Pre-trained Vision and Language Transformers Are Few-Shot Incremental Learners

Keon-Hee Park[1] Kyungwoo Song[2†] Gyeong-Moon Park[1†]

[1]Kyung Hee University, Republic of Korea
[2]Yonsei University, Republic of Korea

{pgh2874, gmpark}@khu.ac.kr kyungwoo.song@yonsei.ac.kr

Abstract

Few-Shot Class Incremental Learning (FSCIL) is a task that requires a model to learn new classes incrementally without forgetting when only a few samples for each class are given. FSCIL encounters two significant challenges: catastrophic forgetting and overfitting, and these challenges have driven prior studies to primarily rely on shallow models, such as ResNet-18. Even though their limited capacity can mitigate both forgetting and overfitting issues, it leads to inadequate knowledge transfer during few-shot incremental sessions. In this paper, we argue that large models such as vision and language transformers pre-trained on large datasets can be excellent few-shot incremental learners. To this end, we propose a novel FSCIL framework called ***PriViLege****,* ***Pr****e-trai****n****ed* ***Vi****sion and* ***L****anguage transformers with prompting functions and knowl****edge*** *distillation. Our framework effectively addresses the challenges of catastrophic forgetting and overfitting in large models through new pre-trained knowledge tuning (PKT) and two losses: entropy-based divergence loss and semantic knowledge distillation loss. Experimental results show that the proposed PriViLege significantly outperforms the existing state-of-the-art methods with a large margin, e.g.,* $+9.38\%$ *in CUB200,* $+20.58\%$ *in CIFAR-100, and* $+13.36\%$ *in miniImageNet. Our implementation code is available at* https://github.com/KHU-AGI/PriViLege.

1. Introduction

We humans have an exceptional ability to quickly comprehend novel concepts from only a small amount of data. To grant this ability for deep neural networks, Few-Shot Class Incremental Learning (FSCIL), introduced in [33] first, imitates a way of learning that closely resembles that of human learning. FSCIL typically comprises a base session and incremental sessions. During the base session, a network learns numerous classes with sufficient training data, while in the incremental sessions, it trains novel classes with few-shot training data per each class. Given the restricted amount of data in incremental sessions, an effective transfer of diverse knowledge learned in the base session is crucial in FSCIL.

In FSCIL, there are two significant challenges: catastrophic forgetting [20] and overfitting [33]. Catastrophic forgetting occurs while the network learns new classes sequentially, *i.e.*, the network severely forgets the previously learned knowledge. On the other hand, overfitting arises when the network overly focuses on a limited set of training data, resulting in a degradation of overall performance. To address these challenges, previous studies mainly have utilized shallow models like Resnet-18 [12, 39, 40]. The advantage of adopting a shallow model lies in its limited number of learnable parameters, making it effective for mitigating forgetting through partial freezing and curbing overfitting. However, the limited capacity of the shallow model hinders capturing and transferring sufficient domain knowledge from the base session to the incremental sessions.

Recently, large pre-trained models like Vision Transformer (ViT) [3] and Contrastive Language-Image Pre-training (CLIP) [26] are widely used in computer vision due to their promising adaptability and performance. In that sense, large pre-trained models can effectively learn and transfer domain knowledge from the base session, overcoming limitations in transferability associated with shallow models. However, finetuning large pre-trained models is prone to forget the useful pre-trained knowledge, while freezing the models hinders the acquisition of domain-specific knowledge during the base session. This inherent trade-off between preserving the pre-trained knowledge and acquiring new domain-specific knowledge hinders their use in FSCIL.

To investigate the challenges and applicability of large models in FSCIL, we conducted 5-way 5-shot experiments

†Corresponding authors

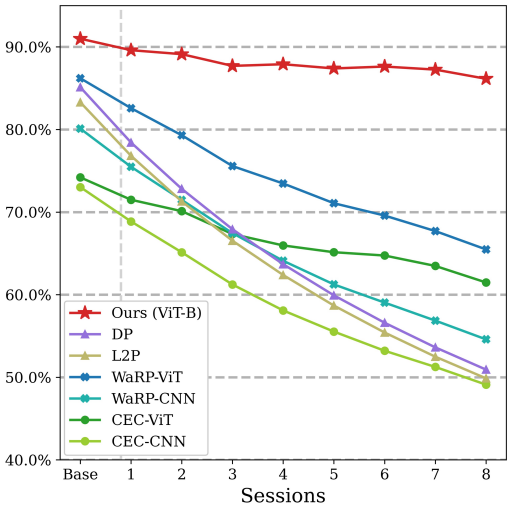


Figure 1. 5-way 5-shot FSCIL experiments on CIFAR-100.

on CIFAR-100 [14], using a Vision Transformer base (ViT-B) pre-trained on ImageNet-21K [28]. We applied this pre-trained model to existing FSCIL methods [12, 40]. As shown in Figure 1, we find that directly applying the ViT backbone to existing methods is ineffective in FSCIL. In detail, selectively freezing parameters (WaRP [12]) leads to severe forgetting during incremental sessions. On the other hand, freezing the entire network (CEC [40]) somewhat alleviates forgetting during the incremental sessions, but struggles to capture the useful knowledge in all the sessions. To sum up, existing FSCIL methods based on the large pre-trained model still show large performance drop due to catastrophic forgetting and the loss of transferability.

