OOD detection. After that, we present our method in Sec. 4,

the analysis of our method in Sec. 5, and experimental results in Sec. 6. Finally, we conclude our paper in Sec. 7.

2 Related Work

OOD Detection. Being an important task that helps detect OOD data during deployment, OOD detection has received lots of research attention, and most of the OOD detection methods fall into three categories: methods need retraining (DeVries & Taylor, 2018; Huang & Li, 2021; Zaeemzadeh et al., 2021), methods need extra OOD data (Hsu et al., 2020; Hendrycks et al., 2018; Dhamija et al., 2018; Ming et al., 2022; Lee et al., 2017; Yu & Aizawa, 2019; Wu et al., 2021), and post-hoc methods. Among methods that need retraining, an extra branch is introduced by (DeVries & Taylor, 2018), and MOS makes use of a group-based feature space, and (Zaeemzadeh et al., 2021) incorporates angular distance into their method. In the category of methods that need extra OOD data, (Hendrycks et al., 2018) is the first to propose this category of method, (Dhamija et al., 2018) proposed to regularize extra image data from different backgrounds, and (Lee et al., 2017) proposed to generate OOD data on the boundary of OOD data and ID data. Besides these two categories of method, the category of post-hoc method have also attracted a lot of attention recently since it need neither retraining nor extra OOD data.

In the category of post-hoc methods, (Hendrycks & Gimpel, 2016) observe that a neural model tends to produce higher softmax values for ID data and lower ones for the OOD data. Therefore, they introduce a score function, the maximum softmax probability (MSP), to achieve OOD detection. To improve the OOD detection performance, (Liang et al., 2017) puts forward ODIN, which enlarges the gap between ID and OOD data by using large temperature scaling and adding perturbations on inputs. Lee uses the features and the class-wise centroids to calculate the Mahalanobis distance (Lee et al., 2018). The energy-based score function is introduced by (Liu et al., 2020). Such a function gives high energy to the OOD data and low energy to the ID data. (Sun et al., 2021) exploits the characteristics of the neural network to the OOD data and leverages the OOD detection performance by removing abnormal activate values.

Different from the existing post-hoc OOD detection methods, this paper takes a different view of the OOD detection problem. Specifically, we propose to connect the last FC layer and the overconfidence problem, and we propose to replace the original weight of the last FC layer with small values instead.

The Last FC Layer. The last FC layer, an important component that appears in many network structures, has been studied in various areas (Basha et al., 2020a;b; Zhao et al., 2020; Zhou et al., 2020) over the year, such as transfer learning (Basha et al., 2020a), continual learning (Zhao et al., 2020), and long tail problem (Zhou et al., 2020). In this paper, from a novel perspective, we build a connection between the last FC layer and the overconfidecne problem in OOD detection. Specifically, we find that the weight of the last FC layer trained on ID data can be an important source of the over confidence problem and propose to assign the weight of the last FC layer with small values instead.

3 Background

Following most previous OOD detection works (Liang et al., 2017; Hendrycks & Gimpel, 2016; Hendrycks et al., 2019; Liu et al., 2020; Sun et al., 2021; Huang & Li, 2021; Huang et al., 2021; Lee et al., 2018), this paper considers OOD detection in image classification. Denote $D_{in} := \mathcal{X}^{in} \times \mathcal{Y}^{in}$ drawn from P_{in} the in-distribution dataset, where P_{in} denotes the in-distribution, $\mathcal{X}^{in}$ denotes the in-distribution input space, and $\mathcal{Y}^{in} = \{1, 2, \cdots, C\}$ denotes the in-distribution label space corresponding to $\mathcal{X}^{in}$. Similarly, denote $D_{out} := \mathcal{X}^{out} \times \mathcal{Y}^{out}$ the out-of-distribution dataset, where $\mathcal{X}^{in}$ denotes the out-of-distribution input space, and $\mathcal{Y}^{in}$ denotes the corresponding out-of-distribution label space. Moreover, denote $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ an image classifier trained on D_{in}. OOD detection can then be treated as a binary classification problem to distinguish whether the input data $\langle x, y \rangle \in D_m$ belongs to D_{in} or D_{out}, where x is an image, and y is its corresponding ground true label. In other words, given a certain neural network $\mathcal{F}$ and a random test input x, the goal of OOD

detection is to define a score function $G(x; \mathcal{F})$ such that:

$$G(x; \mathcal{F}) = \begin{cases} 1, & \text{if } x \in D_{in} \\ 0, & \text{if } x \in D_{out} \end{cases} \tag{1}$$

where $D_{in} \cap D_{out} = \mathcal{X}^{in} \cap \mathcal{X}^{out} = \mathcal{Y}^{in} \cap \mathcal{Y}^{out} = \varnothing$. Note that the D_{out} is inaccessible during the training stage of $\mathcal{F}$.

