OpenMix: Exploring Outlier Samples for Misclassification Detection Supplementary Materials

Fei Zhu[1,2], Zhen Cheng[1,2], Xu-Yao Zhang[1,2], Cheng-Lin Liu[1,2]
[1]MAIS, CASIA, Beijing 100190, China
[2]School of Artificial Intelligence, UCAS, Beijing, 100049, China
{zhufei2018, chengzhen2019}@ia.ac.cn, {xyz, liucl}@nlpr.ia.ac.cn

A. More Details on Experimental Setups

A.1. Experiments on CIFAR-10 and CIFAR-100

Auxiliary datasets. We use `300K Random Images` [6] as the auxiliary outlier dataset for experiments with CIFAR-10 and CIFAR-100. Specifically, `300K Random Images` is a cleaned and debiased dataset with 300K natural images. In this dataset, images that belong to CIFAR classes are removed so that in-distribution (ID) training set and outlier dataset are disjoint. In Sec. 5.2 and Fig. 8, we conduct experiments on CIFAR-10 to study the effectiveness of other outlier datasets, *i.e.*, `Gaussian`, `Rademacher`, `Blob`, `CIFAR-100`. Following [13], `Gaussian` noises are sampled from an isotropic Gaussian distribution. `Rademacher` noises are sampled from a symmetric Rademacher distribution. `Blob` noises consist of algorithmically generated amorphous shapes with definite edges.

Hyperparameters. For OE [6], we set $\lambda = 0.5$ in Eq. 2 as recommended in the original paper [6]. For Mixup [16], the coefficient of linear interpolation λ is sampled as $\lambda \sim \text{Beta}(\alpha, \alpha)$, and we set $\alpha = 0.3$ as recommended in the original paper [16]. For RegMixup [11], we set $\alpha = 10$ as recommended in the original paper [11]. For our OpenMix, the λ in Eq. 4 is sampled as $\lambda \sim \text{Beta}(\alpha, \alpha)$, and we set $\alpha = 10$ in our experiments. The γ in Eq. 5 is set as 1.

A.2. Experiments on ImageNet

For experiments on ImageNet, the backbone is ResNet-50 [2] and we perform automatic mixed precision to accelerate the training by using the open-sourced code at `https://github.com/NVIDIA/apex/tree/master/examples/imagenet`. For each experiment, we train 90 epochs. Three settings which consist of random 100, 200, and 500 classes from ImageNet are conducted after shuffling the class order with the fixed random seed 1993. For each experiment, we use another set of disjoint classes from ImageNet as outliers, and the outlier dataset has the same number of classes as that of training set. The λ in Eq. 4 is sampled as $\lambda \sim \text{Beta}(\alpha, \alpha)$, and we set $\alpha = 10$. The γ in Eq. 5 is set to be 0.5.

B. Additional Experimental Results

B.1. More results for MisD under distribution shift

Table 1 presents more results of MisD under distribution shift. The models trained on clean datasets (CIFAR-10 and CIFAR-100) are evaluated on corrupted dataset CIFAR-10/100-C [4]. the corruption dataset contains copies of the original validation set with 15 types of corruptions of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. In Table 1, we can observe that OpenMix performs the best.

B.2. More results for long-tailed MisD

Besides the results in Table. 6, we also compared the MisD performance of our method with TLC [7] under the same experimental setup. The results of TLC and others are from [7]. As can be observed from Fig. 1, our method has the best performance, *i.e.*, highest AUROC and lowest FPR95.

Table 1. MisD performance under distribution shift. The averaged results for 15 kinds of corruption under five different level perturbation severity are reported.

Method	**AUROC ↑**			**AURC ↓**			**FPR95 ↓**			**ACC ↑**		
	ResNet	WRN	DenseNet	ResNet	WRN	DenseNet	ResNet	WRN	DenseNet	ResNet	WRN	DenseNet
	CIFAR-10-C											
MSP [5]	79.92	83.34	81.82	154.58	120.36	154.72	70.23	64.48	68.56	72.27	75.57	71.08
CRL [10]	82.57	85.86	83.86	143.19	100.27	135.46	68.26	62.86	66.93	71.19	76.24	71.82
OpenMix	**84.98**	**90.38**	**85.62**	**65.51**	**27.78**	**71.86**	**62.11**	**48.07**	**60.65**	**82.03**	**88.33**	**81.38**
	CIFAR-100-C											
MSP [5]	77.39	79.70	75.86	356.87	299.82	376.37	76.70	72.77	76.88	45.27	51.38	44.92
CRL [10]	79.00	80.71	78.15	340.48	273.60	346.73	74.68	71.13	75.25	45.91	53.38	46.56
OpenMix	**78.56**	**84.05**	**79.00**	**303.82**	**176.15**	**299.71**	**74.61**	**66.24**	**74.45**	**50.61**	**62.09**	**51.29**

