上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

Contrastive Learning with Boosted Memorization

Zhihan Zhou

CMIC Shanghai Jiao Tong University

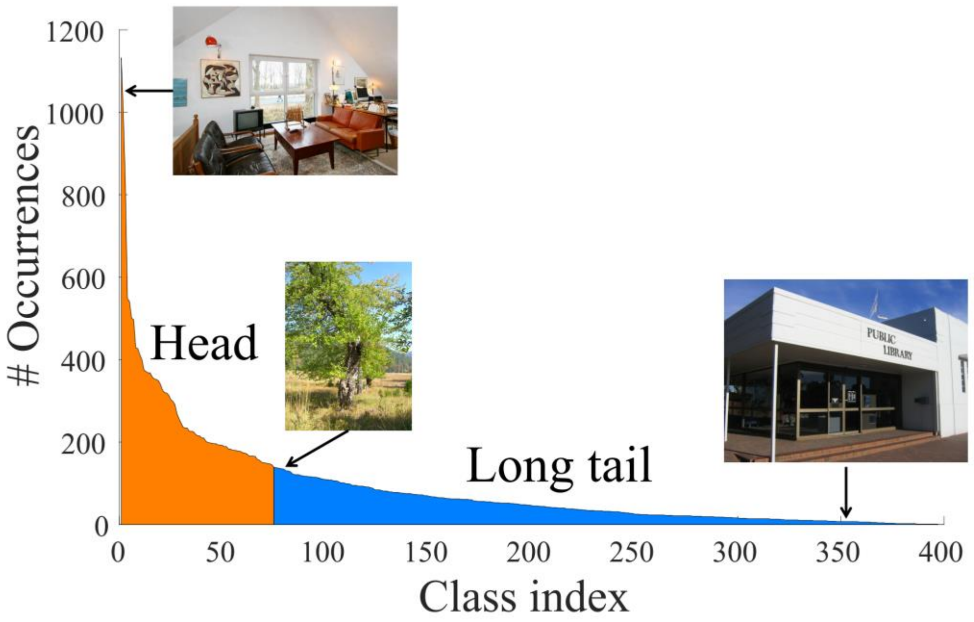
Coauthor with Jiangchao Yao, Yanfeng Wang, Bo Han, Ya Zhang

