Dissecting Out-of-Distribution Detection and Open-Set Recognition: A Critical Analysis of Methods and Benchmarks
–Supplementary Material–

Hongjun Wang[1], Sagar Vaze[2], Kai Han[1*]

[1]The University of Hong Kong [2]University of Oxford

Contents

*Corresponding author.

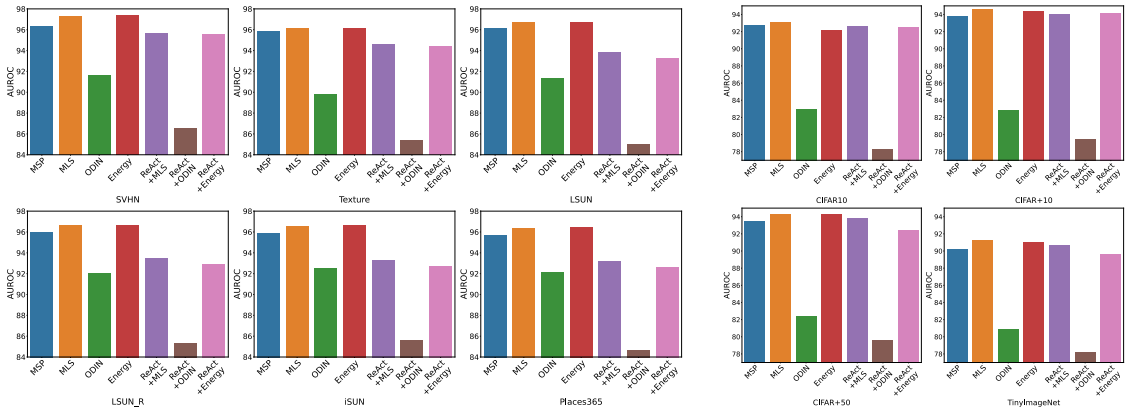
S1 Bar charts for cross-benchmarking results

In this section, we provide more experimental results for different benchmarks in different ways.

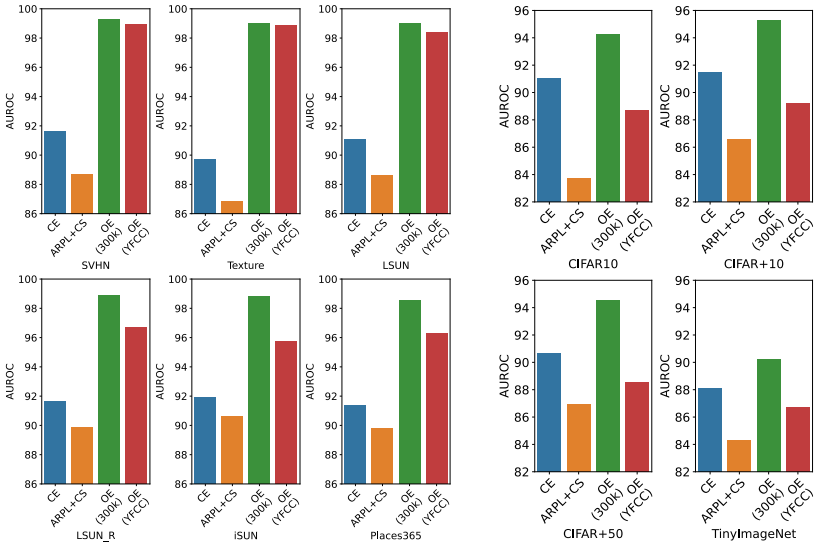
Apart from the numerical results in Table 2 in the main paper, we also present the bar charts for the comparison among different (1) scoring rules and (2) training methods in Figure S1.

In Figure S1 (a), we can observe that magnitude-aware methods (*i.e.*, MLS and Energy) are stable among different training methods on both benchmarks while ODIN is encumbered by being combined with ARPL+CS method.

In Figure S1 (b), we can see that the OE method with the help of 300K random auxiliary images outperforms others by a large margin. We also notice that the choice of auxiliary data is very important in OE performance since OE with YFCC-15M cannot achieve as excellent performance as the one with 300K random images.



