Uncertainty-Guided Semi-Supervised Few-Shot Class-Incremental Learning with Knowledge Distillation

Yawen Cui, Wanxia Deng, Xin Xu, Zhen Liu, Zhong Liu, Matti Pietikäinen, and Li Liu

***Abstract*—Class-Incremental Learning (CIL) aims at incrementally learning novel classes without forgetting old ones. This capability becomes more challenging when novel tasks contain one or a few labeled training samples, which leads to a more practical learning scenario, *i.e.*, Few-Shot Class-Incremental Learning (FSCIL). The dilemma on FSCIL lies in serious overfitting and exacerbated catastrophic forgetting caused by the limited training data from novel classes. In this paper, excited by the easy accessibility of unlabeled data, we conduct a pioneering work and focus on a Semi-Supervised Few-Shot Class-Incremental Learning (Semi-FSCIL) problem, which requires the model incrementally to learn new classes from extremely limited labeled samples and a large number of unlabeled samples. To address this problem, a simple but efficient framework is first constructed based on the knowledge distillation technique to alleviate catastrophic forgetting. To efficiently mitigate the overfitting problem on novel categories with unlabeled data, uncertainty-guided semi-supervised learning is incorporated into this framework to select unlabeled samples into incremental learning sessions considering the model uncertainty. This process provides extra reliable supervision for the distillation process and contributes to better formulating the class means. Our extensive experiments on CIFAR100, miniImageNet and CUB200 datasets demonstrate the promising performance of our proposed method, and define baselines in this new research direction.**

***Index Terms*—Few-shot learning, class-incremental learning, object classification, computer vision, semi-supervised learning, deep learning, knowledge distillation, uncertainty estimation**

I. Introduction

In the past decade, deep learning has achieved startling progress in various computer vision tasks, such as image classification [1], [2], [3], object detection [4], face recognition [5], semantic segmentation [6], image synthesis [7], etc. Very often, the high performance of deep learning networks has been achieved by the standard form of offline supervised learning (batch learning) which usually assumes that the training data contains all interested classes with each class having abundant labeled samples. This is problematic in many realistic applications such as robotics and streaming data. Under this setting, the learned model is only as good as the static training data it builds on, may be unable to recognize any object that is not in the training data. The model needs to be retrained from scratch with enough additional data relevant to the new categories, thus is very expensive to adapt and robustly generalize [8]. This, however, is in stark contrast with common real-world conditions that the model operating in the deployment scenario may encounter new categories that were not present in the initial training [9]. Therefore, it is of great importance to enable the model to have the ability to learn new categories incrementally without forgetting the previously learned categories. To this end, Class Incremental Learning (CIL) which aims at incrementally learning deep learning networks to recognize new categories has been receiving increasing attention in recent years [10], [11].

CIL is certainly a challenging problem, as it has to learn a number of different tasks sequentially while still perform well on old ones by reserving the knowledge learned from previous tasks. Current CIL works are mainly divided into replay methods [10], regularization-based methods [11] and parameter-isolation methods [12], [13]. Replay methods aim to store part of samples belonging to previous categories in raw format or generate pseudo samples with generative models. Regularization-based methods focus on introducing an additional regularization term in the loss function, and the most popular line for this family method is knowledge-distillation-based [14]. Parameter-isolation methods try to preserve significant parameters of the model. However, these three kinds of methods all require the model to be well trained with a massive number of labeled samples in order to select preserved samples or important parameters. However, in CIL scenario, with more and more classes arriving, it is costly and time-consuming to collect abundant labeled data for each new class. It is very realistic that one can only obtain very limited labeled samples but lots of unlabeled data due to various factors such as privacy or security issues. Thus, to overcome the aforementioned limitations, label-efficient techniques like FSL techniques are useful, leading to several works on a

This work was partially supported by National Key Research and Development Program of China No. 2021YFB3100800, the Academy of Finland under grant 331883, the National Natural Science Foundation of China under Grant 61872379, 62022091 and 62022091, and the China Scholarship Council (CSC) under grant 201903170129.

Li Liu and Zhong Liu are with the Laboratory for big data and decision, the College of System Engineering, National University of Defense Technology (NUDT), Changsha, Hunan, China. Li Liu is also with the Center for Machine Vision and Signal analysis (CMVS), University of Oulu, Oulu, Finland. Li Liu is the corresponding author. (email: dreamliu2010@gmail.com; liuzhong@nudt.edu.cn)

Yawen Cui and Matti Pietikäinen are with CMVS, University of Oulu, Oulu, Finland. (email: yawen.cui@oulu.fi; matti.pietikainen@oulu.fi)

Wanxia Deng is with the School of Meteorology and Oceanography, NUDT, Changsha, Hunan, China. (email: dengwanxia14@nudt.edu.cn)

Xin Xu is with the College of Intelligent Science, NUDT, Changsha, Hunan, China. (email:xinxu@nudt.edu.cn)

Zhen Liu is with the College of Electronic Science, NUDT, Changsha, Hunan, China. (email: zhen _liu@nudt.edu.cn))

Manuscript received December 8, 2021; revised April 9, 2022 and September 9, 2022; accepted September 17, 2022.

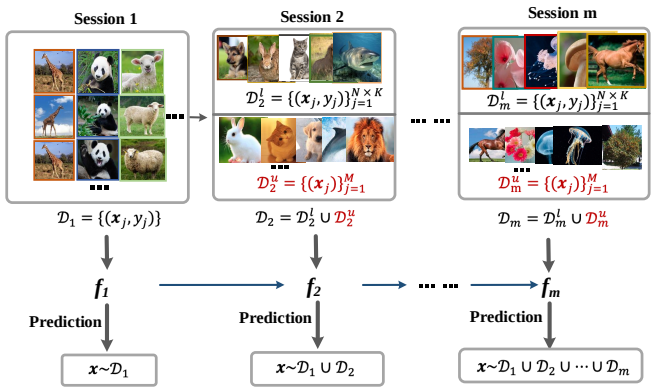


Fig. 1. The task configuration. The first session's training set is a large-scale labeled training set $\mathcal{D}_1$. The sets of following sessions are all N-way K-shot semi-supervised few-shot task settings with labeled data $\mathcal{D}^l$ and unlabeled data $\mathcal{D}^u$.

very realistic setting, *i.e.*, Few-Shot CIL (FSCIL)[15], [16]. There is the biological inspiration for this learning paradigm [17], as human learning is knowledge-driven and we maintain knowledge learned from all previous tasks, apply them to help us solve new tasks with one or several examples (few-shot learning ability) [18], [19], [20] without forgetting old tasks, and incrementally learn and accumulate new knowledge.

There are only several works on the FSCIL setting [15], [21], [22], [23]. Tao *et al.*[15] first proposed to use neural gas to model the topology of each category in feature space. Zhu *et al.*[23] proposed a novel incremental prototype learning scheme to solve FSCIL tasks. Although these methods can accomplish remembering seem categories before and adapting fast to new categories with limited samples, the model still suffers from class-imbalance and overfitting problems. Moreover, the radical problem of FSCIL, the inadequate labeled samples in each incremental session, has not been solved. Fortunately, in the field of natural object classification, a huge number of unlabeled images (*e.g.*, the massive amounts of unlabeled images available from the Internet) are significantly cheap to obtain. Making good use of such unlabeled data to complement the insufficient of labeled data becomes crucial, which is also the target of Semi-Supervised Learning (SSL). As we discussed above, semi-supervised approaches [24], [25] that can leverage unlabeled data are of particular interest. Therefore, in this paper, we present a study aiming to go beyond the aforementioned limitations by considering a very practical learning setting, *i.e.*, Semi-Supervised FSCIL (Semi-FSCIL), where new classes are registered incrementally and with only a few labeled samples and many unlabeled data. To the best of our knowledge, we are the first to study this Semi-FSCIL setting.

The task configuration of Semi-FSCIL is illustrated in Fig. 1. In the first session, the model is trained with a large-scale training set. In the following session, few-shot labeled data with unseen categories and an available unlabeled dataset arrive. Significantly, the model can be guaranteed to skillfully memorize all previously seen categories in each session. Compared to FSCIL, considering that the imported unlabeled dataset can facilitate the training process, the proposed Semi-FSCIL setting can alleviate the negative effect caused by the class imbalance between old and new categories.

To address the Semi-FSCIL problem, with the **U**ncertainty-guided **s**emi-supervised learning method, we propose a simple **K**nowledge **D**istillation-based FSCIL framework termed **Us-KD**. The distillation-based method, a more popular strategy for addressing CIL tasks, can memorize the old knowledge by forcing the network to maintain the same output logits as the previous model. Nevertheless, Tao *et al.* [15] point out that existing distillation-based CIL methods are inappropriate for solving FSCIL tasks where limited labeled samples of new categories bring difficulty in maintaining the output logits for old classes because of a larger learning rate and stronger gradients. Since we introduce unlabeled data into each incremental session of FSCIL and it can be approximately treated like general CIL where a massive number of labeled samples belonging to novel categories are contained, our proposed distillation-based framework for Semi-FSCIL setting can perform better.

In each incremental learning session, the semi-supervised learning paradigm is implemented via pseudo-labeling-based strategies. To decrease the prediction uncertainty on unlabeled samples, we propose to apply a novel uncertainty-guided semi-supervised learning method for the new Semi-FSCIL setting. Specifically, we select unlabeled samples and the corresponding pseudo labels using the uncertainty-guided module in the semi-supervised learning process.

This paper is an extended version of our prior publication [26] in ICIP 2021. In comparison with our previous ICIP version, this submission features several significant improvements extending the preliminary conference version as follows. (1) Deep learning networks express uncertainty (like under-confident or over-confident) in their prediction, especially when they are trained with noisy data, or very limited data like FSL and FSCIL, or when they face novel categories. Understanding this can help get better performance out of it. However, the preliminary conference version [26] did not consider the impact of noisy pseudo-labels of unlabelled data, leading to suboptimal performance. In this study, we take this issue into consideration and propose an effective solution by exploring uncertainty guidance to improve the performance of Semi-FSCIL. (2) This work is a much more comprehensive study of the novel Semi-SFCIL problem in terms of several aspects, including motivations, literature review, approach analysis, and experiments. Finally, the main contributions of this paper are summarized as follow.

