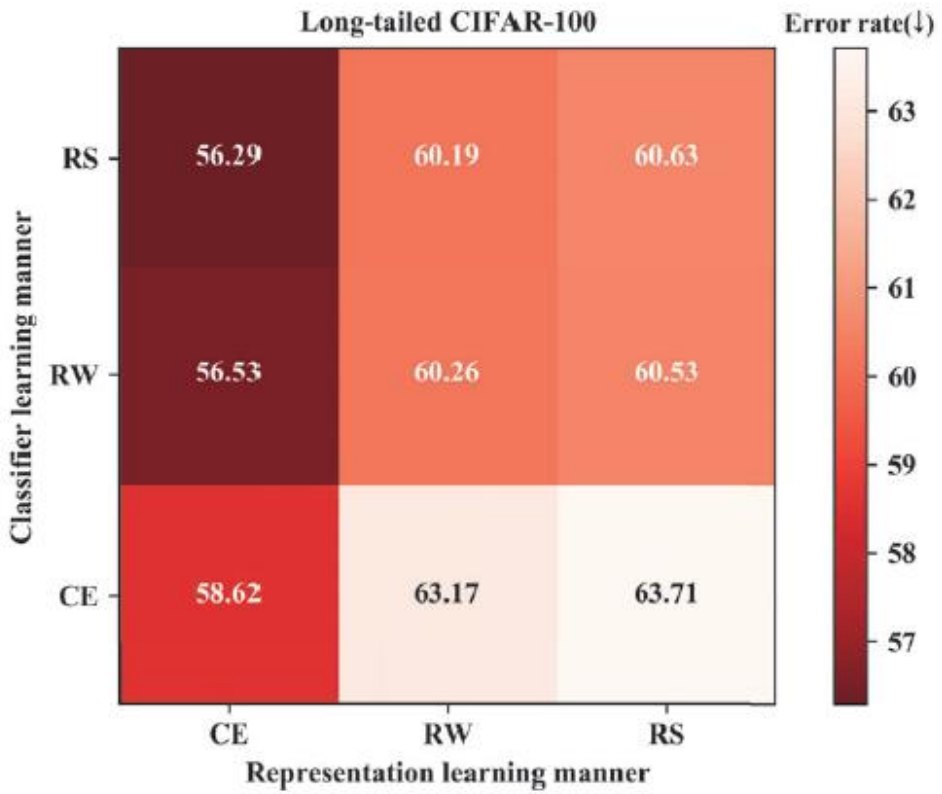
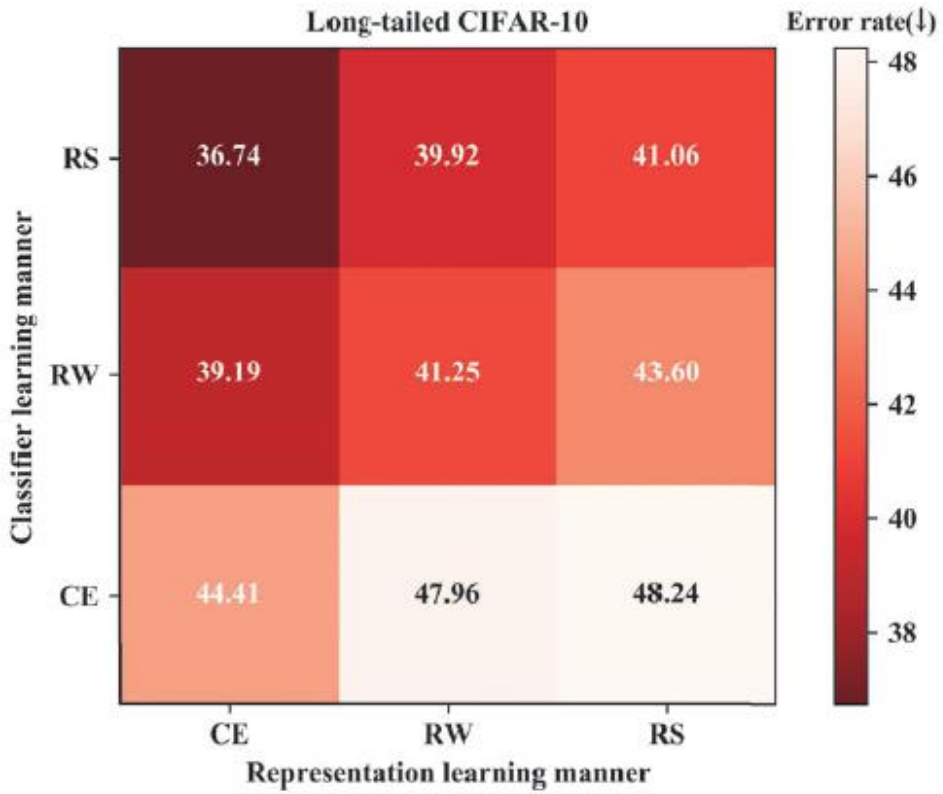
Improving Calibration for Long-Tailed Recognition

