




Class-Incremental Learning with CLIP: Adaptive Representation Adjustment and Parameter Fusion

Linlan Huang¹, Xusheng Cao¹, Haori Lu¹, and Xialei Liu^{1,2}^(✉)

¹VCIP, CS, Nankai University

²NKIARI, Shenzhen Futian

{huanglinlan, caoxusheng, luhaori}@mail.nankai.edu.cn
{xialei}@nankai.edu.cn

Abstract. Class-incremental learning is a challenging problem, where the goal is to train a model that can classify data from an increasing number of classes over time. With the advancement of vision-language pre-trained models such as CLIP, they demonstrate good generalization ability that allows them to excel in class-incremental learning with completely frozen parameters. However, further adaptation to downstream tasks by simply fine-tuning the model leads to severe forgetting. Most existing works with pre-trained models assume that the forgetting of old classes is uniform when the model acquires new knowledge. In this paper, we propose a method named Adaptive **R**epresentation **A**justment and **P**arameter **F**usion (**RAPF**). During training for new data, we measure the influence of new classes on old ones and adjust the representations, using textual features. After training, we employ a decomposed parameter fusion to further mitigate forgetting during adapter module fine-tuning. Experiments on several conventional benchmarks show that our method achieves state-of-the-art results. Our code is available at <https://github.com/linlany/RAPF>.

Keywords: Class-incremental Learning · Visual Language Model

1 Introduction

The real world is constantly changing, which requires the model to adapt to new knowledge while retaining old knowledge. If not updated, the models can become obsolete, degrading their performance over time [10]. Privacy and storage constraints may limit access to old data, leading to a severe imbalance in data distribution. This imbalance, when models are updated, causes models to become biased towards current data and forget previously acquired knowledge, a phenomenon known as catastrophic forgetting [19]. Therefore, the challenge of continuous learning is to balance plasticity and stability [26], allowing models to learn new knowledge without forgetting old knowledge and to reuse and expand knowledge from experience across different tasks.

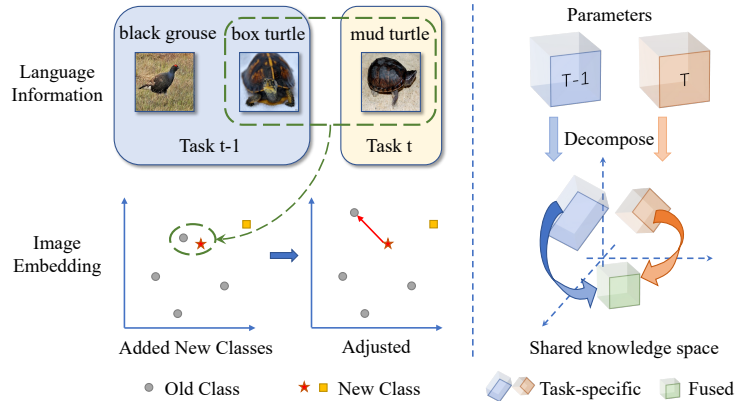


Fig. 1: Semantically similar categories pose significant challenges in CIL across tasks, in which language information can help to pick out the adjacent old and new classes when new data are encountered. Then the image feature representation of the old category can be adjusted accordingly. Additionally, a decomposed parameter fusion strategy is further adapted to reduce forgetting. We decompose the parameters learned from two consecutive tasks into shared knowledge and task-specific knowledge. Then, we fuse the parameters based on this decomposition.

Class-incremental learning, a challenging aspect of continual learning, has garnered increasing attention recently. It involves learning from a data stream where new classes are added over time. To cope with this scenario, three main categories of methods have been proposed: regularization-based, replay-based, and parameter isolation-based methods [4]. These methods aim to preserve previous knowledge by adding regularization terms to the loss function, replaying old data samples, or allocating dedicated parameters for each class. However, most of these methods rely on models that are trained from scratch, which may not be optimal for incremental learning. Thus, an important research direction is to reduce forgetting by applying various methods to models pre-trained on large-scale datasets.

Pre-trained models on large-scale datasets have shown remarkable generalization ability and robustness against catastrophic forgetting in downstream tasks [18, 28, 43, 46]. Moreover, the visual language pre-trained model (CLIP) [30] has demonstrated powerful zero-shot ability in continual learning [38] and high adaptability to downstream tasks [8, 41, 54]. Leveraging the excellent feature extraction ability of the pre-trained model, each incremental step requires updating only a small number of parameters, reducing the risk of forgetting. In contrast, models trained from scratch lack this advantage and may suffer from severe performance degradation. Therefore, pre-trained models have achieved impressive results in continual learning. There are two main strategies for continual learning with pre-trained models: fine-tuning the model [48] or continuously ex-

pand a small number of parameters on the model, such as methods based on prompts [33, 43–45] or adding adapters [52].

Fine-tuning models may compromise the feature extraction ability of the original model and cause catastrophic forgetting, even with regularization constraints. Expanding parameters may mitigate interference with the original model but increase time and space costs over time. For the visual language pre-trained model, the language encoder provides rich information beneficial for continual learning. However, most existing methods only use text features for classification, not fully exploring their potential to reduce forgetting.

In this paper, we propose a method that uses text features to enhance the classification ability of neighboring categories in class-incremental learning. It is based on the observation that CLIP relies on the fixed text features for classification. The text features determine the decision boundary. When a new category arrives, the new decision boundary may divide part of the old category sample into the new category. To address it, we need to enhance the separation of neighboring categories. Using the text features, we can infer the relationship between the old and new categories. We select the pairs of neighboring categories by calculating the distance between the textual features of the new and old categories. Since the new category has sufficient data for learning, we do not need to modify its representation. Instead, we focus on adjusting the representation of old categories that are affected by new categories, as illustrated in Fig. 1. We train a single linear layer to prevent compromising the feature extraction capability of the pre-trained model.

