Parametric Classification for Generalized Category Discovery: A Baseline Study

by Xin Wen* (HKU), Bingchen Zhao* (UoE), and Xiaojuan Qi (HKU)

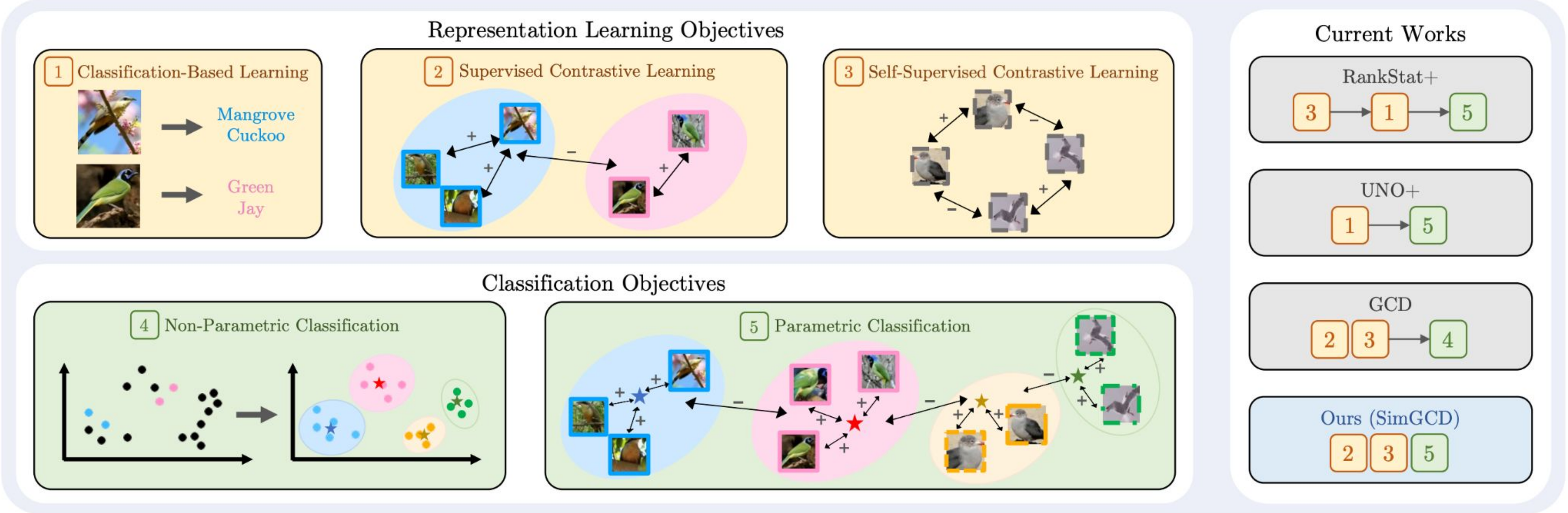
ICCV23

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

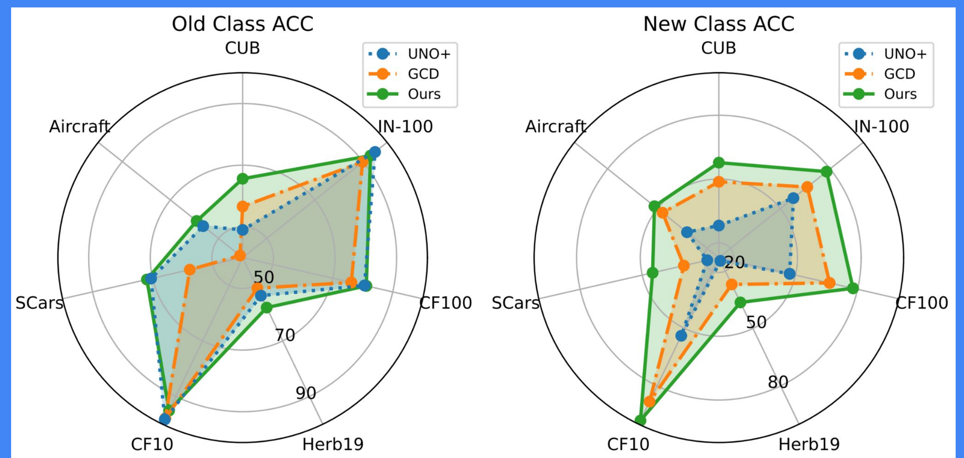
1. Setting and Background

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

Overview of current works: 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? (GOTO next col.)



