FineFMPL: Fine-grained Feature Mining Prompt Learning for Few-Shot Class Incremental Learning

Hongbo Sun[1] , **Jiahuan Zhou**[1] , **Xiangteng He**[1] , **Jinglin Xu**[2] and **Yuxin Peng**[1]*
[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Intelligence Science and Technology, University of Science and Technology Beijing
{sunhongbo, jiahuanzhou, hexiangteng}@pku.edu.cn, xujinglinlove@gmail.com,
pengyuxin@pku.edu.cn

Abstract

Few-Shot Class Incremental Learning (FSCIL) aims to continually learn new classes with few training samples without forgetting already learned old classes. Existing FSCIL methods generally fix the backbone network in incremental sessions to achieve a balance between suppressing forgetting old classes and learning new classes. However, the fixed backbone network causes insufficient learning of new classes from a few samples. Benefiting from the powerful visual and textual understanding ability of Vision-Language (VL) pre-training models, we propose a Fine-grained Feature Mining Prompt Learning (FineFMPL) approach to adapt the VL model to FSCIL, which comprehensively learns and memorizes fine-grained discriminative information of emerging classes. Concretely, the visual probe prompt is firstly proposed to guide the image encoder of VL model to extract global-level coarse-grained features and object-level fine-grained features, and visual prototypes are preserved based on image patch significance, which contains the discriminative characteristics exclusive to the class. Secondly, the textual context prompt is constructed by cross-modal mapping of visual prototypes, feeding into the text encoder of VL model to memorize the class information as textual prototypes. Finally, integrating visual and textual prototypes based on fine-grained feature mining into the model improves the recognition performance of all classes in FSCIL. Extensive experiments on three benchmark datasets demonstrate that our FineFMPL achieves new state-of-the-art. The code is available at https://github.com/PKU-ICST-MIPL/FineFMPL_IJCAI2024.

1 Introduction

Recently, deep networks have achieved remarkable performance in numerous computer vision tasks owing to massive data and computational resources. In practice, their performance is severely limited when dealing with a continual data stream from unseen new classes [Ji *et al.*, 2023]. To handle this problem, Class Incremental Learning (CIL) is investigated to continually learn new classes with abundant labeled data while alleviating the catastrophic forgetting of old classes. However, the strict requirement of sufficient training samples of new classes is impractical in many scenarios when annotated data are hard to obtain [Tao *et al.*, 2020]. Therefore, the Few-Shot Class Incremental Learning (FSCIL) in Figure 1, i.e., the model is trained on abundant labeled samples in the base session and learns new classes from few labeled samples in incremental sessions without seeing old classes, has attracted more and more attention recently [Song *et al.*, 2023]. FSCIL encompasses the catastrophic forgetting problem of inaccessible old class data and the overfitting problem caused by scarce new class samples.

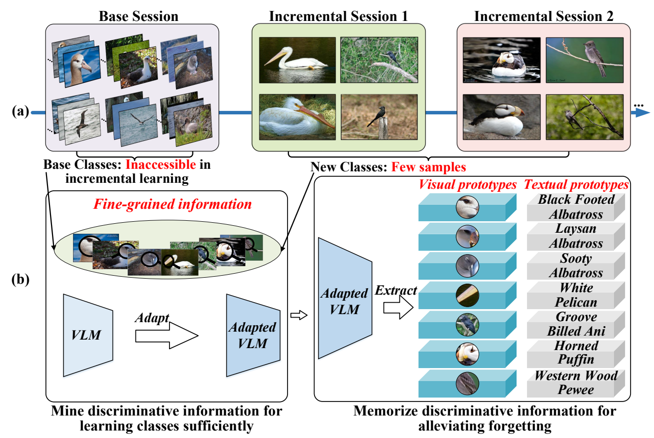

Figure 1: Illustrations of (a) Few-Shot Class Incremental Learning (FSCIL) and (b) the proposed FineFMPL in learning and memorizing fine-grained discriminative information of classes for FSCIL. VLM is short for the Vision-Language Model.

A popular FSCIL paradigm is to well-train the backbone network in the base session which is frozen for fine-tuning the classifiers to learn new classes via knowledge distillation [Dong *et al.*, 2021; Cheraghian *et al.*, 2021; Zhao *et al.*, 2023], attempting to achieve a balance between suppressing forgetting old classes and learning new classes. Though promising progress has been achieved, the above methods are still struggling due to the limited learning ability of the back-

*Corresponding author.

bone network on new data. As the feature extractor, the backbone network is trained on abundant data in base sessions and frozen in incremental sessions, which causes a lower learning ability for new classes compared with the base classes [Tao *et al.*, 2020; Zhang *et al.*, 2021]. In addition, the classifiers are fine-tuned by very few training samples from the entire images, which generally lack attention to the discriminant parts of distinguishing different classes.

