Self-Cooperation Knowledge Distillation for Novel Class Discovery

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

Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology, Fudan University
yzwang20@fudan.edu.cn

Abstract. Novel Class Discovery (NCD) aims to discover unknown and novel classes in an unlabeled set by leveraging knowledge already learned about known classes. Existing works focus on instance-level or class-level knowledge representation and build a shared representation space to achieve performance improvements. However, a long-neglected issue is the potential imbalanced number of samples from known and novel classes, pushing the model towards dominant classes. Therefore, these methods suffer from a challenging trade-off between reviewing known classes and discovering novel classes. Based on this observation, we propose a Self-Cooperation Knowledge Distillation (SCKD) method to utilize each training sample (whether known or novel, labeled or unlabeled) for both review and discovery. Specifically, the model's feature representations of known and novel classes are used to construct two disjoint representation spaces. Through spatial mutual information, we design a self-cooperation learning method to encourage model learning from the two feature representation spaces. Extensive experiments on six datasets demonstrate that our method can achieve significant performance improvements, achieving state-of-the-art performance.

Keywords: Novel class discovery · Self-cooperation learning

1 Introduction

Deep learning, as a powerful and reliable tool, is increasingly popular in established and emerging fields of artificial intelligence [14, 26–28, 33, 49]. However, a considerable part of this success comes from the high-quality labeled training data for each class [9, 47, 48, 50], which is expensive [4, 42] or even impossible [41, 44]. For lower cost, an essential subject is using unlabeled data and training scalable models [3, 13, 24, 43, 45]. In this situation, Novel Class Discovery (NCD) task [16, 17] has been proposed. NCD aims to train scalable models to discover unknown classes in unlabeled sets with the help of semantic representations learned in known classes in labeled sets (ideally the models should

† Corresponding Author

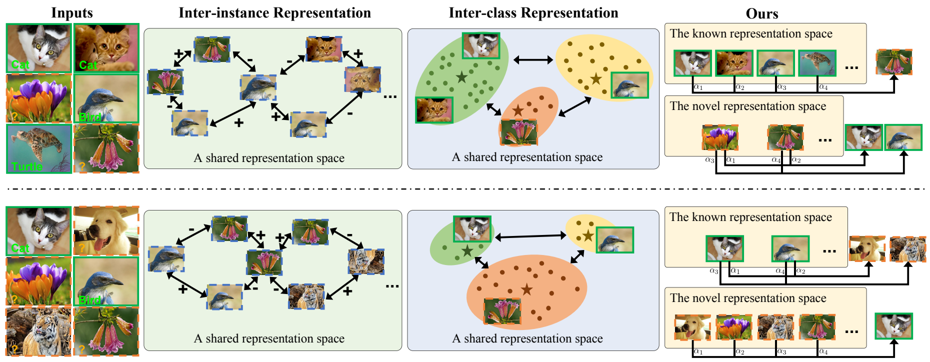

Fig. 1: Diagram of more known samples than novel samples (upper) & more novel samples than known samples (lower). **Inter-instance methods** aim to explore relationships among instances via contrastive learning (KCL [16], MCL [17], NCL [55], GCD [37], DCCL [32]), rank statistics (RS [11], DualRank [54]), consistency and regularization (DTC [12], ComEx [51], SimGCD [46]), and example mixing (OpenMix [56]). **Inter-class methods** aim to explore relationships among multiple classes (IIC [25], rKD [10]). **Our method** builds a self-cooperation pipeline for model learning.

not forget known classes). Note that this task is similar but different from Semi-Supervised Learning (SSL) [35] and Zero-Shot Learning (ZSL) [40]. The former assumes the labeled and unlabeled data share the same distribution and label space. The latter relies on additional descriptions (usually derived from human annotation) of unknown classes. With fewer dependencies and the advantages of open-world practicality, NCD has recently gained significant attention [36].

Most NCD methods focus on establishing a shared representation space for known and novel instances or classes. However, a long-neglected issue is that the imbalanced number of samples from known and novel classes pushes the model toward dominant classes, making it difficult to trade off between reviewing known classes and discovering novel classes. As shown in Fig. 1, we give a diagram of the mainstream inter-instance and inter-class NCD methods respectively. On the one hand, few novel samples and massive known samples lead to the pseudo-supervised information about the novel classes being overlooked due to a slight proportion. On the other hand, massive novel samples and few known samples lead to the information of the known classes being continuously marginalized and gradually approaching unsupervised learning. In either case, it makes the learned feature representations biased toward one or the other.

More intrigued, we evaluate the performance of the competitive NCD methods when dealing with different numbers of known and novel classes and samples on CIFAR100. As shown in Fig. 2, as the number of novel classes increases, the prediction accuracy of the current methods for known classes drops significantly. The performance of novel classes directed by known classes is also inevitably affected. Such results show that imbalanced samples affect existing methods.

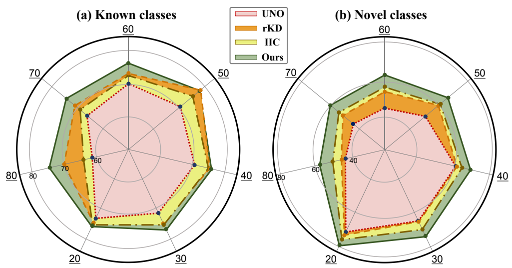

Fig. 2: Performance overview on testing subsets of CIFAR100 with an increasing number of unlabeled classes. The underline numbers represent the number of **novel classes**. We report the accuracy of both (a) known and (b) novel classes.