(a) Performance of different scoring rules across OOD and OSR tasks



(b) Performance of different training methods across OOD and OSR tasks

Figure S1: Analysis of different scoring rules and training methods on standard benchmarks. (a) Among various scoring rules, MLS and Energy show their reliability across OOD and OSR datasets. (b) For different training methods, Outlier Exposure using different auxiliary data obtains an obvious performance boost compared with others on OOD detection and have slight gains on OSR.

S2 OOD detection using CIFAR-100 as the ID training data

To further verify our findings in Section 3.2 in the paper, we also perform experiments using CIFAR-100 as the ID training data. Results are shown in Table S1, Table S2 and Table S3. While ODIN outperforms all the methods with the CE loss (due to superior performance on SVHN and LSUN), the magnitude-aware methods consistently perform well (especially when combined with post-processing techniques). This aligns with the findings observed on CIFAR-10.

Table S1: Results of various methods on OOD detection benchmarks, using CIFAR-100 as ID. The results are averaged from five independent runs. We train ResNet-18 with the CE loss and report the in distribution accuracy as 'ID'.

Training Method	Scoring Rule	OOD detection benchmarks						
		SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=78.69
CE	MSP	83.56	78.38	78.21	79.60	79.07	73.56	78.73
	MLS	85.21	79.33	77.75	81.63	81.12	73.48	79.75
	ODIN	**97.47**	77.05	**90.49**	77.59	79.38	71.96	**82.32**
	GODIN	52.70	58.40	59.64	76.34	76.59	62.09	64.29
	SEM	65.54	31.85	83.58	35.81	37.38	51.01	50.86
	Energy	85.71	79.50	77.12	**82.19**	**81.72**	73.26	79.92
	GradNorm	65.73	62.82	61.13	64.41	64.60	61.08	63.30
	MLS+ReAct	84.51	**83.94**	84.96	80.30	79.98	**78.32**	82.00
	ODIN+ReAct	96.39	78.74	89.62	75.88	77.28	71.78	81.62
	Energy+ReAct	84.36	83.48	77.93	77.09	77.26	75.55	79.28

Table S2: Results of various methods on OOD benchmarks, using CIFAR-100 as ID. The results are averaged from five independent runs. We train ResNet-18 with the ARPL+CS loss and report the in distribution accuracy as 'ID'.

Training Method	Scoring Rule	OOD detection benchmarks						
		SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=78.86
ARPL+CS	MSP	79.95	79.00	80.28	88.16	87.71	74.69	81.63
	MLS	81.27	79.73	79.93	89.63	89.23	74.77	82.43
	ODIN	**87.72**	64.90	78.21	82.03	82.60	65.12	76.76
	GODIN	73.81	62.55	59.30	66.42	75.00	51.41	64.75
	SEM	55.94	33.81	**86.74**	32.23	34.34	57.19	50.04
	Energy	81.69	79.84	78.91	**90.57**	**90.21**	74.58	82.63
	GradNorm	50.79	52.91	49.03	54.68	56.20	50.14	52.29
	MLS+ReAct	77.71	**80.32**	85.93	88.44	87.78	**76.41**	**82.77**
	ODIN+ReAct	20.02	33.07	20.25	36.67	36.74	37.82	30.76
	Energy+ReAct	62.27	63.78	56.36	80.73	81.33	60.05	67.42

Table S3: Results of various methods on OOD detection benchmarks, using CIFAR-100 as ID. The results are averaged from five independent runs. We train ResNet-18 with the OE loss and report the in distribution accuracy as 'ID'.

