UNIVERSITEIT VAN AMSTERDAM

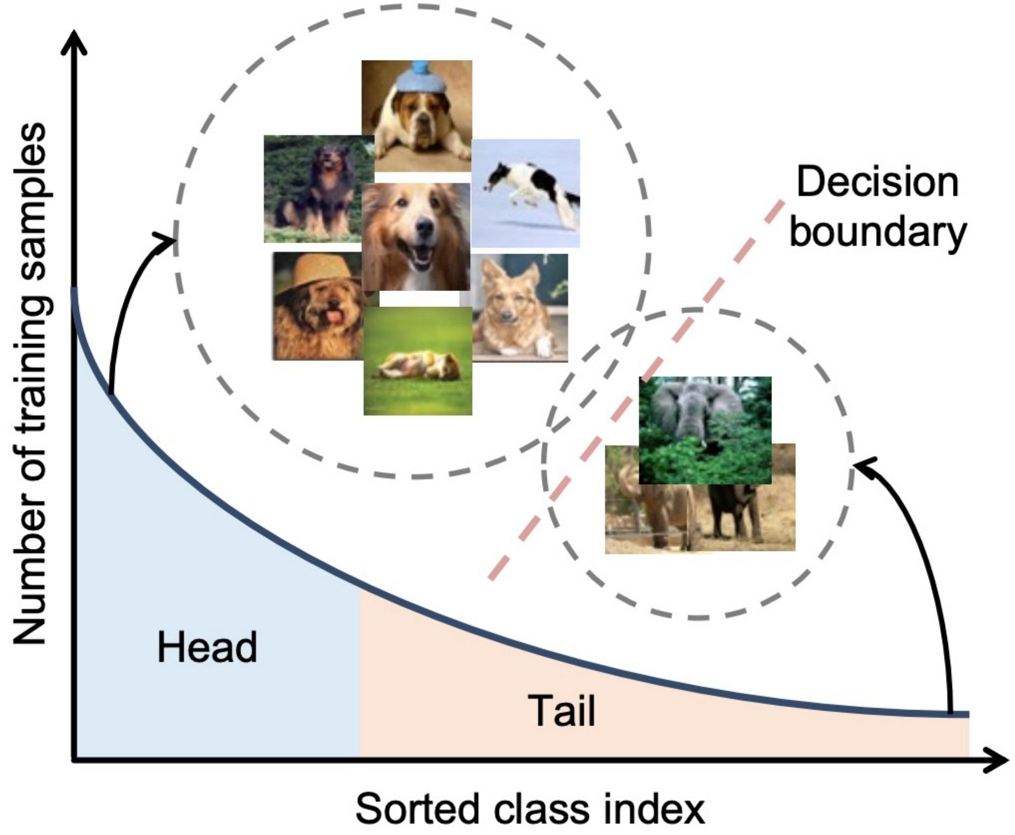
SuperDisco: Super-Class Discovery Improves Visual Recognition for the Long-Tail

Yingjun Du[1], Jiayi Shen[1], Xiantong Zhen[1,2]*, Cees G. M. Snoek[1]
[1]AIM Lab, University of Amsterdam [2]Inception Institute of Artificial Intelligence

