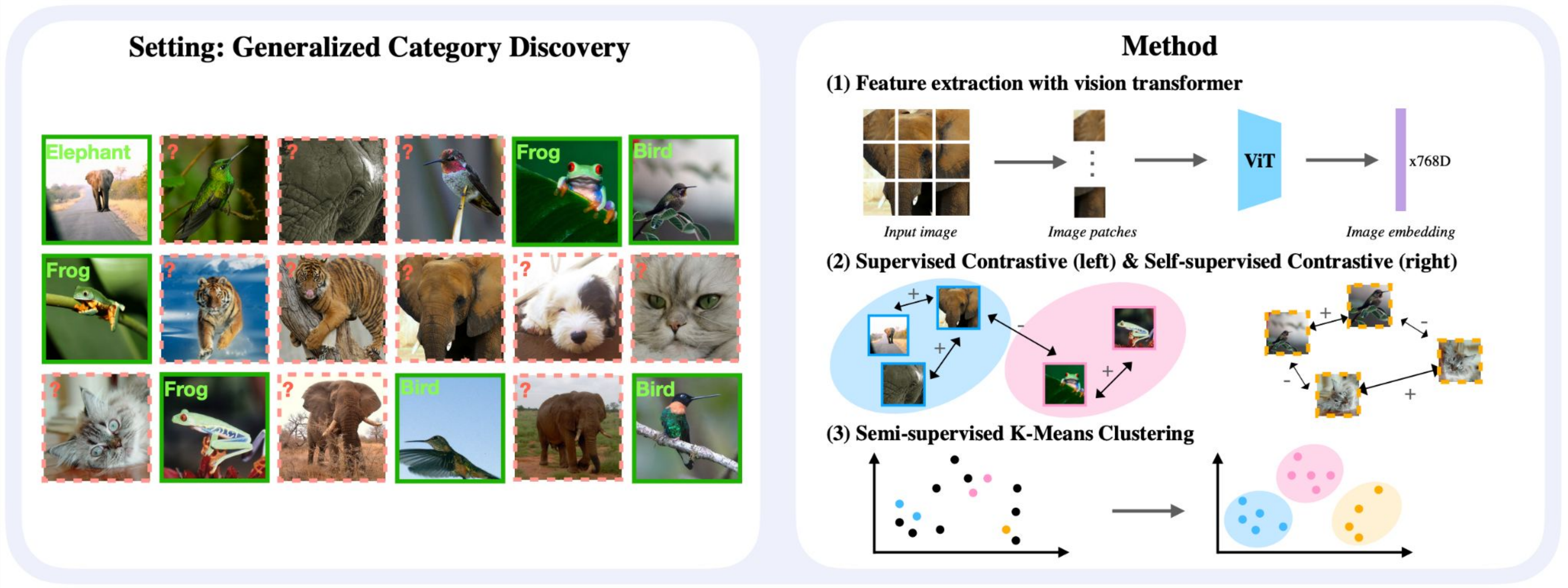
Parametric Classification for Generalized Category Discovery: A Baseline Study

Xin Wen[1*], Bingchen Zhao[2*], and Xiaojuan Qi[1]
[1]The University of Hong Kong, [2]University of Edinburgh
ICCV23

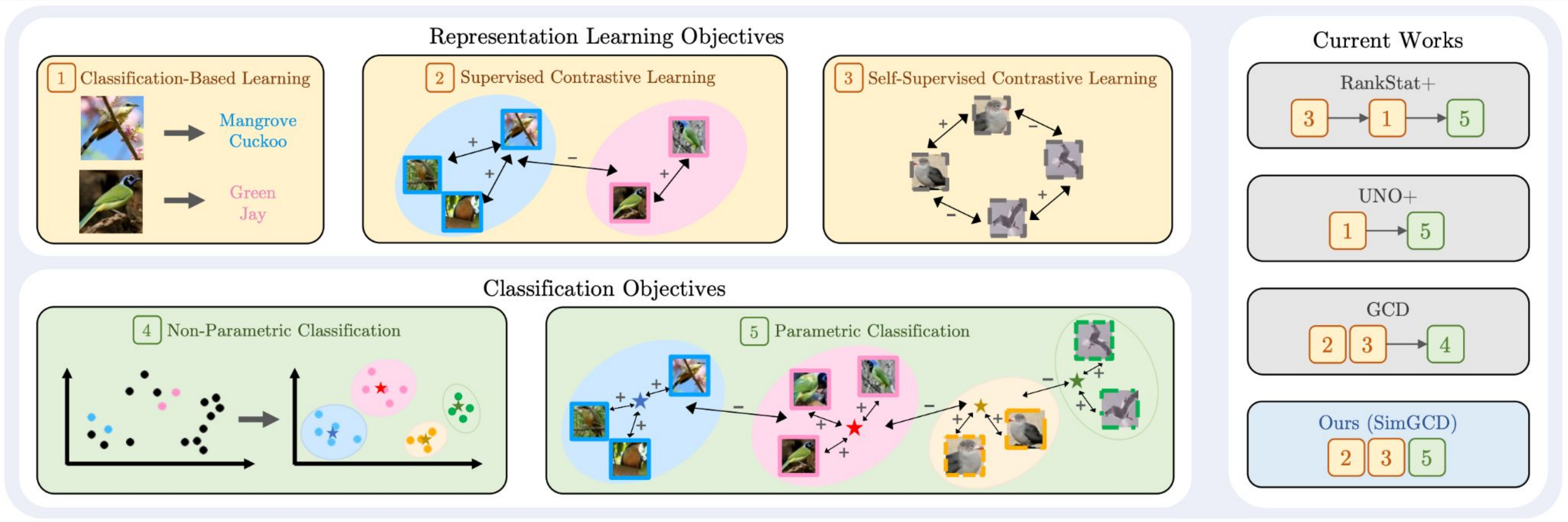
香 港 大 學
THE UNIVERSITY OF HONG KONG
THE UNIVERSITY of EDINBURGH

Generalized Category Discovery aims to recognise **novel** categories from **unlabelled** data using knowledge learned from labelled samples.

