Does Prior Data Matter?
Exploring Joint Training in the Context of Few-Shot Class-Incremental Learning

Shiwon Kim[1*] **Dongjun Hwang**[2*†] **Sungwon Woo**[2*] **Rita Singh**[3†]

[1]Yonsei University [2]Sogang University [3]Carnegie Mellon University

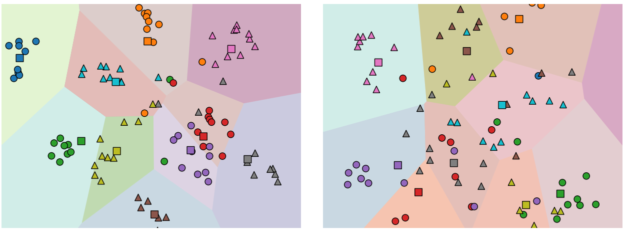
shiwon1998@yonsei.ac.kr, {djhwang, swwoo}@sogang.ac.kr, rsingh@cs.cmu.edu

Abstract

Class-incremental learning (CIL) aims to adapt to continuously emerging new classes while preserving knowledge of previously learned ones. Few-shot class-incremental learning (FSCIL) presents a greater challenge that requires the model to learn new classes from only a limited number of samples per class. While incremental learning typically assumes restricted access to past data, it often remains available in many real-world scenarios. This raises a practical question: should one retrain the model on the full dataset (i.e., joint training), or continue updating it solely with new data? In CIL, joint training is considered an ideal benchmark that provides a reference for evaluating the trade-offs between performance and computational cost. However, in FSCIL, joint training becomes less reliable due to severe imbalance between base and incremental classes. This results in the absence of a practical baseline, making it unclear which strategy is preferable for practitioners. To this end, we revisit joint training in the context of FSCIL by incorporating imbalance mitigation techniques, and suggest a new imbalance-aware joint training benchmark for FSCIL. We then conduct extensive comparisons between this benchmark and FSCIL methods to analyze which approach is most suitable when prior data is accessible. Our analysis offers realistic insights and guidance for selecting training strategies in real-world FSCIL scenarios. Code is available at: https://github.com/shiwonkim/Joint_FSCIL

1. Introduction

Deep neural networks (DNNs) have achieved remarkable progress in various fields, often matching or even surpassing human capabilities [13, 20, 22]. However, when trained on streaming data, they face the challenge of *catastrophic forgetting* (CF) [15, 45], which refers to the loss of previously acquired knowledge when adapting to evolving data distributions. They also struggle with poor *inter-task class separation* (ICS) [26, 31], which leads to ambiguous decision boundaries between previously learned and newly introduced classes. To tackle these issues, class-incremental learning (CIL) has been proposed as a framework that enables models to accommodate new classes over time while maintaining strong performance on all previously observed classes [50, 62]. In this study, we focus on a more practical yet challenging extension of conventional CIL, few-shot class-incremental learning (FSCIL), where new classes emerge with only a few samples [41, 56]. Specifically, the FSCIL task consists of a base session with sufficient training data, followed by multiple incremental sessions where an extremely limited number of samples are provided [42].



(a) Joint training in CIL setting (b) Joint training in FSCIL setting

Figure 1. Feature space visualization of a joint training model on randomly selected 5 base classes (dots) and 5 incremental classes (triangles) from the CIFAR-100 [28] test set. Class centroids are shown as squares. (a) Joint training in CIL obtains well-clustered features. (b) Joint training in FSCIL results in scattered features.

Numerous FSCIL approaches have been proposed to address this challenge under the assumption that *previously seen data are no longer accessible in the following incremental sessions* [25, 48, 60]. However, in many real-world scenarios such as e-commerce applications or industrial deployments, previously collected datasets often remain available [10, 34]—albeit possibly large in size or costly to retrain on. This raises a fundamental question: *If access to previous data is allowed, is it better to retrain a model us-*

[*]Equal contribution.
[†]Corresponding author.

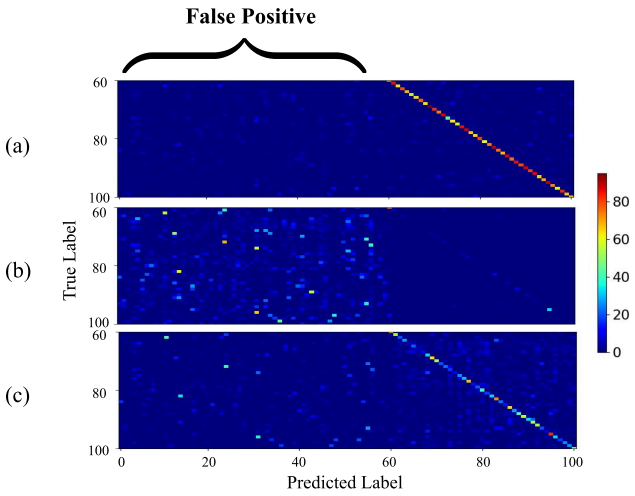


Figure 2. Comparison of confusion matrices on the incremental classes (60-99) of CIFAR-100 [28] test set between standard joint training in (a) CIL setting, (b) FSCIL setting, and (c) imbalance-aware joint training in FSCIL setting. (c) exhibits significantly less false positives for incremental classes than (b).

ing all accumulated data (i.e., joint training), or to update the model solely based on the newly introduced data?

The answer to this question is relatively clear in the context of conventional CIL. Given that each incremental session contains a substantial amount of data, joint training is widely regarded as the ideal upper bound [31, 62]. It serves not only as a comparative baseline for evaluating the performance of CIL methods [29, 44, 45], but also as a methodological benchmark that many CIL algorithms try to emulate [62]. For instance, several studies seek to reduce the inductive bias in CIL models by rectifying their classifier weights [1, 19, 58], output logits [3, 8, 49], or feature embeddings [53, 60] to align with those of the joint training model. The existence of a well-defined upper bound provides a practical guideline: *when access to previous data is permitted, joint training is preferred for maximizing performance, whereas CIL methods are viable alternatives under constraints in training time or computational resources*.

