Combating Representation Learning Disparity with Geometric Harmonization

Zhihan Zhou

CMIC Shanghai Jiao Tong University

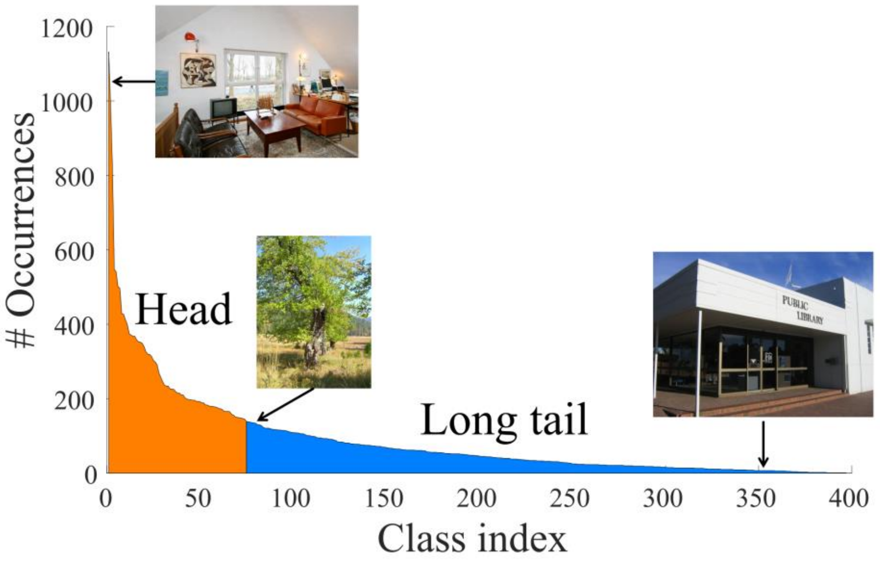
Coauthor with Jiangchao Yao, Feng Hong, Ya Zhang, Bo Han, Yanfeng Wang

NeurIPS 2023 Spotlight

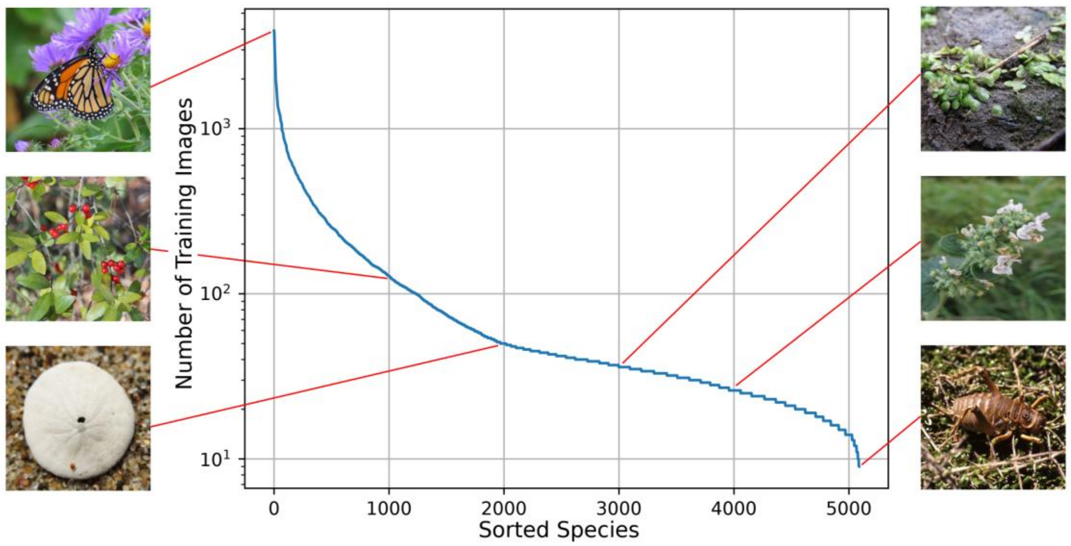
饮 水 思 源 • 爱 国 荣 校

Long-Tailed Distribution

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



SUN-397[1]



iNaturalist[2]

- The importance of long-tailed learning is further emphasized when extended to a range of safety-critical scenarios, including medical intelligence, autonomous driving and criminal surveillance.