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

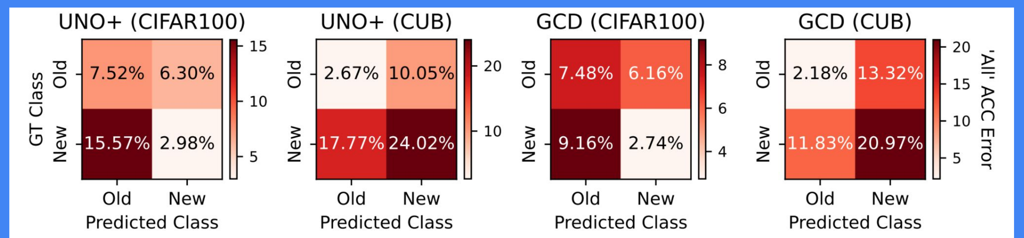


Figure 5. **Prediction bias between 'Old'/'New' classes.** We simplify the setting to binary classification and categorise the errors in 'All' ACC into four types. Both works, especially UNO+, are prone to make "False Old" predictions. In other words, the predictions are biased towards 'Old' classes,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

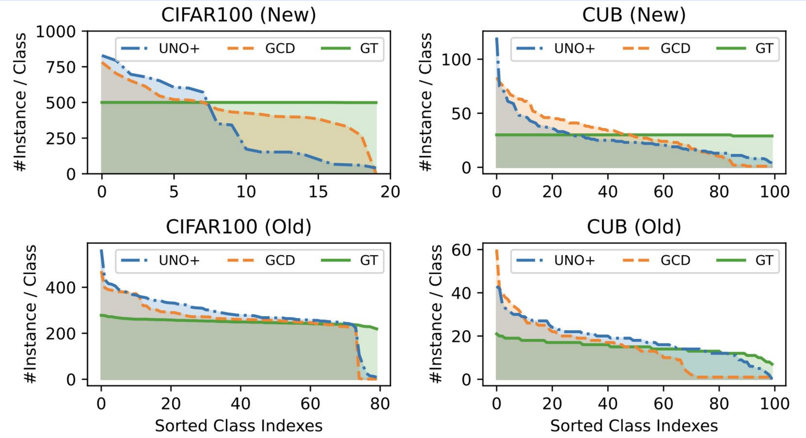
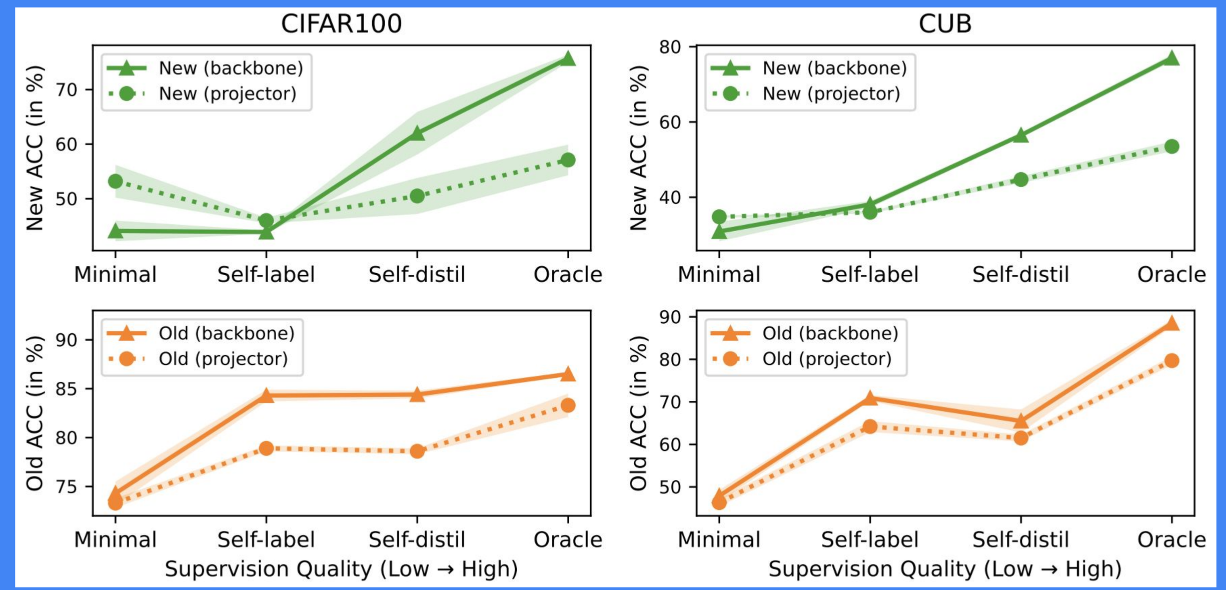