In contrast to conventional CIL, the severe imbalance between base and incremental classes in FSCIL undermines the effectiveness of joint training as a reliable upper bound. Figure 1 depicts the feature space of a ResNet-20 [17] joint training model under both CIL and FSCIL settings using t-SNE [46] and Voronoi diagram [2]. In CIL, incremental class features (triangles) are well separated along decision boundaries (Figure 1a). However, in FSCIL, they are scattered and overlapped, failing to form distinct class regions (Figure 1b). The confusion matrices in Figures 2a and 2b further show that joint training in FSCIL produces notably more false positives for incremental classes than in CIL.

Since joint training proves to be less effective in FSCIL scenarios, it remains unclear whether retraining on the full dataset or incremental learning is preferable. Nevertheless, to the best of our knowledge, no prior work has empirically investigated how to effectively leverage past data in FSCIL settings when it is available.

In this paper, we explore *imbalanced learning* methods as a more realistic joint training benchmark for comparison with FSCIL approaches that do not utilize previous data. Imbalanced learning aims to enhance the representativeness of minority classes, ensuring their contribution to the learning process despite their limited sample size [9, 21, 38]. This objective closely aligns with the fundamental assumption of FSCIL, which involves an imbalanced distribution between base and incremental classes [42, 56].

We categorize eight state-of-the-art imbalanced learning techniques into three taxonomies—resampling-based [12, 16, 33], reweighting-based [7, 11, 35, 51], and optimizer-based [63]—and perform a random search [4, 30] to identify the optimal combination. We present this combination as a new imbalance-aware joint training benchmark for FSCIL. Figure 2c demonstrates the effectiveness of the new benchmark in improving the model's ICS. As shown in the confusion matrix, it significantly reduces false positives for incremental classes. This suggests that imbalance-aware joint training offers a more practical and informative reference than conventional joint training for evaluating different approaches in the FSCIL setting.

Based on this insight, we compare its performance with eight state-of-the-art FSCIL methods to provide guidelines for selecting suitable training strategies in few-shot incremental scenarios under varying resource constraints. To ensure fair and consistent comparison, all methods are reimplemented and integrated into a unified framework instead of relying on disparate codebases. Our framework is made public to support reproducibility and provide a transparent pipeline for future FSCIL research.

Our contributions are three-fold:

- *First*, we initiate a practical discussion on the use of previously observed data in FSCIL. To the best of our knowledge, this is the first empirical study to examine whether retraining or incremental learning is preferable when access to prior data is available in FSCIL settings.
- *Second*, we investigate the effectiveness of joint training with imbalanced learning strategies in FSCIL scenarios. This serves as a more realistic joint training benchmark for FSCIL that reflects the class imbalance.
- *Third*, we conduct a comparative analysis of imbalance-aware joint training and state-of-the-art FSCIL methods under varying resource constraints. Our evaluation offers empirical insights into which training strategy is more effective under such conditions.

2. Related Work

2.1. Few-Shot Class-Incremental Learning

Few-shot class-incremental learning (FSCIL) addresses the challenge of continually adapting to new classes using only a few samples per class while preserving knowledge of previously learned classes [41]. Recent efforts in FSCIL can be broadly categorized into i) *incremental-frozen*, and ii) *fine-tuning* approaches [36]. **Incremental-frozen approaches** keep the feature extractor fixed during incremental learning, thereby maintaining a stable embedding space for base classes even as novel classes are introduced. While this approach consolidates *stability*—the ability to maintain previous knowledge—it can limit *plasticity*—the ability to learn new patterns—thus motivating the use of various techniques to mitigate this trade-off [39, 40, 48, 52, 55, 60, 61]. **Fine-tuning approaches**, on the other hand, update the parameters of the feature extractor partially or entirely in each incremental session, which enhances plasticity at the potential cost of reduced stability [23–25, 59].

In many real-world applications, previous training data often remain accessible even as new data are continuously introduced. However, such scenario has not been considered in existing FSCIL research. This leads to a lack of discussion on proper benchmarks for determining which methods are suitable when prior data is available. In this paper, we explore a new benchmark based on imbalanced learning techniques and compare it against conventional FSCIL methods, providing concrete guidelines for scenarios where past data can be leveraged.

2.2. Imbalanced Learning

Imbalanced learning primarily addresses long-tailed distributions [9], where majority classes significantly outnumber minority classes [57]. Extensive studies have explored strategies to mitigate the resulting model bias, which are commonly categorized into three major approaches: i) *resampling* the training dataset, ii) *reweighting* the objective function, and iii) refining the *optimizer*. **Resampling-based approaches** include techniques such as CMO [33], which employs a CutMix-based augmentation [54] to blend samples from majority and minority classes; DeepSMOTE [12] which applies GAN-based generation of minority samples; and Ghosh et al. [16], which ensure balanced sampling in each training batch. **Reweighting-based approaches** aim to rebalance gradient signals between majority and minority classes [7, 11, 35, 51]. **Optimizer-based approaches** mitigate class imbalance by modifying the optimizer. For example, ImbSAM [63] extends Sharpness-Aware Minimization (SAM) [14] by incorporating class-aware weight updates, enhancing generalization under skewed distributions.

Although these methods are mostly developed for long-tailed datasets, they are not limited to such distributions. Various studies have explored imbalanced learning under different distributional variations. For example, LDAM [7] defines an imbalance ratio and controls the number of samples between major and minor classes accordingly. Buda et al. [6] introduce step and linear imbalance settings, where the number of samples per class decreases linearly. They also consider extreme cases where all classes except one have very few samples.

These variations suggest that imbalanced learning hold potential relevance for the FSCIL problem. In this work, we aim to establish a more realistic benchmark for FSCIL that complements the conventional joint training benchmark, by systematically applying and comparing imbalanced learning methods in the FSCIL setting.

3. Rethinking Joint Training in FSCIL

To provide a practical and informative guideline for the FSCIL community—particularly in scenarios where access to past data is available—we take a step further by rethinking what constitutes a meaningful benchmark for FSCIL. From our findings in Figures 1 and 2, we observe that joint training alone cannot serve as a proper benchmark for FSCIL, as it fails to address the inter-task class separation (ICS) problem under severe class imbalance.

To this end, we further explore class imbalance mitigation strategies to establish a more appropriate joint training benchmark for FSCIL, and investigate whether such an approach can serve as a viable standard. We refer to conventional joint training—the baseline method without any modifications—as *standard joint training* throughout this paper, to clearly distinguish it from joint training schemes with imbalance mitigation techniques discussed in Section 3.1.

