Rethinking Self-Supervision for Few-Shot Class-Incremental Learning

Linglan Zhao
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
llzhao@sjtu.edu.cn

Jing Lu
University of Science and Technology of China
Hefei, China
lujing1@mail.ustc.edu.cn

Zhanzhan Cheng
Zhejiang University & Hikvision Research Institute
Hangzhou, China
11821104@zju.edu.cn

Duo Liu
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
liuduo@sjtu.edu.cn

Xiangzhong Fang
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
xzfang@sjtu.edu.cn

***Abstract*—Few-Shot Class-Incremental Learning (FSCIL) focuses on progressively absorbing new concepts given only limited training data. For tackling this challenge, several recent FSCIL works resort to pre-training models with Self-Supervised Learning (SSL) to obtain features that can generalize well to new classes. However, to avoid overfitting and catastrophic forgetting, previous works only leverage SSL in the base session and keep all or most parameters fixed in incremental sessions, resulting in inadequate adaptation to novel classes. Thus, in this paper, we explore the setting where more parameters can be updated for adapting to novel concepts, and discover that the model pre-trained with SSL leads to degraded performance even compared to that without SSL. It can be attributed to the severer forgetting of base class knowledge. To address this issue, we propose an imprinting-based distillation module for effectively regularizing the adaption process, and a mathematically provable routing strategy for further improved results. The effectiveness of our approach is verified on 3 popular FSCIL benchmarks by significantly outperforming previous methods.**

***Index Terms*—Few-shot learning, incremental learning, self-supervised learning**

I. Introduction

Class-Incremental Learning (CIL) [4]–[6] requires a model to learn novel classes with sufficient data and to resist forgetting previously learned classes. Nevertheless, the demand for abundant novel class training samples in CIL still limits its application when labeled data are prohibitively expensive to acquire. For example, when updating a face recognition model to recognize a new identity, only a single photo corresponding to that person is expected to be uploaded. As a result, Few-Shot Class-Incremental Learning (FSCIL) has become a hot topic of current research [3], [7], [8]. In FSCIL, sufficient training instances are provided only in the base session ($t = 0$) to obtain a stable initial base model, which will progressively incorporate incremental classes given very limited data in the novel sessions ($t \geq 1$). The FSCIL task presents a real challenge in which the scarcity of novel class training samples not only causes severe overfitting, but also exacerbates the notorious catastrophic forgetting.

Several recent FSCIL works [7], [9], [10] resort to Self-Supervised Learning (SSL) [11]–[13] for generalizable representation learning, since SSL explores the intrinsic structure of images without manual annotations. For further analyzing SSL in FSCIL, we conduct detailed experiments on a typical method weight imprinting [1] upon which existing works are built: a nearest-neighbor classifier is constructed using mean feature embeddings [14] as prototypes for each class. We also focus mainly on rotation-based SSL [12] which has been verified [7], [9], [10] to be the most effective in FSCIL.

As shown in the left part of Fig. 1, SSL improves weight imprinting on the joint accuracy (first row), which is consistent with [7], [9], [10]. Notably, the improvement comes mainly from better generalization to novel class (third row) thanks to SSL with similar base class results (second row). However, the performance on novel classes is still limited, since existing works only exploit SSL in the pre-training stage (base session) to enhance the generalization ability of features. Concretely, none [1], [7] or very few parameters [9], [10] can be updated leading to insufficient adaptation to novel classes.

For better adapting to new classes, we set more parameters trainable in incremental sessions and use the widely-adopted knowledge distillation [2], [3] in CIL as a regularization. The right column of Fig. 1 shows that finetuning on SSL pre-trained models even incurs degraded joint performance compared to that without SSL. After analyzing the base and novel results, we find that once fully adapting to novel classes in FSCIL, the model trained with SSL is biased to novel class performance with severe base knowledge forgetting due to data scarcity. Thus, the improvement on novel classes is overwhelmed by the dramatic decrease in base class accuracy.