Moreover, we propose a decomposed parameter fusion method for the linear adaptive layer that does not increase the number of parameters as the task increases. Unlike directly calculating parameter averages, our fusion strategy is more fine-grained and considers the shared knowledge between tasks. To balance stability and plasticity, we merge the parameters before and after an incremental task according to the parameter changes caused by learning the current task. We do not need to add extra distillation loss in the training process to constrain the parameter changes, which lowers the training cost.

The main contributions of this paper are as follows:

- We explore a method to reduce the forgetting of CLIP models by using the textual features of category names;
- We propose a simple but effective method of decomposed parameter fusion for the linear layer adapter of pre-trained models;
- We achieved state-of-the-art results on several datasets.

2 Related Work

2.1 Class-Incremental Learning (CIL)

Incremental learning methods can be classified into three types according to the main strategy they use [4]. The methods based on regularization suppress the forgetting of old categories by imposing some constraints on the model’s parameters

or outputs. MAS [1] and EWC [15] compute the importance of each parameter for the old task and then add a regularization term to avoid catastrophic forgetting. Coil [51] uses knowledge distillation based on optimal transport to share knowledge between tasks. The methods [6, 20, 49] use the loss of knowledge distillation as the regularization term. The methods based on replay preserve the memory of old categories by saving some samples or features of old categories (exemplar), and then replaying them with the new category samples when training new categories. Some are exploring how to reduce the storage occupancy of exemplars [24], while others are looking for better strategies to select samples that will be added to exemplars [22, 23, 36]. There are also methods that save additional models and samples to assist in the current model training [2, 31, 32]. The methods based on dynamic architecture dynamically adjust the model’s structure to adapt to the learning of new categories [7, 42, 46, 47].

CIL with Pre-trained Model. There are two main types of methods for using pre-trained models. One type is to fine-tune the parameters of the model itself to adjust the feature representation. ZSCL [50] utilizes numerous external data to distill the pre-trained model to maintain a stable feature space. Zhang et al. [48] employ different learning rates to update the pre-trained backbone network and the classifier. The other type is to keep the pre-trained model unchanged and add parameters to adjust the feature representation. Liu et al. [21] introduce an adapter to the pre-trained CLIP model to adapt to incremental tasks. PROOF [53] trains an adapter for each task and uses cross-modal attention to fuse the language and vision information of CLIP. RanPac [27] uses a high-dimensional projection to separate features. The recently proposed prompt-based methods [14, 33, 44, 45] select the corresponding prompts to add to the model according to the output of the features by the pre-trained model and then re-obtain the features for classification. Among them, LGCL [14] tries to introduce language guidance. Ostapenko et al. [29] add a classification network to the pre-trained model, and use the pre-trained latent feature space for replay.

CIL without Exemplar. Sometimes it is not possible to store old class samples due to privacy and memory constraints [34]. Some existing works without exemplars use Gaussian distribution to model data and help with classification [12, 37, 48]. Other approaches include oversampling prototypes [56] or augmenting prototypes to simulate replay samples [25, 55]. Some recent works use models to synthesize data from the old task as a substitute for exemplars [3, 9]. Prompt-based methods use the frozen backbone and the relative isolation of prompt parameters to avoid using exemplars [14, 33, 44, 45].

3 Preliminaries

Class Incremental Learning Definition. A class-incremental learning algorithm trains a model M_t at each task t , which can classify data from all classes that the algorithm has seen so far without task id, i.e., $C_1 \cup C_2 \cup \dots \cup C_t$. And the category sets do not intersect, $\forall i \neq j, C_i \cap C_j = \emptyset$. The model M_t only uses

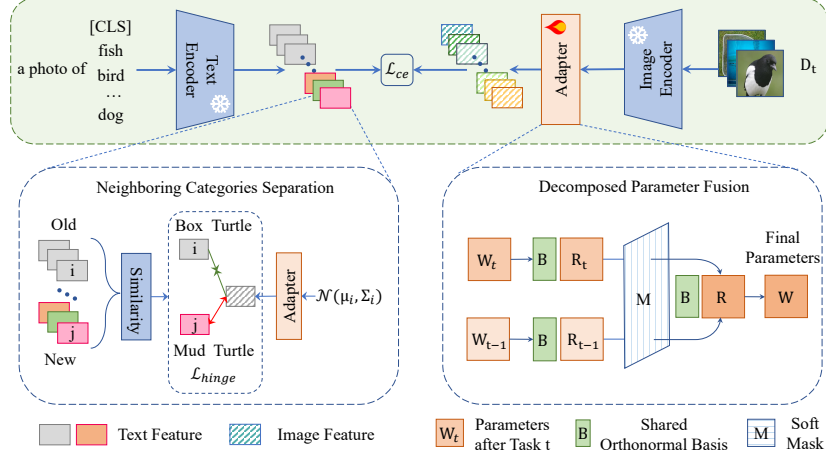


Fig. 2: The framework of our method. The neighboring categories separation module computes the similarity of text features to identify neighboring categories. We sample the distribution of the old class and calculate the hinge loss. In the parameter fusion module, we first decompose \mathbf{W}_t and \mathbf{W}_{t-1} into the same standard orthogonal basis \mathbf{B} . Then, we calculate a soft mask \mathbf{M} from the difference of the decomposed parameters \mathbf{R}_t and \mathbf{R}_{t-1} , which acts as the fusion weight. Finally, we reconstruct the parameter \mathbf{W} from the fused parameter \mathbf{R} and the basis \mathbf{B} .

the data in the current dataset D_t to update from the previous model M_{t-1} , without accessing previous datasets.

