Supplementary material
SURE: SUrvey REcipes for building reliable and robust deep networks

This supplementary material contains the following sections:

- Section 1: Ablation study of λ_{mix} and λ_{crl} for the RegMixup [15] loss and Correctness Ranking Loss (CRL) [14].
- Section 2: Ablation study of τ for the Cosine Similarity Classifier (CSC) [5, 9].
- Section 3: Comparison of the performance of different uncertainty estimation methods on CIFAR10-LT and CIFAR100-LT [12] with imbalance factor 10.
- Section 4: More details about the definition of Area Under the Receiver Operating Characteristic Curve (AUROC) [2] and False Positive Rate at 95% True Positive Rate (FPR95) as mentioned in Section **4.1** (c.f. line 388) in our paper.
- Section 5: More results of failure prediction under distribution shift.
- Section 6: Ablation study of different re-weighting maps.

1. Impact of different λ_{crl} and λ_{mix} in RegMixup [15] loss and Correctness Ranking Loss (CRL) [14]

In this section, we present the results of varying the parameters λ_{crl} and λ_{mix} in the loss function of **SURE**. The experimental results, obtained using a ResNet18 [6] backbone and summarized in Table 1, indicate that different datasets require different optimal weights. Notably, all experiments across various backbones consistently utilized the same values of λ_{crl} and λ_{mix} in our paper. We determined the optimal settings as 0.5 for both λ_{crl} and λ_{mix} on CIFAR10 [12], 1 for CIFAR100 [12], and 2 for Tiny-ImageNet [13]. Specifically, when we fine-tuned DeiT [17], we set λ_{crl} to 0 and λ_{mix} to 0.2 across three datasets. Particularly in our downstream task, we set λ_{crl} to 0 and λ_{mix} to 1 when addressing the challenges of long-tailed distribution data. And we set λ_{crl} to 0.2 and λ_{mix} to 1 when learning with noisy labels.

2. Impact of different τ in Cosine Similarity Classifier (CSC) [5, 9]

In the same vein as the previous ablation study for λ_{crl} and λ_{mix}, we also conducted an analysis of the cosine similarity classifier temperature τ within the **SURE** framework. This study is detailed in Table 2. For CIFAR10 [12] and CIFAR100 [12], the best-performing temperature value was found to be $\tau = 8$, while for Tiny-ImageNet [13], a higher temperature of $\tau = 16$ yielded superior results. Specifically, when we fine-tuned DeiT [17], we set the temperature of $\tau = 16$ on three datasets. Note that across all our downstream tasks, we consistently applied a temperature of $\tau = 8$.

3. More results of failure prediction on CIFAR10-LT and CIFAR100-LT [12]

We evaluate the performance of failure prediction under imbalanced data distribution. The Acc. and AURC are provided in Table 3 for imbalance factor IF = 10. We find that even under imbalanced data distribution, our SURE still significantly outperforms other approaches of failure prediction across different datasets and backbones, demonstrating its robustness under more challenging settings.

Ratios	CIFAR10 [12]		CIFAR100 [12]		Tiny-ImageNet [13]	
	Acc. ↑	AURC ↓	Acc. ↑	AURC ↓	Acc. ↑	AURC ↓
Baseline(MSP)	95.41 ± 0.15	4.89 ± 0.96	74.91 ± 0.25	74.87 ± 0.24	63.27±0.04	134.87±1.14
CRL weight λ_{crl}						
0.1	95.47±0.19	4.60±0.26	75.47±0.46	75.02±2.99	63.32±0.23	135.62±2.56
0.2	95.33±0.26	4.13±0.64	76.04±0.78	73.03±2.04	63.44±0.16	131.62±1.37
0.5	**95.33±0.14**	**3.98±0.20**	75.49±0.39	71.84±1.49	64.86±0.02	124.63±0.49
1	95.13±0.16	4.67±0.40	**76.10±0.43**	**69.05±2.48**	65.29±0.14	117.33±1.08
2	93.99±0.08	6.71±0.28	75.30±0.36	72.40±1.48	**65.59±0.18**	**116.61±0.47**
5	91.58±0.18	13.29±0.33	71.98±0.55	91.42±2.15	62.66±0.17	136.03±0.94
RegMixup regularization weight λ_{mix}						
0.1	95.76±0.08	5.81±0.98	77.59±0.67	66.49±2.09	65.42±0.40	123.37±1.00
0.2	95.85±0.11	4.74±0.41	77.35±0.39	66.59±0.77	65.59±0.20	122.26±0.67
0.5	**96.23±0.10**	**4.68±0.47**	77.21±0.52	66.32±1.96	66.26±0.21	116.50±2.31
1	95.96±0.29	7.04±0.92	**77.64±0.85**	**63.88±5.22**	66.00±0.22	117.79±1.49
2	96.03±0.07	7.03±0.45	77.13±0.31	66.56±0.43	**66.26±0.12**	**113.40±1.31**
5	95.83±0.23	6.17±1.74	77.52±0.95	63.40±6.22	65.40±2.06	119.34±12.49