For effectively preserving base knowledge, we take inspiration from weight imprinting [1] which shows desirable base performance, and propose an imprinting-based knowl-

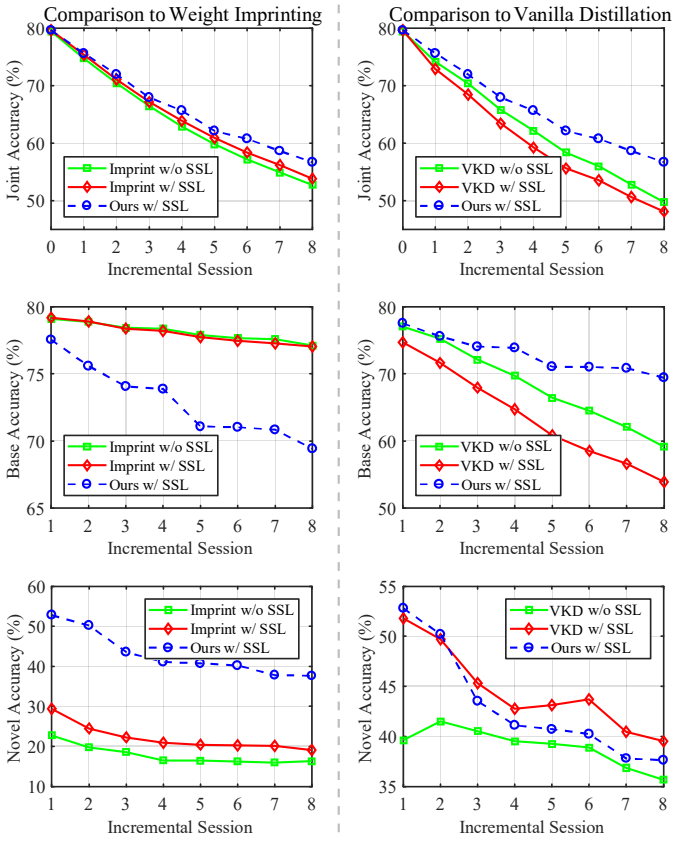

Fig. 1. Comparisons to typical few-shot class-incremental learning (FSCIL) methods: weight imprinting [1] (left) and vanilla knowledge distillation (VKD, right) [2], [3] on CIFAR100. Note that blue curves for our method in the left and right parts are exactly the same and the appeared differences are simply caused by varied coordinate axis scales. Joint accuracy, base class accuracy and novel class accuracy are presented in each row.

edge distillation module. Instead of using the previous model $(t-1)$ as a teacher [2], [3] for regularizing the training of the current model (t), our teacher model is composed of a feature extractor trained on base classes which is fixed afterward, and expandable classification weights based on imprinting [1]. As a result, the updated student model will not deviate too far away from the base model to arrest forgetting, and the generalizable knowledge learned from abundant base class data can be transferred for learning few-shot novel categories without overfitting. To enhance the generalizability of learned features, an additional self-supervision-based regularization loss is included. During inference, a mathematically provable routing strategy that dynamically selects convincing predictions between the student and teacher model is introduced for further improved results. The blue lines of Fig. 1 show that our method surpasses imprinting [1] with a better trade-off between base and novel classes, and outperforms vanilla distillation [2] by better preserving base knowledge.

The contributions of our paper are three-fold: (1) We first provide an in-depth analysis of the effect of SSL on FSCIL. Considering that direct adaptation to novel classes on SSL pre-trained models causes severe base knowledge forgetting, an imprinting-based distillation module is proposed. (2) For further improvements, we introduce a mathematically provable routing strategy that dynamically routes between predictions from the student and teacher model; (3) Extensive experiments on 3 popular FSCIL benchmarks: *mini*-ImageNet, CIFAR100, and CUB200 demonstrate the effectiveness of our method by setting a new state of the art across all the datasets.

II. Related Work

A. Few-Shot Class-Incremental Learning

The task of Few-Shot Class-Incremental Learning (FSCIL) is recently defined by [3] to train models sequentially on few-shot novel tasks while not forgetting previously learned knowledge. Concretely, TOPIC [3] proposes to use a neural gas network for maintaining the topology of features. In addition, ERL++ [15] utilizes an exemplar relation distillation approach to regularize structural relations between the current and previous models. Moreover, semantic-aware distillation [16] resorts to additional word embeddings. Besides, several current works apply self-supervision techniques for better performance. For example, FSLL [9] uses self-supervision as an auxiliary task for better representation learning. CEC [7] views rotated images as fake novel classes to train a graph model in the meta-learning phase. S3C [10] trains a stochastic classifier with rotated images and averages predictions from these images for model ensemble. However, prevailing works often keep all or a majority of model parameters frozen, thereby impeding the model's ability to adapt to novel classes. In this work, we focus on effective model adaptation to few-shot novel classes while balancing stability and plasticity in FSCIL.