[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

Existing works

Method	Aspect	Description
Focal [40]	Sample Reweighting	Hard example mining
rwSAM [41]	Optimization Surface	Data-dependent sharpness-aware minimization
SDCLR [28]	Model Pruning	Model pruning and self-contrast
DnC [59]	Model Capacity	Multi-expert ensemble
BCL [77]	Data Augmentation	Memorization-guided augmentation
GH	Loss Limitation	Geometric harmonization

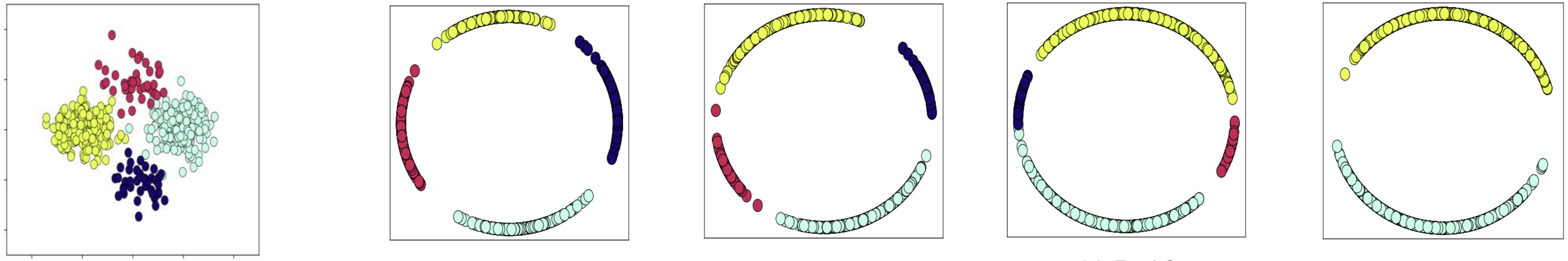
- Few works have considered the *intrinsic limitation* of the widely-adopted contrastive learning loss, which easily leads to *representation learning disparity* where head classes dominate the feature regime but tail classes passively collapse.

Motivation

- Why the conventional contrastive learning underperforms in self-supervised long-tailed context?

Conventional contrastive loss motivates *sample-level uniformity*, that is biased towards the head classes.

- Geometric Harmonization aims at achieving *category-level uniformity*, *i.e.*, equal allocation *w.r.t.* classes.



(a) R=1(Balanced) (b) R=4 (c) R=16 (d) R=64

Contrastive learning causes ***severer representation learning disparity*** when enlarging the imbalance ratios.

Geometric Harmonization

- Challenges I: No guarantee for the desired category-level uniformity
- Challenges II: The latent true labels are not available, while the estimated labels are noisy

Geometric Uniform Structure

$$\mathbf{M}_i^\top \cdot \mathbf{M}_j = C, \quad \forall i, j \in \{1, 2, \ldots, K\}, \; i \neq j,$$

Any two vectors in **M** have the same angle, namely, the unit space are equally partitioned by the vectors.

