# Decoupling MaxLogit for Out-of-Distribution Detection

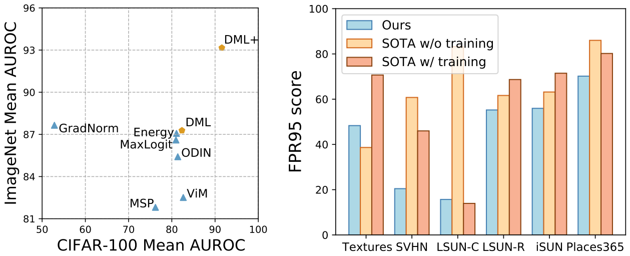
Zihan Zhang [*] and Xiang Xiang [*]
Key Lab of Image Processing and Intelligent Control, Ministry of Education
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology, China

## Abstract

*In machine learning, it is often observed that standard training outputs anomalously high confidence for both in-distribution (ID) and out-of-distribution (OOD) data. Thus, the ability to detect OOD samples is critical to the model deployment. An essential step for OOD detection is post-hoc scoring. MaxLogit is one of the simplest scoring functions which uses the maximum logits as OOD score. To provide a new viewpoint to study the logit-based scoring function, we reformulate the logit into cosine similarity and logit norm and propose to use MaxCosine and MaxNorm. We empirically find that MaxCosine is a core factor in the effectiveness of MaxLogit. And the performance of MaxLogit is encumbered by MaxNorm. To tackle the problem, we propose the Decoupling MaxLogit (DML) for flexibility to balance MaxCosine and MaxNorm. To further embody the core of our method, we extend DML to DML+ based on the new insights that fewer hard samples and compact feature space are the key components to make logit-based methods effective. We demonstrate the effectiveness of our logit-based OOD detection methods on CIFAR-10, CIFAR-100 and ImageNet and establish state-of-the-art performance.*

## 1. Introduction

In real-world applications, the closed-world assumption does not always hold where all the classes in the test phase would be available in the training phase. Out-of-distribution (OOD) detection [11] is a natural and challenging setting, and there is an open space containing outliers not belonging to any training classes. When the model is deployed in practice, OOD data often come from the open world [17]. Thus, it is crucial for a trustworthy model to not only produce accurate predictions on in-distribution (ID) data, but also distinguish the OOD data and reject them. However, the machine-learning model easily produces over-confident wrong predictions on OOD data [25]. For instance, a model may wrongly detect the zebra as a horse with high confidence when the zebra is not in the training set.



(a) AUROC of different methods

(b) FPR95 of SOTA methods

Figure 1. AUROC and FPR95 (in percentage) of previous OOD detection methods on ImageNet and CIFAR-100. (a) shows the AUROC (higher is better) of our methods (orange pentagons) and other methods (blue rectangles). (b) shows the FPR95 (lower is better) of our methods and SOTA methods w/ (LogitNorm [37]) and w/o (ViM [36]) training on CIFAR-100.

A key of the OOD detection algorithm is a scoring function that maps the input to the OOD score, indicating to what extent the sample is an OOD sample. Various scoring functions have been proposed to seek the properties that better distinguish OOD samples. The OOD score is calculated mainly from the output of the model, including features [20, 30, 36], logits [10, 11, 23]. For example, MSP [11] uses the maximum Softmax probabilities, and MaxLogit [10] uses the maximum logits as the OOD score.

MaxLogit and MSP are two simplest scoring functions that do not require extra computational costs. In contrast, other methods require extra storage [30], or extra computational cost [14]. However, the logit-based methods MSP and MaxLogit are not state-of-the-art (SOTA). Intuitively, the simple logit-based method could achieve comparable performance as other complex scoring methods, because logit contains high-level semantic information. We hypothesize some underlying reasons limit the performance.

To revitalize the simple logit-based method, we start the work by analyzing the reasons which cause the performance gap between MSP and MaxLogit. The gap may be due to

[*] Equal contribution, co-first author; also with Nat. Key Lab of MSIIPT. Correspondence to Xiang Xiang (`xex@hust.edu.cn`)

the softmax operation normalizing out much feature norm information of the logits. To delve into the effect of feature norm, we divide the logit into two parts: (1) the cosine similarity between the features and classifier; (2) the feature norm. We discard the classifier weight norm because the norm is identical after the model coverage [27]. We use the top value of the two as the OOD score, named MaxCosine and MaxNorm. Therefore, MaxLogit is a coupled form.

We find that MaxCosine outperforms MaxLogit with the same model and MaxNorm performs much worse than MaxLogit. Thus, MaxLogit (1) is encumbered by MaxNorm, (2) suppresses the effectiveness of MaxCosine, and (3) restricts the flexibility to balance MaxCosine and MaxNorm. The three problems are the bottleneck of MaxLogit. To tackle the problem, we propose Decoupling MaxLogit (DML) for flexibility to balance MaxCosine and MaxNorm. DML decouples the MaxCosine from the equal coefficient with MaxNorm by replacing it with a constant.

The decoupling method solves the second and third problems but still leaves the first problem unsolved. Although MaxNorm helps DML to outperform MaxCosine, the improvement is marginal due to the low performance of MaxNorm. Therefore, we study the role of model training and show that a simple modification to standard training could significantly boost MaxNorm and MaxCosine for OOD detection. Specifically, a feature space with fewer hard samples benefits MaxCosine and a compact feature space benefits MaxNorm. Also, the normalized feature and classifier are the key to the success of the logit-based methods. These findings are not discussed in prior works. We extend DML to DML+ based on the above new insights to further boost the DML performance as shown in Fig. 1.

We summarize our contributions as follows.