Recently, the methods utilizing prompt tuning [16] (L2P [37]) or prefix tuning [17] (Dualprompt [36]) based on the pre-trained ViT show promising performances in Class Incremental Learning (CIL). To further clarify the applicability of these methods in FSCIL, we conducted the experiments under the same setup in Figure 1. We observed that despite of effectively utilizing the large pre-trained model through prompting functions, these methods exhibit inferior performances compared to existing FSCIL methods. We attribute this to the limited number of learnable parameters in the prompt , which hinders effective knowledge transfer to incremental sessions. In other words, the sole utilization of prompting functions to the pre-trained ViT is inadequate for transferring the sufficient domain knowledge in FSCIL.

In this paper, we argue that *large models such as vision and language transformers pre-trained on large datasets can be excellent few-shot incremental learners*. To this end, we propose a new FSCIL framework based on **Pr**e-tra**i**ned **Vi**sion and **L**anguage transformers with prompting functions and knowl**e**d**ge** distillation, called **PriViLege**. Our framework includes newly proposed Pre-trained Knowledge Tuning (PKT), which is a simple yet effective approach to preserve the pre-trained knowledge of large models while learning the domain-specific knowledge effectively. Specifically, our PKT selectively trains specific layers with a new prompt modulation approach to prevent severe forgetting and enhance the knowledge acquisition of prompt. To strengthen the discriminative representation learning during the base session, we introduce a novel entropy-based divergence loss. In addition, we propose a new knowledge distillation, utilizing the pre-trained language model (PLM) to transfer semantic knowledge from the language space to the visual space. Through extensive experiments, we demonstrate that our framework enables the pre-trained large models to effectively serve as few-shot incremental learners with significant improvement.

Our main contributions can be summarized as follows:

- To address the challenges of adopting large pre-trained models in FSCIL, we propose a novel framework **PriViLege**, **Pr**e-tra**i**ned **Vi**sion and **L**anguage transformers with prompting functions and knowl**e**d**ge** distillation.
- We propose a new pre-trained knowledge tuning (PKT), which is designed to obtain domain knowledge effectively during the base session while preserving the useful pre-trained knowledge.
- To enhance the discriminative power during the base session and transfer the knowledge into the incremental sessions, we propose a new entropy-based divergence loss.
- We propose a semantic knowledge distillation loss to enhance representation learning by distilling semantic knowledge captured from the pre-trained language model.
- Comprehensive experiments show that our framework achieves overwhelming performance gains in the FSCIL benchmarks compared to state-of-the-art models.

2. Related Work

Few-Shot Class Incremental Learning. Few-shot class incremental learning (FSCIL) is a sort of class incremental learning but is more challenging since a model is able to learn novel classes with only few training samples. Among many previous approaches for FSCIL [34], dynamic network structure-based methods [29, 33, 38, 40] adjust the network structures themselves during training, while preserving the severe forgetting of the previous knowledge. Feature and feature space-based methods [2, 12, 23, 30, 41, 42] enable the model to adapt to new classes better and improve the generalization ability of feature extractors for new classes. Prototype-based methods [6, 19, 39] aim to align prototypes with classifier weights to enhance the classification performance. However, prior methods primarily address forgetting and overfitting in shallow networks, leading to marginal performance improvements given the limited model capacity. In this paper, we adopt large models such as the pre-trained ViT and CLIP for FSCIL and in-

troduce how to utilize them effectively to overcome major challenges of FSCIL.

Prompt Engineering for Vision Transformer. Prompt tuning [16] and prefix tuning [17] are widely used for prompt engineering in vision transformer. Prompt tuning adds learnable prompts to the input sequence, while prefix tuning directly influences attention patterns by appending prompts to the key and value for task-specific knowledge acquisition. Vision transformers utilizing prompt engineering show remarkable performances in class incremental learning. L2P [37] and Dualprompt [36] utilize prompt and prefix tunings for each to learn new classes while freezing the pre-trained ViT. L2P and Dualprompt leverages randomly initialized prompts for prompt engineering. Some recent approaches [9, 31, 32] have suggested methods for generating prompts to adapt the domain space for effective continual learning. CODA-Prompt [31] requires a collection of prompt components that are combined with input-dependent weights to generate input-specific prompts. APG [32] and DAP [9] utilize prompt generators comprising multiple components, including cross-attention layers, groups of learnable parameters, and linear layers. These prompt-generating methods demand extra components and training costs for the prompt generator. Unlike existing prompt-generating methods that necessitate additional learnable components for prompt generation, we propose a lightweight modulation approach for prompt. This approach significantly reduces the requirement of extra trainable components, and at the same time, it enhances the representation learning of prompt.

Semantic Guidance from Language Models. The utilization of language embeddings for effective learning of novel classes has been extensively explored in the context of generalized zero-shot learning [1, 25]. Recently, a trial by [11] investigates the use of language guidance for enhancing representation learning in continual learning. In the field of FSCIL, various approaches have incorporated additional language information, particularly class names, to improve representation learning in the base session. For instance, [42] addresses the drift of classifier weights by calibrating the encoded language information between existing and new classes. Furthermore, [2] proposes a regularization method leveraging the relational information derived from class word embeddings extracted from the GloVe network [24].