Based on this observation, a natural idea emerged: *Can we use every sample in the training set to review known classes and discover novel classes simultaneously*? As shown in Fig. 1, we establish and maintain non-overlapping representation spaces for known and novel classes respectively. On this basis, we propose a Self-Cooperation Knowledge Distillation (SCKD) method. Specifically, we treat the model's feature representations of known and novel classes as pairwise mutual information and encourage the model to perform self-cooperation learning between the two feature representation spaces from itself. For instance, when there are few known class samples, the model can use a large number of unlabeled samples of novel classes to help review known classes. Based on these, the model can cooperatively discover novel classes and review known classes. Specifically, the primary contributions are summarized as follows:

- 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.

2 Related Work

Novel Class Discovery. The Novel Class Discovery (NCD) task has two main goals: 1) maintaining the model's recognition performance for known classes and 2) promoting the model's discovery of class-disjoint novel classes. KCL [16] and MCL [17] use pairwise semantic similarity learning to distinguish whether instances belong to the same class and reconstruct semantic clusters to achieve knowledge transfer. Han *et al*. [12] standardize the description of the NCD task for the first time and propose the DTC method to observe and learn the temporal ensembling and consistency at different training stages to transfer knowledge from known to novel classes. Subsequent methods also focus on maintaining the same shared representation for known and novel samples and classes. RS [11] learns on labeled and unlabeled samples through self-supervision learning and then proposes ranking statistics to measure the similarity of two data in the representation space. DualRank [54] expands RS and proposes the dual ranking statistics to ensure knowledge transfer and consistency learning at global and local levels. NCL [55] introduces the explicit contrastive learning training objectives to learn the consistency between instances and their related views and neighbors in a shared representation space. OpenMix [56] refers to the Mixup [52] to mix labeled and unlabeled samples to fuse and distinguish known and novel classes in the representation space. Joint [20] uses category discrimination to augment contrastive learning and employs the Winner-Take-All (WTA) hashing algorithm on the shared representation space for cluster assignments. UNO [7] proposes a unified training objective to use labeled samples for known category review and unlabeled samples for novel category discovery. IIC [25] proposes to learn inter-class and intra-class constraints to establish instance-level and inter-class associations in the shared representation space to promote the distinction of novel instances and classes. rKD [10] introduces class relationship representation, transferring knowledge based on the predicted distribution of the pre-trained model on known classes to assist in learning novel classes. In addition, some related tasks are proposed. ComEx [51] focuses on generalized NCD and proposes using multiple regularizations for known and novel classes in the representation space. GCD [37] defines Generalized Category Discovery to relax the settings of NCD and proposes a contrastive learning and category number estimation method. On this basis, DCCL [32] and SimGCD [46] respectively optimize the contrastive learning paradigm and pseudo-label synthesis method through entropy regularization in GCD [37].

The existing methods do not fully consider the potentially imbalanced number of samples from known and novel classes, leading the model toward dominant classes. The model is encouraged to continuously learn novel classes leading to the eventual dilution and marginalization of known knowledge with the popularity of a shared representation space. Instead, we maintain distinct representation spaces and propose cooperative learning to alleviate this issue.

Self Knowledge Distillation. Self Knowledge Distillation (Self-Distillation) uses the knowledge of the target model itself to guide the optimization process, which can be regarded as special knowledge distillation [1]. BAN [8] models the temporal sequence in the model optimization process and uses the knowledge of the model at the previous moment to guide learning at the current moment. SAD [15] allows the model to utilize its attention maps as distillation targets for its lower layers for performance improvements. Zhang *et al*. [53] layer the network into sections and add multiple bottleneck layers to realize the guidance of the deeper layers to the lower layers. FRSKD [19] proposes a feature refinement self knowledge distillation method and guides the establishment of feature map distillation at the same level through integrating and refining each deep feature layer. DLB [34] proposes a training strategy to rearrange sequential sampling by limiting half of each mini-batch to coincide with the previous iteration. Meanwhile, the remaining half coincides with upcoming iterations.

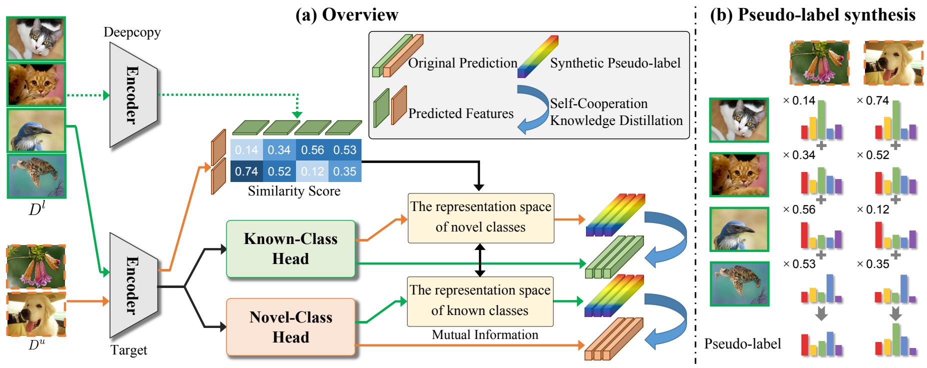

Fig. 3: The overview of (a) our proposed Self-Cooperation Knowledge Distillation and (b) a schematic diagram of pseudo-label synthesis. In (a), we define arrows with different colors and line shapes to distinguish data flows. First, a similarity score matrix is constructed for the spatial mutual information between labeled and unlabeled samples. Then, through the pseudo-label synthesis modules, known samples provide information for discovering novel classes and novel class samples provide information for reviewing known classes. In (b), the pseudo-supervised information comes from the logits prediction and weight from similarity scores. The self-distillation process combines discovering novel classes and reviewing known classes to overcome the potential imbalanced number of samples from known and novel classes.