- To overcome the limitations of MaxLogit, we propose a post-hoc scoring method DML, which decouples MaxLogit for flexibility to balance MaxCosine and MaxNorm. DML outperforms MaxLogit and achieves comparable performance with SOTA methods.
- We offer new insights into the key components to make MaxCosine, MaxNorm and DML effective, including replacing the standard linear classifier with a cosine classifier and different training losses. The findings are supported by empirical results and theoretical analysis. We also prove that the findings could greatly boost the performance of existing OOD scoring methods.
- Based on the insights, we extend DML to DML+ which changes the standard training. Significant improvements on CIFAR and ImageNet have shown its effectiveness.

## 2. Related Work

**OOD Detection** has attracted growing research attention for the safe deployment of the model in the real world. The major branch of OOD detection is classification-based methods [41, 42] including confidence enhancement methods [1,31], outlier exposure [2,7,26,44] and post-hoc detection [10–12,14,21,23,29,36]. The outlier exposure methods require auxiliary OOD data as the outlier to help the model learn the ID or OOD discrepancy. For example, NMD [7] finds that activation means of OOD data and ID data are deviated and trains the OOD detector with OOD and ID data.

An active line of research of OOD detection without extra data is post-hoc detection. It is easy to use the post-hoc methods when given a trained model. MaxLogit [10] experiments with the maximum logit value of the samples. ODIN [21] finds that large temperature scaling and input perturbation help distinguish the OOD samples. Energy [23] proposes using an energy score based on the logsumexp of the logits for OOD detection. GradNorm [14] uses the gradients of KL divergence between the softmax output and a uniform distribution as evidence for ID and OOD distinction. A recent work ViM [36] combines the information from both the feature space and the logits for OOD detection.

Another line of work that detects OOD without extra data involves training. VOS [8] adaptively synthesizes virtual outliers from the low-likelihood region in the feature space to reduce the confidence of the model for the detection task. LogitNorm [37] fixes the widely used cross-entropy loss by normalizing the logits in the training phase.

**Normalization in deep learning.** Normalization is widely adopted in deep learning including face recognition [16, 34, 35, 47], self-supervised learning [3, 18], *etc.* NormFace [35] normalizes the feature and classifier to boost face verification performance. SupCon [18] and SimCLR [3] use cosine similarity to measure the similarity between different samples. In the literature on OOD detection, LogitNorm [37] normalizes the logit value during training time and uses the unnormalized logits to calculate the OOD score. Normalization on logits differs from the cosine similarity because of the cosine value difference. And LogitNorm can be easily combined with common OOD scoring functions.

## 3. Preliminaries

We consider a neural network for the $K$-class classification task, which can be represented as

$$\boldsymbol{f}(\boldsymbol{x};\boldsymbol{W}_{full}) = \boldsymbol{b}_L + \boldsymbol{W}_L\delta(\cdots\delta(\boldsymbol{b}_1 + \boldsymbol{W}_1\boldsymbol{x})\cdots), \quad (1)$$

where $\boldsymbol{W}_{full} = \{\boldsymbol{W}_1, \cdots, \boldsymbol{W}_L\}$ denotes the weights of the $L$ layers, $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L\}$ denotes the biases, and $\delta(\cdot)$ is the nonlinear activation function. Given the data $\boldsymbol{x}_{k,i}$ belonging to class $k$, we define the last-layer features as $\boldsymbol{h}_{k,i} \in \mathbb{R}^d$, $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{W}_{full}) = \boldsymbol{b}_L + \boldsymbol{W}_L\boldsymbol{h}_{k,i}$. The later analysis does not include the bias term for simplicity. Then, the logit is $\boldsymbol{z}_{k,i} = \boldsymbol{W}_L\boldsymbol{h}_{k,i}$ where $\boldsymbol{W}_L = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K]^\top$.

Given a training set $\mathcal{D}_{tr} = \{(\boldsymbol{x}_{k,i}, k)\}_{i=1}^N$ with $N$ training samples and $K$ classes from an underlying distribution

$\mathcal{P}_{tr}$, we first train a model on the training set. The goal of OOD detection is to discriminate if a given sample is from $\mathcal{P}_{tr}$ or another data distribution. Therefore, two keys of OOD detection are (1) training a model robust to OOD samples, *i.e.*, easier to distinguish the ID and OOD samples and (2) designing a score function so that samples with a smaller score are classified as OOD.

We define two metrics to measure feature collapse [48]: Within-class Feature Convergence (WFC) and Class mean Feature Convergence to the corresponding classifier (CFC).

$$\text{WFC} := \frac{\text{trace}(\Sigma_W \Sigma_B^{\dagger})}{K}, \tag{2}$$

$$\text{CFC} := \sum_{k=1}^{K} \left|\left| \frac{\overline{\boldsymbol{h}}_k}{||\boldsymbol{h}||_F} - \frac{\boldsymbol{w}_k}{||\boldsymbol{W}||_F} \right|\right|, \tag{3}$$

where $\dagger$ denotes the pseudo-inverse,$\boldsymbol{h}$ is the feature matrix of all samples. $\overline{\boldsymbol{h}}_k$ and $\overline{\boldsymbol{h}}$ are the mean of class $k$ features and all features respectively, $\Sigma_W = \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}(\boldsymbol{h}_{k,i} - \overline{\boldsymbol{h}}_k)(\boldsymbol{h}_{k,i} - \overline{\boldsymbol{h}}_k)^{\top}$ and $\Sigma_B = \frac{1}{K}\sum_{k=1}^{K}(\overline{\boldsymbol{h}}_k - \overline{\boldsymbol{h}})(\overline{\boldsymbol{h}}_k - \overline{\boldsymbol{h}})^{\top}$. More details are included in the *supplementary materials*.

# 4. Methods

## 4.1. Rethinking MaxLogit

The MSP score of a sample is the maximum Softmax value of the sample: $\max(\text{Softmax}(\boldsymbol{z}_{k,i}))$. The MaxLogit score of a sample is the maximum logit value of the sample: $\max(\boldsymbol{z}_{k,i})$. Recall that $\boldsymbol{z}_{k,i} = [z_{i1}, z_{i2}, ..., z_{iK}]$.