3. Method

Formulation of Few-Shot Class Incremental Learning. In FSCIL, the training dataset $\mathcal{D} = \{\mathcal{D}^0, \mathcal{D}^1, ..., \mathcal{D}^T\}$ are sequentially given, where $\mathcal{D}^0$ is the dataset for the base session, $\mathcal{D}^t = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}^t|}$ is the dataset for the t-th incremental session, $1 \leq t \leq T$, and T is total number of incremental sessions, respectively. Here, $\mathcal{D}^0$ for the base session generally has a large label space $\mathcal{C}^0$ and enough training data for each class $c \in \mathcal{C}^0$. On the other hand, $\mathcal{D}^t$ for the incremental session has only few training samples per each class, *i.e.*, $|\mathcal{D}^t| = k \cdot |\mathcal{C}^t|$, where $|\mathcal{C}^t|$ is total number of novel classes for the t-th task and k is the number of samples per novel class. There is no class overlap between sessions, and at each session, the model can access the current dataset only. In this setting, the goal of FSCIL is to enable the model to incrementally learn new classes from a few samples, while preserving the classification ability of all previously encountered classes.

Method Overview. Figure 2 provides an overview of the proposed method, where the pre-trained Vision Transformer (ViT) serves as the backbone network. To refine the pre-trained knowledge of ViT, we introduce a novel Pre-trained Knowledge Tuning (PKT) (Section 3.1). PKT involves training a base prompt (B-Prompt), a vision-language prompt (VL-Prompt), and selected layers of ViT, thereby enhancing the transferable knowledge for incremental sessions. Additionally, to strengthen the discriminative ability during the base session, we propose an entropy-based divergence loss (Section 3.2) for a vision token in VL-Prompt. Finally, we introduce a semantic knowledge distillation loss (Section 3.3) to transfer the semantic knowledge into a language token in VL-Prompt, improving the representation learning of our model. For stable learning in the few-shot environment, we utilize prototypes of each class as a classifier.

3.1. Pre-trained Knowledge Tuning

Recently, the large pre-trained ViT based on prompting functions [21, 36, 37] shows remarkable performance in class incremental learning. However, the effective integration of large pre-trained ViT into FSCIL remains unexplored. Existing methods adopting the pre-trained ViT struggle with issues such as catastrophic forgetting and overfitting. Furthermore, previous prompt-based methods face challenges in transferring sufficient knowledge to the incremental sessions due to the limited prompt size.

In this paper, we explore how to adapt the powerful pre-trained ViT to the FSCIL task effectively. To this end, we introduce a novel approach termed Pre-trained Knowledge Tuning (PKT), which selectively fine-tunes specific layers using additional prompts to acquire domain-specific knowledge during the base session. In PKT, we specifically update the initial L layers of the pre-trained ViT f_θ, where L is a hyperparameter representing the number of layers to be updated. Through empirical analysis, we determined the optimal number of layers to be updated, *e.g.*, the first two layers ($L = 2$) in ViT-B. Since we freeze most layers in ViT, the pre-trained knowledge remains helping in-

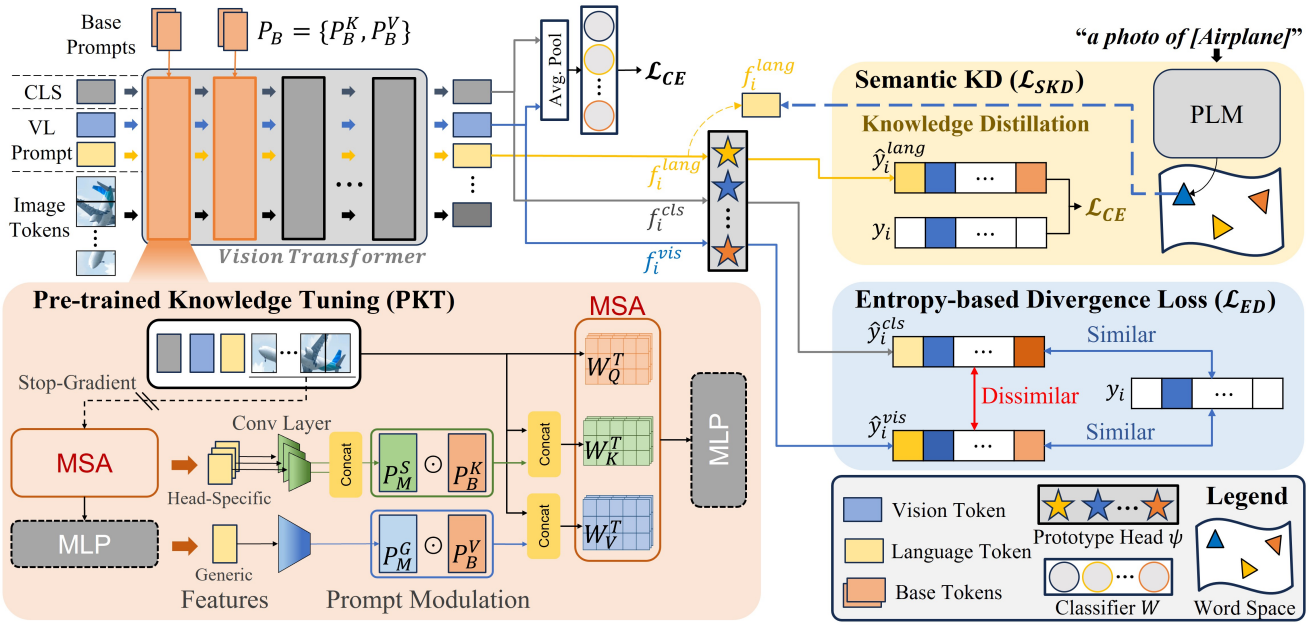


Figure 2. Overall framework of our method. In the base session, the newly proposed PKT trains the B-Prompt, VL-Prompt, and selected layers in the pre-trained ViT. $\mathcal{L}_{ED}$ drives the vision token in VL-Prompt to enhance discriminative ability for better classification. $\mathcal{L}_{SKD}$ leverages language embeddings to provide semantic knowledge to the language token in VL-Prompt.

