Overcoming Bias Towards Base Sessions in Few-Shot Class-Incremental Learning (FSCIL)

Myeongjin Lee[1], Ji-Ae Yoon[2], and Ue-Hwan Kim[2]

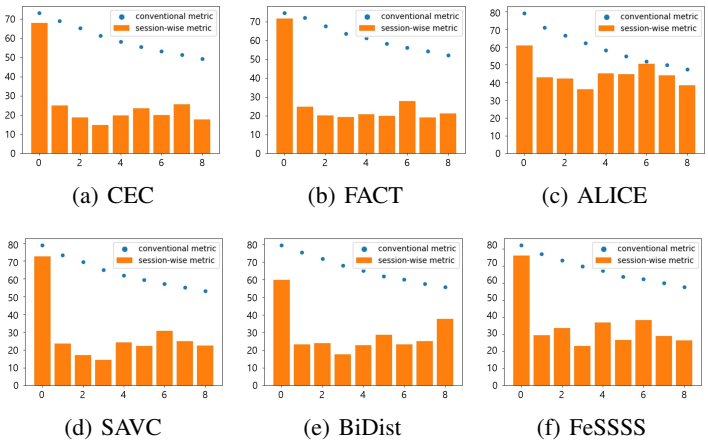
[1] KAIST, Daejeon 34141, Republic of Korea
[2] GIST, Gwangju 61005, Republic of Korea

Abstract. Few-shot class-incremental learning (FSCIL) with a more realistic and challenging problem setting aims to learn a set of novel object classes with a restricted number of training examples in sequence. In the process, striking a balance between not forgetting previously-learned object classes and overfitting to novel ones plays a crucial role. Meanwhile, conventional methods exhibit a significant performance bias towards a base session: excessively low incremental performance compared to base performance. The conventional evaluation metric overshadows the bias. To tackle this, we propose a simple-yet-effective pipeline that achieves a substantial performance margin for incremental sessions. Our approach addresses the inherent bias by employing a dual classification mechanism that balances performance across both base and incremental sessions. Further, we devise and perform comprehensive experiments under diverse conditions—leveraging pretrained representations, various classification modules, and aggregation of the predictions within our pipeline; our findings reveal essential insights towards model design and future research directions. Additionally, we introduce a set of new evaluation metrics and benchmark datasets to address the limitations of the conventional metrics and benchmark datasets which disguise the bias towards a base session. These newly introduced metrics and datasets allow the estimation of the generalization of FSCIL models. As a result of our study, we achieve new state-of-the-art performance with significant margins as a result of our study.

Keywords: few-shot class incremental learning (FSCIL), incremental learning, computer vision, robot vision

1 Introduction

Deep learning has revolutionized various computer vision tasks such as image classification, object detection, instance segmentation and visual relation detection; they have become a standard approach in multiple fundamental areas. However, conventional learning methodologies assume the availability of a fixed set of target classes in advance and a large-scale dataset for effective training. Therefore, these methodologies hardly deal with novel classes for intelligent agents deployed in real-world environments. In real-world settings, learning methodologies need to handle novel classes with a restricted number of training samples.



(a) CEC (b) FACT (c) ALICE

(d) SAVC (e) BiDist (f) FeSSSS

Fig. 1: The comparison between the proposed session-wise accuracy and conventional metric for each previous method on CIFAR100.

The above-mentioned issue naturally leads to the task of class incremental learning (CIL) [16]. CIL aims to update a classifier encountering novel classes in an incremental manner. One of the challenges of CIL is striking a balance between catastrophic forgetting of previously-learned classes and overfitting to novel classes. Moreover, the few-shot class-incremental learning (FSCIL) task [23] has recently emerged to formulate CIL in a more practical and challenging setting—the focus of our work. FSCIL attempts to incrementally learn novel classes by utilizing very few training samples without forgetting formerly learned classes over multiple incremental sessions.