Training Method	Scoring Rule	OOD detection benchmarks						
		SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=77.16
OE	MSP	93.88	87.76	73.66	73.10	74.72	<u>74.49</u>	79.60
	MLS	94.32	88.44	72.53	73.22	75.21	74.10	79.64
	ODIN	**97.20**	83.94	**85.96**	71.61	74.60	71.90	80.87
	GODIN	74.18	83.35	67.85	74.71	77.22	66.61	73.99
	SEM	68.48	47.58	80.55	48.33	46.99	49.15	56.85
	Energy	94.26	88.47	71.19	72.56	74.67	73.70	79.14
	GradNorm	86.54	79.75	53.73	55.55	57.59	63.90	66.18
	MLS+ReAct	94.50	<u>89.04</u>	77.72	80.31	80.88	**76.73**	**83.20**
	ODIN+ReAct	<u>96.76</u>	82.66	<u>85.65</u>	**76.91**	**78.48**	70.50	<u>81.83</u>
	Energy+ReAct	94.07	**89.32**	68.41	<u>75.63</u>	<u>77.25</u>	73.94	79.77

S3 Influence of training configurations for OOD detection

As demonstrated by Vaze et al. [2022], there exists a positive correlation between the closed-set accuracy and open-set performance. Likewise, we would like to verify the correlation between closed-set performance (Accuracy) and OOD detection performance (AUROC). Therefore, we train ResNet-18 with various configurations to investigate the relationship between closed-set accuracy and OOD detection performance Vaze et al. [2022]. For the OOD detection task, we use CIFAR10 as ID data and six commonly used datasets as OOD data: SVHN, Textures, LSUN, LSUN, iSUN, and Places365. For the OSR task, we experiment with CIFAR10, CIFAR+10, CIFAR+50, and TinyImageNet following the class split convention in the OSR literature. (1) For the *ReAct config*, the models are trained with a batch size of 128 for 100 epochs. The initial learning rate is 0.1 and decays by a factor of 10 at epochs 50, 75, and 90. (2) For the *MLS config*, we train the models with a batch size of 128 for 600 epochs with the cosine annealed learning rate, restarting the learning rate to the initial value at epochs 200 and 400. Besides, we linearly increase the learning rate from 0 to the initial value at the beginning. The initial learning rate is 0.1 for CIFAR10/100 but 0.01 for TinyImageNet. Broadly speaking, we train a ResNet-18 on the ID data, with an SGD optimizer and cosine annealing schedule. We train ARPL + CS and OE largely based on the official public implementation. For the auxiliary outlier dataset in the OE loss, we follow Hendrycks et al. [2019] and use a subset of 80 Million Tiny Images Torralba et al. [2008] with 300K images, removing all examples that appear in CIFAR10/100, Places or LSUN classes. The results under different configurations are reported in Tables S4 to S7. Comparing Table S5 and Table S6, we can see that the ID performance in Table S5 is better than that in Table S6 (92.99 vs 91.02). However, the OOD detection performance in Table S5 is inferior to that in Table S6, indicating that the linear correlation for OSR revealed in Vaze et al. [2022] may not hold for the OOD detection task.

Table S4: Results on OOD detection and OSR benchmarks, using ResNet-18 trained with the CE loss, adopting the ReAct config. The results are averaged from five independent runs.

Training Method	Scoring Rule	OOD detection benchmarks							OSR benchmarks					Overall
		SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG ID=92.27	CIFAR10 ID=97.13	CIFAR+10 ID=96.6	CIFAR+50 ID=96.8	TinyImageNet ID=83.4	AVG	
CE	MLS	85.42	**82.20**	86.56	86.40	85.40	84.62	85.1	92.54	95.62	91.81	81.31	90.32	87.19
	ODIN	70.62	65.32	72.49	71.68	70.43	69.10	69.94	59.77	51.37	50.22	80.96	60.58	66.20
	Energy	**85.45**	**82.20**	86.62	86.45	85.44	84.65	85.14	92.52	**95.68**	**91.86**	81.28	**90.34**	87.22
	MLS+ReAct	83.94	79.67	98.18	93.87	92.31	88.42	89.40	92.57	94.92	90.88	81.65	90.01	**89.64**
	ODIN+ReAct	83.84	79.62	**98.31**	**94.00**	**92.44**	**88.48**	**89.45**	56.65	47.76	48.40	81.30	58.53	77.08
	Energy+ReAct	83.08	76.63	95.06	94.17	92.18	85.46	87.76	**92.58**	95.02	90.99	**81.67**	90.07	88.14

Table S5: Results on OOD detection and OSR benchmarks, using ResNet-18 trained with the ARPL+CS loss, adopting the ReAct config. The results are averaged from five independent runs.

