南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

When Noisy Labels Meet Long Tail Dilemmas: A Representation Calibration Method

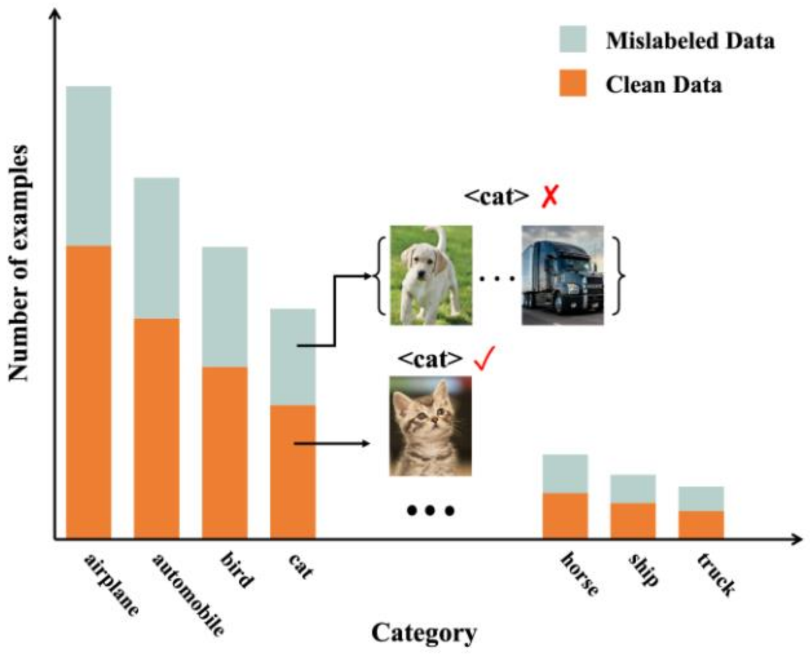
Manyi Zhang[1,*] Xuyang Zhao[2,*] Jun Yao[3] Chun Yuan[1,†] Weiran Huang[4,†]

[1]SIGS, Tsinghua University [2]Peking University [3]Huawei Noah's Ark Lab

[4]Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University

ICCV 2023

Introduction



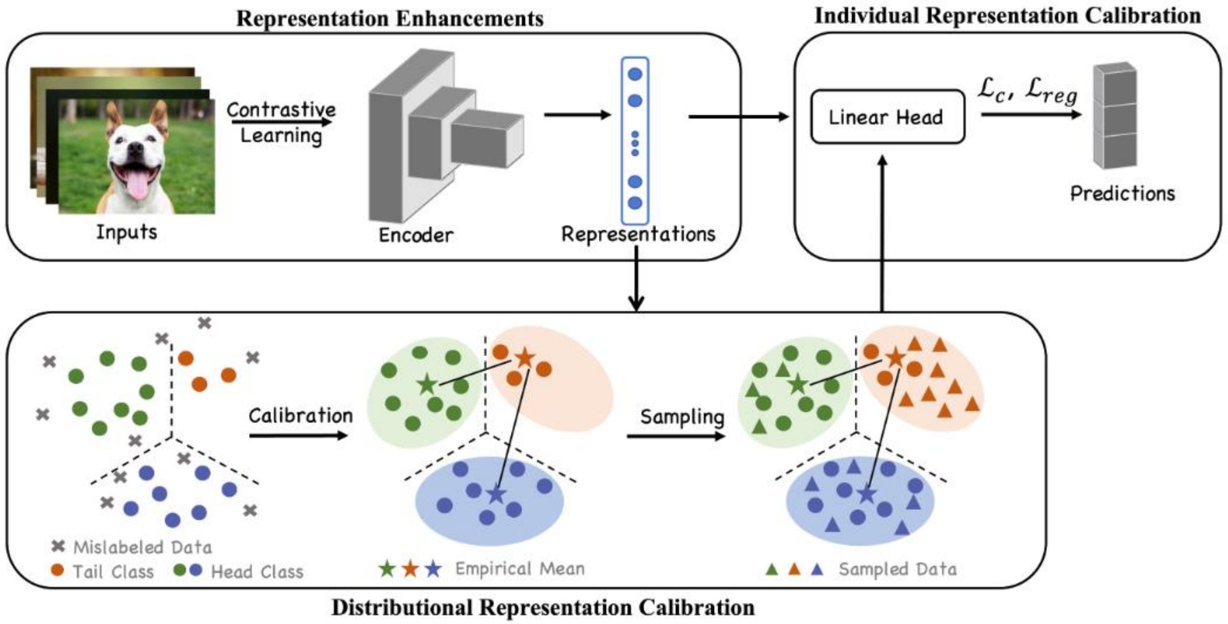
Learning with noisy labels:
1. memorization effect. E.g. Co-teaching
2. the noise transition matrix.