CLIP Adapter. An efficient method to adapt a pre-trained vision-language model to downstream tasks involves incorporating a small network as an adapter [8]. We denote the visual feature extractor of the CLIP backbone network as $f_{img}(\cdot)$, the text feature extractor as $f_{text}(\cdot)$, and the linear adapter as $A(\cdot)$. Given the image input \mathbf{x}_i , the class y_i with the fixed prompt template, e.g., “a photo of a [CLS]”, denoted by \mathbf{t}_i , the output result is as follows:

$$p(y_i|\mathbf{x}_i) = \frac{\exp(\cos(A(f_{img}(\mathbf{x}_i)), f_{text}(\mathbf{t}_i))/\tau)}{\sum_{j=1}^{|y|} \exp(\cos(A(f_{img}(\mathbf{x}_i)), f_{text}(\mathbf{t}_j))/\tau)}, \quad (1)$$

where τ is the temperature. If the text encoder is frozen, we only need to preserve the embeddings of the text instead of class names in CIL. We employ the cross-entropy loss criterion to fine-tune the parameters of the adapter, formulated as follows:

$$\mathcal{L}_{ce}(\mathbf{y}, \mathbf{p}) = - \sum_{i=1}^n y_i \log p_i. \quad (2)$$

Feature Generation. Leveraging the well-learned representations by the pre-trained backbone, we can approximate the feature distribution of each category using a Gaussian distribution [48]. We compute the centroid $\boldsymbol{\mu}_c$ and the covariance matrix $\boldsymbol{\Sigma}_c$ from the image embeddings denoted by $\mathbf{e}_c = f_{img}(\mathbf{x}_c)$ of class

c. Then we can generate image embeddings for class c by sampling from the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.

4 Method

4.1 Overview

As illustrated in Fig. 2, our framework comprises a pre-trained CLIP model and a linear adapter. The image features are derived from passing the images through the image encoder and adapter, while the label features are derived from passing the class labels through the text encoder. The classification result is determined by measuring the similarity between them. The old class features are generated by sampling from their respective Gaussian distributions and fed into the adapter along with the new data to compute classification loss. Simultaneously, we select pairs of neighboring classes based on their text feature similarity. For each old class in these pairs, we sample more data from its Gaussian distribution and feed them into the adapter but only compute the loss for separation. This way, we can adjust the feature representation of the old classes affected by the similar new classes, thereby reducing the forgetting of the old classes caused by learning the new classes. Upon completion of the t -th task training phase, we fuse the parameters of the adapter from the preceding phase and the current adapter to obtain the ultimate adapter for this stage. The adaptive fusion of parameters can better balance plasticity and stability.

4.2 Enhancing Neighboring Categories Separation with Language Guidance

One manifestation of catastrophic forgetting is that the model erroneously identifies old category data as a new category. This phenomenon is more severe when the new category is similar to the old category. These categories belong to neighboring categories. They usually have similar semantic meanings in language and similar appearances, such as ‘macaw’ and ‘lorikeet’. It is challenging for the model to distinguish them, especially when new categories are added.

By using the CLIP text encoder, we can measure the similarity between category names and use it to guide the adapter learning process. To select the nearest pairs of categories, we use the normalized text features of the category names to calculate the distance between the new and old categories:

$$\mathbf{D} = \text{dist}(f_{\text{text}}(\mathbf{t}_{\text{new}}), f_{\text{text}}(\mathbf{t}_{\text{old}})), \quad (3)$$

where \mathbf{D} is the matrix of Euclidean distances between the normalized text features of the new and old categories, $f_{\text{text}}(\mathbf{t}_{\text{new}})$ and $f_{\text{text}}(\mathbf{t}_{\text{old}})$. Learning new categories interferes with the performance of old ones. We focus on measuring the distance between the new and old categories, which reflects the degree of interference. Then, we pick out the neighboring categories set \mathcal{P} :

$$\mathcal{P} = \{(i, j) | \mathbf{D}_{ij} < \alpha\}, \quad (4)$$

where α represents the threshold. \mathbf{D}_{ij} is the distance between the text features of the old category i and the new category j . This criterion reduces the complexity of many-to-many relationships by selecting a subset of one-to-one relationships.

We introduce a hinge loss by sampling the features of the old category from a normal distribution for each pair of neighboring categories. Since we do not have real data on the old classes, excessive adjustments may harm the classification performance of the old classes. Therefore, we only use hinge loss to make small but effective adjustments. The sampled data of the old category c is indicated by $\hat{\mathbf{e}}_c$, and the adapter is indicated by A :

$$\mathcal{L}_{hinge} = \sum_{k=1}^{|\mathcal{P}|} \max(\text{dist}(A(\hat{\mathbf{e}}_c), f_{text}(\mathbf{t}_c)) - \text{dist}(A(\hat{\mathbf{e}}_c), f_{text}(\mathbf{t}_{\ell})) + m, 0), \quad (5)$$

where m is a constant margin, c and ℓ denote the old category and the new category in the k -th pair of neighbor categories which belong to set \mathcal{P} . This loss function can make the adapter focus more on the neighboring categories that are more likely to be confused while minimizing the disturbance to the representation of the old categories.

The final loss function is combined with Cross Entropy loss \mathcal{L}_{ce} as:

$$\mathcal{L} = \mathcal{L}_{hinge} + \mathcal{L}_{ce}. \quad (6)$$

4.3 Increasing Stability with Decomposed Parameter Fusion

We introduce a parameter fusion mechanism to maintain the stability of the linear adapter, thereby reducing forgetting.

We evaluate the importance of each new parameter. The gradient indicates the direction of parameter updates based on new data, and the difference in the parameters between tasks is the weighted sum of the gradients. So the parameters with large changes are more important for learning new knowledge.

To obtain the importance matrix as a fine-grained weight for parameter fusion, we calculate the difference between the parameters after training and the parameters from the previous task and normalize it by the maximum value:

$$\mathbf{M} = \min(1, \frac{|\mathbf{W}_{new} - \mathbf{W}_{old}|}{\max(|\mathbf{W}_{new} - \mathbf{W}_{old}|)} + b), \quad (7)$$

where \mathbf{M} denotes the importance of each new parameter, \mathbf{W}_{new} denotes the parameters obtained from the current task training, \mathbf{W}_{old} denotes the parameters from the previous task and b is constant bias. In order to compare the difference between \mathbf{W}_{new} and \mathbf{W}_{old} under the same standard, we decompose \mathbf{W}_{old} to an orthonormal basis \mathbf{B} by SVD [11] and calculate the projection from matrix \mathbf{W}_{new} to \mathbf{B} :

$$\mathbf{W}_{old} \xrightarrow{decompose} \mathbf{B}\mathbf{R}_{old}, \quad (8)$$

$$\mathbf{R}_{new} = \mathbf{B}^\top \mathbf{W}_{new}. \quad (9)$$

The parameter matrix \mathbf{W}_{old} and \mathbf{W}_{new} is expressed as linear combinations of the orthonormal matrix \mathbf{B} , i.e., \mathbf{R}_{old} and \mathbf{R}_{new} , which represent task-specific knowledge. The matrix \mathbf{B} represents the shared knowledge space across the parameter matrices. Thus, the matrix difference that we calculate can be transformed into calculating the difference of two distinct weights of the same orthonormal basis.