cremental learning without forgetting. The updated ViT is frozen after the base session to preserve domain-specific knowledge and transfers the learned knowledge to incremental sessions. Since we freeze most of the layers, effective learning of domain-specific knowledge becomes a challenge. To address this limitation, we introduce two key prompts: the base prompt (B-Prompt) denoted as $\boldsymbol{P_B} \in \mathbb{R}^{L\times 2\times D}$, where D represents the embedding dimension, and the vision-language Prompt (VL-Prompt) denoted as $\boldsymbol{P_{VL}} \in \mathbb{R}^{2\times D}$. B-Prompt is tailored to capture domain-specific knowledge while selectively fine-tuning some layers at the base session. B-Prompt facilitates the transfer of domain-specific knowledge to incremental sessions. Meanwhile, VL-Prompt, consisting of both vision and language tokens, is designed to transfer positive knowledge from all previous sessions to the next. We train B-Prompt and VL-Prompt using the prefix tuning and the prompt tuning, respectively. Through the utilization of both prefix tuning and prompt tuning, we are able to tailor the training of B-Prompt and VL-Prompt to their respective purposes.

However, since prefix tuning has a limitation of updating B-Prompt as mentioned in [4, 5] which is the slow adaptation speed of learnable prompts, B-Prompt cannot properly learn domain-specific knowledge and can be ignored by the fine-tuned layers. To overcome this, we propose new modulation prompts P_M to assist B-Prompt. Modulation prompts contain a head-specific prompt P_M^S and a generic prompt P_M^G, which are obtained from the multi-head self-attention (MSA) layer and the followed MLP layer of the pre-trained ViT, respectively. The formulation of modulation prompts is as follows:

$$h^{MSA} = \mathrm{MSA}(h_Q, h_K, h_V), \tag{1}$$

$$P_M^S = [g_1^S(h_1^{MSA}); \ldots; g_H^S(h_H^{MSA})], \tag{2}$$

$$h^{MLP} = \mathrm{MLP}(h^{MSA}), \tag{3}$$

$$P_M^G = g^G(h^{MLP}), \tag{4}$$

where H denotes the number of heads in the MSA layer, h^{MSA} and h^{MLP} denote the outputs for each layers, $g^S = \{g_1^S, \cdots, g_H^S\}$ and g^G denote point-wise convolution layers for P_M^S and P_M^G, respectively. We extract the feature vectors from the ViT layers and generate modulation prompts through 1×1 convolution layers to align with the B-Prompt.

Through the pre-trained layers and input data, both the head-specific prompt and generic prompt enhance feature knowledge. These modulation prompts can scale the B-Prompt, enlarging its feature vector depending on the input data. The modulation prompts assist the B-Prompt in capturing domain-specific knowledge by enlarging feature vectors. The process of prefix tuning of PKT is as follows:

$$\bar{P}_K^{'} = P_M^S \odot P_B^K, \tag{5}$$

$$\bar{P}_V^{'} = P_M^G \odot P_B^V, \tag{6}$$

$$\bar{h}^{out} = \mathrm{MSA}([P_{VL}^Q; h_Q], [\bar{P}_K^{'}; h_K], [\bar{P}_V^{'}; h_V]), \tag{7}$$

where $\bar{h}^{out}$ denotes the output of the PKT, and h_Q,h_K, and h_V denote the input query, key, and value, respectively.

In summary, our PKT provides two main advantages: 1) it effectively learns base session knowledge by introducing

additional plasticity in the first L layers and incorporating extra prompts, and 2) by scaling the B-Prompt through the modulation prompts, PKT promotes the update of the B-Prompt. This boosts the B-Prompt to learn useful domain-specific knowledge along with the pre-trained ViT. Empirical results demonstrate that PKT significantly enhances performance, facilitating the positive knowledge transfer in incremental sessions.

3.2. Entropy-based Divergence Loss

During the training of prompts and selected layers using PKT, our model effectively acquires the domain knowledge, for its positive transfer during incremental sessions. To embed multi-perspective knowledge, we involve the vision token along with the [CLS] token for the classification through average pooling. However, since these two tokens share the same objective, their output features become similar as training progresses, which hinders the effective training of the vision token. To strengthen the discriminative power of the vision token itself, we propose a new regularization term, called an entropy-based divergence loss ($\mathcal{L}_{ED}$).

To calculate $\mathcal{L}_{ED}$, we first construct a prototype classifier $\psi \in \mathbb{R}^{|C^0| \times D}$ that consists of prototypes for the base session classes. For each base class c_j, where $c_j \in C^0$, and $j \in \{1, 2, ..., |C^0|\}$, the prototype $proto_{c_j} \in \mathbb{R}^D$ is the average vector of all the output features extracted by the [CLS] token passing through the pre-trained ViT. Therefore, the prototype classifier ψ can be represented as follows:

$$proto_{c_j} = \frac{1}{N_{c_j}} \sum_{k=1}^{N_{c_j}} f_k^{cls}, \tag{8}$$

$$\psi = [proto_{c_1}; proto_{c_2}; ...; proto_{c_{|C^0|}}], \tag{9}$$

where N_{c_j} is the number of training samples for the class c_j. Using the prototype classifier ψ, we then calculate the logits $\hat{y}_i^{cls} = \psi(f_i^{cls})$ and $\hat{y}_i^{vis} = \psi(f_i^{vis})$ corresponding to the [CLS] and vision tokens, respectively. At this time, the prototype classifier ψ is not trainable to serve the stable basis to calculate the loss function. Finally, we use $\hat{y}_i^{cls}$ and $\hat{y}_i^{vis}$ with the label y_i to define the entropy-based divergence loss $\mathcal{L}_{ED}$ as follows:

$$\mathcal{L}_{ED} = \log\left(\frac{\mathcal{L}_{CE}(\hat{y_i}^{vis}, y_i) + \mathcal{L}_{CE}(\hat{y_i}^{cls}, y_i)}{\mathcal{L}_{KL}(\delta(\hat{y}_i^{vis}), \delta(\hat{y}_i^{cls}))} + 1\right), \tag{10}$$

where $\mathcal{L}_{CE}$ is the cross entropy loss, $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence loss [15], and $\delta(\cdot)$ is a softmax function. To minimize $\mathcal{L}_{ED}$, our model learns to minimize both $\mathcal{L}_{CE}(\hat{y_i}^{vis}, y_i)$ and $\mathcal{L}_{CE}(\hat{y_i}^{cls}, y_i)$, and maximize the $\mathcal{L}_{KL}(\delta(\hat{y}_i^{vis}), \delta(\hat{y}_i^{cls}))$. In other words, the proposed entropy-based divergence loss guides the vision token to gain the discriminative knowledge itself, while separating the embedded knowledge of the vision token from the one of the [CLS] token. Through $\mathcal{L}_{ED}$, our model can capture the domain-specific knowledge effectively using the vision token at the base session and provide this transferable knowledge for the incremental sessions.

3.3. Semantic Knowledge Distillation Loss

Even though the transferred knowledge from the base session to the incremental sessions is abundant and useful, it is still challenging to learn the exact representations for novel classes from few-shot training samples. To alleviate this issue, it is required to provide external knowledge related to the novel classes for better adaptation. To this end, we introduce a new semantic knowledge distillation loss ($\mathcal{L}_{SKD}$) to provide additional semantic knowledge by using the pre-trained language model (PLM), e.g., BERT [10]. Through the language embeddings from PLM utilizing the class names given as labels, we can provide useful semantic knowledge to our proposed model.

To achieve the goal of semantic knowledge distillation, we first get the language embedding feature $w_{cn_i} = f_\varphi(word_{cn_i})$, where f_φ is PLM and $word_{cn_i}$ is the class name prompted as "*a photo of* $[cn_i]$", corresponding to the class cn_i. Meanwhile, we also acquire the output feature $f_i^{lang} = f_\theta(x_i)[2]$ corresponding to the language token in VL-Prompt using the ViT backbone. To distill the semantic knowledge from the language embedding feature w_{cn_i} to the language token feature f_i^{lang}, we adopt the knowledge distillation loss ($\mathcal{L}_{KD}$) from [7]. However, these two features come from totally different embedding spaces, *i.e.*, visual feature space, and language embedding space, respectively, solely applying the distillation loss to match two different distributions may be ineffective.

To overcome this issue, we utilize the prototype classifier ψ once again as a stable basis to regulate the output feature of the language token. Specifically, we input the language token feature f_i^{lang} into the prototype classifier ψ, denoted as $\hat{y}_i^{lang} = \psi(f_i^{lang})$, to compute the cross-entropy loss. We utilize the cross-entropy loss ($\mathcal{L}_{CE}$) to minimize the distribution difference between the visual feature space and the language embedding space. We then define semantic knowledge distillation loss $\mathcal{L}_{SKD}$ by adding two losses as follows:

$$\mathcal{L}_{SKD} = \mathcal{L}_{KD}(f_i^{lang}, w_{cn_i}) + \gamma \cdot \mathcal{L}_{CE}(\hat{y}_i^{lang}, y_i), \tag{11}$$

where γ is the balancing hyperparameter for $\mathcal{L}_{CE}$. The second term in $\mathcal{L}_{SKD}$ prevents f_i^{lang} from diverging to the undesirable feature representation using the true label y_i. We used $\gamma = 0.1$ for all of our experiments.

To sum up, the proposed semantic knowledge distillation loss $\mathcal{L}_{SKD}$ enables our model to distill the useful semantic knowledge from the language embedding space into the visual feature space to provide additional information

Dataset	CUB200			CIFAR-100			miniImageNet		
Method	A_{Base}	A_{Last}	A_{Avg}	A_{Base}	A_{Last}	A_{Avg}	A_{Base}	A_{Last}	A_{Avg}
Fine-Tuning + Proto ψ	**84.21±0.13**	3.79±1.47	21.60±1.32	**91.36±0.15**	5.19±0.13	37.04±1.06	93.67±0.02	9.87±5.42	44.60±0.92
CEC [CVPR'21]	75.40±8.01	65.70±8.03	72.41±1.18	74.20±2.03	61.48±3.33	67.10±2.92	87.43±5.90	80.74±7.51	83.06±7.14
L2P [CVPR'22]	44.97±2.32	15.41±3.45	24.99±4.30	83.29±0.50	49.87±0.31	64.08±0.39	94.59±0.21	56.84±0.32	72.97±0.36
DualPrompt [ECCV'22]	53.37±1.83	23.25±2.02	36.30±2.39	85.11±0.29	50.93±0.21	65.45±0.27	95.05±0.20	57.14±0.11	73.31±0.15
NC-FSCIL [ICLR'23]	78.49±2.32	38.80±1.14	57.92±1.71	89.51±0.23	53.70±0.14	68.96±0.17	77.25±0.42	46.35±0.25	59.52±0.33
WaRP [ICLR'23]	67.74±5.57	49.36±6.56	55.85±6.06	86.20±1.46	65.48±1.87	74.55±1.67	83.30±1.06	67.97±1.28	74.13±1.08
PriViLege (Ours)	82.21±0.35	**75.08±0.52**	**77.50±0.33**	90.88±0.20	**86.06±0.32**	**88.08±0.20**	**96.68±0.06**	**94.10±0.13**	**95.27±0.11**

