Revisiting Pool-based Prompt Learning for Few-shot Class-incremental Learning

Supplementary Material

1. Detailed Review of Related Work

1.1. Visual Prompting

Visual Prompt Tuning (VPT) [16] pioneered the adaptation of prompt learning to computer vision by introducing learnable prompt tokens into vision transformers. The core principle involves prepending learnable parameters to the input sequence while keeping the backbone frozen, achieving performance comparable to full fine-tuning with only 0.1% of trainable parameters. This parameter-efficient approach was inspired by prompt learning in Natural Language Processing (NLP) [3, 23], where it evolved from manual text prompts to learnable continuous vectors [21]. The success of prompt learning has also catalyzed various developments in vision-language models, such as CLIP [30] and CoOp [60].

Efficiency and Adaptability: DePT [10] optimizes source-initialized prompts for test-time adaptation, while E2VPT [4] introduces token-wise and segment-wise prompt pruning to reduce parameter overhead. PViT [55] develops task-specific prompts for multi-task scenarios, and LPT [8] addresses long-tailed distribution challenges by combining shared and group-specific prompts.

Continual Learning Applications: Learning to Prompt (L2P) [49] made a significant breakthrough by introducing a pool mechanism to maintain shared prompts across tasks. This approach was further enhanced by DualPrompt [48], which introduced complementary general and expert prompts to capture both task-invariant and task-specific knowledge. These methods have demonstrated superior performance compared to traditional parameter-efficient approaches such as Adapter [29] and LoRA [13]. Recent advances include CODA-Prompt [35], which improves prompt-based continual learning through cross-task knowledge distillation, and S-Prompts [47], which explores similar strategies in domain-incremental learning scenarios. The latest HideP [44] introduces a hierarchical decomposition approach to optimize different components of continual learning separately.

Instance-Level Adaptation: Recent work has focused on addressing the challenge of capturing instance-specific features from different perspectives. InsVP [24] introduces a dual-level instance visual prompting scheme, which leverages both image-level and feature-level instance-specific information to enhance model recognition capability. Similarly, DAM-VP [14] proposes a cluster-level visual prompting method that learns distinct prompt sets for different sample clusters.

Despite these advances, several challenges remain in prompt-based visual learning. Current pool-based prompting methods still struggle with efficient prompt selection and utilization, especially when spatial information plays a vital role in distinguishing fine-grained features. The static nature of prompt pool structures may limit the model's ability to fully leverage task relationships and adapt to novel scenarios. Moreover, the complex interaction between prompts and vision transformer layers requires deeper investigation to optimize knowledge transfer and mitigate catastrophic forgetting. While recent advances have addressed some of these limitations through instance-level adaptation and spatial information utilization, there remains significant room for improvement in prompt pool design and management, particularly in effectively combining local and global spatial information while maintaining model efficiency.

1.2. FSCIL

FSCIL represents a sophisticated integration of Few-Shot Learning (FSL) and Class-Incremental Learning (CIL) [2]. This paradigm inherits the fundamental challenge of FSL in learning from limited samples, while simultaneously addressing the CIL requirement of maintaining model performance on previously learned tasks as new classes are incrementally introduced.

Structure-based and Dynamic Network Methods. Few-Shot Class-Incremental Learning (FSCIL) was pioneered by Tao et al. [40], who introduced a neural gas network to preserve topological relationships between classes. This seminal work laid the foundation for subsequent research, notably CEC [54], which advanced the field by introducing a class-specific classifier architecture integrated with a graph-based model for efficient inter-classifier information propagation and adaptation. Recent developments in this domain have witnessed a significant shift towards dynamic network architectures, encompassing several key directions: dynamic backbone structures (DSN) [51], structure fusion mechanisms (FeSSSS) [1], adaptive sub-networks (SoftNet) [17], and ensemble-based frameworks (LEC-Net) [50]. State-of-the-art methods such as EM [62] have demonstrated the remarkable efficacy of dynamic architectural adaptation in addressing the fundamental challenges of FSCIL.

Knowledge Distillation and Replay-Based Methods. A significant body of research in FSCIL focuses on knowledge preservation through replay buffers or feature generation mechanisms. LCwoF [20] and SKD [32] pioneered this direction by incorporating feature pyramid architectures to enhance knowledge preservation capabilities.

Dataset	Domain	Images	Classes	Base	Base Setup	Novel	Novel Setup	Train/Test	Source
CUB-200	Fine-grained Birds	11,788	200	100	100w-30s (3K)	100 (10/ses)	10×(10w-5s)	30/50	Caltech
iNF200	Fungi Species	10,000	200	100	100w-50s (5K)	100 (10/ses)	10×(10w-5s)	50/50	iNat
FGVCAircraft	Aircraft Variants	10,000	100	50	50w-30s (1.5K)	50 (5/ses)	10×(5w-5s)	30/50	FGVC

Table 1. Detailed dataset configurations for fine-grained visual recognition. Base indicates the number of categories used in initial training (training samples in parentheses), while Novel shows how remaining categories are distributed across 10 incremental sessions. For base training, Train/Test denotes samples per class for training/testing respectively. In novel sessions, we consistently use 5 samples for training and the remaining for testing. The setup format (N-way-K-shot) specifies N categories with K samples per category for both base and novel phases.

Building upon this foundation, [34] employed Generative Adversarial Networks to synthesize features, effectively addressing the data scarcity challenge inherent in few-shot scenarios. Recent advances have further refined these approaches: CABD [56] introduced class-aware bilateral distillation, facilitating knowledge transfer between base and novel classes through a sophisticated distillation framework, while UaD-CE [6] enhanced performance through an innovative combination of reference model distillation and uncertainty-aware adaptive mechanisms. While these methods demonstrate promising results, they typically necessitate additional computational resources for storing historical samples or maintaining teacher models.

