Decoupling MaxLogit for Out-of-Distribution Detection
- *Supplementary Material*

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
xex@hust.edu.cn

A. The definition of WFC and CFC

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

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

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 $\boldsymbol{W}_L = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K]^\top$. There are two conclusions of feature collapse [15]: (1) The within-class variation of the features becomes negligible as the features collapse to their class means $\bar{\boldsymbol{h}}_k = \frac{1}{n}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i}$. (2) The vectors of the class-means converge to having equal length, forming equal-size angles between any given pair, and being the maximally pairwise-distanced configuration constrained to the previous two properties: for $i \neq j$, $||\bar{\boldsymbol{h}}_i|| = ||\bar{\boldsymbol{h}}_j||$ and $< \bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_j >= -\frac{1}{K-1}||\bar{\boldsymbol{h}}_i||^2$ where $\bar{\boldsymbol{h}}_i$ is the mean of class i features.

We define two metrics to measure feature collapse as 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 $\Sigma_B = \frac{1}{K}\sum_{k=1}^{K}(\overline{\boldsymbol{h}}_k - \overline{\boldsymbol{h}})(\overline{\boldsymbol{h}}_k - \overline{\boldsymbol{h}})^\top$. 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.

B. Proof for Proposition 1

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*

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

When the equality holds, WFC and CFC reach the lower bound: WFC, CFC $\geq$0.

Proof. The cross-entropy can be lower bounded by

$$\begin{aligned}
\mathcal{L}_{CE} &= \sum_{k=1}^{K}\sum_{i=1}^{n} -\log\frac{\exp((\boldsymbol{z}_{k,i})_k)}{\exp((\boldsymbol{z}_{k,i})_k) + \sum\limits_{j\neq k}\exp(\boldsymbol{z}_{k,i})_j} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} \log\Big(1 + \frac{\sum\limits_{j\neq k}\exp(\boldsymbol{z}_{k,i})_j}{\exp(\boldsymbol{z}_{k,i})_k}\Big) \\
&\geq \sum_{k=1}^{K}\sum_{i=1}^{n} \log\Big(1 + (K-1)\frac{\exp\big(\sum\limits_{j\neq k}\frac{1}{K-1}(\boldsymbol{z}_{k,i})_j\big)}{\exp(\boldsymbol{z}_{k,i})_k}\Big) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} \log\Big(1 + (K-1)\exp\Big(\sum_{j\neq k}\frac{(\boldsymbol{z}_{k,i})_j}{K-1} - (\boldsymbol{z}_{k,i})_k\Big)\Big),
\end{aligned} \tag{4}$$

where the inequality follows from Jensen's inequality that

$$\sum_{j\neq k}\exp(\boldsymbol{z}_{k,i})_j = (K-1)\sum_{j\neq k}\frac{1}{K-1}\exp(\boldsymbol{z}_{k,i})_j \geq (K-1)\exp\Big(\sum_{j\neq k}\frac{(\boldsymbol{z}_{k,i})_j}{K-1}\Big),$$

which achieves the equality only when $(\boldsymbol{z}_{k,i})_m = (\boldsymbol{z}_{k,i})_n$ for all $m, n \neq k$.

Let $\bar{\boldsymbol{w}} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{w}_k$, then we have

$$\begin{aligned}
\mathcal{L}_{CE} &\geq \sum_{k=1}^{K}\sum_{i=1}^{n} \log\Big(1 + (K-1)\exp\Big(\sum_{j\neq k}\frac{\boldsymbol{w}_j^\top\boldsymbol{h}_{k,i}}{K-1} - \boldsymbol{w}_k^\top\boldsymbol{h}_{k,i}\Big)\Big) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} \log\Big(1 + (K-1)\exp\Big(\frac{K(\bar{\boldsymbol{w}} - \boldsymbol{w}_k)^\top\boldsymbol{h}_{k,i}}{K-1}\Big)\Big) \\
&\geq \sum_{k=1}^{K}\sum_{i=1}^{n} \log\Big(1 + (K-1)\exp\Big(-\frac{K||\bar{\boldsymbol{w}} - \boldsymbol{w}_k||_2||\boldsymbol{h}||_2}{K-1}\Big)\Big) \\
&= n\sum_{k=1}^{K} \log\Big(1 + (K-1)\exp\Big(-\frac{K||\bar{\boldsymbol{w}} - \boldsymbol{w}_k||_2||\boldsymbol{h}||_2}{K-1}\Big)\Big),
\end{aligned} \tag{5}$$

where we use the facts that $(\bar{\boldsymbol{w}} - \boldsymbol{w}_k)^\top\boldsymbol{h}_{k,i} \geq -||\bar{\boldsymbol{w}} - \boldsymbol{w}||_2||\boldsymbol{h}||_2$, which becomes an equality if $\boldsymbol{h}_{k,i} = -||\boldsymbol{h}||_2\frac{\bar{\boldsymbol{w}} - \boldsymbol{w}_k}{||\bar{\boldsymbol{w}} - \boldsymbol{w}_k||_2}$.