B. Self-Supervised Learning

Self-Supervised Learning (SSL) utilizes predefined auxiliary tasks from unlabeled data for better feature learning. Prevailing SSL techniques include solving jigsaw puzzles [11], predicting image rotation [12] and multiview contrastive learning [13]. SSL has achieved improved performance on downstream tasks like incremental learning [17], few-shot learning [18], and FSCIL [7], [9], [10]. Following these works, we focus on rotation-based SSL [12] which has been verified to be the most effective in few-shot tasks, and provide an in-depth analysis of SSL on FSCIL which is ignored in existing works. Finally, two novel modules are proposed to handle the unique challenges when deploying SSL in FSCIL.

III. Methodology

A. Preliminary

Problem Definition. Few-Shot Class-Incremental Learning (FSCIL) [3], [7], [19] task includes a stream of training sets $\{\mathcal{S}^0, \mathcal{S}^1, \cdots, \mathcal{S}^t\}$. In each training set $\mathcal{S}^t = \{(\mathbf{x}_i, y_i)\}_i$ of session t, $\mathbf{x}_i$ is a data point with label $y_i \in \mathcal{C}^t$, and $\mathcal{C}^t$ is the label space of $\mathcal{S}^t$ which contains no overlap between classes in different sessions. Usually, a training set $\mathcal{S}^t$ $(t \geq 1)$ in incremental session is formulated as a N-way K-shot support set consisting of N incremental classes and K (*e.g.*, $K = 5$) training samples per these classes. The only exception is

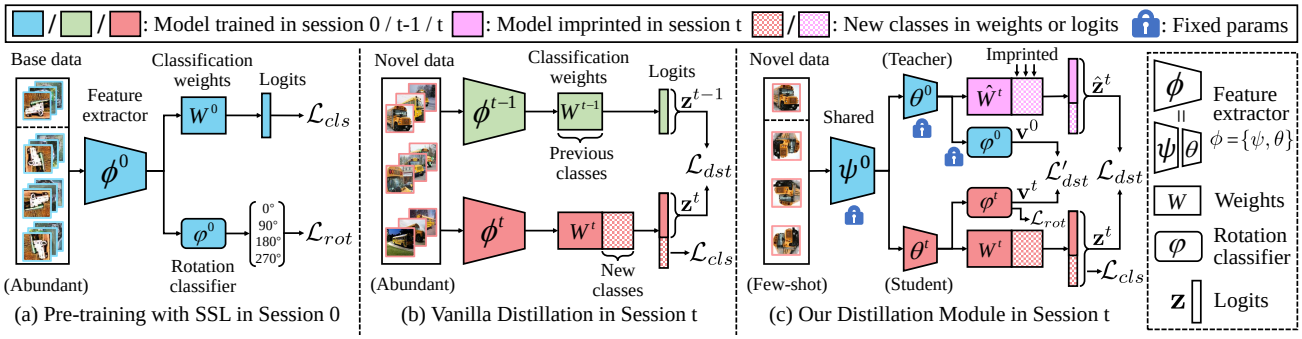

Fig. 2. Comparisons of existing approaches to our work: (a) rotation-based self-supervised training in session 0; (b) vanilla knowledge distillation in session t; (c) our imprinting-based knowledge distillation in session t. In the upper part, we use different colors to denote various components in the frameworks. In the right part, different shapes are used to explain parameterized modules and the output logits. Superscripts are used to highlight parameters and logits of the model from each session.

the training set $\mathcal{S}^0$ which is composed of abundant training samples from categories $\mathcal{C}^0$ for base model initialization. As in standard CIL, during the incremental learning in session t, only the training set $\mathcal{S}^t$ and possibly a small exemplar set $\mathcal{M}$ containing few preserved examples from previous sessions $(0, 1, \cdots, t-1)$ are provided. After training in session t, the evaluation is conducted on queries from all the ever seen categories $\tilde{\mathcal{C}}^t = \mathcal{C}^0 \cup \mathcal{C}^1 \cdots \cup \mathcal{C}^t$.

Rotation-based Self-Supervision. Rotation-based transform [12] has been verified to be the most effective SSL technique in few-shot scenarios [9], [10], [18]. As shown in Fig. 2 (a), besides classification loss $\mathcal{L}_{cls}$, a rotation classifier g_φ is trained to predict the rotated angles from input images which enforces learned features to reduce the bias towards up-right oriented images (*e.g.*, ImageNet-like) and learn more diverse features to disentangle classes for better generalization:

$$\mathcal{L}_{rot} = \mathbb{E}[-\sum_{r\in\mathcal{R}} \text{softmax}\left(g^r_\varphi\left(f_\phi(\mathbf{x}^r)\right)\right)], \tag{1}$$

where $\mathcal{R}$ is the rotation augmentation operation which transforms an input image into four possible 2D rotations $\{0°, 90°, 180°, 270°\}$[1], g^r_φ is the predicted score corresponding to rotation r, and f_ϕ is the feature extractor.