ICML 2022

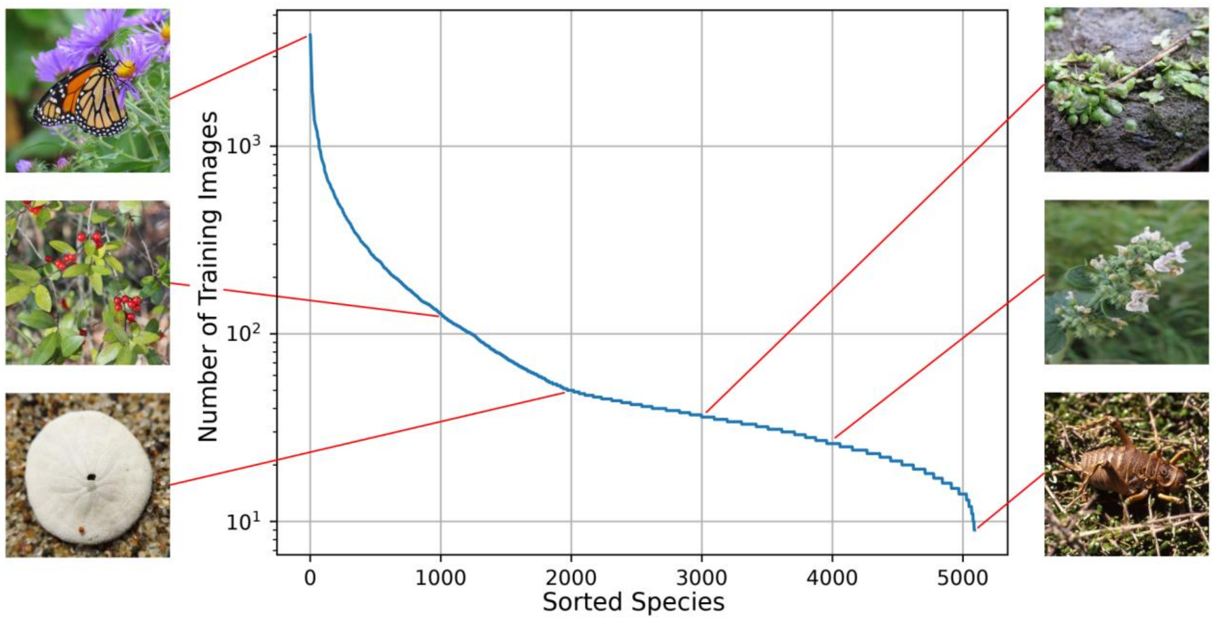
饮 水 思 源 • 爱 国 荣 校

Long-Tailed Distribution

Real-world natural sources usually follow a long-tailed distribution.



SUN-397[1]



iNaturalist[2]

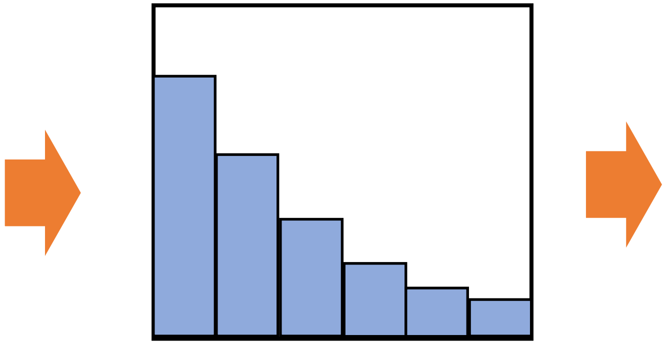
[1]Wang et al. “Learning to model the tail.” NeurIPS 2017

[2]Van Horn et al. “The inaturalist species classification and detection dataset.” CVPR 2018

Supervised Long-Tailed Learning

Supervised methods mainly depend on label information.

iNaturalist



Label guidance

- Resampling[1]
- Reweighting[2]
- Logit Adjustment[3]
- Transfer Learning[4]

……

[1]Kang et al. "Decoupling representation and classifier for long-tailed recognition." ICLR 2019
[2]Cui et al. "Class-balanced loss based on effective number of samples." CVPR 2019
[3]Menon et al. "Long-tail learning via logit adjustment." ICLR 2020
[4]Yin et al. "Feature transfer learning for face recognition with under-represented data." CVPR 2019