THU-AM-329

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

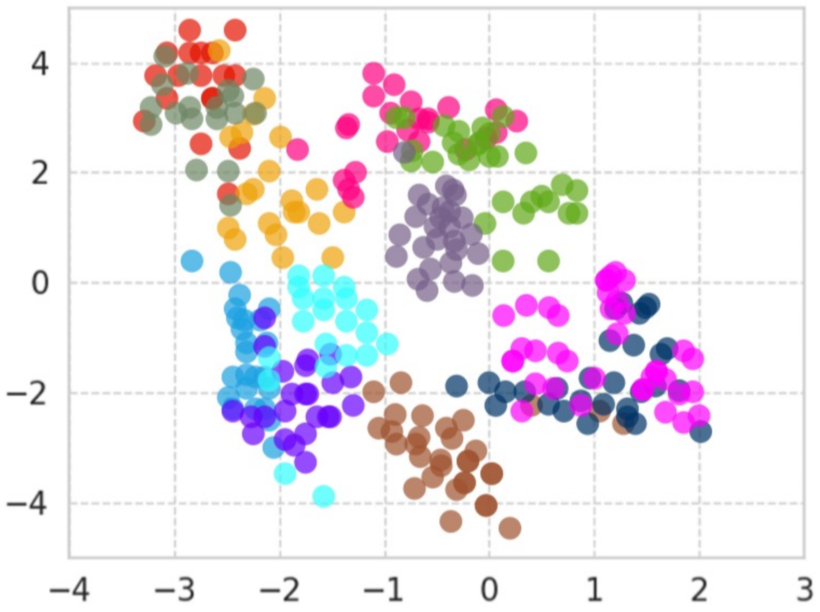
Long-tailed problem

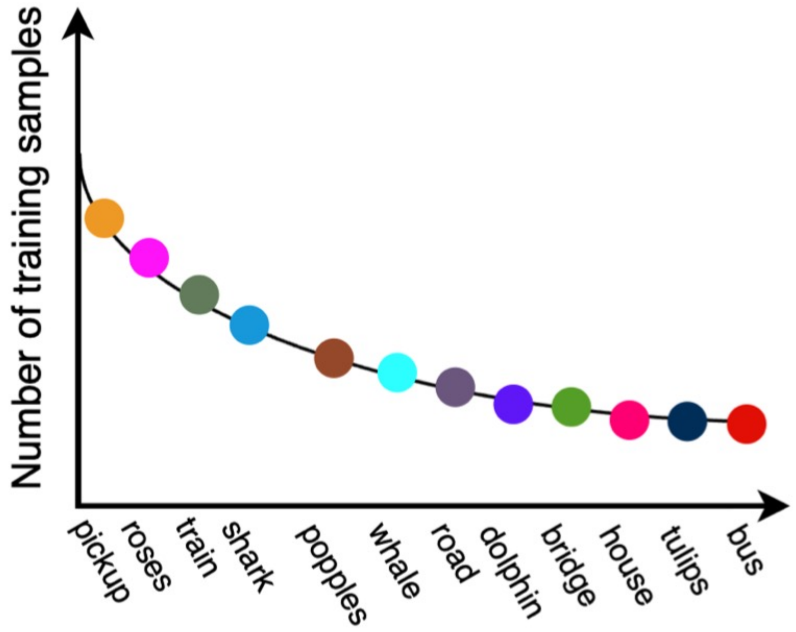


The head-class feature space learned on these sampled is often larger than tail classes, while the decision boundary is usually biased towards dominant classes.

Source from Zhang et.al., 2021

Long-tailed features





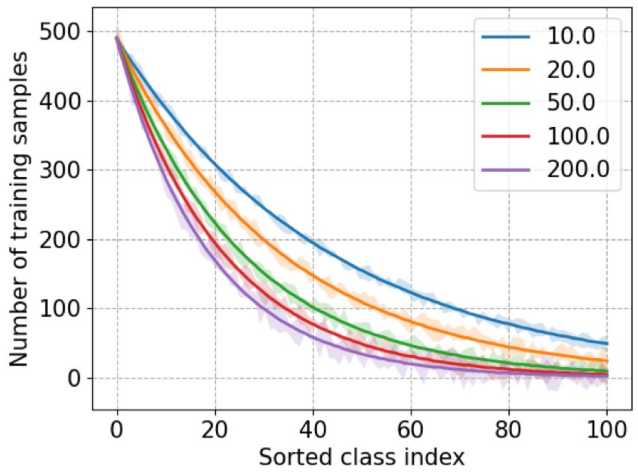
Head data dominate the distribution of features, which causes the tail features to fall within the head feature space.