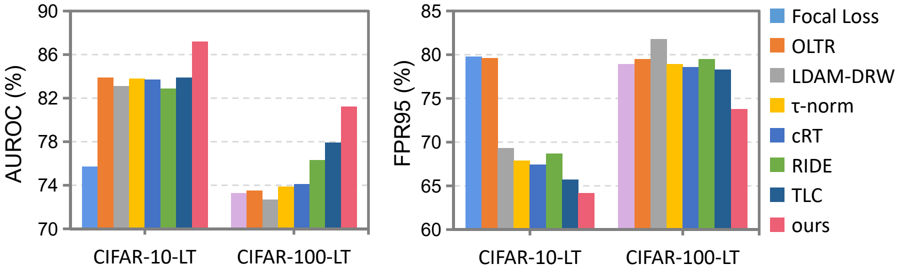

Figure 1. More comparison on long-tailed MisD.

Table 2. OOD detection performance. All values are percentages and are averaged over six OOD test datasets.

Method	**FPR95 ↓**			**AUROC ↑**			**AUPR ↑**		
	ResNet	WRN	DenseNet	ResNet	WRN	DenseNet	ResNet	WRN	DenseNet
	ID: CIFAR-10								
MSP [5]	51.69	40.83	48.60	89.85	92.32	91.55	97.42	97.93	98.11
LogitNorm [14]	29.72	12.97	19.72	94.29	97.47	96.19	98.70	99.47	99.11
ODIN [8]	35.04	26.94	30.67	91.09	93.35	93.40	97.47	97.98	98.30
Energy [9]	33.98	25.48	30.01	91.15	93.58	93.45	97.49	98.00	98.35
MaxLogit [3]	34.61	26.72	30.99	91.13	93.14	93.44	97.46	97.78	98.35
OE [6]	**5.28**	**3.49**	**5.25**	**98.04**	**98.59**	**98.20**	**99.55**	**99.71**	**99.62**
CRL [10]	51.18	40.83	47.28	91.21	93.67	92.37	98.11	98.67	98.35
FMFP [17]	39.50	26.83	35.12	93.83	96.22	94.88	98.73	99.23	98.95
OpenMix (ours)	39.72	16.86	32.75	93.22	96.92	94.85	98.46	99.34	98.84
	ID: CIFAR-100								
MSP [5]	81.68	77.53	77.03	74.21	77.96	76.79	93.34	94.36	93.94
LogitNorm [14]	63.49	57.38	61.56	82.50	86.60	82.10	95.43	96.80	95.16
ODIN [8]	74.30	76.03	69.44	76.55	79.57	80.53	93.54	94.59	94.78
Energy [9]	74.42	74.93	68.36	76.43	79.89	80.87	93.59	94.66	94.86
MaxLogit [3]	74.45	75.27	69.85	76.61	79.75	80.48	93.66	94.67	94.77
OE [6]	**59.85**	**49.02**	**53.03**	**86.33**	**90.07**	**88.51**	**96.47**	**97.67**	**97.25**
CRL [10]	81.67	79.08	75.77	72.72	76.81	76.41	92.69	94.22	93.85
FMFP [17]	80.19	70.98	72.87	72.92	81.54	77.56	92.94	95.71	94.19
OpenMix (ours)	74.66	68.87	66.63	75.95	84.88	81.23	93.56	96.55	95.30

B.3. More results for OOD detection

Table 2 presents the detailed results of OOD detection performance on CIFAR-10 and CIFAR-100. From the results, we show that our method can yield strong OOD detection performance. In addition, since Openmix is a training-time method, it can combine with any other post-processing OOD detection methods such as ODIN, Energy and MaxLogit to get higher OOD detection performance.

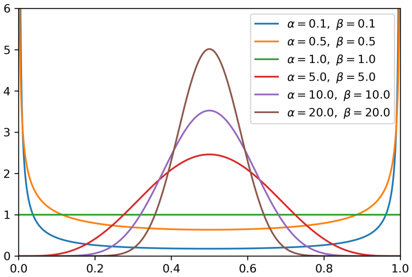

Figure 2. Beta(α, α) pdf for varying α.