- To the best of our knowledge, we are the first to focus on the Semi-Supervised Few-Shot Class-Incremental Learning (Semi-FSCIL), which is devoted to the under-explored challenging semi-supervised form of FSCIL. The detailed problem description and configuration are provided.
- With the unlabeled data, the knowledge distillation-based CIL framework is applicable for the FSCIL task. Based on the developed distillation-based framework, we embed an uncertainty-guided module to guarantee an efficient semi-supervised learning procedure.

- To better model the data distributions of classes with limited labeled samples, we use the popular nearest-mean-of-exemplars classification in the proposed framework and propose a weighted class-mean calculation with labeled data and incorporated unlabeled data.
- Extensive experiments on the CIFAR100, CUB200, *mini*ImageNet benchmark datasets are provided to demonstrate the proposed Us-KD achieves remarkable results. In addition, we conduct careful ablation studies on benchmark FSCIL datasets, verifying the efficacy of the proposed method.

The remainder of the paper is organized as follows. In Section II, we review the related work. Section III introduces the formulation of the novel Semi-FSCIL. We present our proposed framework in Section IV. Section V reports the experimental results and analysis on CIFAR100, *mini*ImageNet, and CUB200 datasets. Our conclusion and future work are presented in Section VI.

II. Related Work

This paper conducts a pioneering work, *i.e.*, embedding the semi-supervised learning into the FSCIL process. In this section, closely-related works on semi-supervised learning, few-shot learning, and class-incremental learning are discussed in detail.

A. Few-Shot Learning

Few-shot learning (FSL) aims to solve the target task with limited labeled samples per class and a related source dataset whose knowledge can be transferred to the few-shot target task. Vanilla FSL can be categorized into data-augmentation-based methods [27], [28] targeting at enlarging the limited labeled dataset in the instance level or the feature level, and prior-knowledge-based methods containing two series of methods: meta-learning-based methods and transfer learning-based methods [29], [30], [31], [32]. The paradigm of transfer learning-based methods simply pretrain a model with a large-scale dataset and further finetune the model on the FSL task with the strategies of alleviating overfitting.

Recent efforts on meta-learning-based FSL develop toward the following three directions. (i) Metric learning-based methods [33], [34], applying the metrics to measure the distance or similarity among support images and query images. Vinyals *et al.*[33] first introduce the metric learning into the FSL task, where the training image and the test image are mapped into embedding space. Then, the attention mechanism is used to obtain the similarity of images. (ii) Optimization-based methods [35], [36] focus on optimizing parameter configurations of a given neural network such that it can effectively be fine-tuned on FSL tasks within a few gradient-descent update steps. MAML [35] aims at obtaining optimal initialization parameters of the model through training. (iii) Memory-based methods [37], [38] model the support set of the FSL task as a sequence and formulating it as a sequence learning task. The query samples are required to match with the previously obtained knowledge. Santoro *et al.* [37] demonstrate the ability of a memory-augmented neural network to rapidly assimilate new data and leverage this data to make accurate predictions after being trained on only a few samples.

Recently, several challenging settings of FSL have been proposed, *e.g.*, semi-supervised FSL and unsupervised FSL. In the semi-supervised FSL scenario, some unlabeled samples of the source/target dataset are also available. The work [39] extends the prototypical network by incorporating unlabeled data to update the original prototypes. Liu *et al.*[40] use the transductive propagation network to propagate labels from labeled images to unlabeled images along with the constructed graph. In unsupervised FSL, the source dataset is totally unlabeled. CACTUs [41] applies clustering on source data to obtain pseudo labels and constructed meta-tasks from clusters randomly by regarding every cluster as a specific class. UMTRA [42] selects N samples from unlabeled training set randomly, and the probability that these N pictures belonged to different categories is very high, so these N pictures are constituted as one N-way 1-shot meta-task. Progressive clustering and episodic training are used in UFLST [43] to implement unsupervised meta-training.

B. Class-Incremental Learning

The Class-Incremental Learning (CIL) aims at incrementally learning a unified classifier to recognize all encountered categories. The main challenge of CIL is the catastrophic forgetting [44] on previous categories when learning new ones. To overcome the difficulty, great efforts have been devoted to the following two directions. The first direction is to identify and preserve significant parameters of the original model [12], [13], and the second one is preserving the knowledge of the old model through some strategies like knowledge distillation [45], [46].

We name the methods following the first direction as parameter-based approaches which try to assess the importance of each parameter in the previous model and set high penalty values of significant parameters learning. The work [12] remembers old tasks by selectively learning the weights that are important for those tasks. Inspired by biological neural networks, Zenke *et al.*[13] introduce intelligent synapses that bring some biological complexity into artificial neural networks. Moreover, each synapse accumulates task-relevant information over time and exploits this information to rapidly store new memories without forgetting old ones. Memory Aware Synapses (MAS) [47] computes the importance of the parameters of a neural network in an unsupervised and online manner. Nevertheless, these methods are difficult to make a reasonable metric to evaluate the significance of parameters.

Transferring knowledge [48], [49] from one network to another is a solution to retain the previous capabilities of the original network. Knowledge distillation technique [14] is an effective and simple method to transfer the knowledge. Another type of CIL method [50], [45], [46], [11] is to memorize via the distillation technique. Expert gate [50] uses only new task data to train the network while preserving the original capabilities. In this work, a set of gating autoencoders are introduced to learn a representation for the current task,

and automatically forward the test sample to the relevant expert at the test stage. iCaRL [10] learns strong classifiers and a data representation simultaneously. Hou *et al.*[11] aims to solve the class-imbalance problem in CIL by introducing three constraints.

To bypass the difficulty of labeled data acquisition, some works are devoted to introducing few-shot learning into CIL, and lead to a new realistic scenario, *i.e.*, Few-Shot CIL (FSCIL) [15], [16]. Existing methods for FSCIL mainly follow two strategies. The first one is knowledge representation and refinement. Tao *et al.*[15] propose TOPIC to model the topology of feature space using neural gas. TOPIC preserves the old knowledge by stabilizing the topology of neural gas, and adapts the feature space by pushing the new class training sample towards a correct new neural gas node with the same label and pulling the new nodes of different labels away from each other. Zhang *et al.*[51] adopts a simple but effective decoupled learning strategy of representations, and Continually Evolved Classifier (CEC) is proposed by employing a graph model to propagate context information between classifiers for adaptation. Zhu *et al.*[23] proposes a novel incremental prototype learning scheme to solve the FSCIL task. The other strategy is via knowledge distillation. Cheraghian *et al.*[21] employs the word embedding as the semantic information during training and an attention mechanism on multiple parallel embeddings of visual data is proposed to align visual and semantic vectors, which reduces issues related to catastrophic forgetting.

C. Semi-Supervised Learning

Semi-Supervised Learning (SSL) focuses on learning a robust model by leveraging a few labeled data and a large amount of unlabeled data. There are many SSL approaches which are mainly divided into consistency-regularization methods [24], [25], pseudo-labeling [52], [53], transductive model [54], generative modeling [55] and graph-based methods [56], [57].

Consistency-regularization methods target making the network's outputs invariant to small input perturbations. However, this method always relies on an extensive set of data augmentations requiring domain-specific knowledge. Miyato *et al.*[24] applies perturbations to the input that changes the output predictions. A recent work [58] incorporates a time-consuming metric to choose time-consuming examples for consistency regularization effectively. Pseudo-labeling is to generate pseudo labels for unlabeled data by using the model trained on labeled data. The pseudo labels of unlabeled data can be created by the predictions of trained neural network [53] or assigned based on neighborhood graph [59].

Transductive models aim to optimize the models using the training set and the testing set for the prediction process. TSVMs [60] propose transductive support vector machines by taking into account a particular test set and minimizing misclassifications of just those particular examples. Joachims *et al.*[54] presents a transductive version of k nearest-neighbor classifier. Graph-based methods aim to build a graph connecting similar observed information, and the label information propagates through the graph from labeled to unlabeled nodes. In [56], labeled and unlabeled data are represented as vertices in a weighted graph, with edge weights encoding the similarity between instances. Liu *et al.*[57] presents that the label propagation scheme could be highly effective when the similarity metric used for propagation was transferred from other related domains.

Our proposed scenario, a very practical learning setting, *i.e.*, Semi-Supervised FSCIL (Semi-FSCIL), where new classes are memorized incrementally with only a few labeled samples and many unlabeled data, extends semi-supervised learning into FSCIL. For the Semi-FSCIL task, by incorporating pseudo-labeling-based semi-supervised learning method into the FSCIL learning framework, the unlabeled data will participate in distillation-based class-incremental learning to alleviate overfitting and class-imbalance problems of FSCIL.

III. The Semi-FSCIL Problem Formulation

The goal of Semi-FSCIL is to incrementally learn novel classes in a semi-supervised manner on top of a base session initializing the model. Once the training is completed, the model $\mathcal{F}(\cdot)$ will be able to classify samples from all the seen classes so far. As for the model $\mathcal{F}(\cdot)$, it usually contains the backbone $\Theta(\cdot)$ for extracting features and the classification model $\Gamma(\cdot)$.