MaxLogit outperforms MSP on different datasets. The monotonically-increasing function transforms on the score function (*e.g.*, $\log(\cdot)$ and $\exp(\cdot)$) do not affect the OOD detection performance. Therefore, the only difference between MSP and MaxLogit is the sum item $\Sigma_{j=1}^{K}\exp(z_{ij})$. After the model coverage, the sum item is primarily affected by the feature norm. Therefore, the difference between MSP and MaxLogit mainly comes from the feature norm. This motivates us to investigate how the cosine similarity and the feature norm affect OOD detection performance.

We decouple the MaxLogit into two parts: MaxCosine and MaxNorm. Given a sample $\boldsymbol{x}_{k,i}$, MaxCosine and MaxNorm can be formulated as

$$\text{MaxCosine} : \max(\cos < \boldsymbol{h}_{k,i}, \boldsymbol{w}_j >)_{j=1}^{K}, \tag{4}$$

$$\text{MaxNorm} : ||\boldsymbol{h}_{k,i}||. \tag{5}$$

The MaxLogit score equals the MaxCosine score multiplied by the MaxNorm score. As we explained, applying an increasing function transform on the score does not affect the OOD detection performance. Thus, MaxLogit can be formulated with two individual parts $\log(\max(\boldsymbol{z}_{k,i})) = \log(\max(\cos < \boldsymbol{h}_{k,i}, \boldsymbol{w}_j >)) + \log|\boldsymbol{h}_{k,i}| + \log|\boldsymbol{w}|$, which is a coupled form of MaxCosine and MaxNorm. Note that the norm of the classifier weight $\boldsymbol{w}_j$ is almost identical after model coverage [27], so we use a constant $|\boldsymbol{w}|$ to replace it.

| Methods | Model1 | Model2 | Model3 |
|---|---|---|---|
| MSP | 76.21 | 82.06 | 74.92 |
| MaxCosine | 81.78 | 81.52 | 78.96 |
| MaxNorm | 56.93 | 43.85 | 64.88 |
| MaxLogit | 80.96 | 80.18 | 78.98 |
| DML | 82.34 (+1.38) | 83.14 (+2.96) | 80.34 (+1.36) |

Table 1. The OOD detection AUROC results on CIFAR-100 with WRN-40-2. The model with a linear classifier is trained with different losses, the same as that in Fig. 6. All values are percentages.

By the reformulation of MaxLogit, we empirically find that MaxCosine outperforms MaxLogit with the same model and MaxNorm performs much worse than MaxLogit. As Table 1 shows, MaxCosine outperforms MaxLogit by around 0.8% on different models. Meanwhile, MaxNorm performs worse than MaxCosine by around 30%. As a coupled form, MaxLogit is dragged down by MaxNorm. When the MaxNorm performs worse on Model2, MaxCosine outperforms MaxLogit by over 1.3%, which is the largest performance gap between the three models. When MaxNorm performs better on Model3, MaxLogit outperforms MaxCosine by 0.03%, which indicates that MaxNorm is complementary to MaxCosine.

Based on the above analysis, we propose a new OOD detection method termed Decoupling MaxLogit (DML). DML can be written as

$$\text{DML} = \lambda\text{MaxCosine} + \text{MaxNorm}, \tag{6}$$

where $\lambda$ is a hyper-parameter. We normalize MaxCosine and MaxNorm scores by their sum on ID data respectively so that their importance can be separately considered. The weight $\lambda$ is tuned on Gaussian noise. As Table 1 shows, DML outperforms MaxLogit by over 1.36%.

## 4.2. Improving MaxCosine and MaxNorm

Although MaxNorm helps DML to outperform MaxCosine, the improvement is marginal due to the low performance of MaxNorm. Therefore, we study the role of model training and show that a simple modification to standard training could significantly boost the performance of MaxNorm and MaxCosine. The experimental setting is the same as in Table 4. We report the mean area under the ROC curve (AUROC) on six OOD test datasets.

*Cosine classifier leads to better MaxCosine and MaxNorm, also logit-based methods.* When considering the norm and cosine similarity, an intuitive idea is to use a cosine classifier. The cosine classifier optimizes the cosine similarity in the training phase. As Table 2 shows, the cosine classifier substantially increases the performance of all

| Classifier | DML | MaxL | MaxC | MaxN | CFC↓ | WFC↓ |
|---|---|---|---|---|---|---|
| Linear | 82.34 | 80.96 | 81.78 | 56.93 | 0.789 | 3778 |
| Cosine | 85.34 | 82.53 | 84.14 | 76.39 | 0.618 | 396 |

Table 2. The AUROC of OOD detection and CFC / WFC on CIFAR-100 with WRN-40-2 and CE loss. MaxL: MaxLogit, MaxC: MaxCosine and MaxN: MaxNorm.

| CLS | Loss | DML | MaxL | MaxC | MaxN | CFC↓ | WFC↓ |
|---|---|---|---|---|---|---|---|
| L | Center | 80.34 | 78.98 | 78.96 | 64.88 | 0.69 | 2106 |
| | Focal | 83.14 | 80.18 | 81.52 | 43.85 | 0.75 | 4115 |
| C | Center | 89.86 | 89.62 | 77.40 | 89.85 | 1.17 | 251 |
| | Focal | 89.38 | 83.41 | 90.90 | 69.85 | 0.47 | 554 |

Table 3. The AUROC of OOD detection of and CFC/WFC of different loss and classifiers on CIFAR-100 with WRN-40-2. CLS: classifier, L: linear classifier, C: cosine classifier.

four methods. Note that the improvement of MaxNorm is much larger than that of MaxCosine. This may be due to the linear classifier could minimize the loss by increasing the feature norm of easy samples (*i.e.*, samples with high confidence) and ignoring the hard samples. This tendency is also noted in [28].