SJTU

Drawbacks of existing works

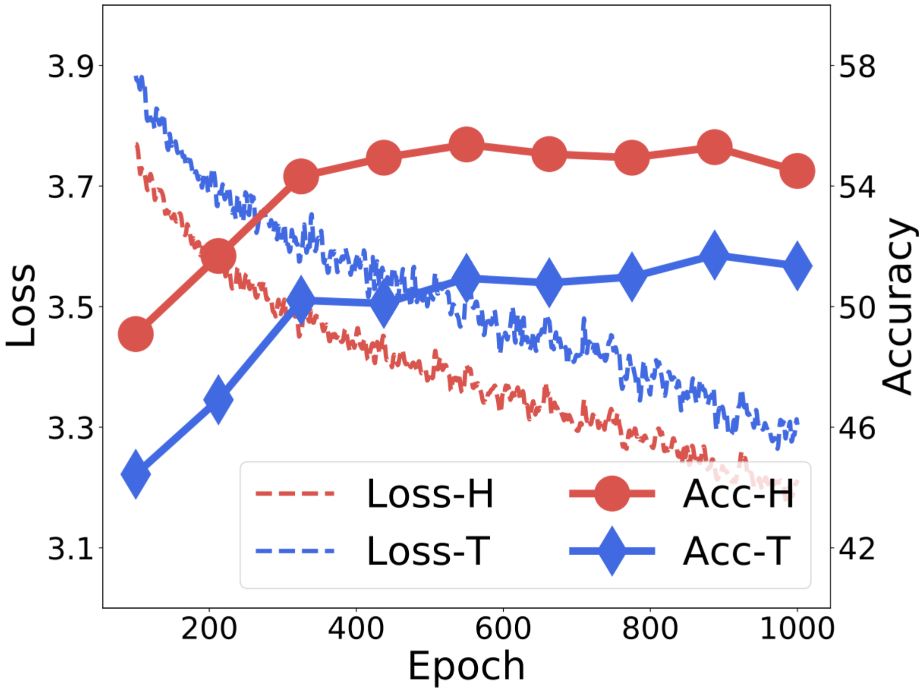
- *Loss perspective*: Focal loss[1], rwSAM[2]
  - sensitive to the accuracy of the tail sample discovery
- *Model perspective*: DnC[3], SDCLR[4]
  - require empirical heuristic and are black-box to understand

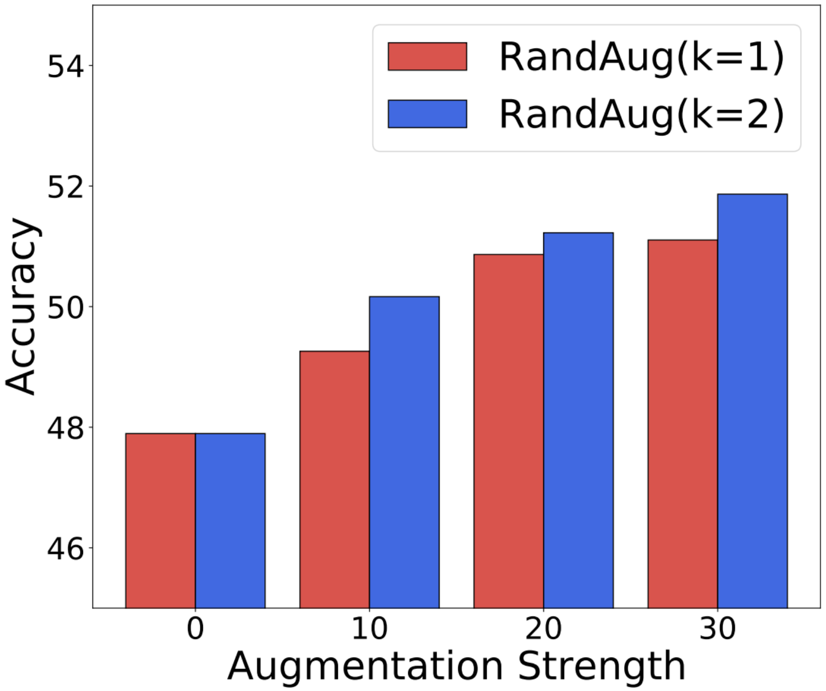
Label guidance

These works have not shown the expected promise due to *noisy tail sample discovery*.

[1]Lin et al. “Focal loss for dense object detection.” ICCV 2017
[2]Liu et al. “Self-supervised learning is more robust to dataset imbalance.” ICLR 2022
[3]Tian et al. “Divide and contrast: self-supervised learning from uncurated data.” ICCV 2021
[4]Jiang et al. “Self-damaging contrastive learning.” ICML 2021