Fig. 1 presents an illustration of the Semi-FSCIL task configuration. Commonly, we present a sequence of disjoint datasets by $\mathcal{D} = \{D_1, D_2, ..., D_m\}$, where $\mathcal{D}_1$ is the large-scale base dataset used in the first base session and the followings are all novel few-shot datasets. We first conduct the base session using a large-scale labeled base dataset $\mathcal{D}_1 = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{|D_1|}$ where $y_j \in C_1$ and C_1 denotes the category set. After that, we incrementally inject novel data with new categories to the model in the following incremental sessions. In the i-th session, we train the model on the dataset $\mathcal{D}_i = \mathcal{D}_i^l \cup \mathcal{D}_i^u$, $\mathcal{D}_i^l$ and $\mathcal{D}_i^u$ denote the labeled training data and unlabeled training data, respectively. The labeled training data $\mathcal{D}_i^l = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{N\times K}$ consists of N classes C_i with K labeled examples per class, *i.e.*, a N-way K-shot problem. The unlabeled training data $\mathcal{D}_i^u = \{\boldsymbol{x}_j\}_{j=1}^{M}$ only comprises unlabeled samples, where $M \gg K$. Noteworthy, there is no overlap between the categories of different sessions, *i.e.*, $C_i \cap C_{i'} = \emptyset$, where $i \neq i'$. With this configuration, we give the distribution formulation of Semi-FSCIL. In traditional supervised learning, the domain required to learn encompasses the feature space $\mathcal{X}$ and marginal probability distribution $p(\boldsymbol{x})$. The task is made of the label space $\boldsymbol{y}$ and conditional probability distribution $p(y|\boldsymbol{x})$. For any session i and i' in Semi-FSCIL, $\mathcal{X}_i = \mathcal{X}_{i'}$, while $p_i(\boldsymbol{x}) \neq p_{i'}(\boldsymbol{x})$, $\boldsymbol{y}_i \neq \boldsymbol{y}_{i'}$ and $p_i(y|\boldsymbol{x}) \neq p_{i'}(y|\boldsymbol{x})$. Moreover, the unlabeled dataset shares the same feature space and label space with the labeled dataset. This configuration is inspired by real-world applications: lots of unlabeled samples can be more easily collected, while annotations are only assigned for a few samples due to the costly labeling.

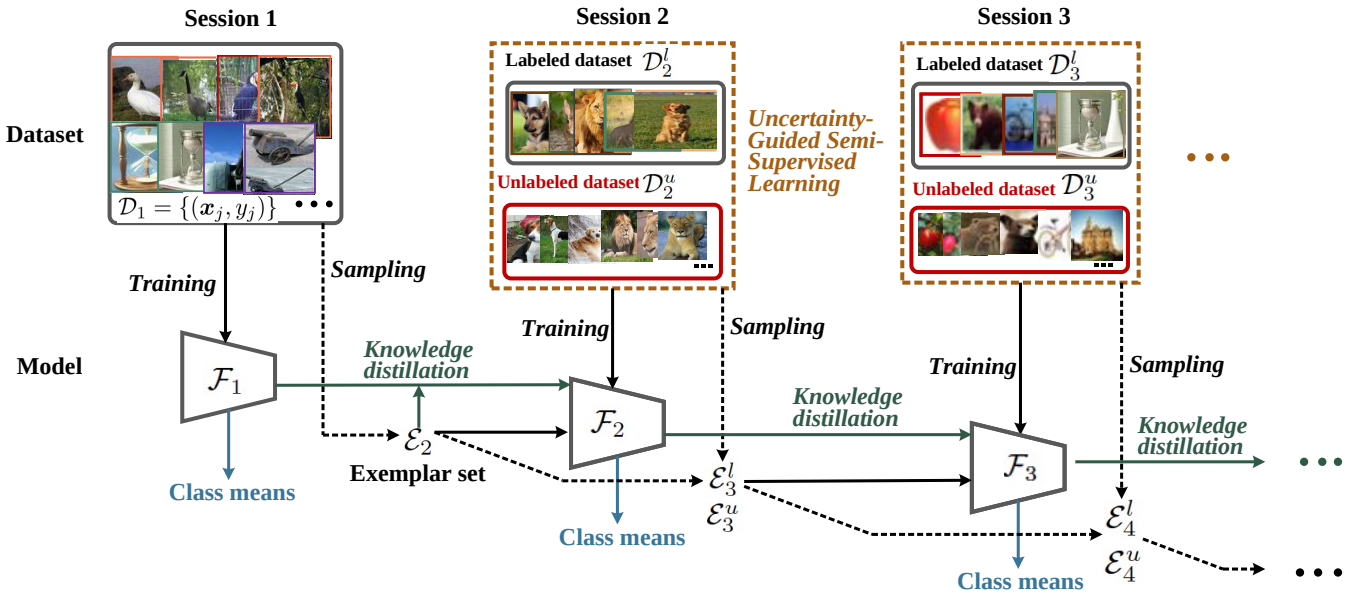


Fig. 2. The proposed Semi-FSCIL framework. Our incremental learning framework is knowledge-distillation-based. We apply the nearest-mean-of-exemplars rule on classification and prioritize exemplars selection based on herding. Moreover, we combine the uncertainty-guided module for selecting unlabeled samples, conducting knowledge distillation and updating prototypes with labeled and unlabeled samples. Only classification loss is computed in the base session, and an extra distillation loss is performed in the following incremental sessions.

IV. Knowledge Distillation-Based Semi-FSCIL Framework

In this section, we introduce the proposed Us-KD framework for Semi-FSCIL. We first present the overview of the framework, then the uncertainty-guided semi-supervised learning and knowledge distillation process in Us-KD, and lastly the computation of class-means in the Semi-FSCIL.

A. Overview

Our Us-KD is illustrated in Figure. 2. First, the model is trained with $\mathcal{D}_1$ by computing the classification loss to obtain $\mathcal{F}_1$. At the end of session 1, we select an exemplar set $\mathcal{E}_2$ used in session 2 from $\mathcal{D}_1$ following [10] for distilling the knowledge. When it comes to the second session, $\mathcal{F}_2$ is initialized by $\mathcal{F}_1$ and $\mathcal{F}_1$ is regarded as the reference model, *i.e.*, the teacher. With the labeled dataset and unlabeled dataset together, uncertainty-guided semi-supervised learning is conducted to learn novel categories. First, $\mathcal{D}_2^l$ is fed to the model for supervised learning. After that, unlabeled iterations are conducted to add unlabeled samples together with the obtained pseudo labels into the training process of $\mathcal{F}_2$. During training in this session, we compute the classification loss and the distillation loss on the exemplar set $\mathcal{E}_2$. Finally, by adding samples from the current dataset, $\mathcal{E}_2$ is updated into the labeled exemplar set $\mathcal{E}_3^l$ with the ground-truth and the unlabeled exemplar set $\mathcal{E}_3^u$ with the obtained pseudo labels. The procedure of session 2 will be iteratively implemented for the rest sessions to accomplish the FSCIL task. The pseudo-code about the whole procedure of Semi-FSCIL with Us-KD is illustrated in Algorithm 1.

B. Uncertainty-Guided Semi-Supervised Learning

To alleviate the overfitting problem on new categories, we propose to add unlabeled data into each incremental learning session, which is the semi-supervised learning setting. Pseudo-labeling-based and consistency-regularization-based methods are the two main series of methods for SSL. One issue with consistency-regularization-based methods is that they often rely on a rich set of augmentations that require domain knowledge. In the few-shot learning setting, domain knowledge is too limited to conduct the augmentation process. Pseudo-labeling-based SSL methods do not inherently require augmentation and can be generally applied to most domains. In this way, we apply the pseudo-labeling-based SSL method on the proposed Semi-FSCIL framework.

There are several methods to generate pseudo-labels. We adopt the approach in which pseudo labels are obtained directly from network predictions. Assuming an unlabeled dataset $\mathcal{D}^u = \{\boldsymbol{x}_j\}_{j=1}^M$ in a specific incremental learning session and the pseudo labels set is $\widetilde{Y} = \{\widetilde{y}_j\}_{j=1}^M$. The probability outputs of a trained model on the unlabeled sample $\boldsymbol{x}_j$ is set to p_j, and p_j^c represents the probability of the sample $\boldsymbol{x}_j$ belonging to class c. Based on the probability outputs, the pseudo labels of $\boldsymbol{x}_j$ can be created as:

$$\widetilde{y}_j = \mathbb{1}\left[p_j^c == \boldsymbol{max}(p_j)\right]_{c=0}^{|p_j|}, \tag{1}$$

where $\boldsymbol{max}(p_j)$ is the maximum of p_j.

The common assumption in SSL is that the decision boundary should move from high-density regions to low-density regions in the learning process [61]. Pseudo-labeling-based methods select unlabeled samples with high confidence predictions, and this indeed makes decision boundaries move in the right direction. Nevertheless, consistency-regularization-based methods outperform pseudo-labeling-based methods in

recent works. The reason is that incorrect predictions can have high confidence scores in poor calibration neural networks [62]. In the Semi-FSCIL setting, the model is introduced with a few labeled data of novel classes and updated for several epochs in each incremental learning session. Then the prediction on unlabeled data is conducted. In this way, the prediction uncertainty is very high since the distribution of novel categories can not be modeled with very limited labeled data. For the Semi-FSCIL, unlabeled data is used to complement the inadequate of novel categories; thus the quality of pseudo labels is significant for improving the overall performance on seen categories. Although confidence-based selection reduces the error rates of pseudo labels, it is not satisfied for Semi-FSCIL since the number of unlabeled samples is much larger than that of labeled samples. Unlabeled data contributes more to update the model, and the quality of pseudo labels is of great significance. To tackle this problem of high uncertainty, we combine an uncertainty-guided module into our Semi-FSCIL.

Assuming $\boldsymbol{u}(\cdot)$ is the uncertainty function of a prediction for a specific unlabeled data. In order to compute the uncertainty of prediction in a specific session i, we obtain the predictions for a certain unlabeled x_j with random crop and random horizontal flip for L times represented as $[p_j^1, p_j^2, ..., p_j^L]$, and final p_j is defined as

$$p_j = \frac{\left(\sum_{l=1}^{L} p_j^l\right)}{L}, \tag{2}$$

and $\boldsymbol{u}(p_j)$ is defined as

$$\boldsymbol{u}(p_j) = \boldsymbol{std}\left(\{p_j^l\}_{l=1}^{L}, p_j\right), \tag{3}$$

where $\boldsymbol{std}$ is the standard deviation function.

Let $g_j = [g_j^1, g_j^2, ..., g_j^{C_i}] \subseteq \{0,1\}^{C_i}$ be a one-hot vector representing the pseudo label of sample x_j, where $g_j^c = 1$ when $\widetilde{y}_j^c$ is selected as the pseudo label and otherwise $g_j^c = 0$. With the definitions above, we set g_j as

$$g_j = \mathbb{1}\left[p_j^c == \boldsymbol{max}(p_j)\right]_{c=0}^{|p_j|} \mathbb{1}[\boldsymbol{u}(p_j) \leq \kappa_u], \tag{4}$$

where κ_u is the uncertainty threshold. We add the additional uncertainty-guided module $\boldsymbol{u}(p_j)$ into the pseudo label selection process to ensure that the model is sufficiently stable to select the correct pseudo labels.