Different from previous self knowledge distillation methods, we propose a customized self knowledge distillation technology for the NCD task, which aims to use the knowledge of all training samples of a mini-batch for both known class review and novel class discovery. Our distillation process attempts to link the training objectives from two subsets for cooperative learning rather than maintaining separate training objectives like previous methods. Notably, our method requires no additional storage overhead to retain training information for multiple training stages (epochs).

3 Methodology

3.1 Preliminaries

In the NCD task, a mini-batch training set contains a labeled subset of known classes $D^l = \{(\boldsymbol{x}_1^l, \boldsymbol{y}_1^l), \ldots, (\boldsymbol{x}_N^l, \boldsymbol{y}_N^l)\}$ and an unlabeled subset of novel classes $D^u = \{\boldsymbol{x}_1^u, \ldots, \boldsymbol{x}_M^u\}$. Each $\boldsymbol{x}_i^l$ in D^l or $\boldsymbol{x}_i^u$ in D^u is an input image. And $\boldsymbol{y}_i^l \in \mathcal{Y}^l = \{1, \ldots, C^l\}$ is the real categorical label of $\boldsymbol{x}_i^l$. Following the mainstream setting [7,25], the set of C^l labeled classes is assumed to be disjoint from the set of C^u unlabeled classes, and the number of latent unlabeled classes C^u is assumed to be known as a priori. The core purpose of NCD is to predict test instances as labeled classes or unlabeled clusters to learn matching composite label sets $\mathcal{Y} = \{1, \ldots, C^l, C^l + 1, \ldots, C^l + C^u\}$. Notably, there may be significant differences between C^l and C^u, making it difficult to review known classes and discover novel classes simultaneously. Based on the above setting, various methods build semantic knowledge and leverage the knowledge about labeled known classes to discover entangled latent classes.

Under the above settings, we propose a Self-Cooperation Knowledge Distillation (SCKD) method to associate the discovery of novel classes with the review of known classes (see Fig. 3) to alleviate the adverse effects of imbalanced number of samples from known and novel classes. In the following sections, we first introduce the SCKD process and the distillation training objective in Sec. 3.2. Then, we summarize the overall training objective in Sec. 3.3.

3.2 Self-Cooperation Knowledge Distillation

Our target model consists of a widely accepted architecture for NCD, *i.e.*, an encoder E and two classification heads h^l and h^u. The encoder E converts the input image into encoded feature vectors. The known-class head h^l is implemented as a linear classifier with C^l output dimensions. The novel-class head h^u is implemented as a multilayer perceptron (MLP) and a linear classifier with C^u output dimensions. Following existing methods [7, 10], we first conduct the standard supervised training on labeled known classes. In addition, we deepcopy the trained encoder E named replica encoder E^r to ensure pure and unbiased feature information of the known classes. Thereafter, the encoder E remains trainable while the replica encoder E^r remains frozen. The total logits prediction for the input sample $\boldsymbol{x}_i$ comes from the concat representation predicted by the two classification heads as $\boldsymbol{l}_i = \left[h^l(E(\boldsymbol{x}_i)), h^u(E(\boldsymbol{x}_i))\right], \boldsymbol{l}_i \in \mathbb{R}^{C^l+C^u}$. The probability distribution can be obtained as $\boldsymbol{p}_i = \sigma(\boldsymbol{l}_i/\tau)$, where σ is a softmax layer and τ is the softmax temperature.

For our SCKD method, based on the disjoint characteristic, we first divide the labeled and unlabeled samples into two groups to maintain separate representation spaces for known and novel classes respectively. Then, we build information bridges from one sample to all samples in another group through a similarity score matrix. Based on the score matrix, we obtain the soft pseudo labels about unlabeled samples from the known representation space and simultaneously obtain pseudo labels about labeled samples from the novel representation space. Finally, we design a self knowledge distillation process to achieve cooperative learning of novel class discovery and known class review for unbiased representation learning.

Similarity Score Matrix. To ensure the information balance between known and novel classes, we establish associations between labeled and unlabeled samples with a similarity score matrix so that all samples can contribute to discovering novel classes and reviewing known classes. The similarity score measures the semantic correlation between labeled and unlabeled samples. First, all labeled samples are fed to replica encoder E^r to get the set of predicted features as $\boldsymbol{v}^l = E^r(\boldsymbol{x}^l), \boldsymbol{v}^l \in \mathbb{R}^{N \times k}$, where k is the output dimension of the encoder. All unlabeled samples are fed to the target encoder E to get the predicted features as $\boldsymbol{v}^u = E(\boldsymbol{x}^u), \boldsymbol{v}^u \in \mathbb{R}^{M \times k}$. The similarity matrix S is calculated via the cosine similarity metric as follows:

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

Then, the score matrix is obtained by the maximum value for normalization as follows:

$$S = \mathrm{Norm}(S) = \frac{S}{|max\,\{S\}|}, S \in \mathbb{R}^{N \times M}. \tag{2}$$

It is worth noting that even if there is a serious imbalanced number of samples from the known and novel classes, marginal classes can still learn effectively from samples of dominant classes.