Table 1. **Ablation Study of hyper-parameters λ_{crl} and λ_{mix} in the loss function of SURE.** Experiments are implemented on CIFAR10, CIFAR100 [12] and Tiny-ImageNet [13] datasets.

Ratios	CIFAR10 [12]		CIFAR100 [12]		Tiny-ImageNet [13]	
	Acc. ↑	AURC ↓	Acc. ↑	AURC ↓	Acc. ↑	AURC ↓
Baseline(MSP)	95.41±0.15	4.89±0.96	74.91±0.25	74.87±0.24	63.27±0.04	134.87±1.14
cosine similarity classifier temperature τ						
4	96.29±0.01	2.44±0.04	79.73±0.22	53.71±0.16	64.86±0.14	128.28±1.76
8	**96.65±0.07**	**2.13±0.03**	**80.37±0.07**	**48.20±0.73**	68.26±0.05	99.76±0.59
16	96.17±0.10	2.52±0.07	79.90±0.35	50.28±1.29	**69.03±0.05**	**94.63±0.74**
32	96.20±0.10	2.51±0.06	79.07±0.32	53.14±1.82	67.44±0.29	103.51±1.89

Table 2. **Ablation Study of hyper-parameters τ in Cosine Similarity Classifier (CSC) of SURE.** Experiments are implemented on CIFAR10, CIFAR100 [12] and Tiny-ImageNet [13] datasets.

4. Definition of AUROC [2] and FPR95

AUROC The area under the receiver operating characteristic curve (AUROC) measures the area under the curve drawn by plotting the true positive(TP) rate against the false positive(FP) rate.

FPR95 FPR95 is the abbreviation of FPR-at-95%-TPR that measures the false positive rate (FPR) = FP/(FP+TN) when the true positive rate (TPR) = TP/(TP+FN) is 95%, where TP, TN, FP, and FN denotes true positives, true negatives, false positives, and false negatives, respectively. It can be interpreted as the probability that an example predicted incorrectly is misclassified as a correct prediction when TPR is equal to 95%.

5. More results of failure prediction under distribution shift

In this section, we present the detailed performances of each corruption in Figure 1. We can observe that **SURE** outperforms the other methods in almost all corruption types. This consistent superiority across various corruption types indicates the robustness of **SURE**.

6. Impact of different re-weighting maps

In this section, we investigate the impact of different re-weighting maps on our uncertainty-aware re-weighting strategy in Table 4. Specifically, we explore four methods: exponential (exp), threshold, power, and linear. Let s_i represent the confidence scores. We define these re-weighting methods with tuning parameters t, α, and p as follows:

Backbones	Methods	CIFAR10-LT [1] IF=10 Acc. ↑	CIFAR10-LT [1] IF=10 AURC ↓	CIFAR100-LT [1] IF=10 Acc. ↑	CIFAR100-LT [1] IF=10 AURC ↓
ResNet18 [6]	MSP [8]	88.49±0.18	40.96±3.19	59.39±0.23	196.28±3.57
	RegMixup [15]	91.28±0.15	17.74±0.99	62.51±1.13	156.56±4.06
	CRL [14]	88.21±0.14	38.78±2.24	60.33±0.29	181.33±3.63
	SAM [3]	88.56±0.38	27.44±1.39	60.24±0.44	183.68±3.17
	SWA [11]	90.37±0.15	20.88±0.90	63.86±0.11	157.43±1.63
	FMFP [19]	90.46±0.06	18.55±0.35	63.20±0.44	153.88±1.91
	SURE	**92.65±0.11**	**14.68±0.86**	**66.83±0.38**	**122.18±0.93**
VGG16-BN [16]	MSP [8]	86.65±0.16	84.26±4.55	57.96±0.28	257.81±1.84
	RegMixup [15]	89.53±0.30	26.75±0.39	61.75±0.08	200.65±4.04
	CRL [14]	86.45±0.21	87.05±1.79	57.69±0.25	255.38±5.34
	SAM [3]	88.24±0.51	40.77±3.57	59.17±0.48	223.72±6.66
	SWA [11]	89.23±0.05	25.02±0.66	60.95±0.51	188.60±5.36
	FMFP [19]	89.23±0.22	21.55±0.34	61.12±0.22	179.68±1.90
	SURE	**90.47±0.23**	**19.51±0.59**	**62.31±0.36**	**158.17±2.43**
DenseNetBC [10]	MSP [8]	87.75±0.53	37.94±7.71	58.61±0.03	225.57±2.51
	RegMixup [15]	91.73±0.16	17.07±0.12	65.14±0.10	131.85±1.81
	CRL [14]	88.11±0.21	38.65±1.47	60.06±0.15	188.90±3.69
	SAM [3]	88.79±0.29	27.02±1.23	61.14±0.34	188.08±3.77
	SWA [11]	90.76±0.40	16.77±1.06	64.52±0.75	149.15±5.80
	FMFP [19]	90.72±0.49	15.80±1.37	**65.62±0.24**	136.10±1.03
	SURE	**91.76±0.23**	**13.72±0.72**	65.34±0.08	**130.95±2.23**
WRNet28 [18]	MSP [8]	89.44±0.10	37.28±1.34	62.46±0.05	185.31±0.83
	RegMixup [15]	92.44±0.29	14.66±1.96	65.99±0.60	144.91±3.02
	CRL [14]	89.57±0.28	37.63±2.31	63.22±0.24	159.26±2.60
	SAM [3]	90.86±0.13	21.11±0.72	65.27±0.13	145.33±2.15
	SWA [11]	92.17±0.27	12.70±0.83	68.73±0.17	122.27±1.09
	FMFP [19]	92.04±0.07	11.35±0.17	69.12±0.40	111.44±1.31
	SURE	**93.91±0.01**	**9.40±0.41**	**70.92±0.27**	**102.64±1.85**

Table 3. **Comparison of the performance of failure prediction on CIFAR10-LT and CIFAR100-LT [1] with imbalance factor 10.** We keep 10% training data as the validation set to select the best model. The means and standard deviations over *three* runs are reported. ↓ and ↑ indicate that lower and higher values are better respectively. For each experiment, the best result is shown in boldface. AURC [4] values are multiplied by 10^3 and all remaining values are in percentage. On datasets with long-tailed distributions, SURE outperforms other methods in almost all cases.

- **Exponential:** The weights are defined using the exponential function:

$$\text{weights} = e^{-t \times s_i}$$

  where t is a scaling factor affecting the influence of confidence scores.
- **Threshold :**

$$\text{weights} = \begin{cases} 1.0 - s_i, & \text{if } s_i < \alpha \\ 0, & \text{otherwise} \end{cases}$$

  Here, α is the threshold value.
- **Power:** The weights are determined by raising the term to a power:

$$\text{weights} = (1.0 - s_i)^p$$

  In this case, p is the exponent applied to the term $1.0 - s_i$.