3.1. Imbalance-Aware Joint Training in FSCIL

A fundamental challenge in FSCIL lies in the severe class imbalance between base and incremental classes. Since incremental classes are introduced with only a few samples, the model tends to be biased toward well-represented base classes, resulting in performance degradation. This closely resembles the problem of *imbalanced learning*, which aims to adjust models to learn meaningful representations for underrepresented classes [9, 21, 38].

Motivated by such conceptual similarity, we systematically explore imbalanced learning techniques to develop a reliable joint training benchmark for FSCIL. We first examine how prior studies have addressed class imbalance, and find that many existing works recommend combining independently functioning strategies from different categories of imbalanced learning, such as i) resampling, ii) reweighting, and iii) optimizer-based methods. For instance, Park et al. [33] highlight that combining resampling and reweighting techniques leads to significant improvements in the performance of minority classes. In addition, Zhou et al. [63]

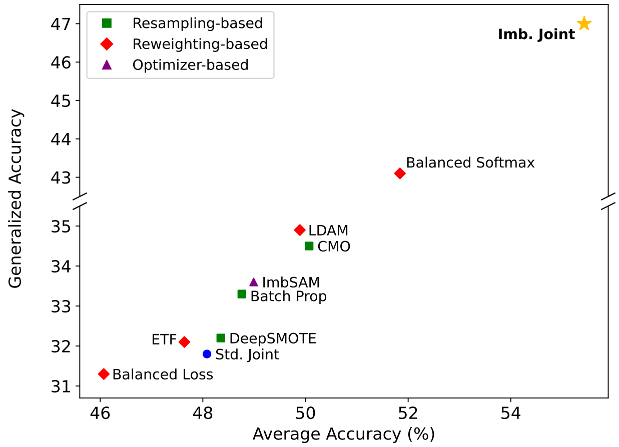


Figure 3. Performance comparison of imbalance-aware joint training (*Imb. Joint*), standard joint training (*Std. Joint*), and 8 imbalanced learning techniques based on *aAcc* and *gAcc* using the CIFAR-100 [28] test set in the last session of a 5-way 5-shot FSCIL setting. Imbalanced learning techniques are presented in three categories: resampling-based, reweighting-based, and optimizer-based. *Imb. Joint* consists of CMO [33], Balanced Softmax [35], and ImbSAM [63], outperforming others by a large margin.

Table 1. Ablation study of each component in the imbalance-aware joint training benchmark on CIFAR-100 [28] test set. All values are measured in the last session of a 5-way 5-shot FSCIL setting. Each component contributes to the performance improvement.

CMO	*BalancedSoftmax*	*ImbSAM*	*aAcc*	*gAcc*
			48.1	31.8
✓			50.1	34.5
✓	✓		55.5	45.9
✓	✓	✓	**55.8**	**46.8**

point out that using only resampling or reweighting without an explicit optimization strategy may cause overfitting or unstable training due to imbalanced gradient updates. These insights suggest that integrating different imbalanced learning methods can yield more stable and robust performance than relying on a single approach.

Building on these findings, we combine methods from three categories of imbalanced learning and search for the most effective configuration in the FSCIL setting. Specifically, we classify eight state-of-the-art imbalanced learning methods into resampling [12, 16, 33], reweighting [7, 11, 35, 51], and optimizer-based [63] approaches, and conduct 30 random search trials per method [4, 30]. We then select the top-performing method from each category and combine them to form the new imbalance-aware joint training benchmark for FSCIL. Note that, for evaluation, we adopt both average accuracy (*aAcc*), a standard metric in FSCIL, and generalized average accuracy (*gAcc*) proposed by Tang et al. [40], which offers a more balanced assessment with explicit emphasis on incremental class performance.

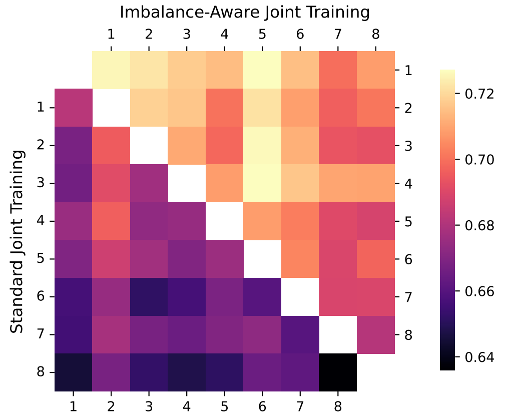


Figure 4. Feature similarity between joint training models under CIL and FSCIL settings on CIFAR-100 [28] test set based on Centered Kernel Alignment (CKA) [27]. The x- and y-axes represent the incremental sessions. The upper triangular matrix shows the similarity between standard joint training in CIL and imbalance-aware joint training in FSCIL, while the lower triangular matrix presents the similarity between standard joint training in CIL and standard joint training in FSCIL. The brighter coloration in the upper triangle indicates that imbalance-aware joint training in FSCIL yields features more similar to those of standard joint training in CIL than standard joint training in FSCIL does.

Based on our experiments, we find that combining *CMO* (resampling-based) [33], *Balanced Softmax* (reweighting-based) [35], and *ImbSAM* (optimizer-based) [63] achieves the best overall performance. It improves *aAcc* by 7%p and *gAcc* by 15%p over standard joint training (Table 1), and also outperforms all individual imbalanced learning methods (Figure 3). These results indicate that imbalance-aware joint training can serve as a more meaningful reference than standard joint training when developing practical guidelines for real-world FSCIL scenarios. Therefore, we suggest this approach as a new joint training benchmark for FSCIL. A detailed analysis is provided in Section 3.2.

3.2. Analysis of Imbalance-Aware Joint Training

To evaluate the individual effectiveness of different types of imbalanced learning techniques, we conduct ablation experiments on the imbalance-aware joint training benchmark using the CIFAR-100 test set. As demonstrated in Table 1, the standard joint training model achieves 48.1% on *aAcc* and 31.8% on *gAcc*. We then independently apply representative methods from each category—CMO, Balanced Softmax, and ImbSAM. Each method yields performance improvements, with the best results reaching 55.8% on *aAcc* and 46.8% on *gAcc*. These findings indicate that each tech-

Table 2. Comparison of base session training setups of 8 FSCIL methods [23, 25, 39, 40, 48, 55, 60, 61].

