Supplementary Materials of LoD

Chuanxing Geng[1,2,3] , **Qifei Li**[1] , **Xinrui Wang**[1] , **Dong Liang**[1,3] , **Songcan Chen**[1,3] and **Pong C. Yuen**[2*]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]Department of Computer Science, Hong Kong Baptist University
[3]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
{gengchuanxing, liqifei, wangxinrui, liangdong, s.chen}@nuaa.edu.cn, pcyuen@comp.hkbu.edu.hk

A Detailed Proof

To prove Proposition 1, we first reintroduce Lemma 1 from [Liu *et al.*, 2020a] and Proposition 1 as follows:

Lemma 1 (Early-learning succeeds)**.** *Denote by $\{\boldsymbol{\theta}_t\}$ the iterates of gradient descent with step size η. For any $\Delta \in (0, 1/2)$, there exists a constant δ_Δ, depending only on Δ, such that if $\delta \leq \delta_\Delta$, then with high probability $1 - o(1)$, there exists a $T = \Omega(1/\eta)$ such that: for all $t < T$, we have $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| \leq 1$ and*

$$-\nabla \mathcal{L}_{CE}(\boldsymbol{\theta}_t)^T \boldsymbol{v} / \|\nabla \mathcal{L}_{CE}(\boldsymbol{\theta}_t)\| \geq 1/6.$$

Proposition 1. *Let l_i denote the loss value of each sample in $\mathcal{D}_{wild}$, which is bounded by R. $\overline{l_{in}} = \frac{1}{|\mathcal{D}_{in}^{wild}|} \sum_{i \in \mathcal{D}_{in}^{wild}} l_i$ and $\overline{l_{out}} = \frac{1}{|\mathcal{D}_{out}^{wild}|} \sum_{i \in \mathcal{D}_{out}^{wild}} l_i$ respectively denote the mean losses of ID and OOD sets from unlabeled wild data $\mathcal{D}_{wild}$, and $n = |\mathcal{D}_{in}^{wild}| + |\mathcal{D}_{out}^{wild}|$. Under the Lemma 1, with high probability, we have*

$$\overline{l_{in}} - \overline{l_{out}} \geq 1 - 2e^{-\boldsymbol{\theta}^T \boldsymbol{v} + \frac{1}{2}\|\boldsymbol{\theta}\|^2 \delta^2} - \mathcal{O}(\frac{R}{\sqrt{n}}).$$

Proof. Lemma 1 indicates that under the condition of noise level Δ, the model parameters $\boldsymbol{\theta}$ update along the proper gradient direction during the early learning stage. This means, during this period, the loss curves of ID (*label-noise*) and OOD (*label-clean*) samples in test-set will have significantly different characteristics, with larger loss values and greater fluctuations for ID samples versus smaller loss values and smaller fluctuations for OOD ones. Next, we analyze the mean loss gap between ID (label-noise) samples in $\mathcal{D}_{\text{in}}^{\text{wild}}$ and OOD (label-clean) samples in $\mathcal{D}_{\text{out}}^{\text{wild}}$ during this stage. Following [Yue and Jha, 2024], we adopt sigmoid function as the activation function for the network outputs. For each sample $(\boldsymbol{x_i}, y_i)$, we have

$$p(y_i = 1) = \text{sig}(\boldsymbol{\theta}^T \boldsymbol{x_i}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x_i}}},$$
$$p(y_i = -1) = 1 - p(y_i = 1).$$

Let $\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{z}_i$, where $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{d\times d})$. For each sample $\boldsymbol{x_i} \in \mathcal{D}_{\text{out}}^{\text{wild}}$ (label-clean), we use $\log$ for its loss, and have

$$l_i(\boldsymbol{\theta}) = \log(1 + e^{-\boldsymbol{\theta}^T(\boldsymbol{v}+\boldsymbol{z}_i)}) \leq e^{-\boldsymbol{\theta}^T(\boldsymbol{v}+\boldsymbol{z}_i)}.$$