Recently, the large-scale Vision-Language (VL) pre-training models, such as the well-known CLIP [Radford *et al.*, 2021], have shown powerful feature extraction ability, which can be adapted to various vision tasks by prompt learning [Zhou *et al.*, 2022d; Sun *et al.*, 2023b]. Inspired by the above observations and analyses, we propose the Fine-grained Feature Mining Prompt Learning (FineFMPL) of VL model to adapt it to the FSCIL task, which learns and memorizes fine-grained discriminative information of emerging classes to achieve promising performance, as shown in Figure 1. Concretely, a visual probe prompt is proposed to induce the image encoder of VL model to scale and gather discriminative image patch information from visual objects. Then, visual prototypes of classes are constructed based on the image patch significance analyses and weighted features aggregation to memorize the class from the visual side. Next, the textual context prompt is proposed conditioned on cross-modal mapping of visual prototypes, which contains implicit object attribute information. It is then input into the text encoder of VL model to generate textual prototypes to depict and memorize the class information from the textual side. Finally, classes' visual and textual prototypes are comprehensively utilized for the few-shot class incremental learning. To sum up, the main contributions can be summarized as follows:

- We propose a Fine-grained Feature Mining Prompt Learning (FineFMPL) method to guide the vision-language model to learn and memorize discriminative information of classes as visual and textual prototypes for few-shot class incremental learning.
- The visual probe prompt is proposed to scale and gather object-level information, and visual prototypes of emerging classes are preserved. Based on the cross-modal mapping of visual prototypes that contain implicit object attribute information, the textual context prompt is constructed to depict and memorize classes as textual prototypes.
- Extensive experiments are conducted on three widely used FSCIL benchmark datasets to demonstrate that our proposed FineFMPL approach surpasses existing state-of-the-art FSCIL methods significantly.

2 Related Work

In this section, we briefly review related works about class incremental learning, few-shot class incremental learning, and prompt learning.

2.1 Class Incremental Learning

In the Class Incremental Learning (CIL) task, the critical challenge is to learn new classes without forgetting the old knowledge. Existing CIL works can be roughly classified into three kinds to address this problem. The first kind of CIL works regularize the model's predictions [Hinton *et al.*, 2015; Li and Hoiem, 2017; Liu *et al.*, 2023b] between the old and current models, where knowledge distillation is commonly used. The second kind of CIL works [Castro *et al.*, 2018; Hou *et al.*, 2019] select a few representative samples of old classes, which are utilized for rehearsal in learning new classes. The third kind of CIL works [Abati *et al.*, 2020; Han *et al.*, 2023] attempt to expand the network for learning new classes. Some prompt learning-based CIL methods [Wang *et al.*, 2022; Smith *et al.*, 2023] recently have been proposed to achieve promising performance. However, it is noted that all the above CIL methods need abundant training samples of both the old and new classes, whose performance drops sharply when there are only few training samples of new classes in realistic scenarios. Therefore, it spurred the research of few-shot class incremental learning.

2.2 Few-shot Class Incremental Learning

Few-Shot Class Incremental Learning (FSCIL) continually recognizes new classes of a few training samples, which face both the catastrophic forgetting problem in CIL and the overfitting problem in few-shot learning [An *et al.*, 2023]. FSCIL was first proposed in [Tao *et al.*, 2020], which utilized a neural gas network to preserve topologies of classes. In the following works, CEC [Zhang *et al.*, 2021] proposed to utilize an independent classifier for each class, where the graph model conducted the information interaction between classifiers. F2M [Shi *et al.*, 2021] proposed finding flat minima to alleviate the catastrophic forgetting problem. Recently, SAVC [Song *et al.*, 2023] proposed introducing semantic knowledge by imaging virtual classes to help divide the classification space during training. CABD [Zhao *et al.*, 2023] proposed the class-aware bilateral distillation to transfer the knowledge from base classes to new classes for alleviating the forgetting and overfitting problem. However, existing FSCIL works generally fix the model's parameter in incremental sessions, which attempt to balance the stability for base classes and plasticity for new classes. Due to the data volume disparity, they are prone to base classes of abundant data, which limits the model's ability to recognize new classes.

2.3 Prompt Learning

Prompt learning is proposed to adapt the large-scale pre-trained models to downstream tasks with limited data, which was first proposed in natural language processing. For example, general knowledge is extracted from GPT [Radford *et al.*, 2019] and BERT [Devlin *et al.*, 2018] for various downstream language tasks with prompt designs, such as utilizing learnable vectors as prompts [Li and Liang, 2021]. Recently, prompt learning has been introduced into the computer vision area, which attempts to adapt the Vision-Language (VL) pre-training model to downstream vision tasks, such as image classification. As the pioneering work, Zhou et al. [Zhou *et al.*, 2022d] proposed CoOp to introduce learnable vectors into the text prompt as the context, which obtained more adaptive classification weights with the text encoder of CLIP. As an