Super-classes distributions

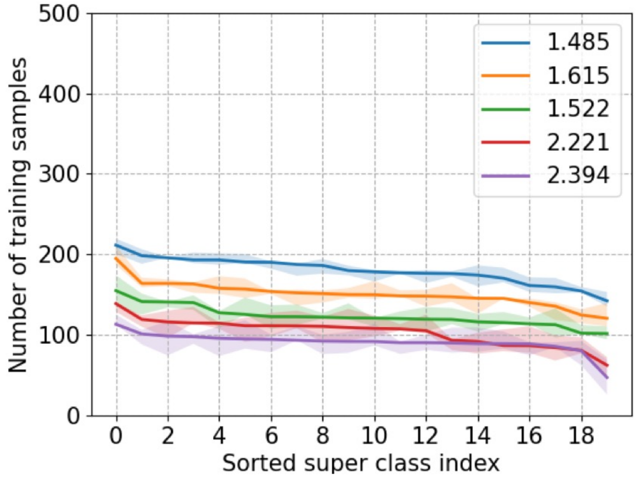
- In the real world, each category has a corresponding superclass, e.g., bus, taxi, and train all belong to the vehicle superclass.

- This observation raises the question: are super-classes of categories also distributed along a long-tail?

Original classes vs. Super-classes in imbalance ratios



(a) 100 original classes



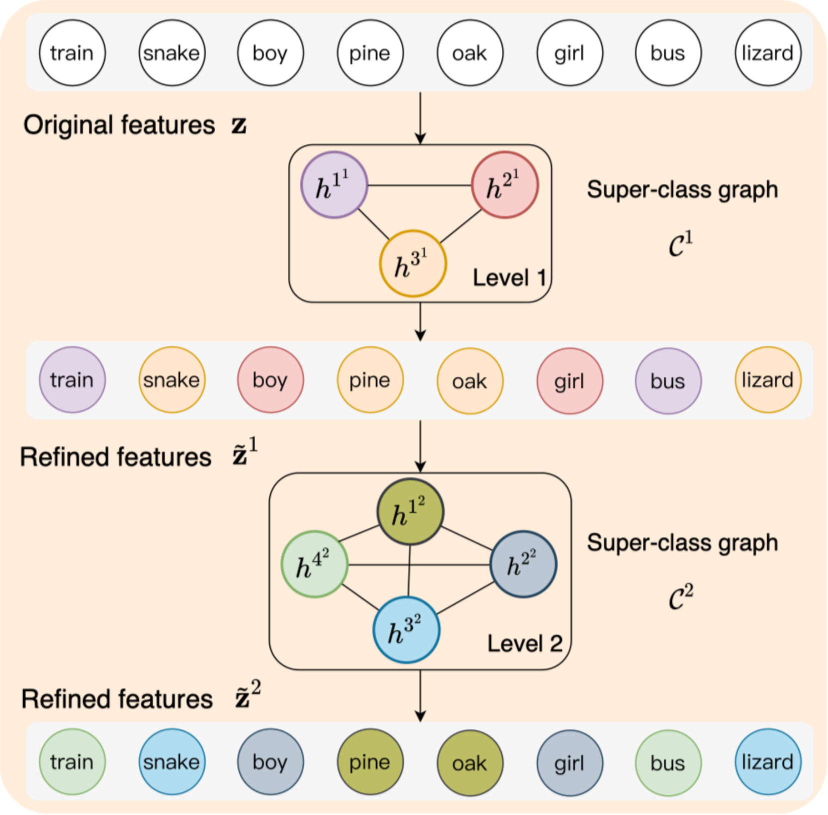
(b) 20 ground truth super-classes

The data imbalance of super-classes is considerably lower than those of the original classes.

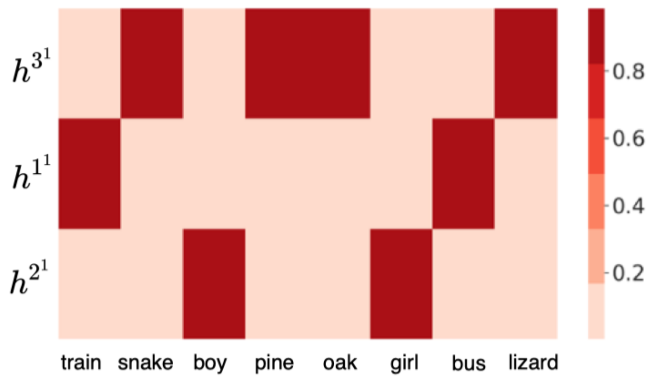
SuperDisco

- We propose a SuperDisco to learn to discover the **super-class graph** for long-tailed visual recognition.
- We construct a learnable graph that discovers the super-class in a hierarchy of semantic abstraction to guide feature representation learning.
- These balanced super-class features could be used to guide the original tail data away from the dominant role of the head data, thus making the tail data more discriminative.