Applying Jensen's inequality, we have

$$\begin{aligned}
\mathcal{L}_{CE} &\geq n\sum_{k=1}^{K} \log\Big(1 + (K-1)\exp\Big(-\frac{K||\bar{\boldsymbol{w}} - \boldsymbol{w}_k||_2||\boldsymbol{h}||_2}{K-1}\Big)\Big) \\
&\geq n\log\Big(1 + (K-1)\exp\Big(-\frac{K}{K-1}\sum_{k=1}^{K}||\bar{\boldsymbol{w}} - \boldsymbol{w}_k||_2||\boldsymbol{h}||_2\Big)\Big) \\
&\geq n\log\Big(1 + (K-1)\exp\Big(-\frac{K\sqrt{K}}{K-1}||W||_F||\boldsymbol{h}||_2\Big)\Big).
\end{aligned} \tag{6}$$

The second inequality is obtained by following

$$\begin{aligned}\sum_{k=1}^{K} -\frac{K||\bar{\boldsymbol{w}}-\boldsymbol{w}_k||_2}{K-1} &\geq -\frac{K}{K-1}\sqrt{K\sum_{k=1}^{K}||\bar{\boldsymbol{w}}-\boldsymbol{w}_k||_2^2} \\ &= -\frac{K}{K-1}\sqrt{K\sum_{k=1}^{K}\left(||\bar{\boldsymbol{w}}||_2^2-2||\bar{\boldsymbol{w}}||_2^2+||\boldsymbol{w}_k||_2^2\right)} \\ &= -\frac{K}{K-1}\sqrt{K\sum_{k=1}^{K}\left(||\boldsymbol{w}_k||_2^2-||\bar{\boldsymbol{w}}||_2^2\right)} \\ &\geq -\frac{K\sqrt{K}}{K-1}||W||_F.\end{aligned} \tag{7}$$

For the first inequality, we use Jenson's inequality for the convex function $\sqrt{x}$ with equality if and only if $\forall\ k$, $||\bar{\boldsymbol{w}}-\boldsymbol{w}_k||_2$ is equal. Then second inequality achieves the equality only when $\bar{\boldsymbol{w}}=0$.

According to the above derivation, the minimal of the cross-entropy loss achieves if and only if $\forall \boldsymbol{h}_{k,i}, \boldsymbol{w}_1^\top \boldsymbol{h}_{k,i} = \cdots = \boldsymbol{w}_{k-1}^\top \boldsymbol{h}_{k,i} = \boldsymbol{w}_{k+1}^\top \boldsymbol{h}_{k,i} = \cdots = \boldsymbol{w}_K^\top \boldsymbol{h}_{k,i}, \boldsymbol{h}_{k,i} = -||\boldsymbol{h}||_2 \frac{\bar{\boldsymbol{w}}-\boldsymbol{w}_k}{||\bar{\boldsymbol{w}}-\boldsymbol{w}_k||_2}, \bar{\boldsymbol{w}}=0, and\ ||\bar{\boldsymbol{w}}-\boldsymbol{w}_1||_2 = \cdots = ||\bar{\boldsymbol{w}}-\boldsymbol{w}_K||_2$.

Recall that WFC and CFC can be formulated as

$$\text{WFC} := \frac{\text{trace}(\Sigma_W \Sigma_B^\dagger)}{K}, \quad \text{CFC} := \sum_{k=1}^{K}\left|\left|\frac{\overline{\boldsymbol{h}}_k}{||\boldsymbol{h}||_F} - \frac{\boldsymbol{w}_k}{||\boldsymbol{W}||_F}\right|\right|. \tag{8}$$