Figure 6. **Prediction bias across 'Old'/'New' classes.** We show the per-class prediction distributions. Both works, especially UNO+, are prone to make long-tailed predictions. In other words, across all classes, the predictions are unexpectedly long-tailed and biased towards the head classes.

2. On the Failure of Parametric Classification

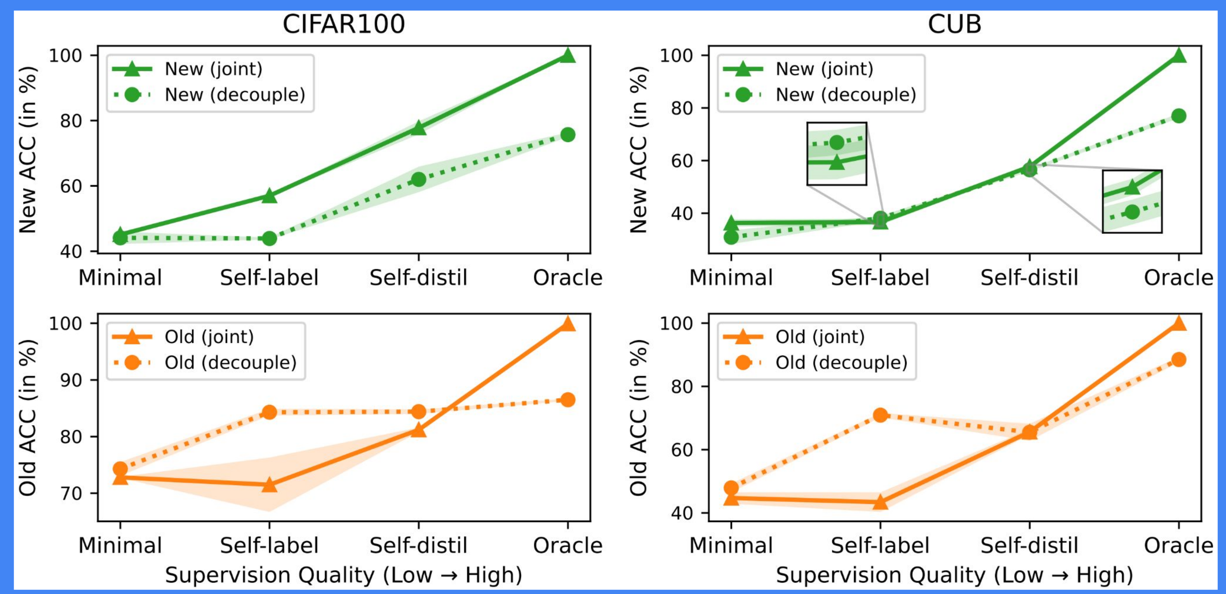
1) Which feature space to build your classifier?