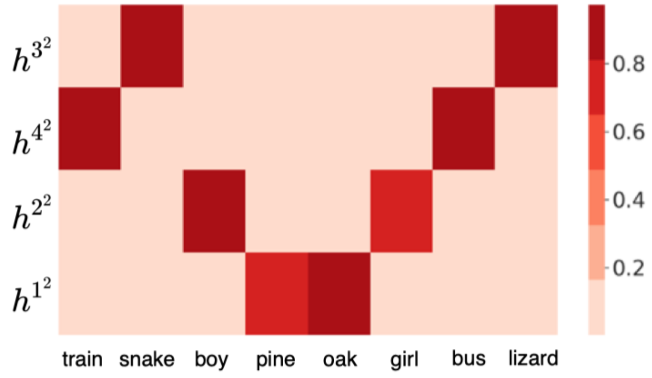
SuperDisco



(a) Illustration of proposed SuperDisco



(b) Visualization of discovered super-class graph $\mathcal{C}^1$



(c) Visualization of discovered super-class graph $\mathcal{C}^2$

The color in each graph represents the discovered super-class. Our SuperDisco can discover the potential super-class at different levels hidden in each category from (b) and (c).

SuperDisco: Learning to discover the super-class graph

- Define super-class graph: $\mathcal{C}^l = (\mathbf{H}_{\mathcal{C}}^l, \mathbf{A}_{\mathcal{C}}^l)$
  - Vertex:

$$\mathbf{H}_{\mathcal{C}}^l = \{\mathbf{h}^{i^l} | \forall i^l \in [1, C^l]\} \in \mathbb{R}^{C^l \times d}$$

  - Edge:

$$A_{\mathcal{C}}^l(\mathbf{h}^{i^l}, \mathbf{h}^{j^l}) = \sigma(\mathbf{W}_c^l(|\mathbf{h}^{i^l} - \mathbf{h}^{j^l}|/\gamma_c^l) + \mathbf{b}_c^l),$$

SuperDisco

- Message passing
  - the original features are rectified and refined, which attend to the most relevant entities according to the similarity between the original image features and super-classes.

$$\mathbf{H}_{\mathcal{R}}^{(m+1)} = \mathrm{MP}(\mathbf{A}_{\mathcal{R}}^{l}, \mathbf{H}_{\mathcal{R}}^{(m)}; \mathbf{W}^{(m)}),$$

$$A_{\mathcal{R}}^{l}(\mathbf{h}^{i^l}, \mathbf{z}^{\mathbf{l}}) = \sigma(\mathbf{W}_{r}^{l}(|\mathbf{h}^{i^l} - \mathbf{z}^{\mathbf{l}}| / \gamma_{r}^{l}) + \mathbf{b}_{r}^{l}),$$

$$\mathbf{H}_{\mathcal{R}}^{l} = \{[\mathbf{z}^{l}, \mathbf{h}^{i^l}] | \forall i^l \in [1, C^l]\} \in \mathbb{R}^{(C^l+1)\times d}$$

Meta-SuperDisco

- We propose a **meta-learning** variant of our SuperDisco algorithm to discover the super-class graph, enabling the model to achieve **more balanced** image representations.
  - We have a small amount of balanced data as validation data.

Prototype graph

- prototype graph: $\mathcal{P}=(\mathbf{C}_{\mathcal{P}}, \mathbf{A}_{\mathcal{P}})$
  - Edge weight:

$$A_{\mathcal{P}}(\mathbf{c}^i, \mathbf{c}^j) = \sigma(\mathbf{W}_p(|\mathbf{c}^i - \mathbf{c}^j|/\gamma_p) + \mathbf{b}_p),$$