4 METHOD

In this section, we introduce our proposed OOD detection methods. The idea behind our method is to better cope with the overconfidence problem by replacing the last FC layer of the model with a new linear layer filled with a small constant value. We consider the input x, the last FC layer f, and the well trained neural network without the last FC layer g. We denote a d dimension feature vector from the penultimate layer of the model as $(z_1, z_2, ..., z_d) = \mathbf{z} := g(x) \in \mathbf{R}^d$, the output of the model as $f(\mathbf{z})$ where matrix $f \in \mathbf{R}^{d \times K}$ and K is the number of classes. Following most of the recent OOD detection methods (Liu et al., 2020; Sun et al., 2021; Wang et al., 2022; 2021; Tonin et al., 2021; Du et al., 2022; Elflein et al., 2021; Wang et al., 2020; Joshi et al., 2022; Chen et al., 2022; Ouyang et al., 2021; Ming et al., 2022), we first define the original measure of OOD uncertainty S_{ori} before incorporating our proposed method as:

$$S_{ori} = \log \sum_{i=1}^{K} e^{f_i(\mathbf{z})} \tag{2}$$

where f_i indicates the i-th column of the matrix f. Note that a larger S_{ori} indicates more confidence that x belongs to the in-distribution.

We then describe how the measure of OOD uncertainty S looks like after incorporating our proposed method. Specifically, let's denote the matrix $f'_i \in \mathbf{R}^{d \times K}$ filled with a value α, and then we replace the f with $f'_i \in \mathbf{R}^d$ to compute S. Since all entries of f'_i are same, all columns of f'_i are identical i.e. $f'_1 = f'_2 = ... = f'_K$. Therefore, S can be denoted as:

$$\begin{aligned}
S &= \log \sum_{i=1}^{K} e^{f'_i(\mathbf{z})} \\
&= \log \left(e^{f'_1(\mathbf{z})} + e^{f'_2(\mathbf{z})} + ... + e^{f'_K(\mathbf{z})}\right) \\
&= \log K e^{f'_1(\mathbf{z})} && \text{where } f'_1 = \alpha J_{d,1} \\
&= \log K e^{f'^T_1 \mathbf{z}} && \text{where } f'_1(\mathbf{z}) = f'^T_1 \mathbf{z} \\
&= \log K e^{\alpha \sum_{i=1}^{d} z_i} && \text{where } f'^T_1 \mathbf{z} = \alpha J_{d,1} \mathbf{z} = \alpha \sum_{i=1}^{d} z_i \\
&= \log K e^{d\alpha \bar{z}} && \text{where } \bar{z} = \frac{1}{d} \sum_{i=1}^{d} \mathbf{z}_i \\
&= d\alpha \bar{z} + \log K
\end{aligned} \tag{3}$$

where $\bar{z} := \mathbb{E}(z)$ and $J_{d,1}$ indicates a $d \times 1$ all-ones matrix.

To perform OOD detection using our proposed method, we further define the score function $G(x; \mathcal{F})$ as:

$$G(x; \mathcal{F}) = \begin{cases} 1, & \text{if } S \geq \lambda \\ 0, & \text{if } S < \lambda \end{cases} \tag{4}$$

where λ is a threshold. In our experiments, we set λ to be a value such that 95% ID data can be detected correctly, which is the same setting following most previous OOD detection methods (Hendrycks & Gimpel, 2016; Liu et al., 2020; Sun et al., 2021; Wang et al., 2022; Liang et al., 2017; Hendrycks et al., 2019).

5 ANALYSIS

Below, we perform theoretical analysis to show the effectiveness of our method. Specifically, we first explain why replacing the trained weight of the last FC layer with a constant value α separates the distributions of ID and OOD data. After that, we further explain why a smaller α can make the ID and OOD data to be more separable.