Learning with long-tailed data:
1. re-sampling and re-weighting techniques.

Learning with noisy label on long-tailed data:
1. distinguish mislabeled data from the data of tail classes for follow-up procedures. E.g. RoLT

2. reduce the side-effects of mislabeled data and long-tailed data in a unified way, relying on strong assumptions. E.g. HAR-DRW

Introduction

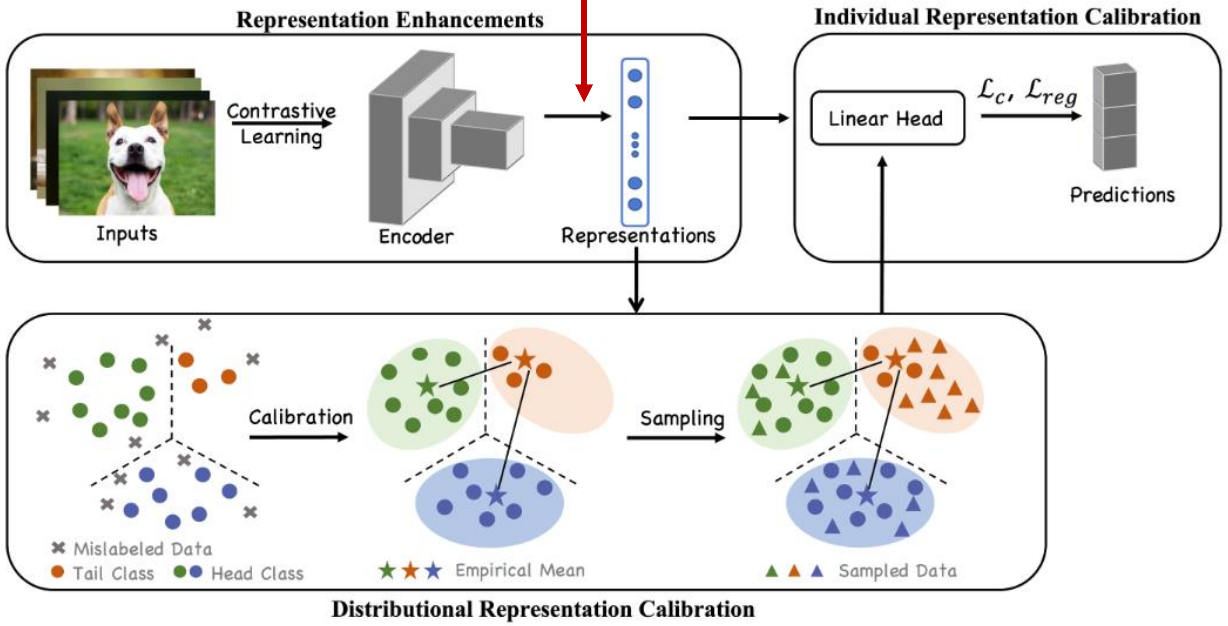


Step1: contrastive learning to achieve representations for all training instances.