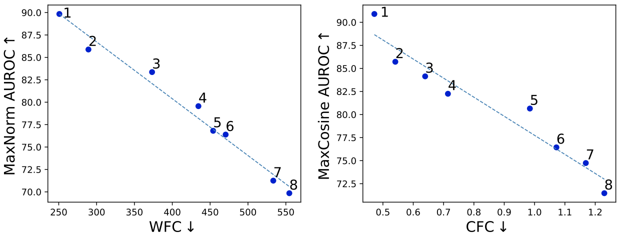
We also analyze it from a statistical view. As Table 2 and 3 show, both the WFC and CFC scores decrease when using the cosine classifier. WFC indicates the with-class feature convergence and WFC is lower when the features are more compact. CFC measures the extent of the features' convergence to the corresponding classifier. Intuitively, the features having similar norms could lead to better MaxNorm. And the features closer to the corresponding classifier could lead to better MaxCosine. All the WFC and CFC are calculated on the training set.

**Lower WFC leads to better MaxNorm.** One approach to improve the WFC of the model is Center loss [38], naturally. The Center loss can be formulated as

$$\mathcal{L}_{center} = \sum_{k=1}^{K} \sum_{i=1}^{n} ||\boldsymbol{h}_{k,i} - \mathcal{C}_k||_2, \tag{7}$$

where $\mathcal{C}_k$ is the mean feature of corresponding class $k$. Center loss is combined with CE loss with a weighting hyper-parameter as in [38]. Center loss urges the features to be more clustered and the features of identical classes have more similar feature norms. We use WFC to quantify that and suppose the model with lower WFC has better MaxNorm performance. To verify the assumption, we take different loss functions with different hyper-parameter and train the WRN-40-2 on CIFAR-100.

We can improve MaxNorm's performance by decreasing the WFC. Fig. 2 (a) shows the correlation between WFC and MaxNorm. For example, CE loss (scatter point 6); Focal loss with different $\gamma$ (5,7,8); Center loss with different weight (1-4). The figure also shows that Center loss helps the model to gain a lower WFC score than both CE and Focal loss. Details can be found in *supplementary materials*.



(a) MaxNorm and WFC. (b) MaxCosine and CFC.

Figure 2. Gains in OOD detection performance of WRN-40-2 (cosine classifier) as CFC or WFC score increases on CIFAR-100.

**Lower CFC leads to better MaxCosine.** MaxCosine uses cosine similarity to identify OOD samples. Thus, when there are fewer ID samples in the low-likelihood region (*i.e.*, fewer hard samples), the performance of MaxCosine could be better. A method to tackle hard samples is hard sample mining. The Focal loss [22] is one of the leading works in this area and can be written as

$$\mathcal{L}_{focal} = -\sum_{k=1}^{K} \sum_{i=1}^{n} (1 - p_{k,i})^{\gamma} \log(p_{k,i}), \tag{8}$$

where $\gamma$ is a hyper-parameter and $p_{k,i}$ is the softmax score.

Fig. 2 (b) shows the correlation between CFC and MaxCosine. For example, CE loss (scatter point 3); Focal loss with different $\gamma$ (1,2); Center loss with different weight (4-8). The figure shows that lower CFC leads to better MaxCosine. Also, training with Focal loss could lead to smaller CFC than both Center loss and CE loss.

In general, we can conclude: (1) compared with a linear classifier, a cosine classifier is more robust to OOD samples; (2) there exists correlations between WFC and MaxNorm, and between CFC and MaxCosine; (3) training with Center loss leads to larger CFC and smaller WFC, while the opposite is true for Focal loss.

### 4.3. Theoretical analysis of WFC and CFC

Below, we reveal the lower bound of WFC and CFC along with the loss optimization.

**Proposition 1** (Lower Bound of WFC and CFC) *For the normalized* $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_K$ *and* $\boldsymbol{h} \in \mathbb{R}^d, (\boldsymbol{z}_{k,i})_j = \boldsymbol{w}_j^\top \boldsymbol{h}_{k,i} \in \mathbb{R}$, *CE loss is bounded by*

$$\mathcal{L}_{CE} \geq n \log\Big(1 + (K-1)\exp\Big(-\frac{K\sqrt{K}}{K-1}||\boldsymbol{W}||_F||\boldsymbol{h}||_2\Big)\Big).$$

*When the equality holds, WFC and CFC reach the lower bound: WFC, CFC ≥0.*

The proof is provided in *supplementary materials*. From Proposition 1, we find that the optimum of WFC and CFC

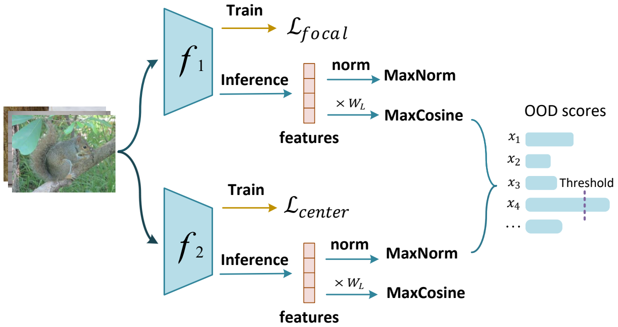


Figure 3. The pipeline of our method. We train two individual models with Focal loss and Center loss. In inference, we calculate MaxNorm on the second model and MaxCosine on the first model. Then we couple the two scores into DML to detect OOD samples.

and CE loss are obtained at the same condition. That indicates the benefit of lower CFC and WFC, which is also illustrated empirically in Table 3.

When CE loss reaches the optimum, Center loss and Focal loss also reach the optimum, *i.e.*, the features collapse to the center and the weighting strategy of Focal loss does not change the optimum. However, during the training process, the optimum is hard to reach. Thus, the two loss functions improve the OOD performance by facilitating the training process. For example, training with Center loss will have lower WFC but higher CFC than CE loss. Similarly, cosine classifier focus on the optimization of cosine similarity. The model can not decrease the loss by increasing the feature magnitude of easy samples to 'escape' the hard samples, which leads to lower WFC and CFC as shown in Table 2.