Overall objective

$$\min_{\theta, \hat{\mathbf{Q}}} \mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + w_{\text{GH}} \mathcal{L}_{\text{GH}},$$

Surrogate Label Allocation

$$\min_{\hat{\mathbf{Q}}=[\hat{\boldsymbol{q}}_1,\ldots,\hat{\boldsymbol{q}}_N]} \mathcal{L}_{\text{GH}} = -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x}_i \sim \mathcal{D}} \hat{\boldsymbol{q}}_i \log \boldsymbol{q}_i,$$

$$\text{s.t.} \;\; \hat{\mathbf{Q}} \cdot \mathbb{1}_N = N \cdot \boldsymbol{\pi}, \;\; \hat{\mathbf{Q}}^\top \cdot \mathbb{1}_K = \mathbb{1}_N,$$

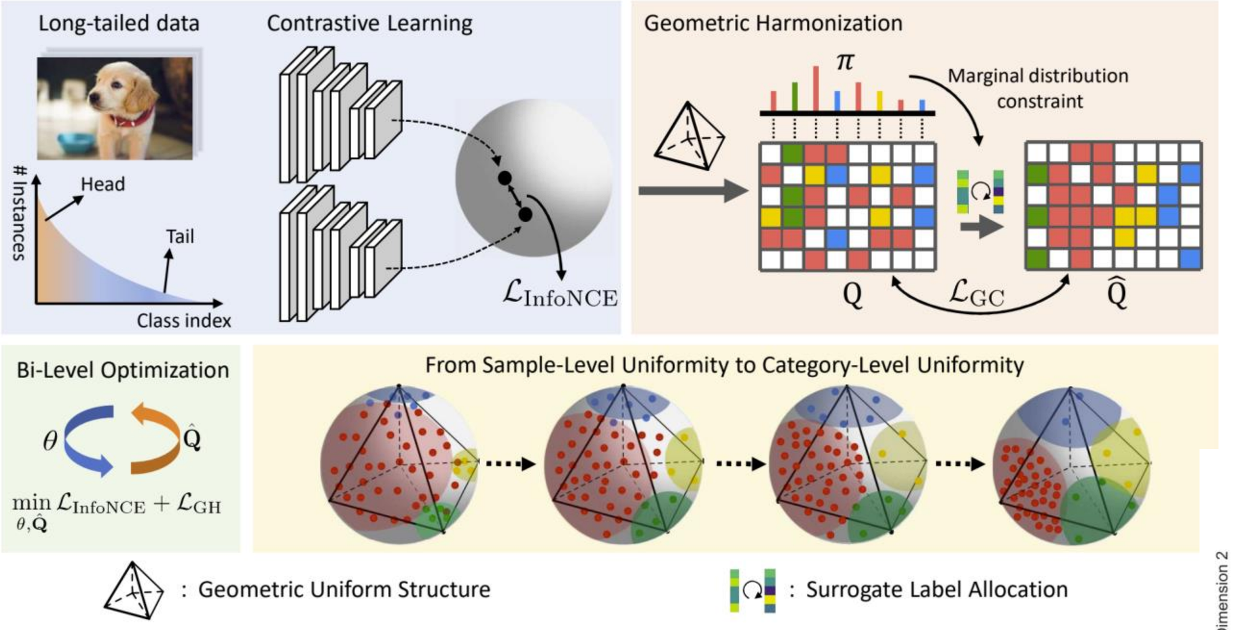
Algorithm 1 Surrogate Label Allocation.

Input: geometric cost matrix $\exp(\lambda \log \mathbf{Q})$ with $\mathbf{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N]$, marginal distribution constraint $\boldsymbol{\pi}$, Sinkhorn regularization coefficient λ, Sinkhorn iteration step E_s

Output: Surrogate label matrix $\hat{\mathbf{Q}}$

Set scaling vectors $\boldsymbol{u} \leftarrow \frac{1}{K} \cdot \mathbb{1}_K, \boldsymbol{v} \leftarrow \frac{1}{N} \cdot \mathbb{1}_N$.
Set distribution constraints $\boldsymbol{r} \leftarrow \frac{1}{N} \cdot \mathbb{1}_N, \boldsymbol{c} \leftarrow \boldsymbol{\pi}$.
for iteration $i = 0, 1, \ldots, E_s$ **do**
    $\boldsymbol{u} \leftarrow \log \boldsymbol{c} - \log \left((\exp(\lambda \log \mathbf{Q})) \cdot \exp(\boldsymbol{v}) \right)$.
    $\boldsymbol{v} \leftarrow \log \boldsymbol{r} - \log \left((\exp(\lambda \log \mathbf{Q}))^\top \cdot \exp(\boldsymbol{u}) \right)$.
end for
return $\hat{\mathbf{Q}} = N \cdot \text{diag}(\boldsymbol{u}) \exp(\lambda \log \mathbf{Q}) \text{diag}(\boldsymbol{v})$