α	**AURC ↓**	**AUROC ↑**	**FPR95 ↓**	**ACC ↑**
None	9.52±0.49	90.13±0.46	43.33±0.59	94.30±0.06
0.1	9.30±2.01	92.04±0.81	46.69±5.72	92.76±0.91
0.5	7.56±1.51	91.87±1.43	44.20±3.76	94.00±0.25
1	6.58±0.61	92.19±0.36	39.17±1.77	94.59±0.19
5	6.08±0.88	92.46±1.02	37.44±0.79	94.89±0.15
10	6.31±0.32	92.09±0.36	39.63±2.36	94.98±0.20
20	5.96±0.90	92.45±0.22	35.70±0.72	95.12±0.17

Table 3. Ablation study on α.

B.4. Ablation study on different distribution of interpolation coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$

We conduct experiments to compare the effectiveness of different interpolation coefficient, which is drawn from the Beta distribution (refer Fig. 2). Specifically, high values of α would encourage $\lambda \approx 0.5$. As shown in Table 3, large values of α (strong interpolations) lead to good performance. Since we aim to improve the exposure of low density regions, the interpolation should be strong to yield low confidence samples.

B.5. Using Cutmix to transform outliers in OpenMix

In our main manuscript, linear interpolation is applied to transform the outliers. An alternative way is to use non-linear strategy like CutMix [15]. From the results in Table 4, we observe that Cutmix based outlier transformation can yield comparable performance as mixup based.

Table 4. Comparison between linear (Mixup) and non-linear (Cutmix) based outlier transformation.

Network	**Method**	**CIFAR-10**				**CIFAR-100**			
		AURC ↓	**AUROC ↑**	**FPR95 ↓**	**ACC ↑**	**AURC ↓**	**AUROC ↑**	**FPR95 ↓**	**ACC ↑**
ResNet110	MSP [ICLR17] [5]	9.52±0.49	90.13±0.46	43.33±0.59	94.30±0.06	89.05±1.39	84.91±0.13	65.65±1.72	73.30±0.25
	OpenMix (w/Mixup)	**6.31±0.32**	92.09±0.36	39.63±2.36	**94.98±0.20**	**73.84±1.31**	85.83±0.22	**64.22±1.35**	**75.77±0.35**
	OpenMix (w/CutMix)	6.74±1.07	**93.45±0.44**	**36.82±3.65**	93.73±0.72	76.28±1.83	**86.49±0.17**	64.78±0.93	74.15±0.41
WRNet	MSP [ICLR17] [5]	4.76±0.62	93.14±0.38	30.15±1.98	95.91±0.07	46.84±0.90	88.50±0.44	56.64±1.33	80.76±0.18
	OpenMix (w/Mixup)	**2.32±0.15**	**94.81±0.34**	**22.08±1.86**	**97.16±0.10**	**39.61±0.54**	**89.06±0.11**	**55.00±1.29**	**82.63±0.06**
	OpenMix (w/CutMix)	3.11±0.50	94.14±0.17	28.25±2.25	96.60±0.40	43.22±1.01	89.16±0.16	55.62±1.67	80.94±0.31
DenseNet	MSP [ICLR17] [5]	5.66±0.45	93.14±0.65	38.64±4.70	94.78±0.16	66.11±1.56	86.20±0.04	62.79±0.83	76.96±0.20
	OpenMix (w/Mixup)	**4.68±0.72**	93.57±0.81	**33.57±3.70**	**95.51±0.23**	**53.83±0.93**	**87.45±0.18**	**62.22±1.15**	**78.97±0.31**
	OpenMix (w/CutMix)	5.44±0.50	**93.80±0.13**	37.28±2.15	94.48±0.39	68.80±6.96	86.46±0.57	63.99±2.41	75.92±1.24

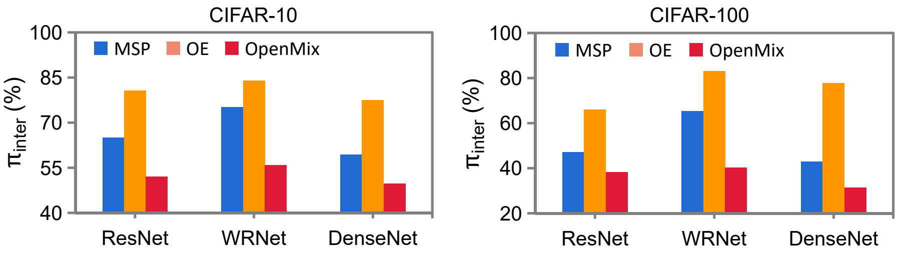

Figure 3. Comparison of the inter-distance π_{inter} of the deep feature space.