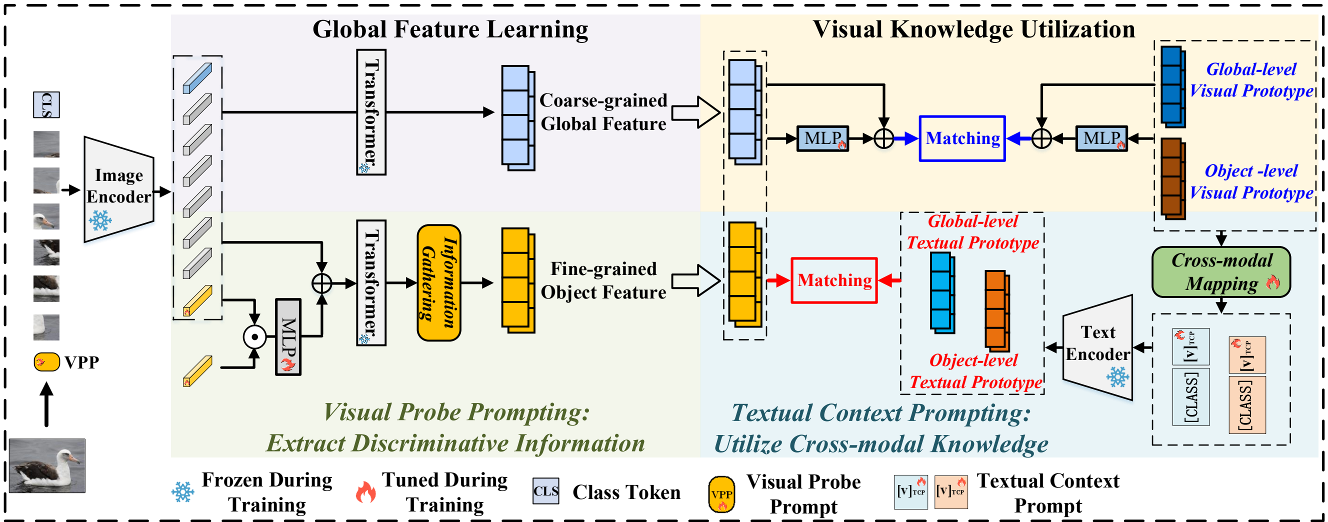

Figure 2: The framework of our FineFMPL model.

extension work of CoOp, Zhou et al. [Zhou *et al.*, 2022c] further proposed an input-conditional text prompt with a mapping neural network, which injected the visual information into the text prompt. Besides, some adapter methods have also been researched for adapting the VL model. Zhang et al. [Zhang *et al.*, 2022] proposed Tip-Adapter-F to utilize the few-sample training set to construct the cache model, which was fine-tuned to adapt VL models to image classification.

In summary, VL models own powerful feature extraction and generalization ability. However, the above methods cannot achieve satisfactory results in FSCIL because of the catastrophic forgetting problem of old classes and the adaptation problem to new classes with only a few samples. Inspired by the importance of discriminative features in classification [He *et al.*, 2022; Sun *et al.*, 2022; Sun *et al.*, 2023a], we propose a fine-grained feature mining prompt learning method to induce the VL model to sufficiently learn and memorize discriminative information of objects as visual and textual prototypes, which benefits the continual learning of new classes with limited data while not forgetting old classes in FSCIL.

3 Approach

The overview of our FineFMPL model is shown in Figure 2. There are two branches to extract coarse-grained global features and fine-grained object features where the proposed visual probe prompting is utilized to scale and gather discriminative information. Global-level and object-level visual prototypes are constructed based on the patch significance calculation for memorizing the class's visual knowledge. Textual context prompt is built by cross-modal mapping of the visual prototypes that contain implicit object attribute information, which generates textual prototypes to depict classes from the textual side for utilizing the cross-modal knowledge in VL models. Finally, the image classification is conducted in a two-pathway recognition way, i.e., visual-visual matching calculation and visual-textual matching calculation.

3.1 Visual Probe Prompt

Discriminative information learning is essential for learning differences among different classes, which plays an important role in image classification. Thus, we propose introducing a Visual Probe Prompt (VPP) into the image encoder of the VL model, which induces the model to scale and gather significant image patch information of visual objects for feature extraction, as shown in Figure 2.

Concretely, the visual probe prompt (VPP) is inserted into the input sequence of the image encoder of the VL model. Then, two branches are constructed to capture global information and visual object information, respectively. The original branch of the image encoder is utilized for extracting the global-level feature in the class token (abbreviated as $\mathbf{CLS}$ in Figure 2), denoted as $\mathbf{f}^{'}(\mathbf{glo})$. A new branch for scaling and gathering discriminative image patch information is constructed with the proposed VPP prompt. Specifically, we denote the output of the $L-1$ layer of the image encoder as $\mathbf{z_{L-1}} = [\mathbf{CLS^{L-1}}, ..., \mathbf{F^{L-1}}(\mathbf{x_p^i}), ..., \mathbf{F^{L-1}}(\mathbf{VPP})]$. Through the learnable VPP token, we scale the image patch token features adaptively as follows:

$$\mathbf{z_{L-1}^{new}} = \mathbf{F^{L-1}}(\mathbf{VPP}) \odot \mathbf{z_{L-1}}\,, \tag{1}$$

$$\mathbf{z_{L-1}^{new}} = \mathbf{z_{L-1}} + \mathrm{MLP}(\mathbf{z_{L-1}^{new}})\,, \tag{2}$$

where $\odot$ demotes the element-wise multiplication. Then, $\mathbf{z_{L-1}^{new}}$ is input into the last transformer layer, which conducts the information interaction to obtain $[\mathbf{f}^{'}(\mathbf{CLS}), ..., \mathbf{f}^{'}(\mathbf{x_p^i}), ..., \mathbf{f}^{'}(\mathbf{VPP})]$. Then, the $\mathbf{f}^{'}(\mathbf{VPP})$ is utilized as a probe to gather discriminative information from image patch tokens based on the probe attention as the complement of $\mathbf{f}^{'}(\mathbf{CLS})$:

$$w_i = \frac{e^{\mathbf{f}^{'}(\mathbf{x_p^i}) \times \mathbf{f}^{'}(\mathbf{VPP})}}{\sum_{i=1}^{N} e^{\mathbf{f}^{'}(\mathbf{x_p^i}) \times \mathbf{f}^{'}(\mathbf{VPP})}}\,, \tag{3}$$

$$\mathbf{f}^{'}(\mathbf{obj}) = \mathbf{f}^{'}(\mathbf{CLS}) + \sum_{i=1}^{N} w_i \times \mathbf{f}^{'}(\mathbf{x_p^i})\,, \tag{4}$$