We substitute the \mathbf{R}_{new} and \mathbf{R}_{old} obtained from the decomposition into the \mathbf{W}_{new} and \mathbf{W}_{old} of Eq. (7) to calculate \mathbf{M} . Then we compute the matrix \mathbf{R} and the final parameter \mathbf{W} :

$$\mathbf{R} = (\mathbf{J} - \mathbf{M}) \odot \mathbf{R}_{old} + \mathbf{M} \odot \mathbf{R}_{new}, \quad (10)$$

$$\mathbf{W} = \mathbf{B}\mathbf{R}, \quad (11)$$

where \mathbf{J} denotes a matrix of all ones and \odot indicates the element-wise product.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct our experiments using three datasets: CIFAR00 [16], ImageNet1K [5], ImageNet100, ImageNet-R [13] and CUB200 [40]. The CIFAR100 dataset consists of 100 categories. Each category contains 600 color images with a resolution of 32×32 pixels. Of these, 500 images are assigned to the train set, and 100 images are assigned to the test set. The ImageNet1K dataset consists of 1000 categories and the ImageNet100 dataset is a subset of the ImageNet1K dataset that consists of 100 selected classes. The ImageNet-R dataset is a collection of diverse image categories derived from the ImageNet dataset. The ImageNet-R dataset includes images in various styles such as art, cartoons, graffiti, embroidery, video games, and so on. They are different expressions of the 200 categories in the ImageNet dataset. We follow the work [38, 44] to split the dataset into a training set and a test set. The CUB200 [40] dataset is widely used for fine-grained visual categorization tasks. It comprises 11,788 images of 200 subcategories belonging to various species of birds.

Competing methods. We compare to CIL methods: L2P ++ [45], Dual-Prompt [44], CODA [33], SLCA [48], ADAM-Adapter [52] and PROOF [53]. Continual-CLIP [38] refers to the zero-shot performance of the CLIP model. PROOF [53] is the method that is based on CLIP with exemplar. To ensure fairness of comparison, all methods use the same OpenAI CLIP pre-trained weights [30]. The results of the three methods DualPrompt, L2P ++, and CODA are obtained by running the publicly available code implementation of method CODA. The results of SLCA, ADAM-Adapter, PROOF and Continual-CLIP methods are obtained from their respective public code.

Evaluation metrics. The average precision after training the t -th task in the test data for the first to t -th tasks is denoted as A_t . Avg is the average of the accuracies of all tasks. $Last$ is the average accuracy after the last tasks.

Table 1: Experimental results for continual learning on CIFAR100. B denotes the number of base classes, and Inc denotes the number of incremental classes. All baseline-based results are reproduced according to their published code with CLIP pre-trained weights for ViT-B/16. We run our experiments for several different shuffles of the class order and report the mean of these orders.

Method	Exemplar	B0 Inc5		B0 Inc10		B0 Inc20		B50 Inc5		B50 Inc10	
		Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last
PROOF [53]	✓	<u>85.12</u>	<u>76.13</u>	<u>84.88</u>	<u>76.29</u>	84.11	76.86	83.22	76.25	83.17	76.5
L2P ++ [45]	✗	79.18	68.67	81.90	73.08	84.39	77.37	58.57	18.04	76.51	48.52
DualPrompt [44]	✗	79.74	69.91	81.45	72.51	85.19	<u>77.47</u>	58.55	15.26	72.00	45.05
CODA [33]	✗	69.78	41.98	76.98	62.25	78.65	65.29	58.45	15.99	67.88	28.77
Continual-CLIP [38]	✗	75.93	66.68	75.15	66.68	74.01	66.68	70.79	66.68	70.77	66.68
SLCA [48]	✗	78.96	66.84	80.53	67.58	<u>85.25</u>	76.99	86.99	76.8	86.55	79.92
ADAM-Adapter [52]	✗	70.18	58.12	75.76	65.50	77.28	67.89	83.38	<u>76.94</u>	83.21	76.94
ours	✗	86.87	79.26	86.19	79.04	85.73	79.24	<u>85.03</u>	79.64	<u>84.73</u>	<u>79.36</u>

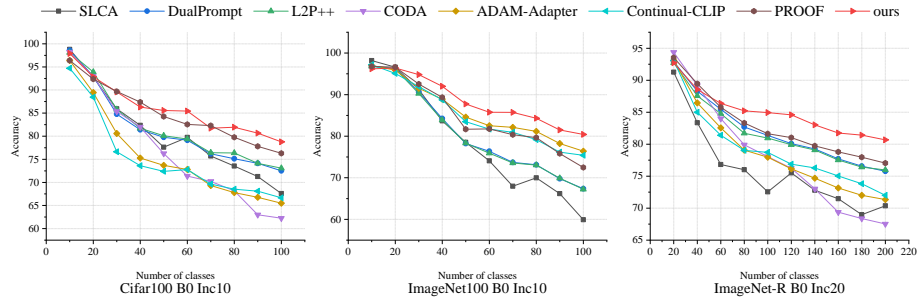
Table 2: Experimental results for continual learning on ImageNet100.