Feature Space-based Methods. Early explorations in feature space optimization for FSCIL include LIMIT [58] and WaRP [18]. iCaRL [33] introduced a practical strategy for simultaneously learning classifiers and feature representations in the class-incremental setting. FOSTER [43] advanced this direction through a two-step paradigm of feature boosting and compression, effectively addressing the performance decline in new classes. Building upon these foundations, FACT [57] introduced Forward Compatible Training, an innovative approach that reserves representation space through strategic prototype placement. TEEN [45] proposed a training-free prototype calibration strategy to address the misclassification of new classes into similar base classes. RDI [59] further investigated the base-novel confusion problem by decoupling label-irrelevant redundancies in both feature and pixel spaces. ALICE [28] advanced this direction by emphasizing both feature space compactness and diversity, while SAVC [37] incorporated semantic knowledge through virtual class generation.

Meta Learning-based and Optimization-based Methods. Meta learning and optimization techniques have emerged as powerful approaches in FSCIL, incorporating various regularization strategies and adaptive loss functions. SPPR [61] enhances feature representation through an innovative prototype refinement strategy, while CLOM [63] introduces a novel cosine loss with negative margin to promote shareable feature learning. C-FSCIL [11] proposes quasi-orthogonal prototype alignment, and Comp-FSCIL [64] introduces a cognitive-inspired approach utilizing primitive composition and reuse modules based on set similarities. NC-FSCIL [52] employs dot-regression loss to optimize feature-prototype alignment. Earlier contributions include MetaFSCIL [5] and FSLL [26], which established fundamental meta-learning frameworks for incremental few-shot scenarios.

Pretrained Model-based Methods. Recent advances in FSCIL have increasingly leveraged the remarkable capabilities of pre-trained models. Several pioneering works [9, 15, 53] have explored multi-modal approaches by incorporating CLIP [31], establishing novel paradigms for feature alignment between visual and textual modalities. In parallel, significant progress has been made in utilizing Vision Transformers for FSCIL tasks. ASP [22] demonstrates the effectiveness of this approach by employing a fixed pre-trained backbone while encoding task-invariant knowledge in learnable prompts, effectively addressing the stability-plasticity dilemma. Yourself [39] further advances this direction through an innovative feature rectification module, while Approximation [46] establishes comprehensive guidelines for mitigating transfer and consistency risks. A significant breakthrough came with PriViLege [27], which effectively addresses catastrophic forgetting through its novel PKT module and semantic knowledge distillation mechanism. Our work builds upon these recent advances, further exploring the potential of pre-trained models in incremental learning scenarios.

2. Empirical Setup

2.1. Experimental Datasets

We first discuss our dataset selection rationale, focusing on distribution overlap concerns with pre-trained weights. Then we present our three benchmark datasets and their partitioning protocols, followed by evaluation metrics.

2.1.1. Dataset Selection Rationale

CIFAR-100 [19], CUB-200 [42], and miniImageNet [41] have served as standard benchmark datasets in the field. However, we identify a critical limitation in their evaluation capacity for our context. Given that our approach leverages ImageNet pre-trained weights—a common practice shared by [27], [46], [39], [22], and [38]—we find that CIFAR-100 and CUB-200 share substantial distribution similarities with

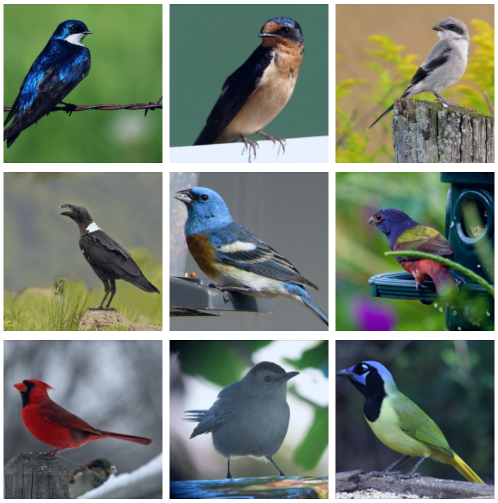

Figure 1. Samples of CUB-200.

ImageNet. This overlap raises concerns about potential data leakage, particularly when applying pre-trained weights to miniImageNet (a subset of ImageNet), which could lead to artificially enhanced performance metrics.

To ensure a more rigorous evaluation of our proposed method, we extend our experiments to two challenging fine-grained datasets: FGVCAircraft [25] and a carefully selected subset from iNaturalist (Fungi) [12]. Specifically, we work with the first 200 fungal species classes from iNaturalist, referred to as iNF200. These datasets offer distinct visual distributions from ImageNet, enabling a more comprehensive assessment of our method's generalization abilities. Table 1 provides the detailed configurations of our experimental datasets, including class distribution, training/testing splits, and the session setup for both base and novel classes.

2.1.2. Benchmark Datasets.

CUB-200 (Caltech-UCSD Birds) is a fine-grained visual classification dataset comprising 200 bird species with a total of 11,788 images, with sample images shown in Figure 1. In our incremental learning setup, we designate the first 100 species as base classes. The remaining classes are evenly distributed across 10 incremental sessions, with each session introducing 10 novel species. Following the few-shot learning paradigm, we utilize 5 samples per class during training, establishing a 10-way 5-shot incremental learning scenario.

FGVCAircraft presents a challenging collection of 10,000 aircraft images spanning 100 distinct aircraft variants, with representative samples shown in Figure 2. We structure our experiments by allocating 50 variants as base classes, with the remaining 50 variants distributed across 10 incremental learning sessions. Each session incorporates 5 new aircraft variants, and similar to our other experiments, we maintain a 5-shot learning protocol, resulting in a 5-way 5-shot incremental learning framework.