C. Knowledge Distillation in Semi-FSCIL

Class-incremental learning aims to incrementally learn new categories that are never seen previously. When new categories are encountered, the classification model is updated by learning the novel samples. The discriminative ability of old categories is undermined, which refers to catastrophic forgetting. To mitigate this forgetting issue, the knowledge distillation technique is employed in class-incremental learning framework with the achieved state-of-the-art results [10], [11]. The core of knowledge distillation-based methods is to maintain the network's output logits corresponding to old categories in two neighboring sessions. Specifically, with new classes arriving, an extra distillation loss is introduced to the existing cross-entropy loss to ensure that the current model mimics the teacher model's performance in the previous session. In this way, the incremental learning loss function is computed as:

$$\mathcal{L}(\mathcal{D}_i, \mathcal{E}_i, \mathcal{F}) = \mathcal{L}_{ce}(\mathcal{D}_i, \mathcal{E}_i, \mathcal{F}) + \lambda \mathcal{L}_{dl}(\mathcal{E}_i, \mathcal{F}), \tag{5}$$

where $\mathcal{L}_{ce}$ means the cross-entropy loss and $\mathcal{L}_{dl}$ denotes the distillation loss. $\mathcal{D}_i$ is the training set of current session i, and $\mathcal{E}_i$ is old class exemplars drawn from previous sets $\{\mathcal{D}_1, \ldots, \mathcal{D}_{i-1}\}$. $\mathcal{L}_{dl}$ can be implemented in various ways and holds the form generally:

$$\mathcal{L}_{dl}(\mathcal{E}_i, \mathcal{F}) = -\sum_{j=0}^{|\mathcal{E}_i|} \tau(\mathcal{F}_{i-1}(\boldsymbol{x}_j)) log(\tau(\mathcal{F}_i(\boldsymbol{x}_j))), \tag{6}$$

where $\mathcal{F}_i$ is the model obtained in the i^{th} session, and $\tau(\nu_i) = \nu_i^{1/\Omega} / \sum_j \nu_j^{1/\Omega}$ is the rescaling function, where Ω is usually set to be greater than 1. To remember the performance of old classes, the distillation loss mainly measures variations on the predictions of old samples obtained by the models in two neighboring sessions. To summarize, knowledge distillation encounters between two neighboring sessions with the exemplar set containing the chosen samples of all seen categories in previous sessions, which is the dependencies among all sessions.

In general CIL tasks, the distillation loss defined above is the common scenario where predictions are prone to new classes. This bias is caused by the class-imbalance problems, since only part of samples of previous sessions will be introduced to the next sessions while the incoming new categories contain large-scale sets. Many novel distillation-based CIL works [10], [11] target at solving this class-imbalance problem. However, in FSCIL, a few labeled samples of new classes are provided in the following session, so the class-imbalance problem is different from general CIL tasks. In this way, the current state-of-the-art distillation-based class-incremental learning methods are redundant for the Semi-FSCIL setting. In this paper, we propose a simple yet efficient knowledge distillation-based framework for Semi-FSCIL.

As described above, knowledge distillation-based incremental learning approaches target keeping the prediction results on old categories in the current session the same as the reference model obtained in the last learning session. In this way, to compute the distillation loss in each incremental session i, part of samples of previous sessions, termed the exemplar set $\mathcal{E}_i$, are chosen from the datasets of all previous sessions. $\mathcal{E}_i$ takes part in not only the distillation process but also the supervised training process together with the current training dataset.

In our proposed framework, we employ the prioritized based on herding strategy [10]. To adapt to our framework, we further put forward to prioritize the labeled and selected unlabeled samples of previous sessions together to generate the exemplar set. Consequently, labeled exemplars and unlabeled exemplars all contribute to the knowledge distillation process. With the standard knowledge distillation formula 6, we define

Algorithm 1 Us-KD Framework for Semi-FSCIL

Input: $\mathcal{D}_1$, $\mathcal{D}_2$, ..., $\mathcal{D}_m$
Output:$\mathcal{F}$ that can classify all seen categories

for n in m **do**
  if n==1 **then**
    $F(\cdot)$ updates with $\mathcal{D}_1^l$ by computing classification loss to obtain F_1;
    Sample exemplars $\mathcal{E}_2^l$ from $\mathcal{D}^1$;
    $\mathcal{E}_2^u$=$\varnothing$;
  else
    F_{ref} = F_{n-1};
    F_{target} = F_{ref};
    for supervised epochs **do**
      F_{target} updates with $\mathcal{E}_i \cup \mathcal{D}_n^l$ by compute classification loss and distillation loss;
    $\hat{D}_n^u$ = $\varnothing$;
    for unlabeled iteration **do**
      Select unlabeled samples and add them to $\hat{D}_n^u$ based on uncertainty-guided manner;
      Remove them from D_n^u;
      Feed D_n^l and $\hat{D}_n^u$ to F_{target};
      Compute the classification loss and the distillation loss;
  F_n = F_{n-1};
  Update $\mathcal{E}_n$ to $\mathcal{E}_{n+1}$ by sampling from D_n^l and $\hat{D}_n^u$;
  Compute class-means for classification;
return $F(\cdot)$ obtained after m sessions.

the new distillation loss as

$$\begin{aligned}\mathcal{L}_{dl}(\mathcal{E}_i, \mathcal{F}) = -\lambda^l \sum_{j=0}^{|\mathcal{E}_i^l|} \tau(\mathcal{F}_{i-1}(\boldsymbol{x}_j)) log(\tau(\mathcal{F}_i(\boldsymbol{x}_j))) \\ -\lambda^u \sum_{j=0}^{|\mathcal{E}_i^u|} \tau(\mathcal{F}_{i-1}(\boldsymbol{x}_j)) log(\tau(\mathcal{F}_i(\boldsymbol{x}_j))),\end{aligned} \tag{7}$$

where $\mathcal{E}_i^l$ and $\mathcal{E}_i^u$ represent the labeled exemplars and unlabeled exemplars, respectively. λ^u and λ^u are the weights for the two parts of distillation loss.

In the supervised set of FSCIL, since the dataset is large-scale in the first session and each category in the incremental session contains a few labeled samples, the class-imbalance problem may exist in the exemplar set. This problem will generate an extra bias in the distillation process, and the target model only better memorizes the categories with the more significant number of samples in the exemplar set. In our proposed Semi-FSCIL task, unlabeled data is added in each incremental session, and the class imbalance between old and new categories is well alleviated. Moreover, when constructing the exemplar set, few-shot categories can contain adequate samples to be prioritized and selected. By the two parts of distillation loss, the target model can fairly handle the categories with the class-balanced exemplar set.

D. Weighted Class Mean Computation in Semi-FSCIL

In our proposed framework, we use nearest-mean-of-exemplars classification, *i.e.*, computing the class-means (*i.e.*, prototypes) in the first step and assigning the labels to the test samples based on the distances with class-means in the feature space. The class-mean is a prototype by averaging features of all reserved samples for each class. In the FSCIL scenario, there are only a few labeled samples per class in incremental sessions. If the prototype of a particular category is generated based on the features of a few samples in each class, the classification performance would not be satisfied. Since a few samples can not model the data distributions or the averaging features of a few examples can not stand for the class means, the overfitting problems will be too severe to perform classification tasks better.

In the proposed Semi-FSCIL, unlabeled samples together with generated pseudo labels are selected, and join the training process with the labeled dataset in each incremental session. The exemplar set is sampled from the labeled datasets and selected unlabeled samples of previous sessions after being prioritized by herding. In this way, we compute the prototypes with labeled data and unlabeled data together. Assuming $\overline{\mathcal{P}}_c$ as the prototype of class c, the $\overline{\mathcal{P}}_c$ is defined as

$$\overline{\mathcal{P}}_c = \frac{\gamma^l \sum_{i=0}^{|\mathcal{E}_c^l|} \Theta(\boldsymbol{x}_i) + \gamma^u \sum_{i=0}^{|\mathcal{E}_c^u|} \Theta(\boldsymbol{x}_i)}{|\mathcal{E}_c^l| + |\mathcal{E}_c^u|}, \tag{8}$$

where $\mathcal{E}_c^l$ and $\mathcal{D}_c^u$ are the labeled exemplars and unlabeled exemplars of class c respectively, and $\Theta(\cdot)$ is the feature extractor model. γ^l and γ^u are the weights for the two parts of the contributions for prototypes. When we conduct classification tasks, the distances $\mathcal{O}$ between a particular unlabeled sample $\boldsymbol{x}^u$ and the prototypes are computed first as follows:

$$\mathcal{O} = [\boldsymbol{f_{dis}}(\Theta(\boldsymbol{x}^u), \overline{\mathcal{P}}_c)]_{c=0}^{|\overline{\mathcal{P}}|}, \tag{9}$$

where $\overline{\mathcal{P}}$ is the set of prototypes, and $\boldsymbol{f_{dis}}$ is the distance function. The label of $\boldsymbol{x}^u$ is assigned by

$$\hat{y} = \mathbb{1}\left[O^c == \boldsymbol{min}(\mathcal{O})\right]_{c=0}^{|\mathcal{O}|}, \tag{10}$$

where $\boldsymbol{min}$ is the minimum function. By combining the selected unlabeled data into computing prototypes, the overfitting problem can be alleviated since more data can better model the distributions. After computing the prototypes, the framework completes the training and testing process of a specific session.

E. Discussion

The proposed Semi-FSCIL is a novel task that is not explored by existing methods. We solve this task by a distillation-based incremental learning framework, in which one of the important parts is uncertainty-guided semi-supervised learning. The uncertainty-guided strategy used for pseudo-labeling-based semi-supervised learning first appeared in [62]. In this paper, we employ a simple implementation of the uncertainty-guided module.

In our proposed framework, we apply nearest-mean-of-exemplars for classification, and update class means by labeled and unlabeled samples together. In this way, it can approximate to the average features of all samples in a total dataset and largely alleviate the overfitting problem. Moreover, labeled data and unlabeled data all contribute to the distillation process; thus the bias can become less severe when the model tries to memorize old categories.

