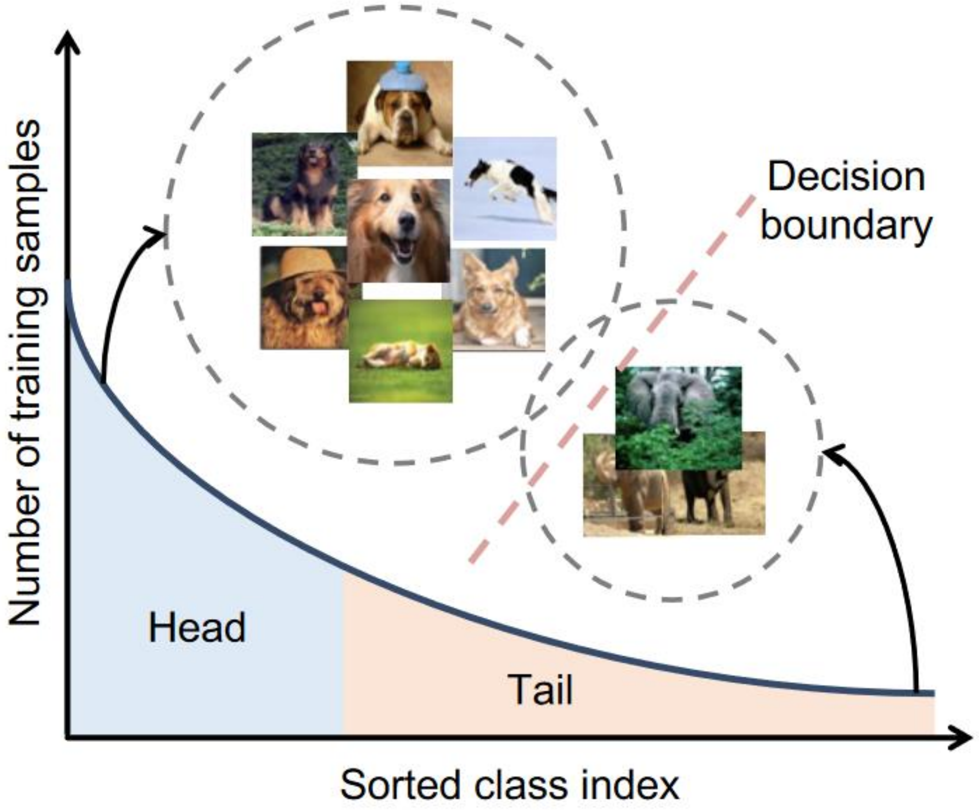
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Open-Sampling: Exploring Out-of-Distribution Data for Re-balancing Long-tailed Datasets

Hongxin Wei [1] Lue Tao [2] Renchunzi Xie [1] Lei Feng [3] Bo An [1]