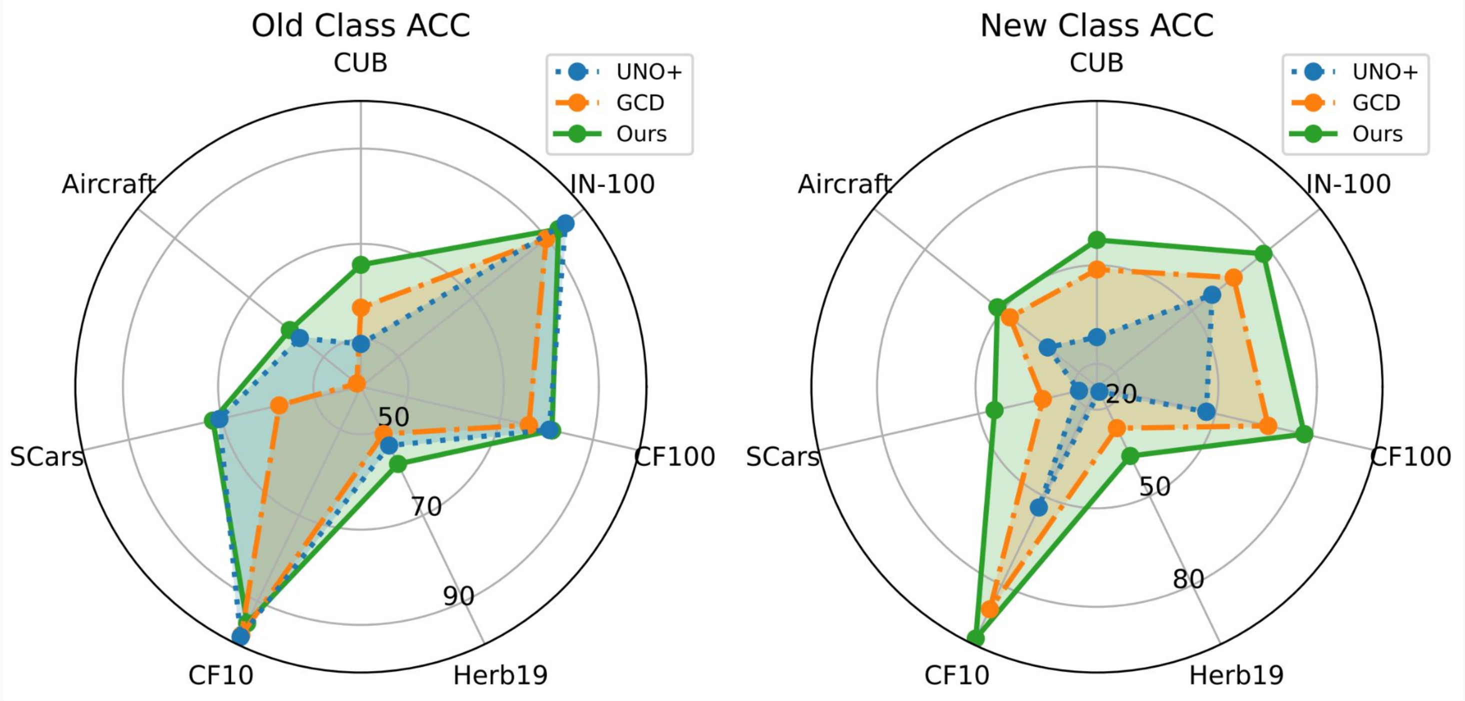


Introduction

Overview of current works: current SOTA is still semi-supervised k-means, and we target on **parametric classification**.



Prior parametric SOTA (UNO+) suffers from **over-fitting to seen ('Old') categories**.
But why?



On the Failure of Parametric Classification

Investigating into the failures of parametric cls. We validate the performance of different design choices under varying supervision qualities.

Representation Learning

- Follows GCD
- Supervised contrastive learning
- Self-Supervised contrastive learning

Classifier

- Prototypical classifier

$$\boldsymbol{l} = \frac{1}{\tau}(\boldsymbol{w}/||\boldsymbol{w}||)^\top (f(\boldsymbol{x})/||f(\boldsymbol{x})||)$$

Training Settings

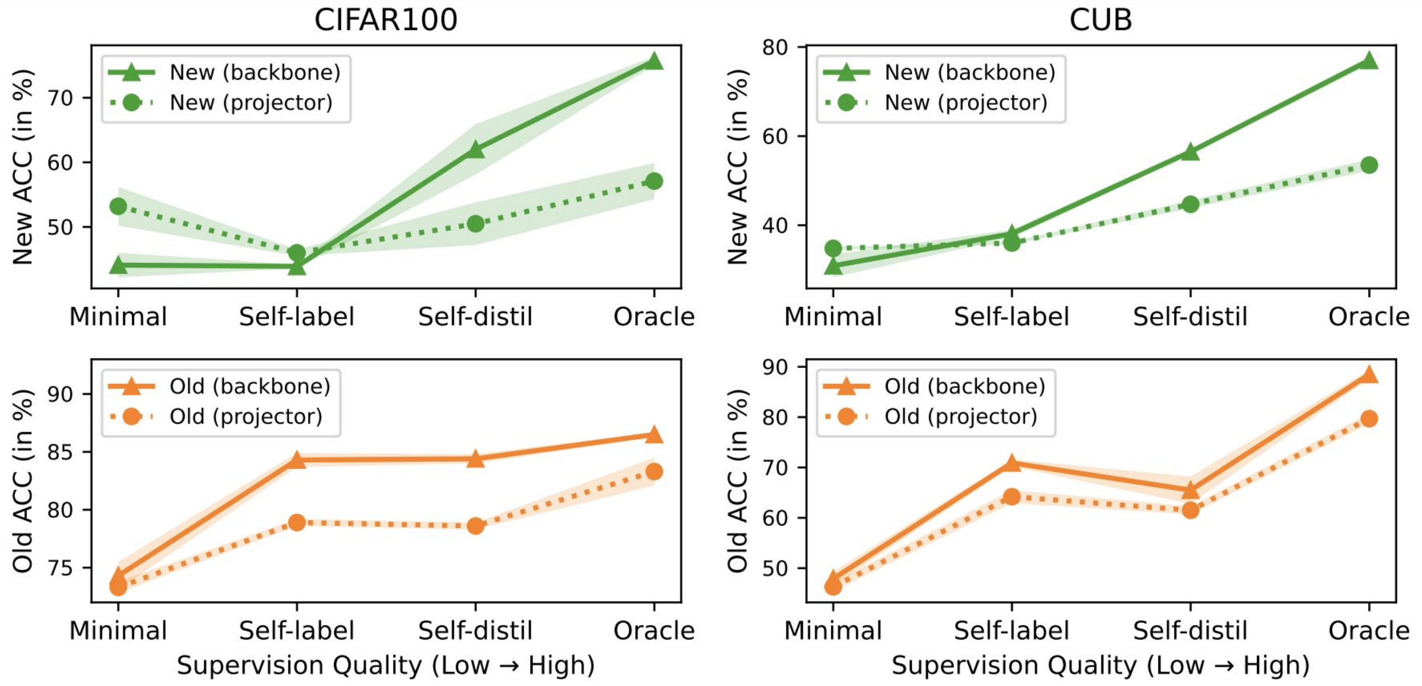
- Cross-entropy loss for classification
- Decouple classification from representation learning