When $\forall \boldsymbol{h}_{k,i}, \boldsymbol{w}_1^\top \boldsymbol{h}_{k,i} = \cdots = \boldsymbol{w}_{k-1}^\top \boldsymbol{h}_{k,i} = \boldsymbol{w}_{k+1}^\top \boldsymbol{h}_{k,i} = \cdots = \boldsymbol{w}_K^\top \boldsymbol{h}_{k,i}, \boldsymbol{h}_{k,i} = -||\boldsymbol{h}||_2 \frac{\bar{\boldsymbol{w}}-\boldsymbol{w}_k}{||\bar{\boldsymbol{w}}-\boldsymbol{w}_k||_2}, \bar{\boldsymbol{w}}=0$, we have $\Sigma_W = 0$ and WFC$= 0$. For CFC, $\overline{\boldsymbol{h}}_k = \boldsymbol{w}_k$ and CFC$= 0$. Note that WFC, CFC≥ 0, so WFC and CFC reach the lower bound: WFC, CFC $\geq$0 when the loss reach the minimum.

□

C. Experimental Setting

C.1. Descriptions of OOD Datasets

For CIFAR-10 and CIFAR-100, we use six common benchmarks as OOD datasets following the work [8] including Textures [1], SVHN [5], LSUN-Crop and LSUN-Resize [13], iSUN [11] and Places365 [14]. These six datasets are called far-OOD datasets for CIFAR [12]. The near-OOD datasets for CIFAR-10 are CIFAR-100 and Tiny-ImageNet. And the near-OOD datasets for CIFAR-100 are CIFAR-10 and Tiny-ImageNet. Textures is a dataset consisting of images of described Textures. SVHN dataset contains 32×32 color images of digits zero to nine. LSUN-Crop and LSUN-Resize are the cropped and resized version of LSUN dataset respectively. LSUN is a scene recognition dataset and iSUN is a large-scale eye-tracking dataset. Places365 consists of images for scene recognition containing 365 scene categories.

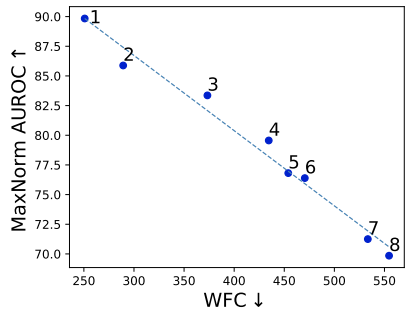
For ImageNet, we use four common benchmarks as OOD datasets following the work [6] including iNaturalist [7], SUN [10], Places365, and Textures. iNaturalist is a real-world dataset containing 8,142 fine-grained species. SUN is a scene recognition dataset containing 397 categories and 108,754 images.

C.2. Training Details

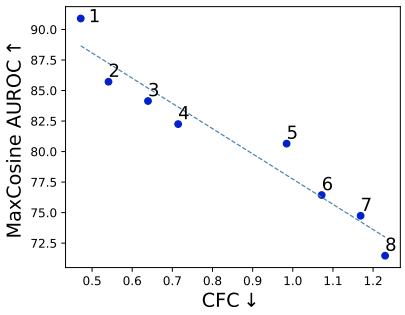
λ in DML and DML+. In DML, we tune λ to balance the effects of MaxNorm and MaxCosine based on the OOD detection performance on Gaussian noise. In DML+, the performance of MCF and MNC is similar. As we explained in the main paper, we set $\lambda = 1$ for all experiments for efficiency.

Scaling parameter. The scale parameter of the cosine classifier is set to 40 for all experiments on CIFAR-10, CIFAR-100, and ImageNet.

Hyper-parameter of loss functions. We set $\gamma = 2$ for Focal loss [3] for all experiments. The weighted Center loss [9] can be formulated as $\mathcal{L} = \omega\mathcal{L}_{Center} + \mathcal{L}_{CE}$. For Center loss on CIFAR, we change the weight from 0 to 0.001 at the 60th epoch and set the weight as 0.005 at the 80th epoch. On ImageNet, we change the weight from 0 to 0.001 at the 50th epoch and set the weight as 0.005 at the 70th epoch.



(a) The correlation between MaxNorm and WFC.



(b) The correlation between MaxCosine and CFC.

Figure 1. Gains in OOD detection performance of WRN-40-2 (cosine classifier) as CFC or WFC score increases on CIFAR-100.

Details of Fig. 2 in the main paper. We present eight experiments for WFC and CFC in Fig. 1 which is the same as Fig. 2 in the main paper. Specifically, we train the cosine classifier model with different loss functions with different hyper-parameter. In Fig. 1, we train the model with CE loss, Focal loss ($\gamma = 1, 2, 3$) and Center loss (weight $\omega = 0.0005, 0.003, 0.005, 0.01, 0.0015$).