Zhisheng Zhong Jiequan Cui Shu Liu Jiaya Jia
Chinese University of Hong Kong SmartMore

Code: https://github.com/Jia-Research-Lab/MiSLAS

motivation

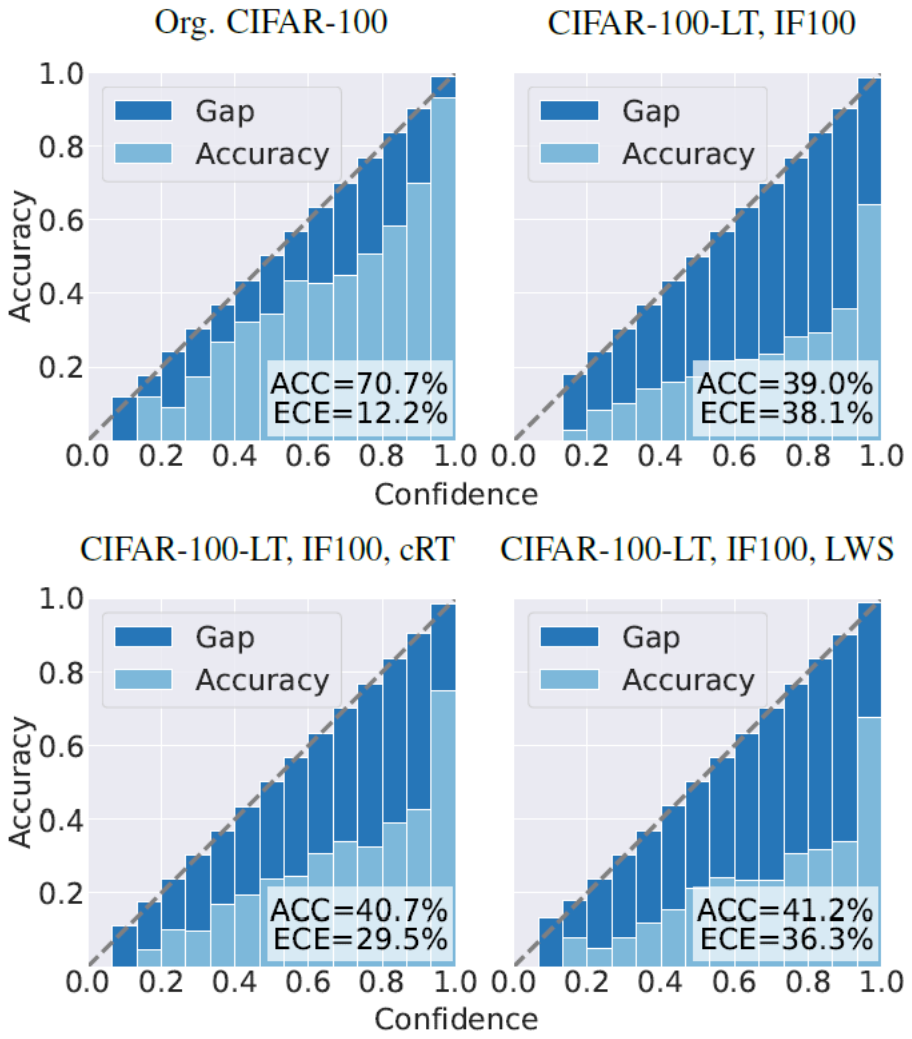

Long-tailed CIFAR-100
Error rate(↓)
RS
RW
CE
56.29
60.19
60.63
56.53
60.26
60.53
58.62
63.17
63.71
63
62
61
60
59
58
57
Classifier learning manner
CE
RW
RS
Representation learning manner