Similarly, for each sample $\boldsymbol{x_j} \in \mathcal{D}_{\text{in}}^{\text{wild}}$ (label-noise), we have

$$l_j(\boldsymbol{\theta}) = \log(1 + e^{\boldsymbol{\theta}^T(\boldsymbol{v}+\boldsymbol{z}_i)}) \geq 1 - e^{-\boldsymbol{\theta}^T(\boldsymbol{v}+\boldsymbol{z}_i)}.$$

Taking the expectation on the difference between OOD (label-clean) and ID (label-noise), we have

$$\mathbb{E}[l_i(\boldsymbol{\theta}) - l_i(\boldsymbol{\theta})] = \mathbb{E}[l_i(\boldsymbol{\theta})] - \mathbb{E}[l_j(\boldsymbol{\theta})] \geq 1 - 2\cdot\mathbb{E}[e^{-\boldsymbol{\theta}^T(\boldsymbol{v}+\boldsymbol{z})}].$$

Note that the term $1-2\cdot\mathbb{E}[e^{-\boldsymbol{\theta}^T(\boldsymbol{v}+\boldsymbol{z})}]$ bounds the loss gap between OOD (label-clean) and know-class (label-noise) samples, and it is independent of the label type. Since

$$\mathbb{E}[e^{-\boldsymbol{\theta}^T(\boldsymbol{v}+\boldsymbol{z})}] = e^{-\boldsymbol{\theta}^T\boldsymbol{v}} \cdot \mathbb{E}[e^{-\boldsymbol{\theta}^T\boldsymbol{z}}] = e^{-\boldsymbol{\theta}^T\boldsymbol{v}} \cdot e^{\frac{1}{2}\|\boldsymbol{\theta}\|^2\sigma^2}. \quad (1)$$

Eq.(1) indicates that the smaller the σ or the projection $\boldsymbol{\theta}$ has on $\boldsymbol{v}$, the larger the expected loss gap. Interestingly, Lemma 1 ensures that we can obtain a good $\boldsymbol{\theta}$ at least within T epochs. Define the mean losses of ID (label-noise) samples and OOD (label-clean) samples as follows:

$$\overline{l_{\text{in}}} = \frac{1}{|\mathcal{D}_{\text{in}}^{\text{wild}}|} \sum_{i \in \mathcal{D}_{\text{in}}^{\text{wild}}} l_i, \quad \overline{l_{\text{out}}} = \frac{1}{|\mathcal{D}_{\text{out}}^{\text{wild}}|} \sum_{i \in \mathcal{D}_{\text{out}}^{\text{wild}}} l_i.$$

By Hoeffding's Inequality on bounded variables and the Union Bound, with probability $\geq 1 - \delta$, we have

$$\overline{l_{\text{in}}} \geq \mathbb{E}[\overline{l_{\text{in}}}] - \mathcal{O}(\frac{R}{\sqrt{|\mathcal{D}_{\text{in}}^{\text{wild}}|}} \sqrt{\log \frac{1}{\delta}}). \quad (2)$$

and

$$\overline{l_{\text{out}}} \leq \mathbb{E}[\overline{l_{\text{out}}}] + \mathcal{O}(\frac{R}{\sqrt{|\mathcal{D}_{\text{out}}^{\text{wild}}|}} \sqrt{\log \frac{1}{\delta}}). \quad (3)$$

According to Eq.(2) and Eq.(3), we have

$$\overline{l_{\text{in}}} - \overline{l_{\text{out}}} \geq 1 - 2e^{-\boldsymbol{\theta}^T\boldsymbol{v} + \frac{1}{2}\|\boldsymbol{\theta}\|^2\delta^2} - \mathcal{O}(\frac{R}{\sqrt{n}}).$$

□

B Details in Hard Benchmarks

To further demonstrate the advantages of our LoD, we conduct experiments on curated hard OOD benchmarks including CIFAR10, CIFAR+10, CIFAR+50, and TinyImageNet. The details of these benchmarks are as follows:

*Corresponding author

Methods	OOD Dataset												ACC
	SVHN		Places		LSUN-Crop		LSUN-Resize		Textures		Average		
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	
With $\mathbb{P}_{\text{in}}$ only													
MSP (ICLR'17)	48.49	91.89	59.48	88.20	30.80	95.65	52.15	91.37	59.28	88.50	50.04	91.12	94.84
ODIN (ICLR'18)	33.35	91.96	57.40	84.49	15.52	97.04	26.62	94.57	49.12	84.97	36.40	90.61	94.84
Mahalanobis (NeurIPS'18)	12.89	97.62	68.57	84.61	39.22	94.15	42.62	93.23	15.00	97.33	35.66	93.34	94.84
Energy (NeurIPS'20)	35.59	90.96	40.14	89.89	8.26	98.35	27.58	94.24	52.79	85.22	32.87	91.73	94.84
CSI (NeurIPS'20)	17.30	97.40	34.95	93.64	1.95	99.55	12.15	98.01	20.45	95.93	17.36	96.91	94.17
ReAct (NeurIPS'21)	40.76	89.57	41.44	90.44	14.38	97.21	33.63	93.58	53.63	86.59	36.77	91.48	94.84
KNN (ICML'22)	24.53	95.96	25.29	95.69	25.55	95.26	27.57	94.71	50.90	89.14	30.77	94.15	94.84
KNN+ (ICML'22)	2.99	99.41	24.69	94.84	2.95	99.39	11.22	97.98	9.65	98.37	10.30	97.99	93.19
DICE (ECCV'22)	35.44	89.65	46.83	86.69	6.32	98.68	28.93	93.56	53.62	82.20	34.23	90.16	94.84
ASH (ICLR'23)	6.51	98.65	48.45	88.34	0.90	99.73	4.96	98.92	24.34	95.09	17.03	96.15	94.84
With $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{wild}}$													
OE (ICLR'19)	0.85	99.82	23.47	94.62	1.84	99.65	0.33	99.93	10.42	98.01	7.38	98.41	94.07
Energy(w/OE) (NeurIPS'20)	4.95	98.92	17.26	95.84	1.93	99.49	5.04	98.83	13.43	96.69	8.52	97.95	94.81
WOODS (ICML'22)	0.15	99.97	12.49	97.00	0.22	99.94	0.03	99.99	5.95	98.79	3.77	99.14	94.84
SAL (ICLR'24)	0.02	99.98	2.57	99.24	0.07	99.99	0.01	99.99	0.90	99.74	0.71	99.78	93.65
LoD (Ours)	**0**	**100**	**1.72**	**99.52**	**0**	**100**	**0**	**100**	**0.66**	**99.90**	**0.48**	**99.88**	94.06

Table 1: Evaluation results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) on standard benchmarks. CIFAR10 is ID dataset, and bold numbers highlight the best results.

Methods	OOD Dataset														ACC
	SVHN		Places		LSUN-Crop		LSUN-Resize		Textures		25K RAND.IMG.		Average		
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	
π=0.1															
OE (ICLR'19)	77.74	82.84	60.70	84.02	31.06	93.99	55.74	88.45	57.39	88.27	50.95	87.44	56.53	87.51	83.04
Energy(w/OE) (NeurIPS'20)	55.89	90.19	49.08	88.03	22.74	94.94	34.10	93.42	39.33	90.63	48.91	88.12	40.23	91.44	90.02
WOODS (ICML'22)	4.90	98.70	18.53	96.27	1.94	99.53	5.73	**98.78**	17.71	96.17	10.37	96.92	9.76	97.89	94.50
SAL (ICLR'24)	5.83	97.63	17.96	96.23	2.50	98.77	5.67	98.56	8.44	97.93	8.95	97.40	8.08	97.82	93.65
LoD (Ours)	**4.41**	**98.96**	**11.82**	**97.50**	**1.84**	**99.56**	**5.61**	98.63	**4.66**	**99.10**	**8.68**	**97.59**	**5.67**	**98.75**	93.99