Training Method	Scoring Rule	OOD detection benchmarks							OSR benchmarks					Overall
		SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG ID=92.99	CIFAR10 ID=97.13	CIFAR+10 ID=97.75	CIFAR+50 ID=97.83	TinyImageNet ID=85.1	AVG	
ARPL+CS	MLS	79.18	86.85	95.59	95.00	94.47	89.98	90.18	**93.23**	97.93	96.27	**82.50**	**92.48**	91.10
	ODIN	50.14	38.35	74.25	42.96	43.46	53.99	50.53	92.57	97.37	95.23	53.20	84.59	64.15
	Energy	79.08	86.82	**95.68**	95.07	94.54	90.03	90.20	93.19	97.96	**96.30**	80.45	91.98	90.91
	MLS+ReAct	**84.19**	89.11	94.98	95.52	94.98	90.23	91.50	92.63	97.96	**96.30**	79.78	91.67	91.57
	ODIN+ReAct	49.22	37.19	65.89	37.80	38.77	47.55	46.07	91.00	94.74	91.84	47.47	81.26	60.15
	Energy+ReAct	**84.13**	89.13	95.10	**95.63**	**95.09**	**90.32**	**91.57**	92.59	**98.01**	**96.33**	79.90	91.71	**91.62**

Table S6: Results on OOD detection and OSR benchmarks, using ResNet-18 trained with the ARPL+CS loss, adopting the MLS config. The results are averaged from five independent runs.

Training Method	Scoring Rule	OOD detection benchmarks							OSR benchmarks					Overall
		SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG ID=91.02	CIFAR10 ID=96.96	CIFAR+10 ID=96.77	CIFAR+50 ID=96.69	TinyImageNet ID=86.91	AVG	
ARPL+CS	MLS	96.36	90.20	96.59	96.95	96.88	93.29	95.05	93.16	96.58	94.67	**84.79**	**92.30**	**93.95**
	ODIN	75.92	71.64	86.25	95.14	95.19	75.97	83.35	58.04	74.80	71.52	63.13	66.87	76.76
	Energy	**96.52**	90.11	**96.76**	**97.16**	**97.07**	**93.45**	95.18	**93.22**	**96.74**	**94.82**	82.10	91.72	93.80
	MLS+ReAct	95.87	**92.37**	96.37	96.34	96.30	92.97	95.04	92.70	96.42	94.53	82.05	91.43	93.59
	ODIN+ReAct	71.87	73.36	83.19	92.34	92.36	69.10	80.37	55.71	62.88	61.85	54.29	58.68	71.70
	Energy+ReAct	96.06	92.35	96.59	96.58	96.53	93.17	**95.21**	92.80	96.61	94.70	82.14	91.56	93.75

Table S7: Results on OOD detection and OSR benchmarks, using ResNet-18 trained with the OE loss, adopting the official configuration of OE. The results are averaged from five independent runs.

Training Method	Scoring Rule	OOD detection benchmarks							OSR benchmarks					Overall
		SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG ID=94.77	CIFAR10 ID=97.59	CIFAR+10 ID=97.38	CIFAR+50 ID=97.34	TinyImageNet ID=77.96	AVG	
OE	MLS	**95.94**	**94.57**	76.58	85.20	87.16	88.46	87.99	96.26	**98.95**	**98.20**	**77.88**	**92.82**	89.92
	ODIN	93.32	92.84	65.85	84.50	87.24	82.29	84.34	93.44	96.18	93.69	77.49	90.20	86.68
	Energy	95.84	94.45	75.50	84.55	86.64	88.19	87.54	**96.33**	**98.95**	98.04	77.73	92.76	89.62
	MLS+ReAct	95.46	94.32	**93.57**	**90.66**	90.75	88.97	**92.29**	96.20	98.93	98.18	77.60	92.73	**92.46**
	ODIN+ReAct	87.64	88.52	89.56	86.19	86.57	76.02	85.75	93.03	95.45	92.75	76.90	89.53	87.26
	Energy+ReAct	87.84	89.29	88.66	90.14	**93.69**	**94.67**	90.72	91.04	98.93	98.02	77.48	91.37	90.98