B.6. More results of feature space distance and visualization

In our main manuscript, we only plot the results of FSU due to the space limitation. Here, Fig. 3 plots the inter-class distance of the deep feature space. As can be observed, the inter-class distance with OE is observably enlarged, which indicates excessive feature compression and has negative influence for MisD. Our OpenMix leads to less compact feature distributions. Besides, Fig. 4 presents qualitative visualization to look at the effectiveness of OpenMix. Compared with MSP and OE, the feature distribution is smoother and

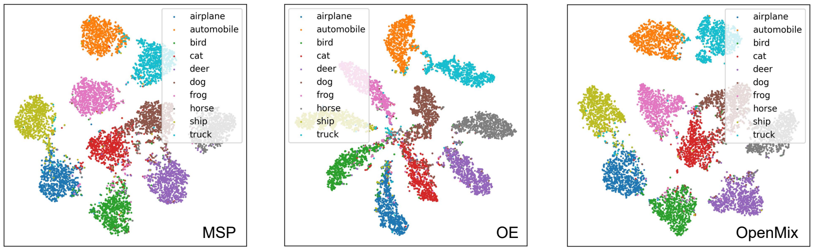

Figure 4. Qualitative visualization of the deep feature space using TSNE [12].

the decision boundary is clearer, and the misclassified samples are mostly mapped to low-density regions in feature distribution.

B.7. Ablation Study of each component in our method

Table 5 presents more results of each component in our method on WRNet and DenseNet.

Table 5. Ablation Study of each component in our method.

Network	Method	CIFAR-10				CIFAR-100			
		AURC	AUROC	FPR95	ACC	AURC	AUROC	FPR95	ACC
ResNet	MSP	9.52	90.13	43.33	94.30	89.05	84.91	65.65	73.30
	+ RC	9.55	91.15	40.03	94.02	94.31	85.53	65.78	71.44
	+ OT	12.38	87.13	61.83	93.84	99.86	82.51	72.94	72.62
	OpenMix	**6.31**	**92.09**	**39.63**	**94.98**	**73.84**	**85.83**	**64.22**	**75.77**
WRNet	MSP	4.76	93.14	30.15	95.91	46.84	88.50	56.64	80.76
	+ RC	4.28	93.95	30.05	95.62	54.32	88.08	60.17	78.69
	+ OT	5.75	90.71	49.69	95.77	54.38	86.24	64.68	80.12
	OpenMix	**2.32**	**94.81**	**22.08**	**97.16**	**39.61**	**89.06**	**55.00**	**82.63**
DenseNet	MSP	5.66	93.14	38.64	94.78	66.11	86.20	62.79	76.96
	+ RC	6.04	93.07	37.55	94.56	70.73	86.78	64.36	75.21
	+ OT	10.46	87.76	62.85	94.29	76.92	84.09	70.55	75.78
	OpenMix	**4.68**	**93.57**	**33.57**	**95.51**	**53.83**	**87.45**	**62.22**	**78.97**

C. Additional Analysis

C.1. More insights: Impact of feature space uniformity for OOD detection and MisD

We provide more insights about the connection between feature space uniformity (FSU, refer to Sec. 3.2 for detailed definition) and OOD detection, MisD performance. According to the familiarity hypothesis [1], the features are less activated for OOD samples from unknown classes than that for ID samples. Therefore, MisD is more difficult than OOD detection, and the FSU has different impact on those two tasks. In what follows, we provide more illustration based on Fig. 5. Specifically,

- For OOD detection, at the baseline state (`state 0`), the OOD distribution has some overlap with ID distribution. ① When decreasing the FSU, the distribution of known classes is compressed and the overlap between OOD and ID samples could be reduced (`state -1`). However, when further decreasing the FSU, the ID distribution could be much over-compact and the model maps most of the OOD samples to the ID region (`state -2`), leading to worse OOD detection performance. ② When increasing the FSU, more OOD samples could be mapped to low density regions (`state 1`). However, when further increasing the FSU, the ID distribution would be under-activated (`state 2`), leading to worse separation between ID and OOD distribution.