V. Experiments

A. Dataset

To evaluate the effectiveness of the proposed Us-KD for Semi-FSCIL, we conducted extensive experiments on three datasets: CIFAR100 [64], *mini*ImageNet [33] and CUB200 [65]. Figure 3 gives examples of these three datasets.

CIFAR100 [64] is widely used in class-incremental learning. It includes 100 classes with 600 RGB images per class. For each category, 500 images are used for training and 100 images for testing. The size of the image is 32×32. ***mini*ImageNet** [33] is a subset of the ImageNet with small number of classes. It includes 600 images for each of 100 classes. These images are in the size of 84×84. This dataset is also widely-used in few-shot learning tasks. **CUB200** [65] contains about $6,000$ training images and $6,000$ test images of over 200 bird categories. The images are resized to 256×256 and then cropped to 224×224 for training.

For CIFAR100 and *mini*ImageNet, we set 60 and 40 classes as the base and novel categories, respectively, and chose a 5-way 5-shot setting. In total, we had 9 training sessions, *i.e.*, one session for base classes and 8 sessions for novel classes. While for CUB200, we adopted the 10-way 5-shot setting by choosing 100 classes as base classes and splitting the remaining 100 classes into 10 incremental learning sessions. For the sessions of learning novel categories, each session's training set was constructed by randomly choosing 5 training instances per class from the original training set to construct 5/10-way 5-shot tasks, and some instances were also picked randomly from the rest of the training set, and their labels were discarded to construct the unlabeled set. In each incremental learning session, the size of unlabeled dataset is 500. We used the whole test set for the evaluation purpose, which was enough to evaluate the model's generalization ability.

B. Network Setting and Implementation Details

In our proposed framework, we applied ResNet-18 [2] for CIFAR100, *mini*ImageNet and CUB200, which follows the same setting as [15]. For CIFAR100 and *mini*ImageNet, we load the ResNet-18 model without pre-training. While for CUB200, pre-trained ResNet-18 is applied. This configuration also follows previous FSCIL methods for the fair comparison. We implemented the models using Pytorch and trained on GeForce RTX 2080 GPUs. For the first session of CIFAR100 and *mini*ImageNet, the learning rate started from 0.1 and was divided by 10 after 80 and 120 epochs (160 epochs in total). For the rest sessions of CIFAR100, the learning rate was 0.01, and we used early stopping to avoid overfitting. For the rest sessions of *mini*ImageNet, the learning rate was 0.001, and the supervised epoch number was 160. For CUB200, the base learning rate in the first session was 0.0005, and divided by 10 after 80 and 120 epochs (160 epochs in total). The learning rate of the following sessions was 0.001 used in 160 supervised epochs.

The models were trained by SGD with the training batch size of 128 for CIFAR100 and *mini*ImageNet, and 32 for CUB200. The test batch size for CIFAR100 and *mini*ImageNet is 100, and for CUB200, this batch size is 50. For the class means, the dimension was set as 512. The learning process for each dataset was repeated 10 times, and we reported the average overall test accuracy. In the uncertainty-guided module of semi-supervised learning, the threshold κ_u was assigned dynamically. In our experiments for the three datasets, we selected unlabeled samples from an unlabeled set with a size of 500. The uncertainties of the prediction on 500 unlabeled samples were ranked in ascending order. κ_u was set as the 100-th value in the rank. For each unlabeled iteration, we chose 5 unlabeled samples to join the training procedure. For CIFAR100, we selected 150 unlabeled samples in the 30 unlabeled iterations. For *mini*ImageNet, we selected 150 unlabeled samples in around 30 unlabeled iterations. For CUB200, 75 unlabeled samples were added into the labeled dataset during around 15 unlabeled iterations. In each unlabeled iteration, the model was trained on the labeled dataset and the selected unlabeled dataset with pseudo labels for extra 10 epochs. The model was not trained on labeled data and selected unlabeled data from the reference model with more epochs to guarantee that labeled samples contributed more to the training process. Moreover, during the 10 epochs, the training dataset was not set randomly to ensure that the model was trained on labeled data first and losses could be generated from unlabeled samples. In Equation 7 and Equation 8, the weights were all set to 1. We did not give different weights, since the unlabeled exemplars selected in the uncertainty-guided semi-supervised learning process and prioritized based on herding ensures that the quality of unlabeled data can have an equal contribution with labeled samples to the knowledge distillation process.

For CIFAR100, we preserve 20 exemplars for each previous category, which maintains the same as class-incremental learning task [11]. *mini*ImageNet contains the same number of categories, *i.e.*, 100 categories, so we also preserve 20 exemplars per category. CUB200 has 200 categories, so we store 10 exemplars per category.

C. Comparative studies

In this section, we conducted extensive experiments to demonstrate the performance of the proposed Us-KD for Semi-FSCIL. For the three datasets, the results of most comparison methods are directly quoted from their original papers to facilitate fair comparison. As for the evaluation indicators, we used the overall accuracy (%) on the test set of all seen categories in the current sessions and a performance dropping rate (PD) [51] that measures the absolute accuracy drops in the last session w.r.t. the accuracy in the first session.

CIFAR100 Figure 4 (a) shows the detailed comparison results of the proposed approaches and many existing

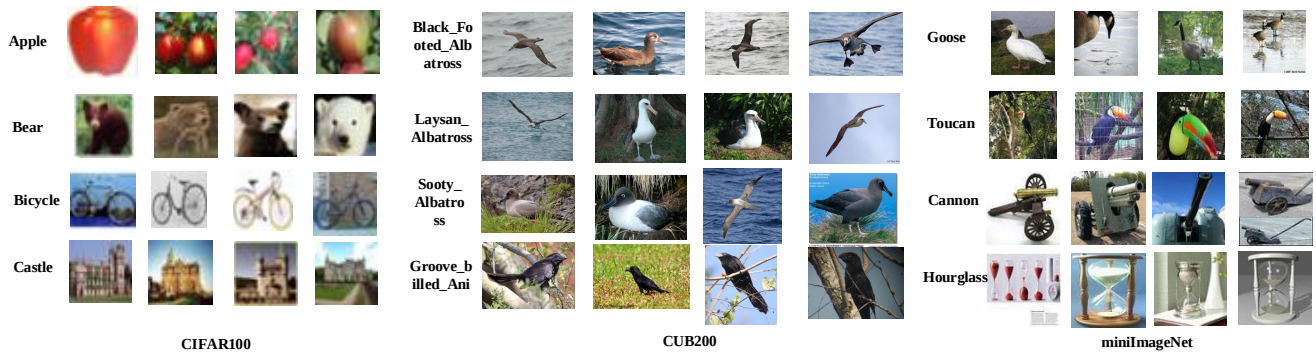


Fig. 3. Example images from the CIFAR100, CUB200 and *mini*ImageNet datasets used in our experiments.

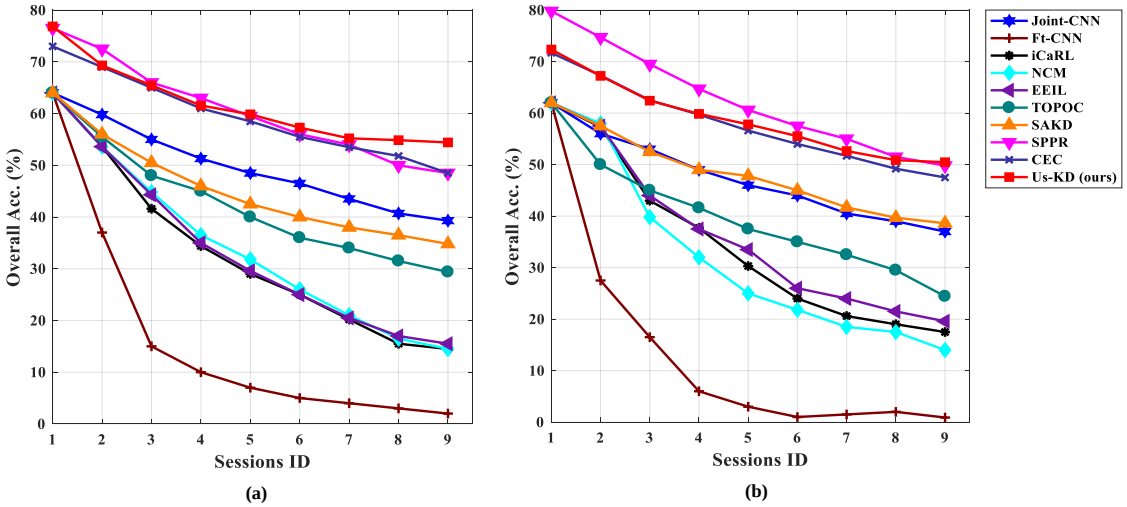


Fig. 4. Comparative study on CIFAR100 dataset (a) and *mini*ImageNet (b). We obtain the state-of-the art results with the proposed methods.

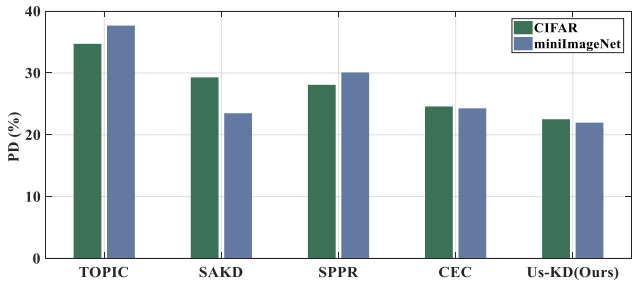


Fig. 5. Performance dropping rate of CIFAR100 and *mini*ImageNet. The deterioration from the first session to the last session in our method is much slighter than other related methods.

approaches in 9 learning sessions. It shows that the proposed Us-KD achieves the highest overall accuracy in the last incremental learning session, which is superior to the recently proposed SPPR [23] by around 5.92% in the final session. Compared to the baselines SPPR [23] and CEC [51], the performance of Us-KD is sometimes inferior in the first 5 sessions. In our experiments, we introduced an extra indicator PD to demonstrate the efficiency of our proposed method. Figure 5 illustrates the PD results, and our methods suffer light deterioration.