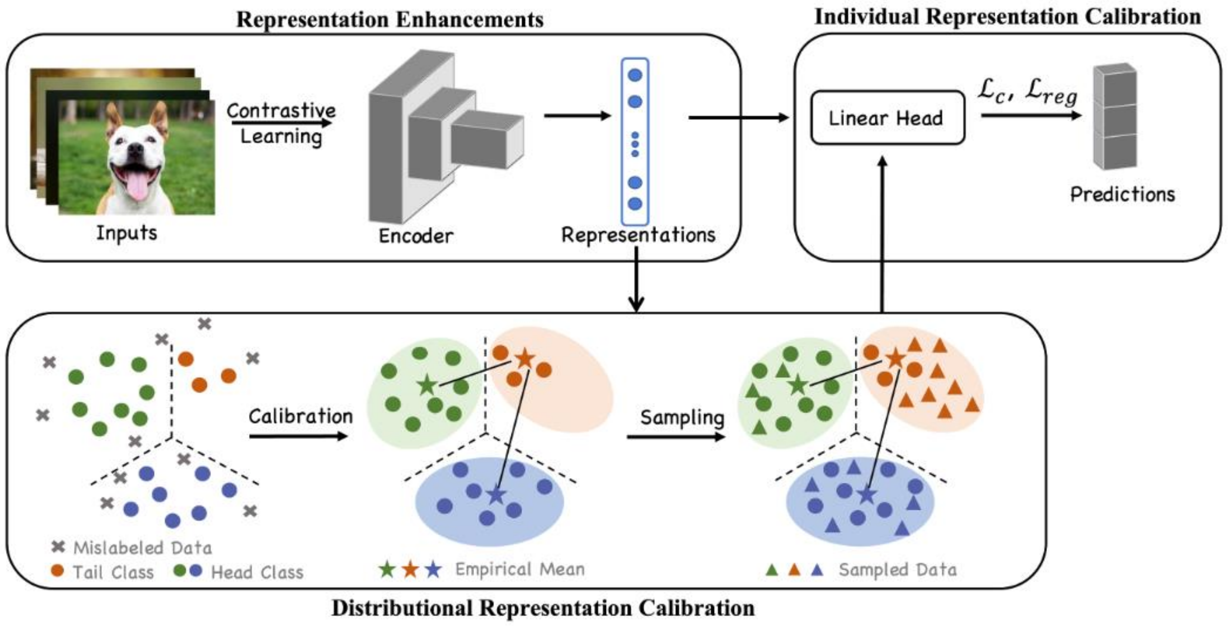
Step2: two representation calibration strategies are performed: distributional and individual representation calibrations.

Method/Enhancing Representations Through Contrastive Learning

$$\mathcal{L}_{con}(\boldsymbol{x}_i) = -\log \frac{\exp(\hat{\boldsymbol{z}}_i^q \cdot \hat{\boldsymbol{z}}_i^k / \tau)}{\sum_{\hat{\boldsymbol{z}}^{k'} \in \mathcal{A}} \exp(\hat{\boldsymbol{z}}_i^q \cdot \hat{\boldsymbol{z}}^{k'} / \tau)},$$



Method/ Distributional Representation Calibration



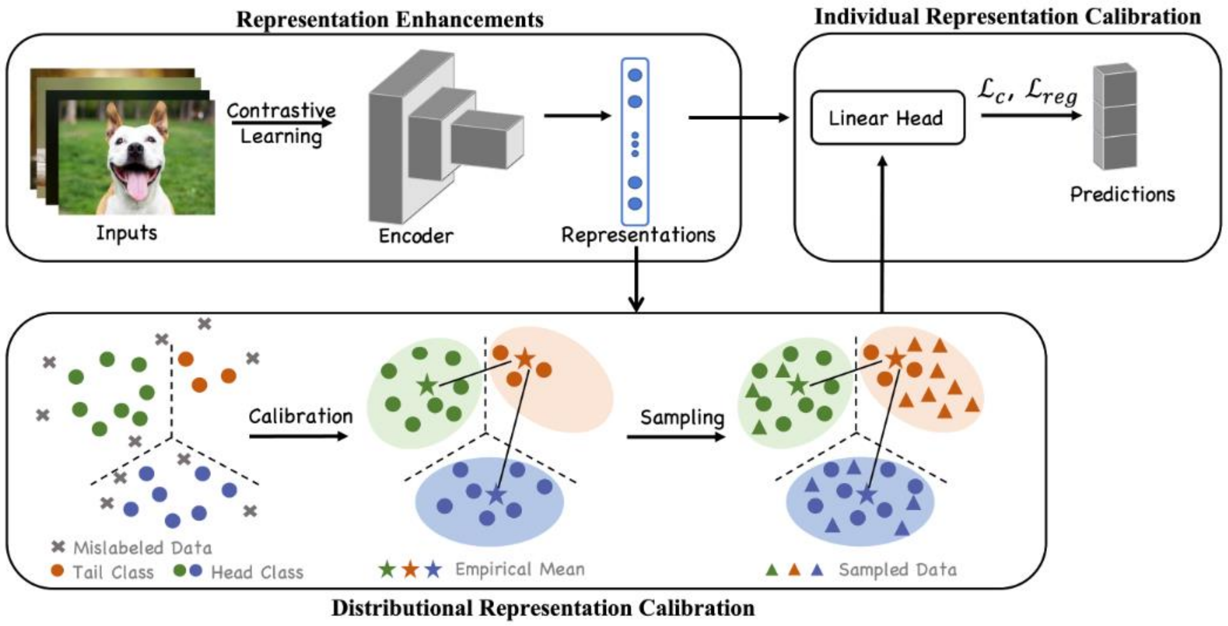
Step1: given the learned representations z.
Step2: employ LOF and remove outliers.
Step3: estimate the multivariate Gaussian distribution