In this study, we uncover a critical issue in current FSCIL methods. They exhibit a strong performance bias towards the base session—resulting in significantly lower accuracy in incremental sessions. As Fig. 1 shows, current state-of-the-art (SoTA) FSCIL methods yield poor performance when evaluated solely on each incremental session. Thus, current FSCIL methods exhibit challenges in adapting to novel classes—not aligning with the fundamental objective of FSCIL. Especially, the commonly-used metric in FSCIL, the average accuracy on all the encountered classes, has a bias towards a base session having the largest number of test samples [20]; high performance on base session classes can easily inflate the metric value, and overshadow the underlying problems of poor adaptation to incremental sessions and the presence of performance imbalance. Furthermore, conventional datasets bear analogous characteristics to ImageNet and the inductive bias of convolution would lead to unexpected negative consequences; potentially biasing the model design process. Therefore, we claim the standard evaluation benchmark is not thorough enough to reveal such consequences and results in poor adaptation to incremental sessions.

To address these issues, we propose a novel pipeline, a set of new metrics, and new benchmark datasets for FSCIL. First, we develop a simple-yet-effective pipeline for FSCIL that can effectively learn incremental classes—mitigating the performance bias towards a base session. The pipeline consists of three key components: 1) Feature extraction, 2) Classification modules, and 3) Prediction aggregation. Each compo-

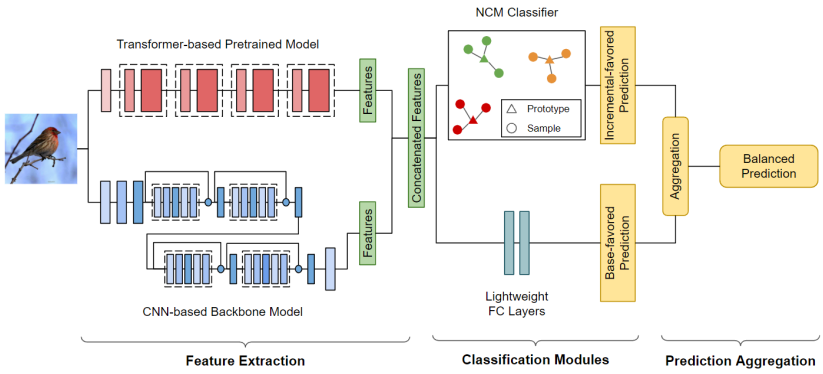


Fig. 2: Overall workflow of our proposed pipeline.

nent leverages pretrained representations, various classification modules, and appropriate aggregation of the predictions, respectively. In order to disclose essential insights for the construction of each component, we conduct comprehensive experiments; our meticulously-designed experiments provide invaluable guidance for designing a strong FSCIL model. Moreover, the proposed pipeline demonstrates superior session-level generalization compared to current SoTA methods.

Next, we introduce a set of novel performance metrics that allow granular session-level evaluations. Specifically, our metrics assess the accuracy of each session individually and quantify the extent of performance bias towards a base session. Unlike the conventional metric, our metrics represent how effectively an FSCIL method generalizes over novel classes and attains balanced performance between base and incremental sessions. Furthermore, we collect two new benchmark datasets for evaluation in the process of designing our study. These datasets facilitate the evaluation of model robustness across various target distributions and the propensity of neural architectures to adapt to different target distributions, thereby aiding in the development of unbiased model design guidelines.

2 Related Work

2.1 Few-Shot Class-Incremental Learning (FSCIL)

The FSCIL task [23] investigates the CIL task in a more practical and challenging setting; an FSCIL method needs to handle novel classes with an extremely small number of training samples although the method can utilize abundant training data for each object class in the base session. Various research efforts have led to the advancement of FSCIL. The beginning work on FSCIL has proposed the TOPIC framework which employs a neural gas network to preserve the topology of features across previous and novel classes [23]. CEC has employed a pseudo incremental learning paradigm and a graph neural network to learn context information for classifiers [31]. FACT has concentrated on forward compatibility and utilized virtual prototypes and virtual instances for securing room for incremental classes in the feature space [33]. ALICE has made

use of the angular penalty loss [26], data augmentation, and class augmentation techniques for learning both discriminative and transferable features [20]. FeSSSS [2] has leveraged self-supervised feature representations and a Gaussian generator to deal with overfitting and catastrophic forgetting [2]. SAVC [22] has aimed to achieve better separation between base classes and virtual classes generated by predefined transformations to capture diverse information. BiDist [32] has presented a distillation structure with two teachers,each alleviating overfitting and forgetting, respectively.