Table 2: Evaluation results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) on unseen datasets. We use CIFAR10 as ID and a subset (25K images) of 300K Random Images as wild OOD data. Bold numbers highlight the best results

- **CIFAR10.** CIFAR10 [Krizhevsky, 2009] contains 10 classes, where 6 classes are randomly selected as in-distribution (ID) classes, and the remaining 4 classes are used as out-of-distribution (OOD) classes.
- **CIFAR+10 & CIFAR+50.** In this set of experiments, 4 classes from CIFAR10 are randomly selected as ID classes, and 10/50 non-overlapping classes randomly selected from CIFAR100 [Krizhevsky, 2009] are OOD classes.
- **TinyImageNet.** TinyImageNet is a subset derived from ImageNet [Deng *et al.*, 2009] with a total of 200 classes, of which 20 classes are randomly selected as ID classes and the rest 180 classes are treated as OOD classes.

Please note that, since ID and OOD are randomly divided, to mitigate the effects of randomness, each dataset is evaluated across five distinct "ID/OOD" splits following [Neal *et al.*, 2018; Vaze *et al.*, 2022], and the results are averaged. Moreover, similar to standard benchmarks [Katz-Samuels *et al.*, 2022; Du *et al.*, 2024], we use 70% of data from the OOD classes as the OOD part of the unlabeled wild data.

C Additional Results on CIFAR10

In this part, we utilize CIFAR10 as the ID dataset to evaluate our LoD under $\pi = 0.1$. In addition to the four methods utilizing wild data compared in the main paper, we here also evaluate methods that rely solely on labeled ID data ($\mathbb{P}_{\text{in}}$ only) including MSP [Hendrycks and Gimpel, 2016], ODIN [Liang *et al.*, 2017], Mahalanobis [Lee *et al.*, 2018], Energy [Liu *et al.*, 2020b], CSI [Tack *et al.*, 2020], ReAct [Sun *et al.*, 2021], KNN and KNN+ [Sun *et al.*, 2022], DICE [Sun and Li, 2022] and ASH [Djurisic *et al.*, 2022]. The detailed results are presented in Table 1, which demonstrate that methods trained using both ID and wild data exhibit significantly better performance compared to those trained solely with ID data. Additionally, compared with methods utilizing $\mathbb{P}_{\text{wild}}$, LoD continues to exhibit superior performance, outperforming other SOTA methods in terms of FPR95 and AUROC metrics. Furthermore, LoD achieves competitive ID classification accuracy, either matching or exceeding the performance of leading SOTA methods such as SAL and WOODS.

D Additional Results on Unseen OOD Datasets

In this part, we follow [Du *et al.*, 2024] and evaluate our LoD on unseen OOD datasets, which are different from the OOD data we use in the wild. Table 2 and Table 3 report the results.

In Table 2, we employ CIFAR10 as the ID dataset. As for the wild OOD data, [Du *et al.*, 2024] utilizes the full 300K-image dataset. However, we argue that this setting seems inappropriate due to a significant imbalance: the ID data in the wild data contains only 25K images, while the OOD counterpart comprises 300K images–12 times larger than the ID

data. Therefore, we randomly sample a subset of 25K images from the 300K as the wild OOD data. The detailed results presented in Table 2 demonstrate that our LoD consistently outperforms SOTA baselines such as SAL and WOODS on the unseen OOD datasets, highlighting the effectiveness of our method.

In Table 3, we employ CIFAR100 as ID data. As for the wild OOD data, we follow [Du *et al.*, 2024] and utilize TinyImageNet-crop (TINc)/TinyImageNet-resize (TINr) dataset as the wild OOD data using during training and TINr/TINc as the unseen OOD data during testing. The results in Table 3 demonstrate the advantages of our LoD.

Methods	OOD Dataset			
	TINr		TINc	
	FPR95	AUROC	FPR95	AUROC
STEP (NeurIPS'21)	72.31	74.59	48.68	91.14
TSL (MM'23)	57.52	82.29	29.48	94.62
SAL (ICLR'24)	43.11	89.17	19.30	96.29
LoD (Ours)	**23.54**	**92.81**	**9.67**	**98.10**

Table 3: Evaluation results of FPR95↓ (%), AUROC↑ (%) on unseen datasets. CIFAR100 is ID, and bold numbers highlight the best results.