$$\hat{\boldsymbol{\mu}}_k = \sum_{\{i|(\boldsymbol{z}_i,\tilde{y}_i)\in\tilde{\mathcal{S}}'_k\}} \frac{\boldsymbol{z}_i}{|\tilde{\mathcal{S}}'_k|},$$

$$\hat{\boldsymbol{\Sigma}}_k = \sum_{\{i|(\boldsymbol{z}_i,\tilde{y}_i)\in\tilde{\mathcal{S}}'_k\}} \frac{(\boldsymbol{z}_i-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{z}_i-\hat{\boldsymbol{\mu}}_k)^\top}{|\tilde{\mathcal{S}}'_k|-1},$$

Purpose: recover representation distributions.
Assumption: multivariate Gaussian distribution.

Method/Distributional Representation Calibration



$$\mathcal{B}_k = \left\{-||\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_k||^2 \mid i \in \mathcal{G}_h\right\},$$
$$\mathcal{C}_k^q = \left\{i \mid -||\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_k||^2 \in \mathtt{topq}(\mathcal{B}_k)\right\}.$$

$$\omega_c^k = \frac{n_c||\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}}_k||^2}{\sum_{j\in\mathcal{C}_k^q} n_j||\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k||^2},$$
$$\hat{\boldsymbol{\mu}}_k' = \gamma \sum_{c\in\mathcal{C}_k^q} \omega_c^k \hat{\boldsymbol{\mu}}_c + (1-\gamma)\hat{\boldsymbol{\mu}}_k,$$
$$\hat{\boldsymbol{\Sigma}}_k' = \gamma \sum_{c\in\mathcal{C}_k^q} \omega_c^k \hat{\boldsymbol{\Sigma}}_c + (1-\gamma)\hat{\boldsymbol{\Sigma}}_k + \alpha\mathbf{1},$$

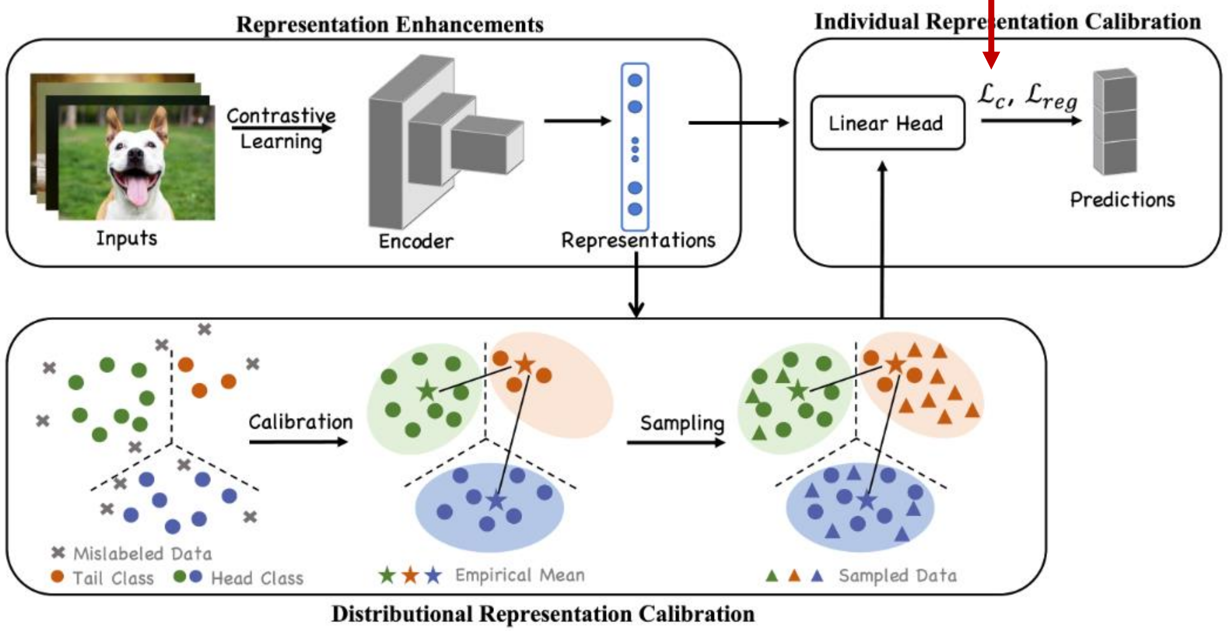
Problem: the data of tail classes may not be enough to estimate.
Motivation:' Free Lunch for Few-shot Learning: Distribution Calibration'
(similar classes having similar means and covariance on representations)