- For MisD, compared with OOD samples, the misclassified samples are ID and closer to the correct samples. Therefore, MisD performance is more sensitive to the change of FSU. As a result, ① decreasing the FSU would easily lead to more overlap between correct and misclassified ID samples (`state -1`). To improve the separation, it more helpful to ② increase the FSU (`state 1`), making the features be less activated for misclassified samples. However, when further increasing the FSU,

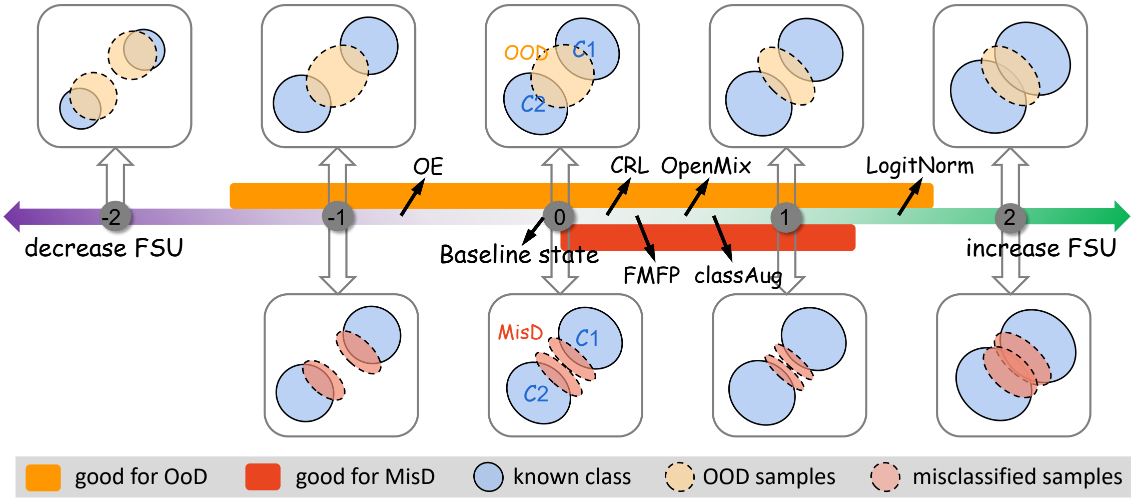

Figure 5. Illustration of how the change of FSU affects the OOD detection and MisD performance.

the distribution of correct samples would be under-activated (`state 2`), leading to worse separation between correct and wrong data.

In conclusion, both over-compact and over-dispersive feature distributions are harmful for OOD detection and MisD. To effectively detect OOD and misclassified samples, it is better to increase the FSU to a proper level. In Fig. 5, the orange region is good for OOD detection, while the red region is good for MisD. The common region between the orange and red is desirable for detecting OOD and misclassified samples in a unified manner. In addition, we compute the FSU of several representative methods (OE [6], MSP [5], CRL [10], FMFP [17], OpenMix, classAug [18], LogitNorm [14]) and mark the corresponding position in Fig. 5. The effect of them is consistent with our analysis.

C.2. Theoretical analysis: OpenMix increases the exposure of low density regions

In standard training, with cross-entropy loss and one-hot label, there are few uncertain samples are mapped to low density regions. An intuitive interpretation of the effectiveness of OpenMix is it increases the exposure of low density regions in feature space by synthesizing and learning the mixed samples. We provide a theoretical justification showing that our method can increase the sample density in the original low-density regions.

Suppose we have a known class consisting of samples drawn from probability density function $f(x)$, and an outlier distribution $f_{\text{ood}}(x)$ that is farther away from $f(x)$. By applying linear interpolation (*i.e.*, Mixup) between ID distribution $f(x)$ and outlier distribution $f_{\text{ood}}(x)$, we can get a mixed set. Denote $f_{\text{mix}}(x)$ the bimodal distribution that represents the probability density function of mixed samples. With integration of $f(x)$ and $f_{\text{mix}}(x)$, the new data probability density function is denoted as $\bar{f}(x) = \frac{1}{2}\left(f(x) + f_{\text{mix}}(x)\right)$. The following theorem shows that the probability density on the subset $S = \{x||x^\tau v| > C, x \in \mathcal{R}^d\}$ is enlarged, in which C is a sufficiently large constant and $v \in \mathcal{R}^d$ is the certain direction. For example, for the single dimensional case with variance σ^2, the density is guaranteed to be enlarged in the set $S' = \{x||x| > 1.5\sigma, \mu = 1\}$, which is exactly the low-density area for the Guassian distribution.