Knowledge Distillation. Knowledge distillation [2] has been widely adopted in standard CIL for preserving learned knowledge when incorporating novel classes. As shown in Fig. 2 (b), in addition to classification loss $\mathcal{L}_{cls}$, knowledge distillation loss $\mathcal{L}_{dst}$ aligns the output logits between the current model and previous model on already learned classes:

$$\mathcal{L}_{dst} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{M}}[\sum_{k=1}^{|\tilde{\mathcal{C}}^{t-1}|} -\tau_k(\mathbf{z}^{t-1};T)\log\tau_k(\mathbf{z}^t;T)], \tag{2}$$

where $\mathbf{z}^t$ and $\mathbf{z}^{t-1}$ denote logits from the current and previous models, $\tau_k(\mathbf{z};T) = \text{softmax}(\mathbf{z}_k/T)$ is the softened probability of the k-th class, and T is the distillation temperature.

[1]Using these four angles can get the best empirical results [9], [10], [12], [18] which can be efficiently implemented by transpose and flip operations.

B. Imprinting-based Distillation Module

As discussed in Section. I, vanilla knowledge distillation [2] is not suitable for FSCIL, especially on models trained with SSL. It is because when fully adapting to novel classes, the model is biased to these new classes with severe forgetting of knowledge from base classes due to data scarcity. Contrarily, weight imprinting [1] can preserve base class knowledge at the sacrifice of model's plasticity to novel classes. This inspires us to design an imprinting-based distillation module that better regularizes the adaption process without severe forgetting.

As shown in Fig. 2 (c), instead of directly using the model from previous session $(t-1)$ as guidance in vanilla distillation, the teacher model in our distillation module is composed of a fixed feature extractor pre-trained on abundant base class samples (session 0), and classification weights $\hat{W}^t$ expanded by imprinting [1]. For limiting computation cost and alleviating overfitting, we set the lower layers ψ of f_ϕ which capture basic visual patterns [20] shared, while the last residual layer θ [21] and classification weights W^t of the student can be fully updated for learning novel concepts. Hereinafter, we use f_{ϕ^0} and f_{ϕ^t} to denote the feature extractor of the teacher and the student model in session t.

Concretely, at the beginning of incremental session t, we first expand the classification weights of both teacher and student $\{\hat{W}^{t-1}, W^{t-1}\}$ in the previous session from $\mathbb{R}^{d\times|\tilde{\mathcal{C}}^{t-1}|}$ to $\mathbb{R}^{d\times|\tilde{\mathcal{C}}^t|}$ (d represents the dimension of features) for coarse novel classes accommodation, obtaining $\{\hat{W}^t, W^t\}$. The expanded weight $\boldsymbol{p}_c$ for each novel class c from $\mathcal{C}^t$ is computed by mean feature embeddings (prototype) of training examples from the corresponding class [1], [14]:

$$\boldsymbol{p}_c = \frac{\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}^t}\mathbb{I}\left[y_i=c\right]f_\phi\left(\mathbf{x}_i\right)}{\sum_{(\mathbf{x}_i,y_i)\in\mathcal{S}^t}\mathbb{I}\left[y_i=c\right]}, \tag{3}$$

where $\mathbb{I}$ is an indicator function and f_ϕ represents f_{ϕ^0} ($f_{\phi^{t-1}}$) when expanding $\hat{W}^{t-1}$ (W^{t-1}) for the teacher (student) model. After that, $\hat{W}^t$ is also fixed in the current session t, while W^t is learnable for further adaptation to novel classes.

For efficiently adapting to novel classes with limited training data and resisting the catastrophic forgetting of base

knowledge, we propose to regularize the model training by transferring the general knowledge learnt in the teacher model. Formally, the loss function for our imprinting-based knowledge distillation is defined as:

$$\mathcal{L}_{dst} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{S}^t\cup\mathcal{M}}[\sum_{k=1}^{|\tilde{\mathcal{C}}^t|} -\tau_k(\hat{\mathbf{z}}^t;T)\log\tau_k(\mathbf{z}^t;T)], \quad (4)$$

where $\mathbf{z}^t = (W^t)^\top f_{\phi^t}(\mathbf{x})$ is output logits[2] from the student model, and $\hat{\mathbf{z}}^t = (\hat{W}^t)^\top f_{\phi^0}(\mathbf{x})$ is the guidance from the teacher to force the student not deviate too far away from the stable point of base classes. Compared to Eq. 2, our distillation is not only applied to $\mathcal{M}$ for resisting forgetting of $|\tilde{\mathcal{C}}^{t-1}|$ previous classes, but also imposed on $\mathcal{S}^t$ to regularize overfitting of newly occurred classes (*i.e.*, distilling on $|\tilde{\mathcal{C}}^t|$ classes).