S4 Hyperparameters investigation for ReAct

To ensure fairness and assess stability in our cross-evaluation, we report the averaged results from five independent trials, carefully selected around the optimal values. We identify the best hyper-parameters of ReAct which truncates activations above threshold to limit the effect of noise. Therefore, we investigate the choice of the percentile of activations for truncation. Our experiments involve varying the percentile of pruning activations. As shown in Figure S2, ReAct's performance exhibits high sensitivity to the choice of the activation pruning percentile, and the optimal value of the percentile differs for each dataset.

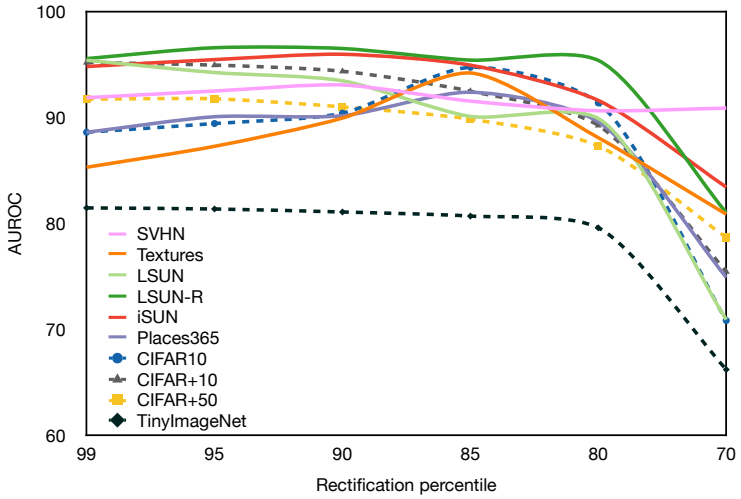


Figure S2: Investigation on the optimal hyper-parameters of ReAct for different datasets across OOD and OSR. ReAct is sensitive to the choice of the activation pruning percentile, and the optimal value of the percentile differs for each dataset.

S5 Correlation of different metrics

To demonstrate that our takeaways (in Section 3.2 in the main paper) are metric-agnostic, we further evaluate using the following metrics: (1) the area under the precision-recall curve (AUPR); (2) OSCR Dhamija et al. [2018] which summarises the trade-off between closed-set accuracy and out-distribution / open-set performance as the threshold on the open-set score is varied. Results are shown in Table S8. The same takeaways key findings remain consistent when utilizing AUROC (see also Table 2 in the main paper).

Table S8: Results of various methods on OOD detection benchmarks under different metrics (AUPR/OSCR), using CIFAR10 as ID. We train ResNet-18 with the OE loss. The results are averaged from five independent runs.