In spite of the effectiveness of conventional FSCIL methods, they happen to focus on improving base session performance and display unsatisfactory performance for each incremental session, which gets hidden by the conventional performance metrics.

2.2 Self-Supervised Pretrained Representations

Self-supervised representation learning obtains labels from the input data using semi-automatic processes and predicts part of the data from other parts [30]—removing the need for manual annotation. Thus, SSL could leverage the massive amount of unlabeled data for feature representation learning. SSL methods fall into either generative or discriminative approaches although there exist other categorizations. Generative approaches require substantial computations and a number of discriminative approaches; especially contrastive approaches are currently displaying state-of-the-art performance. Among recent advanced SSL methods, DeepCluster-v2 [5] clusters previously-learned representations for updating new representations, SeLa-v2 [3] combines clustering and representation learning through principled learning formulation that averts degeneracy, SwAV [5] exploits contrastive methods without requiring to compute pairwise comparisons, Moco-v2 [8] implements the design improvements of SimCLR [7] in the MoCo framework [13], MoBY [29] tunes advances in SSL towards ViTs [12], and DINO [6] reveals the properties of ViTs learned by self-distillation with no labels.

3 Method

In this section, we describe the proposed pipeline for FSCIL that alleviates the bias toward a base session and achieves high performance in both base and incremental sessions. Fig. 2 depicts the overall pipeline.

3.1 Problem Formulation

The FSCIL task includes two types of learning sessions: 1) a single base session and 2) a set of multiple incremental sessions.

Base Session: In the base session, an FSCIL method takes in the training dataset $\mathcal{D}^0_{\text{train}} = \{(\mathbf{x}^0_i, y^0_i)\}_{i=1}^{|D^0_{\text{train}}|}$, where $\mathbf{x}^0_i$ and y^0_i represent the input, *i.e.*, an image, and the ground-truth object class label, respectively. The training dataset of the base session delivers a *sufficient* number of training samples and the FSCIL method measures its performance on the test dataset $\mathcal{D}^0_{\text{test}} = \{(\mathbf{x}^0_i, y^0_i)\}_{i=1}^{|D^0_{\text{test}}|}$.

Incremental Sessions: After the base session, the FSCIL method undergoes a set of incremental sessions in sequence; the FSCIL method accepts a series of datasets $\{\mathcal{D}^1, ..., \mathcal{D}^s, ..., \mathcal{D}^N\}$,

where $\mathcal{D}^s = (\mathcal{D}^s_{\text{train}}, \mathcal{D}^s_{\text{test}})$ representing training and test datasets and N denotes the total number of incremental sessions. There is no overlapping labels between the label sets of object classes for incremental sessions, *i.e.*, $\mathcal{C}^i \cap \mathcal{C}^j = \varnothing$ for $\forall i, j$ and $i \neq j$, where $\mathcal{C}^s$ stands for the set of class labels of s-th session. Moreover, the training datasets of consecutive incremental sessions encompass *insufficient* numbers of training data samples. The C-way K-shot setting illuminates that each incremental session deals with N object classes with K training samples per class. The performance evaluation protocol after the last incremental session handles entire object classes that have been considered over all sessions $\mathcal{C}^0 \cup \mathcal{C}^1 \cup ... \cup \mathcal{C}^N$.

3.2 Feature extraction: Leveraging pretrained representations

Motivation and process In the context of FSCIL, training the feature extractor only in the base session is typically preferred since incremental sessions have small amount of data. However, exclusive training of the feature extractor in the base session exhibits a critical limitation: poor novel class adaptation, which results in a significant performance drop for incremental sessions. To address this, we propose to leverage recent advancements in pretrained representations. We note that instead of utilizing excessively large datasets for pretraining, which can lead to overly simplistic performance improvement, we employ ImageNet, akin to previous methods when training on CUB200 [3].

One straightforward idea for utilizing pretrained representations is fine-tuning. However, updating model parameters in incremental sessions would cause a detrimental impact on base performance. On the other hand, freezing representations for the whole sessions would restrict plasticity. To cope with this dilemma, our architecture leverages two types of features from a frozen pretrained backbone and an FSCIL baseline through concatenation [2].

Design space of feature extraction We investigate the following design space to disclose fundamental insights towards selection of pretrained features—we are the first to study the effect of various feature representations in the context of FSCIL to the best of our knowledge:

(1) Backbone architectures.