Moreover, since the teacher model is pre-trained with rotation-based SSL, an additional self-supervision-based regularization loss is introduced to further enhance the generalizability of learned features for the student:

$$\mathcal{L}'_{dst} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{S}^t\cup\mathcal{M}}[\sum_{k=1}^{|\mathcal{R}|} -\tau_k(\mathbf{v}^0;T')\log\tau_k(\mathbf{v}^t;T')], \quad (5)$$

where $\mathbf{v}^t = g_{\varphi^t}\left(f_{\phi^t}(\mathbf{x})\right)$ is the rotation prediction from the student model, $\mathbf{v}^0 = g_{\varphi^0}\left(f_{\phi^0}(\mathbf{x})\right)$ is that from the teacher, and $|\mathcal{R}| = 4$ denotes the four possible 2D rotations.

Finally, combined with the cross-entropy classification loss $\mathcal{L}_{cls}$ on $\mathbf{z}^t$ and the rotation prediction loss $\mathcal{L}_{rot}$ (Eq. 1) on $\mathbf{v}^t$, the total loss for student model training is derived as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rot} + w_{dst}\cdot\mathcal{L}_{dst} + w'_{dst}\cdot\mathcal{L}'_{dst}, \quad (6)$$

where w_{dst} and w'_{dst} represent two balancing weights.

C. Mathematically Provable Routing Strategy

Although solely utilizing the student model with our imprinting-based distillation for prediction can obtain promising results, the base class performance is inevitably sacrificed more or less due to adaptation to novel classes. For better retaining base class knowledge, we design a mathematically provable routing strategy that can be efficiently applied during inference.

Given a test sample $\mathbf{x}$ which is predicted into base classes by the student model: $y' = \arg\max_k \mathbf{z}^t_k \in \mathcal{C}^0$, the probability to correctly classify it into the ground-truth label y is:

$$\begin{aligned}
P_{(y'=y)} &= P_{(y\in\mathcal{C}^0)}P_{(y'=y|y\in\mathcal{C}^0)} + P_{(y\notin\mathcal{C}^0)}\underline{P_{(y'=y|y\notin\mathcal{C}^0)}} && (7)\\
&= P_{(y\in\mathcal{C}^0)}P_{(y'=y|y\in\mathcal{C}^0)} && (8)\\
&\le P_{(y\in\mathcal{C}^0)}P_{(\hat{y}'=y|y\in\mathcal{C}^0)} && (9)\\
&= P_{(y\in\mathcal{C}^0)}P_{(\hat{y}'=y|y\in\mathcal{C}^0)} + P_{(y\notin\mathcal{C}^0)}\underline{P_{(\hat{y}'=y|y\notin\mathcal{C}^0)}} && (10)\\
&= P_{(\hat{y}'=y)} = P_{\left(\arg\max_{1\le k\le|\mathcal{C}^0|}\hat{\mathbf{z}}^t_k=y\right)}, && (11)
\end{aligned}$$

where $\hat{y}'$ is the teacher's prediction only on base classes. Eq. 9 holds because, given a test sample from base classes, the teacher model can make better predictions as it is only trained on base classes and kept fixed afterward without forgetting. Eq. 8 (Eq. 10) holds since the underlined term equals zero: $y'(\hat{y}') \neq y$ when $y'(\hat{y}') \in \mathcal{C}^0$ and $y \notin \mathcal{C}^0$. Based on the above proof, the final prediction $\tilde{y}$ can be formulated:

$$\tilde{y} = \begin{cases} \arg\max_{1\le k\le|\mathcal{C}^0|}\hat{\mathbf{z}}^t_k & \text{if } y'\in\mathcal{C}^0 \\ \arg\max_{(|\mathcal{C}^0|+1)\le k\le|\tilde{\mathcal{C}}^t|}\mathbf{z}^t_k & \text{if } y'\notin\mathcal{C}^0. \end{cases} \quad (12)$$

Namely, for inputs whose pseudo labels y' from the student model belong to base classes, predictions from the teacher are used. Otherwise, we directly adopt the student's predictions.