Methods	Acc.
w/o. re-weighting	87.72
exp	
$t = 0.5$	89.73
$t = 1$	**90.22**
$t = 2$	88.96
threshold	
$\alpha = 0.5$	89.35
$\alpha = 0.6$	89.50
$\alpha = 0.7$	89.01
$\alpha = 0.8$	89.60
$\alpha = 0.9$	89.87
power	
$p = 2$	89.82
$p = 3$	89.44
$p = 4$	89.60
$p = 5$	89.25
linear	89.60

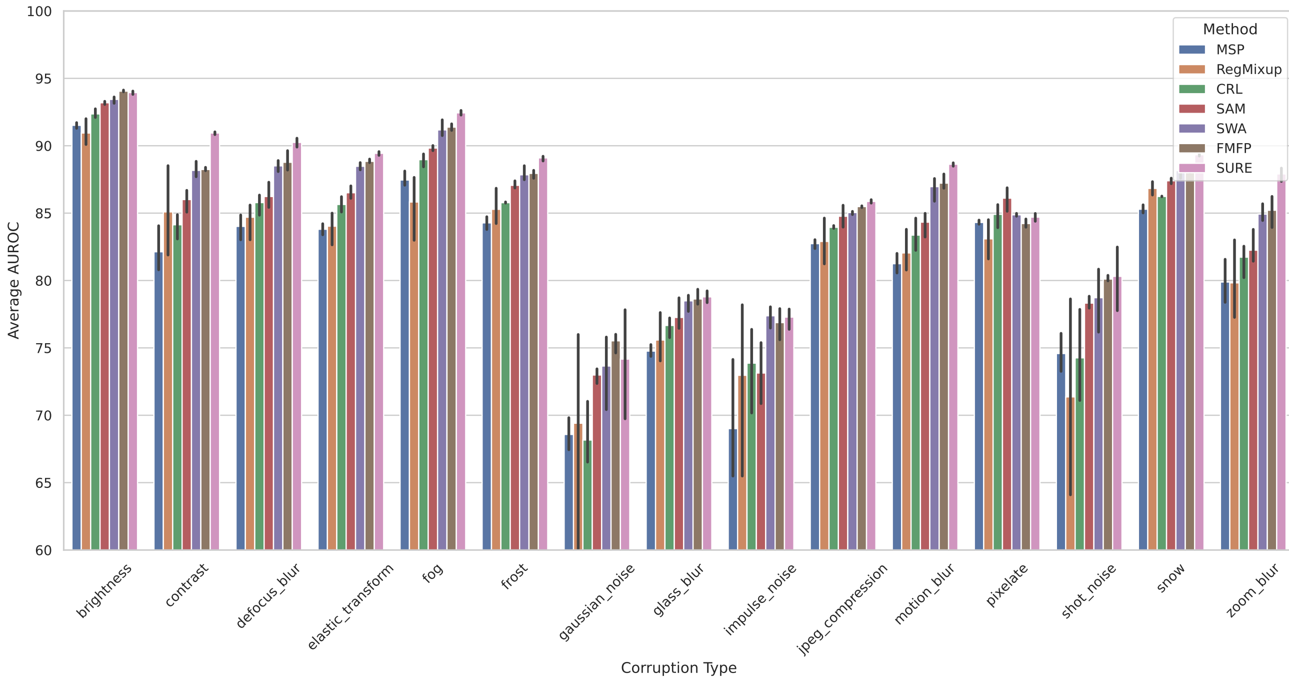
Table 4. **Impact of different re-weighting maps.** We have investigated the impact of different re-weighting maps on our uncertainty-aware re-weighting strategy on CIFAR10-LT [1] with an Imbalance Factor (IF) of 50. Based on our findings, 'exp' (exponential) method with $t = 1$ was selected as the re-weighting map for all our long-tailed classification experiments.

