Self-Cooperation Knowledge Distillation for Novel Class Discovery

Yuzheng Wang, Zhaoyu Chen, Dingkang Yang,
Yunquan Sun, Lizhe Qi

Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology,
Fudan University, Shanghai, China

Outline

- 1. Background
- 2. Motivations
- 3. Proposed Method
- 4. Experimental Results
- 5. Conclusion

1. Background

The purpose of the Novel Class Discovery task:

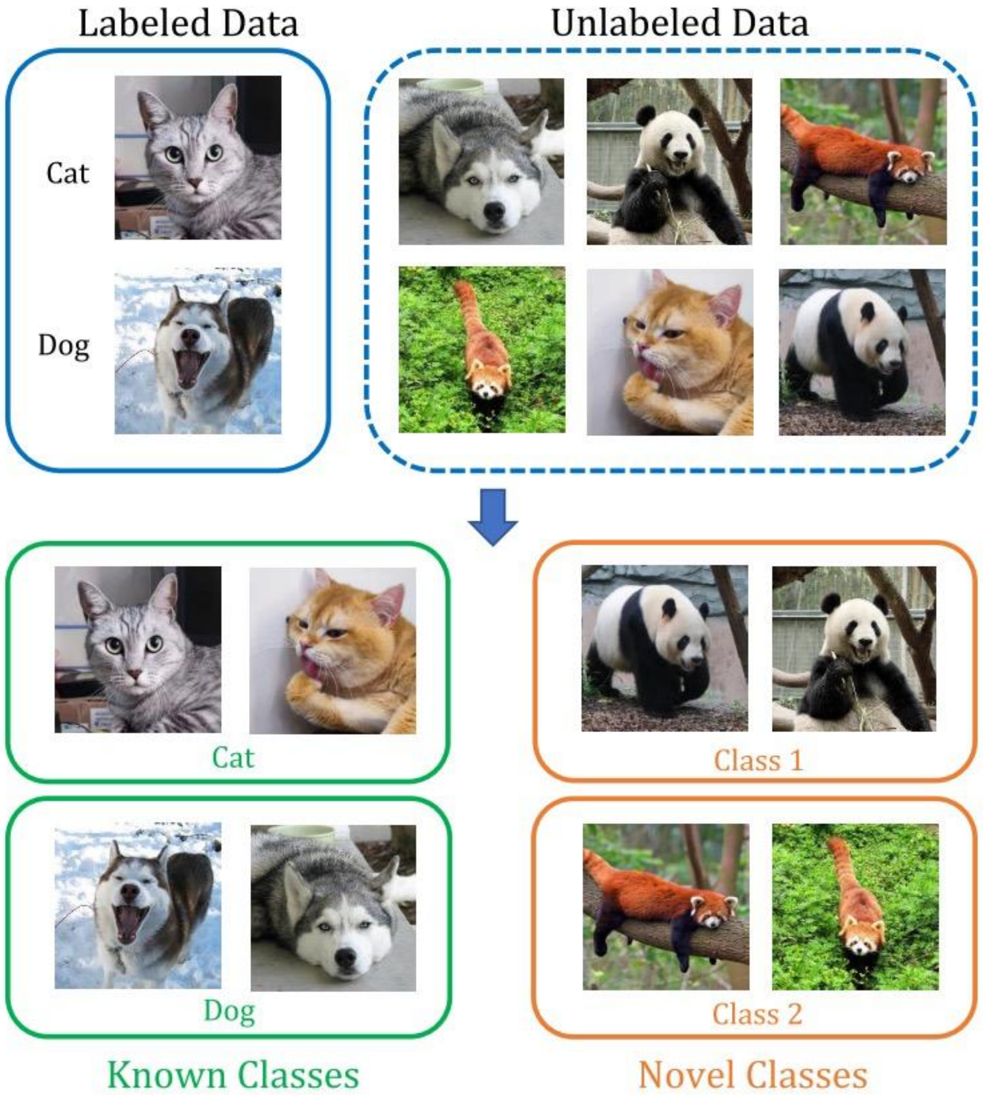
- ➢ Retain the model's cognition of known class samples
- ➢ Promote the model's learning of unfamiliar novel class samples

Training samples:

- ➢ Known class samples with labels
- ➢ Novel class samples without labels

Evaluation:

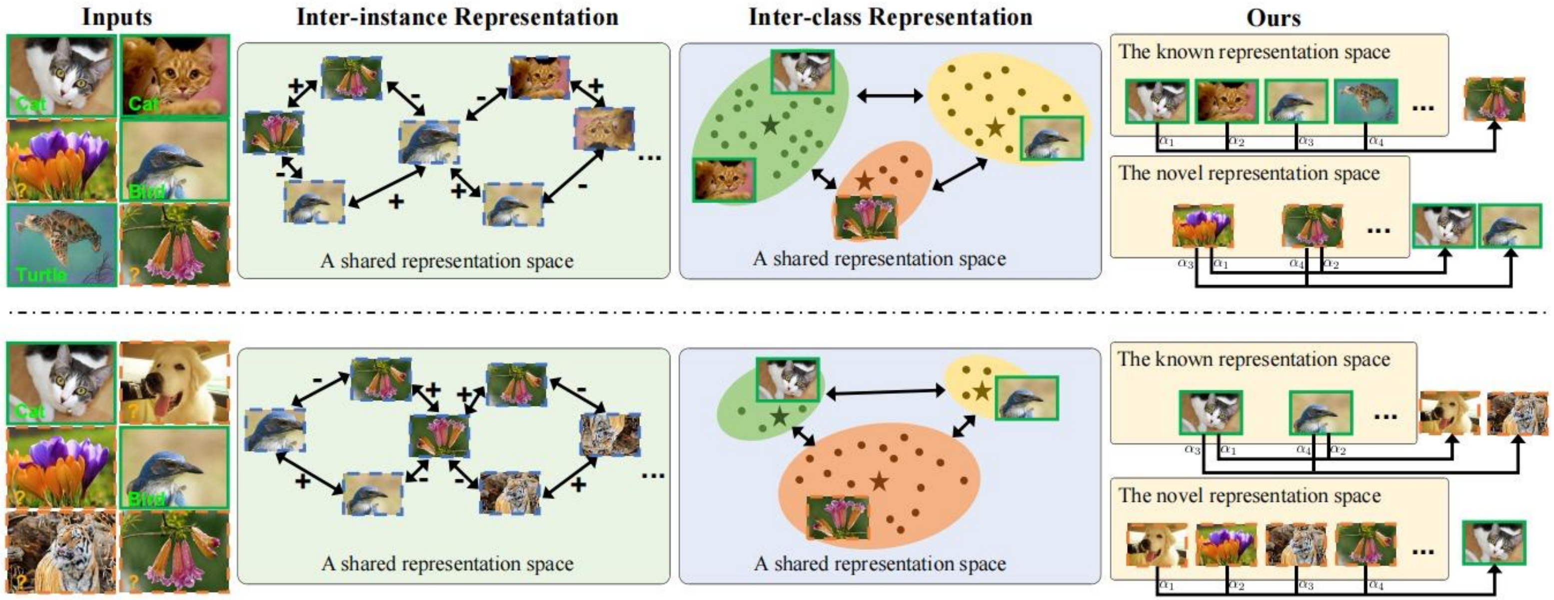
- ➢ Accuracy
- ➢ Clustering accuracy



2. Motivations

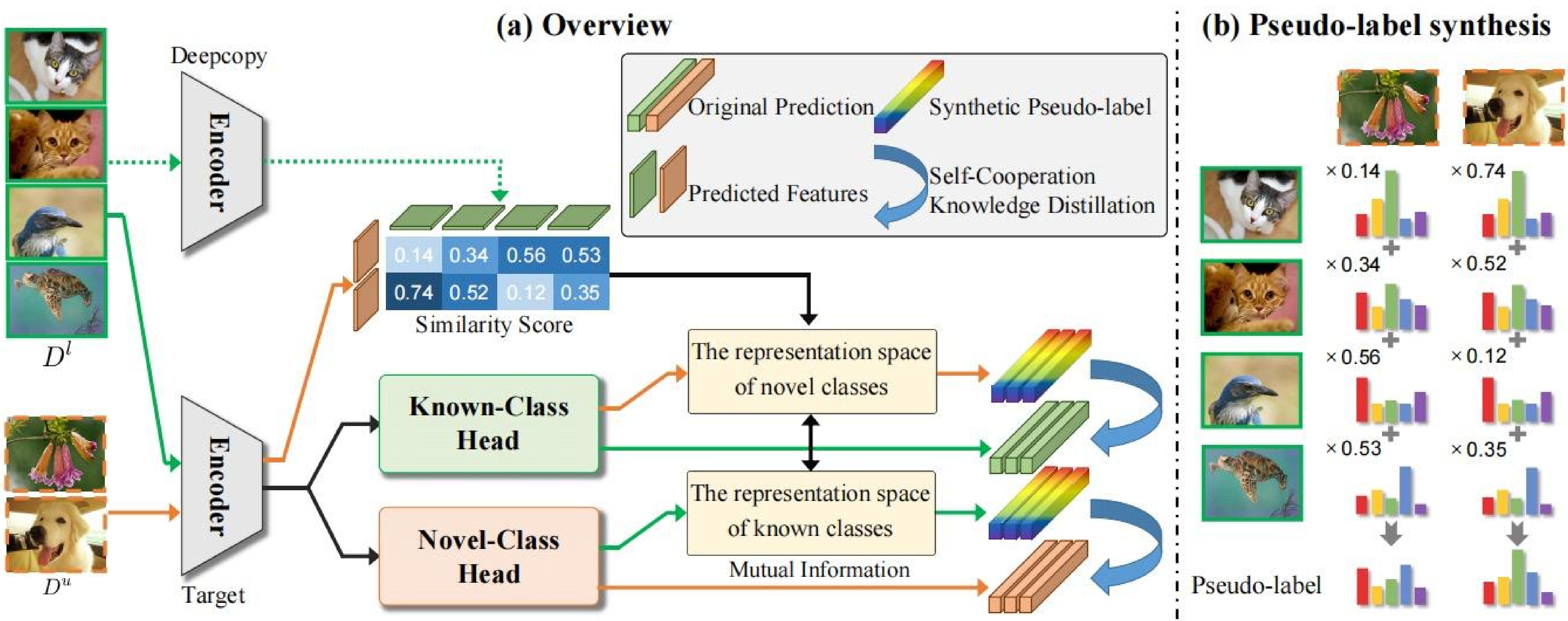
Motivations: Existing NCD methods focus on establishing a shared representation space for known and novel instances or classes. However, a long-neglected issue is the imbalanced number of samples from known and novel classes pushes the model toward the dominant party, making it challenging to trade-off between reviewing known classes and discovering novel classes.

Analysis:
Inter-instance methods aim to explore relationships among instances via contrastive learning, rank statistics, consistency and regu larization, and example mixing.
Inter-class methods aim to explore relationships among multiple classes.



Can we use all sample information to review known classes and discover novel classes simultaneously?

3. Proposed Method



3. Proposed Method

Preliminaries:

A mini-batch training set:

$$D^l = \left\{(\boldsymbol{x}_1^l, \boldsymbol{y}_1^l), \ldots, (\boldsymbol{x}_N^l, \boldsymbol{y}_N^l)\right\}$$