Varied Supervision Quality

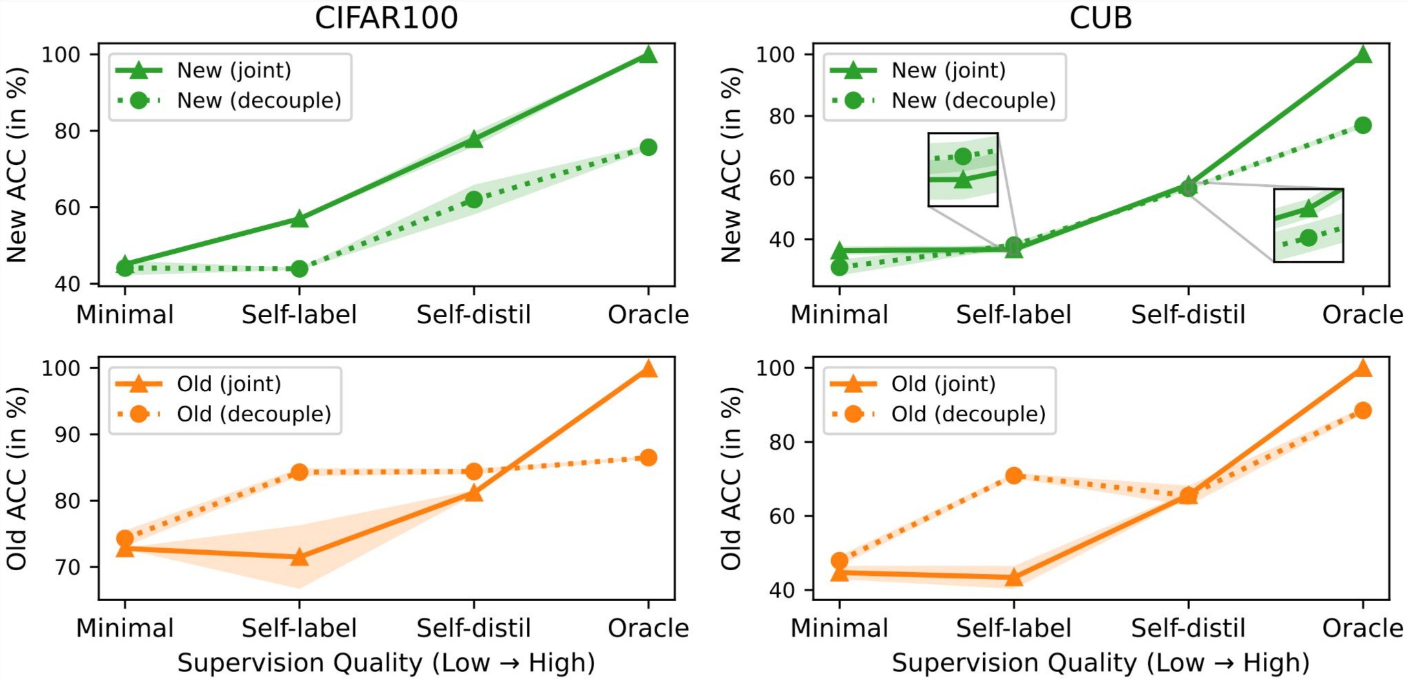
- Minimal (lower bound setting)
- Pseudo-labelling on unlabelled samples
  - With different pseudo-labelling strategies
  - i.e., self-labelling and self-distillation
- Oracle (upper bound setting)

Which feature space to build your classifier?

The *post-backbone* representations consistently benefit classification performance.



Decoupled or joint representation learning?
Guiding rep. learning with cls. objective can be helpful, *only when high-quality sup. is available.*



So what's wrong with UNO+'s pseudo labels?
The devil is in the biased predictions.

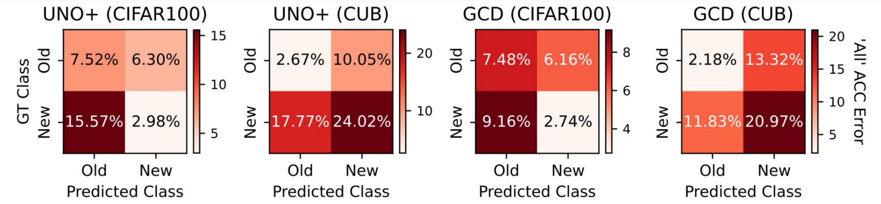


Figure 5. **Prediction bias between 'Old'/'New' classes.** We simplify the setting to binary classification and categorise the errors in 'All' ACC into four types. Both works, especially UNO+, are prone to make "False Old" predictions. In other words, the predictions are biased towards 'Old' classes, and many samples corresponding to 'New' classes are misclassified as an 'Old' class.