	CEC	S3C	WaRP	FACT	TEEN	SAVC	LIMIT	Yourself
P1: Exposure of test set during training	✓	✓	✗	✗	✓	✓	✓	✗
P2: Unfair usage of pre-trained encoders	✗	✗	✗	✗	✗	✗	✗	✓

Table 3. Comparison of methods across multiple sessions on CIFAR-100 [28]. **S0** presents the base session and **S1-S8** denote incremental sessions. The best and second-best results are **bolded** and underlined. All methods are reproduced within our unified codebase. *Std. Joint* and *Imb. Joint* denote standard joint training and imbalance-aware joint training, respectively.

						***aAcc* in each session (%)**					***aAcc* S8**			
Method	**Architecture**	**S0**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**Base**	**Inc.**	*aAcc*	*gAcc*
Std. Joint	ResNet-20	77.5	73.1	68.1	63.2	59.7	56.3	53.1	51.2	48.1	**78.8**	1.9	61.1	46.9
Imb. Joint		78.4	75.0	71.6	66.9	64.0	62.5	60.0	59.1	**55.3**	70.5	32.5	**65.9**	**58.0**
CEC [55]	ResNet-20	76.4	57.0	53.1	50.0	47.6	45.3	43.7	41.7	39.7	50.5	23.5	50.5	45.9
FACT [60]		68.2	63.0	58.4	54.7	51.6	48.9	46.6	44.3	42.4	62.5	12.1	53.1	43.6
TEEN [48]		67.0	62.3	58.1	54.5	51.2	48.7	46.1	43.8	41.7	63.7	8.8	52.6	42.4
S3C [23]		56.6	54.7	52.4	49.2	47.5	46.1	44.7	43.5	41.3	47.4	32.1	48.4	45.5
WaRP [25]		70.0	66.2	62.5	58.5	55.4	52.7	51.2	49.1	47.1	64.2	21.6	57.0	49.2
SAVC [39]		80.5	76.0	71.5	67.4	64.1	61.3	59.2	57.0	54.7	76.5	22.2	65.8	55.6
LIMIT [61]		74.1	69.9	66.2	62.2	59.1	56.3	54.2	52.1	49.7	68.9	21.1	60.4	51.2
YourSelf [40]	DeiT-S	71.6	67.2	64.1	60.4	57.4	54.6	52.8	51.2	48.5	56.0	**37.3**	58.7	54.6

nique independently enhances performance, confirming that imbalanced learning methods from different categories provide complementary benefits.

Furthermore, to evaluate how effectively the imbalance-aware joint training approach addresses the ICS problem in FSCIL, we use Centered Kernel Alignment (CKA) [27] to measure the similarity of network representations between joint training models in CIL and FSCIL settings (Figure 4). The upper triangular matrix shows the similarities between imbalance-aware joint training in FSCIL and standard joint training in CIL, while the lower one represents the similarities between standard joint training in FSCIL and CIL. As indicated by the brighter coloration, the upper triangle exhibits consistently higher similarity across all sessions. This observation suggests that imbalance-aware joint training in FSCIL produces representations more closely aligned with those of standard joint training in CIL than standard joint training in FSCIL does.

4. Towards a Practical Guideline for FSCIL

4.1. Experimental Setup

4.1.1. General Settings

Dataset. Following prior works, we conduct experiments on CIFAR-100 [28], *mini*ImageNet [37], and CUB-200 [47] datasets using the data splits in Tao et al. [41]. For CIFAR-100 and *mini*ImageNet, 60 classes are allocated to the base session, and each of the 8 incremental sessions contains 5 classes. For CUB-200, 100 classes are used for the base session, followed by 10 incremental sessions with 10 classes each. All datasets are evaluated under the 5-shot setting.

Evaluation Metrics. Following common practice in FSCIL research [41, 55], we use average accuracy (*aAcc*) as our primary evaluation metric. Additionally, unlike prior works that report only the overall average accuracy, we separately report the average accuracy of base classes and incremental classes. We also adopt generalized average accuracy (*gAcc*) proposed by Tang et al. [40], which balances the evaluation of base and incremental classes using a tunable parameter α that controls their respective weights.

Implementation Details. As the backbone feature extractor, all methods except for YourSelf [40] utilize ResNet-20 [18] for CIFAR-100, and ResNet-18 for *mini*ImageNet and CUB-200. YourSelf employs DeiT-S [43], a ViT-based architecture, across all three datasets. All our experiments are conducted on a single NVIDIA A5000 GPU. To ensure a consistent environment, we incorporate all methods into a unified codebase. All reimplementations are based on publicly available GitHub repositories.

4.1.2. A Standardized Evaluation Protocol for FSCIL

In this paper, we evaluate existing FSCIL methods and joint training approaches after resolving inconsistencies in their experimental setups. Although most methods follow a similar training pipeline, subtle but unfair differences undermine the reliability of performance comparisons. To address this, we identify two major inconsistencies as shown in Table 2, and standardize them to enable a unified and fair comparison of eight FSCIL methods and joint training.

Exposure of test set during training (P1). A major issue in prior FSCIL research is the use of the test set as a validation set. Many methods select the best-performing epoch

Table 4. Comparison of methods across multiple sessions on *mini*ImageNet [37]. **S0** represents the base session and **S1-S8** correspond to incremental sessions. The best and second-best results are **bolded** and underlined. All methods are reproduced within our unified codebase. *Std. Joint* and *Imb. Joint* denote standard joint training and imbalance-aware joint training, respectively.