Long-tailed CIFAR-10
Error rate(↓)
RS
RW
CE
36.74
39.92
41.06
39.19
41.25
43.60
44.41
47.96
48.24
48
46
44
42
40
38
Classifier learning manner
CE
RW
RS
Representation learning manner

introduction

Expected calibration error (ECE)

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{S}_b|}{N} \Big| \mathrm{acc}(\mathcal{S}_b) - \mathrm{conf}(\mathcal{S}_b) \Big| \times 100\%$$



Methods-mixup

Mixup

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j, \quad \text{where } y_i, y_j \text{ are one-hot label encodings}$$

Mark	Stg.-1	Stg.-2	ResNet-50	ResNet-101	ResNet-152
CE	☒		45.7 / 13.7	47.3 / 13.7	48.7 / 14.5
CE	☑		45.5 / 7.98	47.7 / 10.1	48.3 / 10.2
cRT	☒	☒	50.3 / 8.97	51.3 / 9.34	52.7 / 9.05
cRT	☒	☑	50.2 / 3.32	51.3 / 3.38	52.8 / 3.60
cRT	☑	☒	**51.7** / 5.62	**53.1** / 6.86	**54.2** / 6.02
cRT	☑	☑	51.6 / **3.13**	53.0 / **2.93**	54.1 / **3.37**

Mark	Stg.-1	Stg.-2	ResNet-50	ResNet-101	ResNet-152
CE	☒		45.7 / 13.7	47.3 / 13.7	48.7 / 14.5
CE	☑		45.5 / 7.98	47.7 / 10.1	48.3 / 10.2
LWS	☒	☒	51.2 / 4.89	52.3 / 5.10	53.8 / 4.48
LWS	☒	☑	51.0 / 5.01	52.2 / 5.38	53.6 / 5.50
LWS	☑	☒	52.0 / **2.23**	**53.5** / **2.73**	**54.6** / **2.46**
LWS	☑	☑	52.0 / 8.04	53.3 / 6.97	54.4 / 7.74

Table 1: Top-1 accuracy (%) and ECE (%) of the plain cross-entropy (CE) model, and decoupling models of cRT (left) and LWS (right), for ResNet families trained on the ImageNet-LT dataset. We vary the augmentation strategies with (☑), or without (☒) mixup $\alpha = 0.2$, on both of the stages.