***mini*ImageNet** Figure 4 (b) lists the accuracy of the proposed approach and existing several representative methods on the *mini*ImageNet dataset. Our method Us-KD outperforms the other baseline methods. The overall accuracy of the proposed Us-KD (50.47%) outperforms the state-of-the-art [23] by around 0.67% in the last session. Moreover, from Figure 5, we can see the proposed method dramatically alleviate the deterioration of incremental learning sessions, demonstrating the efficiency of the proposed Us-KD in the distillation process.

CUB200 Table I presents the accuracies over 11 learning sessions on CUB200. Our proposed Us-KD outperforms the other related methods with only one strategy for incremental learning. Although our method did not achieve the state-of-the-art result (57.81%) [22] in the last session, superior results of PD are achieved by our algorithm Us-KD.

D. Ablation studies

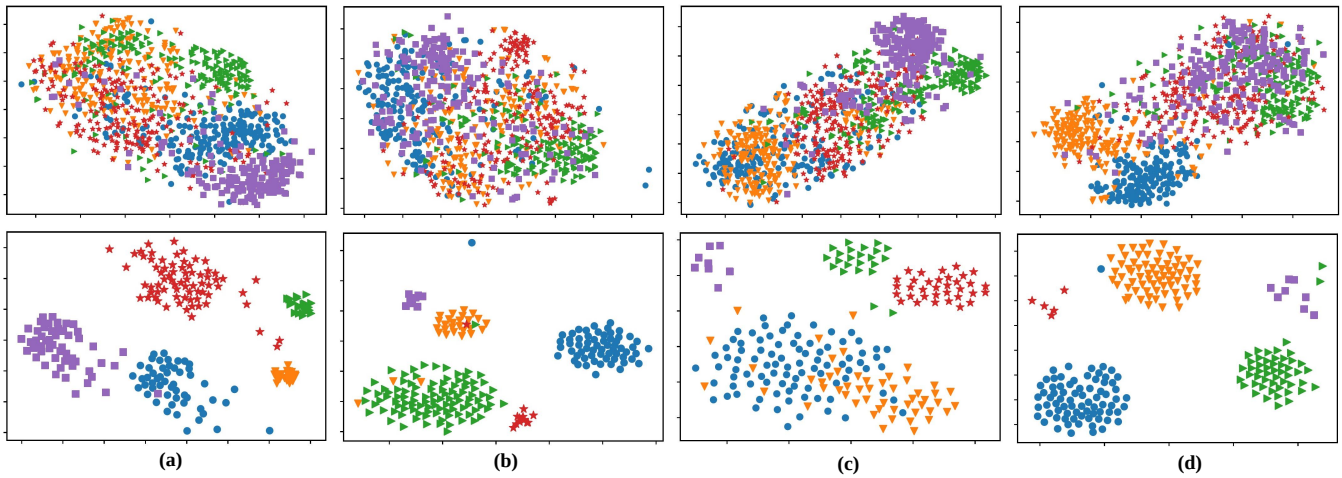
To demonstrate the efficiency of Us-KD, we conducted a series of ablation experiments.

The impact of the uncertainty-guided module. Table II show the ablation study results. Ss-KD stands for the framework obtained by replacing uncertainty-guided semi-supervised learning with the standard one (*i.e.*, Equation 1). As shown in Table II, Us-KD can not perform better without the uncertainty-guided module in the semi-supervised learning process, because much noise information was added when unlabeled samples were combined into incremental

TABLE I

THE CLASSIFICATION ACCURACIES (%) OF CUB200. PD IS THE PERFORMANCE DROPPING RATE FROM THE FIRST SESSION TO THE LAST SESSION. S1 REPRESENTS THE KNOWLEDGE REPRESENTATION AND REFINEMENT STRATEGY, AND S2 STANDS FOR THE KNOWLEDGE DISTILLATION STRATEGY. OUR RESULTS ACHIEVE LESS PD THAN OTHER RELATED METHODS ONLY WITH THE S2.

Method	S1	S2	Session ID											PD↓
			1	2	3	4	5	6	7	8	9	10	11	
Ft-CNN [15]			68.68	44.81	32.26	25.83	25.62	25.22	20.84	16.77	18.82	18.25	17.18	51.50
Joint-CNN [15]			68.68	62.43	57.23	52.80	49.50	46.10	42.80	40.10	38.70	37.10	35.60	33.08
EEIL [63]		✓	68.68	53.63	47.91	44.20	36.30	27.46	25.93	24.70	23.95	24.13	22.11	46.57
iCaRL [10]		✓	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	47.52
NCM [11]		✓	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87	48.81
TOPIC [15]	✓		68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28	42.40
SAKD [21]		✓	68.23	60.45	55.70	50.45	45.72	42.90	40.89	38.77	36.51	34.87	32.96	35.27
SPPR [23]	✓		68.68	61.85	57.43	52.68	50.19	46.88	44.65	43.07	40.17	39.63	37.33	31.35
CEC [51]	✓		**75.85**	**71.94**	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	23.57
Us-KD (ours)		✓	74.69	71.71	**69.04**	**65.08**	**63.60**	**60.96**	**59.06**	**58.68**	**57.01**	**56.41**	**55.54**	**19.15**



(a) (b) (c) (d)

Fig. 6. (a)-(d) The t-SNE visualization of network activations on *mini*ImageNet. We present four visualization examples in the figure. The tops illustrate the visualization results on unlabeled samples when assigned the labels only based on the confidence predictions. The bottoms show the visualization results on the unlabeled samples when the pseudo labels are assigned and filtered based on the proposed uncertainty-guided module.

TABLE II

THE ABLATION STUDY ON THE UNCERTAINTY-GUIDE SEMI-SUPERVISED LEARNING. PD IS THE PERFORMANCE DROPPING RATE FROM THE FIRST SESSION TO THE LAST SESSION. SS-KD STANDS FOR THE FRAMEWORK WITH THE STANDARD SEMI-SUPERVISED LEARNING (*i.e.*, EQUATION 1). FROM THE TABLE, WE CAN CONCLUDE THAT THE MODEL CAN PERFORM BETTER WITH THE UNCERTAINTY-GUIDED OPERATION IN THE SEMI-SUPERVISED LEARNING PROCESS.

Dataset	Method	Session ID											PD↓
		1	2	3	4	5	6	7	8	9	10	11	
CIFAR100	Ss-KD	76.85	69.82	65.43	62.25	59.65	**57.46**	**55.73**	54.89	53.47	-	-	23.38
	Us-KD	76.85	**69.87**	**65.46**	**62.36**	**59.86**	57.29	55.22	**54.91**	**54.42**	-	-	**22.43**
***mini*ImageNet**	Ss-KD	72.35	67.05	61.97	58.45	55.77	53.48	50.56	50.15	49.26	-	-	23.09
	Us-KD	72.35	**67.22**	**62.41**	**59.85**	**57.81**	**55.52**	**52.64**	**50.86**	**50.47**	-	-	**21.88**
CUB200	Ss-KD	74.69	70.92	66.73	63.68	61.94	58.74	56.94	55.75	53.65	53.15	52.95	21.74
	Us-KD	74.69	**71.71**	**69.04**	**65.08**	**63.60**	**60.96**	**59.06**	**58.68**	**57.01**	**56.41**	**55.54**	**19.15**

TABLE III

COMPARATIVE RESULTS OF HARD AND SOFT PSEUDO LABELS.

Type	Session ID									PD↓
	1	2	3	4	5	6	7	8	9	
Soft	76.85	64.77	61.49	57.63	54.20	52.13	51.14	49.77	48.83	28.02
Soft (with sharpening)	76.85	67.40	62.06	59.47	57.81	55.71	53.51	51.08	50.02	26.83
Hard	**76.85**	**69.87**	**65.46**	**62.36**	**59.86**	**57.29**	**55.22**	**54.91**	**54.42**	**22.43**

TABLE IV

THE ABLATION STUDY OF CUB200 ON THE NUMBER OF UNLABELED SAMPLES ADDED IN EACH INCREMENTAL LEARNING SESSION. PD IS THE PERFORMANCE DROPPING RATE FROM THE FIRST SESSION TO THE LAST SESSION.

Num_U	Session ID											PD↓
	1	2	3	4	5	6	7	8	9	10	11	
25	74.69	70.31	67.06	63.17	60.90	57.65	56.44	54.77	52.50	52.99	51.45	23.24
50	74.69	71.33	68.43	64.14	61.84	59.36	57.30	55.67	54.65	55.10	54.45	20.24
75	**74.69**	71.71	**69.04**	65.08	63.60	**60.96**	**59.06**	**58.68**	**57.01**	**56.41**	**55.54**	**19.15**
100	74.69	**72.13**	68.80	64.99	63.51	61.37	58.65	58.26	56.63	55.82	54.97	19.72
125	74.69	71.59	68.98	**65.67**	**63.68**	60.43	59.05	58.36	56.42	55.53	54.94	19.75

learning sessions, which demonstrates the importance of the uncertainty-guided module. Moreover, in Figure 6, we show the t-SNE [66] visualization results on *mini*ImageNet. When we assigned and filtered pseudo labels based on our uncertainty-guided module (bottoms) in semi-supervised learning, the decision boundary is much clearer than that of only based on the confidence prediction (tops).