Although the cosine classifier and loss functions do not require extra training costs, the OOD detection performance greatly benefits from decreasing WFC and CFC.

### 4.4. Decoupling MaxLogit + (DML+)

DML is a post-hoc scoring function that is model agnostic and training scheme agnostic. In this section, we improve DML based on the key insights which make MaxCosine and MaxNorm effective in practice.

To further improve DML, a robust method is to apply MaxCosine on the cosine classifier model trained with Focal loss and apply MaxNorm on the cosine classifier model trained with Center loss. This method does not require tuning $\lambda$ in DML on each model. As we explained, the MaxNorm and MaxCosine are complementary. Thus, as shown in Fig. 3, we extend DML to DML+ as

$$\text{DML+} = \lambda \text{MaxCosine}_F + \text{MaxNorm}_C, \qquad (9)$$

where $\text{MaxCosine}_F$ means MaxCosine applied on the Focal model and $\text{MaxNorm}_C$ means MaxNorm applied on the Center model. We use MCF to represent $\text{MaxCosine}_F$ and MNC to represent $\text{MaxNorm}_C$ for convenience.

Notably, DML+ offers several compelling advantages: (1) The MCF model and MNC model have similar performance. Thus, DML+ is free from hyper-parameter tuning, where we set $\lambda = 1$ for all experiments. (2) DML+ is robust to different OOD samples because DML+ draws on the strengths of both models. (3) DML+ is OOD agnostic and does not rely on the information of OOD samples.

## 5. Experiments

### 5.1. Experimental Settings

**In-distribution datasets.** We use three common benchmarks CIFAR-10 and CIFAR-100 [19] and ImageNet [5].
**Out-of-distribution datasets.** For CIFAR, we use six benchmarks as OOD datasets following the work [37] including Textures [4], SVHN [24], LSUN-Crop and LSUN-Resize [43], iSUN [40] and Places365 [46]. For ImageNet, we use four common benchmarks as OOD datasets following the work [30] including iNaturalist [32], SUN [39], Places365, and Textures. The evaluations include a diverse range of domains. Details of datasets can be found in *supp*.
**Evaluation metrics.** We use two commonly-adopted metrics to evaluate our methods. The AUROC is a threshold-free metric to describe the performance of a model. Notably, a higher AUROC is better. FPR95 is the false-positive rate of OOD data when the true-positive rate of ID data is 95%, and a smaller FPR95 is better. For all experiments, we report both scores in percentage. All reported scores are averaged over ten runs.
**Hyper-parameters.** In DML, we tune $\lambda$ based on the OOD detection performance on Gaussian noise. In DML+, we set $\lambda = 1$ for all experiments.
**Training details.** The baseline methods we reproduced are all experimented on the model trained with CE loss and linear classifier, as in the original paper report. The scale parameter of the cosine classifier is set to 40 for all experiments. We set $\gamma = 2$ for Focal loss.

We experiments several model architectures including WRN-40-2 [45], ResNet34 and ResNet50 [9] and DenseNet [13]. For CIFAR-10 and CIFAR-100, all the models are trained for 200 epochs using SGD optimizer with a momentum of 0.9 and the cosine learning rate scheduler, which gradually decays the learning rate from 0.1 to 0. The weight decay is $5 \times 10^{-4}$, and the batch size is 128. For ImageNet, we train the ResNet50 from scratch for 90 epochs using the same SGD and learning rate scheduler as on CIFAR. The weight decay is $5 \times 10^{-4}$, and the batch size is 256.

### 5.2. Comparison with SOTA

Yang *et al.* [41] experiment on 22 OOD detection methods with identical settings. We choose the SOTA methods based on that result for different datasets. For OOD detection methods without extra data and with training, we choose LogitNorm [37] and GODIN [12]. For OOD de-

| Methods | Textures |  | SVHN |  | LSUN-C |  | LSUN-R |  | iSUN |  | Places365 |  | Average |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ |
| MSP | 74.24 | 84.43 | 77.17 | 80.66 | 84.26 | 66.30 | 73.37 | 81.98 | 73.04 | 82.49 | 75.20 | 82.69 | 76.21 | 79.76 |
| ODIN* | 75.60 | 80.23 | 79.60 | 83.52 | 93.01 | 37.45 | 83.51 | 69.69 | 81.01 | 74.47 | 75.55 | 78.93 | 81.38 | 70.71 |
| Energy | 75.80 | 82.23 | 83.92 | 75.72 | 93.53 | 37.25 | 79.05 | 76.02 | 78.48 | 78.38 | 75.71 | 82.58 | 81.08 | 72.03 |
| ViM | 91.25 | **38.65** | 86.38 | 60.76 | 79.08 | 83.24 | 85.02 | 61.64 | 84.21 | 63.17 | 70.19 | 86.00 | 82.68 | 65.58 |
| MaxLogit | 76.55 | 82.30 | 83.67 | 76.50 | 92.86 | 42.50 | 79.08 | 76.50 | 78.05 | 78.50 | 75.52 | 82.30 | 80.96 | 73.09 |
| ours (DML) | 79.57 | 82.63 | 83.85 | 76.21 | 87.57 | 60.28 | 82.88 | 71.31 | 82.25 | 73.38 | 77.91 | 80.13 | 82.34 | 73.98 |
| LogitNorm* | 78.65 | 70.67 | 92.48 | 45.98 | 97.56 | 13.93 | 84.77 | 68.68 | 83.79 | 71.47 | 77.14 | 80.20 | 85.73 | 58.49 |
| ours (MCF) | **91.74** | 40.15 | 95.60 | 26.93 | 92.30 | 36.90 | **95.78** | **22.74** | **94.58** | **27.39** | 75.40 | 81.59 | 90.90 | 39.28 |
| ours (MNC) | 85.52 | 58.41 | 94.92 | 32.21 | 97.53 | 13.54 | 88.98 | 50.37 | 88.69 | 49.51 | **83.41** | 68.56 | 89.84 | 45.43 |
| ours (DML+) | 88.56 | 49.24 | **96.51** | **21.69** | **97.84** | **12.56** | 91.85 | 37.01 | 91.50 | 37.67 | 83.31 | **68.31** | **91.57** | **37.75** |