ICML 2022

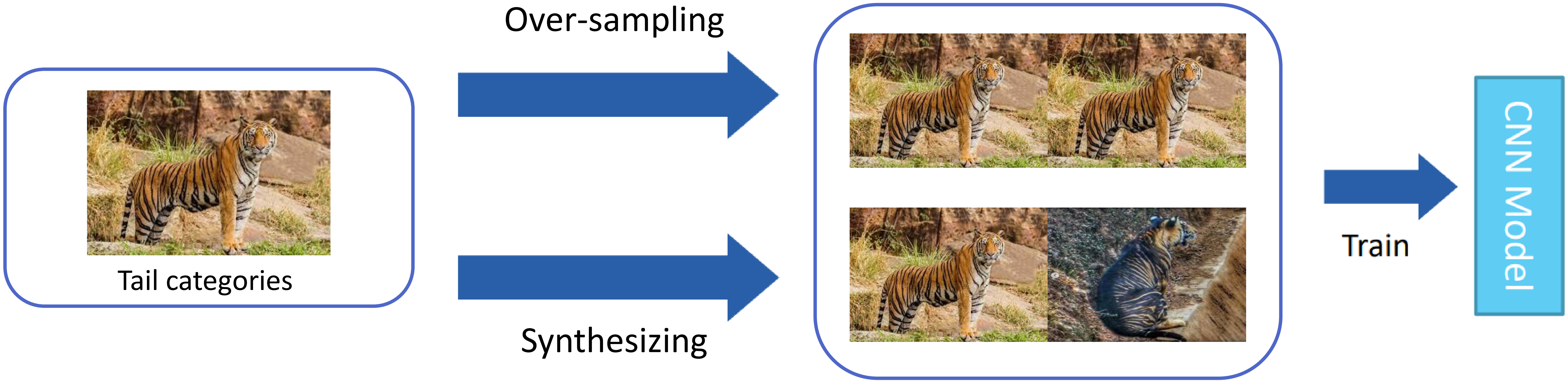
Long-Tailed Learning



- **Long-tailed Learning**
  - **Class Re-balancing**
    - Re-sampling
    - Cost-sensitive Learning
    - Logit Adjustment
  - **Information Augmentation**
    - Transfer Learning
    - Data Augmentation
  - **Module Improvement**
    - Representation Learning
    - Classifier Design
    - Decoupled Training
    - Ensemble Learning

Motivation

- Re-Sampling: class-balanced sampling
  - – Over-sampling for the tail categories.
  - – Synthesizing samples for the tail categories.
- Drawbacks
  - – Overfitting to the tail.
  - – Noise in the synthesized samples.



Can open-set data be used for re-balancing the class priors of the training dataset?

Theoretical Motivation

- Bayes classifier:

$$y^* = \arg\max_{y \in \mathcal{Y}} P(y|\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} P(\boldsymbol{x}|y)P(y)$$

- For long-tailed learning:

$$P_{\mathrm{s}}(Y = i) \neq P_{\mathrm{t}}(Y = i)$$

Theorem 2.1. *Assume that $P_{\mathrm{out}}(Y)$ is the discrete uniform distribution over the label space $\mathcal{Y}$. Let the augmented dataset be $\mathcal{D}_{\mathrm{mix}} = \mathcal{D}_{\mathrm{train}} \cup \mathcal{D}_{\mathrm{out}}$, and $P_{\mathrm{mix}}(X, Y)$ be the underlying data distribution of $\mathcal{D}_{\mathrm{mix}}$, then we have*

$$\arg\max_{y \in \mathcal{Y}} P_{\mathrm{mix}}(\boldsymbol{x}|y)P_{\mathrm{mix}}(y) = \arg\max_{y \in \mathcal{Y}} P_{\mathrm{s}}(\boldsymbol{x}|y)P_{\mathrm{s}}(y).$$

$$P_{\mathrm{mix}}(\boldsymbol{x}, y) = \frac{N}{M+N} P_{\mathrm{s}}(\boldsymbol{x}, y) + \frac{M}{M+N} P_{\mathrm{out}}(\boldsymbol{x}, y)$$

$$\Downarrow$$

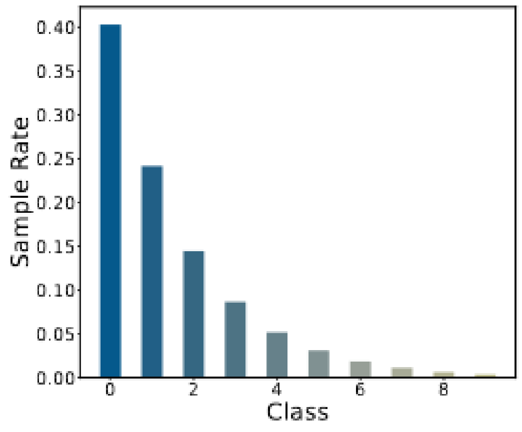
$$\begin{aligned}
& P_{\mathrm{mix}}(\boldsymbol{x}|y)P_{\mathrm{mix}}(y) \\
= & \frac{N}{M+N} P_{\mathrm{s}}(\boldsymbol{x}|y)P_{\mathrm{s}}(y) + \frac{M}{M+N} P_{\mathrm{out}}(\boldsymbol{x}|y)P_{\mathrm{out}}(y) \\
= & \frac{N}{M+N} P_{\mathrm{s}}(\boldsymbol{x}|y)P_{\mathrm{s}}(y) + \frac{M}{M+N} P_{\mathrm{out}}(\boldsymbol{x})P_{\mathrm{out}}(y) \\
= & \frac{N}{M+N} P_{\mathrm{s}}(\boldsymbol{x}|y)P_{\mathrm{s}}(\boldsymbol{x}, y) + \frac{1}{K} \cdot \frac{M}{M+N} P_{\mathrm{out}}(\boldsymbol{x})
\end{aligned}$$