Motivations of Boosted Contrastive Learning



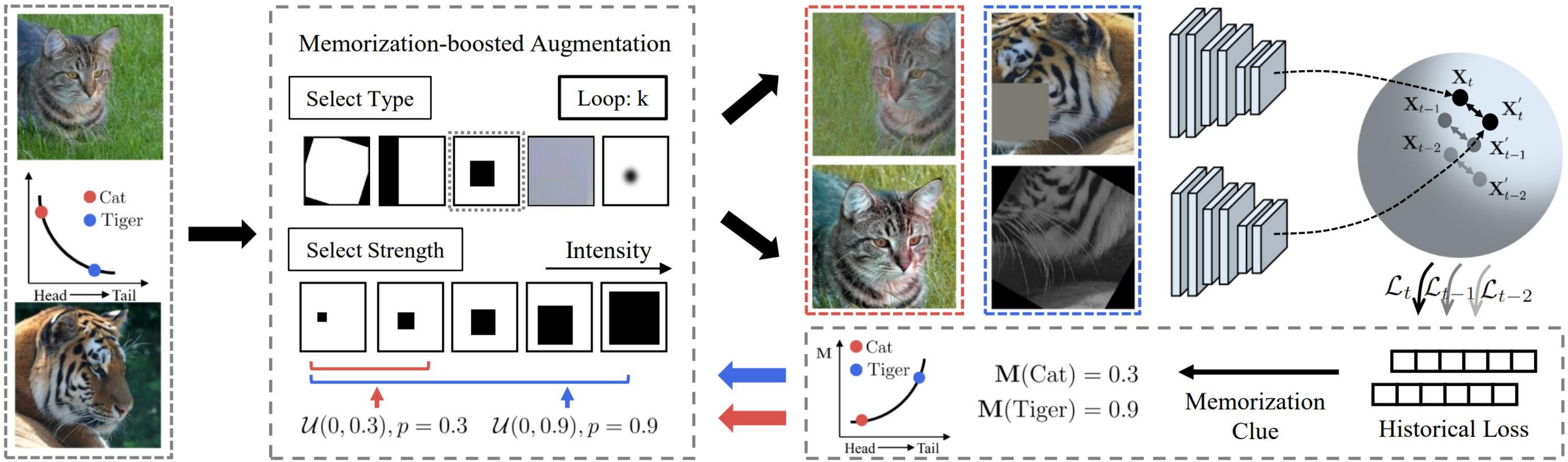


SimCLR Pretrained on CIFAR-LT

(*Left*) Memorization effect still holds under long-tailed distribution.

(*Right*) Stronger information discrepancy motivates tail samples mining.

Boosted Contrastive Learning



- Calculate *memorization scores* based on historical statistics to detect tail.
- Construct *instance-wise augmentations* to enhance representation learning.

Memorization-Guided Tail Discovery

Recent advances in *memorization* definition [1]:

$$\mathtt{mem}(\mathcal{A}, S, i) := \underset{h\sim\mathcal{A}(S)}{\mathbf{Pr}}[h(x_i) = y_i] - \underset{h\sim\mathcal{A}(S^{\setminus i})}{\mathbf{Pr}}[h(x_i) = y_i]$$

- Drawbacks: computationally *expensive* and limited to *supervised learning*.
- Inspired by the *learning speed proxy* explored in [2], we extend the memorization estimation to *self-supervised learning*.

$$\mathcal{L}^m_{i,0} = \mathcal{L}_{i,0}, \quad \mathcal{L}^m_{i,t} = \beta\mathcal{L}^m_{i,t-1} + (1-\beta)\mathcal{L}_{i,t}$$

$$\mathbf{M}_{i,t} = \frac{1}{2}\left(\frac{\mathcal{L}^m_{i,t} - \bar{\mathcal{L}}^m_t}{\max\left\{\left|\mathcal{L}^m_{i,t} - \bar{\mathcal{L}}^m_t\right|\right\}_{i=0,\ldots,N}} + 1\right)$$