Method	Architecture	*aAcc* in each session (%)									*aAcc* S8		*aAcc*	*gAcc*
		S0	S1	S2	S3	S4	S5	S6	S7	S8	Base	Inc.		
Std. Joint	ResNet-18	71.0	66.7	62.1	58.5	54.8	51.5	48.7	46.3	44.2	72.3	2.1	56.0	43.0
Imb. Joint		71.0	69.4	65.3	62.9	59.0	57.1	55.1	53.6	51.7	66.7	29.1	60.6	53.5
CEC [55]	ResNet-18	70.9	65.0	61.1	58.1	55.5	52.7	50.1	48.2	46.7	65.2	18.9	56.5	47.9
FACT [60]		69.5	64.7	60.4	57.0	53.8	50.8	48.0	46.0	44.1	66.8	9.9	54.9	44.4
TEEN [48]		64.9	60.7	56.9	54.4	52.0	49.4	47.0	45.2	43.8	58.0	22.4	52.7	45.8
S3C [23]		57.7	53.9	51.1	49.0	47.6	45.1	42.8	41.4	40.7	51.4	24.9	47.7	42.3
WaRP [25]		71.5	66.7	63.0	60.2	57.7	55.1	52.5	50.9	49.7	65.9	25.4	58.6	50.8
SAVC [39]		80.0	75.4	71.2	67.5	64.5	61.1	58.1	56.0	**54.1**	**76.2**	21.1	**65.3**	**55.6**
LIMIT [61]		72.9	66.4	62.3	59.0	56.0	53.3	50.5	48.6	47.1	64.2	21.6	57.3	49.2
YourSelf [40]	DeiT-S	71.8	66.0	62.3	59.4	57.4	54.5	52.0	50.5	49.4	60.9	**32.2**	58.2	52.6

Table 5. Comparison of methods across multiple sessions on CUB-200 [47]. **S0** presents the base session and **S1-S10** denote incremental sessions. The best and second-best results are **bolded** and underlined. All methods are reproduced within our unified codebase. *Std. Joint* and *Imb. Joint* denote standard joint training and imbalance-aware joint training, respectively.

Method	Architecture	*aAcc* in each session (%)											*aAcc* S10		*aAcc*	*gAcc*
		S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Base	Inc.		
Std. Joint	ResNet-18	76.8	73.0	69.4	66.5	64.8	63.8	60.5	59.7	60.6	59.5	58.3	**77.4**	40.1	62.6	57.0
Imb. Joint		77.5	74.9	71.9	68.1	67.9	65.3	63.6	62.6	62.5	62.3	61.9	73.2	51.4	65.1	62.8
CEC [55]	ResNet-18	72.6	68.2	63.3	58.0	57.4	53.1	51.0	49.3	46.5	46.2	44.6	66.2	24.1	52.2	48.1
FACT [60]		77.8	73.7	70.0	65.6	64.3	61.2	59.8	58.9	57.6	56.5	55.2	73.0	38.1	61.0	57.1
TEEN [48]		78.6	73.7	69.8	64.8	64.0	60.6	59.7	58.8	57.6	56.0	54.9	70.5	40.0	60.7	57.8
S3C [23]		62.1	60.6	57.8	54.3	55.3	52.5	51.9	51.0	50.7	50.4	49.9	54.5	44.1	52.7	52.1
WaRP [25]		77.0	73.4	70.0	66.0	64.8	61.8	60.7	57.9	58.2	57.3	56.2	72.3	40.9	61.4	57.9
SAVC [39]		78.0	75.0	71.9	67.6	67.1	64.4	63.8	61.9	61.0	60.5	59.9	75.1	45.4	64.3	60.4
LIMIT [61]		67.3	63.1	58.7	54.4	53.3	49.4	46.9	44.8	43.4	42.3	40.3	58.1	23.3	48.2	44.2
YourSelf [40]	DeiT-S	80.8	77.8	74.7	72.0	68.9	65.5	64.6	64.1	62.5	63.1	**62.5**	73.4	**52.3**	**66.4**	**64.3**

in the base session using the test set. Some methods even use the test set from the last incremental session—which covers the entire label space of the dataset—for hyperparameter tuning [55, 64]. This leakage can lead to overfitting to the test data and result in unreliable evaluations of generalization performance [5]. WaRP [25] acknowledges this issue and avoids the usage of the test set by selecting the checkpoint from the final epoch instead.

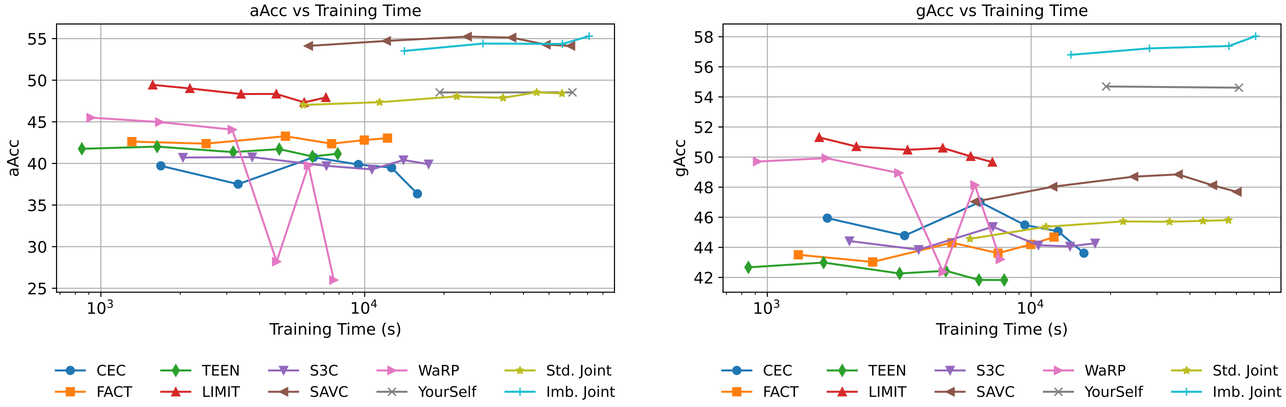
Such inconsistent usage of the test set prevents fair comparisons across methods. To resolve this, we create a new validation set by splitting the original training set in a 9:1 ratio. In addition, for methods that retrain the entire model during incremental sessions [40], we standardize the evaluation by using model weights from the final epoch, since reserving a separate validation set is impractical due to the limited size of incremental data.

Unfair usage of pre-trained encoders (P2). Another issue is that some FSCIL methods leverage additional information from pre-trained encoders. For example, YourSelf [40] employs knowledge distillation from a CNN-based state-of-the-art teacher model [52] to accelerate the convergence of a ViT-based encoder. This teacher model requires prior knowledge of the total number of classes, which may compromise the fairness of the comparison. To ensure consistency, we modify YourSelf to perform knowledge distillation only from a model trained under our standardized evaluation protocol, without such additional information. Likewise, we exclude Park et al. [32] from our comparison, as it uses large-scale pre-trained encoders like CLIP that already demonstrate strong zero-shot classification performance.