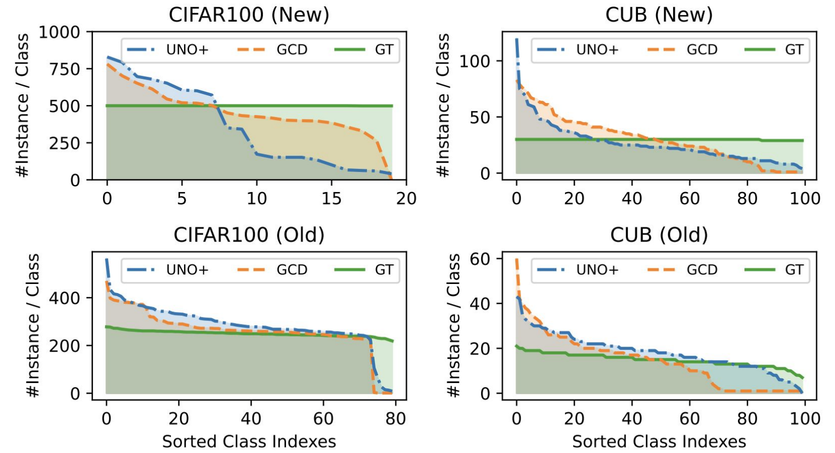
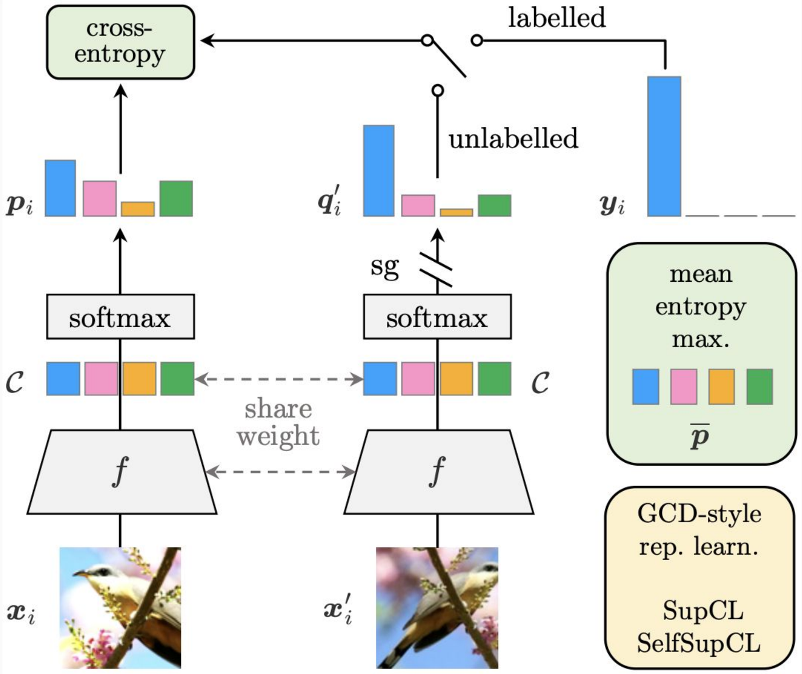


Figure 6. **Prediction bias across 'Old'/'New' classes.** We show the per-class prediction distributions. Both works, especially UNO+, are prone to make long-tailed predictions. In other words, across all classes, the predictions are unexpectedly long-tailed and biased towards the head classes.

SimGCD:
Our Simple Yet Strong Solution

We present a simple yet effective pseudo-labelling replacement that features **self-distillation** and **entropy regularisation**.



Supervised Classification Objective: $\mathcal{L}^s_{\text{cls}} = \frac{1}{|B^l|} \sum_{i \in B^l} \ell(\boldsymbol{y}_i, \boldsymbol{p}_i)$

Entropy

Self-Supervised Classification Objective: $\mathcal{L}^u_{\text{cls}} = \frac{1}{|B|} \sum_{i \in B} \ell(\boldsymbol{q}'_i, \boldsymbol{p}_i) - \varepsilon H(\overline{\boldsymbol{p}})$

Average Predictions: $\overline{\boldsymbol{p}} = \frac{1}{2|B|} \sum_{i \in B} (\boldsymbol{p}_i + \boldsymbol{p}'_i)$

Supervised Contrastive Learning: $\mathcal{L}^s_{\text{rep}} = \frac{1}{|B^l|} \sum_{i \in B^l} \frac{1}{|\mathcal{N}_i|} \sum_{q \in \mathcal{N}_i} - \log \frac{\exp(\boldsymbol{z}_i^\top \boldsymbol{z}'_q / \tau_c)}{\sum_i^{i \neq n} \exp(\boldsymbol{z}_i^\top \boldsymbol{z}'_n / \tau_c)}$

Self-Supervised Contrastive Learning: $\mathcal{L}^u_{\text{rep}} = \frac{1}{|B|} \sum_{i \in B} - \log \frac{\exp(\boldsymbol{z}_i^\top \boldsymbol{z}'_i / \tau_u)}{\sum_i^{i \neq n} \exp(\boldsymbol{z}_i^\top \boldsymbol{z}'_n / \tau_u)}$

Main Experiments

Our experiments cover all current GCD benchmarks that are **coarse/fine-grained**, **balanced/long-tailed**, or **small/large-scale**.

