RCL: Reliable Continual Learning for Unified Failure Detection Appendix

Fei Zhu[1], Zhen Cheng[2,3], Xu-Yao Zhang[2,3], Cheng-Lin Liu[2,3], Zhaoxiang Zhang[1,2,3,4]
[1]Centre for Artificial Intelligence and Robotics, HKISI-CAS
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[3]School of Artificial Intelligence, UCAS
[4]Shanghai Artificial Intelligence Laboratory
{zhufei2018, chengzhen2019, zhaoxiang.zhang}@ia.ac.cn, {xyz, liucl}@nlpr.ia.ac.cn

A. Joint Learning of MisD and OOD Detection

When jointly learning objectives of CRL and LogitNorm, the OOD detection ability could be improved, but the MisD performance is decreased observably. The results in Table A1 show that a similar phenomenon can be observed when combining MisD method FMFP [14, 16] and LogitNorm.

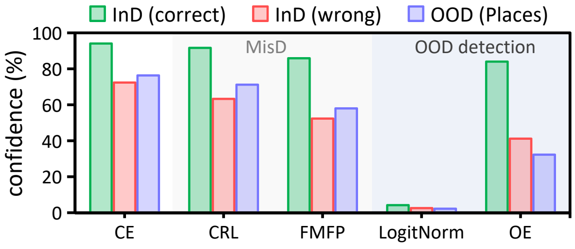

Figure A1. Average confidence of InD (correct), InD (wrong) and OOD examples with different training objectives. The dataset is CIFAR-100 and the model is ResNet110.

Different levels of sensitivity. Why jointly learning CRL or FMFP with LogitNorm is useless for unified failure detection? To understand this problem, we identify that misclassified and OOD examples might have different levels of sensitivity, leading to conflict during training. Figure A1 plots the average confidence of InD (correct), InD (wrong) and OOD examples with different training objectives. As can be observed, for the CE baseline, the relation of average confidence is InD (correct) > OOD > InD (wrong); for MisD methods like CRL and FMFP, the relation of average confidence is the same as that of CE baseline; while for OOD detection methods such as LogitNorm and OE [8], the relation of average confidence is InD (correct) > InD (wrong) > OOD. The above results indicate that misclassified examples are more sensitive than OOD examples when using MisD learning objectives, and OOD examples are more sensitive than misclassified examples when using OOD detection learning objectives. Therefore, combining them might lead to conflicts.

Table A1. Combining MisD and OOD detection objectives is useless for unified failure detection. The model used is ResNet110.

Dataset	Method	MisD			OOD Detection	
		AURC↓	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10	CE	8.96	45.72	90.43	39.54	89.86
	FMFP	5.26	20.29	94.01	20.03	94.13
	FMFP+LN	13.29	44.06	84.12	33.60	91.38
CIFAR-100	CE	90.38	52.07	84.80	70.14	73.29
	FMFP	67.91	40.86	86.86	68.46	72.91
	FMFP+LN	126.87	75.35	73.16	49.47	84.12

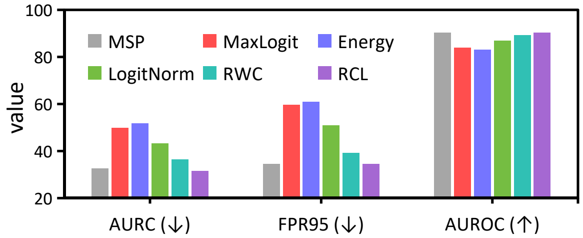

Figure A2. MisD performance on ImageNet-200 with ResNet50. The proposed RCL can maintain the performance of MisD when improving OOD detection performance.

B. Evaluation on the ImageNet

Setup. We train on ImageNet-200 [3] with resolution 224 × 224 using a ResNet50 backbone [5]. We train 90 epochs using SGD with momentum 0.9, weight decay 1e-4 and batch-size 256. The start learning rate is 0.1 and decays by a factor of 10 at epochs 30 and 60, respectively. For OOD detection, the OOD datasets include iNaturalist [9], SUN [12], Places [13] and Textures [2] with non-overlapping cat-

Table A2. MisD, OOD detection and unified failure detection performance on ImageNet-200 using a ResNet50 backbone.