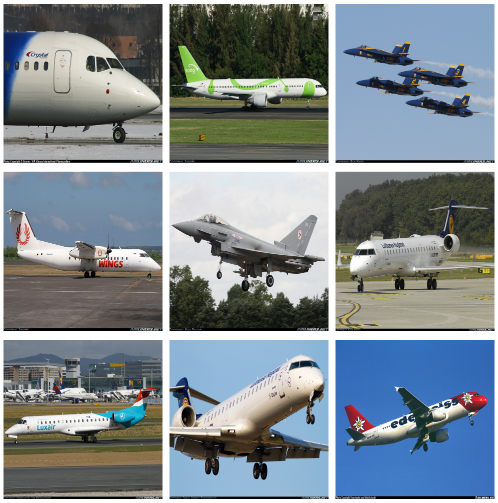

Figure 2. Samples of FGVCAircraft.

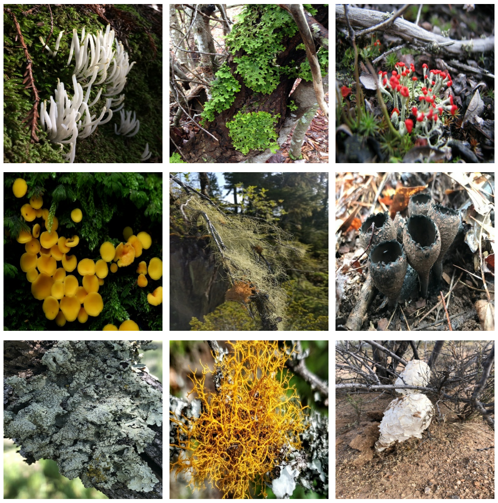

Figure 3. Samples of iNF200.

iNF200 (iNaturalist Fungi-200) represents a carefully curated subset of the iNaturalist dataset's fungi category, with diverse examples illustrated in Figure 3. From the extensive collection of fungal species available in iNaturalist, we select the first 200 classes, each containing 50 images, yielding a total of 10,000 samples. Our experimental protocol assigns the initial 100 species to the base training set. The subsequent 100 species are systematically distributed across 10 incremental sessions, with each session introducing 10 new species. Maintaining consistency with our experimental design, we employ 5 examples per class, establishing a 10-way 5-shot incremental learning configuration.

2.1.3. Dataset Partitioning Protocol

In our experimental design, we adhere to established dataset partitioning protocols to maintain consistency and reproducibility. The CUB-200 dataset follows the division scheme outlined in CEC [54], while FGVCAircraft and iNF200 are organized according to the protocol described in [7]. This careful consideration in dataset selection and organization provides a robust evaluation framework that effectively addresses the limitations of ImageNet-aligned datasets.

2.1.4. Evaluation Protocol

We employ three complementary metrics to comprehensively evaluate our model's performance. For each incremental session t, we compute the classification accuracy A_t. The overall model performance is assessed through the average accuracy across all sessions, computed as Avg $= \frac{1}{T}\sum_{i=1}^{T} A_i$.

2.2. Model Implementation

2.2.1. Model Architecture and Training Strategy

Our framework builds upon the Vision Transformer (ViT) architecture, leveraging a model pre-trained on ImageNet-21K as our backbone network. We adopt and extend the Visual Prompt Tuning (VPT) paradigm [16], specifically following the VPT-deep configuration, which has demonstrated superior performance in visual transfer learning tasks. In this setup, we strategically insert learnable prompt tokens between consecutive transformer blocks, enabling fine-grained feature adaptation while maintaining the pre-trained model's robust representational capacity. The prompt parameters are initialized from a normal distribution, following standard practices in transformer-based architectures.

Following the established practice in few-shot class-incremental learning [22, 27, 38, 39, 46], we keep the pre-trained ViT backbone frozen during the entire learning process. This conventional strategy has proven effective in preventing catastrophic forgetting and maintaining stable feature representations. During training, we only optimize the inserted visual prompts and the prototype-based classifier [36], which aligns with the field's standard approach of selective parameter updating when adapting to novel classes.

2.2.2. Implementation Details

The training process consists of two main phases. In the base session, we train the model for 16 epochs with a learning rate of 0.02. This is followed by the novel session, where we fine-tune for 5 epochs with a reduced learning rate of 0.002. For the Local Spatial Pool, we maintain 30 learnable spatial prompts (CNN & Local Spatial Prompts), each initialized with Kaiming initialization and employing dropout (rate = 0.1) for robust feature extraction. These prompts are trained with a learning rate of 0.001 during the base session.

In the Global Spatial Pool, we design frequency-based prompts organized in concentric regions. These prompts, represented as trainable frequency masks, are optimized throughout both sessions with learning rates of 0.1 and 0.005 for base and novel sessions, respectively. The mask weights are initialized using a normal distribution to ensure balanced initial frequency responses. For adaptive feature integration between local and global spatial information, we introduce two learnable parameters α_l and α_g, constrained by α_l + α_g = 1, with both initially set to 0.5.

Integration Strategy	Acc (%)	Δ
Ours (Learnable α_l and α_g)	**81.392**	-
Architecture Variants		
GSP after LSP	81.242	-0.150
GSP before LSP	81.175	-0.217
Linear layer fusion	81.260	-0.132
Weight Constraints		
Fixed sum constraint ($\alpha_l + \alpha_g = 1$)	80.179	-1.213
Only LSP weight learnable	80.159	-1.233
Only GSP weight learnable	80.139	-1.253
Advanced Fusion Mechanisms		
Neural network weights	80.152	-1.240
Gating mechanism (Sigmoid)	80.503	-0.889
Non-linear fusion (MLP)	80.682	-0.710

Table 2. Ablation study on different feature integration strategies. Our proposed learnable weights achieve the best performance (81.392%) compared to various alternative fusion mechanisms.

These fusion parameters are exclusively optimized during the base session with a learning rate of 0.005, allowing the model to learn optimal combinations of local and global spatial features. To ensure stable convergence, we adopt a CosineAnnealingLR schedule for all learning rate adjustments. All experiments are conducted on a single GPU setup.