$$D^u = \{\boldsymbol{x}_1^u, \ldots, \boldsymbol{x}_M^u\}$$

$$\mathcal{Y}^l = \left\{1, \ldots, C^l\right\}$$

Two classification heads:

$$\boldsymbol{l}_i = \left[h^l(E(\boldsymbol{x}_i)), h^u(E(\boldsymbol{x}_i))\right], \boldsymbol{l}_i \in \mathbb{R}^{C^l+C^u}$$

Similarity Score Matrix:

$$\boldsymbol{v}^l = E^r(\boldsymbol{x}^l), \boldsymbol{v}^l \in \mathbb{R}^{N\times k}$$

$$\boldsymbol{v}^u = E(\boldsymbol{x}^u), \boldsymbol{v}^u \in \mathbb{R}^{M\times k}$$

$$S_{ij} = \cos\left(\boldsymbol{v}_i^l,\ \boldsymbol{v}_j^u\right),\ i = 1, \ldots, N \text{ and } j = 1, \ldots, M.$$

$$S = \mathrm{Norm}(S) = \frac{S}{|max\,\{S\}|}, S \in \mathbb{R}^{N\times M}.$$

Pseudo-Label Synthesis:

$$\hat{\boldsymbol{l}}_{uh}^u = \alpha \cdot S^T \cdot \boldsymbol{l}_{uh}^l,\ \ \hat{\boldsymbol{l}}_{uh}^u \in \mathbb{R}^{M\times C^u},$$

$$\hat{\boldsymbol{l}}_{kh}^l = \alpha \cdot S \cdot \boldsymbol{l}_{kh}^u,\ \ \hat{\boldsymbol{l}}_{kh}^l \in \mathbb{R}^{N\times C^l}.$$