		Labelled		Unlabelled	
Dataset	Balance	#Image	#Class	#Image	#Class
CIFAR10 [27]	✓	12.5K	5	37.5K	10
CIFAR100 [27]	✓	20.0K	80	30.0K	100
ImageNet-100 [35]	✓	31.9K	50	95.3K	100
CUB [39]	✓	1.5K	100	4.5K	200
Stanford Cars [26]	✓	2.0K	98	6.1K	196
FGVC-Aircraft [29]	✓	1.7K	50	5.0K	50
Herbarium 19 [33]	✗	8.9K	341	25.4K	683
ImageNet-1K [13]	✓	321K	500	960K	1000

SimGCD reaches **state-of-the-art** performance on all benchmarks: fine-grained classification

	CUB			Stanford Cars			FGVC-Aircraft		
Methods	All	Old	New	All	Old	New	All	Old	New
k-means [28]	34.3	38.9	32.1	12.8	10.6	13.8	16.0	14.4	16.8
RS+ [20]	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2
UNO+ [16]	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2
ORCA [6]	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1
GCD [37]	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9
SimGCD	**60.3**	**65.6**	**57.7**	**53.8**	**71.9**	**45.0**	**54.2**	**59.1**	**51.8**
Δ	**+9.0**	**+9.0**	**+9.0**	**+14.8**	**+14.3**	**+15.1**	**+9.2**	**+18.0**	**+4.9**

SimGCD reaches **state-of-the-art** performance on all benchmarks: generic object recognition

	CIFAR10			CIFAR100			ImageNet-100		
Methods	All	Old	New	All	Old	New	All	Old	New
k-means [28]	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RS+ [20]	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [16]	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	**95.0**	57.9
ORCA [6]	81.8	86.2	79.6	69.0	77.4	52.0	73.5	92.6	63.9
GCD [37]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
SimGCD	**97.1**	95.1	**98.1**	**80.1**	**81.2**	**77.8**	**83.0**	93.1	**77.9**
Δ	**+5.6**	**-2.8**	**+9.9**	**+7.1**	**+5.0**	**+11.3**	**+8.9**	**+3.3**	**+11.6**

SimGCD reaches **state-of-the-art** performance on all benchmarks: more challenging datasets

	Herbarium 19			ImageNet-1K		
Methods	All	Old	New	All	Old	New
k-means [28]	13.0	12.2	13.4	-	-	-
RS+ [20]	27.9	55.8	12.8	-	-	-
UNO+ [16]	28.3	53.7	14.7	-	-	-
ORCA [6]	20.9	30.9	15.5	-	-	-
GCD [37]	35.4	51.0	27.0	52.5	72.5	42.2
SimGCD	**44.0**	**58.0**	**36.4**	**57.1**	**77.3**	**46.9**
Δ	+8.6	+7.0	+9.4	+4.6	+4.8	+4.7

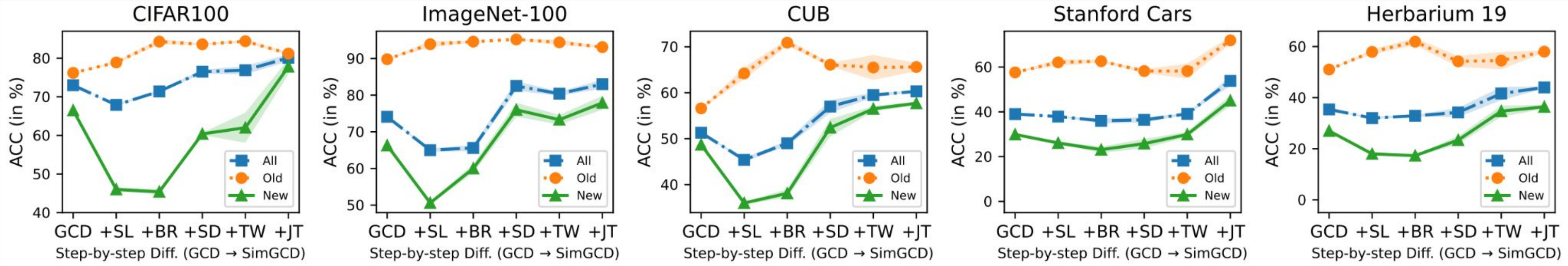
Also, notably **faster inference** since time for semi-supervised k-means is reduced.

Methods	CF100	CUB	Herb19	IN-100	IN-1K
GCD [37]	7.5m	9m	2.5h	36m	7.7h
SimGCD	1m	18s	3.5m	9.5m	0.6h

Table 5. Inference time over the unlabelled split.

Analytical Experiments

Step-by-step ablation study (GCD→SimGCD) shows **consistent benefit from gradually stronger pseudo-labels**.



SL: self-labelling, BR: post-backbone representation

SD: self-distillation, TW: teacher temperature warm-up, JT: joint training

Entropy regularisation shows notable benefit in **alleviating the prediction biases** between and within seen and novel categories.

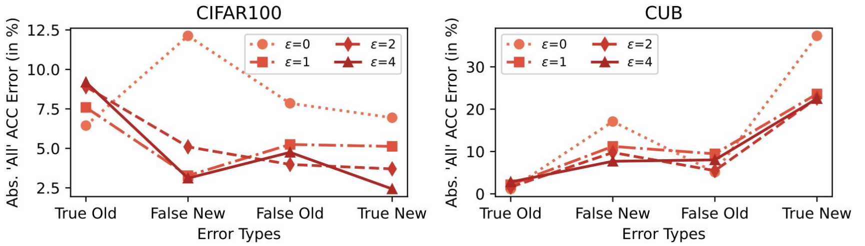


Figure 9. **Effect of entropy regularisation on four types of classification errors.** Appropriate entropy regularisation helps overcome the bias between 'Old'/'New' classes (see "False New" and "False Old", lower is better).