3. Hyperparameter Experiments

3.1. Feature Integration Mechanism Study

We evaluate various feature integration strategies to validate our design choice of using learnable weights α_l and α_g. As shown in Table 2, our experiments explore three aspects of the integration mechanism:

First, for architectural design, different choices for combining LSP and GSP, including sequential integration and linear fusion, show competitive but slightly inferior performance (∼81.2%) compared to our parallel processing with learnable weights. When examining weight constraints, experiments with fixed sum or single learnable weight show significant performance degradation (∼80.1-80.2%), validating our choice of independent weight optimization. Moreover, more sophisticated approaches like neural net-

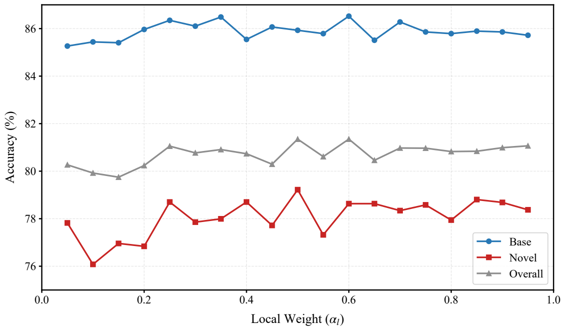

Figure 4. Impact of adaptive weighting mechanism: Analysis of accuracy variations with different initial weight combinations between Local (α_l) and Global (α_g) Spatial Prompting branches. The model demonstrates robust performance across various weight settings, with optimal performance achieved near balanced weights.

works, gating mechanisms, and MLPs show inferior performance (80.15-80.68%), suggesting that complex fusion strategies might introduce unnecessary optimization difficulties.

We further investigate the impact of different weight initializations between LSP and GSP branches. As shown in Fig. 4, our analysis reveals that the best performance is achieved with balanced weight distributions (α_l = 0.5–0.6), suggesting complementary contributions from both local and global spatial information. The model maintains robust performance (above 79.5%) across various weight initializations, demonstrating strong stability. Notably, base classes show slight preference for higher local weights (α_l = 0.6), while novel classes achieve optimal performance with balanced weights ($\alpha_l = 0.5$).

These comprehensive experiments yield several important insights for our framework design:

- Simple learnable weights outperform complex fusion mechanisms, emphasizing flexibility over complexity
- Independent optimization of LSP and GSP weights is crucial for optimal performance
- The framework shows remarkable robustness to initialization while maintaining effective weight learning
- Balanced contribution from both spatial perspectives is essential for overall performance

3.2. Detailed Analysis of LSP's Effectiveness

We conduct extensive experiments to validate the effectiveness of our proposed Local Spatial Prompting (LSP) module across different architectures. Table 3 presents a detailed breakdown of the performance metrics in the final session, where we evaluate the model's capability in both within-space and cross-space classification scenarios.

Our experimental results reveal several significant findings:

Model	**B→B+N**	**N→B+N**	**B→B**	**N→N**
VPT-Deep	81.18	73.14	84.71	73.92
VPT-Deep + LSP	**83.41**	**73.41**	**84.77**	**75.77**
VPT-Shallow	78.25	67.44	81.60	69.01
VPT-Shallow + LSP	**82.12**	**70.27**	**83.28**	**72.46**

Table 3. Detailed performance comparison of LSP across different architectures in the final session. B→B+N denotes the accuracy of base classes' logits mapped to both base and novel classes space; N→B+N represents novel classes' logits mapped to the complete class space; B→B indicates base classes' accuracy within base class space; N→N shows novel classes' accuracy within novel class space.

Universal Effectiveness. LSP demonstrates consistent performance improvements across both deep and shallow architectures. Notably, in the final session, both architectures show enhanced performance across all evaluation metrics, validating LSP's architectural versatility.

Enhanced Cross-space Recognition. The B→B+N and N→B+N metrics show significant improvements with LSP. In VPT-Deep configuration, B→B+N accuracy improved by 2.23% (81.18% → 83.41%) while N→B+N improved by 0.27%. More substantial gains are observed in VPT-Shallow, with B→B+N increasing by 3.87% and N→B+N by 2.83%.

Robust Within-space Performance. The within-space metrics (B→B and N→N) also show consistent improvements. VPT-Deep + LSP achieves 84.77% B→B accuracy and 75.77% N→N accuracy, while VPT-Shallow + LSP shows more significant gains, particularly in N→N performance with a 3.45% improvement.

Stability Across Sessions. As shown in Figure 5, models equipped with LSP exhibit more stable performance trajectories, particularly in later incremental stages. This stability can be attributed to the spatial dimension exploitation that effectively mitigates token-dimension saturation.

These empirical results strongly support our theoretical framework:

- The comprehensive improvements in both cross-space (B→B+N, N→B+N) and within-space (B→B, N→N) metrics validate our hypothesis about token-dimension saturation and the effectiveness of spatial-dimension prompting as a solution.
- The enhanced stability in later sessions, coupled with improved N→N performance, confirms that LSP successfully maintains discriminative feature learning capability throughout the incremental learning process.
- The balanced improvements across all metrics suggest that LSP effectively manages the stability-plasticity trade-off inherent in FSCIL scenarios, particularly benefiting the challenging cross-space recognition tasks.