5.1 EFFECTIVENESS OF ASSIGNING THE LAST FC LAYER WITH A CONSTANT VALUE

In this section, we analyze why assigning the last FC layer with a constant value α can separate ID and OOD data. Following the settings in Sec. 3, we denote the neural network trained on the ID data as $\mathcal{F}$. Besides, we further denote the output of its penultimate layer is $\boldsymbol{z} = (z_1, z_2, ..., z_n) \in \boldsymbol{R^n}$. Then, we can rewrite the score S produced by our method further as:

$$\begin{aligned} S &= \log \sum_{i=1}^{K} e^{f_i'(\boldsymbol{z})} \\ &= \log k e^{d\alpha\bar{z}} \\ &= d\alpha\bar{z} + \log K \\ &\propto \bar{z} \\ &= \mathbb{E}[\boldsymbol{z}] \end{aligned} \tag{5}$$

We denote the $\boldsymbol{z}$ corresponding to the in-distribution data as $\boldsymbol{z^{in}} = (z_1^{in}, z_2^{in}, ..., z_n^{in})$. Following the same assumption from (Ming et al., 2022; Sun et al., 2021; 2022), we assume that each z_i^{in} obeys the rectified Gaussian distribution i.e. $z_i^{in} \sim max(0, \mathcal{N}(\mu, \sigma_{in}^2))$. Then, we can model z_i^{in} with a random variable x as:

$$z_i^{in} = \frac{1}{\sigma_{in}\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma_{in}^2}} \tag{6}$$

We denote the corresponding expectation of $\boldsymbol{z^{in}}$ as $\mathbb{E}_{in}$, and it can be written as:

$$\begin{aligned} \mathbb{E}_{in}[\boldsymbol{z}] &= \int_0^{+\infty} \frac{x}{\sigma_{in}\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{\sigma_{in}^2}} dx \\ &= \frac{1}{\sigma_{in}\sqrt{2\pi}} \int_0^{+\infty} x e^{-\frac{(x-\mu)^2}{\sigma_{in}^2}} dx \\ &= \frac{\mu}{\sqrt{2\pi}} \int_{-\frac{\mu}{\sigma_{in}}}^{+\infty} e^{-\frac{v^2}{2}} dv + \frac{\sigma_{in}}{\sqrt{2\pi}} \int_{-\frac{\mu}{\sigma_{in}}}^{+\infty} v e^{-\frac{v^2}{2}} dv \qquad \text{where } v = \frac{x-\mu}{\sigma_{in}} \\ &= \frac{\mu}{\sqrt{2\pi}} (1 - \int_{-\infty}^{-\frac{\mu}{\sigma_{in}}} e^{-\frac{v^2}{2}} dv) + \frac{\sigma_{in}}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{-\mu}{\sigma_{in}})^2} \\ &= \mu[1 - \Phi(\frac{-\mu}{\sigma_{in}})] + \sigma_{in}\varphi(\frac{-\mu}{\sigma_{in}}) \end{aligned} \tag{7}$$

where Φ and φ are Cumulative distribution function(cdf) and Probability density function(pdf) respectively. And then, we are going to model the expectation corresponding to the out-of-distribution data. Following the same observation from (Sun et al., 2021) that the output of the penultimate layer of the network corresponding to the OOD data,$\boldsymbol{z^{out}}$, is positively skewed. Specifically, we cam denote $\boldsymbol{z^{out}} = (z_1^{out}, z_2^{out}, ..., z_n^{out})$, so we can model each $z_i^{out} \sim$ESN$(\mu, \sigma_{out}^2, \epsilon)$, where μ, σ_{out}^2, ϵ indicate the mean, the deviation and the degree of skewness of the ESN distribution. Therefore, following the theorem in (Mudholkar & Hutson, 2000), the expectation of $\boldsymbol{z^{out}}$, $\mathbb{E}_{out}[\boldsymbol{z}]$, can be modeled as:

$$\mathbb{E}_{out}[\boldsymbol{z}] = \mu - (1+\epsilon)\Phi(\frac{-\mu}{(1+\epsilon)\sigma_{out}})\mu + (1+\epsilon)^2\varphi(\frac{-\mu}{(1+\epsilon)\sigma_{out}}) - \frac{4\epsilon}{\sqrt{2\pi}}\sigma_{out} \tag{8}$$

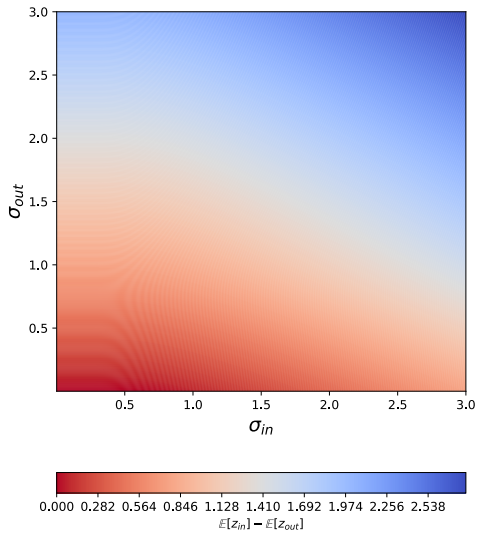


Figure 2: The relationship of Δ, σ_{in}, and σ_{out}, where Δ represents the difference of the $\mathbb{E}_{in}[\boldsymbol{z}]$ and $\mathbb{E}_{out}[\boldsymbol{z}]$. Observe that Δ is positive here, which indicates that our method can generate a higher score for the in-distribution data than the out-distribution ones, in other words, separate the distributions of ID and OOD data.