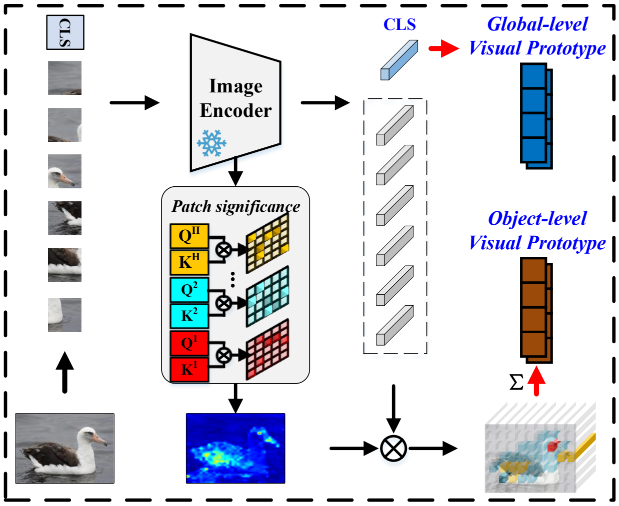

Figure 3: Illustrations of visual prototypes construction.

where N denotes the number of image patches. Thus, the proposed visual probe prompt aggregates the information from significant image patches adaptively to obtain the discriminative object-level feature $\mathbf{f}^{'}(\mathbf{obj})$.

3.2 Visual Prototypes Construction

In FSCIL, there are only a few training samples when learning new classes in incremental sessions. It needs to memorize discriminative information of new classes to distinguish them from old classes. Thus, we extract global-level and object-level visual prototypes from the training samples, as shown in Figure 3. The visual object in the image generally contains distinct image patch information crucial to the final classification. Thus, the patch significance calculation is utilized to induce the model to extract and memorize discriminative visual object information.

The image is first split into image patches as the input of the image encoder of the VL model. Concretely, the image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ with height H and width W is split with sliding stride S. Thus, the number of image patches is $N = \lfloor \frac{H}{S} \rfloor \times \lfloor \frac{W}{S} \rfloor$. The image patch is then projected by linear mapping $\mathrm{F}(\cdot)$ and combined with the class token $\mathbf{CLS}$. Next, the position embeddings are added to introduce the position information. The input image patch sequence is denoted as $\mathbf{z_0} = [\mathbf{CLS}, \mathbf{F}(\mathbf{x_p^1}), \mathbf{F}(\mathbf{x_p^2}), ..., \mathbf{F}(\mathbf{x_p^N})]$. It is noted that the $\mathbf{CLS}$ token represents the whole image for interacting with all the image patches in the transformer layers, which is utilized for the final classification. Specifically, a multi-head self-attention (MSA) module and a feed-forward neural network (FFN) in the transformer layer propagate information among image patches and the $\mathbf{CLS}$ token. We denote the input of k_{th} transformer layer as $\mathbf{z_{k-1}}$, and its output can be obtained as follows:

$$\mathbf{z_k^{'}} = \mathrm{LN}(\mathrm{MSA}(\mathbf{z_{k-1}}) + \mathbf{z_{k-1}}), \tag{5}$$

$$\mathbf{z_k} = \mathrm{LN}(\mathrm{FFN}(\mathbf{z_k^{'}}) + \mathbf{z_k^{'}}), \tag{6}$$

where $\mathrm{LN}(\cdot)$ denotes the layer normalization. The above calculation in the transformer layer enhances the global representation ability of $\mathbf{CLS}$ token to cover comprehensive context information, which is then saved as the global-level visual prototype $\mathbf{C_j^G}$ of the j_{th} class by the averaging operation on all the training samples of the j_{th} class.

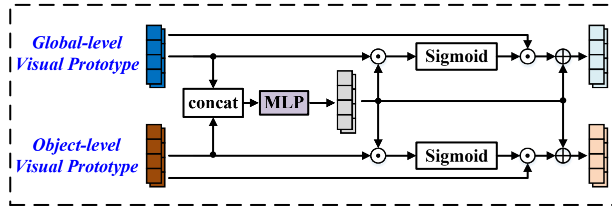

Figure 4: Cross-modal mapping of visual prototypes, which is then added to the learnable text prompt, i.e., textual context prompt.

In the self-attention calculation of the transformer, the higher the impact of the image patch on the $\mathbf{CLS}$ token, the higher its significance on the final classification, which can reveal the visual object in the image. Assume H self-attention heads exist in the l_{th} transformer layer. We utilize the $\mathbf{Q}$ and $\mathbf{K}$ to denote the query vectors and key vectors of the image patches and $\mathbf{CLS}$ token. d denotes the dimension of the above vectors. The self-attention weights are calculated as follows:

$$\mathbf{A_h^l} = \mathrm{softmax}(\frac{\mathbf{QK^T}}{\sqrt{(d/H)}}), \tag{7}$$

where $\mathbf{A_h^l} \in \mathbb{R}^{(N+1)\times(N+1)}$ $(h = 1, 2, ..., H)$ depicts the mutual significance of image patches and $\mathbf{CLS}$ token. There are L transformer layers in total. We adopt the recursive matrix multiplication to calculate the total attention for h_{th} attention head, following [He *et al.*, 2022].

$$\mathbf{A_h} = \prod_{l=1}^{L} \mathbf{A_h^l}. \tag{8}$$

The attention weight between the $\mathbf{CLS}$ token and other image patch tokens is extracted from $\mathbf{A_h}$ and denoted as $\mathbf{AM_h^{cls}} \in \mathbb{R}^{N\times 1}$. Taking all the H attention heads into account, and then the final significance of each image patch token (Figure 3) can be calculated as follows:

$$\mathbf{AM} = \sum_{h=1}^{H} \mathbf{AM_h^{cls}}. \tag{9}$$

For extracting the object-level visual features, we adopt the following weighted sum way:

$$\mathbf{f}(\mathbf{obj}) = \sum_{i=1}^{N} \mathbf{AM^i} \times \mathbf{z_L^i}. \tag{10}$$

Then, we can get the object-level prototype $\mathbf{C_j^O}$ for the j_{th} class by utilizing the averaging calculation on all the training samples of the j_{th} class.