Merits

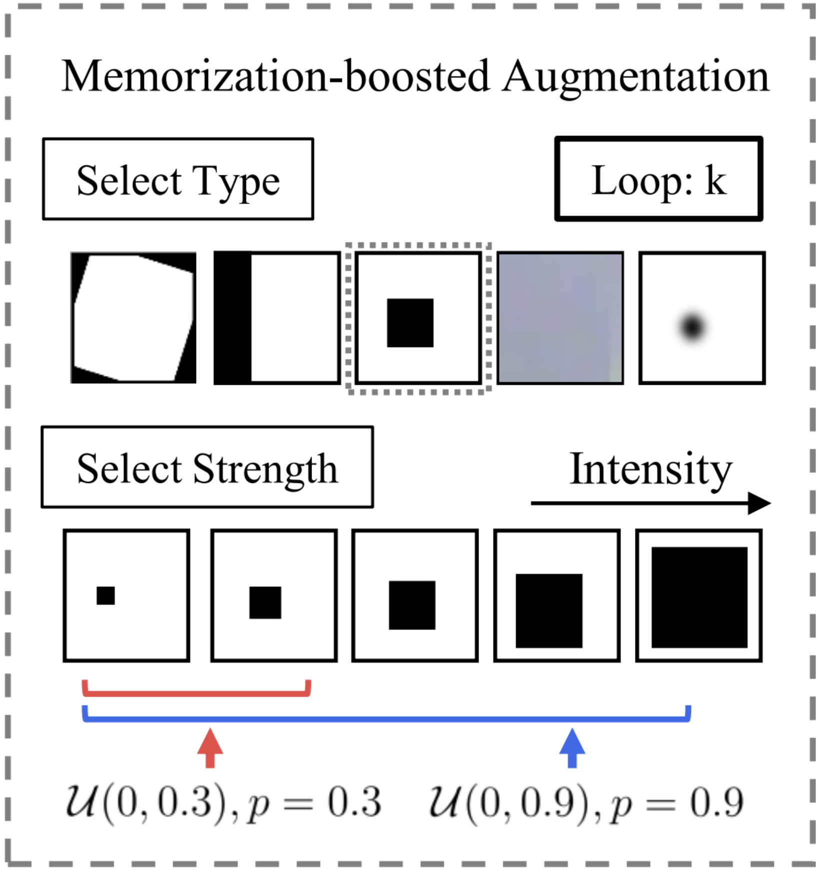
✓Computationally *efficient*

✓ Robust to the randomness issue

[1] Feldman et al. "Does learning require memorization? a short tale about a long tail." STOC 2020

[2] Jiang et al. "Characterizing structural regularities of labeled data in overparameterized models." ICML 2021

Memorization-boosted Augmentation



- Head
- Tail

- "InfoMin Principle"[1]

Good view set: share the *minimal* information *necessary* to perform well at downstream task.

- Dynamical information discrepancy

$$\Psi(x_i; \mathcal{A}, \mathbf{M}_i) = a_1(x_i) \circ \ldots \circ a_k(x_i),$$

$$a_j(x_i) = \begin{cases} A_j(x_i; \mathbf{M}_i\zeta) & u \sim \mathcal{U}(0,1) \ \& \ u < \mathbf{M}_i \\ x_i & \text{otherwise} \end{cases}$$

✓ Strong(Tail): Enhance tail representation

✓ Weak(Head): Avoid task-irrelevant noise

[1] Tian et al. "What makes for good views for contrastive learning? " NeurIPS 2020

Experiments

Table 1. Fine-grained analysis for various methods pre-trained on CIFAR-100-LT, ImageNet-LT and Places-LT. Many/Medium/Few corresponds to three partitions on the long-tailed data. Std is the standard deviation of the accuracies among Many/Medium/Few groups.