Therefore, the difference of the $\mathbb{E}_{in}[\boldsymbol{z}]$ and $\mathbb{E}_{out}[\boldsymbol{z}]$ is:

$$
\begin{aligned}
\Delta =& \mathbb{E}[\boldsymbol{z}_{in}] - \mathbb{E}[\boldsymbol{z}_{out}] \\
=& \mu[1 - \Phi(\frac{-\mu}{\sigma_{in}})] + \sigma_{in}\varphi(\frac{-\mu}{\sigma_{in}}) - \mu - (1+\epsilon)\Phi(\frac{-\mu}{(1+\epsilon)\sigma_{out}})\mu \\
&+ (1+\epsilon)^2\varphi(\frac{-\mu}{(1+\epsilon)\sigma_{out}}) - \frac{4\epsilon}{\sqrt{2\pi}}\sigma_{out} \\
=& -[(1+\epsilon)^2\phi(\frac{-\mu}{(1+\epsilon)\sigma_{out}}) + \frac{4\epsilon}{\sqrt{2\pi}}]\sigma_{out} \\
&- [\Phi(\frac{-\mu}{\sigma_{in}}) - (1+\epsilon)\Phi(\frac{-\mu}{(1+\epsilon)\sigma_{out}})]\mu + \phi(\frac{-\mu}{\sigma_{in}})\sigma_{in}
\end{aligned} \tag{9}
$$

Given $\mu = 1.0$ and $\epsilon = -0.5$, we can plot Δ in Fig. 2, and we can find out that it is greater than 0, i.e $S_{in} > S_{out}$. Therefore, we conclude that our method can produce greater confidence scores to in-distribution data than for the out-distribution data.

5.2 Effectiveness of a small α

In this section, we further explain why a smaller α can make the ID and OOD data to be more separable. We denote the norm difference of S_{in} and S_{out} as:

$$
\begin{aligned}
\frac{S_{in} - S_{out}}{||\alpha||_2} &= \frac{1}{\alpha^2}(\log\sum_{i=1}^{K} e^{f_i'(\boldsymbol{z^{in}})} - \log\sum_{i=1}^{K} e^{f_i'(\boldsymbol{z^{out}})}) \\
&= \frac{d}{\alpha}(\mathbb{E}[\boldsymbol{z}_{in}] - \mathbb{E}[\boldsymbol{z}_{out}])
\end{aligned} \tag{10}
$$

As shown in Eq. 10, to make the ID and OOD data to be more separable, we actually hope to make the norm difference of S_{in} and S_{out} to be larger Recall that $\mathbb{E}[\boldsymbol{z}_{in}] - \mathbb{E}[\boldsymbol{z}_{out}]$ is a positive number as we discuss above. Therefore, a smaller α can make the norm difference of S_{in} and S_{out} larger, and thus make the ID and OOD data to be more separable.

6 EXPERIMENTS

In this section, we evaluate the effectiveness of our method on ImageNet and CIFAR OOD detection benchmarks. All experiments are conducted on NVIDIA Tesla V100 GPUs.

6.1 IMAGENET BENCHMARK

Setup. We use ReAct (Sun et al., 2021) as a baseline of our method and follow it. We use both a ResNet50 (He et al., 2016) model and a MobileNet-v2 (Sandler et al., 2018) model pre-trained on ImageNet (Deng et al., 2009) as the image classifier. Note that for fair comparison, we directly use the models trained by (Sun et al., 2021). Moreover, we set α in Eq. 3 to be a small number 0.01 in our experiments.

Evaluation Metric. We evaluate our OOD detection method on the following two common metrics: (1) **FPR95** measures the FPR (False Positive Rate) of the OOD data when the recall (Positive Rate of the ID data) is at 95%. Note that a lower FPR95 indicates better performance of OOD detection. (2) **AUROC** measures the area under the TPR (True Positive Rate) and FPR (False Positive Rate). Note that a higher AUROC indicates better performance of OOD detection.