Method	MisD			OOD Detection		Failure Detection			ID-ACC
	AURC↓	FPR95↓	AUROC↑	FPR95↓	AUROC↑	AURC↓	FPR95↓	AUROC↑	
MSP	32.69	34.50	90.34	42.36	86.29	243.05	32.60	91.67	83.49
MaxLogit	49.86	59.77	84.04	38.01	89.47	246.47	37.00	91.83	83.49
Energy	51.77	61.01	83.13	38.16	89.64	248.05	37.72	91.55	83.49
LogitNorm	43.20	51.00	86.91	29.91	91.80	232.90	27.28	94.11	83.13
PwF	36.57	39.13	89.31	35.66	89.65	234.47	28.11	93.40	83.51
RCL	31.52	34.56	90.43	39.49	87.88	235.62	30.42	92.61	83.94

Table A3. Comparison of MisD, OOD detection and unified failure detection performance when tuning different parts of the network. Results are obtained on ResNet110 trained on CIFAR-100.

Method	MisD			OOD Detection		Failure Detection			ID-ACC
	AURC↓	FPR95↓	AUROC↑	FPR95↓	AUROC↑	AURC↓	FPR95↓	AUROC↑	
MSP	90.38	52.07	84.80	70.14	73.29	143.68	56.82	83.60	73.04
CRL	76.93	41.40	86.92	73.05	71.97	139.17	56.25	84.16	73.92
w/ BN	76.46	42.08	87.16	70.67	72.85	136.27	54.06	84.90	73.81
w/ Block 3	85.93	45.85	86.16	62.26	77.34	130.78	48.94	87.24	72.59
w/ Full	80.56	46.68	86.14	56.82	80.95	123.90	46.98	87.73	74.04
FMFP	67.25	40.86	86.86	68.46	72.91	132.87	52.94	84.18	75.87
w/ BN	67.04	41.45	87.03	67.54	72.98	132.52	52.68	84.26	75.82
w/ Block 3	71.21	41.29	86.89	62.90	75.70	128.59	49.59	85.44	74.98
w/ Full	67.91	43.09	87.11	59.63	79.12	122.69	47.46	86.75	75.71

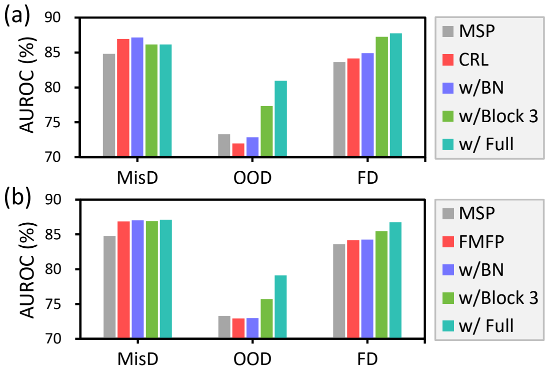

Figure A3. Selective layer tuning on (a) CRL and (b) FMFP.

egories *w.r.t.* ImageNet. The averaged result over those four OOD datasets is reported. For unified failure detection, we keep the same number of misclassified and OOD examples. Since the MSP score with CE training objective has been verified to be the state-of-the-art failure detection approach [1, 10], we directly tuning the CE baseline.

Results. In Figure A2 and Table A2, we can find that MaxLogit, Energy and LogitNorm ruin the performance of CE baseline with MSP. Our method can successfully improve OOD detection performance while maintaining the MisD ability. Since we keep the same number of OOD and misclassified examples when evaluating unified failure detection, strong OOD detection method LogitNorm could achieve good failure detection performance. However, its MisD ability is much worse than CE baseline, which is undesirable in practice because misclassified InD examples widely exist and should be detected and rejected effectively. Overall, our approach achieves a good balance between MisD and OOD detection and outperforms the previous state-of-the-art method (i.e., CE baseline) consistently.

C. Selective Layer Tuning

In Section 5.2, we show that the proposed reliable continual learning paradigm can be performed in a more efficient way by only tuning selective informative layers in a deep neural network based model. Here we provide more results. Figure A3 compares the performance (AUROC) of MisD, OOD detection and failure detection (FD) when tuning different layers of ResNet110 on CIFAR-100. We can find that: (1) Only tuning Batch-Normalization (BN) layers benefits MisD, while leaving little freedom to acquire OOD detection ability; (2) Tuning Black 3 that contains those layers with rich semantic knowledge (near the output of a model) is effective for improving OOD detection performance. Detailed results are provided in Table A3.

Table A4. Detecting failures in distribution shifts scenario. Models are trained on clean training dataset with ResNet110.