Implementation details. For CIFAR-10 and CIFAR-100, all the models including the linear classifier and cosine classifier, 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 5×10^{-4}, and the batch size is 128. For ImageNet, we train the ResNet50 from scratch for 90 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 5×10^{-4}, and the batch size is 256.

D. Model calibration results

For calibration performance, as Table 1 shows, DML+ and MCF have similar ECE scores with CE except for MNC. When combined with temperature scaling, Focal loss yields better-calibrated models [4]. The calibration results of MCF model perform similarly to [4]. We report two ECE scores of DML+, one higher than CE and one lower than CE. Interestingly, the only difference between them is the combination methods of the logits from MCF and MNC. Specifically, we add the logits of MCF and MNC to calculate the softmax score (DML+(1)) or we first normalize the logits of MCF and MNC and then add them to calculate the softmax score (DML+(2)).

methods	CE	LogitNorm	MCF	MNC	DML+(1)	DML+(2)
ECE	0.77	0.67	0.63	2.09	0.57	1.41

Table 1. ECE in percentage after temperature scaling on CIFAR-10.

E. The effect of ensembling

DML+ uses two models to produce the final OOD score. In this section, we investigate how ensembling influences OOD detection. we add another baseline named MaxAvgLogit (the score function is the max of the averaged logits from the two models) as one reviewer suggested. As Table 2 shows, we train two models with different seeds. MaxLogit(1) and MaxLogit(2) are MaxLogit with the two models and MaxAvgLogit is the max of the averaged logits from the two models. The results show that ensembling boosts AUROC by 1.29%. However, DML+ is still 5.46% higher than that which indicates the effectiveness of DML+.

methods	MaxLogit(1)	MaxLogit(2)	MaxAvgLogit	DML	DML+
AUROC	89.17	89.34	91.63	91.32	97.09

Table 2. Mean AUROC results on two near-OOD and six far-OOD datasets. The WRN-40-2 models are trained on CIFAR-10.

F. Limitation

Although DML has competitive performance, DML does not always outperform other complex SOTA methods. In addition, more training methods could be explored to improve the performance of OOD detection. Also, OOD detection with the CLIP model which uses naturally a cosine-based classifier could be explored in future work.

G. More Experimental Results

This section presents the complete baselines' results and our methods on different datasets. Table 3 shows the results of our methods and baseline methods on CIFAR-100 with WRN-40-2. We report the full results on near-OOD and far-OOD datasets. Table 4 shows the results of our methods and baseline methods on ImageNet with ResNet50. iNaturalist dataset is the near-OOD dataset and others are far-OOD datasets. Table 5 shows the results of our methods and baseline methods on CIFAR-10 with WRN-40-2. We also report the full results on near-OOD and far-OOD datasets. Table 7 shows the OOD detection results of our methods and baseline methods on CIFAR-100 with ResNet34.

In the main paper, we report the OOD detection performance on far-OOD datasets due to space limitations. In Table 3 and 5, we report the performance on near-OOD datasets additionally. Near-OOD datasets only have semantic shift compared with ID datasets, while far-OOD further contains obvious covariate (domain) shift [12]. As the tables show, our methods outperform other methods on both near- and far-OOD datasets.

Table 6 shows the OOD detection results of different models. We train the model with a linear classifier or a cosine classifier with CE loss, Focal loss, or Center loss. As Table 6 shows, 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 the SOTA method LogitNorm by 5%. Also, we notice that Focal(U) and Center(U) have similar OOD detection performance with CE(U). However, when trained with a cosine classifier, the 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. In addition, we can observe that DML outperforms MaxLogit on different models, which illustrates the effectiveness of decoupling.