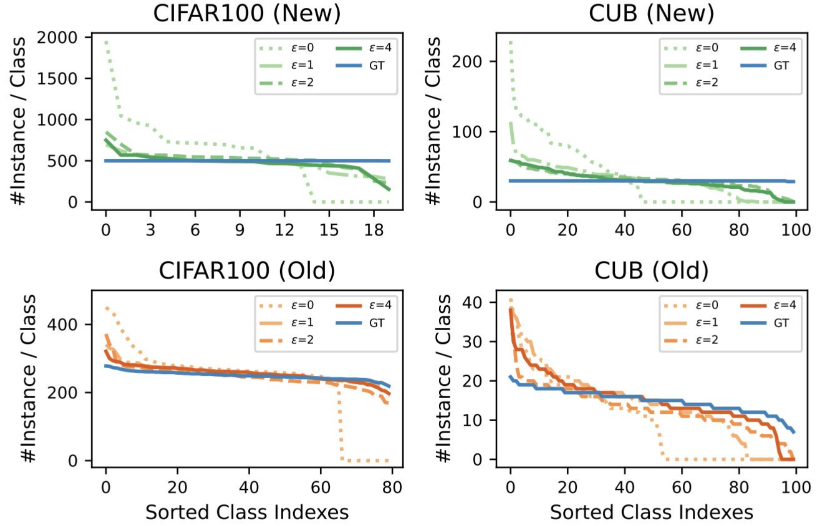
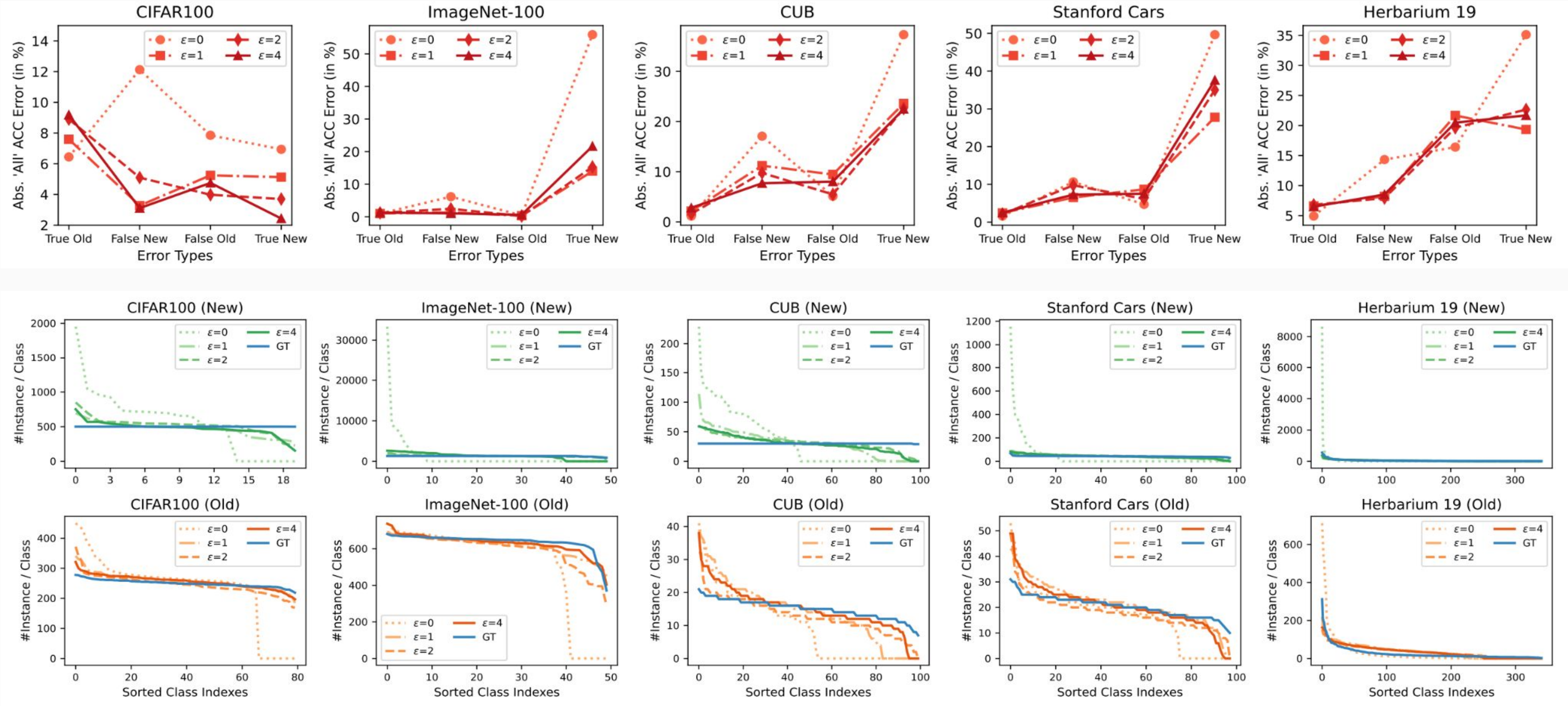


Figure 10. **Per-class prediction distributions with different entropy regularisation weights.** Proper entropy regularisation helps overcome the bias across 'Old'/'New' classes, and approach the GT class distribution.

And the benefit is also **consistent across multiple datasets**.



Take a closer look!