In summary, based on the image patch significance analysis, we get the global-level prototype $\mathbf{C_j^G}$ to memorize the comprehensive context information of classes. The object-level prototype $\mathbf{C_j^O}$ is obtained to save the discriminative information in visual objects of classes.

3.3 Textual Context Prompt

The textual feature of class information extracted by the VL model can benefit the recognition of images by the text prompt design [Zhou *et al.*, 2022d]. Intuitively, the more comprehensive context information the prompt covers, the more detailed class information that the textual feature describes. Thus, we propose textual context prompts conditioned on cross-modal mapping of visual prototypes, which contains implicit object attribute information, such as bird wing texture, based on the image patch significance analyses, as shown in Figure 2 and Figure 4.

Concretely, We can get the global-level visual prototype $\mathbf{C_j^G}$ and object-level visual prototype $\mathbf{C_j^O}$ of the j_{th} class in Section 3.2. To comprehensively use the visual features, we first learn the mutual feature by concatenating the features and mapping as follows:

$$\mathbf{t_j} = \mathrm{MLP}([\mathbf{C_j^G}, \mathbf{C_j^O}])\,, \tag{11}$$

where MLP conducts the cross-modal mapping. $\mathbf{t_j}$ is extracted from the two kinds of visual prototypes, which contain different context information. $\mathbf{C_j^G}$ contains the global context, including the situation, and $\mathbf{C_j^O}$ contains the object attribute information, including the object's color and texture. For obtaining stronger context information, the attention-based feature enhancement is conducted as follows:

$$\begin{aligned}
&\mathbf{a_G} = \mathrm{sigmoid}(\mathbf{t_j} \odot \mathbf{C_j^G})\,, \\
&\mathbf{t_j^G} = \mathbf{a_G} \odot \mathbf{C_j^G} + \mathbf{t_j}\,, \\
&\mathbf{a_O} = \mathrm{sigmoid}(\mathbf{t_j} \odot \mathbf{C_j^O})\,, \\
&\mathbf{t_j^O} = \mathbf{a_O} \odot \mathbf{C_j^O} + \mathbf{t_j}\,,
\end{aligned} \tag{12}$$

where $\odot$ demotes the element-wise multiplication. $\mathbf{t_j^G}$ and $\mathbf{t_j^O}$ are added to learnable prompt tokens to form textual context prompts, which are input into the text encoder of the VL model to get the global-level textual prototype $\mathbf{T_j^G}$ and object-level textual prototype $\mathbf{T_j^O}$ to memorize the information of classes from the textual side. It benefits the model's final classification performance by utilizing the cross-modal alignment knowledge in VL models.

3.4 Inference

As shown in Figure 2, we construct two pathways for the model's recognition based on the dual-modality prompting. The first pathway utilizes the visual information of the training samples memorized by visual prototypes, and the second pathway utilizes the cross-modal knowledge contained in the textual prototypes.

In the first pathway, we utilize global-level visual prototype $\mathbf{C_j^G}$ and object-level visual prototype $\mathbf{C_j^O}$ for cosine similarity matching calculation with corresponding extracted features of $\mathbf{f}'(\mathbf{glo})$ and $\mathbf{f}'(\mathbf{obj})$. A simple MLP network is added before the similarity calculation to narrow the gap between the pre-trained data of the VL model and downstream data. Finally, we get the prediction logit for the global level $\mathbf{p_1^I}$ and object level $\mathbf{p_1^O}$ and the total prediction logit for the first pathway is calculated:

$$\mathbf{p_1} = \alpha \mathbf{p_1^O} + \mathbf{p_1^I} \tag{13}$$

Dataset	C^{base}	C^{inc}	#Inc	N-way-K-shot	Image Size
CUB-200-2011	100	100	10	10-way-5-shot	224×224
CIFAR100	60	40	8	5-way-5-shot	32×32
miniImageNet	60	40	8	5-way-5-shot	84×84

Table 1: Dataset setup in the FSCIL task.

where α is a tunable hyper-parameter. We obtain the prediction logit for the second pathway by calculating the cross-modal similarity directly, which utilizes global-level textual prototype $\mathbf{T_j^G}$ and object-level textual prototype $\mathbf{T_j^O}$ for matching calculation with $\mathbf{f}'(\mathbf{glo})$ and $\mathbf{f}'(\mathbf{obj})$. The prediction logit are denoted as $\mathbf{p_2^I}$ and $\mathbf{p_2^O}$, respectively. The total prediction logit for the second pathway is $\mathbf{p_2} = \alpha \mathbf{p_2^O} + \mathbf{p_2^I}$. Thus, we can get the final prediction logit $\mathbf{p}$ of the image:

$$\mathbf{p} = \beta \mathbf{p_1} + \mathbf{p_2}, \tag{14}$$

where β is a tunable hyper-parameter. Our FineFMPL approach mines and memorizes discriminative information of classes as visual prototypes and textual prototypes, which learns the classes sufficiently and alleviates the forgetting to benefit FSCIL.

4 Experiments

We conduct extensive comparison experiments and ablation studies on three standard few-shot class incremental learning (FSCIL) benchmark datasets, i.e., CUB-200-2011 [Wah *et al.*, 2011], CIFAR 100 [Krizhevsky *et al.*, 2009], and miniImageNet [Russakovsky *et al.*, 2015], which shows the effectiveness of our proposed FineFMPL approach.