- **Linear:** A linear relationship is used to calculate the weights:

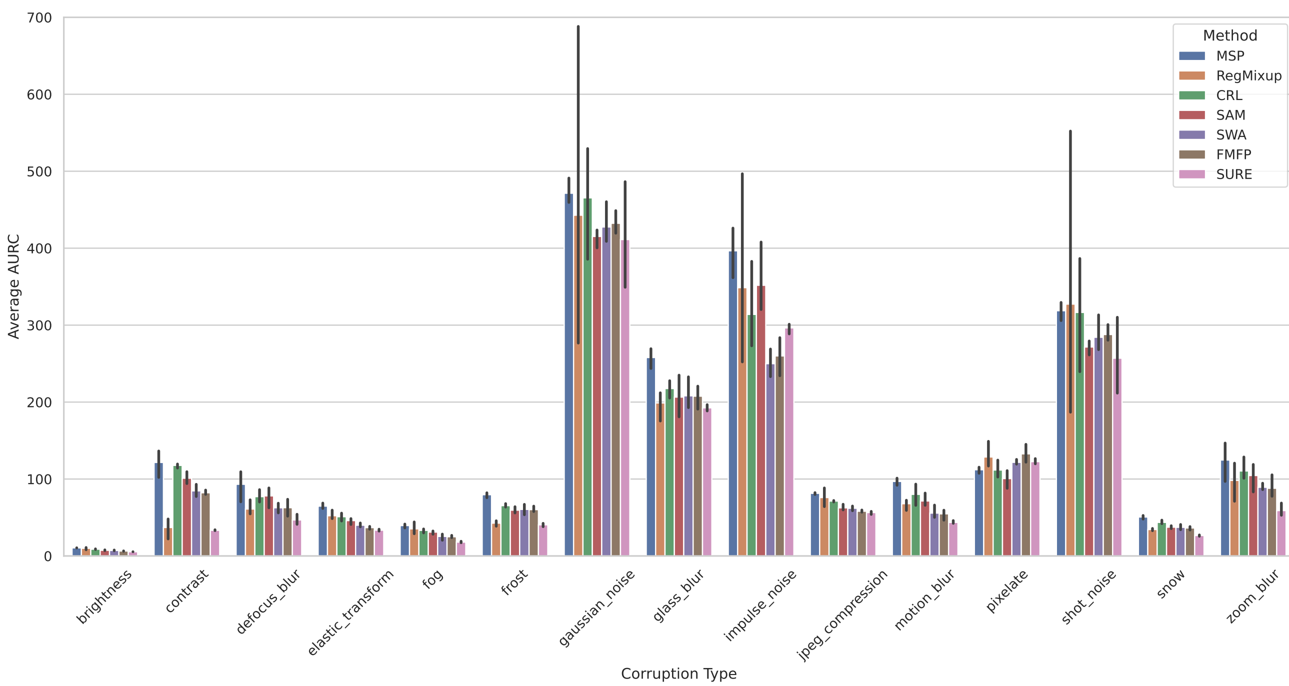
$$\text{weights} = 1.0 - s_i$$

This method directly subtracts the confidence scores from 1.0.

Based on the best result in Table 4, we choose "exp" (exponential) with t = 1 as the re-weighting map for all our long-tail classification experiments.



(a) AUROC



(b) AURC

Figure 1. **Comparison of the average AUROC [2] (higher is better) and AURC [2] (lower is better) on CIFAR10-C [7].** We choose DenseNet [10] as the backbone and CIFAR-10 as the training set. The evaluation results are averaged across the images with 5 severity levels under 15 types of corruption.

References

[1] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 3, 4

[2] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006. 1, 2, 5

[3] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2020. 3

[4] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *ICLR*, 2018. 3

[5] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 3

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 5

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

[9] Shell Xu Hu, Pablo G Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil D Lawrence, and Andreas Damianou. Empirical bayes transductive meta-learning with synthetic gradients. In *ICLR*, 2020. 1

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 3, 5

[11] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv*, 2018. 3

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009. 1, 2

[13] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 1, 2

[14] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, 2020. 1, 3

[15] Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *NeurIPS*, 2022. 1, 3

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *CVPR*, 2014. 3

[17] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1

[18] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 3

[19] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *ECCV*, 2022. 3