4.2. Comparison of FSCIL and Joint Training

In this section, we conduct a comprehensive comparison of existing FSCIL methods and joint training approaches (*i.e.*, standard joint training and imbalance-aware joint training) across three datasets. We then provide an in-depth analysis and discussion on which approach is most suitable, depending on the availability of previously used training data.

First, we present the experimental results on CIFAR-100 in Table 3. The results show that the standard joint training approach achieves strong performance only on the base classes, while its incremental accuracy is extremely low at 1.9%. In contrast, when incorporating imbalanced learning techniques to joint training, we observe a significant improvement in the incremental accuracy (32.5%), which in turn leads to improvements in *aAcc* and *gAcc*.



(a) Average Accuracy (*aAcc*)

(b) Generalized Average Accuracy (*gAcc*)

Figure 5. Performance comparison across different training times on CIFAR-100 [28] test set. Training times are presented on a log scale. *Std. Joint* and *Imb. Joint* denote standard joint training and imbalance-aware joint training, respectively.

Notably, this imbalance-aware joint training approach also outperforms existing FSCIL methods in both *aAcc* and *gAcc*. While YourSelf yields the highest incremental class accuracy among all methods, it shows relatively low performance on the base classes, resulting in lower overall performance compared to imbalance-aware joint training. These findings suggest that, when the previous training dataset is accessible, imbalance-aware joint training may be a more effective strategy than FSCIL methods.

However, unlike in CIFAR-100, FSCIL methods outperform imbalance-aware joint training on the other datasets. On *mini*ImageNet, for example, the FSCIL method SAVC achieves the best overall results with an *aAcc* of 65.3% and a *gAcc* of 55.6%, surpassing imbalance-aware joint training by 4.7%p and 2.1%p, respectively (Table 4). Similarly, on CUB-200, YourSelf demonstrates a *gAcc* of 64.3%, outperforming the joint training approach (62.8%) as shown in Table 5. The fact that FSCIL methods—without access to previous training data—can perform better than the imbalance-aware joint training approach that leverages such data challenges the conventional belief that more information necessarily leads to better performance.

Discussion. The results in this section suggest that, despite having access to additional data, current imbalanced learning techniques may perform worse than FSCIL approaches under extreme data distributions—particularly when only a few classes have very limited samples. This highlights the need for future research on imbalanced learning approaches that can better handle such challenging scenarios. Since FSCIL methods have shown strong performance in these cases, their strategies could be effectively adapted for imbalanced learning. In particular, since previous training datasets are typically accessible in imbalanced learning, techniques in FSCIL that simulate past data can be replaced with actual use of previous dataset for direct application to imbalanced learning problems. For example, while YourSelf stores only the distribution of past data, this strategy can be extended to directly utilize the full previous training dataset.

4.3. Resource-Aware Comparison

In this section, we analyze each method based on its training time to provide practical guidelines for users on which approach to adopt depending on available training resources. Training efficiency is assessed using two metrics (*aAcc* and *gAcc*). To account for varying training times, we evaluate the performance of each model at [100, 200, 400, 600, 800, 1000] epochs. However, due to their longer training times, the imbalance-aware joint training and YourSelf methods are evaluated only at [100, 200, 400, 500] epochs and [100, 200] epochs, respectively.

Figure 5a shows the *aAcc* values for each method over training time. We observe that FSCIL methods with longer training durations, such as SAVC and YourSelf, generally achieve higher *aAcc* than those with shorter training times. Figure 5b presents the *gAcc* values for each method, which reveal a different trend. Unlike the results in *aAcc*, methods with longer training times, including SAVC and standard joint training, achieve lower *gAcc*. This suggests that they maintain strong base class performance but struggle with incremental class learning. In contrast, the imbalance-aware joint training approach consistently records the highest *gAcc* across all epochs, outperforming FSCIL methods regardless of training duration.

Interestingly, some methods with shorter training times, such as LIMIT and TEEN, show a declining trend in *gAcc* as training progresses. This pattern suggests that prolonged training on the base session can lead to overfitting to base classes. Conversely, FACT exhibits the opposite tendency, with *gAcc* improving over time. This indicates that FACT retains greater flexibility during training and benefits from longer training durations.

These results suggest that when users possess sufficient computational resources and access to the previous training dataset, the imbalance-aware joint training method can be a viable choice. However, in cases where resources are sufficient but access to prior data is restricted, SAVC or YourSelf are strong alternatives, despite their longer training times. When both computational resources and access to previous data are limited, LIMIT offers a better balance between efficiency and overall performance.

5. Conclusion

Few-shot class-incremental learning (FSCIL) is particularly challenging, as models must continually accommodate new classes with only a few samples per class. In this paper, we highlight a practical but relatively underexplored problem in the FSCIL literature: *the lack of an established benchmark for evaluating whether leveraging previously learned data, when available, is beneficial in the FSCIL setting*. We point out that standard joint training, which serves as the upper bound in conventional CIL, is unsuitable as a benchmark in FSCIL due to its instability under imbalanced data distributions. To address this issue, we explore imbalanced learning techniques that enhance the performance of joint training in FSCIL and suggest a new joint training benchmark. We then conduct extensive experiments to compare this imbalance-aware joint training benchmark with state-of-the-art FSCIL methods. Based on these comparisons, we offer practical guidelines for determining whether utilizing past data is beneficial in FSCIL scenarios.

Limitations. We acknowledge that we are unable to reproduce a broader range of recent FSCIL methods and therefore cannot include them in our comparisons. Additionally, while our experiments allow for a performance comparison between FSCIL and joint training, we cannot provide a detailed analysis of why certain methods outperform others due to the limited number of datasets used in the evaluation. Future work focuses on covering a more diverse set of FSCIL approaches and conducting experiments on a broader range of real-world datasets, thereby providing more comprehensive and practical insights.

Acknowledgement

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program).