4.1 Dataset and Metric

For fair comparisons with state-of-the-art (SOTA) FSCIL methods, the same benchmark datasets and FSCIL setting [Tao *et al.*, 2020] are adopted, as shown in Table 1.

CUB-200-2011. It is a fine-grained image classification dataset comprising 11,788 images from 200 bird classes. Subtle differences among different bird classes make this dataset very challenging. In FSCIL, 100 classes are selected as base classes, and the remaining classes are split into 10 sessions, where each session learns 10 classes with 5 examples for each class (10-way-5-shot).

CIFAR100. This is composed of 60,000 images from 100 classes. In the few-shot continual learning process, 60 classes are selected as the base class set, and the remaining 40 classes are incremental classes. There are 8 continual sessions, learning 5 new classes with 5 examples each class in each session, i.e., the 5-way-5-shot setting.

miniImageNet. It covers 60,000 images from 100 classes. 60 classes are set as base classes, and the remaining 40 classes are divided into 8 sessions, which learns 5 new classes with 5 examples each class, also in a 5-way-5-shot manner.

The classification accuracy is adopted as the evaluation metric, which is calculated after each session.

4.2 Implementation Details

Our FineFMPL approach adopts the widely-used public VL model, i.e., CLIP [Radford *et al.*, 2021], as the backbone,

Methods	Publications	Accuracy (%) in each session											Avg
		0	1	2	3	4	5	6	7	8	9	10	
F2M [Shi *et al.*, 2021]	NeurIPS 2021	81.1	78.2	75.6	72.9	70.9	68.2	67.0	65.3	63.4	61.8	60.3	69.5
CEC [Zhang *et al.*, 2021]	CVPR 2021	75.9	71.9	68.5	63.5	62.4	58.3	57.7	55.8	54.8	53.5	52.3	61.3
FACT [Zhou *et al.*, 2022a]	CVPR 2022	75.9	73.2	70.8	66.1	65.6	62.2	61.7	59.8	58.4	57.9	56.9	64.4
MCNet [Ji *et al.*, 2023]	TIP 2023	77.6	74.0	70.5	65.8	66.2	63.8	62.1	61.8	60.4	60.1	59.1	65.6
DSN [Yang *et al.*, 2023]	TPAMI 2023	76.1	72.2	69.6	66.7	64.4	62.1	60.2	58.9	57.0	55.1	54.2	63.3
LIMIT [Zhou *et al.*, 2022b]	TPAMI 2023	75.9	73.6	72.0	68.1	67.4	63.6	62.4	61.4	59.9	58.7	57.4	65.5
LDC [Liu *et al.*, 2023a]	TPAMI 2023	77.9	76.9	74.6	70.1	68.9	67.2	64.8	64.2	63.0	62.4	61.6	68.3
GKEAL [Zhuang *et al.*, 2023]	CVPR 2023	78.9	75.6	72.3	68.6	67.2	64.3	63.0	61.9	60.2	59.2	58.7	66.4
CABD [Zhao *et al.*, 2023]	CVPR 2023	79.1	75.4	72.8	69.1	67.5	65.1	64.0	63.5	61.9	61.5	60.9	67.3
SAVC [Song *et al.*, 2023]	CVPR 2023	81.9	77.9	75.0	70.2	70.0	67.0	66.2	65.3	63.8	63.2	62.5	69.4
TEEN [Wang *et al.*, 2023]	NeurIPS 2023	77.3	76.1	72.8	68.2	67.8	64.4	63.3	62.3	61.2	60.3	59.3	66.6
CLIP* [Radford *et al.*, 2021]	ICML 2021	64.9	62.9	61.6	57.9	58.1	58.2	56.9	55.7	54.3	54.4	55.1	58.2
Coop*[Zhou *et al.*, 2022d]	IJCV 2022	83.9	79.7	77.5	72.5	70.4	69.2	67.6	66.1	63.9	63.5	63.4	70.7
IOS [Yoon *et al.*, 2023]	arXiv 2023	81.3	77.4	75.8	73.3	72.6	70.4	68.7	67.3	65.9	64.4	63.8	71.0
Our FineFMPL method	This paper	**86.7**	**84.2**	**83.2**	**79.8**	**80.0**	**79.0**	**78.1**	**77.5**	**76.2**	**76.1**	**76.4**	**79.7**

Table 2: Comparison with SOTA methods on the CUB-200-2011 dataset for FSCIL. * denotes the reproduced results with the officially released codes. **Bold value** indicates the optimal classification accuracy and the underlined value indicates the suboptimal accuracy.