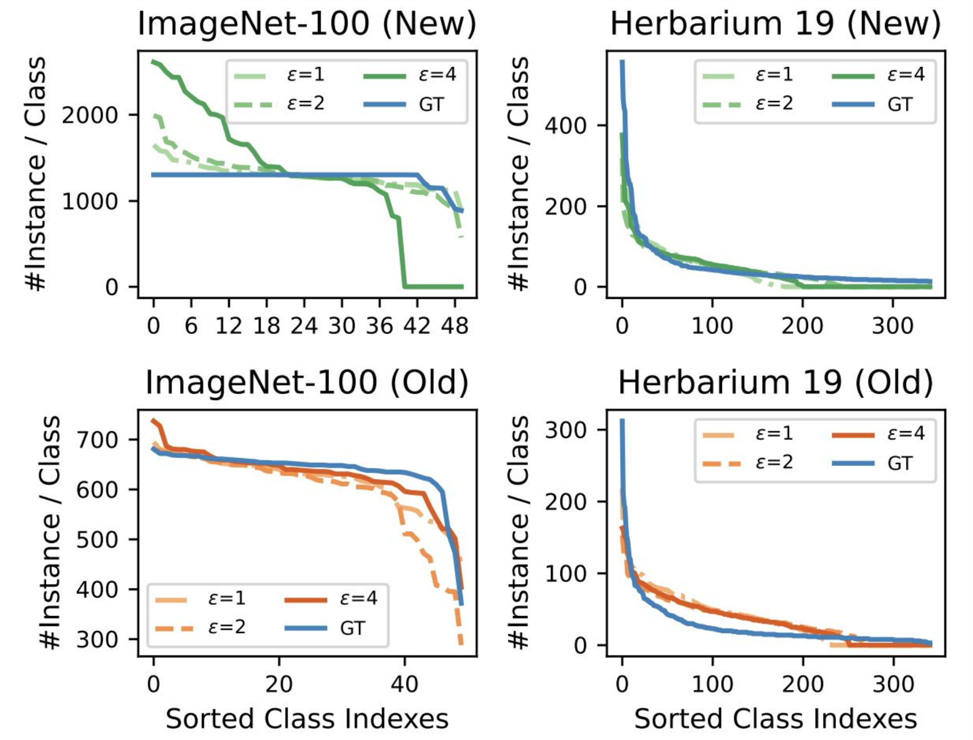
Though the regulariser enforces **uniform** predictions…

On **class-balanced** ImageNet-100:

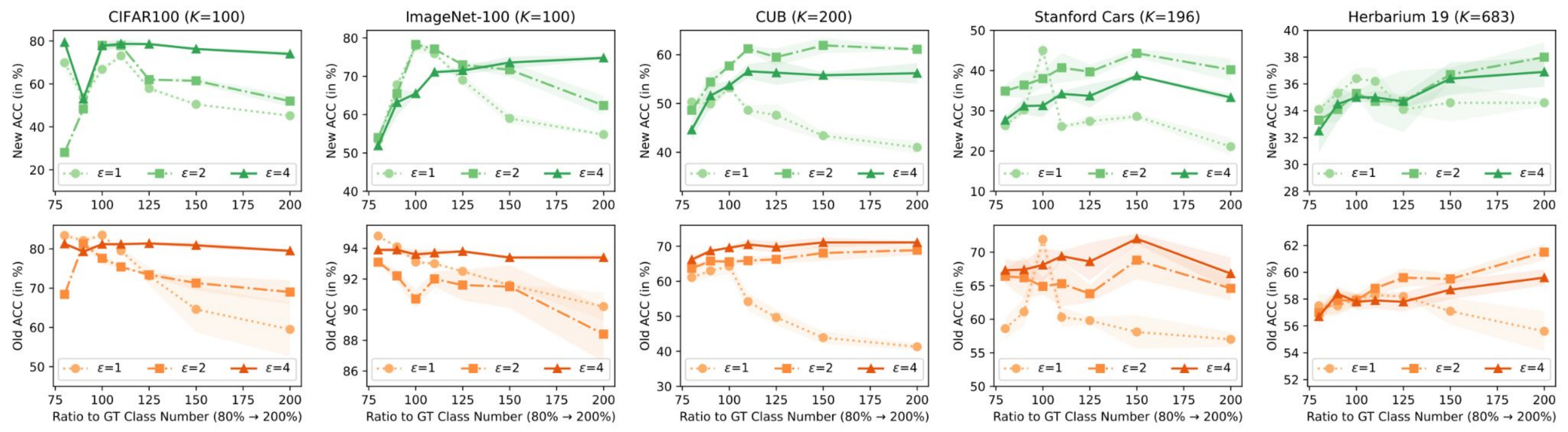
Over-regularisation could make the predictions **more biased**.

On **long-tailed** Herbarium 19:

Such regularisation could also **help fit long-tailed distribution**.



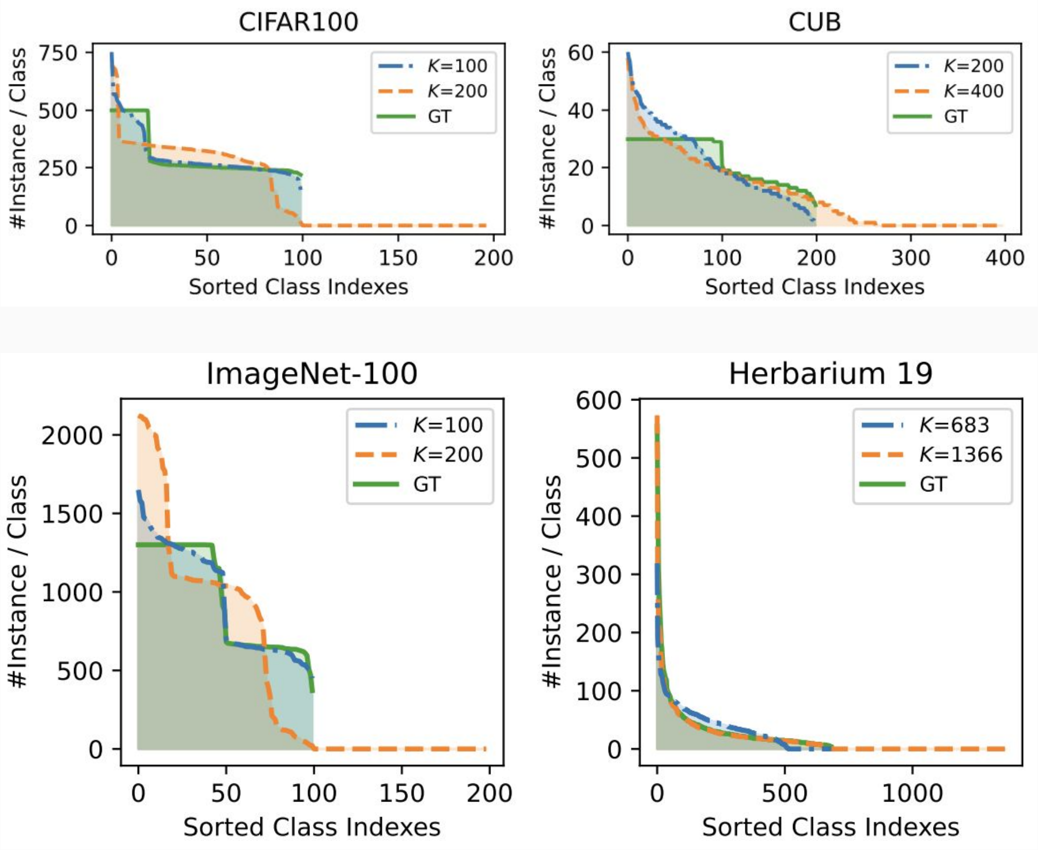
Entropy regularisation also enforces robustness to unknown class numbers, but over-regularisation could harm recognising 'New' classes under GT class numbers.