Method/Individual Representation Calibration

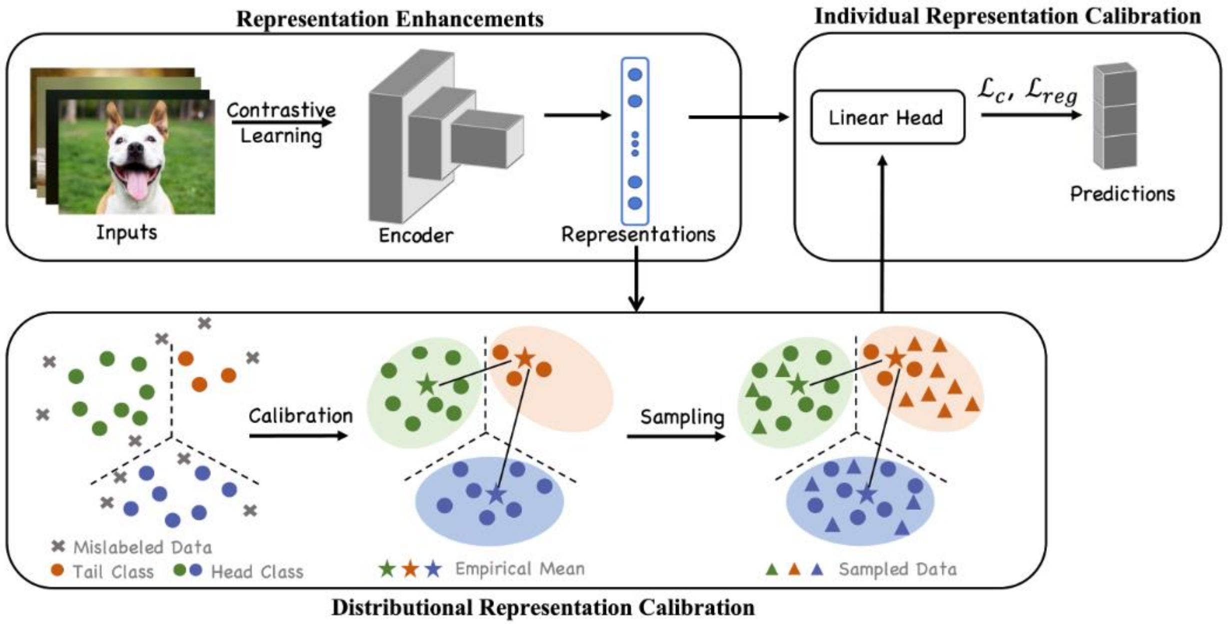
Purpose: restrict the distance.

$$\mathcal{L}_{reg}(\boldsymbol{x}) = ||\boldsymbol{z} - \boldsymbol{z}^0||^2 = ||f(\boldsymbol{x}) - \boldsymbol{z}^0||^2.$$

$$\mathcal{L} = \mathcal{L}_c + \beta\mathcal{L}_{reg},$$



Method



Algorithm 1 Algorithm of the proposed method RCAL

Require: the training dataset $\tilde{\mathcal{S}} = \{(\boldsymbol{x}_i, \tilde{y}_i\})_{i=1}^{n}$, regularization strength β, scalar temperature τ, confidence weight γ, the pre-training epochs T_p, max epochs T_m.