References

[1] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 844–853, 2021. 2

[2] Franz Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM computing surveys (CSUR)*, 23(3):345–405, 1991. 2

[3] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 583–592, 2019. 2

[4] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011. 2, 4

[5] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pages 1006–1014. PMLR, 2015. 6

[6] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018. 3

[7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2, 3, 4

[8] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018. 2

[9] Wuxing Chen, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and CL Philip Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6):137, 2024. 2, 3

[10] Dongkyu Cho, Taesup Moon, Rumi Chunara, Kyunghyun Cho, and Sungmin Cha. Cost-efficient continual learning with sufficient exemplar memory. *arXiv preprint arXiv:2502.07274*, 2025. 1

[11] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2, 3, 4

[12] Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE transactions on neural networks and learning systems*, 34(9):6390–6404, 2022. 2, 3, 4

[13] Quan Feng and Songcan Chen. Learning multi-tasks with inconsistent labels by using auxiliary big task. *Frontiers of Computer Science*, 17(5):175342, 2023. 1

[14] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 3

[15] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 1

[16] Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, 113(7):4845–4901, 2024. 2, 3, 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[19] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 2

[20] Bong Kyung Jang, Shiwon Kim, Jae Yong Yu, JaeSeong Hong, Hee Woo Cho, Hong Seon Lee, Jiwoo Park, Jeesoo Woo, Young Han Lee, and Yu Rang Park. Classification models for arthropathy grades of multiple joints based on hierarchical continual learning. *La radiologia medica*, pages 1–13, 2025. 1

[21] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of big data*, 6 (1):1–54, 2019. 2, 3

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021. 1

[23] Jayateja Kalla and Soma Biswas. S3c: Self-supervised stochastic classifiers for few-shot class-incremental learning. In *European Conference on Computer Vision*, pages 432–448. Springer, 2022. 3, 5, 6

[24] Haeyong Kang, Jaehong Yoon, Sultan Rizky Hikmawan Madjid, Sung Ju Hwang, and Chang D Yoo. On the soft-subnetwork for few-shot class incremental learning. *arXiv preprint arXiv:2209.07529*, 2022.

[25] Do-Yeon Kim, Dong-Jun Han, Jun Seo, and Jaekyun Moon. Warping the space: Weight space rotation for class-incremental few-shot learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3, 5, 6

[26] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. *Advances in neural information processing systems*, 35: 5065–5079, 2022. 1

[27] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 4, 5

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 2, 4, 5, 7

[29] Wojciech Łapacz, Daniel Marczak, Filip Szatkowski, and Tomasz Trzciński. Exploring the stability gap in continual learning: The role of the classification head. *arXiv preprint arXiv:2411.04723*, 2024. 2

[30] Rafael G Mantovani, André LD Rossi, Joaquin Vanschoren, Bernd Bischl, and André CPLF De Carvalho. Effectiveness of random search in svm hyper-parameter tuning. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–8. Ieee, 2015. 2, 4

[31] Saleh Momeni and Bing Liu. Achieving upper bound accuracy of joint training in continual learning. *arXiv preprint arXiv:2502.12388*, 2025. 1, 2

[32] Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23881–23890, 2024. 6

[33] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6887–6896, 2022. 2, 3, 4

[34] Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet K Dokania, Philip HS Torr, Ser-Nam Lim, Bernard Ghanem, and Adel Bibi. Computationally budgeted continual learning: What does matter? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3698–3707, 2023. 1

[35] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 2, 3, 4

[36] Shuvendu Roy, Chunjong Park, Aldi Fahrezi, and Ali Etemad. A bag of tricks for few-shot class-incremental learning. *arXiv preprint arXiv:2403.14392*, 2024. 3

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5, 6

[38] Ravid Shwartz-Ziv, Micah Goldblum, Yucen Li, C Bayan Bruss, and Andrew G Wilson. Simplifying neural network training under class imbalance. *Advances in Neural Information Processing Systems*, 36:35218–35245, 2023. 2, 3

[39] Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24183–24192, 2023. 3, 5, 6

[40] Yu-Ming Tang, Yi-Xing Peng, Jingke Meng, and Wei-Shi Zheng. Rethinking few-shot class-incremental learning: Learning from yourself. In *European Conference on Computer Vision*, pages 108–128. Springer, 2025. 3, 4, 5, 6

[41] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12183–12192, 2020. 1, 3, 5

[42] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning,

and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324, 2024. 1, 2

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 5

[44] Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):4069, 2020. 2

[45] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022. 1, 2

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2

[47] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5, 6

[48] Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 5, 6

[49] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 2

[50] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021. 1

[51] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35:37991–38002, 2022. 2, 3, 4

[52] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023. 3, 6

[53] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6982–6991, 2020. 2

[54] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[55] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2021. 3, 5, 6

[56] Jinghua Zhang, Li Liu, Olli Silvén, Matti Pietikäinen, and Dewen Hu. Few-shot class-incremental learning for classification and object detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 2

[57] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023. 3

[58] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13208–13217, 2020. 2

[59] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow versus fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1576–1588, 2021. 3

[60] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9046–9056, 2022. 1, 2, 3, 5, 6

[61] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12816–12831, 2022. 3, 5, 6

[62] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[63] Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11345–11355, 2023. 2, 3, 4

[64] Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. *Advances in neural information processing systems*, 35:27267–27279, 2022. 6

Appendix

A. Additional Analysis

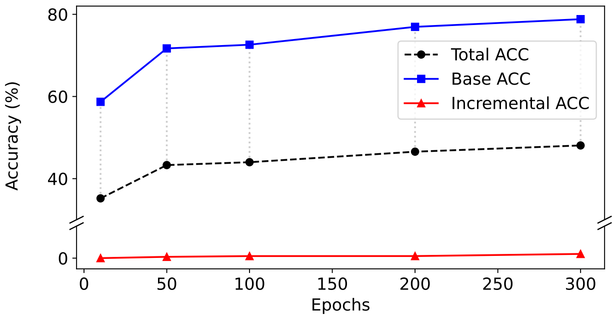


Figure A1. Average accuracies of base and incremental classes in standard joint training on the CIFAR-100 test set by the number of base session training epochs. All results are reported from the last incremental session.

Performance Bias in Standard Joint Training. Figure A1 illustrates the average accuracies of base and incremental classes in the FSCIL setting for standard joint training. We observe that standard joint training achieves near-zero accuracy on incremental classes, and its overall performance is heavily biased toward base classes. Such gap between base and incremental accuracies grows linearly as the number of training epochs in the base session increases.