  - Vertex:

$$\mathbf{C}_{\mathcal{P}_i}=\{\mathbf{c}^i|\forall i \in [1, K]\}$$

Super graph

- Connecting the prototype graph to the super-class graph

$$\mathcal{S}^l = (\mathbf{A}^l, \mathbf{M}^l)$$

  - Edge weight: $\mathbf{A}^l = (\mathbf{A}_{\mathcal{P}}, \mathbf{A}_{\mathcal{S}}^l; {\mathbf{A}_{\mathcal{S}}^l}^{\mathrm{T}}, \mathbf{A}_{\mathcal{C}}^l)$

where $A_{\mathcal{S}}^l(\mathbf{c}^i, \mathbf{h}^{j^l}) = \dfrac{\exp(-\|(\mathbf{c}^i - \mathbf{h}^{j^l})/\gamma_s^l\|_2^2/2)}{\sum_{j^{l'}=1}^{J} \exp(-\|(\mathbf{c}^i - \mathbf{h}^{j^{l'}})/\gamma_s^l\|_2^2/2)},$

  - Vertex:

$$\mathbf{M} = (\mathbf{C}_{\mathcal{P}^l}^l; \mathbf{H}_{\mathcal{C}^l}^l)$$

- Message passing $\mathbf{M}^{(m+1)} = \mathrm{MP}(\mathbf{A}^l, \mathbf{M}^{(m)}; \mathbf{W}^{(m)}).$

Benefit of SuperDisco and Meta-SuperDisco

	Imbalance ratio				
	10	20	50	100	200
Baseline	60.3	57.3	47.5	44.9	39.3
SuperDisco	65.9	60.7	57.2	50.9	45.2
Meta-SuperDisco	68.5	63.1	58.3	53.8	47.5

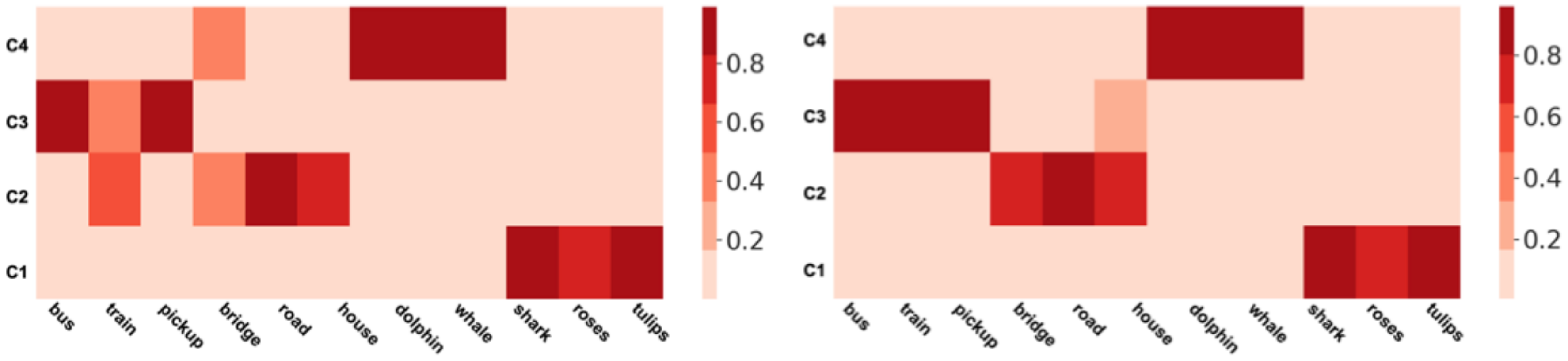
SuperDisco achieves better performance, while Meta-SuperDisco is even better for the long-tailed challenges.

Effect of number of super-class levels

	Imbalance ratio				
	10	20	50	100	200
Baseline	60.3	57.3	47.5	44.9	39.3
(20)	61.2	60.1	49.9	47.3	41.9
(2, 4, 8)	65.3	62.7	53.1	49.8	43.2
(4, 8, 16)	**69.1**	**64.2**	55.2	52.3	45.9
(4, 8, 16, 32)	68.5	63.1	58.3	**53.8**	**47.5**
(4, 8, 16, 32, 64)	66.9	62.7	**58.9**	52.9	46.3
Oracle super-classes	66.9	63.2	54.7	51.4	43.2

SuperDisco provides higher performance gains with more complex hierarchies.