Table 1. Comparison of the performance on CUB200, CIFAR-100, and miniImageNet. CUB200 has a 10-way 5-shot incremental setup, and CIFAR-100 and miniImageNet have a 5-way 5-shot incremental setup. We report the best as **bold** and the second-best as underlined.

during the few-shot incremental sessions. It is beneficial to mitigate the challenge of representation learning from the few-shot data. Moreover, $\mathcal{L}_{SKD}$ drives the network to learn abundant base knowledge using enough classes in the base session, which leads to positive knowledge transfer for the incremental sessions.

The total loss for the base session ($\mathcal{L}_{base}$) can be summerized as follows:

$$\mathcal{L}_{base} = \mathcal{L}_{CE}(\hat{y}_i, y_i) + \alpha \cdot \mathcal{L}_{ED} + \beta \cdot \mathcal{L}_{SKD}, \quad (12)$$

where α and β are the scaling factors for entropy-based divergence loss and semantic knowledge distillation loss. In the incremental sessions, we do not use the entropy-based divergence loss since a few samples are not enough to learn discriminative features. Therefore, the total loss for the incremental sessions (L_{inc}) can be expressed as follows:

$$\mathcal{L}_{inc} = \mathcal{L}_{CE}(\hat{y}_i, y_i) + \beta \cdot \mathcal{L}_{SKD}. \quad (13)$$

4. Experiments

4.1. Experimental Settings

Datasets and Metrics. We evaluated our method with the SOTA FSCIL methods on three datasets: CIFAR-100 [14], miniImageNet [27], and CUB200 [35]. As shown in Table 2, we followed the same split configuration proposed by [33] in all datasets. We evaluated the performance by measuring the accuracy of the base session A_{Base}, last session A_{Last}, and the average accuracy of all the sessions A_{Avg}. We conducted 5 simulations under different random seeds and reported the averages.

Baselines and Implementation Details. We considered the following recent FSCIL methods as baselines: CEC [40], WaRP [12], and NC-FSCIL [39]. We also set L2P [37] and DualPrompt [36] as baselines to compare the methods using ViT. For the backbone network, we used a ViT-B/16 [3] pre-trained by ImageNet-21K [28] for all the methods including ours. We used BERT-base [10] to extract the word class embeddings. We set the first two layers as trainable for the pre-trained knowledge tuning. We set 0.5 for both α and β. We used Adam optimizer [13], cosine annealing scheduler [18], and the learning rate as 2e-4. We trained our method using an RTX 3090 GPU and set the batch size as 128. We trained 5 epochs for the base session and 3 epochs for the incremental sessions.

Session	CUB200	CIFAR-100	miniImageNet
Base	100	60	60
Incremental	10-way 5-shot	5-way 5-shot	5-way 5-shot
# of sessions	1+10	1+8	1+8

Table 2. Configuration settings for FSCIL benchmarks on CUB-200, CIFAR-100, and miniImageNet.

4.2. Main Experimental Results

We reported the base, last, and average accuracy of CUB200, CIFAR-100, and miniImageNet, respectively, in Table 1. As shown in Table 1, our method, PriViLege, surprisingly overwhelmed all the baselines with a large margin on all the datasets compared with SOTA methods in FSCIL. Our proposed methods using ViT-B/16 reported an about $+9.38\%$ performance enhancement in A_{Last} and about $+5.09\%$ in A_{Avg} against CEC on CUB200. Our method also showed outstanding performance in CIFAR-100 where our proposed methods reported about $+20.58\%$ in A_{Last} and about $+13.53\%$ in A_{Avg} against WaRP. Also, compared with prompt-based methods such as L2P and DualPrompt, our novel method, PriViLege, reported powerful performance enhancement. Our experiments consistently revealed notable enhancements in A_{Last} and A_{Base} across all datasets. The emphasis on effective domain knowledge learning and transferability enhancement through PriViLege, incorporating PKT, $\mathcal{L}_{ED}$, and $\mathcal{L}_{SKD}$, contributes to its outstanding performance.

It is noteworthy that our method, PriViLege, showed

Dataset	CUB 200		
Ablation	A_{Base}	A_{Last}	A_{Avg}
Baseline	**84.21±0.13**	3.79±1.47	21.60±1.32
PKT	79.06±0.77	70.81±0.76	73.36±0.77
PKT + $\mathcal{L}_{ED}$	80.31±0.54	72.70±0.45	75.04±0.40
PKT + $\mathcal{L}_{SKD}$	82.10±0.57	73.44±0.40	76.27±0.30
Ours	82.21±0.35	**75.08±0.52**	**77.50±0.33**

Table 3. Ablation experiment on CUB 200. The baseline denotes fine-tuning pre-trained ViT with prototype classifier ψ.