- CNN: We utilize the extensively-studied ResNet50 backbone architecture [14].
- ViT: We employ the DeiT-small (DeiT-s) [24] and Swin Transformer-tiny (Swin-t) [17] backbones which have similar numbers of parameters as ResNet50. Further, ViT could have different patch sizes and we denote it as the number at the end such as DeiT-s8 and DeiT-s16.

(2) Learning methodologies for pretraining.

- Supervised learning (SL): Supervised representations have extracted features from ImageNet [21] using the CrossEntropy loss.

[3] Current methods initiate training from ImageNet-pretrained representations due to relatively small amount of base session data in CUB200

– Self-supervised learning (SSL): Self-supervised representations learn from semi-automatic processes and various methods exist. For CNN, we examine DeepCluster-v2 [4], Moco-v2 [8], SeLa-v2 [3], SwAV [5] and DINO [6]; for ViT, DINO [6] and MoBY [29].

3.3 Classification modules: How can we make the most of the extracted representations?

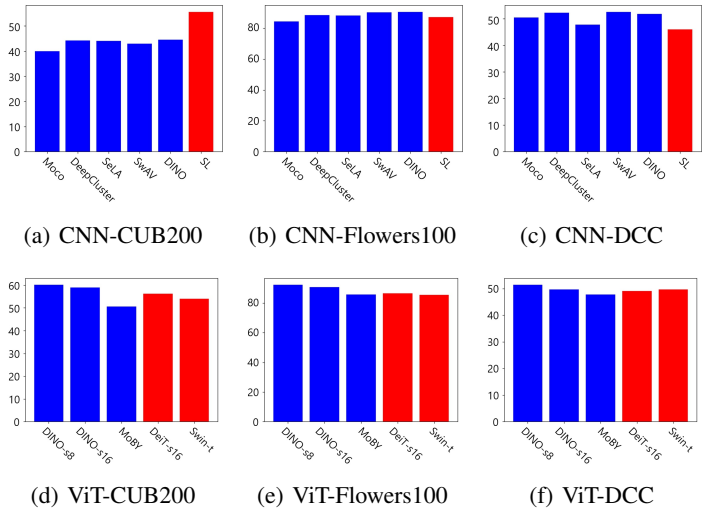
The majority of previous methods have endeavored to attain high performance using a single classifier—leading to complicated yet biased algorithms. However, achieving favorable performance in both base and incremental sessions solely with a single module poses significant challenges due to the plasticity-stability dilemma [1]. In contrast, we propose to decouple the classifier into two distinct entities and exploit two types of classification modules in combination: one designed for base-favored prediction and the other for incremental (inc)-favored prediction. Specifically, the two types are as follows:

Lightweight network (LN) [2]: LN is a 2-layer multi-layer perceptron (MLP) and gets updated by the usual back-propagation. To overcome the catastrophic forgetting, old class features are sampled from the corresponding Gaussian distribution estimated on each session. LN exhibits performance bias toward the base session due to the abundance of samples from base classes. It enables a more accurate estimation of feature distribution, providing more precise base class data compared to other classes during training.

Nearest class mean (NCM) [15]: Metric-based classifier with cosine similarity is another commonly-used classification module in FSCIL. Since the classifier is non-parameteric, it is less susceptible to overfitting, especially in scenarios with limited data such as incremental sessions of FSCIL. We employ the nearest class mean (NCM) classifier with balanced testing [20]. NCM directly utilizes diverse features from the pretrained backbone, resulting in less bias towards a base session but superior incremental session-favored features.

3.4 Prediction aggregation: Enhancing both base and incremental performance

We propose to aggregate the predictions from the two classification modules as the overall prediction. Since the two types of classifiers complement each other, the aggregation of their predictions can yield a synergistic effect—the combination of base-favored and inc-favored predictions for balanced performance. We perform the aggregation by summing the softmax outputs of each module. However, naive summation of the outputs does not lead to performance enhancement. Since LN tends to yield significantly sharper distribution than that of NCM, we use a temperature scheme [28] to control the peakiness of the output. Consequently, the base-favored module slightly dominates when getting evaluated on test data for the base session while the inc-favored module has a slightly stronger influence when evaluation for the incremental sessions, which enhances overall performance balance.