Method	Exemplar	B0 Inc5		B0 Inc10		B0 Inc20		B50 Inc5		B50 Inc10	
		Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last
PROOF [53]	✓	<u>86.92</u>	75.52	84.71	72.48	81.92	68.56	84.16	74.44	82.78	71.04
L2P ++ [45]	✗	75.43	62.10	80.51	67.22	84.12	73.70	62.00	22.15	74.11	49.46
DualPrompt [44]	✗	75.40	61.10	80.65	67.38	84.65	74.24	62.10	22.36	74.20	49.78
CODA [33]	✗	51.64	24.94	64.13	34.76	69.78	43.96	57.33	19.95	65.14	28.80
Continual-CLIP [38]	✗	85.74	75.40	84.98	75.40	84.03	75.40	81.35	75.40	81.09	75.40
SLCA [48]	✗	78.40	63.36	78.63	59.92	84.08	71.08	<u>86.47</u>	72.22	<u>86.26</u>	71.18
ADAM-Adapter [52]	✗	85.78	<u>75.72</u>	<u>85.84</u>	<u>76.40</u>	<u>85.85</u>	<u>77.08</u>	84.90	<u>78.58</u>	84.60	<u>78.58</u>
ours	✗	87.59	79.87	87.51	80.23	86.72	80.10	86.53	80.16	86.36	80.22

Implementation detail. We developed our method using PyTorch and ran it on an RTX 3090 GPU. Our backbone is the ViT-B/16 version of CLIP. We train the model with the Adam optimizer for 15 epochs, starting with a learning rate 0.001. We use a MultiStepLR scheduler that reduced the learning rate by a factor of 0.1 at epochs 4 and 10. We use 0.65 as the default threshold for the text feature distance to select the adjacent category pairs. We used approximately 2000 sampled data per epoch to simulate replay samples, which matched the amount of data added by the conventional setting with replay. In each iteration, we sampled 20 additional features for every category that was selected by threshold. Due to insufficient data for some classes, we cannot obtain a full-rank covariance matrix. We follow previous works [17, 39] and use covariance shrinkage to obtain a full-rank matrix. We run our experiments for several different shuffles of the class order and report the mean of these orders.

Table 3: Experimental results for continual learning on ImageNet-R.

Method	Exemplar	B0 Inc10		B0 Inc20		B0 Inc40		B100 Inc10		B100 Inc20	
		Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last
PROOF [53]	✓	<u>82.69</u>	<u>77.25</u>	<u>82.83</u>	<u>77.05</u>	82.63	77.12	81.61	77.10	81.78	77.17
L2P ++ [45]	✗	76.87	68.78	81.67	75.98	82.81	77.87	56.17	17.90	67.73	43.28
DualPrompt [44]	✗	77.07	69.41	82.01	75.77	<u>83.77</u>	78.64	57.37	19.18	69.18	45.37
CODA [33]	✗	75.23	64.53	78.00	67.52	78.80	71.27	56.62	17.64	65.62	35.06
Continual-CLIP [38]	✗	79.84	72.00	79.12	72.00	77.59	72.00	76.93	72.00	76.76	72.00
SLCA [48]	✗	80.18	73.57	75.92	70.37	83.35	<u>79.1</u>	<u>82.85</u>	<u>78.57</u>	<u>83.50</u>	<u>79.67</u>
ADAM-Adapter [52]	✗	76.71	68.75	78.65	71.35	79.87	73.02	79.87	75.37	79.75	75.37
ours	✗	86.28	79.62	85.58	80.28	84.69	80.18	84.12	80.04	83.99	80.35

**Fig. 3:** Accuracy curve of our method with other SOTA baselines on CIFAR100, ImageNet100 and ImageNet-R.

5.2 Comparison Results

Table 1, Tab. 2 and Tab. 3 show the comparison results of our method and existing methods on three datasets, CIFAR100, ImageNet100 and ImageNet-R. Our method outperforms various competing methods in most settings by a significant margin. On the ImageNet100 dataset, our method achieves a final accuracy that is at least 1.58% higher than other methods. Our method outperforms other methods by at least 1.08% in final accuracy in the base0 setting experiment on the ImageNet-R dataset. This shows that we have good effects in reducing forgetting and maintaining model stability by using textual information and parameter fusion. Compared with zero-shot Continual-CLIP, our method only adds a learnable linear layer to the model structure, but it also has a significant improvement. This shows that our method does not rely solely on the generalization of the pre-trained model to achieve good performance. Figure 3 illustrates the decreasing trend of average accuracy as the number of classes increases for different datasets and settings. Our method can significantly mitigate forgetting.

The three prompt-based methods suffer from a severe imbalance of prompt usage in the base50 and base100 settings because they use the same number of prompts for each task by default. This leads to predictions of the model being

Table 4: Results on CUB200 (B0 Inc20) and ImageNet1K (B0 Inc100).

		PROOF	L2P ++	DualPrompt	CODA	Continual-CLIP	SLCA	ADAM-Adapter	ours
CUB200	Avg	83.11	71.90	71.74	66.61	60.60	73.30	78.80	<u>83.04</u>
	Last	<u>75.53</u>	62.99	62.14	50.88	51.16	60.39	70.61	76.34
ImageNet1K	Avg	76.23	79.30	<u>79.39</u>	76.99	72.96	79.10	76.60	81.73
	Last	65.26	69.60	<u>69.79</u>	66.96	64.44	68.27	68.74	72.58

biased towards the base classes, making it hard to inject new knowledge. L2P and DualPrompt have surpassed the performance of the original model pre-trained on ImageNet21k on the imageNet-R dataset. However, CODA performs poorly on the three datasets, suggesting that its method may be highly coupled with a specific pre-trained model, such as ViT pre-trained on Imagenet21k.

SLCA achieves better performance on the initial task by fine-tuning the backbone and the classifier. Consequently, its experimental results in the base50 and base100 settings are relatively good. However, the accumulated forgetting caused by fine-tuning the backbone becomes more obvious in long-sequence settings.