$$\Downarrow$$

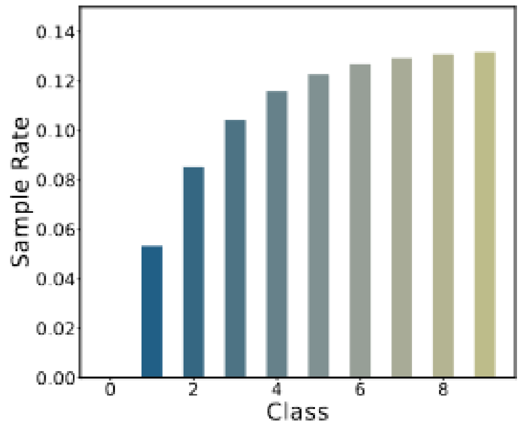
$$\begin{aligned}
& \arg\max_{y \in \mathcal{Y}} P_{\mathrm{mix}}(\boldsymbol{x}|y)P_{\mathrm{mix}}(y) \\
= & \arg\max_{y \in \mathcal{Y}} \left\{ \frac{N}{M+N} P_{\mathrm{s}}(\boldsymbol{x}|y)P_{\mathrm{s}}(\boldsymbol{x}, y) + \boxed{\frac{M/K}{M+N} P_{\mathrm{out}}(\boldsymbol{x})} \right\} \\
= & \arg\max_{y \in \mathcal{Y}} \frac{N}{M+N} P_{\mathrm{s}}(\boldsymbol{x}|y)P_{\mathrm{s}}(\boldsymbol{x}, y) \\
= & \arg\max_{y \in \mathcal{Y}} P_{\mathrm{s}}(\boldsymbol{x}|y)P_{\mathrm{s}}(\boldsymbol{x}, y)
\end{aligned}$$

Open-sampling

Definition 2.2 (Complementary Distribution). *Complementary Distribution (CD) is a label distribution for the auxiliary dataset to re-balance the class priors of the original dataset. In particular, Minimum Complementary Distribution (MCD) is the complementary distribution that requires the smallest number of auxiliary instances to re-balance the original training set.*



(a) Original



(c) MCD

Complementary Sampling Rate

- Denote CD as Γ, MCD as Γ^m, Γ_j for class j
- $\Gamma_j = \frac{\alpha - \beta_j}{K \cdot \alpha - 1}$, where $\beta_j = \frac{n_j}{\sum_{i=1}^{K} n_i}$
- $\sum_{i=1}^{K} \Gamma_i = 1$
- $\alpha \in \mathbb{R}^+ \geq \max_j \beta_j$
- $\Gamma = \Gamma_m$, if $\alpha = \max_j \beta_j$ ⇨ MCD
- $\Gamma_j \to \frac{1}{K}$ as $\alpha \to \infty$ ⇨ Uniform distribution
- Default $\alpha = \max_j \beta_j + \min_j \beta_j$

Open-sampling

- Training objective:

$$\mathcal{L} = \mathbb{E}_{P_{\text{mix}}(X,Y)}[\ell(f(\boldsymbol{x};\boldsymbol{\theta}, y)] = \mathbb{E}_{P_{\text{mix}}(X,Y)}[-\mathbf{e}^{y} \log f(\boldsymbol{x};\boldsymbol{\theta})]$$

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\widetilde{\boldsymbol{x}} \sim P_{\text{out}}(X)} \left[\omega_{\widetilde{y}} \cdot \ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right]$$

$$\Downarrow$$

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{((\boldsymbol{x},y)\sim P_{\text{s}}(X,Y))} \left[\ell\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)\right] + \eta \cdot \mathbb{E}_{(\widetilde{\boldsymbol{x}})\sim P_{\text{out}}(X)} \left[\omega_{\widetilde{y}} \cdot \ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right]$$

Algorithm 1 Open-sampling

Require: Training dataset $\mathcal{D}_{\text{train}}$. Open-set auxiliary dataset $\mathcal{D}_{\text{out}}^{(x)}$;

1: **for** each iteration **do**
2: Sample a mini-batch of original training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=0}^{n}$ from $\mathcal{D}_{\text{train}}$;
3: Sample a mini-batch of open-set instances $\{\widetilde{\boldsymbol{x}}_i\}_{i=0}^{m}$ from $\mathcal{D}_{\text{out}}^{(x)}$;
4: Generate random noisy label $\widetilde{y}_i \sim \Gamma$ for each open-set instance $\widetilde{\boldsymbol{x}}_i$;
5: Perform gradient descent on f with $\mathcal{L}_{\text{total}}$ from Equation (2);
6: **end for**

Experiment

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Table 1. Test accuracy (%) of ResNet-32 on long-tailed CIFAR-10 and CIFAR-100 with various imbalance ratios. "†" indicates the reported results from (Kim et al., 2020). The bold indicates the improved results by integrating our regularization.