Dataset. In this benchmark, we consider ImageNet (Deng et al., 2009) as the ID dataset, and following (Sun et al., 2021; Hsu et al., 2020; Huang & Li, 2021), we evaluate our method on four commonly-used OOD datasets, including iNaturalist, SUN, Places365, and Textures. Note that all of these four datasets have non-overlapping classes w.r.t ImageNet. Below, we introduce each of them in more detail: (1) **iNaturalist** (Van Horn et al., 2018) contains 5,000 categories of plants and animals images, and the resolution of each image is 800×800. To conduct OOD detection on this dataset, following the setting of (Sun et al., 2021; Huang & Li, 2021; Huang et al., 2021), 110 classes that is non-overlapping with classes of ImageNet are first picked up, and 10,000 images from these 110 classes are then randomly selected. (2) **SUN** (Xiao et al., 2010) contains 397 classes of natural images, and the resolution of each image is larger than 200×200. To conduct OOD detection on this dataset, following the setting of (Sun et al., 2021; Huang & Li, 2021; Huang et al., 2021), 50 classes that is non-overlapping with classes of ImageNet are first picked up, and 10,000 images from these 110 classes are then randomly selected. (3) **Places** (Zhou et al., 2017) contains 205 categories of scene images whose resolutions are 512×512. To conduct OOD detection on this dataset, following the setting of (Sun et al., 2021; Huang & Li, 2021; Huang et al., 2021), 50 classes that is non-overlapping with classes of ImageNet are first picked up, and 10,000 images from these 110 classes are then randomly selected. (4) **Textures** (Cimpoi et al., 2014) contains 47 classes of textural images whose resolutions are either 300×300 or 640×640. Following (Sun et al., 2021; Huang & Li, 2021; Huang et al., 2021), the whole dataset with 5,640 images is used for OOD detection evaluation.

Results. In Tab. 1, we compare our method with the existing post-hoc OOD detection methods on all the four OOD datasets. As shown, our method demonstrates the best averaged result compared with common post-hoc OOD detection methods on both ResNet50 and MobileNet-V2, which demonstrates the effectiveness of our method.

6.2 CIFAR BENCHMARK

Setup. We use ReAct (Sun et al., 2021) as a baseline of our method and follow it. We use the ResNet18 (He et al., 2016) model as the image classifier for both CIFAR-10 and CIFAR-100. Note that for fair comparison, we directly use the models trained by (Sun et al., 2021). Moreover, we set α in Eq. 3 to be a small number 0.01 in our experiments.

Evaluation metric & Dataset. Following (Hendrycks & Gimpel, 2016; Liu et al., 2020; Liang et al., 2017; Sun et al., 2021; Huang et al., 2021; Lee et al., 2018), we use the **FPR95** and **AUROC** metrics elaborated in Sec. 6.1 to evaluate our OOD detection method. In this benchmark, we use CIFAR-10 and CIFAR-100 as the ID datasets (Krizhevsky et al., 2009). With respect to the OOD datasets, following (Liu et al., 2020; Sun et al., 2021; Huang et al., 2021; Cimpoi et al., 2014), besides using the Places dataset and the Textures dataset that we have introduced above, we also