ADAM mainly trains an adapter model on the initial task, and then does not train the model on the subsequent tasks. Therefore, its performance depends on the proportion of data in the initial task. Since the model is not trained on the subsequent tasks, it maintains the model stability but does not learn new knowledge. The low plasticity of the model affects the performance. Due to the large image style difference in ImageNet-R, the dataset requires the model to learn more knowledge to adapt to the rich image style, and this method performs worse on this dataset than on the unified style cifar100 and imagenet100.

PROOF, designed for CLIP, expands the adapter with each task and uses cross-modal attention to fuse information from text and images. However, it does not consider the effect of neighboring categories in CLIP classification. Even without exemplar and extended parameters, our approach still has advantages.

We also evaluate our approach on the large dataset ImageNet1K and the fine-grained dataset CUB200. The results are shown in Tab. 4. The experimental results show that our method still has advantages in the large data set. Especially on the CUB200, a more difficult fine-grained dataset, the performance of our method is much improved compared to the zero-shot performance of the CLIP.

In summary, our method does not expand the model as the task increases, while maintaining the learning of new data and alleviating forgetting in the incremental tasks. More results are provided in the supplementary material.

5.3 Ablation Studies and Other Analysis

Module ablation. Table 5 shows the results of different components in our method. Random means random selection of class pairs, rather than using the distance of the textual feature. In this case, the accuracy is very close to the baseline, while using the full module with textual features improves the accuracy by 2.24% over the baseline. This indicates that the text-guided selection of

Table 5: Module ablation results on B0 Inc10 ImageNet100. SG denotes sampling a few old category features from the Gaussian distribution. $\mathcal{L}_{hinge}(\text{random})$ is random selection of class pairs for \mathcal{L}_{hinge} . The meaning of PF w/o MD is to use only Eq. (7) for parameter fusion and MD means matrix decomposition.

Ablation	Last Accuracy \uparrow
Adapter-finetune + SG (Baseline)	73.80
Baseline + $\mathcal{L}_{hinge}(\text{random})$	74.08
Baseline + \mathcal{L}_{hinge}	76.04
Baseline + \mathcal{L}_{hinge} + PF w/o MD	79.28
Baseline + \mathcal{L}_{hinge} + PF w/ MD (Full)	80.23

Table 6: The prediction results of the model under different ablation experiment settings for 50 test images of “kingsnake”. Three other snakes are the adjacent new categories that are selected by text feature similarity. More similar results are provided in the supplementary material.

	w/o \mathcal{L}_{hinge}	w/ \mathcal{L}_{hinge}
kingsnake	25	35
night snake	11	8
worm snake	2	1
eastern hog-nosed snake	10	5
others	2	1
accuracy	0.5	0.7

the nearest class is effective. The performance of the parameter fusion method without matrix decomposition and the full parameter fusion method with matrix decomposition shows that the parameter fusion method with fine-grained importance matrix is effective and matrix decomposition further enhances it.

Training Cost Analysis. Figure 4 illustrates the comparison of our incremental parameter size with other methods on ImageNet100. Using the same backbone, freezing the backbone parameters has much lower update costs than fine-tuning the entire model, as done by SLCA. Among the methods that freeze the backbone and add learnable parameters, our method adds the fewest parameters. Although we add a text encoder to the pure visual encoder method, because the label text is fixed, we only need to calculate the text features of the labels once during the entire training process and the number of labels is far lower than the data volume. Therefore, the cost of the text encoder

is negligible compared to the visual encoder, which processes many images and multiple iterations. For decomposed parameter fusion, we decompose the matrix only once after each task training. It is negligible compared to the training time.

Neighboring categories classification. As shown in Fig. 5, our approach effectively corrects old categories misclassified into new categories. Figure 5c shows that the number of samples misclassified to the new category decreases, the number of samples correctly classified increases, and the negative effect on other categories is small. Table 6 shows some examples of the affected categories.

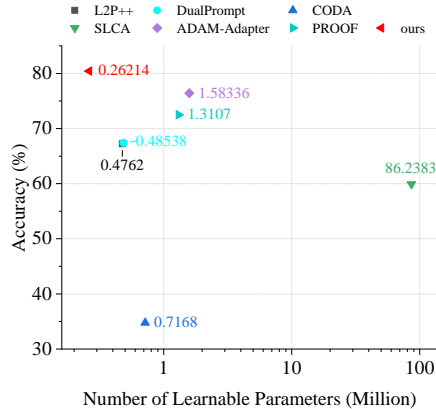


Fig. 4: Comparison of methods in terms of accuracy and learnable parameters.

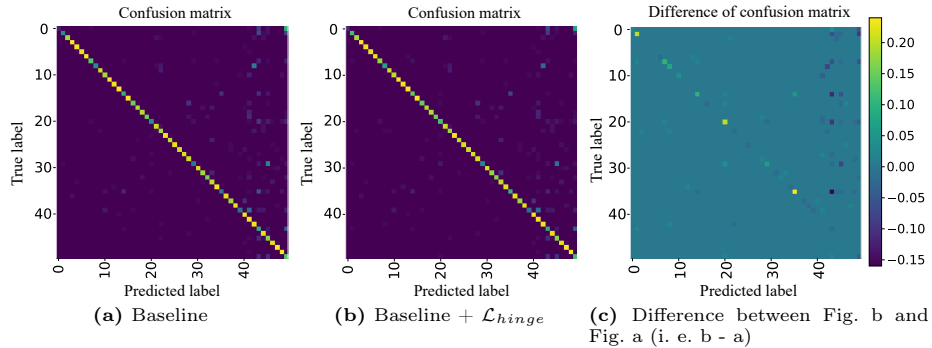


Fig. 5: The confusion matrixes of the first 5 tasks in the ImageNet100 B0 Inc10 experiment and their difference. We only show the first 5 tasks for better readability.

It shows that the new categories that are close to the old categories selected by the similarity of the features of the text are the main cause of the wrong prediction of the old affected categories. The selected categories account for 23 of 25 erroneous predictions. See the supplementary material for more similar examples. The \mathcal{L}_{hinge} used for the selected categories can effectively reduce the erroneous prediction of the affected old categories.