Table 4. OOD detection for our methods and baseline methods on CIFAR-100. AUR represents AUROC and FPR95 for FPR, and all values are percentages. The best model is emphasized in bold, while the 2nd and 3rd are underlined. * means the results are from [37]. The methods *above the line are post-hoc methods* while *under the line are methods with improved training*.

| Dataset | CIFAR-10 |  | CIFAR-100 |  |  |  |
|---|---|---|---|---|---|---|
| Methods | WRN-40-2 |  | R34 |  | DenseNet |  |
|  | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ |
| MSP | 90.47 | 51.60 | 81.62 | 74.69 | 78.84 | 75.98 |
| Energy | 90.48 | 35.85 | 83.65 | 69.83 | 83.21 | 66.35 |
| ViM | 94.98 | 23.01 | 84.96 | 58.85 | 83.37 | 59.84 |
| MaxLogit | 90.45 | 36.35 | 83.28 | 70.82 | 82.96 | 67.52 |
| ours (DML) | 92.22 | 37.02 | 84.82 | 69.21 | 83.17 | 67.55 |
| LogitNorm | 96.91 | 15.65 | 77.79 | 71.18 | 81.82 | 67.00 |
| ours (MCF) | 97.25 | 13.49 | **89.06** | **54.06** | 90.89 | 41.07 |
| ours (MNC) | 97.26 | 13.85 | 84.28 | 65.94 | 82.76 | 68.68 |
| ours (DML+) | **98.00** | **10.08** | 87.88 | 57.16 | **92.19** | **31.60** |

Table 5. Results on CIFAR-10, and different model architectures on CIFAR-100. All numbers are average and in percentage.

tection methods without extra data and training, we choose MSP [11], ODIN [21], Mahalanobis [20], Energy [23], ReAct [29], GradNorm [14], MaxLogit [10], and ViM [36]. These methods are traditional or SOTA as reported in [41]. Our methods rank top on the *near-OOD* datasets too and complete results including the *near-OOD* results are available in *supplementary materials*.

**Results on CIFAR.** We present our results of CIFAR-100 in Table 4. We emphasize the best model in bold while the second and third ones are in underlines. On all six OOD datasets, our methods perform the best on AUROC. On three datasets, including SVHN, LSUN-R and iSUN, our methods achieve the best, second and third AUROC and FPR95. On Textures, ViM [36] which is a SOTA post-hoc method, performs the best on FPR95. On LSUN-C and Places365, LogitNorm [37] takes the second or third place on AUROC. LogitNorm is a training method for OOD detection which uses LogitNorm loss. LogitNorm performs best in our baselines, whose AUROC is 85.73%, 3.04% higher than SOTA post-hoc method ViM, which also illustrates the necessity of focusing on model training.

We also observe that MCF and MNC are complementary: on Textures, LSUN-R and iSUN, MCF outperforms MNC by around 7% on AUROC while MNC performs better on the other three datasets. As a result, our DML is more robust on different OOD datasets, and the average AUROC is 5.84% higher than the prior SOTA method LogitNorm. In addition, the left part of Table 5 shows the results of CIFAR-10 and our methods outperform ViM and LogitNorm by more than 1% in terms of AUROC.

**Results on ImageNet.** ImageNet is a large-scale dataset and the OOD detection performance degrades as explained in [15]. The results are shown in Tabel 6. Our methods perform best on three datasets, excluding Textures. On Textures, feature distance-based methods perform better, including ViM, KNN and Mahalanobis. We hypothesize that our non-best performance relates to the feature norm not being distinguishable for the OOD samples, because MNC has the smallest AUROC among four OOD datasets. For the other three datasets, our methods take the first, second and third places on AUROC, excluding iNaturalist where KNN (w/ CL) takes the third place. KNN (w/ CL) is the KNN OOD detection method applied on the ResNet50 trained with the SupCon loss [18]. However, KNN used on the conventional model (w/o CL) performs worse than MaxLogit. Thus, the performance boost mainly comes from the better model. More details of SupCon will be discussed next. Our method has 93.16% AUROC and 29.79% FPR95, surpassing the second place (excluding KNN with CL) by 5.51% and 10.91%, respectively.

### 5.3. Ablation Study

**Different architectures.** In the right part of Table 5, we show that our methods are effective on different model architectures. We choose DenseNet and ResNet34 which are commonly used in CIFAR-100. It is shown that our meth-