Dataset	Long-tailed CIFAR-10			Long-tailed CIFAR-100		
Imbalance Ratio	100	50	10	100	50	10
Standard	71.61 ± 0.21	77.30 ± 0.13	86.74 ± 0.41	37.59 ± 0.19	43.20 ± 0.30	56.44 ± 0.12
SMOTE †	71.50 ± 0.57	-	85.70 ± 0.25	34.00 ± 0.33	-	53.80 ± 0.93
CB-RW	72.57 ± 1.30	78.19 ± 1.79	87.18 ± 0.95	38.11 ± 0.78	43.26 ± 0.87	56.40 ± 0.40
CB-Focal	70.91 ± 0.39	77.71 ± 0.57	86.89 ± 0.21	37.84 ± 0.80	42.96 ± 0.77	56.09 ± 0.15
Ours	$\mathbf{77.62 \pm 0.28}$	$\mathbf{81.76 \pm 0.51}$	$\mathbf{89.38 \pm 0.46}$	$\mathbf{40.26 \pm 0.65}$	$\mathbf{44.77 \pm 0.25}$	$\mathbf{58.09 \pm 0.29}$
LDAM-RW	74.21 ± 0.61	78.86 ± 0.65	86.44 ± 0.78	29.02 ± 0.34	36.41 ± 0.84	54.23 ± 0.72
+ Ours	$\mathbf{75.19 \pm 0.34}$	$\mathbf{79.76 \pm 0.44}$	$\mathbf{87.28 \pm 0.61}$	$\mathbf{35.85 \pm 0.62}$	$\mathbf{42.18 \pm 0.82}$	$\mathbf{55.48 \pm 0.59}$
LDAM-DRW	78.08 ± 0.38	81.88 ± 0.44	87.49 ± 0.18	42.84 ± 0.25	47.13 ± 0.28	57.18 ± 0.47
+ Ours	$\mathbf{79.82 \pm 0.31}$	$\mathbf{82.22 \pm 0.45}$	$\mathbf{87.83 \pm 0.38}$	$\mathbf{44.07 \pm 0.75}$	$\mathbf{47.5 \pm 0.24}$	$\mathbf{57.43 \pm 0.31}$
Balanced Softmax	78.03 ± 0.28	81.63 ± 0.39	88.10 ± 0.32	42.11 ± 0.70	46.79 ± 0.24	58.06 ± 0.40
+ Ours	$\mathbf{79.05 \pm 0.20}$	$\mathbf{82.76 \pm 0.52}$	$\mathbf{88.89 \pm 0.21}$	$\mathbf{42.86 \pm 0.27}$	$\mathbf{47.28 \pm 0.58}$	$\mathbf{58.80 \pm 0.72}$
SSP	74.58 ± 0.16	79.20 ± 0.43	88.50 ± 0.24	43.00 ± 0.51	47.04 ± 0.60	59.08 ± 0.46
+ Ours	$\mathbf{79.38 \pm 0.65}$	$\mathbf{82.18 \pm 0.33}$	$\mathbf{88.80 \pm 0.43}$	$\mathbf{43.57 \pm 0.29}$	$\mathbf{48.66 \pm 0.57}$	$\mathbf{59.78 \pm 0.91}$

CB： $(1-\beta^{n_i})/(1-\beta)$

Experiment

Table 3. Classification accuracy (%) on CelebA-5 with ResNet-32. "†" indicates the reported results are obtained from (Kim et al., 2020). The shadow indicates the improved results.

Method	Accuracy	Method	Accuracy	Method	Accuracy
Standard	72.7	M2m †	75.6	LDAM-DRW	74.5
SMOTE †	72.8	**Ours**	**76.8**	**LDAM-DRW + Ours**	**76.9**
CB-RW	73.6	LDAM-RW	73.1	Balanced Softmax	76.4
CB-Focal	74.2	**LDAM + Ours**	**75.8**	**Balanced Softmax + Ours**	**78.6**

Table 4. Top-1 accuracy (%) on Places-LT with an ImageNet pre-trained ResNet-152. Baseline results are taken from original papers. "DT" indicates decoupled training.