Methods-mixup

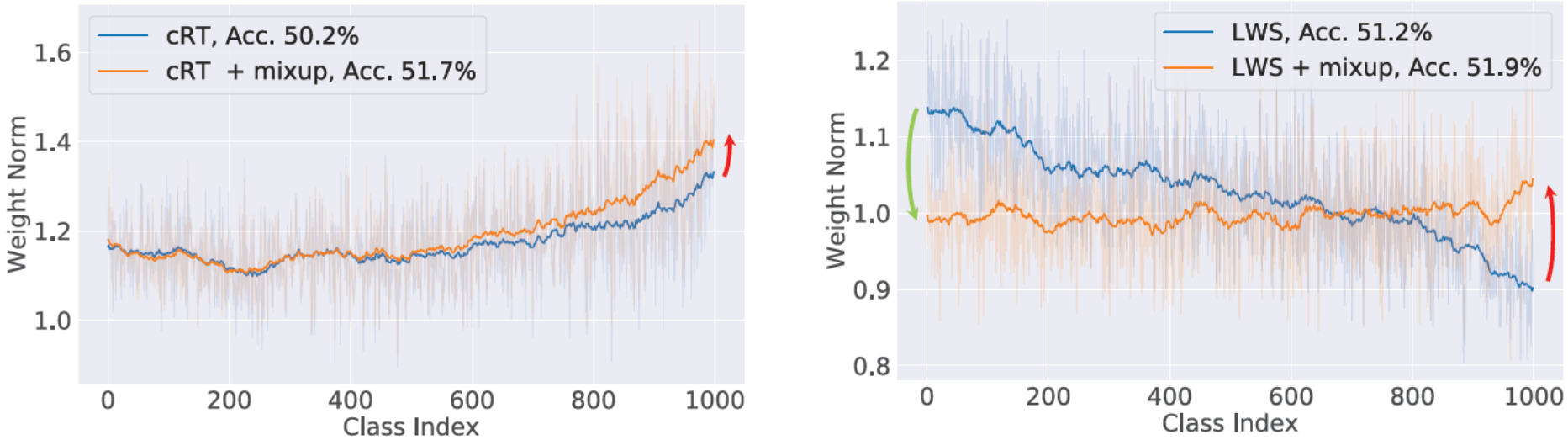


Figure 2: Classifier weight norms for the ImageNet-LT validation set where classes are sorted by descending values of N_j, where N_j denotes the number of training sample for Class-j. Left: weight norms of cRT with or without mixup. Right: weight norms of LWS with or without mixup. Light shade: true norm. Dark lines: smooth version. *Best viewed on screen.*

Methods-Label aware Smoothing

The cross-entropy loss after the softmax activation is

$$l(y, \boldsymbol{p}) = -\log(\boldsymbol{p}_y) = -\boldsymbol{w}_y^\top \boldsymbol{x} + \log(\sum \exp(\boldsymbol{w}_i^\top \boldsymbol{x}))$$

The optimal solution is $\boldsymbol{w}_y^{*\top} \boldsymbol{x} = \inf$ while other $\boldsymbol{w}_i^\top \boldsymbol{x}, i \neq y$ are small enough

After label aware smoothing

$$l(\boldsymbol{q}, \boldsymbol{p}) = -\sum_{i=1}^{K} \boldsymbol{q}_i \log \boldsymbol{p}_i,$$

$$\boldsymbol{q}_i = \begin{cases} 1 - \epsilon_y = 1 - f(N_y), & i = y, \\ \frac{\epsilon_y}{K-1} = \frac{f(N_y)}{K-1}, & \text{otherwise}, \end{cases}$$

ϵ_y is a small label smoothing factor for Class-y.

The optimal solution is $\boldsymbol{w}_i^{*\top} \boldsymbol{x} = \begin{cases} \log\left(\frac{(K-1)(1-\epsilon_y)}{\epsilon_y}\right) + c, & i = y, \\ c, & \text{otherwise} \end{cases}$

Methods-Label aware Smoothing

$f(N_y)$ 的选取?

- Concave form:

$$f(N_y) = \epsilon_K + (\epsilon_1 - \epsilon_K) \sin\left[\frac{\pi(N_y - N_K)}{2(N_1 - N_K)}\right];$$

- Linear form:

$$f(N_y) = \epsilon_K + (\epsilon_1 - \epsilon_K)\frac{N_y - N_K}{N_1 - N_K};$$

- Convex form:

$$f(N_y) = \epsilon_1 + (\epsilon_1 - \epsilon_K) \sin\left[\frac{3\pi}{2} + \frac{\pi(N_y - N_K)}{2(N_1 - N_K)}\right],$$