The impact of the pseudo label type. In the semi-supervised learning process, we compare the overall classi-

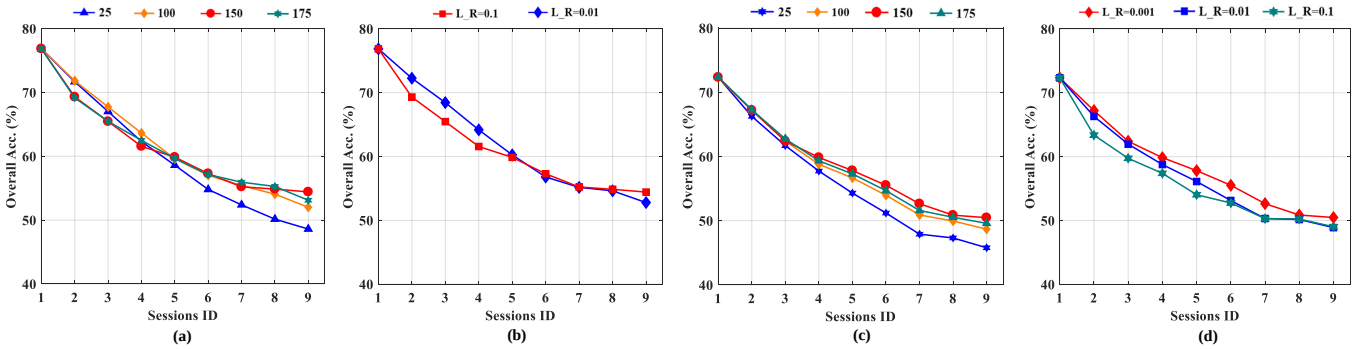


Fig. 7. Ablation study results on the number of unlabeled samples (a, c) and the learning rate (b, d). (a) Ablation study on the number of unlabeled samples by CIFAR100. With 150 unlabeled samples in each incremental learning session, we obtain the highest overall classification accuracy. (b) Ablation study on the learning rate by CIFAR100. When the learning rate is 0.1, the model performs better than that when the learning rate is 0.01. (c) Ablation study on the number of unlabeled samples by *mini*ImageNet. With 150 unlabeled samples in each incremental learning session, we obtain the highest overall classification accuracy. (d)Ablation study on the learning rate by *mini*ImageNet. When the learning rate is 0.001, the model performs better than that when the learning rate is 0.01 or 0.1.

fication performance with hard pseudo labels and soft pseudo labels. The results are illustrated in Table III. We also sharpen the pseudo labels by assigning the temperature as 0.5. Soft pseudo labels contains more semantic information than one-hot labels, while they may also be noisy. In this way, if we use the soft pseudo labels directly, the overall classification performance is inferior to that of using hard pseudo labels. Then, we sharpen the soft pseudo labels to enhance the quality of soft pseudo labels, and the overall classification performance is promoted. When hard pseudo labels is used in our proposed method, the superior performance is obtained.

The impact of the number of unlabeled samples added in each incremental learning session. From different curves in Figure 7 (a) and Figure 7 (c), we can conclude that the model achieves the best performance on CIFAR100 and *mini*ImageNet dataset when 150 unlabeled samples added in each incremental learning sessions. The ablation study results of CUB200 is illustrated in Table IV. With 75 unlabeled samples added, our framework can achieve the best performance.

The impact of the learning rate in each incremental learning session. Figure 7 (b) presents the results of CIFAR100. When setting the learning rate as 0.1, the performance of the model suffered from a sharp decrease first, and then reduced slightly; thus the overall accuracy of the last session is higher than that of when setting the learning rate as 0.01. The ablation study results on the learning rate with *mini*ImageNet is shown in Figure 7 (d). When the learning rate is 0.001 in the incremental learning session, the model outperforms in all sessions.

The impact of τ. τ is the temperature for the distillation loss in Equation 6. As described in the paper, $\tau(\nu_i) = \nu_i^{1/\Omega} / \sum_j \nu_j^{1/\Omega}$, and the hyperparameter is Ω in this function. In our paper, Ω is termed 3. In this way, we conduct the experiment on the impact of Ω, which results of CIFAR100 are illustrated in Table V. We can conclude that the framework has the superior performance when Ω is 3.

The impact of γ^l and γ^u. γ^l, and γ^u are the weights that labeled data and unlabeled data contribute to the prototype, respectively. Table VI shows the ablation study results on the weights. When γ^l and γ^u are all assigned to 1, superior results are obtained.

TABLE V
ABLATION STUDY RESULTS ON Ω.

Ω	Session ID									PD↓
	1	2	3	4	5	6	7	8	9	
2	76.85	68.03	64.26	61.00	58.46	56.98	55.62	54.44	52.72	24.13
3	**76.85**	**69.87**	65.46	62.36	59.86	**57.29**	55.22	**54.91**	**54.42**	**22.43**
4	76.85	69.46	66.40	62.57	60.33	57.21	55.87	55.43	53.36	23.49

TABLE VI
ABLATION STUDY RESULTS ON γ^l AND γ^u WITH CIFAR100.

γ^l	γ^u	Session ID									PD↓
		1	2	3	4	5	6	7	8	9	
1.0	**0.8**	76.85	69.18	65.50	61.93	**60.06**	**58.16**	**56.49**	**55.32**	53.24	23.61
0.8	**1.0**	76.85	66.91	63.46	60.68	58.11	56.34	55.24	54.49	53.02	23.83
1.1	**0.9**	76.85	69.09	**65.87**	62.19	59.17	57.22	56.29	55.15	53.98	22.87
0.9	**1.1**	76.85	68.83	65.93	61.72	59.67	57.09	55.90	54.80	53.29	23.56
1	**1**	**76.85**	**69.87**	65.46	**62.36**	59.86	57.29	55.22	54.91	**54.42**	**22.43**

The impact of the backbone. In our existing experiments, ResNet-18 [2] was applied as the backbone for the fair comparison with the existing FSCIL methods. In Table VII, we present the performance of CUB200 with the ResNet-50 [2] and Vit-Tiny [67] as the backbone, respectively. The same number of unlabeled samples are combined into the training process. With our proposed framework, no matter which backbone you use, FSCIL can be benefited from unlabeled samples with the semi-supervised learning technique. By comparing the results, we can find that the accuracy experiences a slightly decreasing trend by using ResNet-50 as the backbone than ResNet-18 and ViT-Tiny. Moreover, with ViT-Tiny, FSCIL can gain more from the incorporated unlabeled samples.

The impact of the number of exemplars per class. We show the ablation study results on the number of exemplars with CUB200 in Table VIII. Theoretically, storing more exemplars can better preserve old knowledge. However, the performance trade-off between old and new categories should be handled, and there may be inadequate samples for being selected as the exemplar set.

VI. CONCLUSION AND FUTURE WORK

In this paper, we conduct a novel task named semi-supervised few-shot class-incremental learning (Semi-FSCIL), aiming at incrementally learning few-shot novel tasks by

TABLE VII
ABLATION STUDY ON THE BACKBONE WITH CUB200.

Backbone	Unlabeled data	Session ID											PD↓
		1	2	3	4	5	6	7	8	9	10	11	
ResNet-50 [2]	✗	78.19	75.13	72.14	68.82	67.84	66.10	65.13	64.22	62.34	62.66	62.10	16.09
	✓	78.19	**76.07**	**73.17**	**71.27**	**70.26**	**68.47**	**66.02**	**66.12**	**65.54**	**64.86**	**64.21**	**13.98**
ViT-Tiny [67]	✗	81.84	77.95	74.52	70.92	69.04	66.39	64.54	64.04	61.35	60.65	60.22	21.62
	✓	81.82	77.82	**75.19**	**72.08**	**69.22**	**67.52**	**66.84**	**66.32**	**65.10**	**64.52**	**63.93**	**17.89**
ResNet-18 [2]	✗	74.69	69.17	66.28	62.70	60.51	57.78	56.63	54.30	53.05	52.93	52.02	22.67
	✓	74.69	**71.71**	**69.04**	**65.08**	**63.60**	**60.96**	**59.06**	**58.68**	**57.01**	**56.41**	**55.54**	**19.15**

TABLE VIII
RESULTS OF CUB200 WITH DIFFERENT NUMBERS OF EXEMPLARS PER CATEGORY.

#exemplars	Session ID											PD↓
	1	2	3	4	5	6	7	8	9	10	11	
5	74.69	69.62	65.57	60.94	59.71	55.64	53.12	51.88	50.25	49.42	47.93	26.76
10	74.69	71.71	69.04	65.08	63.60	60.96	59.06	58.68	57.01	56.41	55.54	19.15
20	74.69	73.05	70.23	66.87	65.29	63.36	62.25	61.23	59.74	58.76	57.37	17.32
30	**74.69**	**73.57**	**72.01**	**68.64**	**67.80**	**64.29**	**63.00**	**62.82**	**61.28**	**60.63**	**59.67**	**15.02**

incorporating unlabeled data in a semi-supervised manner. Few-shot class-incremental learning tasks always suffer from catastrophic forgetting old categories and overfitting on few-shot novel categories. This paper also provides a simple but efficient solution to tackle these two problems in Semi-FSCIL. Technically, we implement Semi-FSCIL by proposing a knowledge distillation framework with the uncertainty-guided semi-supervised learning (Us-KD). The uncertainty-guided operation assigns pseudo labels and filters unlabeled samples by taking the prediction uncertainty into consideration. Unlabeled samples can alleviate the class-imbalance problem first, then contribute to the distillation process and class means for better memorizing old categories and model the data distribution, respectively. Comparative experiments on three gold FSCIL datasets demonstrate that Us-KD yields remarkable results compared with many existing methods in overall classification accuracy and performance dropping rate. Besides, extensive ablation experiments clearly show the effectiveness of Us-KD in significantly improving the performance of the proposed novel Semi-FSCIL task.