Methods	CIFAR-10		TinyImageNet		Textures		SVHN		LSUN-C		LSUN-R		iSUN		Places365	
	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓
MSP	75.90	82.11	72.73	80.73	74.24	84.43	77.17	80.66	84.26	66.30	73.37	81.98	73.04	82.49	75.20	82.69
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
Energy	77.29	79.92	79.05	74.91	75.80	82.23	83.92	75.72	93.53	37.25	79.05	76.02	78.48	78.38	75.71	82.58
ReAct	70.06	81.05	72.52	76.40	68.06	83.33	78.80	77.87	91.80	39.19	72.50	77.62	71.78	80.02	67.99	83.85
Mahalanobis	64.60	93.79	82.92	67.22	91.13	38.20	81.33	71.44	58.44	94.78	84.45	65.78	83.95	64.74	68.54	89.89
GradNorm	52.91	94.90	43.66	97.87	55.76	86.51	54.65	97.71	90.72	43.43	34.18	87.34	32.47	99.41	49.29	96.67
ViM	68.76	88.16	85.92	59.10	91.25	**38.65**	86.38	60.76	79.08	83.24	85.02	61.64	84.21	63.17	70.19	86.00
MaxLogit	76.88	80.43	76.00	77.80	76.55	82.30	83.67	76.50	92.86	42.50	79.08	76.50	78.05	78.50	75.52	82.30
ours (DML)	**77.76**	79.70	81.03	73.97	79.57	82.63	83.85	76.21	87.57	60.28	82.88	71.31	82.25	73.38	77.91	80.13
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
ours (MCF)	74.56	82.85	**94.79**	**26.97**	**91.74**	40.15	95.60	26.93	92.30	36.90	**95.78**	**22.74**	**94.58**	**27.39**	75.40	81.59
ours (MNC)	73.20	84.33	84.68	55.78	85.52	58.41	94.92	32.21	97.53	13.54	88.98	50.37	88.69	49.51	**83.41**	68.56
ours (DML+)	76.69	**79.35**	88.30	44.14	88.56	49.24	**96.51**	**21.69**	**97.84**	**12.56**	91.85	37.01	91.50	37.67	83.31	**68.31**

Table 3. OOD detection for our methods and baseline methods on CIFAR-100 with WRN-40-2. 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 [8]. The methods above the line are post-hoc methods while under the line are methods with improved training.

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
ReAct	53.92	91.17	45.21	93.39	42.18	93.66	53.69	84.06	48.75	90.57	72.18
GODIN*	85.40	61.91	85.60	60.83	83.81	63.70	73.27	77.85	82.02	66.07	70.43
Mahalanobis	53.46	97.48	37.40	99.18	38.30	99.17	88.47	44.33	54.41	85.07	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 4. 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 [2]. * means the results are from [6]. The methods above the line are post-hoc while under the line are with improved training.

Methods	CIFAR-100		TinyImageNet		Textures		SVHN		LSUN-C		LSUN-R		iSUN		Places365	
	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓
MSP	87.32	59.33	87.88	57.54	87.14	59.29	90.85	61.61	95.34	32.43	91.58	48.84	89.82	52.67	88.10	54.75
ODIN*	-	-	-	-	76.35	59.86	82.98	53.92	97.14	13.31	93.52	27.21	92.03	33.31	81.25	54.32
Energy	84.82	49.88	86.64	46.20	82.31	53.47	91.02	48.45	97.92	9.43	92.79	29.58	90.60	35.73	88.22	38.46
ReAct	78.46	57.94	80.84	53.40	75.50	61.54	86.51	59.57	96.98	12.96	89.38	36.08	86.23	42.61	83.45	45.44
GradNorm	43.16	91.09	53.90	84.38	49.18	82.41	52.00	88.23	93.81	21.78	63.37	74.93	58.87	80.84	54.21	82.80
ViM	80.13	70.34	88.34	53.39	94.59	23.70	97.77	10.65	95.03	28.78	98.12	9.14	97.84	10.09	86.52	55.73
MaxLogit	84.70	50.45	86.42	46.48	82.48	53.36	91.02	48.75	97.80	10.09	92.70	30.31	90.53	36.55	88.18	39.06
ours(DML)	87.91	49.16	89.33	44.93	87.54	51.29	92.00	48.28	96.49	17.37	93.97	30.39	92.42	36.09	90.88	38.74
LogitNorm*	50.13	94.84	97.25	15.33	94.28	28.64	98.47	8.03	99.42	2.37	97.87	10.93	97.73	12.28	93.66	31.64
ours (MCF)	81.36	63.86	94.50	27.10	97.04	**14.65**	99.27	**3.36**	99.07	4.91	98.32	8.77	98.09	10.11	91.74	39.76
ours (MNC)	91.14	44.82	96.31	21.49	95.77	21.80	98.60	7.66	99.56	1.95	97.90	11.29	97.32	14.25	94.40	26.16
ours (DML+)	**91.36**	**42.55**	**97.34**	**14.70**	**97.05**	15.31	**99.38**	3.37	**99.72**	**1.11**	**98.58**	**7.57**	**98.40**	**8.75**	**94.87**	**24.34**

Table 5. OOD detection performance comparison with various methods on CIFAR-10 with WRN-40-2. 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 [8]. The methods above the line are post-hoc while under the line are with improved training.