Pseudo-Label Synthesis. After constructing two disjoint representation space and calculating the similarity matrix, we aim to synthesize pseudo-labels to form self-supervised information and encourage the model to utilize the spatial mutual information to overcome the imbalanced number of samples from known and novel classes. All labeled and unlabeled samples are fed to the target model. We collect the labeled and unlabeled predictions at each classification head. For the novel-class head h^u, we extract information about labeled samples from the known representation space and collect the predicted logits $\boldsymbol{l}_{uh}^l = h^u(E(x^l)),\ \boldsymbol{l}_{uh}^l \in \mathbb{R}^{N \times C^u}$. Combined with the similarity score matrix S, we map the information of the known representation space into unlabeled pseudo-label representation, *i.e.*, the process of novel class discovery using sample knowledge of known classes (see Fig. 3b). The formalized mathematical expression can be summarized as follows:

$$\hat{\boldsymbol{l}}_{uh}^u = \alpha \cdot S^T \cdot \boldsymbol{l}_{uh}^l,\ \ \hat{\boldsymbol{l}}_{uh}^u \in \mathbb{R}^{M \times C^u}, \tag{3}$$

where α is a label smoothing coefficient used to adjust the degree of mapping intervention between known and novel classes. Similarly, for the known-class head h^l, we extract information about unlabeled samples from the novel representation space and collect the predicted logits $\boldsymbol{l}_{kh}^u = h^l(E(x^u)),\ \boldsymbol{l}_{kh}^u \in \mathbb{R}^{M \times C^l}$. The information of the novel-class representation space is mapped into labeled pseudo-label representation as follows:

$$\hat{\boldsymbol{l}}_{kh}^l = \alpha \cdot S \cdot \boldsymbol{l}_{kh}^u,\ \ \hat{\boldsymbol{l}}_{kh}^l \in \mathbb{R}^{N \times C^l}. \tag{4}$$

Self Knowledge Distillation Objectives. After obtaining the pseudo-label representation, we encourage models to learn consistent knowledge representations that connect known and novel representation spaces. To utilize the information of known class samples in the novel-class discovery process, we collect the original predictions $\boldsymbol{l}_{uh}^u = h^u(E(x^u)),\ \boldsymbol{l}_{uh}^u \in \mathbb{R}^{M \times C^u}$ of the novel-class head h^u for unlabeled samples. The pseudo-label representation from the spatial mutual information is expected to guide updating the original prediction, *i.e.*, a self-distillation process between labeled and unlabeled samples. The self-distillation loss for consistency constraints can be computed as follows:

$$\mathcal{L}_{k \to n} = \frac{1}{M} \sum_{1}^{M} KL(\boldsymbol{l}_{uh}^u, \hat{\boldsymbol{l}}_{uh}^u), \tag{5}$$

where KL denotes the Kullback-Leibler divergence loss. With the help of $\mathcal{L}_{k\to n}$, the target model can fully use all labeled and unlabeled samples to discover novel classes. Similarly, to utilize the information of novel class samples in the known-class review process, we constrain the original predictions $\boldsymbol{l}_{kh}^{l} = h^{l}(E(x^{l}))$, $\boldsymbol{l}_{kh}^{l} \in \mathbb{R}^{N\times C^{l}}$ of the known-class head h^{l} for labeled samples and pseudo-labels from the unlabeled sample $\hat{\boldsymbol{l}}_{kh}^{l}$ as follows:

$$\mathcal{L}_{n\to k} = \frac{1}{N}\sum_{1}^{N} KL(\boldsymbol{l}_{kh}^{l}, \hat{\boldsymbol{l}}_{kh}^{l}). \tag{6}$$

$\mathcal{L}_{n\to k}$ encourages the target model to integrate all labeled and unlabeled samples for reviewing known classes.

Notably, previous methods focus on using labeled samples to review known classes and unlabeled samples to discover novel classes. In contrast, our method builds a cooperative process to comprehensively use labeled samples (known classes) to guide the learning of unlabeled samples (discover novel classes). Simultaneously, it can use unlabeled samples (novel classes) to guide the learning of labeled samples (review known classes). Therefore, our self-distillation method can cope with the challenge of the imbalanced number of samples from known and novel classes. In addition, our method encourages the model to joint optimization of the known and novel classes, thus helping to avoid catastrophic forgetting of known classes and over-conservatism in novel class learning. Finally, we combine the two parts of the training objectives to form our overall Self-Cooperation Knowledge Distillation loss as follows:

$$\mathcal{L}_{SCKD} = \mathcal{L}_{k\to n} + \mathcal{L}_{n\to k}. \tag{7}$$

3.3 Overall Objective

In addition to the proposed distillation loss, we also use the cross-entropy loss as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N+M}\sum_{1}^{N+M} \boldsymbol{y}_i \log \boldsymbol{p}_i,\ \boldsymbol{y}_i \in \mathcal{Y}, \tag{8}$$

where $\boldsymbol{y}_i$ denotes label of image $\boldsymbol{x}_i$ and $\boldsymbol{p}_i$ denotes the probability distribution of the model for $\boldsymbol{x}_i$. Following existing baselines [7], $\boldsymbol{y}_i$ is one-hot-like ground truth for the labeled image. For the unlabeled image, $\boldsymbol{y}_i$ is the model's pseudo label after Sinkhorn-Knopp regularization [5], following mainstream NCD methods [7, 10]. Ultimately, our overall training objective is

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{SCKD}, \tag{9}$$

where β is the loss trade-off parameter.