Models	Methods	Conferences	iNaturalist		SUN		Places		Textures		Average	
			FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet50	MSP	ICLR 2017	53.40	88.01	73.68	79.83	76.12	78.74	68.88	80.54	68.02	81.78
	Mahalanobis	NeurIPS 2018	93.90	62.04	96.83	51.88	97.02	52.99	37.50	91.64	81.31	64.64
	ODIN	ICLR 2018	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
	Energy	NeurIPS 2020	50.60	90.96	65.03	84.52	70.53	81.87	56.16	86.63	60.58	86.00
	ReAct	NeurIPS 2021	19.41	96.41	27.44	93.94	37.43	91.42	51.36	89.34	33.91	92.78
	GradNorm	NeurIPS 2021	24.54	94.06	43.71	87.49	56.60	82.17	35.87	89.44	40.18	88.29
	ViM	CVPR 2022	73.10	87.12	83.08	79.23	83.14	77.10	**13.55**	**97.18**	63.22	85.16
	MaxLogit	ICML 2022	48.30	91.31	66.34	84.40	70.85	81.97	58.43	86.34	60.98	86.00
	KL Matching	ICML 2022	50.68	88.91	79.62	78.47	79.89	77.23	67.19	83.08	69.34	81.92
	Ours	/	**11.67**	**97.71**	**22.05**	**95.22**	**32.79**	**92.58**	29.42	93.63	**23.98**	**94.79**
MobileNet-V2	MSP	ICLR 2017	62.70	85.75	79.22	76.11	81.14	75.60	73.16	78.64	74.05	79.03
	Mahalanobis	NeurIPS 2018	99.42	26.67	99.33	24.10	99.02	27.49	64.26	77.08	90.51	38.83
	ODIN	ICLR 2018	55.39	87.62	54.07	85.88	57.36	84.71	49.96	85.03	54.20	85.81
	Energy	NeurIPS 2020	51.68	90.75	64.89	84.01	70.87	80.98	57.40	86.52	61.21	85.57
	ReAct	NeurIPS 2021	41.87	92.68	58.77	85.69	65.53	82.74	44.71	90.60	52.72	87.93
	GradNorm	NeurIPS 2021	33.46	92.59	41.86	89.84	56.24	84.23	31.34	92.70	40.72	89.84
	ViM	CVPR 2022	93.27	73.57	94.89	66.25	94.35	65.10	23.41	**95.31**	76.48	75.06
	MaxLogit	ICML 2022	53.16	90.75	68.53	83.42	72.96	80.75	60.43	85.96	63.77	85.22
	KL Matching	ICML 2022	56.60	87.11	84.44	73.93	83.13	73.74	69.53	81.58	73.43	79.09
	Ours	/	**26.77**	**95.07**	**40.67**	**90.63**	**53.77**	**86.59**	**22.69**	95.11	**35.98**	**91.85**

Table 1: ImageNet benchmark comparison results. Comparison with existing post-hoc OOD detection methods. With respect to each model, the model weight used by all methods are the same. The best performance is bold and the second best is underlined. ↑ means that larger values are better and ↓ indicates that smaller values are better.

ID data	Methods	Conferences	iSUN		LSUN (crop)		LSUN (resize)		SVHN		Textures		Places		Average	
			FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR-10	MSP	ICLR 2017	54.57	92.11	45.27	93.79	51.96	92.72	59.51	91.25	62.35	89.22	62.35	89.22	56.00	91.39
	ODIN	ICLR 2018	99.42	26.67	99.33	24.10	99.02	27.49	64.26	77.08	64.26	77.08	64.26	77.08	81.76	51.58
	Energy	NeurIPS 2020	27.52	95.59	**10.19**	**98.05**	23.47	96.14	53.96	91.32	46.65	91.15	46.65	91.15	34.74	93.90
	ReAct	NeurIPS 2021	**21.15**	96.47	23.03	95.96	18.22	96.98	46.50	92.44	**45.10**	**91.95**	**45.10**	**91.95**	33.18	94.29
	GradNorm	NeurIPS 2021	54.14	89.34	22.55	96.21	48.94	90.67	82.44	79.85	71.25	80.69	71.25	80.69	58.43	86.24
	ViM	CVPR 2022	27.52	95.59	**10.19**	**98.05**	23.47	96.14	53.96	91.32	46.65	91.15	46.65	91.15	34.74	93.90
	MaxLogit	ICML 2022	28.37	95.51	10.88	97.93	24.32	96.06	53.51	91.39	47.00	91.13	47.00	91.13	35.18	93.86
	KL Matching	ICML 2022	51.93	88.95	44.14	93.01	49.35	90.24	59.39	88.20	61.10	84.43	61.10	84.43	54.50	88.21
	Ours		20.76	**96.54**	21.26	96.24	**17.69**	**97.06**	**45.89**	**92.50**	45.40	91.89	45.40	91.89	**32.73**	**94.35**
CIFAR-100	MSP	ICLR 2017	81.90	76.56	81.90	76.56	81.90	76.56	81.70	77.80	81.90	76.56	81.90	76.56	81.87	76.77
	ODIN	ICLR 2018	76.66	**83.51**	**28.72**	**94.51**	79.61	82.13	40.94	93.29	83.63	72.37	87.71	71.46	66.21	82.88
	Energy	NeurIPS 2020	80.05	79.19	80.05	79.19	80.05	79.19	81.24	84.59	80.05	79.19	80.05	79.19	80.25	80.09
	ReAct	NeurIPS 2021	73.00	81.74	73.00	81.74	73.00	81.74	70.64	88.24	73.00	81.74	73.00	81.74	72.61	82.82
	GradNorm	NeurIPS 2021	80.85	71.25	80.85	71.25	80.85	71.25	57.61	87.77	80.85	71.25	80.85	71.25	76.98	74.00
	ViM	CVPR 2022	80.05	79.19	80.05	79.19	80.05	79.19	81.24	84.59	80.05	79.19	80.05	79.19	80.25	80.09
	MaxLogit	ICML 2022	79.60	79.23	79.60	79.23	79.60	79.23	80.31	84.45	79.60	79.23	79.60	79.23	79.72	80.10
	KL Matching	ICML 2022	80.00	76.74	80.00	76.74	80.00	76.74	75.37	79.63	80.00	76.74	80.00	76.74	79.23	77.22
	Ours		**66.75**	83.17	66.75	83.17	**66.75**	**83.17**	**26.42**	**95.52**	**66.75**	**83.17**	**66.75**	**83.17**	**60.03**	**85.23**