Threshold ablation. In Fig. 6, we study the effect of threshold α for adjacent categories. Since most distances are greater than 0.5, we gradually increase the threshold from 0.5. As the threshold increases, more adjacent categories are selected, improving performance. However, a threshold that is too large will select many pairs of dissimilar categories, which will interfere with the classification and increase the computation. There is no benefit to choosing more different classes. So we use 0.65 as the threshold for all experiments.

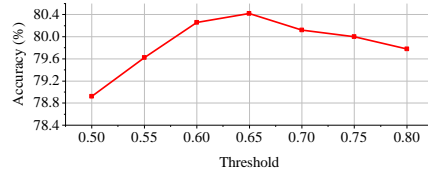


Fig. 6: Experiment with different thresholds α on the ImageNet100 B0 Inc10.

Prompt-based methods with CLIP text encoder. For a fair comparison, we add the text encoder of CLIP to the classifier-independent prompt-based methods as a classifier. The results are shown in Tab. 7. The key and value of different tasks are relatively independent in L2P++ and DualPrompt, resulting in minimal change in the representation of the old classes. Using fixed CLIP text features instead of a trainable linear layer as the classifier may cause misclassification of the samples of old classes when the new class text feature is similar to the old class. Hence, the performance of these two methods drops when using the CLIP text encoder. CODA uses a weight to combine all prompt components enabling the old class to benefit from the newly expanded parameters. Performance is improved when using the information-rich text classifier. Compared with these methods, ours can make better use of the CLIP text encoder.

Table 7: Last accuracy under the B0 INC20 setting on ImageNet-R.

Backbone	CODA	DualPrompt	L2P++
w/o text encoder	67.52	75.77	75.98
w text encoder	69.93	72.30	71.97

Table 8: Results of different sampling under the B0 INC20 setting on CIFAR100.

Method	PASS	Napa-vq	ours
Avg	84.38	84.23	86.19
Last	77.18	77.22	79.04

Table 9: Comparison results of conventional methods on ImageNet-R B0 Inc20. The results of conventional methods are reproduced from PILOT [35] with the pre-trained weights on ImageNet21K.

Method	Coil [51]	DER [47]	iCaRL [31]	FOSTER [42]	ours
Avg	80.48	81.16	72.76	82.49	85.58
Last	73.12	75.10	61.62	76.00	82.28

Different ways to sample features for training. PASS [55] and Napa-vq [25] are not designed based on pre-trained models and cannot be directly applied to pre-trained models. Therefore, we replace the feature sampling and class separation process in our method with the PASS [55] and Napa-vq [25] sampling methods as shown in Tab. 8. PASS overlooks adjacent classes, while Napa-vq depends more on Na-vq modeling from scratch. Our strategy utilizes the CLIP model more effectively.

5.4 Comparison with Conventional Methods

Conventional methods, which did not use pre-trained models but with exemplars, also achieved good results in class incremental learning. In Tab. 9, we compare conventional methods that use the ViT-B/16 pre-trained on ImageNet21K for initialization. We have also evaluated these baselines with the same initialization as ours, but the performance is much worse. Our method still has a significant advantage over these state-of-the-art methods even without exemplar.

6 Conclusion

In this paper, we study the problem of incremental learning with pre-trained vision-language models. We find that introducing textual features of classes to adjust the representations of classes that are greatly affected by new data can effectively alleviate forgetting. Moreover, a simple linear adapter with a parameter fusion strategy can efficiently maintain model stability and reduce forgetting. Experiments demonstrate the effectiveness of our method. Manual threshold selection is a limitation. Future work can design a mechanism to dynamically adapt the threshold and a mechanism to fuse the parameters more efficiently. The mutual influence of text and image can also be further explored.

Acknowledgements

This work is funded by NSFC (NO. 62206135), Young Elite Scientists Sponsorship Program by CAST (NO. 2023QNR001), Tianjin Natural Science Foundation (NO. 23JCQNJC01470), and the Fundamental Research Funds for the Central Universities (Nankai University). Computation is supported by the Supercomputing Center of Nankai University.

References

1. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: *Eur. Conf. Comput. Vis.* pp. 139–154 (2018)
2. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H., Ranzato, M.: On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486* (2019)
3. Choi, Y., El-Khamy, M., Lee, J.: Dual-teacher class-incremental learning with data-free generative replay. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 3543–3552 (2021)
4. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(7), 3366–3385 (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 248–255. *Ieee* (2009)
6. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: *Eur. Conf. Comput. Vis.* pp. 86–102. *Springer* (2020)
7. Douillard, A., Ramé, A., Couairon, G., Cord, M.: Dytox: Transformers for continual learning with dynamic token expansion. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 9285–9295 (2022)
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.* pp. 1–15 (2023)
9. Gao, Q., Zhao, C., Ghanem, B., Zhang, J.: R-dfcil: Relation-guided representation learning for data-free class incremental learning. In: *Eur. Conf. Comput. Vis.* pp. 423–439. *Springer* (2022)
10. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10), 3614–3631 (2020)
11. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. In: *Handbook for Automatic Computation: Volume II: Linear Algebra*, pp. 134–151. *Springer* (1971)
12. Hayes, T.L., Kanan, C.: Lifelong machine learning with deep streaming linear discriminant analysis. In: *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.* pp. 220–221 (2020)
13. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. *Int. Conf. Comput. Vis.* (2021)