for $t = 1, ..., T_p$ **do**
  Pre-train the encoder network f with MoCo [20].
end for
Extract deep representations of instances with $\boldsymbol{z} = f(\boldsymbol{x})$.
for $c = 1, ..., K$ **do**
  Perform the LOF algorithm for the c-th class and obtain preserved examples $\tilde{\mathcal{S}}'_c$.
  Build the multivariate Gaussian distribution $\mathcal{N}(f(\boldsymbol{x})|\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c)$ for c-th class using $\tilde{\mathcal{S}}'_c$.
end for
Calibrate the multivariate Gaussian distributions of tail classes with the statistics of head classes.
Sample data points from achieved multivariate Gaussian distributions of all classes.
for $t = T_p + 1, ..., T_m$ **do**
  Add distance constraints between learned representations and representations brought by contrastive learning.
  Adopt the mixup technology to original examples.
  Train the encoder f and the linear head h simultaneously on the training dataset and sample data points with the training loss in Eq. (2).
end for
return The robust classifier $h(f(\boldsymbol{x}))$ for testing.

Experiments/Simulated noisy and class-imbalanced datasets.

Dataset	Imbalance Ratio	10					100				
CIFAR-10	Noise Rate	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
	ERM	80.41	75.61	71.94	70.13	63.25	64.41	62.17	52.94	48.11	38.71
	LDAM	84.59	82.37	77.48	71.41	60.30	71.46	66.26	58.34	46.64	36.66
	LDAM-DRW	85.94	83.73	80.20	74.87	67.93	76.58	72.28	66.68	57.51	43.23
	CRT	80.22	76.15	74.17	70.05	64.15	61.54	59.52	54.05	50.12	36.73
	NCM	82.33	74.73	74.76	68.43	64.82	68.09	66.25	60.91	55.47	42.61
	MiSLAS	87.58	85.21	83.39	76.16	72.46	75.62	71.48	67.90	62.04	54.54
	Co-teaching	80.30	78.54	68.71	57.10	46.77	55.58	50.29	38.01	30.75	22.85
	CDR	81.68	78.09	73.86	68.12	62.24	60.47	55.34	46.32	42.51	32.44
	Sel-CL+	86.47	85.11	84.41	80.35	77.27	72.31	71.02	65.70	61.37	56.21
	HAR-DRW	84.09	82.43	80.41	77.43	67.39	70.81	67.88	48.59	54.23	42.80
	RoLT	85.68	85.43	83.50	80.92	78.96	73.02	71.20	66.53	57.86	48.98
	RoLT-DRW	86.24	85.49	84.11	81.99	80.05	76.22	74.92	71.08	63.61	55.06
	RCAL (Ours)	88.09	86.46	84.58	83.43	80.80	78.60	75.81	72.76	69.78	65.05

Dataset	Imbalance Ratio	10					100				
CIFAR-100	Noise Rate	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
	ERM	48.54	43.27	37.43	32.94	26.24	31.81	26.21	21.79	17.91	14.23
	LDAM	51.77	48.14	43.27	36.66	29.62	34.77	29.70	25.04	19.72	14.19
	LDAM-DRW	54.01	50.44	45.11	39.35	32.24	37.24	32.27	27.55	21.22	15.21
	CRT	49.13	42.56	37.80	32.18	25.55	32.25	26.31	21.48	20.62	16.01
	NCM	50.76	45.15	41.31	35.41	29.34	34.89	29.45	24.74	21.84	16.77
	MiSLAS	57.72	53.67	50.04	46.05	40.63	41.02	37.40	32.84	26.95	21.84
	Co-teaching	45.61	41.33	36.14	32.08	25.33	30.55	25.67	22.01	16.20	13.45
	CDR	47.02	40.64	35.37	30.93	24.91	27.20	25.46	21.98	17.33	13.64
	Sel-CL+	55.68	53.52	50.92	47.57	44.86	37.45	36.79	35.09	31.96	28.59
	HAR-DRW	51.04	46.24	41.23	37.35	31.30	33.21	26.29	22.57	18.98	14.78
	RoLT	54.11	51.00	47.42	44.63	38.64	35.21	30.97	27.60	24.73	20.14
	RoLT-DRW	55.37	52.41	49.31	46.34	40.88	37.60	32.68	30.22	26.58	21.05
	RCAL (Ours)	57.50	54.85	51.66	48.91	44.36	41.68	39.85	36.57	33.36	30.26

Methods for long-tailed data

Methods for noisy labels

Methods for both

Experiments/Results on Real-world Noisy and Imbalanced Datasets

Table 2: Top1 and Top5 test accuracy on Webvision and ImageNet validation sets. Partial numerical results come from [5, 61]. The best results are in **bold**.