IV. Experiments

A. Experiment Setup

Dataset Statistics. We evaluate on 3 FSCIL datasets defined by [3]: (1) regular dataset *mini*-ImageNet [22]: consisting of 60,000 84×84 images for 100 categories; (2) low-resolution dataset CIFAR100 [23]: containing 100 classes with 600 32×32 small images per class. (3) fine-grained dataset CUB200 [24]: comprising of 11,788 samples of size 224×224 from 200 bird categories. 60 classes in *mini*-ImageNet and CIFAR100 are considered as base categories and the others are divided into 8 incremental steps where 5 labeled samples (5-way 5-shot) are provided for each novel category. Moreover, 100 base classes in CUB200 are chosen and the others are divided into 10 10-way 5-shot incremental tasks.

Implementation Details. We use ResNet18 [21] as the feature extractor and optimizer SGD with a Nesterov momentum 0.9 for training. The results are obtained by averaging 10 different runs. In session $t = 0$, the model is trained on base class training data with rotation-based SSL (Fig. 2 (a)) as [9], [10], [12]. The training consists of 200 epochs from scratch with initial learning rate 0.1 which is reduced by a factor of 0.1 at the 120/160-th epoch for *mini*-ImageNet and CIFAR100. Considering that the model for CUB200 is pre-trained on ImageNet [3], [7], [8], the model is trained for 120 epochs with an initial learning rate of 0.01 reduced by a factor of 0.1 at the 50/70/90-th epoch. We use color jitter, left-right flip, and random resized crop for data augmentation. In sessions $t \ge 1$, the student model is updated for 100 iterations with learning rate 0.001, temperature hyper-parameters $T = 16$, $T' = 8$ and balancing weights $w_{dst} = 100$, $w'_{dst} = 1$.

B. Comparisons to Existing Methods

We compare our approach to recent state-of-the-art methods on 3 public FSCIL benchmarks. Following previous works [4], [19], so far we assume an exemplar set $\mathcal{M}$ can be accessed with one single exemplar reserved for each class by default. It is observed in Table I that our method outperforms the second-best result on *mini*-ImageNet dataset by 1.76% for the average accuracy and 3.64% for the final improvement. Our method is also flexible to undertake cases where none or more exemplars are available. Given 5 reserved exemplars per class [15], [19], the final accuracy can be improved by 1.01%. In a more challenging case where no exemplar is provided

[2]To alleviate the imbalance of base and novel classes, L2 normalized feature embeddings and weights [6] are used. To streamline notations, normalization is omitted in the equations.

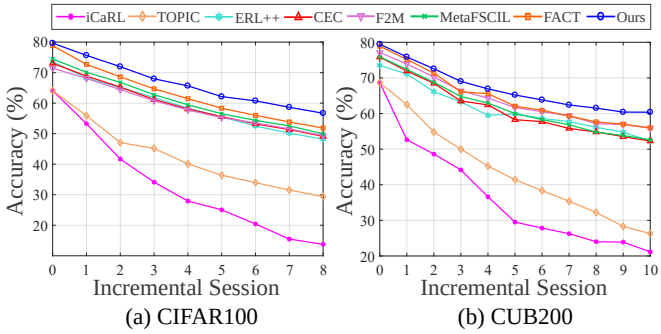

Fig. 3. Comparisons with other state-of-the-art methods on (a) CIFAR100 and (b) CUB200 datasets.

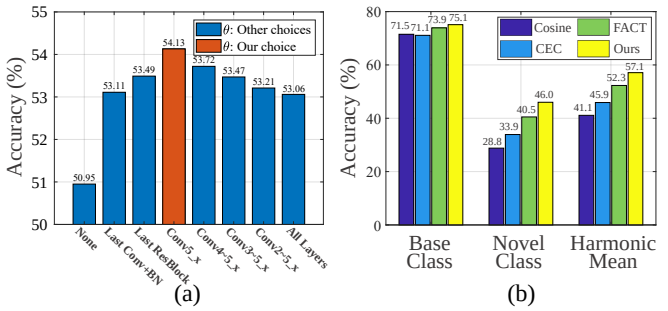

Fig. 4. Analyses of (a) selection of θ on *mini*-ImageNet dataset; (b) balancing between base and novel class performance on CUB200 dataset.

(*i.e.*, setting $\mathcal{M} = \varnothing$ in Eq. 4-6), we still acquire the final result of 52.15% which outperforms other state-of-the-arts. Moreover, Fig. 3 shows that our approach can also surpass existing approaches on the other two datasets.

C. Further Analysis

Ablation Studies. As shown in row (1)-(2) of Table II, weight imprinting [1] in row (1) is improved by 1.1% after applying rotation-based SSL (Rot). The improvement comes mainly from better generalization to novel class with 2.5% increased accuracy. However, novel category performance is still very limited since none of the parameters can be updated for new concepts. After adapting to novel classes with more parameters trainable using vanilla knowledge distillation (VKD in Eq. 2, row (3)), we observe the boosting in novel accuracy but the dramatic (about 19%) base class forgetting, which still results in degraded joint accuracy.