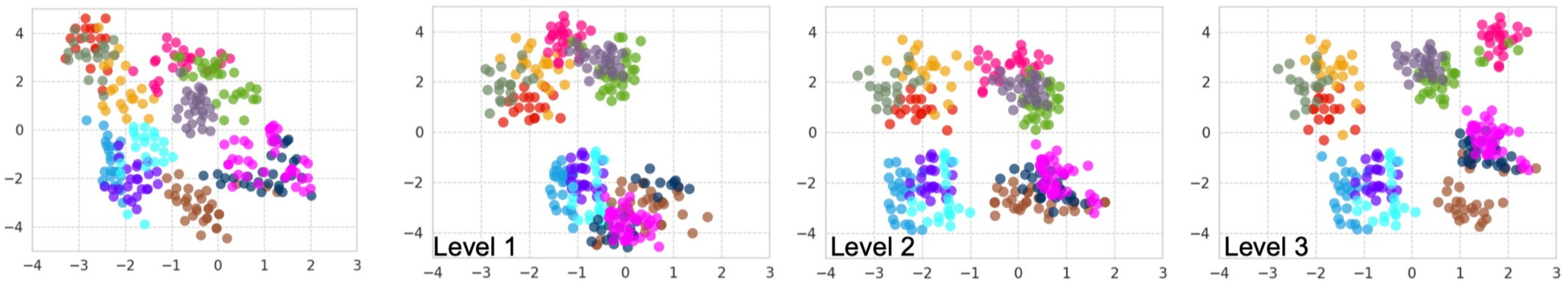
Visualization of SuperDisco



(a) SuperDisco (b) Meta-SuperDisco

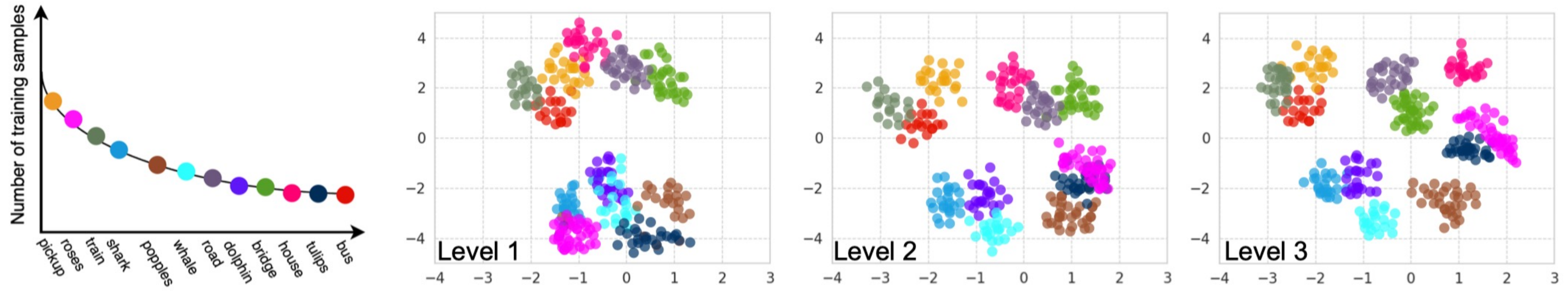
Our SuperDisco discovers super-classes hidden in each category, while Meta-SuperDisco discovers more accurate super-classes.