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

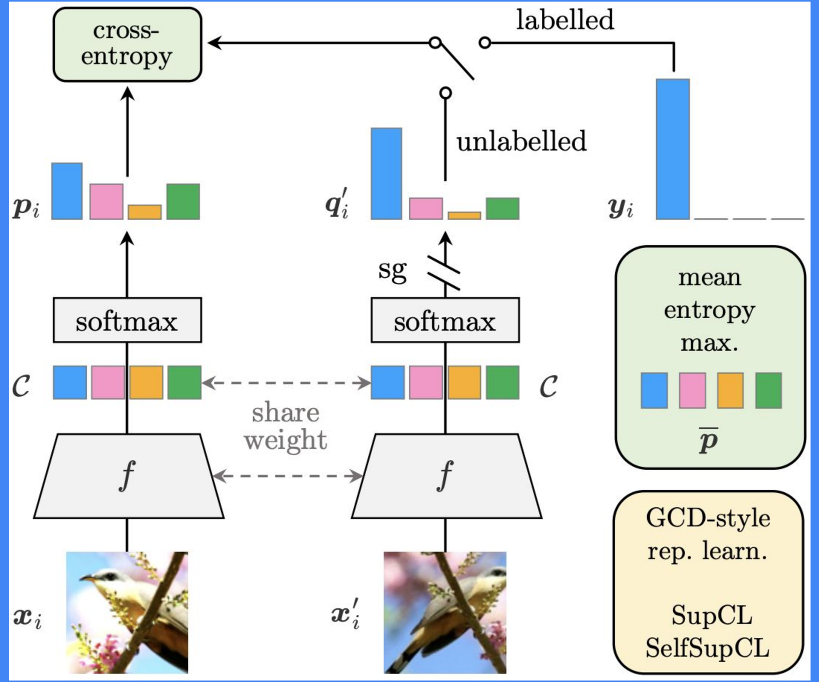


2) Decoupled or joint representation learning?

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



3. A Frustratingly Simple Yet Strong Baseline



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

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

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

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

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

4. Experiments

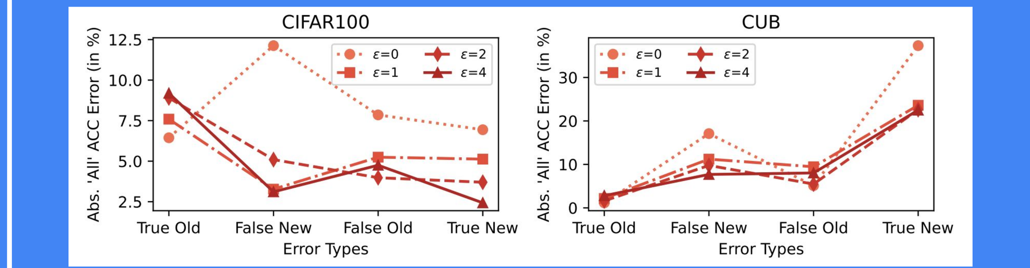
SOTA (⬆️~10%) over all current GCD benchmarks that are **coarse/fine-grained**, **balanced/long-tailed**, or **small/large-scale**.

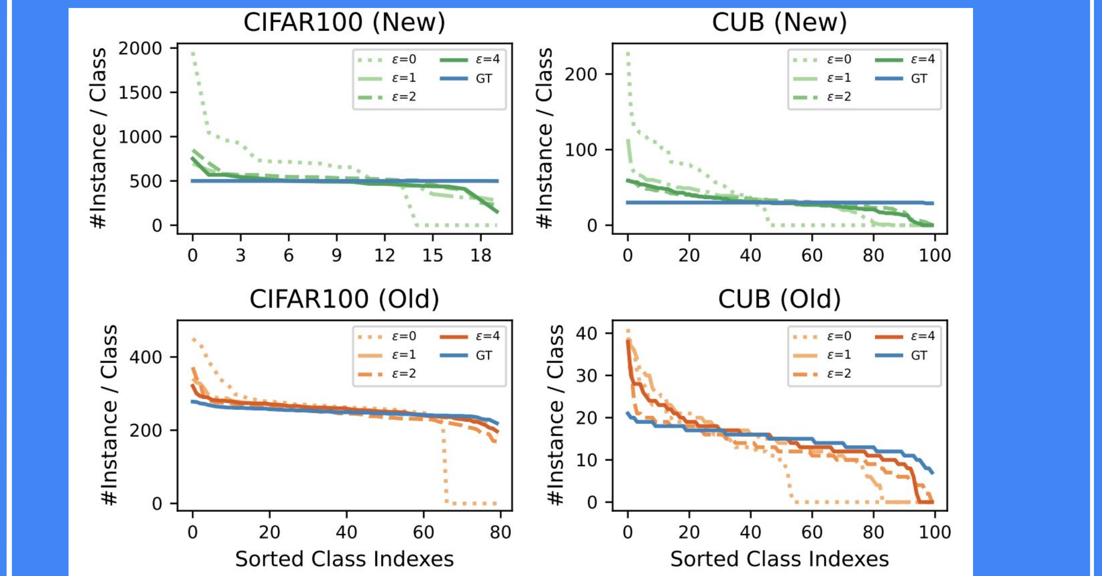
Methods	CUB			Stanford Cars			FGVC-Aircraft		
	All	Old	New	All	Old	New	All	Old	New
k-means [28]	34.3	38.9	32.1	12.8	10.6	13.8	16.0	14.4	16.8
RS+ [20]	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2
UNO+ [16]	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2
ORCA [6]	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1
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