Table A1. Average false positive (FP) rate and false negative (FN) rate of standard joint training on CIFAR-100 test set under both the CIL and FSCIL settings.

Method	FP rate	FN rate
Standard Joint Training (CIL)	0.254	0.255
Standard Joint Training (FSCIL)	0.488	0.521

Inter-Task Class Separation (ICS) in Joint Training. To highlight the ICS problem in standard joint training under the FSCIL setting, we quantitatively evaluate the quality of decision boundary formation using false positive (FP) and false negative (FN) rates. As shown in Table A1, the joint training model exhibits higher FP and FN rates in the FSCIL setting than in the CIL setting on average. This shows that the ICS problem is more severe in standard joint training under the FSCIL setting compared to the CIL setting.

Exploring Imbalanced Learning in FSCIL. Figure A2 compares the base and incremental class accuracies across the top-5 trials for each method. The imbalance-aware joint training approach achieves markedly higher incremental accuracy, whereas most other approaches—including the standard joint training—show near-zero accuracy on incremental classes, indicating severe overfitting to base classes.

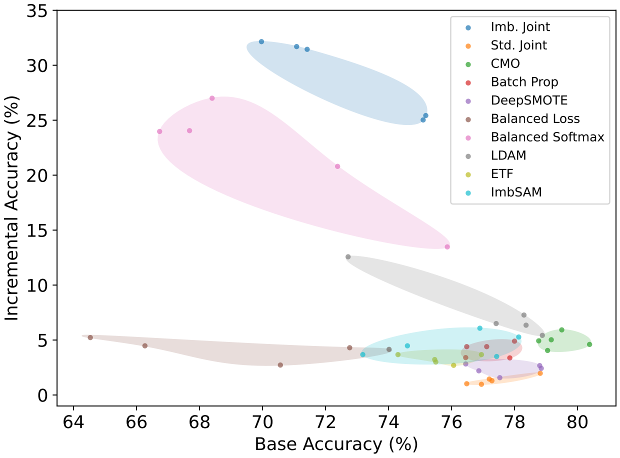


Figure A2. Base and incremental class performance comparison of imbalance-aware joint training (*Imb. Joint*), standard joint training (*Std. Joint*), and 8 imbalance learning techniques on CIFAR-100 test set in the last session of a 5-way 5-shot FSCIL setting. The figure shows the base and incremental class accuracies of the top-5 trials for each method after 30 iterations of hyperparameter random search. The best trials are determined according to the *aAcc*. *Imb. Joint* consists of CMO, Balanced Softmax, and ImbSAM.

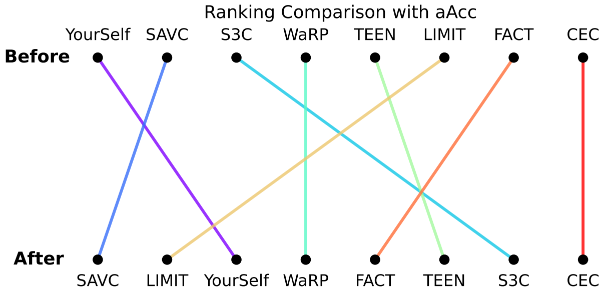
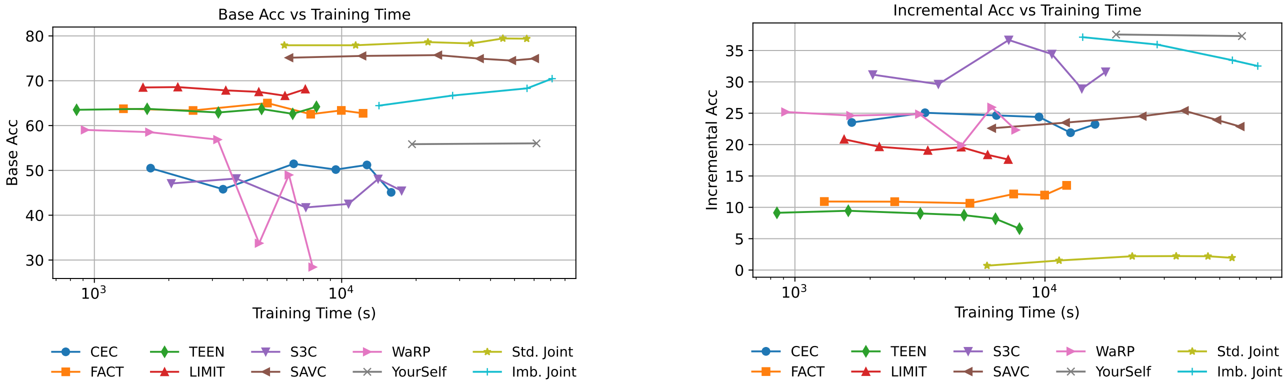


Figure A3. Comparison of FSCIL method performance rankings on CIFAR-100 test set before and after applying our standardized evaluation protocol. Rankings are presented in descending order from left to right.

Performance Rankings Before and After Protocol Standardization. Figure A3 illustrates the ranking shifts of existing FSCIL methods in terms of *aAcc* before and after applying our standardized evaluation protocol. The "**Before**" rankings are based on the values reported in prior studies, whereas the "**After**" rankings reflect the results reproduced under our proposed protocol. All methods except for CEC and WaRP show changes in ranking. Particularly, S3C and LIMIT exhibit the largest shifts, each moving by four positions. These ranking shifts highlight inconsistencies in previous FSCIL experiments and evaluation, underscoring the need for a standardized protocol in future studies.



(a) Base Classes

(b) Incremental Classes

Figure A4. Performance comparison of base and incremental classes across different training times on CIFAR-100 test set. Training times are presented on a log scale. *Std. Joint* and *Imb. Joint* denote standard joint training and imbalance-aware joint training, respectively.

Base and Incremental Performance by Training Time. Figure A4 shows the *aAcc* of base and incremental classes with respect to training time across different methods. For base classes, standard joint training and SAVC achieve the highest performance, benefiting from their longer training durations. In contrast, YourSelf and imbalance-aware joint training mark the best results for incremental classes. Overall, imbalance-aware joint training maintains strong performance on both the base and incremental classes. Notably, TEEN and LIMIT exhibit a trend in which extended training improves base class performance but degrades incremental class performance, likely due to overfitting to base classes as the number of training epochs increases.