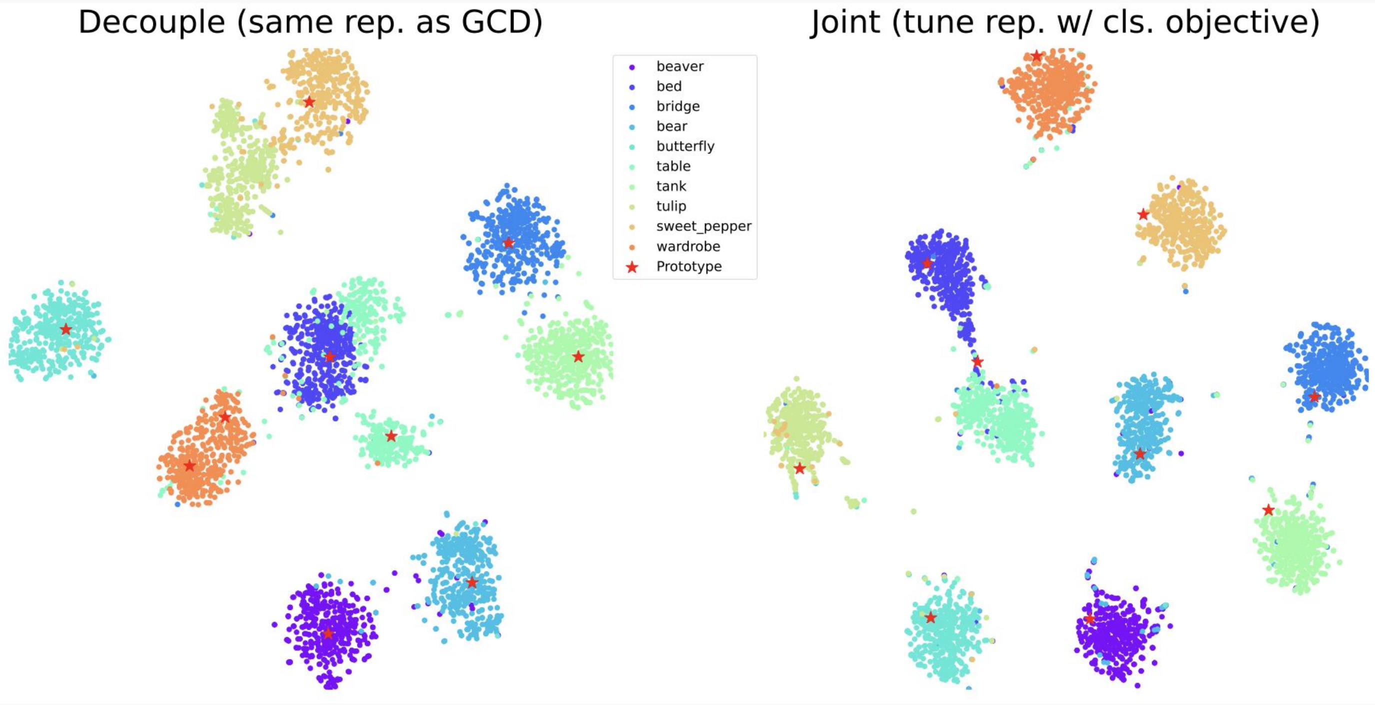
What makes for such robustness?

Our method identifies the criterion for ‘New’ classes, thus **keeping the number of active prototypes close to the ground-truth class number**.

A loose *K* greater than the ground truth may harm fitting the class-balanced ImageNet-100, but can help fit the long-tailed Herbarium 19.



Jointly supervising representation learning with a classification objective helps disambiguate (e.g., bed & table) and forms compacter clusters.



Limitations And Future Works

Representation Learning for An Open World

Could the features induced by a cat/dog classifier recognise table/bed, husky/beagle, or vise-versa?

Neural nets always spare no effort to find a short cut, thus **representations induced by closed-set classifiers easily bias to those predefined classes**, and novel classes could be hard to recognise.

Possible solutions:

- **Use more generalizable features**
  - E.g., self-supervised learning
- **Use weaker classification supervision**
  - E.g., SupCL rather than CE
  - Or even decouple cls. from rep.
- **Use regularisation terms**
  - To penalise possible biases
- **Keep in mind there are sth. out there**
  - E.g., use auxiliary prototypes
- **Make the class set big enough**
  - Thus evth. is in this closed set

Alignment to Human-Defined Categories

Could cat/dog labels help recognise table/bed, husky/beagle, or vise-versa?

In GCD, **human labels in seen categories implicitly define the metric for unseen ones**. E.g., cat/dog labels helps distinguish tiger/bear.

But what if seen/novel categories are of **different granularities**, in **different domains**, or the class set is so big and categories **overlap with each other** (e.g., ImageNet-22K)?

Further, could we **drop the matching process** between discovered clusters and text class names, or even **directly predicting the novel categories in the text space**?

Thanks!

Referenced papers:

[GCD] Sagar Vaze et al., Generalized Category Discovery, In *CVPR*, 2022.
[UNO] Enrico Fini et al., A Unified Objective for Novel Class Discovery, In *ICCV*, 2021.