(a) CNN-CUB200 (b) CNN-Flowers100 (c) CNN-DCC

(d) ViT-CUB200 (e) ViT-Flowers100 (f) ViT-DCC

Fig. 3: The average of session-wise accuracy, i.e. A_s, for various types of pretrained representations; blue indicates SSL representations while red indicates SL representations.

4 Evaluation Benchmark

4.1 Datasets

In our experiments, we consider three datasets: one widely-used FSCIL dataset (CUB200 [25]) and two proposed datasets (Flowers100 and DCC). We omit the *mini*ImageNet [21] dataset for fair comparison since we utilize ImgaeNet-pretrained representations in our experiments and the dataset is a subset of ImageNet—resulting in meaninglessly-superb performance. Moreover, we propose two additional benchmark datasets since conventional datasets display analogous distributions as ImageNet [21]. To discover datasets distinct from ImageNet, we establish a criterion: if self-supervised pretraining shows better transferability than its supervised counterpart, it implies a significant distribution shift in the downstream dataset. Drawing from the previous study on the transferability of various pretraining methods on ImageNet [10], we create two FSCIL evaluation benchmarks based on datasets that meets our criterion as detailed below.

CUB200. CUB200 initially designed for fine-grained image classification provides 224×224 sized 11,788 images of 200 object classes. For incremental learning, we partition the 200 object classes into 100 base session classes and 100 incremental session classes to configure a 10-way 5-shot setting following the standard evaluation protocol [23].

Flowers100. Based on the Oxford 102 Folowers dataset [19], we build the Flowers100 dataset. We use the first 100 classes of the dataset excluding the last two classes. Instead of using the original training-test split, we transfer a part of data from the test set to the training set due to the limited number of data available for each class. Specifically, we set the number of base classes to 60 and transfer 20 images per base class from the test set to the training set, resulting in approximately 40 training images per class. Next, we

construct 8 incremental sessions in a 5-way 5-shot manner, and secure approximately 20 test images per class. Finally, we resize the images to 224×224 following the standard.
DTD-CropDiseases-ChestX (DCC). We construct the DCC dataset combining DTD [9], CropDiseases [18] and ChestX [27]. CropDiseases and ChestX are two of the benchmark datasets of the cross-domain few-shot learning (CD-FSL) [11]—they make DCC significantly different from ImageNet. We organize the DCC dataset in the way that each session includes classes from all three datasets (DTD, CropDiseases, and ChestX). Particularly, the base session consists of 67 classes (37 from DTD, 28 from CropDiseases, and 2 from ChestX) with 80 images per class. DCC contains 5 incremental sessions (5-way 5-shot). Moreover, we combine 2, 2, and 1 classes from each of the three datasets to gather 5 classes in each incremental session. For the test dataset, we collect 40 images per class and we resize the images to 224×224.

4.2 Performance Metrics

Limitations of Conventional Metric. The most commonly used metric in FSCIL calculates top-1 accuracy for all sessions encountered up to the current session. However, the metric is biased towards the base session having the largest test samples, thereby hindering an accurate evaluation of adaptability to incremental sessions [20]. Due to this biased metric, the low incremental performance of most current methods is often overshadowed.
Proposed session-level evaluation metrics. In order to conduct a more detailed evaluation, we propose to use the session-wise accuracy in FSCIL, *i.e.*, i-th session accuracy (a_i). a_i independently assesses performance for each session. The key distinction from the previously proposed metrics in FSCIL is that a_i measures accuracy only after training on the last session and considers classes solely in $\mathcal{C}^i$. In addition to the session-wise accuracy, we introduce additional metrics, *Incremental average* (Inc_{avg}), *Imbalance* (IMB), and *Session Average* (A_{s}) as follows:

$$\text{Inc}_{\text{avg}} = \frac{\sum_{i=1}^{N} a_i}{N}, \quad \text{IMB} = a_0 - \text{Inc}_{\text{avg}} \quad \text{and} \quad \text{A}_{\text{s}} = \frac{\sum_{i=0}^{N} a_i}{N+1}, \tag{1}$$

where a_i is i-th session accuracy and N is the number of incremental sessions. Inc_{avg} and IMB evaluate incremental class adaptation and the degree of the bias towards a base session, respectively. A_{s} indicates the average of session-wise accuracy which ensures equal weighting for all sessions in our evaluation.