This comprehensive analysis demonstrates that LSP not only provides quantitative performance improvements but

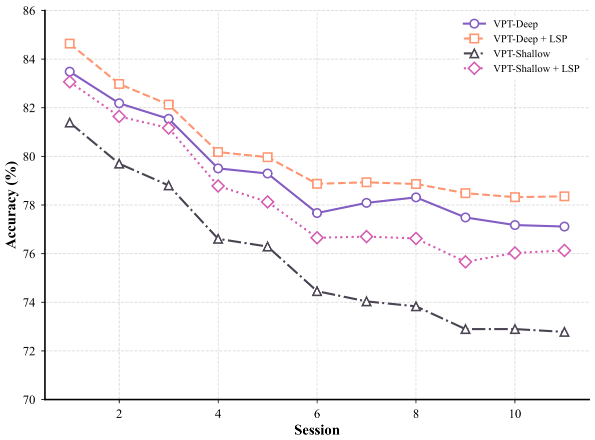

Figure 5. Performance comparison of different VPT variants across sessions. The proposed LSP module consistently improves the performance on both VPT-Deep and VPT-Shallow architectures, demonstrating its effectiveness in addressing the token-dimension saturation problem. Notably, the performance gain is more significant in the VPT-Shallow setting (+3.348%) compared to VPT-Deep (+1.243%), indicating LSP's particular strength in resource-constrained scenarios.

also addresses fundamental challenges in FSCIL through its innovative spatial prompting mechanism, especially in scenarios requiring cross-space knowledge transfer.

3.3. Analysis of Soft Pool vs. Hard Pool Methods

Motivated by our visualization analysis that revealed token-dimension saturation issues, we conducted experiments comparing soft and hard pooling strategies. The soft pooling approach attempts to address the saturation problem by computing similarity-based weighted combinations of prompts, potentially allowing the model to utilize information from multiple prompts while avoiding direct token-dimension conflicts.

As shown in Figure 6, our experimental results with carefully controlled prompt numbers (VPT length = 5, selected pool size = 5) reveal several key findings:

Consistent Performance Gap: Even in this controlled setting where token-dimension saturation is avoided, hard pooling consistently outperforms soft pooling across all configurations. The performance gap (1.30% improvement in Pool-Deep, 1.09% in VPT-Pool-Deep) suggests that the limitations of soft pooling are not merely due to token saturation but rather inherent to the weighted combination approach itself.

VPT Integration Benefits: The integration of VPT (length = 5) with pool-based methods shows improved performance in both variants, with VPT-Pool-Deep achieving the best results (75.70% for soft, 76.79% for hard). This indicates that the initialized VPT prompts provide complementary features to the selected pool prompts when their total number is kept below the saturation threshold.

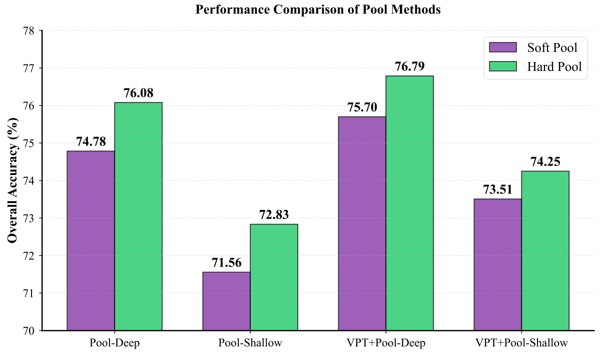

Figure 6. Performance comparison between soft and hard pooling methods. All experiments are conducted with a fixed VPT length of 5 and pool selection size of 5, which is determined based on our previous analysis showing that token-dimension conflicts emerge when the total number of prompts exceeds 10. This controlled setting allows us to: (1) maintain both VPT and pool prompts below the saturation threshold, ensuring both modules can potentially contribute positively to the model's performance, and (2) specifically investigate whether soft pooling can effectively aggregate information from all prompts in the pool without interference from saturation effects.

Architecture Impact: Deep architectures demonstrate superior performance across both pooling strategies, suggesting better capability in utilizing the limited number of prompts effectively.

These findings indicate that while soft pooling was designed to potentially leverage information from all prompts through weighted combinations, it fails to outperform the direct selection mechanism of hard pooling even in scenarios carefully designed to avoid token-dimension saturation. This suggests that the challenge lies not in the quantity of information that can be stored in the token dimension, but rather in how effectively this information can be utilized. The superior performance of hard pooling, even with the same number of prompts, supports our motivation to explore alternative dimensions for prompt learning rather than attempting to optimize information aggregation within the token dimension.

3.4. Analysis of VPT and Pool Architecture Variants

As illustrated in Fig. 7, we first analyze the performance of different Visual Prompt Tuning architectures on the CUB-200 dataset. The results show that VPT-deep consistently outperforms VPT-shallow across all sessions, maintaining higher accuracy (77.2% vs. 72.8% at session 10) and exhibiting better stability in knowledge retention. This performance gap motivates us to extend the deep integration strategy to the pool-based mechanism, leading to the development of Pool-deep architecture for our subsequent investigation into prompt pool mechanisms for FSCIL tasks.

The superior performance of deep variants can be attributed to their hierarchical feature modulation capability,