To alleviate forgetting, our imprinting-based knowledge distillation (IKD in Eq. 4) can better preserve base knowledge, obtaining over 2.6% improvement compared to weight imprinting. Moreover, with our self-supervision-based regularization loss (RKD in Eq. 5), the result is further improved by about 0.6%.

Finally, row (6) of Table II validates the effectiveness of our mathematically provable routing strategy by about 1.4% increase of base class performance and achieves the best overall result. Furthermore, we also compare our routing strategy to other ensemble strategies including directly averaging the two predictions (Pred-Avg) and concatenating features from both models for classification (Feat-Concat). Table III shows that our strategy surpasses all the other approaches thanks to the mathematical proof of the routing strategy.

Selection of Learnable Parameters θ. As shown in the left part of Fig. 4 (a), given only few layers learnable for adapting to novel classes, it is challenging for the model to learn novel classes since the plasticity is hindered. Also, the right part shows that, when updating excessive layers, undesirable results are observed caused by overfitting in FSCIL. Finally, the model obtains the optimal result by finetuning conv5_x in [21] (the last residual layer), validating our choice in Section III.

Balancing Base and Novel Class Performance. For further analyzing new class adaptation and base class preservation in FSCIL, our approach is evaluated on the separate base and novel class performance along with the harmonic mean. In Fig. 4 (b), our method can simultaneously obtain the highest Acc_{base}, Acc_{novel} and harmonic mean compared to other recent state-of-the-art methods. The above observation is consistent with the motivation of our paper by effectively absorbing novel knowledge and resisting the severe forgetting of base classes on SSL pre-trained model in FSCIL.

Results without SSL. Although the main focus of this work is to have a comprehensive study of self-supervised learning (SSL) in FSCIL, our proposed distillation module and the provable routing strategy are not limited to SSL scenarios. To this end, we simply pre-train our model without SSL (w/o $\mathcal{L}_{rot}$ in Eq. 1) and omit $\mathcal{L}'_{dst}$ in Eq. 5 when adapting to novel classes due to the absence of the rotation classifier. We can observe from Table IV that our approach still surpasses existing methods verifying the robustness of the algorithm.

V. Conclusion

In this paper, we reveal that although direct adaptation to novel classes on models pre-trained with self-supervision can obtain improved novel class performance, it causes severe base knowledge forgetting in FSCIL. To solve the above issue, an imprinting-based distillation module and a mathematically provable routing strategy are proposed. Extensive experimental evaluations and detailed ablations validate the effectiveness of our method by significantly outperforming previous approaches across all the datasets.

References

[1] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *CVPR*, 2018.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[3] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *CVPR*, 2020.

[4] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017.

[5] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *ECCV*, 2018.

[6] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *CVPR*, 2019.

[7] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, "Few-shot incremental learning with continually evolved classifiers," in *CVPR*, 2021.

Method	Acc. in each session (%)									Avg.	Final Impro.
	0	1	2	3	4	5	6	7	8		
LUCIR*◇ [6]	61.31	47.80	39.31	31.91	25.68	21.35	18.67	17.24	14.17	30.83	+39.96
iCaRL*◇ [4]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	33.29	+36.92
EEIL*◇ [5]	61.31	46.58	44.00	37.29	33.14	27.12	24.10	21.57	19.58	34.97	+34.55
TOPIC [3]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	39.64	+29.71
ERL++* [15]	61.70	57.58	54.66	51.72	48.66	46.27	44.67	42.81	40.79	49.87	+13.34
IDLVQ* [25]	64.77	59.87	55.93	52.62	49.88	47.55	44.83	43.14	41.84	51.16	+12.29
CEC [7]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	57.75	+6.50
F2M* [19]	72.05	67.47	63.16	59.70	56.71	53.77	51.11	49.21	47.84	57.89	+6.29
CLOM [26]	73.08	68.09	64.16	60.41	57.41	54.29	51.54	49.37	48.00	58.48	+6.13
Replay* [27]	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	58.02	+5.92
MetaFSCIL [28]	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	58.85	+4.94
FACT [8]	72.56	69.63	66.38	62.77	60.60	57.33	54.34	52.16	50.49	60.70	+3.64
Ours (0 exemplar)	74.78	69.80	66.37	62.89	59.98	57.46	54.60	52.84	52.15	61.21	
Ours (1 exemplar)[default]*	**74.78**	**70.66**	**66.73**	**64.13**	**61.84**	**58.84**	**55.86**	**55.20**	**54.13**	**62.46**	
Ours (5 exemplars)*	74.78	70.83	67.14	64.69	62.59	60.02	57.17	56.19	55.14	63.17	

*: method with exemplars. ◇: results from [3].