Method	OOD detection benchmarks																	
	SVHN			Textures			LSUN			LSUN-R			iSUN			Places365		
	AUROC	AUPR	OSRC	AUROC	AUPR	OSRC	AUROC	AUPR	OSRC	AUROC	AUPR	OSRC	AUROC	AUPR	OSRC	AUROC	AUPR	OSRC
OE+MLS	**99.21**	**99.54**	**94.69**	**98.82**	**98.81**	**95.16**	**99.02**	**98.13**	**94.01**	**98.53**	91.91	**90.88**	**98.57**	91.94	**90.94**	**97.32**	**99.42**	**95.25**
OE+ODIN	**99.43**	**99.42**	**94.64**	**98.73**	94.05	93.21	**99.14**	97.12	92.89	**98.78**	78.48	83.85	**98.75**	79.07	84.42	96.41	95.98	93.01
OE+Energy	**99.20**	**99.52**	**94.65**	**98.78**	**98.76**	**95.14**	**99.02**	**98.01**	**93.94**	**98.55**	91.38	90.58	**98.58**	91.52	**90.70**	**97.31**	**99.44**	**95.24**
OE+MLS+ReAct	95.18	**99.50**	94.45	92.22	98.80	95.01	79.46	97.90	93.49	83.34	**92.22**	**91.07**	83.68	**92.26**	91.13	87.46	**99.41**	95.10
OE+ODIN+ReAct	84.16	99.37	94.37	82.92	93.11	92.34	64.00	96.13	91.17	73.90	73.62	79.22	75.45	74.52	80.24	71.65	95.31	92.09
OE+Energy+ReAct	94.41	**99.49**	94.41	91.36	**98.76**	94.98	73.88	97.79	93.41	80.03	91.77	90.80	81.16	91.94	90.92	86.19	**99.43**	95.09

S6 Investigation on applying OE to the strong CLIP model

Pre-trained vision language models like CLIP Radford et al. [2021] have recently shown remarkable *robustness* to distribution shifts. The success of these models suggests that pre-training on diverse and extensive datasets is a promising approach for enhancing robustness. By utilizing pre-trained visual-language models such as CLIP, it is possible to extend OOD detection to various challenging ID datasets.

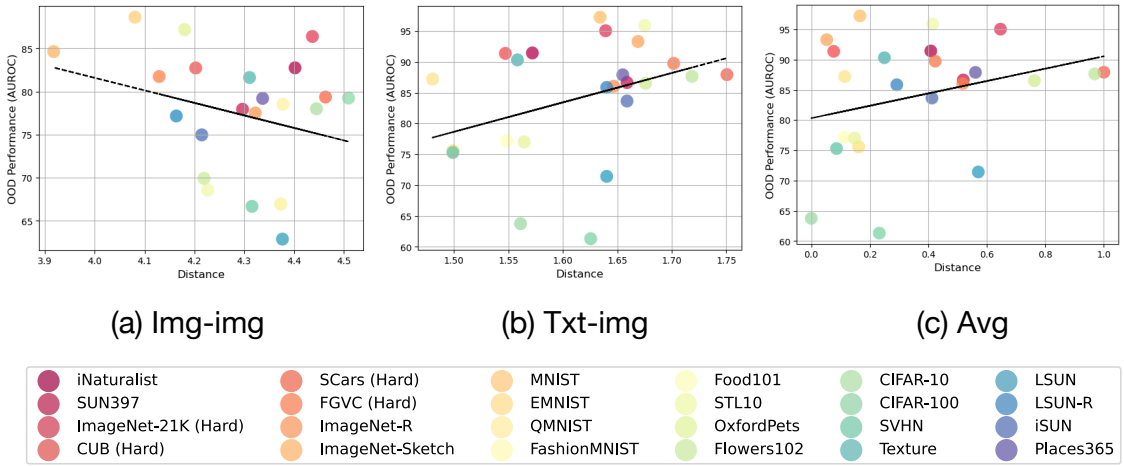


Figure S3: The average performance *vs.* distance to YFCC of all datasets when used as leave-one-out ID datasets in turn for the CLIP model. We compute cosine similarity with both image/text embeddings from ID data and image embeddings from auxiliary data. No clear correlation between the distance to YFCC and OOD detection performance is observed.