Table 1: The details of dataset splits involved in the experiments.

Datasets		CIFAR10	CIFAR100-20	CIFAR100-50	ImageNet-100	Stanford Cars	CUB	FGVC-Aircraft
Known	# Images	25K	40K	25K	≈63.7K	≈4.0K	≈3.0K	≈3.3K
	# Classes	5	80	50	50	98	100	50
Novel	# Images	25K	10K	25K	≈63.4K	≈4.1K	≈3.0K	≈3.3K
	# Classes	5	20	50	50	98	100	50

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed method under NCD settings on both the generic image recognition datasets (including CIFAR10 [22], CIFAR100 [22] and ImageNet-100 [6]) and the fine-grained datasets (including Stanford Cars [21], CUB [39] and FGVC-Aircraft [31]). Following UNO [7], we divide CIFAR100 into two categories: 80/20 known/novel classes and 50/50 known/novel classes. ImageNet-100 denotes randomly sub-sampling 100 classes from the ImageNet dataset. The details of the dataset splits are shown in Tab. 1.

Evaluation Metrics. Following the mainstream NCD evaluation paradigm [7, 25], we conduct our experiments using both **task-aware** and **task-agnostic** evaluation protocols. For task-aware evaluation protocol, it is a priori information that each test sample comes from known or novel classes. For task-agnostic evaluation protocol, above information about classes is unknown. Following the widely recognized project, we use the accuracy measure for labeled samples and the average clustering accuracy for unlabeled samples. The average clustering accuracy is defined as:

$$\text{ClusterAcc} = \max_{perm \in P} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{y_i = perm(\hat{y}_i)\right\}, \tag{10}$$

where y_i is the ground-truth label and $\hat{y}_i$ is predicted clusters of a test sample $\boldsymbol{x}_i^u \in D^u$. P is the set of all permutations computed with the Hungarian algorithm [23]. We report the results over 5 runs for ImageNet-100 and 3 runs for other datasets in subsequent experimental evaluations.

Implementation Details. Following the mainstream NCD method [7, 11], by default, we adopt a two-stage training strategy, *i.e.*, first a supervised training stage, and then a discovery training stage. The replica encoder E^r is derived from the pre-trained model after the first stage and frozen in the second stage. For CIFAR10, CIFAR100 and ImageNet-100, we use ResNet18 [14] as our backbone network. We first pre-train on labeled samples for 100 epochs and then train on all labeled and unlabeled samples for 500 epochs, following UNOv2 [7]. The learning rate initially increases from 0.001 to 0.4 during the first 10 epochs and then decreases to 0.001 at 500 epochs with a cosine annealing schedule. We adopt the SGD optimizer with the momentum as 0.9, weight decay as 1.5×10^{-4}. For a fair comparison, we introduce the multi-head technique same with UNO [7] and

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	FGVC-Aircraft
k-means [30]	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 [56]	95.3	87.2	-	74.76	-	-	-
NCL [55]	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 [54]	91.6±0.6	75.3±2.3	-	-	-	-	-
ComEx [51]	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 [37]	-	-	-	71.75	42.6±0.4	56.4±0.3	49.5±1.0
SimGCD [46]	-	-	-	81.92	50.2±0.5	62.3±0.4	53.6±1.1
IIC [25]	**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**

employ four heads. For Stanford Cars, CUB, FGVC-Aircraft datasets, we utilize DINO pre-trained ViT-B-16 [2] as the backbone, use the output of `[CLS]` token with a dimension of 768 as the encoder, and only finetune its last block. Besides, we use the AdamW optimizer and employ two heads. The learning rate initially increases from 0.0001 to 0.001 during ten epochs and then decreases to 0.0001 at 100 epochs with a cosine annealing schedule. We first pre-train on labeled samples for 50 epochs and then train on all labeled and unlabeled samples for 100 epochs, following UNO [7], IIC [25], and rKD [10]. For all datasets, the softmax temperature τ is set to 0.1, the temperature of knowledge distillation is 1, the hyperparameters α is 0.1, and β is 0.5. Besides, we set the batch size to 512 and apply the multi-crop strategy (*e.g.*, random crop, flip, color jittering, and grey-scale), following UNO [7], IIC [25] and rKD [10]. For the Sinkhorn-Knopp algorithm, we inherit the hyperparameters $n\_iter = 3$ and $\epsilon = 0.05$.

4.2 Performance Comparison

We compare our method with state-of-the-art NCD models, including: KCL [16], MCL [17], DTC [12], RS+ [11], OpenMix [56], NCL [55], Joint [20], DualRank [54], UNO [7], IIC [25], and rKD [10]. Furthermore, we reproduce the training objectives in GCD methods under NCD settings, including ComEx [51], GCD [37], and SimGCD [46].