Geometric Harmonization



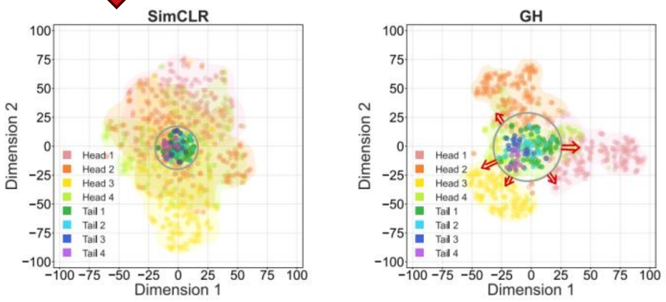
Theorem 3.4. *(Optimal state for $\mathcal{L}$) Given Eq. (4) under the proper optimization strategy, when it arrives at the category-level uniformity (Definition 3.3) defined on the geometric uniform structure* $\mathbf{M}$ *(Definition 3.1), we will achieve the minimum of the overall loss $\mathcal{L}^*$ as*

$$\mathcal{L}^* = -2\sum_{l=1}^{L} \pi_l^{\boldsymbol{y}} \log\left(1/\left(1+(K-1)\exp(C-1)\right)\right) + \log\left(J/L\right), \quad (5)$$

where J denotes the size of the collection of the negative samples and $\pi^{\boldsymbol{y}}$ refers to the marginal distribution of the latent ground-truth labels $\boldsymbol{y}$.

theoretically grounded!

empirically verified!



	Inter-class Uniformity		Neighborhood Uniformity	
	SimCLR	+GH	SimCLR	+GH
C100	1.00	**2.80**	0.72	**2.00**
C50	1.23	**2.73**	0.91	**1.94**
C10	1.18	**2.60**	0.85	**1.83**

➢ Our GH can promote the desired category-level uniformity!