Visualization of refined features



(a) Original features

(b) Different levels refined features by the SuperDisco



(c) Training distribution

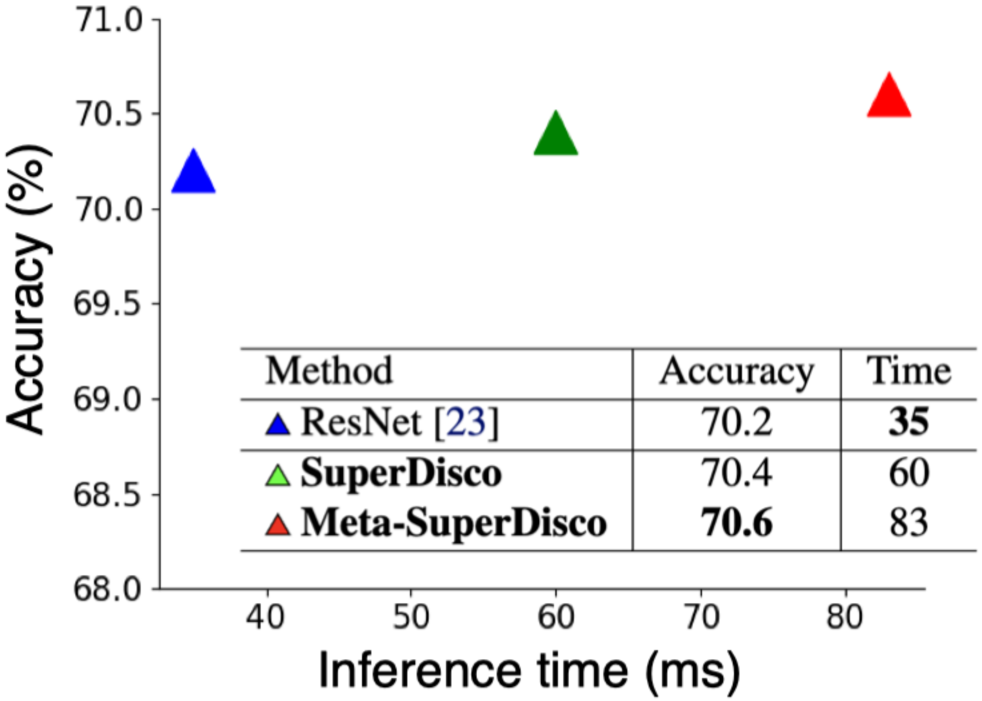
(d) Different levels refined features by the Meta-SuperDisco

Our SuperDisco (b) guides the original features on being clustered into the corresponding super-class space at different levels, while the Meta-SuperDisco (d) obtains more discriminative intra-class features.

Comparison with the state-of-the-art on long-tailed datasets

	Venue	*ImageNet-LT*				*Places-LT*				*iNaturalist*			
		Many	Medium	Few	All	Many	Medium	Few	All	Many	Medium	Few	All
Kang *et al.* [32]	ICLR 19	60.2	47.2	30.3	49.9	40.6	39.1	28.6	37.6	65.0	66.3	65.5	65.9
Kang *et al.* [31]	ICLR 21	61.8	49.4	30.9	51.5	-	-	-	-	-	-	-	68.6
He *et al.* [25]	ICCV 21	64.1	50.4	31.5	53.1	-	-	-	-	70.6	70.1	67.6	69.1
Li *et al.* [40]	CVPR 21	**66.8**	51.1	35.4	56.0	-	-	-	-	-	-	-	69.3
Samuel *et al.* [52]	ICCV 21	64.0	49.8	33.1	53.5	-	-	-	-	-	-	-	69.7
Alshammari *et al.* [1]	CVPR 22	62.5	50.4	41.5	53.9	-	-	-	-	71.2	70.4	69.7	70.2
Zhang *et al.* [75]	CVPR 21	61.3	52.2	31.4	52.9	40.4	***42.4***	30.1	39.3	69.0	71.1	***70.2***	70.6
Parisot *et al.* [47]	CVPR 22	63.2	52.1	***36.9***	54.1	39.7	41.0	***34.9***	***39.2***	-	-	-	-
Park *et al.* [48]	CVPR 22	***66.4***	**53.9**	35.6	***56.2***	-	-	-	-	**73.1**	***72.6***	68.7	***72.8***
This paper		66.1	***53.3***	**37.1**	**57.1**	**45.3**	**42.8**	**35.3**	**40.3**	***72.3***	**72.9**	**71.3**	**73.6**

Limitations



Method	Accuracy	Time
▲ ResNet [23]	70.2	**35**
△ **SuperDisco**	70.4	60
△ **Meta-SuperDisco**	**70.6**	83

Accuracy (%) vs. speed (ms) comparison with different methods on balanced CIFAR-100. SuperDisco has little impact on the performance of balanced datasets at the expense of increased inference time.

Conclusions

- Important observation: the imbalance factor of a super-class is considerably lower than the imbalance factor of the original class;
- We propose a super-class graph to rectify and refine the original features;
- Meta-learn super-class graph;
- State-of-the-art performance on four long-tailed datasets.

Thank you!