5 Experiments

5.1 Ablations and analysis

Analysis on utilization of pretrained representations in FSCIL Fig. 3 shows A_{s} for various pretrained representations on the proposed datasets. CNN-based representations display sensitivity to the target dataset's characteristics. For CUB200 which has analogous characteristics as ImageNet, SL ResNet50 attains the leading performance among other pretrained representations.

Method	CUB200				Flowers100				DCC			
	a_0	Inc_{avg}	IMB	A_s	a_0	Inc_{avg}	IMB	A_s	a_0	Inc_{avg}	IMB	A_s
LN	**81.91**	57.74	24.17	59.94	**97.84**	90.84	7.00	91.62	82.31	44.70	37.61	50.97
NCM	74.62	**62.22**	**12.39**	**63.35**	93.67	**92.85**	**0.82**	**92.94**	77.99	**54.50**	**23.49**	**58.41**
LN + NCM	79.54	**63.77**	15.77	**65.20**	96.42	**94.11**	2.31	**94.37**	80.97	**55.00**	**25.97**	**59.33**

Table 1: Experimental results for individual classification modules and their aggregation. Best results are in bold and second-best are underlined.

Method	CUB200				Flowers100				DCC			
	a_0	Inc_{avg}	IMB	A_s	a_0	Inc_{avg}	IMB	A_s	a_0	Inc_{avg}	IMB	A_s
CEC [31]	71.30	27.47	43.84	31.45	96.00	62.46	33.54	66.19	76.67	29.85	46.82	37.65
FACT [33]	72.84	39.78	33.06	42.78	96.71	77.45	19.26	79.59	79.85	26.64	53.21	35.51
ALICE [20]	68.44	51.65	16.79	53.17	93.25	84.71	8.54	85.66	76.53	44.90	31.63	50.17
SAVC [22]	77.44	47.53	29.91	50.25	97.75	83.45	14.30	85.04	81.53	42.60	38.93	49.09
BiDist [32]	71.44	45.66	25.78	48.00	95.67	82.70	12.97	84.14	75.78	48.60	27.18	53.13
FeSSSS [2]	**81.91**	57.74	24.17	59.94	**97.84**	90.84	7.00	91.62	**82.31**	44.70	37.61	50.97
Ours	79.54	**63.77**	**15.77**	**65.20**	96.42	**94.11**	**2.31**	**94.37**	80.97	**55.00**	**25.97**	**59.33**

Table 2: Comparative study results. The results of a_0 differ from the original papers since we used the proposed metric.

On the other hand, SSL CNNs shows superior performance than SL CNN for Flowers100 and DCC which entail distinctive characteristics compared to ImageNet. In this case, certain SSL CNNs such as DeepCluster-v2, SwAV and DINO even surpass supervised ViTs. Notably, the performance gap between SSL and SL widens as the target distribution diverges further from that of ImageNet—the largest gap for DCC. In the case of ViT, they do not exhibit dependency on the distribution of the target dataset. Especially, DINO-s8 consistently outperforms other categories of pretrained representations. Therefore, we identify that DINO-s8 is powerful representation robust to target distribution shift and adopt it for our feature extraction process.

Analysis on the classification modules Table 1 presents the results of the proposed session-level metrics for LN and NCM. LN achieves high base accuracy but low Inc_{avg}, while NCM shows lower base accuracy and higher Inc_{avg}. These results corroborate our hypothesis that LN and NCM can serve as base-favored and inc-favored modules, respectively.

Analysis on the prediction aggregation Table 1 displays the results for the prediction aggregation process. LN+NCM enhances Inc_{avg} compared to NCM, with only a slight decline in base accuracy compared to LN. Moreover, the aggregated results (LN+NCM) exhibit significantly higher performance in incremental sessions compared to the base-favored module (LN) and in base sessions compared to the inc-favored module (NCM). While NCM entails a lower IMB, it is mainly due to its low base accuracy which is also a crucial metric for FSCIL. Hence, in this context, we could prioritize achieving higher A_s for an effective FSCIL module. Therefore, these results indicate that LN+NCM—our

prediction aggregation process—not only reduces performance bias but also achieves high overall performance, resulting in an effective FSCIL module[4]

5.2 Comparative study

Table 2 presents the comparative results between the previous methods and our pipeline using the proposed metrics. Our pipeline significantly outperforms the previous methods, achieving a new SoTA. Notably, our newly introduced session-level metrics uncover that the baselines display poor adaptability to the incremental sessions (low Inc_{avg}) and A_{s}, which was obscured by the conventional metric. On the other hand, our pipeline effectively resolves these limitations and demonstrates substantial performance superiority.