Method	Many	Medium	Few	Overall
Lifted Loss (Oh Song et al., 2016)	41.1	35.4	24.0	35.2
Focal Loss	41.1	34.8	22.4	34.6
Range Loss (Zhang et al., 2017)	41.1	35.4	23.2	35.1
OLTR (Liu et al., 2019b)	44.7	37.0	25.3	35.9
cRT* (Kang et al., 2020)	42.1	37.6	24.9	36.7
LWS* (Kang et al., 2020)	40.6	39.1	28.6	37.6
Ours*	42.9	**39.3**	26.8	**38.2**

Experiment

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

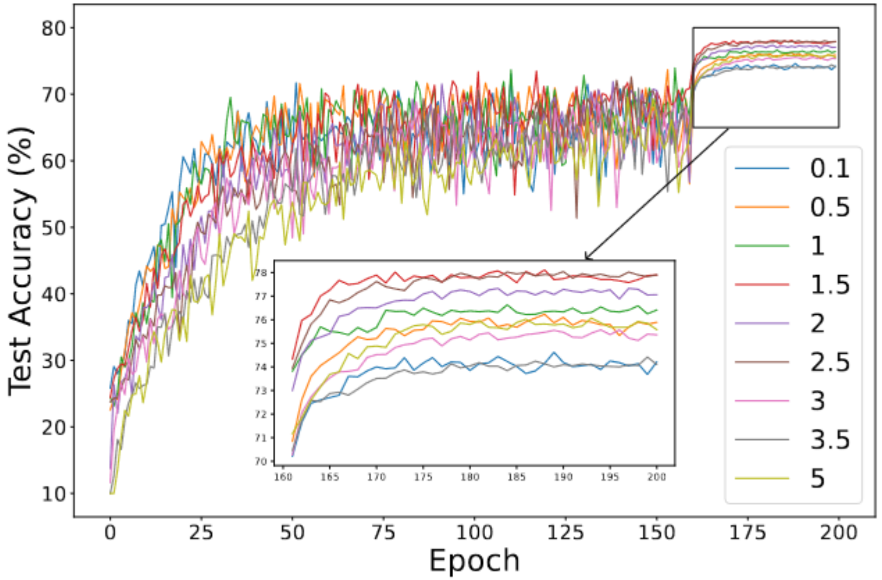
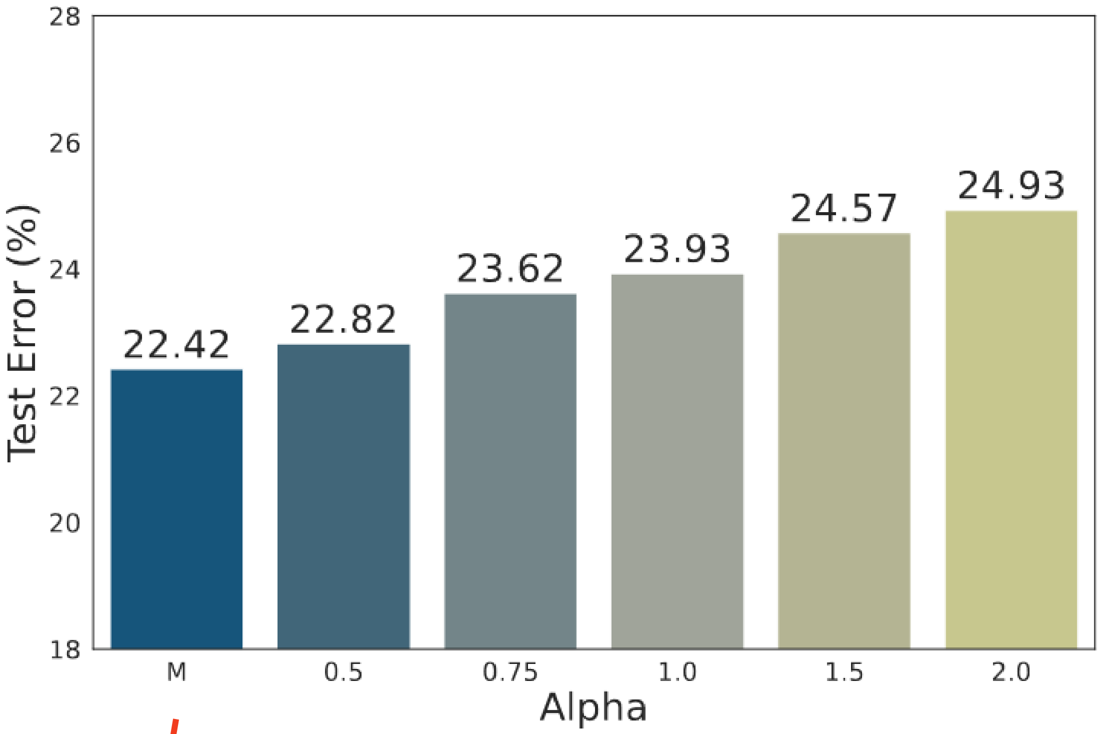
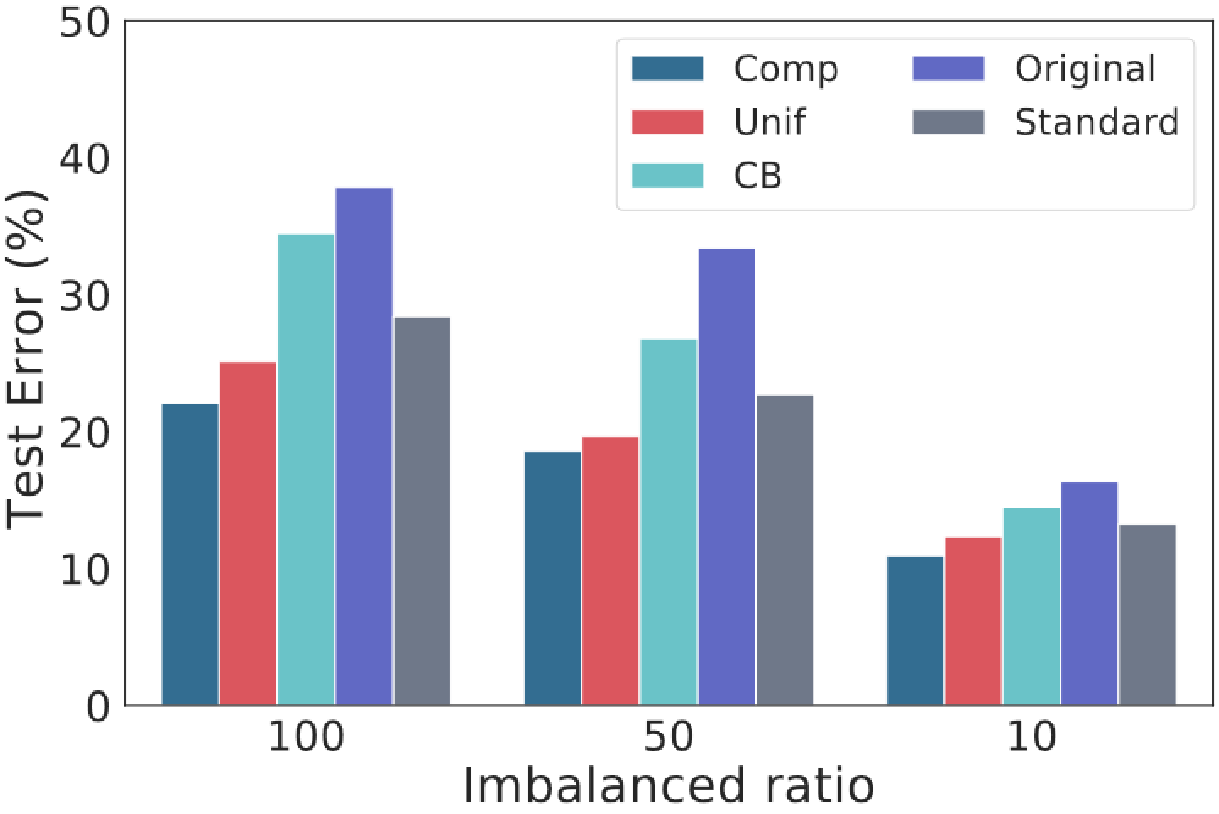


Figure 3. Results of sensitivity analysis on long-tailed CIFAR-10 with various values for η.