Severity	Method	CIFAR-10-C			CIFAR-100-C		
		AURC↓	FPR95↓	AUROC↑	AURC↓	FPR95↓	AUROC↑
#1	MSP	52.97	58.00	85.41	206.49	59.52	80.87
	ours	**39.63**	**33.19**	**89.32**	**188.32**	**56.43**	**82.24**
#5	MSP	311.51	75.64	72.85	537.26	72.57	73.62
	ours	**249.88**	**57.42**	**78.08**	**506.60**	**69.84**	**74.82**

D. Failure detection under Distribution Shift

As suggested by recent works [10, 14–16], a more general failure detection should include detecting failures in distribution shifts scenarios like image corruptions [7] and iWildCam [11]. Table A4 shows that our method performs remarkably better than MSP baseline on distribution shifted dataset CIFAR-10-C and CIFAR-100-C [7], demonstrating its effectiveness in more general failure detection setting.

Table A5. How the order of learning MisD and OOD detection impacts the effectiveness of our method. CIFAR-100 / ResNet110.

Method	AURC↓	FPR95↓	AUROC↑
MSP	143.68	56.82	83.60
OOD⇒MisD	134.48	56.50	85.63
MisD⇒OOD	**123.90**	**46.98**	**87.73**

Table A6. Using OE as OOD detection method in our RCL framework. CIFAR-100 / ResNet110.

Method	AURC↓	FPR95↓	AUROC↑
CRL	139.17	56.25	84.16
ours (FMFP-OE)	**116.64**	44.75	87.61
ours (CRL-OE)	119.07	**43.67**	**88.15**

E. Analysis and Discussion

Order of learning MisD and OOD detection. Table A5 presents the performance of unified failure detection on CIFAR-100 when tuning an OOD detector using MisD method, and our framework still outperforms MSP. We also note that `MisD⇒OOD` performs better than `OOD⇒MisD`. We suspect that this is because OOD methods often have negative effect on MisD [10, 14, 15], which consumes part of the MisD performance when tuning model with MisD later. Nevertheless, our approach provides a simple and effective way to achieve the challenging goal of unified failure detection.

Using OE as OOD detection method in RCL. Table A6 reports the performance of our method when using outlier-based OE [6] as the OOD detection technique in the proposed reliable continual learning framework. As can be observed, our method yields stronger performance.

Table A7. Failure detection performance with different λ. CIFAR-100 / ResNet110.

Metric	$\lambda = 1e2$	$1e3$	$2.5e3$	$5e3$	$1e4$	$5e4$	$1e5$
AURC↓	122.66	124.96	122.65	121.71	123.90	121.90	124.15
FPR95↓	46.30	47.27	45.13	45.36	46.98	45.44	47.13
AUROC↑	87.44	87.13	87.61	87.69	87.73	87.49	86.97

Table A8. Experimental results of training ViT on CIFAR-100.

Method	AURC↓	FPR95↓	AUROC↑
MSP	29.25	29.54	92.68
ours	**28.96**	**28.61**	**93.08**

Failure detection performance with different λ. Although the results of MisD and OOD detection ability vary according to λ (*this also gives us a chance to flexibly control the ability of MisD and OOD detection*), the unified failure detection performance is quite stable as shown in Table A7.

ViT experiments. We conduct experiments by tuning supervised (ImageNet21K) pretrained ViT [4] on CIFAR-100. Particularly, we find that the supervised pretrained ViT model has strong failure detection ability, and the OOD detection ability can not be further improved with LogitNorm. Therefore, we tune it with FMFP [14] and stronger OOD detection method OE [6]. Table A8 reports unified failure detection results, verifying the effectiveness of our reliable continual learning framework.

References

[1] Mélanie Bernhardt, Fabio De Sousa Ribeiro, and Ben Glocker. Failure detection in medical image classification: A reality check and benchmarking testbed. *Transactions on Machine Learning Research*, 2022. 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3606–3613, 2014. 1

[3] 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 Recognition*, pages 248–255, 2009. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[5] 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. 1

[6] Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich.

Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 3

[7] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. 3

[8] 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. In *International Conference on Machine Learning*, 2022. 1

[9] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8769–8778, 2018. 1

[10] Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3

[11] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 3

[12] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3485–3492, 2010. 1

[13] Bolei Zhou, Àgata 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, 2018. 1

[14] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *European Conference on Computer Vision*, pages 518–536. Springer, 2022. 1, 3

[15] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Openmix: Exploring outlier samples for misclassification detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12074–12083, 2023. 3

[16] Fei Zhu, Xu-Yao Zhang, Zhen Cheng, and Cheng-Lin Liu. Revisiting confidence estimation: Towards reliable failure prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3