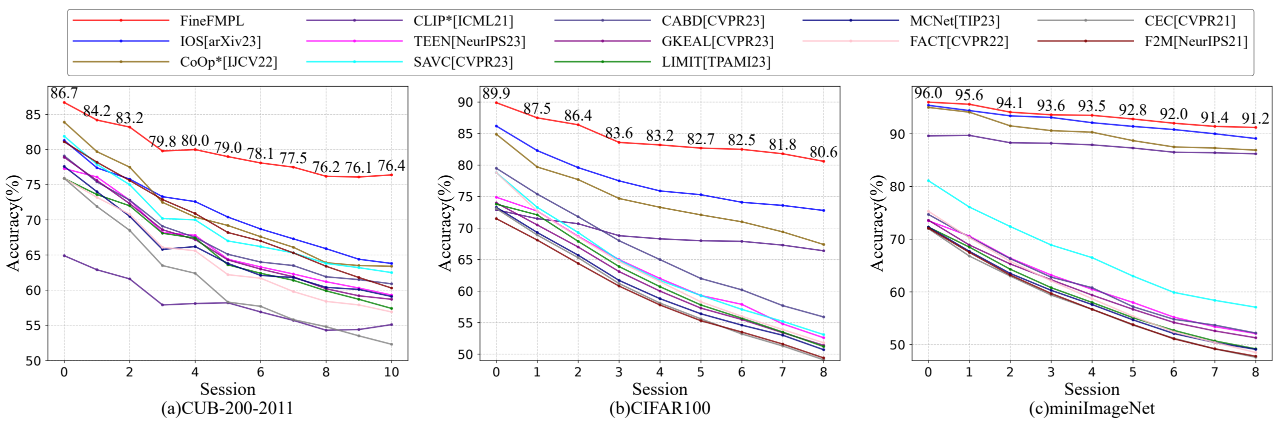

Figure 5: The classification accuracy trends of our FineFMPL method and other SOTA compared methods on the CUB-200-2011, CIFAR100, and miniImageNet datasets for FSCIL.

where the ViT-B_16 is selected as the image encoder, and the transformer network is utilized as the text encoder. The original weights of CLIP are frozen during the whole training stage. We follow the mainstream experimental setting in [Tao *et al.*, 2020] for a fair comparison. In the training process, we set 50 training epochs for CUB-200-2011 in the base session and 10 epochs for each incremental session. For the CIFAR 100 and miniImageNet datasets, we set 30 training epochs in the base session and 10 epochs in each incremental session. α is set to 0.5, 2, 0.5, and β is set to 1.5, 1, and 0.5 for the three datasets, respectively. The batch size is set as 256. The learning rate is initialized as 1e-3 in the base session and 1e-4 in each incremental session, which all adopt the cosine annealing schedule. AdamW [Kingma and Ba, 2014] is utilized as the model optimizer. All the experiments are conducted on one NVIDIA A40 GPU with Pytorch.

4.3 Comparison with State-Of-The-Art Methods

We conduct extensive comparison experiments with state-of-the-art (SOTA) methods on three standard FSCIL benchmark datasets. The comparison results are shown in Table 2 and Figure 5. We can observe that:

- On the challenging CUB-200-2011 dataset, which has subtle differences among different classes, our proposed FineFMPL approach outperforms the compared methods by significant margins in each session, as shown in Table 2. Compared with existing SOTA FSCIL methods using pure vision backbone networks such as ResNet [He *et al.*, 2016], our FineFMPL surpasses the typical F2M, SAVC, CABD, and TEEN by **10.2%**, **10.3%**, **12.4%**, and **13.1%** on the average classification accuracy of all sessions, respectively. Besides, our FineFMPL approach can keep the performance better in the incremental sessions compared with other SOTA FSCIL methods. SAVE proposed to imagine virtual classes to introduce semantic knowledge for enhancing the separation of different classes. By contrast, we propose to transfer the general knowledge of the VL model and mine the discriminative information of classes with the fine-grained dual-modality prompts design. Thus, our FineFMPL has a stronger generalization ability to learn new classes with limited data, which alleviates the overfitting problem in FSCIL to a large extent. Compared with CABD, which utilizes the class-aware bilateral dis-

Visual Probe Prompt		Textual Context Prompt		CUB(%)	CIFAR100(%)	miniImageNet(%)
Use global-level visual prototype	Use object-level visual prototype	Use global-level textual prototype	Use object-level textual prototype			
✗	✗	✗	✗	58.2	69.1	87.8
✓	✗	✓	✗	79.0	83.1	92.8
✗	✓	✗	✓	66.0	82.3	90.9
✓	✓	✗	✗	77.9	79.8	91.4
✓	✓	✓	✓	79.7	84.2	93.4

Table 3: Ablation studies about each component of our FineFMPL method on the three FSCIL datasets. The values in the table denote the average classification accuracy over all sessions.

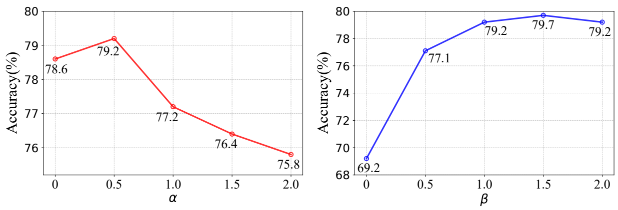

Figure 6: Hyper-parameter experiments about α and β on the CUB-200-2011 dataset.

tillation from base classes, our FineFMPL model directly memorizes discriminative information of classes as the visual and textual prototypes, which helps alleviate the catastrophic forgetting problem of old classes.

- We compare our FineFMPL method with the prompting-based methods, such as CoOp and the recent FSCIL method IOS, which all utilize CLIP as the backbone. Our FineFMPL model achieves **9.0%** and **8.7%** average performance gains, respectively. Compared with the textual prompt pool construction of IOS in different sessions, our FineFMPL method utilizes both the visual prompt and textual prompt for inducing the VL model to gather significant image patch information and depict class information, which enhances the model's learning ability of classes in FSCIL.
- Figure 5 shows the classification accuracy trend on the CUB-200-2011, CIFAR100, and miniImageNet datasets. Our FineFMPL approach can achieve consistent performance gains, bringing **8.7%**, **6.7%**, and **1.2%** average improvements over the sub-optimal method. Our FineFMPL approach declines more slowly in incremental stages. We attribute it to discriminative information memorizing from both the visual and textual sides.