14. Khan, M.G.Z.A., Naeem, M.F., Van Gool, L., Stricker, D., Tombari, F., Afzal, M.Z.: Introducing language guidance in prompt-based continual learning. In: *Int. Conf. Comput. Vis.* pp. 11463–11473 (2023)
15. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* **114**(13), 3521–3526 (2017)
16. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
17. Kumar, S., Zaidi, H.: Gdc-generalized distribution calibration for few-shot learning. *arXiv preprint arXiv:2204.05230* (2022)
18. Lee, K.Y., Zhong, Y., Wang, Y.X.: Do pre-trained models benefit equally in continual learning? In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 6485–6493 (2023)
19. Li, Z., Hoiem, D.: Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(12), 2935–2947 (2017)
20. Li, Z., Hoiem, D.: Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(12), 2935–2947 (2017)
21. Liu, X., Cao, X., Lu, H., Xiao, J.w., Bagdanov, A.D., Cheng, M.M.: Class incremental learning with pre-trained vision-language models. *arXiv preprint arXiv:2310.20348* (2023)
22. Liu, Y., Su, Y., Liu, A.A., Schiele, B., Sun, Q.: Mnemonics training: Multi-class incremental learning without forgetting. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 12245–12254 (2020)
23. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. *Adv. Neural Inform. Process. Syst.* **30** (2017)
24. Luo, Z., Liu, Y., Schiele, B., Sun, Q.: Class-incremental exemplar compression for class-incremental learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 11371–11380 (2023)
25. Malepathirana, T., Senanayake, D., Halgamuge, S.: Napa-vq: Neighborhood-aware prototype augmentation with vector quantization for continual learning. In: *Int. Conf. Comput. Vis.* pp. 11674–11684 (2023)
26. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., Van De Weijer, J.: Class-incremental learning: survey and performance evaluation on image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(5), 5513–5533 (2022)
27. McDonnell, M.D., Gong, D., Parvaneh, A., Abbasnejad, E., van den Hengel, A.: Ranpac: Random projections and pre-trained models for continual learning. *Adv. Neural Inform. Process. Syst.* **36** (2024)
28. Mehta, S.V., Patil, D., Chandar, S., Strubell, E.: An empirical investigation of the role of pre-training in lifelong learning. *arXiv preprint arXiv:2112.09153* (2021)
29. Ostapenko, O., Lesort, T., Rodriguez, P., Arefin, M.R., Douillard, A., Rish, I., Charlin, L.: Continual learning with foundation models: An empirical study of latent replay. In: *Conference on Lifelong Learning Agents*. pp. 60–91. PMLR (2022)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *Int. Conf. Mach. Learn.* pp. 8748–8763. PMLR (2021)
31. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2001–2010 (2017)

32. Sarfraz, F., Arani, E., Zonooz, B.: Error sensitivity modulation based experience replay: Mitigating abrupt representation drift in continual learning. In: *Int. Conf. Learn. Represent.* (2023), <https://openreview.net/forum?id=zlbc17019Z3>
33. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 11909–11919 (2023)
34. Smith, J.S., Tian, J., Halbe, S., Hsu, Y.C., Kira, Z.: A closer look at rehearsal-free continual learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2409–2419 (2023)
35. Sun, H.L., Zhou, D.W., Ye, H.J., Zhan, D.C.: Pilot: A pre-trained model-based continual learning toolbox. *arXiv preprint arXiv:2309.07117* (2023)
36. Sun, Z., Mu, Y., Hua, G.: Regularizing second-order influences for continual learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 20166–20175 (2023)
37. Tang, Y.M., Peng, Y.X., Zheng, W.S.: When prompt-based incremental learning does not meet strong pretraining. In: *Int. Conf. Comput. Vis.* pp. 1706–1716 (2023)
38. Thengane, V., Khan, S., Hayat, M., Khan, F.: Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114* (2022)
39. Van Ness, J.: On the dominance of non-parametric bayes rule discriminant algorithms in high dimensions. *Pattern Recognition* **12**(6), 355–368 (1980)
40. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
41. Wang, E., Peng, Z., Xie, Z., Yang, F., Liu, X., Cheng, M.M.: Unlocking the multi-modal potential of clip for generalized category discovery (2024), <https://arxiv.org/abs/2403.09974>
42. Wang, F.Y., Zhou, D.W., Ye, H.J., Zhan, D.C.: Foster: Feature boosting and compression for class-incremental learning. In: *Eur. Conf. Comput. Vis.* pp. 398–414. Springer (2022)
43. Wang, Y., Huang, Z., Hong, X.: S-prompts learning with pre-trained transformers: An occam’s razor for domain incremental learning. *Adv. Neural Inform. Process. Syst.* **35**, 5682–5695 (2022)
44. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: *Eur. Conf. Comput. Vis.* pp. 631–648. Springer (2022)
45. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 139–149 (2022)
46. Wu, T.Y., Swaminathan, G., Li, Z., Ravichandran, A., Vasconcelos, N., Bhotika, R., Soatto, S.: Class-incremental learning with strong pre-trained models. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 9601–9610 (2022)
47. Yan, S., Xie, J., He, X.: Der: Dynamically expandable representation for class incremental learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 3014–3023 (2021)
48. Zhang, G., Wang, L., Kang, G., Chen, L., Wei, Y.: Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *arXiv preprint arXiv:2303.05118* (2023)
49. Zhang, J., Zhang, J., Ghosh, S., Li, D., Tasci, S., Heck, L., Zhang, H., Kuo, C.C.J.: Class-incremental learning via deep model consolidation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* pp. 1131–1140 (2020)

50. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. In: *Int. Conf. Comput. Vis.* pp. 19125–19136 (2023)
51. Zhou, D.W., Ye, H.J., Zhan, D.C.: Co-transport for class-incremental learning. In: *ACM Int. Conf. Multimedia.* pp. 1645–1654 (2021)
52. Zhou, D.W., Ye, H.J., Zhan, D.C., Liu, Z.: Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *arXiv preprint arXiv:2303.07338* (2023)
53. Zhou, D.W., Zhang, Y., Ning, J., Ye, H.J., Zhan, D.C., Liu, Z.: Learning without forgetting for vision-language models. *arXiv preprint arXiv:2305.19270* (2023)
54. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**(9), 2337–2348 (2022)
55. Zhu, F., Zhang, X.Y., Wang, C., Yin, F., Liu, C.L.: Prototype augmentation and self-supervision for incremental learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 5871–5880 (2021)
56. Zhu, K., Zhai, W., Cao, Y., Luo, J., Zha, Z.J.: Self-sustaining representation expansion for non-exemplar class-incremental learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 9296–9305 (2022)