| Methods | iNaturalist | | SUN | | Places365 | | Textures | | Average | | ID ACC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | AUR ↑ | FPR ↓ | |
| MSP | 87.36 | 59.29 | 79.92 | 73.42 | 79.82 | 73.88 | 80.75 | 68.48 | 81.81 | 68.77 | 72.18 |
| Energy | 91.02 | 55.10 | 85.58 | 62.11 | 83.98 | 65.34 | 87.68 | 52.25 | 87.07 | 58.70 | 72.18 |
| GradNorm | 91.79 | 31.24 | 88.87 | 38.53 | 86.28 | 46.29 | 83.66 | 46.76 | 87.63 | 40.70 | 72.18 |
| ViM | 88.40 | 67.95 | 72.65 | 91.87 | 71.47 | 91.09 | **97.52** | 12.40 | 82.51 | 65.83 | 72.18 |
| KNN(w/o CL)* | 86.20 | 59.08 | 80.10 | 69.53 | 74.87 | 77.09 | 97.18 | **11.56** | 84.59 | 54.32 | 76.65 |
| MaxLogit | 91.05 | 54.49 | 84.96 | 65.45 | 83.69 | 67.60 | 86.71 | 57.09 | 86.60 | 61.16 | 72.18 |
| ours (DML) | 91.61 | 47.32 | 86.14 | 57.40 | 84.68 | 61.43 | 86.72 | 52.80 | 87.28 | 54.74 | 72.18 |
| KNN(w/ CL)* | 94.72 | 30.83 | 88.40 | 48.91 | 84.62 | 60.02 | 94.45 | 16.97 | 90.55 | 39.18 | 79.10 |
| ours (MCF) | 93.77 | 36.29 | 89.50 | 51.18 | 86.78 | 57.38 | 94.35 | 28.46 | 91.10 | 43.33 | 71.98 |
| ours (MNC) | **97.88** | **10.94** | **94.49** | **25.34** | **91.82** | **34.99** | 85.21 | 50.57 | 92.35 | 30.46 | 72.54 |
| ours (DML+) | 97.50 | 13.57 | 94.01 | 30.21 | 91.42 | 39.06 | 89.70 | 36.31 | **93.16** | **29.79** | 72.54 |

Table 6. OOD detection performance comparison with various methods on ImageNet. We train ResNet50 for 90 epochs from scratch for all models. KNN (w/o CL) means KNN method tested on ResNet50 trained with CE loss, while (w/ CL) means the ResNet50 trained with SupCon [18]. * means the results are from [30]. The methods above the line are post-hoc while under the line are with improved training.

| Methods | original | norm+uniform | norm+greedy |
|---|---|---|---|
| AUROC | 87.89 | 89.70 | 89.82 |

Table 7. The results of different coupling methods of DML+.

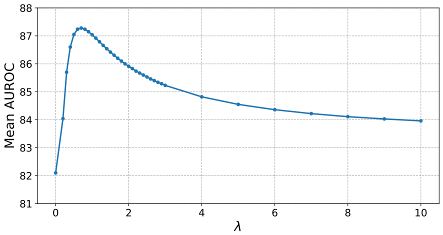
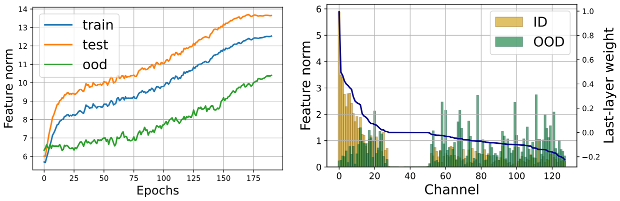


Figure 4. Effect of $\lambda$ in DML with ImageNet.

ods are robust to model architecture changes. For example, DML+ has 92.19% AUROC, which surpasses the second place by 8.82%.

**Different $\lambda$ for DML+.** We experiment with three coupling methods for DML+ on Textures as shown in Table 7. Original means we use the MCF and MNC scores without normalization. However, the cosine similarity has different norms from the feature norm. Thus, the coupling performance is affected mainly by the one with larger norms. We normalize the MCF and MNC by the ID information. We accumulate the score of ID training data and divide the score of test data with the accumulated number. The uniform in the table means $\lambda = 1$, and greedy means we set $\lambda$ based on the OOD detection performance on Gaussian noise. For instance, if MCF performs better than MNC on Gaussian noise, $\lambda$ would be larger than 1. Due to the extra training cost and marginal improvement, we choose norm+uniform for all experiments, which is fast and convenient.

**Different $\lambda$ for DML.** In Fig. 4, We ablate how the hyperparameter $\lambda$ in DML affects the OOD detection performance. When $\lambda = 0$, DML equals to MaxNorm and performs the worst. When $\lambda > 0$, the performance increases much by over 2%, which indicates that MaxCosine and MaxNorm are complementary.



(a) CIFAR-100 feature norm. (b) CIFAR-100 feature activation.

Figure 5. (a) shows how the average feature norm for ID train/test and OOD data evolve as training proceeds. (b) shows the classifier weight (blue line) and average feature norm (bars) for different channels. We choose the 48th class data as the ID data and sort the channels by the descending 48th classifier weight.

**Visualization.** We visualize the feature norm of training, test and OOD data as training proceeds in Fig. 5 (a). The feature norm of OOD data is always less than that of test data, which guarantees MaxNorm's effectiveness. As Fig. 5 (b) shows, the channel activation of ID data is larger when the corresponding classifier weight is larger. As the product of the feature and classifier, the cosine similarity of ID data (orange bars) is larger than the OOD data (green bars).

## 5.4. Discussion

**Our model can improve various scoring methods.** In the above experiments, we test the existing scoring methods with linear classifier and CE loss because the methods do not pay attention to this area. In Fig. 6, we show that our model (Focal(N) and Center(N)) not only improves MaxCosine and MaxNorm but also greatly boosts the performance of existing methods. Our MNC model (Center(N)) improves all the scoring functions by more than 9%. For GradNorm, the cosine classifier and center loss facilitate the AUROC from 52.8% to 90.8%, which is higher than

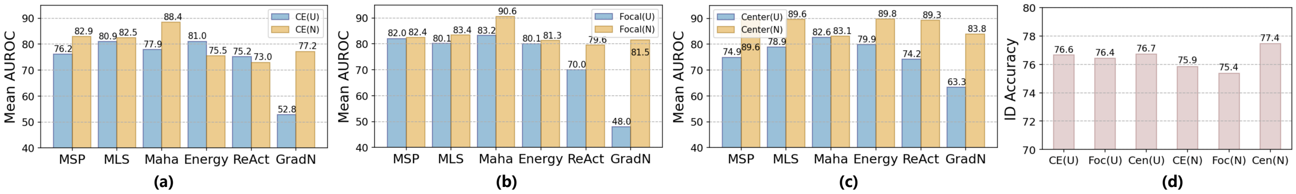


Figure 6. We report the mean AUROC of WRN-40-2 on CIFAR-100 with different models and methods. (a), (b) and (c) show the performance of existing methods on different models. (U) stands for linear classifier and (N) means the cosine classifier. (d) shows the in-distribution classification accuracy. MLS: MaxLogit, Maha: Mahalanobis, GradN: GradNorm, Foc: Focal and Cen: Center.

the SOTA method LogitNorm by 5%. We also observe that MaxCosine and cosine classifiers with CE loss (in Table 3) achieve better results than ViM (84.1% vs. 82.7%), which shows the effectiveness of the cosine classifier on OOD detection. Also, we notice that Focal(U) and Center(U) have similar OOD detection performance with CE(U).