4.4 Ablation Experiments

Experimental results of ablation studies on the three FSCIL datasets are shown in Table 3. We can observe that:

- Compared with CLIP (our baseline), i.e., the first row, our FineFMPL brings significant average accuracy gains of 21.5%, 15.1%, and 5.6% on the CUB-200-2011, CIFAR 100, and miniImageNet datasets, respectively. We attribute the gains to our FineFMPL's inducing CLIP to learn and memorize the discriminative information of classes to alleviate old classes' forgetting and new classes' learning simultaneously.
- Compared with the baseline, our visual probe prompt brings 19.7%, 10.7%, and 3.6% performance gains when using both the global-level and object-level prototypes (as shown in row 4). The proposed visual probe prompt scales and gathers discriminative visual features, which help learn the classes sufficiently to improve the FSCIL performance.
- Compared with only utilizing the fine-grained visual prompt, the recognition accuracy further improves by 1.8%, 4.4%, and 2.0% through introducing the textual context prompt, as shown in the last two rows of the table. The memorization of classes in the textual side brings an important complement to visual information kept in visual prototypes, which benefits alleviating the forgetting problem in FSCIL.

4.5 Hyper-parameter Experiments

Hyper-parameter experiments about α in Eq. 13 and β in Eq. 14 are conducted on the CUB-200-2011 dataset under the FSCIL setting. The experimental results are presented in Figure 6. The best performance is obtained when α is set as 0.5. The trend of the curve indicates the importance of the object branch, which captures crucial discriminative traits for assisting classification. Our proposed FineFMPL approach can perform best when β is set as 1.5. It verifies the effectiveness of discriminative visual information memorized by the global-level and object-level visual prototypes, which alleviates the forgetting problem when learning new classes.

5 Conclusion

In this paper, we propose Fine-grained Feature Mining Prompt Learning (FineFMPL) of the Vision-Language (VL) pre-training model for Few-Shot Class Incremental Learning (FSCIL), which learns and memorizes discriminative information of classes as visual and textual prototypes. We propose the visual probe prompt for inducing the image encoder of the VL model to extract significant image patch information of visual objects, and the visual prototypes of classes are preserved to save visual knowledge of classes. The textual context prompt is then constructed and conditioned on the cross-modal mapping of visual prototypes that contain object attribute information implicitly, which help depict the class information as textual prototypes. Sufficient discriminative information learning and memorizing benefits the understanding of new classes with few training samples while not forgetting old classes, which achieves promising performance in FSCIL.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 61925201, 62132001, and 62373043.

References

[Abati *et al.*, 2020] Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3931–3940, 2020.

[An *et al.*, 2023] Yuexuan An, Xingyu Zhao, and Hui Xue. Learning to learn from corrupted data for few-shot learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3423–3431, 2023.

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

[Cheraghian *et al.*, 2021] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2534–2543, 2021.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dong *et al.*, 2021] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1255–1263, 2021.

[Han *et al.*, 2023] Bing Han, Feifei Zhao, Yi Zeng, Wenxuan Pan, and Guobin Shen. Enhancing efficient continual learning with dynamic structure development of spiking neural networks. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2993–3001, 2023.

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

[He *et al.*, 2022] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 852–860, 2022.

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

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

[Ji *et al.*, 2023] Zhong Ji, Zhishen Hou, Xiyao Liu, Yanwei Pang, and Xuelong Li. Memorizing complementation network for few-shot class-incremental learning. *IEEE Transactions on Image Processing*, 32:937–948, 2023.

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

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Li and Hoiem, 2017] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[Liu *et al.*, 2023a] Binghao Liu, Boyu Yang, Lingxi Xie, Ren Wang, Qi Tian, and Qixiang Ye. Learnable distribution calibration for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Liu *et al.*, 2023b] Chenghao Liu, Xiaoyang Qu, Jianzong Wang, and Jing Xiao. Fedet: a communication-efficient federated class-incremental learning framework based on enhanced transformer. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3984–3992, 2023.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.

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

[Shi *et al.*, 2021] 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*, 34:6747–6761, 2021.

[Smith *et al.*, 2023] 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*, pages 11909–11919, 2023.

[Song *et al.*, 2023] 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.

[Sun *et al.*, 2022] Hongbo Sun, Xiangteng He, and Yuxin Peng. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5853–5861, 2022.

[Sun *et al.*, 2023a] Hongbo Sun, Xiangteng He, and Yuxin Peng. Hcl: Hierarchical consistency learning for webly supervised fine-grained recognition. *IEEE Transactions on Multimedia*, 26:5108–5119, 2023.

[Sun *et al.*, 2023b] Hongbo Sun, Xiangteng He, Jiahuan Zhou, and Yuxin Peng. Fine-grained visual prompt learning of vision-language models for image recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5828–5836, 2023.

[Tao *et al.*, 2020] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12183–12192, 2020.

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

[Wang *et al.*, 2022] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[Wang *et al.*, 2023] 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. *arXiv preprint arXiv:2312.05229*, 2023.

[Yang *et al.*, 2023] 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.

[Yoon *et al.*, 2023] 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. *arXiv preprint arXiv:2309.02833*, 2023.

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

[Zhang *et al.*, 2022] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022.

[Zhao *et al.*, 2023] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11838–11847, 2023.

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

[Zhou *et al.*, 2022b] 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*, 2022.

[Zhou *et al.*, 2022c] Kaiyang Zhou, Jingkang Yang, C. C. Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022d] 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.

[Zhuang *et al.*, 2023] Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. Gkeal: Gaussian kernel embedded analytic learning for few-shot class incremental task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2023.