E Additional Results on Different Networks

To verify the applicability of LoD, the data-centric method, this part conducts experiments on different network structures on CIFAR10 and CIFAR+10. Table 4 reports the results, and we can find these networks mentioned here are all suitable for our LoD. In particular, LoD seems to follow scaling laws: the larger the network, the better it performs.

Networks(♯ params)	CIFAR10			CIFAR+10		
	FPR95	AUROC	ACC	FPR95	AUROC	ACC
WideResNet-40-2 (2.2M)	2.56	99.40	96.34	1.50	99.62	97.29
ResNet18 (11.2M)	2.47	99.44	96.46	0.96	99.72	97.32
ResNet34 (21.3M)	2.29	99.51	96.48	0.90	99.73	97.39

Table 4: Evaluation results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) on different networks on hard benchmarks.

Ratios	Places			Textures		
	FPR95	AUROC	ACC	FPR95	AUROC	ACC
1:6	10.50	98.07	72.9	8.88	97.64	74.10
1:3	8.21	98.36	73.14	8.09	98.00	73.58
1:1	4.08	99.12	73.06	6.17	98.53	74.06
3:1	3.34	99.16	72.21	4.79	98.87	73.30
6:1	3.91	98.95	71.38	3.63	99.14	73.22

Table 5: Detailed results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) across different ratios on standard benchmarks.

Ratios	CIFAR10			CIFAR+10		
	FPR95	AUROC	ACC	FPR95	AUROC	ACC
1:6	3.07	99.30	96.18	8.30	97.66	97.35
1:3	2.78	99.34	96.27	7.04	98.22	97.25
1:1	2.57	99.38	96.28	2.52	99.31	97.25
3:1	2.56	99.40	96.34	1.50	99.62	97.29
6:1	2.33	99.44	96.21	1.13	99.73	97.34

Table 6: Detailed results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) across different ratios on hard benchmarks.

Ratios	Textures			CIFAR10		
	FPR95	AUROC	ACC	FPR95	AUROC	ACC
100	4.79	98.87	73.30	2.56	99.40	96.34
200	4.17	99.07	72.86	2.37	99.45	96.28
300	4.02	99.05	72.14	2.36	99.45	96.28
400	3.82	99.05	72.19	2.33	99.46	96.30
500	3.60	99.09	72.28	2.31	99.56	96.30

Table 7: Detailed results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) in different training epochs.

F Detailed Results on Different Ratios $|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$

Table 5 reports the detailed results in different ratios on representative standard benchmark Places and Textures, while Table 6 reports the detailed results on CIFAR10 and TinyImageNet. As the ratio $|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$ increasing, the performance of our method consistently improves.

G Detailed Results on the Impact of Epoch in Early-learning Succeeds

Table 7 reports the detailed results on the impact of training epochs in early-learning success. The results demonstrate a consistent performance improvement in our LoD model as the number of training epochs increases from 100 to 500.

References

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

[Djurisic *et al.*, 2022] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

[Du *et al.*, 2024] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? *arXiv preprint arXiv:2402.03502*, 2024.

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

[Katz-Samuels *et al.*, 2022] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022.

[Krizhevsky, 2009] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[Lee *et al.*, 2018] 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.

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

[Liu *et al.*, 2020a] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[Liu *et al.*, 2020b] 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.

[Neal *et al.*, 2018] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–628, 2018.

[Sun and Li, 2022] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022.

[Sun *et al.*, 2021] 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.

[Sun *et al.*, 2022] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[Tack *et al.*, 2020] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[Vaze *et al.*, 2022] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *the International Conference on Learning Representations*, abs/2110.06207, 2022.

[Yue and Jha, 2024] Chang Yue and Niraj K Jha. Ctrl: Clustering training losses for label error detection. *IEEE Transactions on Artificial Intelligence*, 2024.