Methods-Label aware Smoothing

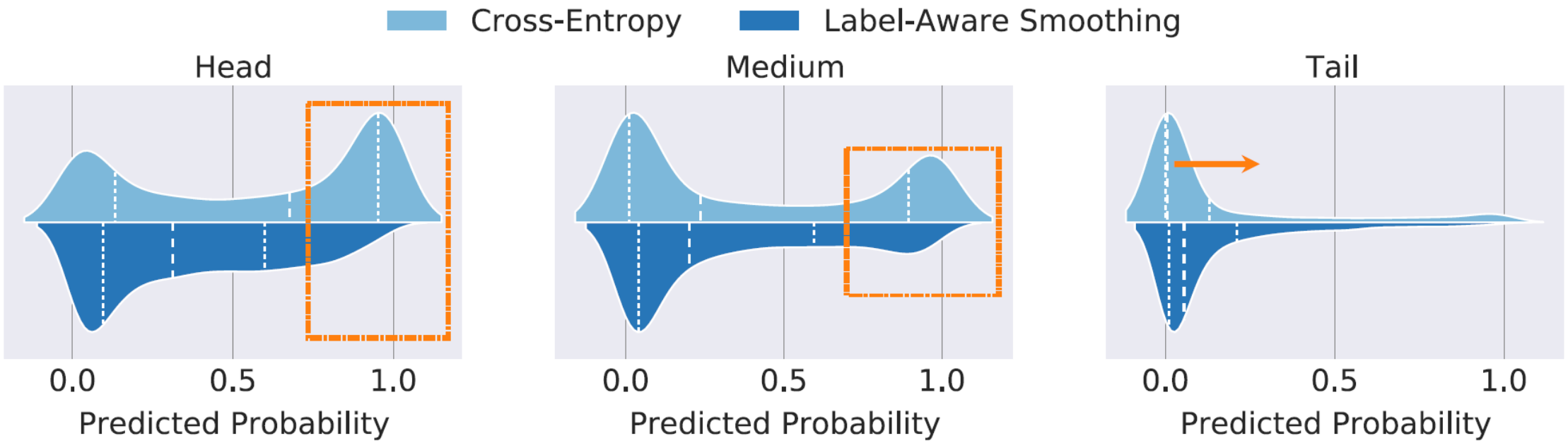


Figure 3: Violin plot of predicted probability distributions for different parts of the classes, head (100+ images per class), medium (20-100 images per class), and tail (less than 20 images per class) on CIFAR-100-LT with IF 100. The upper half part in light blue denotes "LWS + cross-entropy". The bottom half part in deep blue represents "LWS + label-aware smoothing".

To combine the advantages of cRT and LWS, we design the classifier framework in Stage-2 as $\boldsymbol{z} = \operatorname{diag}(\boldsymbol{s})\left(r\boldsymbol{W} + \Delta\boldsymbol{W}\right)^{\top}\boldsymbol{x}$

Methods-Shift Learning on Batch Normalization

Instance-balanced dataset $\mathcal{D}_{\mathrm{I}}$

class-balanced dataset $\mathcal{D}_{\mathrm{C}}$

$$\boldsymbol{x}_i \sim P_{\mathcal{D}_{\mathrm{I}}}(\boldsymbol{x}, y), \quad \boldsymbol{\mu}_{\mathrm{I}}^{(j)} = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{x}_i)^{(j)},$$

$$\boldsymbol{\sigma}_{\mathrm{I}}^{2\,(j)} = \frac{1}{m}\sum_{i=1}^{m}\left[g(\boldsymbol{x}_i)^{(j)} - \boldsymbol{\mu}_{\mathrm{I}}^{(j)}\right]^2,$$

$$\boldsymbol{x}_i \sim P_{\mathcal{D}_{\mathrm{C}}}(\boldsymbol{x}, y), \quad \boldsymbol{\mu}_{\mathrm{C}}^{(j)} = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{x}_i)^{(j)},$$

$$\boldsymbol{\sigma}_{\mathrm{C}}^{2\,(j)} = \frac{1}{m}\sum_{i=1}^{m}\left[g(\boldsymbol{x}_i)^{(j)} - \boldsymbol{\mu}_{\mathrm{C}}^{(j)}\right]^2.$$