Dataset	CUB 200		
# of Layers	A_{Base}	A_{Last}	A_{Avg}
0 Layers	76.07±0.56	60.19±1.11	67.08±0.71
2 Layers	**79.06±0.77**	**70.81±0.76**	**73.36±0.77**
5 Layers	78.42±0.84	68.52±0.84	71.99±0.84
7 Layers	76.96±0.74	63.15±2.73	68.06±1.54
10 Layers	74.95±0.78	57.78±2.30	64.19±1.59
12 Layers	73.62±2.72	56.02±1.47	62.71±2.14

Table 4. Further study on the number of tuned layers on CUB200.

significant performance improvement on CUB200. Given the fewer samples per class in CUB200, learning sufficient knowledge in the base session and transferring it to incremental sessions becomes challenging. However, our proposed method demonstrated improved performance across all metrics on CUB200. This underscores that the ability of PriViLege to capture effective domain-specific knowledge and transfer the knowledge into incremental sessions helps the successful application of ViT in FSCIL.

4.3. Ablation Study

We conducted an ablation study on CUB200 to validate our method. We set fine-tuning pre-trained ViT with a prototype classifier ψ as the baseline. Table 3 illustrates the performance comparison of each component. PKT exhibited notable enhancements in A_{Last} and A_{Avg}, showcasing the effectiveness of our proposed tuning approach in transferring domain-specific knowledge to incremental sessions. Despite limited fine-tuning layers by freezing most layers, our PKT showed only a slight decline in A_{Base} compared to the baseline. This underscores that PKT can effectively capture domain-specific knowledge. Furthermore, our proposed losses, entropy-based divergence loss, and semantic knowledge distillation loss, demonstrated promising performance. Applying the entropy-based divergence loss resulted in a performance enhancement of approximately $+1.89\%$ in A_{Last} and $+1.68\%$ in A_{Avg}. Similarly, the semantic knowledge distillation loss recorded improvements of about $+2.63\%$ in A_{Last} and approximately $+2.91\%$ in A_{Avg} when compared to the sole application of PKT. In conclusion, our method, PriViLege, demonstrated outstanding performance with significant margins compared to the baselines.

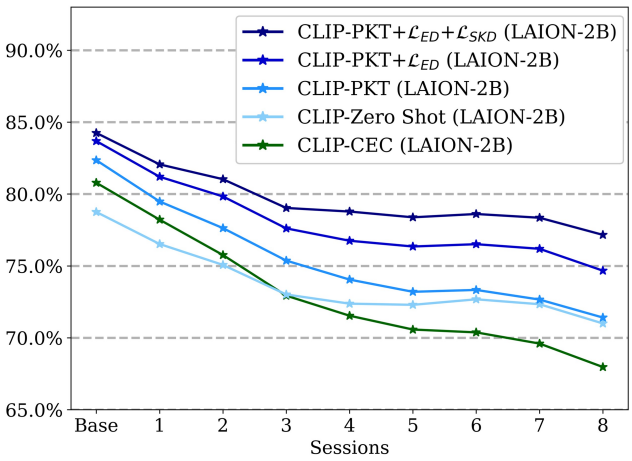


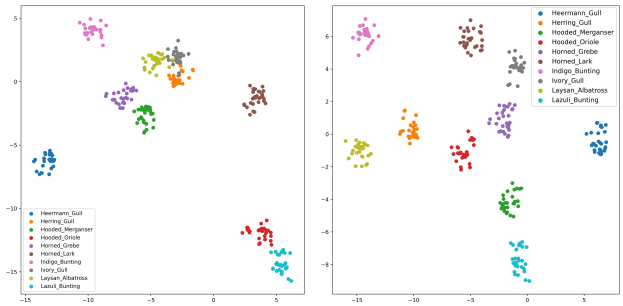
Figure 3. CLIP performance on CIFAR-100. We compare our proposed methods with zero-shot performance of CLIP and CEC.

4.4. Analysis

PriViLege on Pre-trained CLIP Network. To evaluate the adaptability of our proposed method, PriViLege, we compared our method with the zero-shot performance of CLIP [26]. To integrate PriViLege with CLIP, we exclusively trained the vision encoder of CLIP and employed a text encoder to extract language embedding features for the semantic knowledge distillation loss. Additionally, we compared the performance of the existing method, CEC, with the zero-shot performance in CLIP.

As shown in Figure 3, We noticed that while the existing method, CEC, exhibited inferior performance compared to zero-shot performance, our novel method demonstrated outstanding results in contrast to zero-shot performance. Specifically, we observed that the proposed entropy-based divergence loss and semantic knowledge distillation loss significantly contributed to notable performance enhancements. Given that CLIP is pre-trained with a contrastive approach involving both vision and language data, our proposed entropy-based divergence loss can contribute to improved representation knowledge by enhancing discriminative knowledge. Moreover, since CLIP already incorporates language embedding features, the semantic knowledge distillation loss effectively provides external knowledge for better adaptation. Our experiments revealed that our proposed method is applicable to CLIP and can show outstanding performance within this framework.