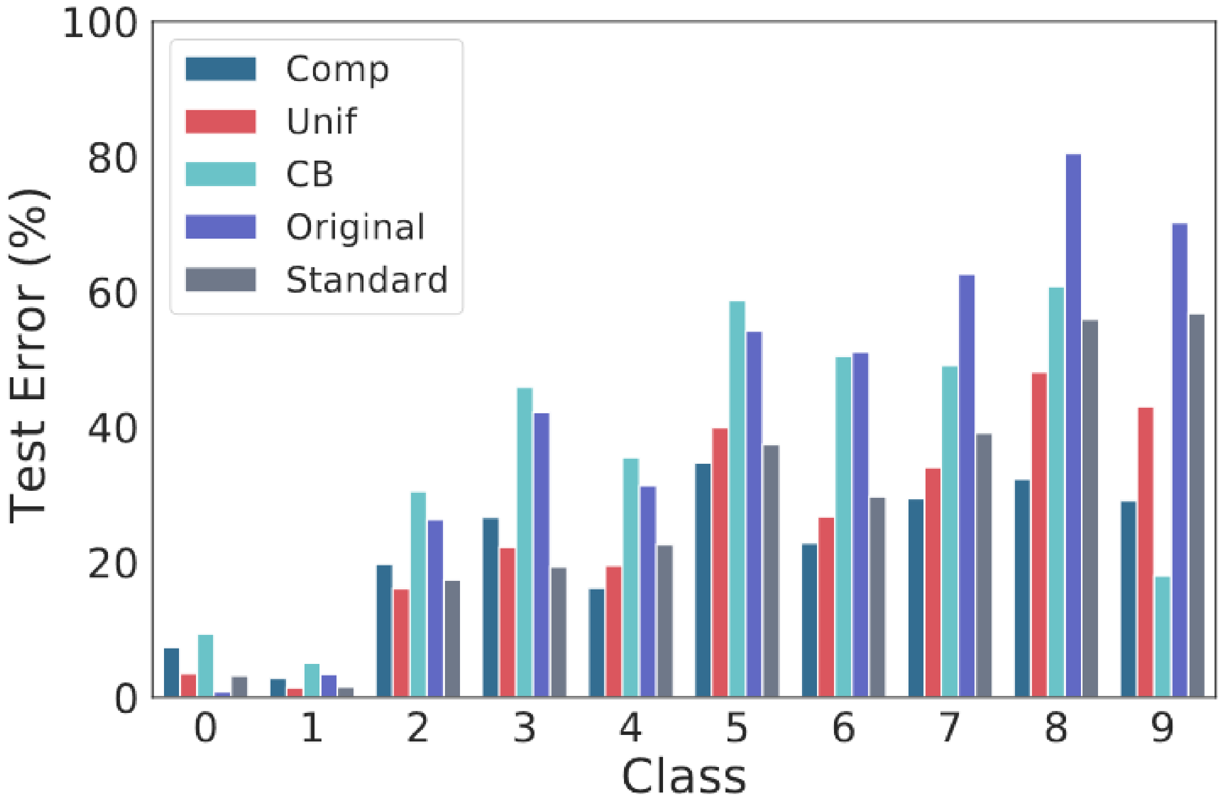


$$\alpha = (\max_j \beta_j + \min_j \beta_j)$$

Experiment

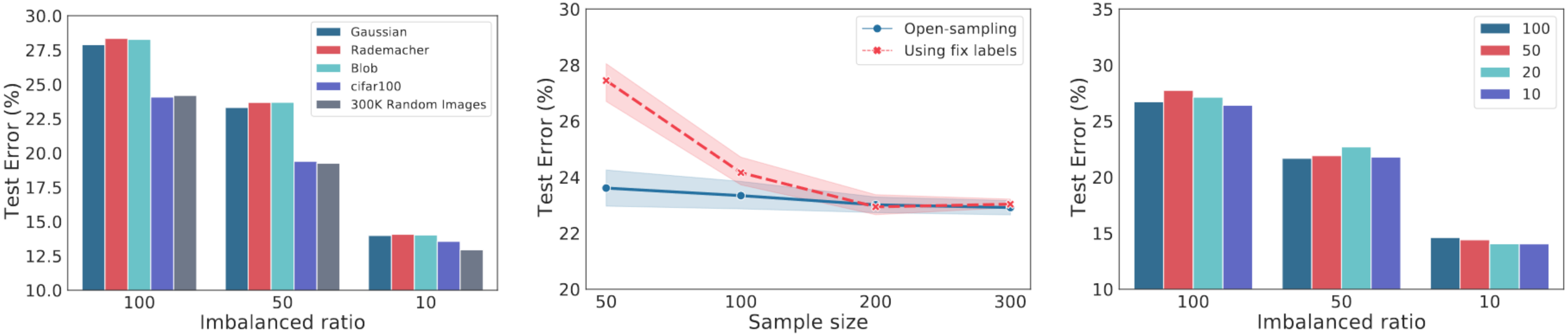


(a) Label distribution.