Methods	CE(U)		Focal(U)		Center(U)		CE(N)		Focal(N)		Center(N)	
	AUROC	ACC	AUROC	ACC	AUROC	ACC	AUROC	ACC	AUROC	ACC	AUROC	ACC
MSP	76.21	76.56	82.06	75.85	74.92	76.42	82.91	75.36	82.42	76.72	89.60	77.45
Energy	81.08	76.56	80.13	75.85	79.91	76.42	75.55	75.36	81.30	76.72	89.83	77.45
ReAct	75.24	76.56	70.00	75.85	74.21	76.42	72.97	75.36	79.58	76.72	89.33	77.45
GradNorm	52.84	76.56	48.02	75.85	63.34	76.42	77.24	75.36	81.55	76.72	83.88	77.45
Mahalanobis	77.93	76.56	**83.25**	75.85	**82.65**	76.42	**88.44**	75.36	**91.21**	76.72	83.15	77.45
MaxLogit	80.96	76.56	80.18	75.85	78.98	76.42	82.53	75.36	83.41	76.72	89.62	77.45
DML	82.34	76.56	83.14	75.85	80.34	76.42	85.34	75.36	89.38	76.72	**89.86**	77.45
MaxCosine	**81.78**	76.56	81.52	75.85	78.96	76.42	84.14	75.36	90.90	76.72	77.40	77.45
MaxNorm	56.93	76.56	43.85	75.85	64.88	76.42	76.39	75.36	69.85	76.72	89.85	77.45

Table 6. In-distribution classification accuracy and average OOD detection performance (Textures, SVHN, LSUN-C, LSUN-R, iSUN and Places365) of post-hoc scoring functions on CIFAR-100 with WRN-40-2. We train all models with identical training settings as in the main paper. All values are in percentage. N means cosine classifier and U means linear classifier. Recall that our DML AUROC is 91.57%. The methods above the line are previous post-hoc scoring methods and the methods under the line are our proposed post-hoc scoring functions.

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	79.34	79.05	81.60	76.54	80.33	77.42	86.07	65.86	84.62	62.55	77.68	79.49	81.62	74.69
Energy	81.27	76.77	84.01	73.34	80.08	78.40	90.06	53.62	88.59	57.34	78.05	79.18	83.65	69.83
ReAct	79.90	76.94	82.99	73.69	78.53	78.73	89.55	53.91	87.92	57.64	76.27	79.56	82.50	70.14
GradNorm	67.78	78.57	69.46	76.38	56.79	87.77	75.11	62.18	73.90	64.81	**86.72**	81.05	68.28	75.35
ViM	86.50	56.30	85.28	61.19	80.23	69.99	**92.31**	**41.57**	**90.96**	**44.38**	73.98	79.70	84.96	58.85
MaxLogit	81.03	76.43	83.63	73.52	80.21	76.88	89.16	58.11	87.70	61.26	78.06	78.25	83.28	70.82
ours (DML)	82.61	74.99	85.11	70.79	83.04	73.65	89.89	57.24	88.94	60.36	79.31	78.18	84.82	69.21
LogitNorm*	74.99	79.57	90.92	**49.86**	**96.05**	**21.88**	63.89	97.11	63.22	97.42	77.64	82.08	77.79	71.18
ours (MCF)	**90.64**	**50.38**	**91.62**	53.18	93.28	36.74	92.19	49.08	90.81	54.33	75.84	80.63	**89.06**	**54.06**
ours (MNC)	81.93	74.88	85.85	70.86	93.41	37.94	82.43	67.50	82.89	67.36	79.16	77.07	84.28	65.94
ours (DML+)	86.92	60.43	89.65	58.42	95.04	29.75	87.16	58.83	87.16	58.99	79.57	**76.57**	87.88	57.16

Table 7. OOD detection performance comparison with various methods on CIFAR-100 with ResNet34. 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 [8]. The methods above the line are post-hoc while under the line are with improved training.

References

[1] 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. 3

[2] 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. 6

[3] 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. 3

[4] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020. 4

[5] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 3

[6] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022. 3, 6

[7] 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. 3

[8] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. 2022. 3, 5, 6, 7

[9] 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. 3

[10] 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. 3

[11] 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. 3

[12] 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. 3, 5

[13] 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. 3

[14] 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. 3

[15] 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. 1