Training Subset. We first report the average clustering accuracy on the training split of the unlabeled subset, using the task-aware evaluation protocol (a common practice in the literature [7,12]). As shown in Tab. 2, our SCKD method outperforms the other SOTA methods on most benchmarks, except CIFAR10. We suspect that a possible reason is that the data distribution in CIFAR10 is relatively simple, and the semantic information is not rich enough, which may

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 [55]	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	<u>69.0</u>	75.2	76.4	75.4	<u>83.9</u>	51.3	<u>67.3</u>	<u>81.1</u>	67.5	74.2	<u>72.2</u>	55.2	<u>63.7</u>
IIC [25]	75.1	<u>61.0</u>	68.1	<u>75.9</u>	<u>78.4</u>	<u>76.4</u>	82.7	<u>51.4</u>	66.8	80.2	<u>69.4</u>	<u>74.8</u>	71.2	<u>55.5</u>	63.3
SCKD	<u>77.9</u>	**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 [25]	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**

reduce the effectiveness of the semantic similarity-based module in SCKD. In addition, for other semantically rich benchmark datasets, our method can achieve stable gains (higher average performance and smaller standard deviation).

Testing Subset. We also compare the performance of the testing subset with labeled and unlabeled samples, using the task-agnostic evaluation protocol. The results are shown in Tab. 3. Our method achieves the best results on almost all benchmarks for both known and novel classes. A well-founded finding is that our method significantly improves the model's performance for novel classes. In NCD tasks, the model is often more difficult to learn novel classes (from the test results). Previous methods often only used unlabeled samples for novel class learning, losing the information of labeled samples. In contrast, our SCKD method additionally incorporates information from labeled samples without introducing additional forward overhead, and the proposed cooperative learning paradigm also helps both known and novel class learning.

Varying the Number of Clusters. To compare the performance of different NCD methods in dealing with the imbalanced number of samples from known and novel classes, we set up a set of experiments with different proportions of the number of known and novel classes to approximate the situation of information imbalance between known and novel classes on CIFAR100. Specifically, we divide the training set into seven ways, from 80/20 known/novel classes to 20/80 known/novel classes, and collect the prediction performance of various methods on the testing set for known and novel classes. The results are shown in Tab. 4. We observe that as the number of unlabeled classes gradually increases ($20 \rightarrow 80$), existing methods gradually ignore the information of known classes, especially when they only account for a small proportion. An interesting phenomenon is that the model's performance for the novel classes also drops significantly even if the number of samples of the novel classes increases. We think a crucial reason is that the learning of novel classes relies on the knowledge of known classes, so the forgetting of known class knowledge also brings difficulties to the transfer of knowledge from known classes to novel ones. In contrast, the performance of our method for known classes does not drop significantly. Our method uses the information of all samples to learn known and novel classes to perform cooperative learning. As a result, the stable and reliable performance proves that our SCKD method can better deal with the imbalance issue.

Unknown Number of Novel Classes. In the NCD task, the number of latent unlabelled classes C^u is assumed to be known as a priori. However, for real application scenarios, the number of classes (clusters) about the novel classes may be unknown. Due to an insufficiently accurate number of novel classes, the model may over- or under-segment unknown unlabeled samples. Considering this need, previous NCD methods (*e.g.*, DualRank [54], UNO [7], and IIC [25]) introduce an estimation algorithm proposed in DTC [12] to estimate the number of unlabelled classes before the discovery stage. Specifically, DTC first extracts a subset from the label set. This subset is then clustered with the unlabelled set. After optimizing the clustering quality, the estimated number of novel classes can be obtained. Following this setting, we test the performance of different methods on the estimated number of classes $\hat{C}^u$ on the CIFAR100-20 dataset split (averaged over 3 runs). The results are shown in Tab. 5. Compared with the accurate number of novel classes, our method achieves more significant gains (from 0.2 to 2.1) and stabler performance (smaller standard deviation) compared to other methods when dealing with the challenge of unknown number. Our synthesized soft pseudo labels fuse the semantic information of multiple samples without depending on the number of novel classes, reducing adverse interference. The results demonstrate that our method also performs best when facing more realistic challenges.

Table 5: Experiments with the estimated number of novel classes on CIFAR100-20, using the task-aware evaluation protocol.

Method	Class number	
	C^u=20	$\hat{C}^u$=23
DTC [12]	67.3±1.2	64.3
RankStats+ [11]	75.2±4.2	71.2
UNO [7]	90.5±0.7	71.5±1.8
IIC [25]	92.4±0.2	85.1±0.9
SCKD	**92.6±0.6**	**87.2±0.4**

4.3 Ablation Study

We perform ablation study experiments on four dataset splits, including CIFAR100-20, CIFAR100-50, Stanford Cars, CUB, and FGVC-Aircraft. In addition to the clustering accuracy introduced previously, we also introduce two commonly used clustering evaluation metrics: normalized mutual information (NMI) [38] and adjusted rand index (ARI) [18], following existing methods [7, 25]. MNI and ARI

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			CIFAR100-50			Stanford Cars			CUB		
				ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	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

can measure the similarity between the clustering results and the ground-truth distributions (larger is better).