We finetune CLIP using OE with auxiliary data and apply MLS to the similarity scores to separate ID and OOD samples. For auxiliary data, we use the subset of YFCC-100M Cheng et al. [2021], known as YFCC-15M. It consists of the 15 million images which were used in CLIP pretraining. Specifically, we consider the following ID datasets: iNaturalist, SUN397, Places, hard splits of SSB according to Vaze et al. [2022] (ImageNet-21K, CUB, SCars and FGVC), variants of ImageNet (ImageNet-R and ImageNet-Sketch), variants of MNIST (MNIST, EMNIST, QMNIST and FashionMNIST), Food101, STL10, OxfordPets, Flowers102, CIFAR-10, CIFAR-100, SVHN, Texture, LSUN, LSUN-R, iSUN, Places365. As for the OOD test datasets, we treat one dataset as the hold-out ID dataset in turn to evaluate the performance of others OOD datasets that do not overlap with the ID dataset. Figure S3 examines different query types of inputs (*e.g.*, images, prompts). We use a set of pre-defined prompts for each class, which are collected from prior works Radford et al. [2021], Cherti et al. [2023]. We compute cosine similarity with the L_2-normalized embeddings of images or the embedding of prompts of each class by averaging over the prompt pool. We present the OOD detection performance *vs.* OOD-AUX distance in Figure S4 and Figure S5 by comparing different features. Interestingly, there is no obvious trend shared among all the cases. This is likely due to the fact that CLIP is pretrained on YFCC-15M, while in OOD detection, YFCC-15M is treated as auxiliary data to mimic OOD data to be pushed away from the ID data, hurting the original pretrained strong embedding space of CLIP.

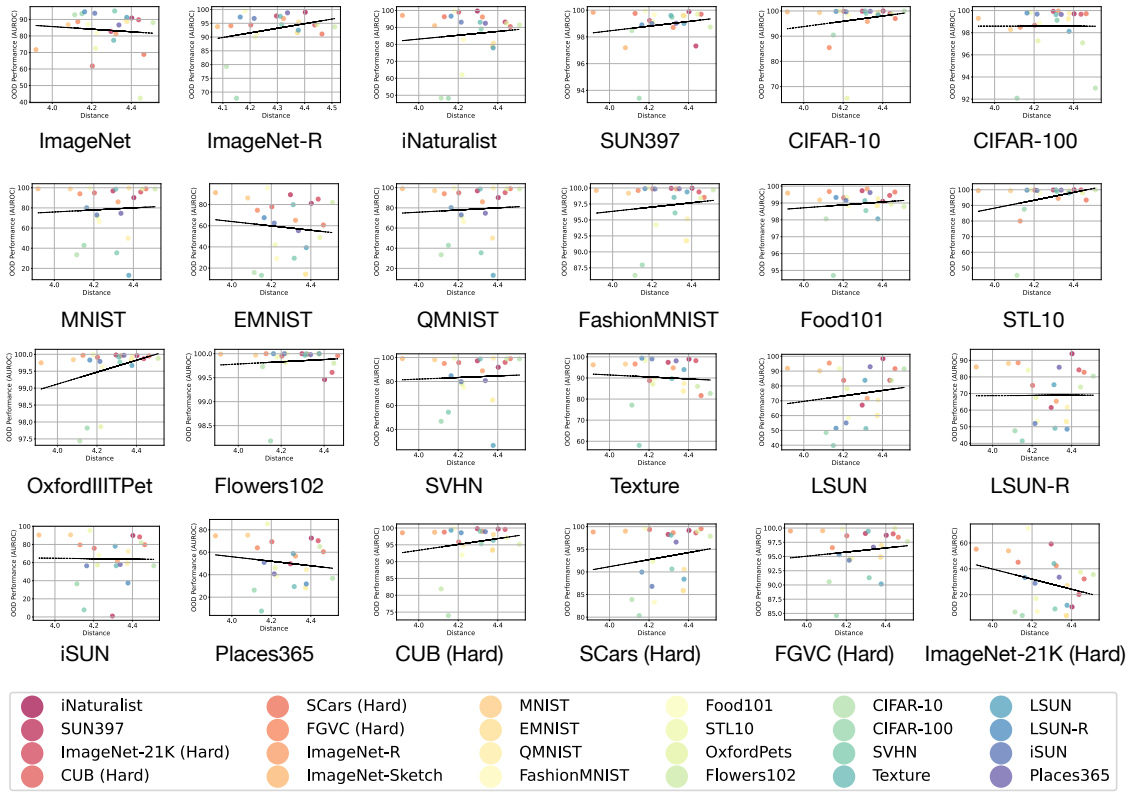


Figure S4: OOD detection performance *vs.* OOD-AUX distance to YFCC, when taking each dataset as ID dataset and others as OOD datasets for CLIP model. We calculate the distance between image features of all OOD samples and those of YFCC-15M.