Methods	CIFAR10			CIFAR100			ImageNet-100		
	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

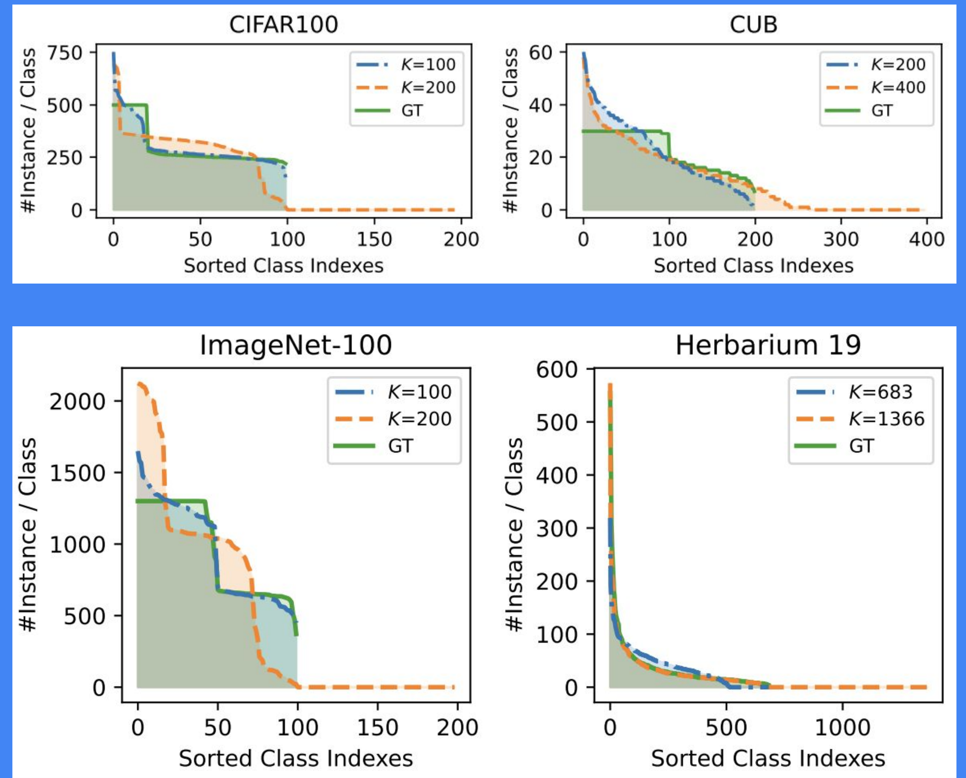
Methods	Herbarium 19			ImageNet-1K		
	All	Old	New	All	Old	New
k-means [28]	13.0	12.2	13.4	-	-	-
RS+ [20]	27.9	55.8	12.8	-	-	-
UNO+ [16]	28.3	53.7	14.7	-	-	-
ORCA [6]	20.9	30.9	15.5	-	-	-
GCD [37]	35.4	51.0	27.0	52.5	72.5	42.2
SimGCD	**44.0**	**58.0**	**36.4**	**57.1**	**77.3**	**46.9**
Δ	+8.6	+7.0	+9.4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

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





The model tends to **keep the number of active prototypes close to the GT class number**.



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