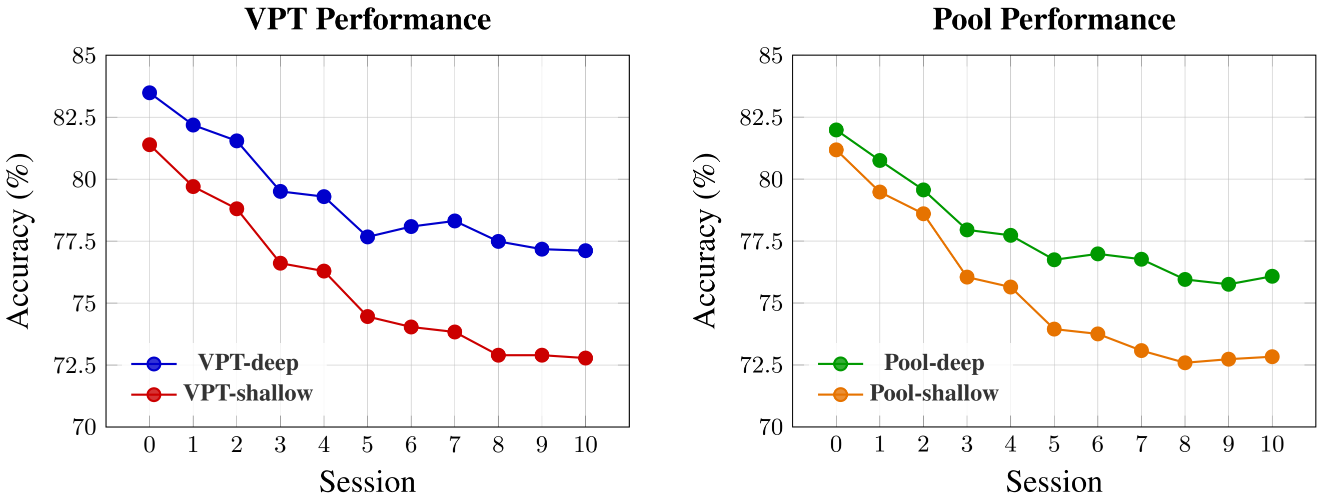

Figure 7. Performance comparison of VPT and Pool methods on the CUB-200 dataset. Both approaches demonstrate superior performance with deep configuration, showing consistent advantages in maintaining accuracy across incremental sessions. The deep variants (VPT-deep and Pool-deep) consistently outperform their shallow counterparts, with VPT-deep achieving the best overall performance and exhibiting more stable accuracy retention.

where prompts are inserted at multiple transformer layers (L_1, L_2, ..., L_N). This design enables the model to capture and modify features at various abstraction levels, from low-level visual patterns to high-level semantic concepts. While the original prompt pool method [49] only utilized shallow integration at the input layer, we enhance it to Pool-deep by extending prompt insertion across multiple layers. As shown in Fig. 7, this improvement leads to consistently better results compared to Pool-shallow (76.1% vs. 72.9% at session 10). The performance curves also indicate that deep architectures are more resilient to catastrophic forgetting, maintaining more stable accuracy across incremental sessions.

4. Extended Attention Pattern Analysis

We provide an extended visualization of attention patterns in Figures 8 and 9 to further demonstrate the effectiveness of our Local Spatial Prompting mechanism across a diverse set of samples. These additional examples consistently show how our approach reshapes the attention distribution, shifting focus from background elements to semantically meaningful regions of the target objects.

In these extended examples, we observe several key patterns:

Consistent Focus Improvement: Across all samples, the prompt attention maps (pmt) show more precise focus on the target objects compared to the baseline class token attention (cls).

Background Suppression: Our method effectively reduces attention to irrelevant background elements, which is particularly evident in samples with complex or cluttered backgrounds.

Feature Highlighting: The prompt attention maps consistently highlight discriminative features of the objects, suggesting that our Local Spatial Prompting mechanism helps the model learn more robust and relevant feature representations.

These additional visualizations further support our main findings and demonstrate the generalizability of our approach across different scenarios and object types.

5. Limitations

While our proposed Local-Global Spatial Prompt Framework advances Few-Shot Class Incremental Learning, it faces notable limitations. The spatial prompts buffer introduces additional memory overhead, challenging deployment in resource-constrained environments. Additionally, the framework's effectiveness heavily depends on adequate pre-training, which may not be available in domains with limited labeled data or rapidly changing environments. These limitations suggest important directions for future research to enhance the framework's practical applicability.