Ablation study

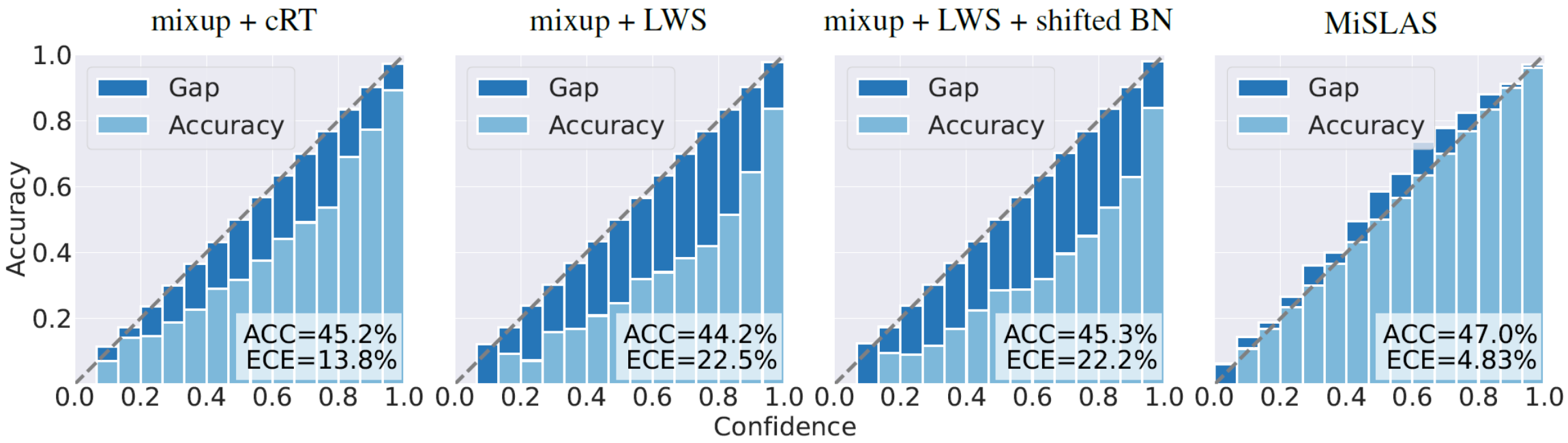


Figure 4: Reliability diagrams of ResNet-32 trained on CIFAR-100-LT with IF 100. From left to right: cRT with mixup, LWS with mixup, LWS with mixup and shifted BN, and MiSLAS (*complying with Fig. 1*).

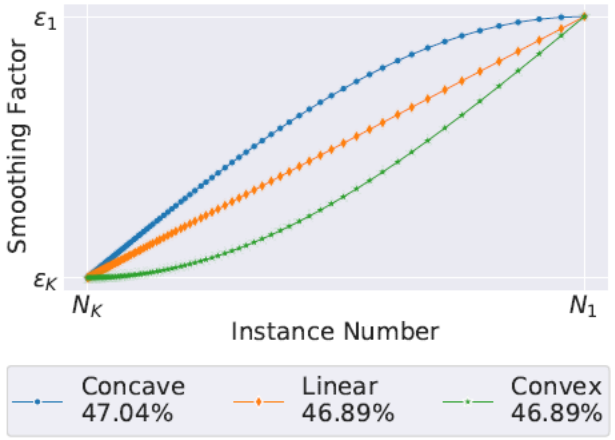
Ablation study

Comparing re-weighting with label-aware smoothing.

Method	100	50	10
CB-CE [7]	44.3 / 20.2	50.5 / 19.1	62.5 / 13.9
LAS	**47.0 / 4.83**	**52.3 / 2.25**	**63.2 / 1.73**

Table 2: Comparison in terms of test accuracy (%) / ECE (%) of label-aware smoothing (LAS) with re-weighting, class-balanced cross-entropy (CB-CE, [7]) in Stage-2. Both models are based on ResNet-32 and trained on CIFAR-100-LT with IF 100, 50, and 10.

How f(·) affects label-aware smoothing?



How ϵ_1 and ϵ_K affect label-aware smoothing?