Training Objectives $\mathcal{L}$. We also verify the proposed knowledge transfer objectives. The results in Tab. 6(1)-(4) show the following observations. First, using one of the proposed self-knowledge distillation objectives alone can also significantly improve the performance of the baseline. Combining all two objectives, the model can establish correlations between labeled and unlabeled samples, thereby achieving better results. In addition, an interesting result is that compared to using known class information to directly guide novel class discovery $\mathcal{L}_{k\to n}$, using only novel class information to guide the review of known classes $\mathcal{L}_{n\to k}$ can significantly improve the model's performance in discovering novel classes on CIFAR100. One reason may be that known and novel heads share the same trainable encoder. The process of using novel samples for learning the known classes also pays attention to the information of the novel classes and improves the encoder's ability to encode features of novel class samples. For the Stanford Cars and CUB dataset, we freeze most encoder parameters. In this case, using known information to guide novel classes can improve the model's performance in discovering novel classes more effectively.

Replica Encoder E^r. Since the known and novel classes share the same encoder. Before discovering the novel stage, we deepcopy the pre-trained encoder to ensure unbiased feature information of the known classes. We test the impact of the replica encoder E^r on the results. When E^r is not used, we use the target encoder itself to obtain the similarity scores between labeled and unlabeled samples. The results in Tab. 6(4)-(5) show the following observations. First, only using the target encoder itself can distinguish and associate labeled and unlabeled samples, making the proposed training target effective compared to the baseline. In addition, the shared target encoder integrates information from all categories and is more likely to suffer from the imbalance of information between known and novel categories, especially when there are significantly more novel categories (*e.g.*, CIFAR100-50). The replica encoder retains pure prior information about known categories, thus bringing about a clearer similarity score calculation and promoting the model's differentiation of novel categories.

Similarity Score Matrix S. The similarity score matrix S is introduced to obtain the target pseudo-label by weighting the similarity of known and novel class features. As shown in Tab. 7, we evaluate the effectiveness of our proposed similarity weighting method. 'Average S' denotes all samples have the same influence coefficients. 'Random S' denotes all samples randomly generate influence coefficients ranging from 0 to 1. Our weighting method significantly improves model performance. 'Average S' causes the model to be unable to focus on the mutual information between categories, causing the model to be unable to fully utilize the sample information. 'Random S' may produce pseudo correlations among samples and interfere with model learning.

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**

4.4 Visualization

To intuitively verify the effectiveness of the proposed SCKD, we provide qualitative results. As shown in Fig. 4, we use the t-SNE tool [29] to observe the model's prediction on the CIFAR100-50 testing set. Compared with the original UNO [7], combining with the proposed SCKD can more tightly group samples of the various classes. In accordance with the visualization, SCKD helps to separate the classes better.

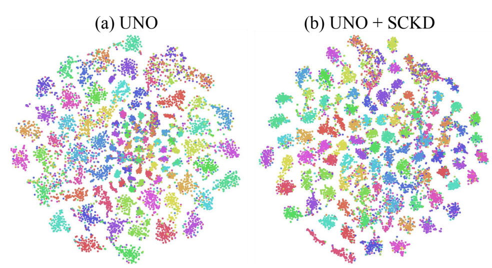

Fig. 4: t-SNE visualization on CIFAR100-50.

5 Conclusion

This paper considers a practical but long-neglected challenge in the Novel Class Discovery (NCD) task, *i.e.*, the imbalanced number of samples from known and novel classes. The model despises the learning of novel classes when the known class dominates while the model forgets the prior knowledge of the known class when the novel class dominates. Compared to transferring knowledge by a shared representation space, we collect the model's feature representations of known and novel classes as two disjoint representation spaces. Then we propose a Self-Cooperation Knowledge Distillation (SCKD) method to utilize every sample for review and discovery. Our method is conceptually simple and intuitive and achieves top-level competitive results in numerous experiments on six datasets.

Acknowledgements. This work is supported by the Shanghai Engineering Research Center of AI & Robotics, Fudan University, China, the Engineering Research Center of AI & Robotics, Ministry of Education, China, and the Green Ecological Smart Technology School-Enterprise Joint Research Center.

References

1. Allen-Zhu, Z., Li, Y.: Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. arXiv preprint arXiv:2012.09816 (2020) 4
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021) 10
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 1
4. Chen, Z., Li, B., Xu, J., Wu, S., Ding, S., Zhang, W.: Towards practical certifiable patch defense with vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15148–15158 (2022) 1
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013) 8
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 9
7. Fini, E., Sangineto, E., Lathuiliere, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9284–9292 (2021) 4, 5, 6, 8, 9, 10, 11, 12, 14
8. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International Conference on Machine Learning. pp. 1607–1616. PMLR (2018) 4
9. Ge, Z., Qi, L., Wang, Y., Sun, Y.: Zoom-and-reasoning: Joint foreground zoom and visual-semantic reasoning detection network for aerial images. IEEE Signal Processing Letters **29**, 2572–2576 (2022) 1
10. Gu, P., Zhang, C., Xu, R., He, X.: Class-relation knowledge distillation for novel class discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 2, 4, 6, 8, 10, 11
11. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Automatically discovering and learning new visual categories with ranking statistics. arXiv preprint arXiv:2002.05714 (2020) 2, 3, 9, 10, 11, 12
12. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8401–8409 (2019) 2, 3, 10, 12
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 1
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1, 9

15. Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning lightweight lane detection cnns by self attention distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1013–1021 (2019) 4
16. Hsu, Y.C., Lv, Z., Kira, Z.: Learning to cluster in order to transfer across domains and tasks. arXiv preprint arXiv:1711.10125 (2017) 1, 2, 3, 10
17. Hsu, Y.C., Lv, Z., Schlosser, J., Odom, P., Kira, Z.: Multi-class classification without multi-class labels. arXiv preprint arXiv:1901.00544 (2019) 1, 2, 3, 10
18. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**, 193–218 (1985) 12
19. Ji, M., Shin, S., Hwang, S., Park, G., Moon, I.C.: Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10664–10673 (2021) 4
20. Jia, X., Han, K., Zhu, Y., Green, B.: Joint representation learning and novel category discovery on single-and multi-modal data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 610–619 (2021) 4, 10
21. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013) 9
22. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 9
23. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955) 9
24. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016) 1
25. Li, W., Fan, Z., Huo, J., Gao, Y.: Modeling inter-class and intra-class constraints in novel class discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3449–3458 (2023) 2, 4, 5, 9, 10, 11, 12
26. Liu, S., Chen, Z., Liu, Y., Wang, Y., Yang, D., Zhao, Z., Zhou, Z., Yi, X., Li, W., Zhang, W., et al.: Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23436–23446 (2023) 1
27. Liu, Y., Xia, Z., Zhao, M., Wei, D., Wang, Y., Liu, S., Ju, B., Fang, G., Liu, J., Song, L.: Learning causality-inspired representation consistency for video anomaly detection. In: Proceedings of the 31st ACM International Conference on Multimedia (ACM MM). pp. 203–212 (2023) 1
28. Liu, Y., Yang, D., Fang, G., Wang, Y., Wei, D., Zhao, M., Cheng, K., Liu, J., Song, L.: Stochastic video normality network for abnormal event detection in surveillance videos. Knowledge-Based Systems **280**, 110986 (2023) 1
29. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 14
30. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967) 10
31. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 9
32. Pu, N., Zhong, Z., Sebe, N.: Dynamic conceptional contrastive learning for generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7579–7588 (2023) 2, 4

33. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 1
34. Shen, Y., Xu, L., Yang, Y., Li, Y., Guo, Y.: Self-distillation from the last mini-batch for consistency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11943–11952 (2022) 4
35. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017) 2
36. Troisemaine, C., Lemaire, V., Gosselin, S., Reiffers-Masson, A., Flocon-Cholet, J., Vaton, S.: Novel class discovery: an introduction and key concepts. arXiv preprint arXiv:2302.12028 (2023) 2
37. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7492–7501 (2022) 2, 4, 10
38. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th annual international conference on machine learning. pp. 1073–1080 (2009) 12
39. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 9
40. Wang, W., Zheng, V.W., Yu, H., Miao, C.: A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(2), 1–37 (2019) 2
41. Wang, Y., Chen, Z., Yang, D., Guo, P., Jiang, K., Zhang, W., Qi, L.: Out of thin air: Exploring data-free adversarial robustness distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5776–5784 (2024) 1
42. Wang, Y., Chen, Z., Yang, D., Liu, Y., Liu, S., Zhang, W., Qi, L.: Adversarial contrastive distillation with adaptive denoising. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 1
43. Wang, Y., Chen, Z., Zhang, J., Yang, D., Ge, Z., Liu, Y., Liu, S., Sun, Y., Zhang, W., Qi, L.: Sampling to distill: Knowledge transfer from open-world data. arXiv preprint arXiv:2307.16601 (2023) 1
44. Wang, Y., Ge, Z., Chen, Z., Liu, X., Ma, C., Sun, Y., Qi, L.: Explicit and implicit knowledge distillation via unlabeled data. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 1
45. Wang, Y., Yang, D., Chen, Z., Liu, Y., Liu, S., Zhang, W., Zhang, L., Qi, L.: De-confounded data-free knowledge distillation for handling distribution shifts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12615–12625 (2024) 1
46. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16590–16600 (2023) 2, 4, 10
47. Yang, D., Chen, Z., Wang, Y., Wang, S., Li, M., Liu, S., Zhao, X., Huang, S., Dong, Z., Zhai, P., Zhang, L.: Context de-confounded emotion recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19005–19015 (June 2023) 1
48. Yang, D., Huang, S., Xu, Z., Li, Z., Wang, S., Li, M., Wang, Y., Liu, Y., Yang, K., Chen, Z., et al.: Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20459–20470 (2023) 1

49. Yang, D., Yang, K., Li, M., Wang, S., Wang, S., Zhang, L.: Robust emotion recognition in context debiasing. arXiv preprint arXiv:2403.05963 (2024) 1
50. Yang, D., Yang, K., Wang, Y., Liu, J., Xu, Z., Yin, R., Zhai, P., Zhang, L.: How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. Advances in Neural Information Processing Systems (NeurIPS) **36** (2024) 1
51. Yang, M., Zhu, Y., Yu, J., Wu, A., Deng, C.: Divide and conquer: Compositional experts for generalized novel class discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14268–14277 (2022) 2, 4, 10
52. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 4
53. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3713–3722 (2019) 4
54. Zhao, B., Han, K.: Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. Advances in Neural Information Processing Systems **34**, 22982–22994 (2021) 2, 4, 10, 12
55. Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10867–10875 (2021) 2, 4, 10, 11
56. Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., Sebe, N.: Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9462–9470 (2021) 2, 4, 10