	CIFAR-100-LT				ImageNet-LT				Places-LT			
Methods	Many	Medium	Few	Std	Many	Medium	Few	Std	Many	Medium	Few	Std
SimCLR	48.70	46.81	44.02	2.36	41.16	32.91	31.76	5.13	31.12	33.85	35.62	2.27
Focal	48.46	46.73	44.12	**2.18**	40.55	32.91	31.29	4.95	30.18	31.56	33.32	1.57
DnC	54.00	46.68	45.65	4.55	29.54	19.62	18.38	6.12	28.20	28.07	28.46	**0.20**
SDCLR	51.22	49.22	45.85	2.71	41.24	33.62	32.15	4.88	32.08	35.08	35.94	2.03
BCL-I	50.45	48.23	45.97	2.24	**42.53**	**35.66**	33.93	4.54	32.27	34.96	**38.03**	2.88
BCL-D	**53.98**	**51.97**	**49.52**	2.23	41.92	35.29	**34.07**	**4.22**	**32.34**	**35.44**	37.75	2.71

- Consistent performance gain on Many/Medium/Few partitions.
- Relative low Std confirms the merits on representation balancedness.

Experiments: Downstream Task

Table 3. The classification accuracy of supervised learning with self-supervised pre-training on CIFAR-100-LT and ImageNet-LT.

Dataset	CE	CE with the following model initialization					
		CL	Focal	DnC	SDCLR	BCL-I	BCL-D
CIFAR-100-LT	41.7	44.4	44.4	44.4	<u>44.6</u>	**45.1**	**45.4**
ImageNet-LT	41.6	45.5	45.4	42.2	<u>45.9</u>	**46.9**	**46.4**
Dataset	**cRT**	**cRT with the following model initialization**					
		CL	Focal	DnC	SDCLR	BCL-I	BCL-D
CIFAR-100-LT	44.1	48.9	48.7	48.6	<u>49.8</u>	**49.9**	**50.0**
ImageNet-LT	46.7	<u>47.5</u>	47.3	43.5	47.3	**48.4**	**48.1**
Dataset	**LA**	**LA with the following model initialization**					
		CL	Focal	DnC	SDCLR	BCL-I	BCL-D
CIFAR-100-LT	45.7	50.1	49.5	49.7	<u>50.4</u>	**50.8**	**50.5**
ImageNet-LT	47.4	<u>48.6</u>	48.4	45.6	48.2	**49.7**	**49.1**

- BCL can potentially further boost the supervised long-tailed representation learning.

Experiments: Downstream Task

Table 4. The linear probing performance of all methods on CUB, Cars, Aircrafts, Dogs and NABirds. We pretrain the backbone ResNet-50 on ImageNet-LT under different methods, and then transfer to these datasets for the linear probing evaluation. The top-1 and top-5 accuracies are reported by computing the highest and top-5 highest predictions to match the ground-truth labels.

	CUB		Cars		Aircrafts		Dogs		NABirds		All	
Methods	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
SimCLR	29.62	57.35	21.45	44.93	30.48	57.01	46.67	79.22	16.52	37.61	28.95	55.22
Focal	29.08	56.89	21.40	44.35	30.99	57.64	46.59	78.14	16.31	36.97	28.87	54.80
DnC	16.97	40.90	8.15	23.79	13.71	33.18	29.83	61.92	8.44	22.75	15.42	36.51
SDCLR	28.98	57.27	22.10	46.13	31.05	58.18	46.69	78.82	16.17	37.10	29.00	55.50
BCL-I	**30.00**	**58.08**	23.67	49.16	32.37	60.31	**48.61**	**79.99**	**17.42**	**38.96**	30.41	57.30
BCL-D	28.79	57.37	**25.90**	**51.34**	**34.95**	**62.77**	47.49	78.86	16.41	37.24	**30.71**	**57.51**

- Considerable improvements on various downstream fine-grained datasets.
- BCL encourages to learn more generalizable and robust representation.

Summary

- BCL builds a momentum loss to capture clues from the memorization effect.
- BCL drives the instance-wise augmentation to enhance long-tailed learning.
- BCL is *simple*, *adaptive*, and *orthogonal* to almost all the SSL methods.

Thanks! Code and models are available at

Github