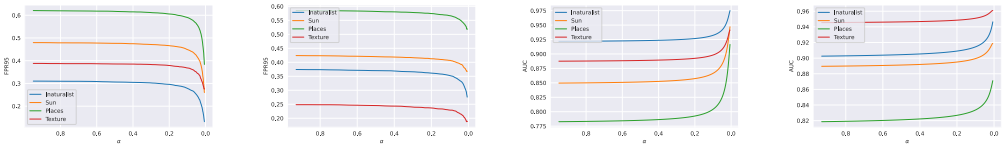
Table 2: CIFAR benchmark comparison results. Comparison with existing post-hoc OOD detection methods. Note that all methods are based on the same weights of ResNet18 (He et al., 2016). The best performance is bold and the second best is underlined. ↑ means that larger values are better and ↓ indicates that smaller values are better.

evaluate our method on three other OOD datasets including iSUN (Xu et al., 2015), LSUN (Yu et al., 2015), and SVHN (Netzer et al., 2011). Below, we introduce each of them in more detail: (1) **LSUN** dataset contains 10,000 images of 10 scene categories. Following (Sun et al., 2021; Liu et al., 2020; Hendrycks & Gimpel, 2016), to conduct OOD detection on this dataset, we randomly crop images in this dataset to size 32×32. Besides, following (Sun et al., 2021; Liu et al., 2020; Hendrycks & Gimpel, 2016), we also conduct OOD detection on a variant of this dataset (**LSUN_Resize**) by resizing images in **LSUN** dataset to size 32×32. (2) **iSUN** dataset is sampled from the SUN (Xiao et al., 2010) dataset, and contains 20,608 images of 397 categories. Following (Sun et al., 2021; Liu et al., 2020; Hendrycks & Gimpel, 2016), the whole dataset is used for OOD detection evaluation. (3) **SVHN** dataset contains 26,032 images of 10 categories for testing. Following (Sun et al., 2021; Liu et al., 2020; Hendrycks & Gimpel, 2016), we use all the 26,032 images for OOD detection evaluation.

Results In Tab. 2, we compare our method with the existing post-hoc OOD detection methods on all the six OOD datasets. As shown, our method demonstrates the best averaged result compared with common post-hoc OOD detection methods, which demonstrates the effectiveness of our method.

6.3 Ablation Studies

Effect of α. In the previous section, we analyzed the effect of α on the performance of OOD detection from a mathematical point of view and concluded that a smaller α has a positive effect on performance. In this subsection, we will experimentally show the impact of α on the performance



(a) FPR95 change with α under ResNet structure (b) FPR95 change with α under MobileNet structure (c) AUROC change with α under ResNet structure (d) AUROC change with α under MobileNet structure

Figure 3: **Ablation results.** The smaller α brings more benefits. Experiments are conducted on ResNet50 trained on the ImageNet dataset. We demonstrate various α on different OOD datasets. A lower FPR95 and a higher AUROC indicate better OOD detection performance.

Methods	iNaturalist		SUN		Places		Textures		Average	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	53.40	88.01	73.68	79.83	76.12	78.74	68.88	80.54	68.02	81.78
MSP + Ours	24.87	94.16	43.22	87.92	56.26	82.90	36.82	89.46	40.29	88.61
Energy	50.60	90.96	65.03	84.52	70.53	81.87	56.16	86.63	60.58	86.00
Energy + Ours	21.96	95.29	40.38	89.84	53.65	85.86	33.53	91.11	37.38	90.52
ReAct	19.41	96.41	27.44	93.94	37.43	91.42	51.36	89.34	33.91	92.78
ReAct + Ours	11.67	97.71	22.05	95.22	32.79	92.58	29.42	93.63	23.98	94.79
ViM	73.10	87.12	83.08	79.23	83.14	77.10	13.55	97.18	63.22	85.16
ViM + Ours	20.78	95.67	39.76	90.09	54.04	84.95	16.30	96.18	32.72	91.72
MaxLogit	48.30	91.31	66.34	84.40	70.85	81.97	58.43	86.34	60.98	86.00
MaxLogit + Ours	27.33	93.48	44.97	87.39%	58.22	81.90	37.64	89.02	42.04	87.95
KL Matching	50.68	88.91	79.62	78.47	79.89	77.23	67.19	83.08	69.34	81.92
KL Matching + Ours	42.32	91.67	73.31	83.06	75.75	80.94	58.84	86.85	62.55	85.63