Train	WebVision-50			
Test	WebVision		ILSVRC12	
Method	Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)
ERM	62.5	80.8	58.5	81.8
Co-teaching [18]	63.58	85.20	61.48	84.70
INCV [7]	65.24	85.34	61.60	84.98
MentorNet [25]	63.00	81.40	57.80	79.92
CDR [63]	-	-	61.85	-
HAR [5]	75.5	90.7	70.3	90.0
RoLT+ [61]	77.64	92.44	74.64	92.48
RCAL (Ours)	76.24	92.83	73.60	93.16
RCAL+ (Ours)	**79.56**	**93.36**	**76.32**	**93.68**

combine semi-supervised learning

Table 3: Test accuracy on the Clothing1M test dataset. Partial numerical results come from [78]. The best results are in **bold**.

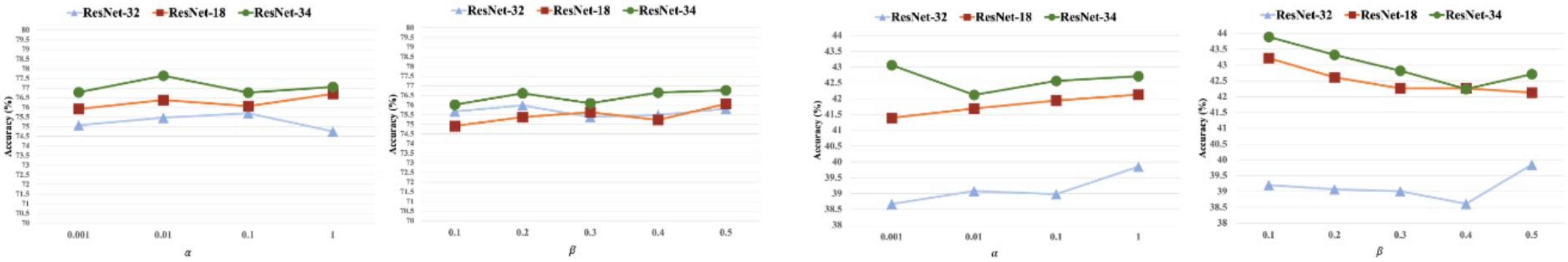
Method	Top1 (%)	Method	Top1 (%)
ERM	68.94	Co-teaching [18]	67.94
MentorNet [25]	67.25	CDR [63]	68.25
Forward [48]	69.84	D2L [45]	69.74
Joint [53]	72.23	GCE [79]	69.75
Pencil [73]	73.49	LRT [82]	71.74
SL [58]	71.02	MLNT [33]	73.47
PLC [78]	74.02	DivideMix [32]	74.76
ELR+ [42]	74.81	**RCAL+ (Ours)**	**74.97**

Experiments/Ablation Study

Table 7: Ablation study results of test accuracy (%) on simulated CIFAR-10 and CIFAR-100. We report the mean. The best results are in **bold**. In the following, “CL” means unsupervised contrastive learning. “DC” means distributional calibration. “REG” means individual calibration by restricting the distance between subsequently learned representations and the representations brought by unsupervised contrastive learning.

Dataset	CIFAR-10				CIFAR-100			
Imbalance Ratio	10		100		10		100	
Noise Rate	0.2	0.4	0.2	0.4	0.2	0.4	0.2	0.4
RCAL	**86.46**	**83.43**	**75.81**	**69.78**	**54.85**	**48.91**	**39.85**	**33.36**
RCAL w/o Mixup	84.08	79.27	72.47	64.83	51.22	45.53	36.78	30.85
RCAL w/o Mixup, REG	83.23	78.12	67.49	58.27	48.74	42.15	34 31	27.14
RCAL w/o Mixup, REG, DC	80.40	74.37	64.02	54.61	47.01	40.85	32.27	25.42
RCAL w/o Mixup, REG, DC, CL	75.61	70.13	62.17	48.11	43.27	32.94	26.21	17.91

Experiments/Ablation Study



(a) (b) (c) (d)

Table 6: Test accuracy (%) of many/medium/few classes on CIFAR-10, where the noise rate and imbalance ratio are 0.5 and 10.

Method	Many	Medium	Few	Overall
ERM	82.71	55.31	57.22	63.25
MiSLAS	67.16	69.52	81.66	72.46
RCAL (Ours)	84.10	84.13	73.98	80.80

2021

Thank you