(b) Label distribution.

Experiment



(d) Auxiliary dataset. (e) Sample size. (f) Number of Classes.

Experiment

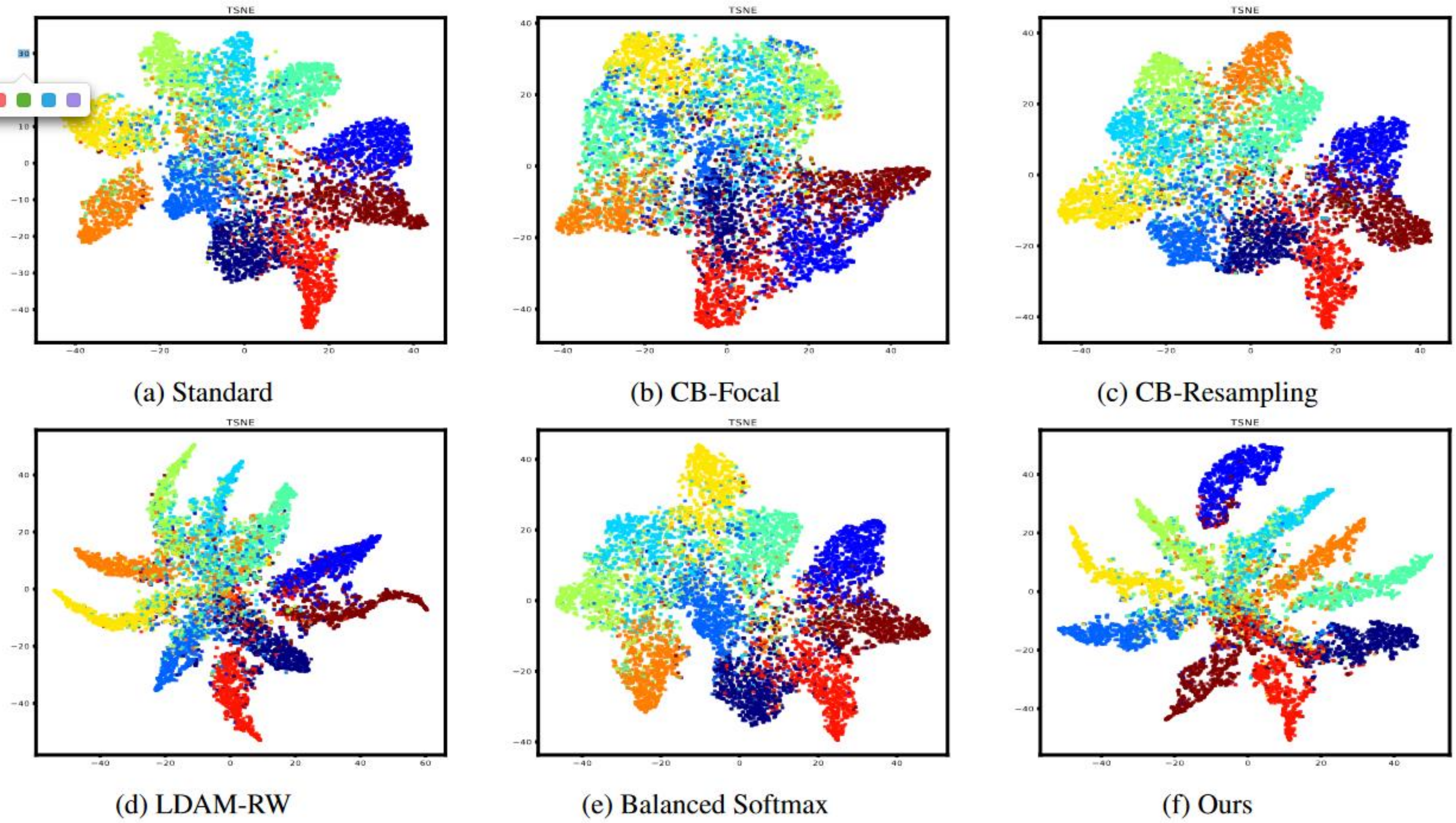


Figure 5. t-SNE visualization of test set on long-tailed CIFAR-10 with imbalance ratio 100. We can observe that LDAM and our method appear to learn more separable representations than Standard training and the other algorithms.

THANKS