TABLE I
RESULTS OF 5-WAY 5-SHOT FSCIL ON *mini*-IMAGENET. METRIC "FINAL IMPRO." IS THE IMPROVEMENT OF ACCURACY IN THE FINAL SESSION.

	Components					Acc. in each session (%)									Final session Acc.	
	Rot	VKD	IKD	RKD	Route	0	1	2	3	4	5	6	7	8	Acc_{base}	Acc_{novel}
(1)						74.65	69.85	65.30	61.67	58.65	55.48	52.74	50.79	48.97	72.50	13.68
(2)	✓					74.78	69.95	65.76	62.40	59.31	56.32	53.60	51.49	50.06	72.67	16.15
(3)	✓	✓				74.78	68.32	63.90	60.13	57.69	54.11	50.19	47.73	46.72	53.60	36.40
(4)	✓		✓			74.78	68.37	64.87	62.23	60.35	57.29	54.57	53.48	52.70	66.88	31.43
(5)	✓		✓	✓		74.78	68.74	65.50	63.05	60.70	57.85	54.92	54.32	53.30	66.35	33.73
(6)	✓		✓	✓	✓	**74.78**	**70.66**	**66.73**	**64.13**	**61.84**	**58.84**	**55.86**	**55.20**	**54.13**	67.73	33.73

TABLE II
ABLATIONS OF OUR OVERALL FRAMEWORK ON *mini*-IMAGENET. METRICS "ACC_{base}" AND "ACC_{novel}" ON THE RIGHTMOST PART GIVE THE ACCURACY OF CLASSIFYING BASE AND NOVEL CLASS TEST SAMPLES IN THE FINAL SESSION.

Strategy	Single	Pred-Avg	Feat-Concat	Routing (ours)
Final Acc.	53.30%	53.05%	53.53%	**54.13**%

TABLE III
ANALYSES OF ENSEMBLE STRATEGIES ON *mini*-IMAGENET.

Method	CEC [7]	FACT [8]	Ours w/o SSL	Ours w/ SSL
Final Acc.	47.63%	50.49%	**51.78**%	**54.13**%

TABLE IV
EFFECTS OF SSL ON OUR METHOD ON *mini*-IMAGENET.

[8] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, L. Ma, S. Pu, and D.-C. Zhan, "Forward compatible few-shot class-incremental learning," in *CVPR*, 2022.
[9] P. Mazumder, P. Singh, and P. Rai, "Few-shot lifelong learning," in *AAAI*, 2021.
[10] J. Kalla and S. Biswas, "S3c: Self-supervised stochastic classifiers for few-shot class-incremental learning," in *ECCV*, 2022.
[11] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.
[12] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.
[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.
[14] J. Snell, K. Swersky *et al.*, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017.
[15] S. Dong, X. Hong, X. Tao *et al.*, "Few-shot class-incremental learning via relation knowledge distillation," in *AAAI*, 2021.
[16] A. Cheraghian, S. Rahman *et al.*, "Semantic-aware knowledge distillation for few-shot class-incremental learning," in *CVPR*, 2021.
[17] F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu, "Prototype augmentation and self-supervision for incremental learning," in *ICCV*, 2021.
[18] S. Gidaris, A. Bursuc, N. Komodakis, P. Perez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *ICCV*, 2019.
[19] G. Shi, J. Chen *et al.*, "Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima," *NeurIPS*, 2021.
[20] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *NeurIPS*, 2014.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[22] O. Vinyals, C. Blundell *et al.*, "Matching networks for one shot learning," in *NeurIPS*, 2016.
[23] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," *Technical report*, 2011.
[25] K. Chen and C.-G. Lee, "Incremental few-shot learning via vector quantization in deep embedded space," in *ICLR*, 2021.
[26] Y. Zou, S. Zhang, Y. Li *et al.*, "Margin-based few-shot class-incremental learning with class-level overfitting mitigation," in *NeurIPS*, 2022.
[27] H. Liu, L. Gu, Z. Chi *et al.*, "Few-shot class-incremental learning via entropy-regularized data-free replay," in *ECCV*, 2022.
[28] Z. Chi, L. Gu, H. Liu *et al.*, "Metafscil: A meta-learning approach for few-shot class incremental learning," in *CVPR*, 2022.