While one might assume the performance improvement is solely due to the larger backbone, our superior balanced and overall performance compared to FeSSSS [2], which uses the same backbones as the proposed pipeline in our experiments, indicates that our enhancement goes beyond the larger backbone, resulting from the harmonious aggregation of two complementary classification modules.

6 Conclusion

In this study, we unveiled the critical bias towards a base session in FSCIL. To address this, we proposed a pipeline consisting of three modules and introduced novel session-level metrics as well as new benchmark datasets for meticulous analysis and evaluation of robustness to the target distribution shift. We conducted a comprehensive range of experiments and showed our pipeline's effectiveness in both mitigating the bias and facilitating novel class adaptation. It successfully achieves a superior balance, enhancing the overall performance on the new benchmark datasets as evidenced by the introduced metrics. We believe our proposed pipeline holds the potential to make a substantial contribution to the advancement of FSCIL methods.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2022-0-00926, Development of Self-directed Visual Intelligence Technology Based on Problem Hypothesis and Self-supervised Methods; No.2022-0-00907, Development of AI Bots Collaboration Platform and Self-organizing AI; No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1C1C1009989).

[4] Note that our prediction aggregation process is more than simple ensemble operations. Simply combining a large number of classification modules does not necessarily lead to improvement. This is because many current methods excel only in the base, and combining such methods does not address the persisting issue of performance bias.

References

1. Abraham, W.C., Robins, A.: Memory retention–the synaptic stability versus plasticity dilemma. Trends in Neurosciences 28(2), 73–78 (2005)
2. Ahmad, T., Dhamija, A.R., Cruz, S., Rabinowitz, R., Li, C., Jafarzadeh, M., Boult, T.E.: Few-shot class incremental learning leveraging self-supervised features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3900–3910 (2022)
3. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371 (2019)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision. pp. 132–149 (2018)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems 33, 9912–9924 (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
9. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3606–3613 (2014)
10. Ericsson, L., Gouk, H., Hospedales, T.M.: How well do self-supervised models transfer? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5414–5423 (2021)
11. Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. pp. 124–141. Springer (2020)
12. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(1), 87–110 (2022)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
15. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 831–839 (2019)
16. Karim, M., Verma, V., Singh, P., Namboodiri, V., Rai, P.: Knowledge consolidation based class incremental online learning with limited data. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 2621–2627 (2021)

17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
18. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Frontiers in plant science 7, 1419 (2016)
19. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
20. Peng, C., Zhao, K., Wang, T., Li, M., Lovell, B.C.: Few-shot class-incremental learning from an open-set perspective. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV. pp. 382–397. Springer (2022)
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3), 211–252 (2015)
22. Song, Z., Zhao, Y., Shi, Y., Peng, P., Yuan, L., Tian, Y.: 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. pp. 24183–24192 (2023)
23. Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., Gong, Y.: Few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12183–12192 (2020)
24. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
25. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
26. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
27. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017)
28. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018)
29. Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H.: Self-supervised learning with swin transformers. arXiv preprint arXiv:2105.04553 (2021)
30. Zhang, C., Zhang, C., Song, J., Yi, J.S.K., Kweon, I.S.: A survey on masked autoencoder for visual self-supervised learning. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 6805–6813 (2023)
31. Zhang, C., Song, N., Lin, G., Zheng, Y., Pan, P., Xu, Y.: Few-shot incremental learning with continually evolved classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12455–12464 (2021)
32. Zhao, L., Lu, J., Xu, Y., Cheng, Z., Guo, D., Niu, Y., Fang, X.: Few-shot class-incremental learning via class-aware bilateral distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11838–11847 (2023)
33. Zhou, D.W., Wang, F.Y., Ye, H.J., Ma, L., Pu, S., Zhan, D.C.: Forward compatible few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9046–9056 (2022)