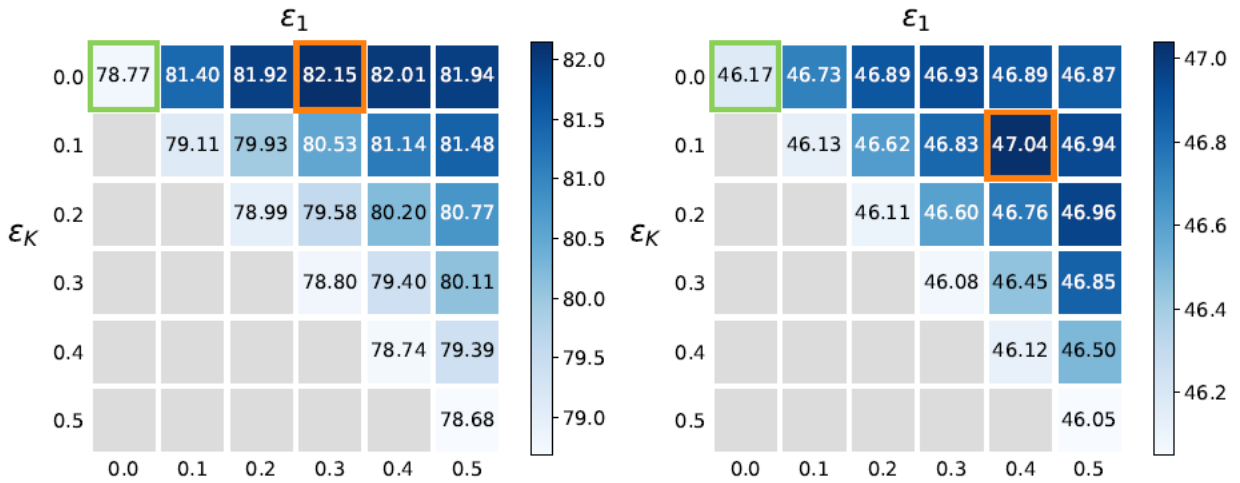


Figure 5: Ablation study of two hyperparameters ϵ_1 and ϵ_K in label-aware smoothing. Heat map visualization on CIFAR-10-LT with IF 100 (left) and on CIFAR-100-LT with IF 100 (right).

Experiment

Method	CIFAR-10-LT			CIFAR-100-LT		
	100	50	10	100	50	10
CE	70.4	74.8	86.4	38.4	43.9	55.8
mixup [37]	73.1	77.8	87.1	39.6	45.0	58.2
LDAM+DRW [4]	77.1	81.1	88.4	42.1	46.7	58.8
BBN(include mixup) [39]	79.9	82.2	88.4	42.6	47.1	59.2
Remix+DRW(300 epochs) [5]	79.8	-	89.1	46.8	-	61.3
cRT+mixup	79.1 / 10.6	84.2 / 6.89	89.8 / 3.92	45.1 / 13.8	50.9 / 10.8	62.1 / 6.83
LWS+mixup	76.3 / 15.6	82.6 / 11.0	89.6 / 5.41	44.2 / 22.5	50.7 / 19.2	62.3 / 13.4
MiSLAS	**82.1 / 3.70**	**85.7 / 2.17**	**90.0 / 1.20**	**47.0 / 4.83**	**52.3 / 2.25**	**63.2 / 1.73**

Table 4: Top-1 accuracy (%) / ECE (%) for ResNet-32 based models trained on CIFAR-10-LT and CIFAR-100-LT.

Method	ResNet-50
CE	44.6
CE+DRW [4]	48.5
Focal+DRW [18]	47.9
LDAM+DRW [4]	48.8
CRT+mixup	51.7 / 5.62
LWS+mixup	52.0 / 2.23
MiSLAS	**52.7 / 1.83**

(a) ImageNet-LT

Method	ResNet-50
CB-Focal [7]	61.1
LDAM+DRW [4]	68.0
BBN(include mixup) [39]	69.6
Remix+DRW [5]	70.5
cRT+mixup	70.2 / **1.79**
LWS+mixup(under-conf.)	70.9 / 9.41
MiSLAS(under-conf.)	**71.6** / 7.67

(b) iNaturalist 2018

Method	ResNet-152
Range Loss [38]	35.1
FSLwF [8]	34.9
OLTR [20]	35.9
OLTR+LFME [35]	36.2
cRT+mixup	38.3 / 12.4
LWS+mixup	39.7 / 11.7
MiSLAS	**40.4 / 3.59**

(c) Places-LT

Table 5: Top-1 accuracy (%) / ECE (%) on ImageNet-LT (left), iNaturalist 2018 (center) and Places-LT (right).

thanks