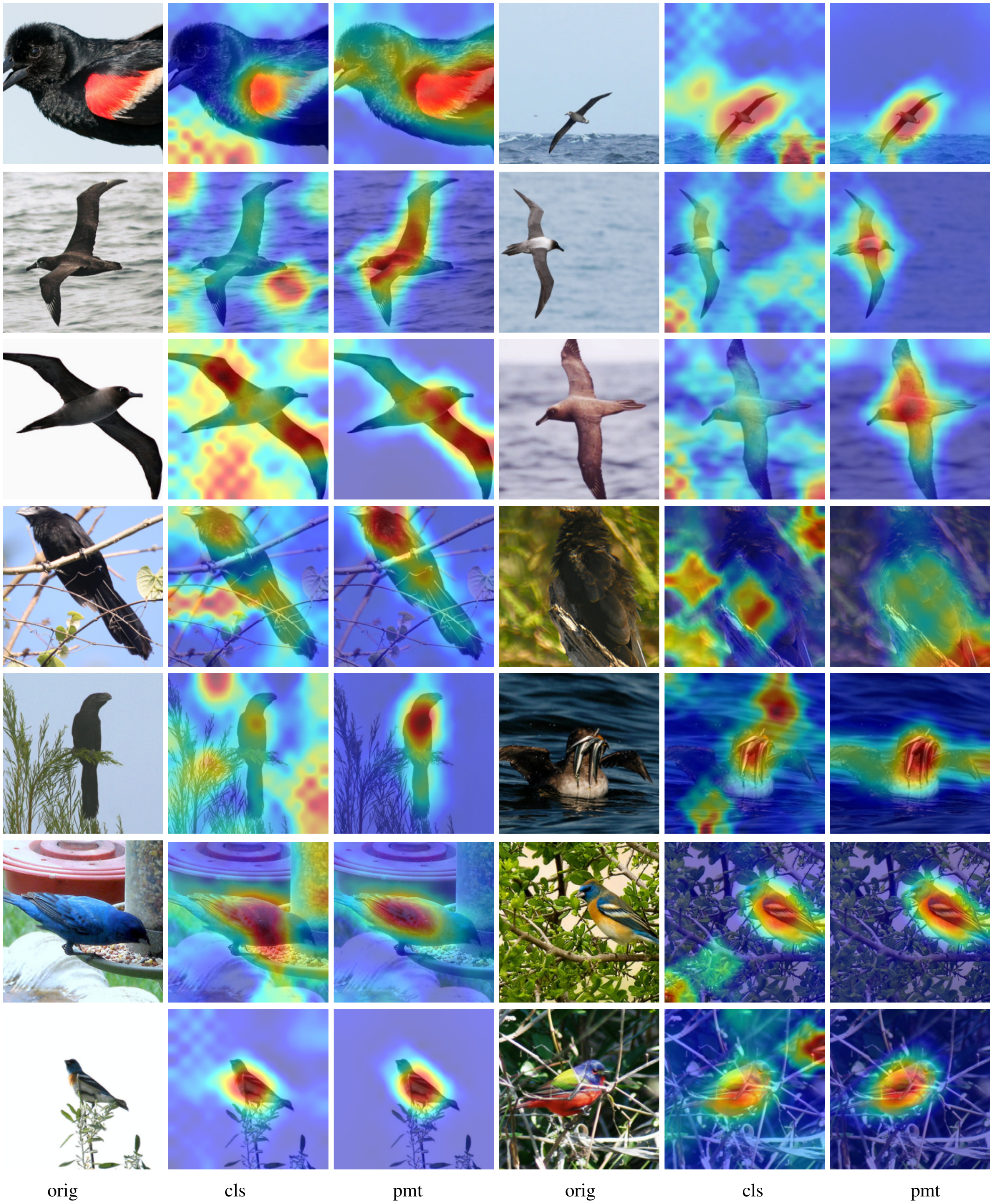

Figure 8. Extended visualization of attention patterns (Part 1). For each group of three images: orig shows the original input image, cls displays the class token attention heat map before applying our method, and pmt shows the prompt attention heat map after applying our Local Spatial Prompting mechanism.

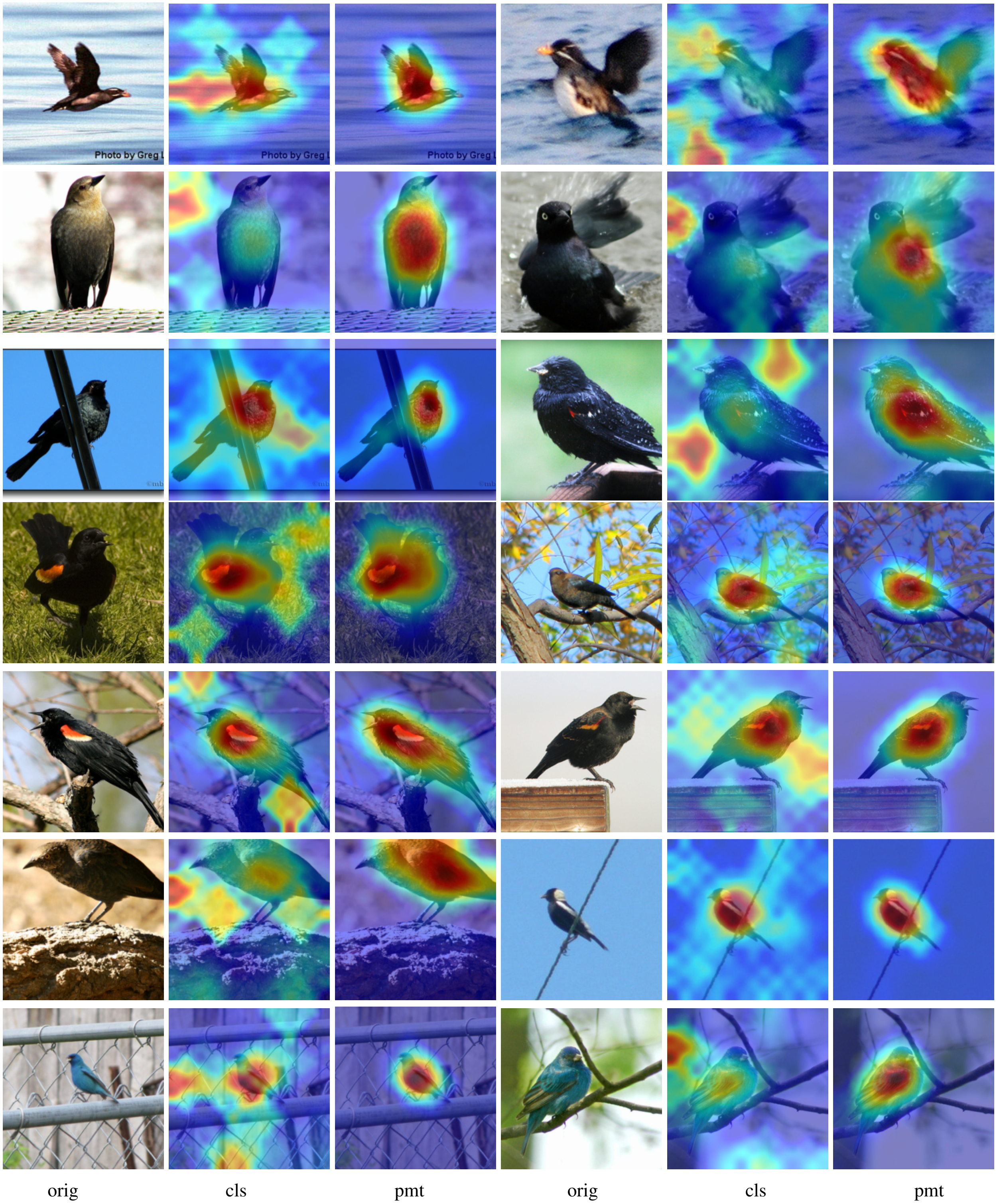

Figure 9. Extended visualization of attention patterns (Part 2). The visualization demonstrates consistent improvement in attention focus across diverse samples, with the model learning to attend to discriminative features rather than background elements.