Table 3: **Ablation results.** The compatibility with the existing post-hoc OOD detection methods. Under the ResNet50 model trained on the ImageNet, we evaluate different existing post-hoc OOD detection methods with and without ABC. ↑ means that larger values are better and ↓ indicates that smaller values are better.

of OOD detection. We randomly sample α from a continuous uniform distribution between 0 and 1 i.e. $\alpha \in U_{[0,1]}$. And then we evaluate FPR95 and AUROC under various α on the iNaturalist, SUN, Places and Textures OOD datasets (Van Horn et al., 2018; Xiao et al., 2010; Zhou et al., 2017; Cimpoi et al., 2014) with ResNet50 (He et al., 2016) and MobileNet-V2 (Sandler et al., 2018) trained on ImageNet (Deng et al., 2009). The result is shown in the Fig. 3. As shown, as long as α decreases, a larger AUROC and a smaller FPR95 are consistently achieved throughout various OOD datasets, demonstrating the effectiveness of our method.

Effect of different baseline methods. To validate the general effectiveness of our proposed method, we apply our method on various different post-hoc OOD detection methods, including MSP, energy, react, vim, MaxLogit, and KL-Matching (Liu et al., 2020; Sun et al., 2021; Wang et al., 2022; Hendrycks et al., 2019). As shown in Tab. 3, our method achieves consistent performance improvement when applied on various different post-hoc OOD detection methods. This demonstrates that our proposed method can be flexibly applied on various post-hoc OOD detection methods to improve their performance.

7 Conclusion

In this paper, we present a simple yet effective OOD detection method, which replaces the trained weight of the last FC layer with a small value. We theoretically analyze that the proposed method can make the ID data and OOD data to be more separable, and thus better cope with the overconfidence problem. We shows two ablation experiments to show that our method is compatible with existing OOD detection methods and achieves consistent performance improvement. Our method achieves superior performance on the ImageNet and CIFAR OOD detection benchmarks.

REFERENCES

S Basha, Sravan Kumar Vinakota, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Autofcl: Automatically tuning fully connected layers for transfer learning. *arXiv preprint arXiv:2001.11951*, 2020a.

SH Shabbeer Basha, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378:112–119, 2020b.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Qichao Chen, Wenjie Jiang, Kuan Li, and Yi Wang. Improving energy-based out-of-distribution detection by sparsity regularization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 539–551. Springer, 2022.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

Sven Elflein, Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. On out-of-distribution detection with energy-based models. *arXiv preprint arXiv:2107.08785*, 2021.

Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

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

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

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

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Abhishek Joshi, Sathish Chalasani, and Kiran Nanjunda Iyer. Semantic driven energy based out-of-distribution detection. *arXiv preprint arXiv:2208.10787*, 2022.

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

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

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

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

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pp. 15650–15665. PMLR, 2022.

Govind S Mudholkar and Alan D Hutson. The epsilon–skew–normal distribution for analyzing near-normal data. *Journal of statistical planning and inference*, 83(2):291–309, 2000.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Yawen Ouyang, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. Energy-based unknown intent detection with data manipulation. *arXiv preprint arXiv:2107.12542*, 2021.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022.

Francesco Tonin, Arun Pandey, Panagiotis Patrinos, and Johan AK Suykens. Unsupervised energy-based out-of-distribution detection using stiefel-restricted kernel machine. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

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

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Energy-based out-of-distribution detection for multi-label classification. 2020.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 34:29074–29087, 2021.

Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: a unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 62–71, 2021.

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

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.

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

Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9518–9526, 2019.

Karina Zadorozhny, Patrick Thoral, Paul Elbers, and Giovanni Cinà. Out-of-distribution detection for medical applications: Guidelines for practical evaluation. *arXiv preprint arXiv:2109.14885*, 2021.

Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9452–9461, 2021.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13208–13217, 2020.

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

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9719–9728, 2020.