Theorem C.1. *Let $f(x)$ and $f_{mix}(x)$ be the probability density functions defined as follows,*

$$f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{x^\tau \Sigma^{-1} x}{2}\right),$$

and

$$f_{mix}(x) = \frac{1}{2(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^\tau \Sigma^{-1}(x-\mu)}{2}\right) + \frac{1}{2(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^\tau \Sigma^{-1}(x-\mu)}{2}\right),$$

where $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$ and Σ are the correspondingly mean vector and positive-definite *covariance matrix. Assume that* $\mu = \Sigma^{1/2}\bar{\mu}$*, where* $\|\bar{u}\| = 1$ *can be chosen arbitrary, then , it follows that*

$$f(x) < \bar{f}(x), \textit{ for any } x \in S' = \{x||x^\tau v| > 1.5, v = \Sigma^{-1/2}\bar{\mu}\}.$$

Proof. Denote $g(x) = \bar{f}(x) - f(x)$, we have

$$\begin{aligned} g(x) =& \frac{1}{2}\left(f(x) + f_{\text{mix}}(x)\right) - f(x) = \frac{1}{2}\left(f_{\text{mix}}(x) - f(x)\right) \\ =& \frac{1}{4(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{x^\tau\Sigma^{-1}x}{2}\right)\left[\exp\left(-\frac{\mu^\tau\Sigma^{-1}\mu}{2}\right)\left(\exp\left(x\Sigma^{-1}\mu\right) + \exp\left(-x\Sigma^{-1}\mu\right)\right) - 2\right] \end{aligned}$$

In what follows, we show that $g(x) > 0$ on the region $x \in S'$. Firstly, it is trivial to see that $g(0) = 2\exp\left(-\frac{\mu^\tau\Sigma^{-1}\mu}{2}\right) - 2 < 0$. To analyze the property of $g(x)$, we need to analyze the following function:

$$\begin{aligned} h(x) &= \exp\left(x^\tau\Sigma^{-1}\mu\right) + \exp\left(-x^\tau\Sigma^{-1}\mu\right) - 2\exp\left(\frac{\mu^\tau\Sigma^{-1}\mu}{2}\right) \\ &= \exp\left(x^\tau\Sigma^{-1/2}\bar{\mu}\right) + \exp\left(-x^\tau\Sigma^{-1/2}\bar{\mu}\right) - 2\exp\left(\frac{\bar{\mu}^\tau\bar{\mu}}{2}\right) \\ &= \exp\left(x^\tau\Sigma^{-1/2}\bar{\mu}\right) + \exp\left(-x^\tau\Sigma^{-1/2}\bar{\mu}\right) - 2\exp\left(\frac{1}{2}\right) \end{aligned}$$

Noticing that $\exp(x) + \exp(-x)$ is an even function and it is increasing with respect to $|x|$, thus for $h(x)$, there exists a positive constant m such that, $h(x) > 0$ when $|x^\tau\Sigma^{-1/2}\bar{\mu}| \geq m$. We can see that $\exp(1.5) + \exp(-1.5) - 2\exp\left(\frac{1}{2}\right) > 0$, which means $m \geq 1.5$. That means $g(x) > 0$ when $|x^\tau\Sigma^{-1/2}\bar{\mu}| \geq 1.5$, thus we complete the proof. □

References

[1] Thomas G Dietterich and Alex Guyer. The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition*, 132:108931, 2022. 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[3] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. 2022. 2

[4] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 1

[5] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2, 3, 5

[6] Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 1, 2, 5

[7] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, pages 6970–6979, June 2022. 1

[8] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 2

[9] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 33:21464–21475, 2020. 2

[10] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, pages 7034–7044, 2020. 2, 5

[11] Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Regmixup: Mixup as a regularizer can surprisingly improve accuracy and out distribution robustness. In *NeurIPS*, 2022. 1

[12] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, pages 2579–2605, 2008. 4

[13] Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An. Open-set label noise can improve robustness against inherent label noise. *NeurIPS*, 34:7978–7992, 2021. 1

[14] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022. 2, 5

[15] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*, pages 6023–6032, 2019. 3

[16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1

[17] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *ECCV*, pages 518–536. Springer, 2022. 2, 5

[18] Fei Zhu, Xu-Yao Zhang, Rui-Qi Wang, and Cheng-Lin Liu. Learning by seeing more classes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5