Layer Tuning Ablation for PKT. We studied further experiment on CUB200 to find out how many layers to be tuned for the proposed PKT. As shown in Table 4, fine-tuning the first 2 layers with additional prompts in PKT showed best performance across all metrics. This approach effectively captures domain-specific knowledge while preserving pre-trained knowledge. It is noteworthy that the absence of layer tuning resulted in inferior performance in A_{Base} and A_{Last} due to the limited capacity to



(a) w/o $\mathcal{L}_{ED}$ (b) w/ $\mathcal{L}_{ED}$

Figure 4. Feature space visualization on CUB200 to validate the efficacy of $\mathcal{L}_{ED}$.

learn domain-specific knowledge. Moreover, an increase in the number of learnable layers led to a decline in both A_{Base} and A_{Last}. It demonstrate that when the number of learnable layers is sufficient for capturing domain-specific knowledge, additional randomly initialized prompts may impede representation learning.

Visualization to Validate Entropy-based Divergence Loss. To evaluate the effectiveness of the entropy-based divergence loss ($\mathcal{L}_{ED}$), we conducted additional experiments to visualize the feature space. The purpose of entropy-based divergence loss is to enhance the discriminative ability of the pre-trained ViT. We compared the feature space with and without the application of entropy-based divergence loss. Figure 4a displays the result without $\mathcal{L}_{ED}$, while Figure 4b shows the result with $\mathcal{L}_{ED}$.

As shown in Figure 4a, we observed certain classes with unclear decision boundaries, such as "Herring Gull" and "Ivory Gull". However, the application of the proposed entropy-based divergence loss clarifies the classification of these similar classes, as illustrated in Figure 4b. This result leads us to conclude that our proposed entropy-based divergence loss enhances the discriminative ability between similar classes. The improved discriminative power of the pre-trained ViT can contribute to the effective capture of domain-specific knowledge in the base session and it helps incremental session learning.

Efficacy of Semantic Knowledge Distillation. We introduced semantic knowledge distillation loss ($\mathcal{L}_{SKD}$), utilizing language embeddings from a pre-trained language model to incorporate additional semantic knowledge. To evaluate the effectiveness of semantic knowledge distillation loss, we compared class-wise performance using a metric Δ_{diff}, which represents the performance difference between results with and without $\mathcal{L}_{SKD}$.

In Figure 5, we observed fluctuated performance with and without $\mathcal{L}_{SKD}$. For simplicity, we reported the top-10 classes with the most significant improvement and the bottom-10 classes with the most notable decline. As shown in Figure 5, the top-10 classes exhibited an enhanced performance of about 19.90%. This improvement is noteworthy, even when considering that the bottom-10 classes showed a decrease of about -6.76%. Particularly, It is noteworthy that the names of the top-10 classes include color-related terms (*e.g.*, red, green, yellow, and white) and shape descriptors (*e.g.*, headed, tailed, palm, billed, pelagic, and fish). This observation underscores the high effectiveness of our $\mathcal{L}_{SKD}$, especially when the class names incorporate characteristic information. This effectiveness is especially valuable in the fine-grained dataset such as a CUB200.

Figure 5. Class-wise accuracy on CUB200 to compare the performance of with and without $\mathcal{L}_{SKD}$.

5. Conclusion

In this study, we introduced a novel Few-Shot Class Incremental Learning (FSCIL) method using large pre-trained vision and language transformers, coined PriViLege. We addressed severe issues of catastrophic forgetting and overfitting in Vision Transformer (ViT) through Pre-trained Knowledge Tuning (PKT), entropy-based divergence loss, and semantic knowledge distillation loss. Our proposed method, PriViLege, achieved significant performance improvements across all benchmarks, and we demonstrated our framework can be applicable to different pre-trained models, including ViT and CLIP. We believe our PriViLege sheds light on a new research direction utilizing large models in FSCIL research. As future work, we will explore how to apply pre-trained large models effectively to a challenging FSCIL scenario without the base session.

Acknowledgements

This work was supported by MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2023-00258649) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and in part by the IITP grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068, and by the IITP grant funded by the Korea government (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)).

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2013. 3

[2] Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 6

[4] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. *International Conference on Computer Vision (ICCV)*, 2023. 4

[5] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022. 4

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

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[8] Zitong Huang et al. Learning prompt with distribution-based feature replay for few-shot class-incremental learning. *arXiv preprint*, 2024. 4

[9] Dahuin Jung, Dongyoon Han, Jihwan Bang, and Hwanjun Song. Generating instance-level prompts for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[10] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. 5, 6

[11] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Didier Stricker, Federico Tombari, and Muhammad Zeshan Afzal. Introducing language guidance in prompt-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[12] 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 (ICLR)*, 2023. 1, 2, 6

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 2, 6

[15] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 1951. 5

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 2, 3

[17] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, 2021. 2, 3

[18] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 6

[19] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[20] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. 1989. 1

[21] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[22] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020. 1

[23] Can Peng, Kun Zhao, Tianren Wang, Meng Li, and Brian C Lovell. 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*, 2022. 2

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014. 3

[25] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, 2022. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1, 7

[27] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations (ICLR)*, 2017. 6

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. 2, 6, 1

[29] Juwon Seo, Ji-Su Kang, and Gyeong-Moon Park. Lfs-gan: Lifelong few-shot image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[30] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[31] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[32] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. When prompt-based incremental learning does not meet strong pretraining. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[33] 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*, 2020. 1, 2, 6

[34] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *arXiv preprint arXiv:2304.08130*, 2023. 2

[35] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[36] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 6, 1

[37] 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. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6, 1

[38] 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 (CVPR)*, 2021. 2

[39] 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 *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 6

[40] 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 (CVPR)*, 2021. 1, 2, 6

[41] 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*, 2022. 2

[42] 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 (PAMI)*, 2022. 2, 3