Self Knowledge Distillation Objectives:

$$\mathcal{L}_{k\to n} = \frac{1}{M}\sum_1^M KL(\boldsymbol{l}_{uh}^u, \hat{\boldsymbol{l}}_{uh}^u), \quad \mathcal{L}_{n\to k} = \frac{1}{N}\sum_1^N KL(\boldsymbol{l}_{kh}^l, \hat{\boldsymbol{l}}_{kh}^l).$$

$$\mathcal{L}_{SCKD} = \mathcal{L}_{k\to n} + \mathcal{L}_{n\to k}.$$

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{SCKD},$$

4. Experimental Results

Performance Comparison

Table 2: Comparison with state-of-the-art methods on the unlabeled training subset, using task-aware evaluation protocol. **Bold** and underline numbers denote the best and the second best results, respectively.

Method	CIFAR10	CIFAR100-20	CIFAR100-50	ImageNet-100	Stanford Cars	CUB	Aircraft
k-means [31]	72.5±0.0	56.3±1.7	28.3±0.7	67.21	13.1±1.0	42.2±0.5	18.5±0.3
KCL [16]	72.3±0.2	42.1±1.8	-	-	-	-	-
MCL [17]	70.9±0.1	21.5±2.3	-	-	-	-	-
DTC [12]	88.7±0.3	67.3±1.2	35.9±1.0	-	-	-	-
RS+ [11]	91.7±0.9	75.2±4.2	44.1±3.7	-	36.5±0.6	55.3±0.8	38.4±0.6
OpenMix [60]	95.3	87.2	-	74.76	-	-	-
NCL [59]	93.4±0.5	86.6±0.4	52.7±1.2	-	43.5±1.2	48.1±0.9	43.0±0.5
Joint [20]	93.4±0.6	76.4±2.8	-	-	-	-	-
DualRank [58]	91.6±0.6	75.3±2.3	-	-	-	-	-
ComEx [54]	93.6±0.3	85.7±0.7	53.4±1.3	-	-	-	-
UNO [7]	93.3±0.4	90.5±0.7	62.3±1.4	79.56	49.8±1.4	59.2±0.4	52.1±0.7
GCD [40]	-	-	-	71.75	42.6±0.4	56.4±0.3	49.5±1.0
SimGCD [49]	-	-	-	81.92	50.2±0.5	62.3±0.4	53.6±1.1
IIC [26]	**99.1±0.0**	92.4±0.2	65.8±0.9	80.24	55.2±0.7	71.3±0.6	56.0±0.8
rKD [10]	93.5±0.3	91.2±0.1	65.3±0.6	80.94	53.5±0.8	65.7±0.6	55.8±0.9
SCKD	95.6±0.2	**92.6±0.6**	**68.2±0.4**	**82.18**	**56.8±0.7**	**73.1±0.4**	**56.5±0.7**

Table 3: Comparison with state-of-the-art methods on the testing subset, using task-agnostic evaluation protocol.

Method	CIFAR100-50			CIFAR100-80			Stanford Cars			CUB			Aircraft		
	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All
RS+ [11]	69.7	40.9	55.3	71.2	56.8	68.3	81.8	31.7	56.3	80.7	51.8	66.1	66.4	36.5	51.5
NCL [59]	72.4	25.7	49.0	72.7	41.6	66.5	83.5	24.4	53.4	79.8	13.1	46.3	62.8	26.5	44.6
UNO [7]	71.5	50.7	61.1	73.2	73.1	73.2	81.7	46.7	63.9	78.7	62.1	70.3	71.2	52.4	61.8
rKD [10]	**78.6**	59.4	69.0	75.2	76.4	75.4	83.9	51.3	67.3	81.1	67.5	74.2	72.2	55.2	63.7
IIC [26]	75.1	61.0	68.1	75.9	78.4	76.4	82.7	51.4	66.8	80.2	69.4	74.8	71.2	55.5	63.3
SCKD	77.9	**62.6**	**70.3**	**76.1**	**79.7**	**76.8**	**84.3**	**52.0**	**67.9**	**81.2**	**72.3**	**76.8**	**72.4**	**56.1**	**64.3**

Table 4: Experimental results with an increasing number of unlabeled classes on CIFAR100. Results are reported on the testing subset about both known and novel classes (averaged over 3 runs), using the task-agnostic evaluation protocol.