References

[1] Touqeer Ahmad, Akshay Raj Dhamija, Steve Cruz, Ryan Rabinowitz, Chunchun Li, Mohsen Jafarzadeh, and Terrance E. Boult. Few-shot class incremental learning leveraging self-supervised features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3900–3910, 2022. 1

[2] Afra Feyza Akyürek, Ekin Akyürek, Derry Tanti Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning, 2022. 1

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[4] Han Cheng, Wang Qifan, Cui Yiming, Cao Zhiwen, Wang Wenguan, Qi Siyuan, and Liu Dongfang. E2vpt: An effective and efficient approach for visual prompt tuning. In *International Conference on Computer Vision (ICCV)*, 2023. 1

[5] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14146–14155, 2022. 2

[6] Yawen Cui, Wanxia Deng, Haoyu Chen, and Li Liu. Uncertainty-aware distillation for semi-supervised few-shot class-incremental learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[7] Thang Doan, Sima Behpour, Xin Li, Wenbin He, Liang Gou, and Liu Ren. A streamlined approach to multimodal few-shot class incremental learning for fine-grained datasets, 2024. 4

[8] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification, 2023. 1

[9] Marco D'Alessandro, Alberto Alonso, Enrique Calabrés, and Mikel Galar. Multimodal parameter-efficient few-shot class incremental learning. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, page 3385–3395. IEEE, 2023. 2

[10] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N. Metaxas. Visual prompt tuning for test-time domain adaptation, 2022. 1

[11] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[12] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018. 3

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 1

[14] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting, 2023. 1

[15] Zitong Huang, Ze Chen, Zhixing Chen, Erjin Zhou, Xinxing Xu, Rick Siow Mong Goh, Yong Liu, Wangmeng Zuo, and Chunmei Feng. Learning prompt with distribution-based feature replay for few-shot class-incremental learning, 2024. 2

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022. 1, 4

[17] 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, 2023. 1

[18] 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. 2

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 2

[20] Anna Kukleva, Hilde Kuehne, and Bernt Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *ICCV*, 2021. 1

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021. 1

[22] Chenxi Liu, Zhenyi Wang, Tianyi Xiong, Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. Few-shot class incremental learning with attention-aware self-adaptive prompt, 2024. 2, 4

[23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. 1

[24] Zichen Liu, Yuxin Peng, and Jiahuan Zhou. InsVP: Efficient instance visual prompting from image itself. In *ACM Multimedia 2024*, 2024. 1

[25] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. 3

[26] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning, 2021. 2

[27] Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners, 2024. 2, 4

[28] Can Peng, Kun Zhao, Tianren Wang, Meng Li, and Brian C. Lovell. Few-shot class-incremental learning from an open-set perspective, 2022. 2

[29] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning, 2021. 1

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[32] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. Self-supervised knowledge distillation for few-shot learning. *https://arxiv.org/abs/2006.09785*, 2020. 1

[33] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning, 2017. 2

[34] Abhilash Shankarampeta and Koichiro Yamauchi. Few-shot class incremental learning with generative feature replay. In *International Conference on Pattern Recognition Applications and Methods*, 2021. 2

[35] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning, 2023. 1

[36] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017. 4

[37] 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. 2

[38] Hongbo Sun, Jiahuan Zhou, Xiangteng He, Jinglin Xu, and Yuxin Peng. Finefmpl: fine-grained feature mining prompt learning for few-shot class incremental learning. In *IJCAI*, 2024. 2, 4

[39] 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*, 2024. 2, 4

[40] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[41] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning, 2017. 2

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 2

[43] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning, 2022. 2

[44] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured suboptimality, 2023. 1

[45] 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, 2023. 2

[46] Xuan Wang, Zhong Ji, Xiyao Liu, Yanwei Pang, and Jungong Han. On the approximation risk of few-shot class-incremental learning. In *European Conference on Computer Vision*, 2024. 2, 4

[47] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning, 2023. 1

[48] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning, 2022. 1

[49] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning, 2022. 1, 7

[50] Boyu Yang, Mingbao Lin, Binghao Liu, Mengying Fu, Chang Liu, Rongrong Ji, and Qixiang Ye. Learnable expansion-and-compression network for few-shot class-incremental learning, 2021. 1

[51] Boyu Yang, Mingbao Lin, Yunxiao Zhang, Binghao Liu, Xiaodan Liang, Rongrong Ji, and Qixiang Ye. Dynamic support network for few-shot class incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2945–2951, 2023. 1

[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. In *ICLR*, 2023. 2

[53] In-Ug Yoon, Tae-Min Choi, Sun-Kyung Lee, Young-Min Kim, and Jong-Hwan Kim. Image-object-specific prompt learning for few-shot class-incremental learning, 2023. 2

[54] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 4

[55] Tianhao Zhang, Zhixiang Chen, and Lyudmila S. Mihaylova. Pvit: Prior-augmented vision transformer for out-of-distribution detection, 2025. 1

[56] Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, and Xiangzhong Fang. Few-shot class-incremental learning via class-aware bilateral distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11838–11847, 2023. 2

[57] 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 *CVPR*, 2022. 2

[58] 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, 2023. 2

[59] Haichen Zhou, Yixiong Zou, Ruixuan Li, Yuhua Li, and Kui Xiao. Delve into base-novel confusion: Redundancy exploration for few-shot class-incremental learning. *arXiv preprint arXiv:2405.04918*, 2024. 2

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1

[61] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning, 2021. 2

[62] Mingli Zhu, Zihao Zhu, Sihong Chen, Chen Chen, and Baoyuan Wu. Enhanced few-shot class-incremental learning via ensemble models, 2024. 1

[63] 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. 2

[64] Yixiong Zou, Shanghang Zhang, Haichen Zhou, Yuhua Li, and Ruixuan Li. Compositional few-shot class-incremental learning. *arXiv preprint arXiv:2405.17022*, 2024. 2