Main results

Dataset		SimCLR	+GH	Focal	+GH	SDCLR	+GH	DnC	+GH	BCL	+GH	**Improv.**
CIFAR-R100	Many	54.97	57.38	54.24	57.01	57.32	57.44	55.41	57.56	59.15	59.50	**+1.56**
	Med	49.39	52.27	49.58	52.93	50.70	52.85	51.30	53.74	54.82	55.73	**+2.35**
	Few	47.67	52.12	49.21	51.74	50.45	54.06	50.76	53.26	55.30	57.67	**+3.09**
	Std	3.82	2.99	2.80	2.76	3.90	2.38	2.54	2.36	2.37	1.89	**-0.61**
	Avg	50.72	53.96	51.04	53.92	52.87	54.81	52.52	54.88	56.45	57.65	**+2.32**
CIFAR-R50	Many	56.00	58.88	55.40	57.97	57.50	58.47	56.03	59.04	59.44	60.82	**+2.16**
	Med	50.48	53.00	51.14	53.55	51.85	53.88	52.68	55.05	54.73	57.58	**+2.44**
	Few	50.12	54.27	50.02	53.58	52.15	53.58	50.83	54.81	57.30	58.55	**+2.87**
	Std	3.30	3.09	2.84	2.54	3.18	2.74	2.64	2.38	2.36	1.66	**-0.38**
	Avg	52.24	55.42	52.22	55.06	53.87	55.34	53.21	56.33	57.18	59.00	**+2.49**
CIFAR-R10	Many	57.85	59.26	58.18	60.06	58.47	59.21	59.82	61.09	60.41	61.41	**+1.26**
	Med	55.06	56.91	55.82	56.79	54.79	56.06	56.67	58.33	57.15	59.27	**+1.57**
	Few	54.03	55.85	54.64	57.24	52.97	55.58	56.21	57.33	59.76	60.30	**+1.74**
	Std	1.98	1.75	1.80	1.77	2.80	1.97	1.96	1.95	1.73	1.07	**-0.35**
	Avg	55.67	57.36	56.23	58.05	55.44	56.97	57.59	58.94	59.12	60.34	**+1.52**
ImageNet-LT	Many	41.69	41.53	42.04	42.55	40.87	41.92	41.70	42.19	42.92	43.22	**+0.44**
	Med	33.96	36.35	35.02	36.75	33.71	36.53	34.68	36.63	35.89	38.16	**+2.23**
	Few	31.82	35.84	33.32	36.28	32.07	36.04	33.58	35.86	33.93	36.96	**+3.25**
	Std	5.19	3.15	4.62	3.49	4.68	3.26	4.41	3.45	4.73	3.32	**-1.39**
	Avg	36.65	38.28	37.49	38.92	36.25	38.53	37.23	38.67	38.33	39.95	**+1.68**
Places-LT	Many	31.98	32.46	31.69	32.40	32.17	32.78	32.07	32.51	32.69	33.22	**+0.55**
	Med	34.05	35.03	34.33	35.14	34.71	35.60	34.51	35.55	35.37	36.00	**+0.87**
	Few	35.63	36.14	35.73	36.49	35.69	36.18	35.84	35.91	37.18	37.62	**+0.45**
	Std	1.83	1.89	2.05	2.08	1.82	1.82	1.91	1.87	2.26	2.23	0.00
	Avg	33.61	34.33	33.65	34.42	33.99	34.70	33.90	34.52	34.76	35.32	**+0.68**

GH provides consistent improvements on top of all the baseline methods in terms of linear probing accuracy and representation balancedness.

Experiments: Transfer Learning

Dataset	LA	Logit adjustment pretrained with the following SSL methods										Improv.
		SimCLR	+GH	Focal	+GH	SDCLR	+GH	DnC	+GH	BCL	+GH	
CIFAR-LT	46.61	49.81	50.84	49.83	51.04	49.79	50.73	49.97	50.84	50.38	51.32	**+1.00**
ImageNet-LT	48.27	51.10	51.67	51.15	51.82	50.94	51.64	51.31	51.88	51.43	52.06	**+0.63**
Places-LT	27.07	32.63	33.86	32.69	33.75	32.55	34.03	32.98	34.09	33.15	34.48	**+1.24**

	Image Classification		Fine-Grained Visual Classification					
	ImageNet	Places	CUB200	Aircraft	StanfordCars	StanfordDogs	NABirds	Average
SimCLR	52.06	37.65	44.61	65.89	57.63	50.99	46.86	53.20
+GH	**53.39**	**38.47**	**45.76**	**68.08**	**60.24**	**52.88**	**47.58**	**54.91**

Pretrained on large-scale long-tailed CC3M!

	Object Detection			Instance Segmentation		
	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
SimCLR	31.7	51.0	33.9	30.2	49.8	32.1
+GH	**32.7**	**52.2**	**35.2**	**31.1**	**50.8**	**33.0**

GH can potentially further boost the supervised long-tailed learning and downstream fine-grained classification, object detection and instance segmentation.

SJTU

Summary

- Conventional contrastive learning fails under long-tailed distribution.
- The intrinsic limitation lies in *pursuing sample-level uniformity*.
- We propose GH to efficiently *promote category-level uniformity* via an instance-wise label calibration based on the geometric statistics.
- GH is theoretically and empirically verified to *tackle representation learning disparity and enhance downstream generalization*.

Thanks! Codes will be available at:

https://github.com/MediaBrain-SJTU/Geometric-Harmonization