However, when trained with a cosine classifier, different scoring functions gain a significant performance boost, especially Center(N). As a result, the simple training scheme could serve as the future baseline for OOD detection.

**The ID classification accuracy is at least maintained.** The ID classification accuracy on ImageNet and CIFAR is shown in Tabel 6 and Fig. 6 (d). The MNC model outperforms the linear classifier with CE loss on both datasets. MCF model achieves comparable ID performance on CIFAR-10 and ImageNet. As the coupling form of MNC and MCF, DML+ chooses the MNC model to output ID prediction. Overall, our DML+ has comparable or better classification performance on the ID data while substantially improving the OOD detection performance.

**How to choose the methods between DML+, MCF and MNC.** DML, as the coupling form, performs the best on average. However, DML+ needs to train two models that cost double the training and memory resource. The MNC model is a better choice when the training resource is limited. Compared to the baseline, it has similar training costs but higher OOD detection performance with different scoring functions, as shown in Fig. 6. In addition, the in-distribution classification accuracy of MNC is higher, as explained.

**Relations to better ID performance [30, 33].** The recent work [33] shows that the closed-set and open-set performance are strongly correlated. The MSP with a stronger model achieves SOTA on open-set benchmarks, including longer training time, stronger model architecture, and better augmentations. We also focus on the model training on OOD detection, but our conclusion is orthogonal. Our MCF and MNC model take similar training costs as the baseline and achieve similar ID classification accuracy. But the OOD detection performance is much higher than the baseline.

Similarly, KNN [30] explores SupCon [18] which has better ID classification accuracy (79.10% vs. 72.18%) as shown in Table 6. SupCon takes more training cost than baseline (20× GPU training cost). However, the performance increment is less than our MCF and MNC models. Our DML+ method has two times the training cost but has much higher OOD performance.

**Relations to other methods**:

1. MaxCosine and GODIN [12]. GODIN decomposes the logit as $f_i(x) = \frac{h_i(x)}{g(x)}$ where $h_i(x)$ is the cosine similarity and $g(x)$ can be viewed as temperature scaling. In inference, the data with perturbation is fed into the model and the output maximum cosine similarity is used for OOD detection. MaxCosine is different: (1) we do not use the perturbation process; (2) we do not contain $g(x)$ item.

2. MaxNorm and Objecto [6]. Objecto [6] uses the regularization method to decrease the unknown feature magnitude to increase separation in deep feature space. Our method is different: (1) Objecto [6] needs unknown data during training, ours is OOD-agnostic; (2) we use MaxNorm as OOD scoring function.

3. Our methods and LogitNorm [37]. As explained in [37], the normalization of logits and features is different. Also, we set $s = 40$ for the cosine classifier, while LogitNorm needs to tune the parameter on OOD dataset several times. Our AUROC is higher than LogitNorm (91.57% vs. 85.73%) and we can further facilitate the existing scoring functions (MSP AUROC 89.60% vs. 81.41%).

## 6. Conclusion

This paper presents the limitations of the simple logit-based scoring function MaxLogit. To overcome the limitations, we propose Decoupling MaxLogit which decouples MaxLogit for flexibility to balance MaxCosine and MaxNorm. Unlike prior works, we provide important insights that fewer hard samples and compact feature space are two key components to make logit-based methods effective. We extend DML to DML+ which changes the standard training to achieve the ideal feature space. The method can be easily implemented with existing networks and does not require sophisticated changes to the training scheme. Extensive experiments show that DML+ can greatly improve OOD detection and maintain in-distribution classification accuracy. We hope our work will inspire future logit-based research, and more training methods will be explored.

**Acknowledgement.** This research was supported by HUST Independent Inno. Res. Fund (2021XXJS096), and Alibaba Inno. Res. program (CRAQ7WHZ11220001-20978282).

## References

[1] Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably adversarially robust detection of out-of-distribution data. *Advances in Neural Information Processing Systems*, 33:16085–16095, 2020. 2

[2] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–445. Springer, 2021. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitionn*, pages 248–255. Ieee, 2009. 5

[6] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 8

[7] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19217–19227, 2022. 2

[8] 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. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[10] 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. 1, 2, 6

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 1, 2, 6

[12] 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*, pages 10951–10960, 2020. 2, 5, 8

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 5

[14] 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. 1, 2, 6

[15] 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*, pages 8710–8719, 2021. 6

[16] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2020. 2

[17] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, June 2021. 1

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2, 6, 7, 8

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[20] 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. 1, 6

[21] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 2, 6

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[23] 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. 1, 2, 6

[24] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[25] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015. 1

[26] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021. 2

[27] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 2, 3

[28] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 4

[29] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 2, 6

[30] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022. 1, 5, 7, 8

[31] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[32] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 5

[33] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021. 8

[34] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 2

[35] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 2

[36] 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*, pages 4921–4930, 2022. 1, 2, 6

[37] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. 2022. 1, 2, 5, 6, 8

[38] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 4

[39] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5

[40] 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. 5

[41] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022. 2, 5, 6

[42] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[43] 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. 5

[44] 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*, pages 9518–9526, 2019. 2

[45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 5

[46] 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. 5

[47] Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Learning towards the largest margins. In *International Conference on Learning Representations*, 2022. 2

[48] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021. 3