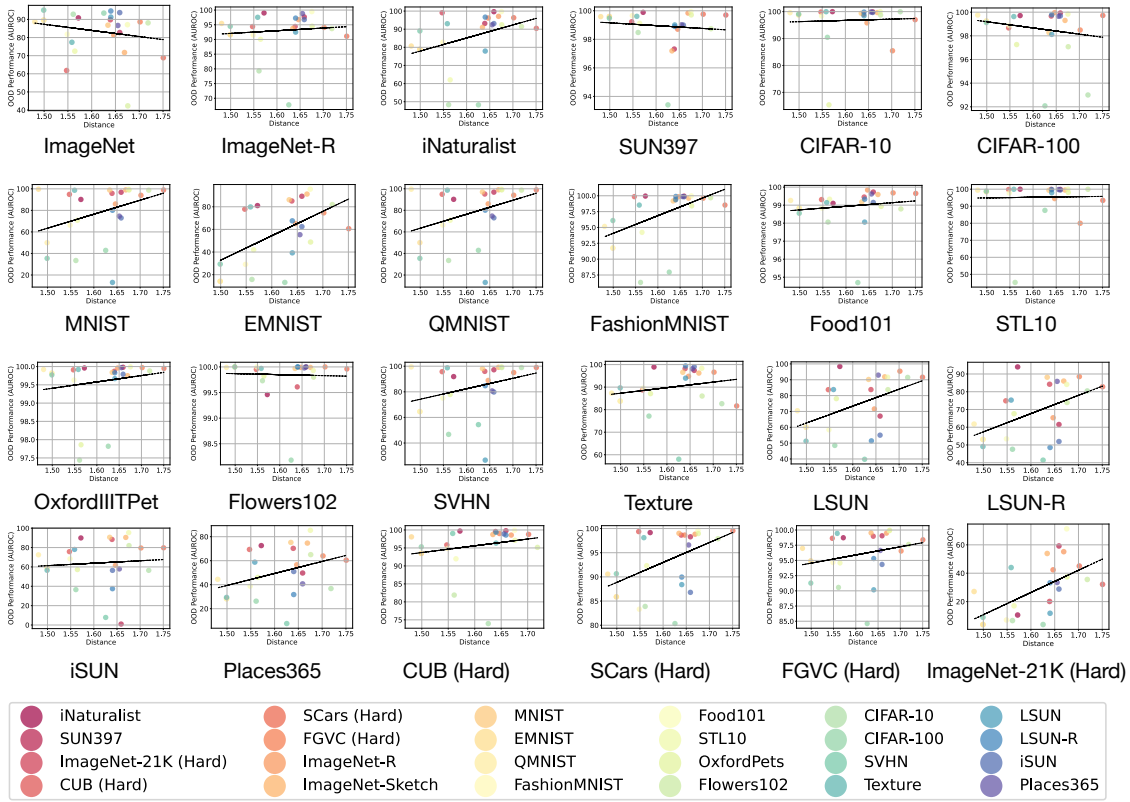


Figure S5: OOD detection performance *vs.* OOD-AUX distance to YFCC, when taking each dataset as ID dataset and others as OOD datasets for CLIP model. We calculate the distance between image features of all OOD samples and prompt features of YFCC-15M.

References

R. Cheng, B. Wu, P. Zhang, P. Vajda, and J. E. Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *CVPR*, 2021.

M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023.

A. R. Dhamija, M. Günther, and T. Boult. Reducing network agnostophobia. In *NeurIPS*, 2018.

D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE TPAMI*, 2008.

S. Vaze, K. Han, A. Vedaldi, and A. Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *ICLR*, 2022.