Despite the effectiveness of the Us-KD, there are still some aspects that can be further improved. (1) The proposed Us-KD aims at solving the novel task Semi-FSCIL, *i.e.*, the training dataset in each incremental learning session contains a labeled set and an unlabeled set, where the few-shot learning task is conducted in a semi-supervised manner. The length of the task sequence in the benchmark dataset is around 10, while our proposed distillation-based framework can not perform well when the task sequence becomes longer. This scenario is donated as Large-Scale FSCIL (LSFSCIL) which is the ultimate goal for FSCIL that has not been explored in the current literature. In the future work, we will dedicate to solving the issues in FSCIL task first, then exploring solutions to the LSFSCIL problem. (2) In Us-KD, unlabeled samples are selected based on the uncertainty-aware module. However, we did not take class imbalance between pseudo labels into consideration. We will extend the Us-KD to solve this class-balance problem. (3) The proposed approach has the potential for adapting to other visual recognition tasks, including medical imagery analysis, but it is beyond the current scope of this paper. We leave this research on the generalization ability as the future work.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105. 1

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 1, 8, 11, 12

[3] X. Wang and G.-J. Qi, "contrastive learning with stronger augmentations," *arXiv preprint arXiv:2104.07713*, 2021. 1

[4] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International Jornal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020. 1

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699. 1

[6] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016, pp. 1060–1069. 1

[8] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring generalization in deep learning," in *NeurIPS*, 2017, pp. 5947–5956. 1

[9] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, no. 2, pp. 245–258, 2017. 1

[10] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017, pp. 2001–2010. 1, 4, 5, 6, 10

[11] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *CVPR*, 2019, pp. 831–839. 1, 3, 4, 6, 8, 10

[12] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017. 1, 3

[13] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *ICML*, 2017, pp. 3987–3995. 1, 3

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS Workshop*, 2015. 1, 3

[15] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *CVPR*, 2020, pp. 12 183–12 192. 2, 4, 8, 10

[16] T. Chen, S. Wu, X. Yang, Y. Xu, and H.-S. Wong, "Semantic regularized class-conditional gans for semi-supervised fine-grained image synthesis," *IEEE Transactions on Multimedia*, 2021. 2, 4

[17] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, 1993. 2

[18] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, no. 33, 2011. 2

[19] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *CVPR*, 2019, pp. 7260–7268. 2

[20] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, 2020. 2

[21] A. Cheraghian, S. Rahman, P. Fang, S. K. Roy, L. Petersson, and M. Harandi, "Semantic-aware knowledge distillation for few-shot class-incremental learning," in *CVPR*, 2021, pp. 2534–2543. 2, 4, 10

[22] K. Chen and C.-G. Lee, "Incremental few-shot learning via vector quantization in deep embedded space," in *ICLR*, 2021. 2, 9

[23] K. Zhu, Y. Cao, W. Zhai, J. Cheng, and Z.-J. Zha, "Self-promoted prototype refinement for few-shot class-incremental learning," in *CVPR*, 2021, pp. 6801–6810. 2, 4, 9, 10

[24] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018. 2, 4

[25] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *NeurIPS*, 2019. 2, 4

[26] Y. Cui, W. Xiong, M. Tavakolian, and L. Liu, "Semi-supervised few-shot class-incremental learning," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1239–1243. 2
[27] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *ECCV*, 2010, pp. 127–140. 3
[28] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, and L. Sigal, "Multi-level semantic feature augmentation for one-shot learning," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4594–4605, 2019. 3
[29] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 441–449. 3
[30] Z. Yu, L. Chen, Z. Cheng, and J. Luo, "Transmatch: A transfer-learning scheme for semi-supervised few-shot learning," in *CVPR*, 2020, pp. 12 856–12 864. 3
[31] Y. Wang, C. Xu, C. Liu, L. Zhang, and Y. Fu, "Instance credibility inference for few-shot learning," in *CVPR*, 2020, pp. 12 836–12 845. 3
[32] G.-J. Qi, W. Liu, C. Aggarwal, and T. Huang, "joint intermodal and intramodal label transfers for extremely rare or unseen classes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1360–1373, 2016. 3
[33] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *NeurPIS*, 2016, pp. 3630–3638. 3, 8
[34] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017, pp. 4080–4090. 3
[35] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, pp. 1126–1135. 3
[36] Y.-X. Wang and M. Hebert, "Learning to learn: Model regression networks for easy small sample learning," in *ECCV*, 2016, pp. 616–634. 3
[37] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *ICML*, 2016, pp. 1842–1850. 3
[38] P. Shyam, S. Gupta, and A. Dukkipati, "Attentive recurrent comparators," in *ICML*, 2017, pp. 3173–3181. 3
[39] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *ICLR*, 2018. 3
[40] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *ICLR*, 2018. 3
[41] K. Hsu, S. Levine, and C. Finn, "Unsupervised learning via meta-learning," in *ICLR*, 2018. 3
[42] S. Khodadadeh, L. Boloni, and M. Shah, "Unsupervised meta-learning for few-shot image classification," in *NeurIPS*, 2019, pp. 10 132–10 142. 3
[43] Z. Ji, X. Zou, T. Huang, and S. Wu, "Unsupervised few-shot feature learning via self-supervised training," *Frontiers In Computational Neuroscience*, vol. 14, 2020. 3
[44] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165. 3
[45] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017. 3
[46] H. Jung, J. Ju, M. Jung, and J. Kim, "Less-forgetful learning for domain expansion in deep neural networks," in *AAAI*, 2018. 3
[47] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *ECCV*, 2018, pp. 139–154. 3
[48] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 35–44. 3
[49] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, "generalized deep transfer networks for knowledge propagation in heterogeneous domains," 2016. 3
[50] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *CVPR*, 2017, pp. 3366–3375. 3
[51] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, "Few-shot incremental learning with continually evolved classifiers," in *CVPR*, 2021, pp. 12 455–12 464. 4, 8, 9, 10
[52] W. Shi, Y. Gong, C. Ding, Z. M. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *ECCV*, 2018, pp. 299–315. 4
[53] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop*, vol. 3, no. 2, 2013. 4
[54] T. Joachims, "Transductive learning via spectral graph partitioning," in *ICML*, 2003, pp. 290–297. 4
[55] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *NeurIPS*, 2016, pp. 2352–2360. 4
[56] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003, pp. 912–919. 4
[57] B. Liu, Z. Wu, H. Hu, and S. Lin, "Deep metric transfer for label propagation with limited annotated data," in *ICCV Workshop*, 2019, pp. 0–0. 4
[58] T. Zhou, S. Wang, and J. Bilmes, "Time-consistent self-supervision for semi-supervised learning," in *ICML*, 2020, pp. 11 523–11 533. 4
[59] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *CVPR*, 2019, pp. 5070–5079. 4
[60] T. Joachims *et al.*, "Transductive inference for text classification using support vector machines," in *ICML*, vol. 99, 1999, pp. 200–209. 4
[61] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *International workshop on artificial intelligence and statistics*. PMLR, 2005, pp. 57–64. 5
[62] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *ICLR*, 2021. 6, 7
[63] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *ECCV*, 2018, pp. 233–248. 10
[64] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009. 8
[65] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-gem," in *ICLR*, 2018. 8
[66] V. D. M. Laurens and G. Hinton, "Visualizing data using t-sne," *The Journal of Machine Learning Research*, vol. 9, p. 2579–2605, 2008. 10
[67] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "an image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020. 11, 12

Yawen Cui received B.E. degree in computer science and technology from Jiangnan University, Wuxi, China, and the M.S. degree in software engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2016 and 2019, respectively. She is currently pursuing the Ph.D. degree in Computer Science from the University of Oulu, Finland. Her research interests include few-shot learning and incremental learning.

Wanxia Deng received the university B.E. degree in electronic information science and technology from Xiamen University, Xiamen, China in 2014. She received the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China in 2016 and 2022, respectively. She is an assistant professor with the School of Meteorology and Oceanography, National University of Defense Technology, Changsha, China. Her research interests include domain adaptation, transfer learning, deep learning and remote image processing.

Xin Xu received the B.S. degree in electrical engineering from the Department of Automatic Control, National University of Defense Technology (NUDT), Changsha, China, in 1996, and the Ph.D. degree in control science and engineering from the College of Mechatronics and Automation, NUDT, in 2002. He has been a Visiting Professor with The Hong Kong Polytechnic University, the University of Alberta, the University of Guelph, and the University of Strathclyde, U.K. He is currently a Full Professor with the Institute of Unmanned Systems, College of Intelligence Science and Technology, NUDT.

Zhen Liu received the Ph.D. degree in Information and Communication Engineering from National University of Defense Technology (NUDT), in 2013. He is currently a professor with the College of Electronic Science, NUDT. He has been awarded the Excellent Young Scientists Fund on his project titled "Intelligent Countermeasure for Radar Target Recognition" in 2020. His current research interests include radar signal processing, radar electronic countermeasure, compressed sensing, and machine learning.

Zhong Liu received the Ph.D. degree in management science from the National University of Defense Technology (NUDT), Changsha, China, in 2000. He is currently a Professor with NUDT. He was a Vice Dean with the College of Systems Engineering, NUDT, where he is a Senior Advisor with the Research Center for Computational Experiments and Parallel Systems. His current research interests include planning systems, computational organization, and intelligent systems.

Matti Pietikäinen received the doctor of science degree in technology from the University of Oulu, Finland. He is an emeritus professor with the Center for Machine Vision and Signal Analysis, University of Oulu. From 1980 to 1981 and from 1984 to 1985, he visited the Computer Vision Laboratory, University of Maryland. He has made fundamental contributions, *e.g.*, to Local Binary Pattern (LBP) methodology, texture based image and video analysis, and facial image analysis. He has authored more than 350 refereed papers in international journals, books, and conferences. His papers have about 71,000 citations citations in Google Scholar (hindex 93). He was associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), the Pattern Recognition, the IEEE Transactions on Forensics and Security, and the Image and Vision Computing journals. Currently, he serves as guest editor for special issues of the IEEE Transactions on Pattern Analysis and Machine Intelligence. He was president of the Pattern Recognition Society of Finland from 1989 to 1992, and was named its honorary member in 2014. From 1989 to 2007, he served as member of the Governing Board of International Association for Pattern Recognition (IAPR), and became one of the founding fellows of the IAPR in 1994. In 2014, his research on LBP-based face description was awarded the Koenderink Prize for fundamental contributions in computer vision. He was the recipient of the IAPR King-Sun Fu Prize 2018 for fundamental contributions to texture analysis and facial image analysis. In 2018, he was named a highly cited researcher by Clarivate Analytics, by producing multiple highly cited papers in 2006-2016 that rank in the top 1 percent by citation for his field in web of science. He is a fellow of the IEEE for contributions to texture and facial image analysis for machine vision.

Li Liu received her Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2012. During her PhD study, she spent more than two years as a Visiting Student at the University of Waterloo, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory at the Chinese University of Hong Kong. From 2016.12 to 2018.11, she worked as a senior researcher at the Machine Vision Group at the University of Oulu, Finland. She was a cochair of nine International Workshops at CVPR, ICCV, and ECCV. She served as the Leading Guest Editor for special issues in *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)* and *International Journal of Computer Vision*. She is serving as the Leading Guest Editor for IEEE TPAMI special issue on "Learning with Fewer Labels in Computer Vision". Her current research interests include computer vision, pattern recognition and machine learning. Her papers have currently over 7400 citations in Google Scholar. She currently serves as Associate Editor for *IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT)*, Pattern Recognition and Pattern Recognition Letters.