Method	Accuracy of known classes (%)							Clustering accuracy of novel classes (%)						
	Number of unlabeled classes							Number of unlabeled classes						
	20	30	40	50	60	70	80	20	30	40	50	60	70	80
UNO [7]	73.2	72.5	71.6	71.5	70.7	67.5	62.5	73.1	65.6	60.4	50.7	48.5	44.5	45.2
IIC [26]	75.9	75.5	75.4	75.1	73.7	69.5	64.7	78.4	69.5	64.7	61.0	57.4	54.2	51.2
rKD [10]	75.2	75.2	75.6	**78.6**	74.2	71.4	69.6	76.4	65.7	61.7	59.4	55.6	52.6	47.5
SCKD	**76.1**	**76.5**	**75.7**	77.9	**75.8**	**74.6**	**74.7**	**79.7**	**72.4**	**67.2**	**64.6**	**62.1**	**60.2**	**57.4**

4. Experimental Results

Ablation study about training losses

Table 6: Ablation study of our method for four dataset splits on the unlabeled training subset, using task-aware evaluation protocol. Results contain clustering accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) that are averaged over 3 runs.

ID	$\mathcal{L}_{k\to n}$	$\mathcal{L}_{n\to k}$	E^r	CIFAR100-20 ACC	CIFAR100-20 NMI	CIFAR100-20 ARI	CIFAR100-50 ACC	CIFAR100-50 NMI	CIFAR100-50 ARI	Stanford Cars ACC	Stanford Cars NMI	Stanford Cars ARI	CUB ACC	CUB NMI	CUB ARI
(1)	✗	✗	-	87.57	0.8361	0.7771	62.62	0.6861	0.4742	48.59	0.6913	0.3480	63.56	0.7869	0.5002
(2)	✓	✗	✓	91.57	0.8617	0.8333	65.59	0.6914	0.4985	54.19	0.7270	0.4029	70.50	0.8177	0.5719
(3)	✗	✓	✓	92.18	0.8695	0.8453	65.98	0.7069	0.5165	49.58	0.6970	0.3597	68.77	0.8040	0.5486
(4)	✓	✓	✓	**92.56**	**0.8754**	**0.8495**	**68.18**	**0.7128**	**0.5415**	**56.84**	**0.7388**	**0.4278**	**73.14**	**0.8262**	**0.5976**
(5)	✓	✓	✗	92.24	0.8713	0.8464	66.35	0.7094	0.5256	53.57	0.7168	0.3952	69.78	0.8074	0.5635

Ablation study about similarity score matrix

Table 7: Ablation study about the similarity score matrix S. All results are evaluated on the unlabeled training set.

Settings	Stanford Cars	CUB	FGVC-Aircraft
Average S	49.3	64.7	52.6
Random S	47.9	62.8	50.2
Our	**56.8**	**73.1**	**56.5**

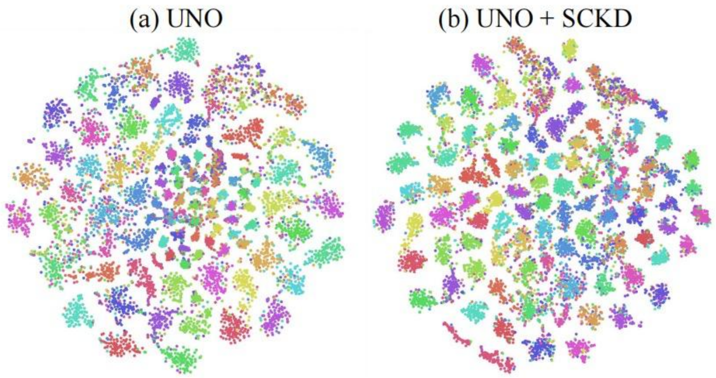


Fig. 4: t-SNE visualization for known and novel classes on CIFAR100-50 testing set, using task-agnostic evaluation protocol.

5. Conclusion

- We consider a practical but long-neglected challenge in the NCD task, i.e., the imbalanced number of samples from known and novel classes, making it difficult to balance reviewing known classes and discovering novel classes.
- We propose a simple yet effective SCKD method. SCKD can associate every sample for simultaneously reviewing known classes and discovering novel classes by building a cooperative learning paradigm.
- Extensive experiments on six benchmark datasets for novel class discovery show that the proposed method performs competitively and outperforms the state-of-the-art methods, demonstrating the effectiveness of SCKD.

Thanks !