Collaborative Adapter Experts for Class-Incremental Learning

Sunyuan Qiang, Xinxing Yu, Yanyan Liang, *Member, IEEE*, Jun Wan, *Senior Member, IEEE*, and Du Zhang, *Senior Member, IEEE*

***Abstract*—Pre-trained models (PTMs) with parameter-efficient fine-tuning (PEFT) techniques have been extensively utilized in class-incremental learning (CIL) scenarios. However, they still remain susceptible to performance degradation as the individual PEFT module operates as an independent learning entity during the incremental process. To this end, this work proposes a novel class-incremental collaborative adapter experts (CICAE) model, which incorporates multiple adapters operating collaboratively to facilitate CIL. Specifically, our model primarily consists of two phases. Initially, multiple adapters are employed to establish a multi-expert system aimed at acquiring diverse incremental knowledge. Through the collaborative knowledge sharing (CKS) mechanism, the expertise of each adapter expert is transferable, promoting collaborative development and mutual advancement. Subsequently, with the category prototype distributions, collaborative classifier alignment (CCA) is proposed to further align the classifiers with the representation space in a cooperative manner. Extensive experiments on CIL benchmarks validate the superior performance of our model.**

***Index Terms*—Class-incremental learning, pre-trained models, parameter-efficient fine-tuning, collaborative learning.**

I. Introduction

REMARKABLE successes have been achieved via deep learning in diverse areas such as computer vision and natural language processing. Nevertheless, in real-world scenarios [1], [2] with dynamic changes, learning is not a one-time process but is instead an ongoing one. Class-incremental learning (CIL) [3], [4], [5] specifically addresses the scenario where a model is trained to learn new classes over time while retaining the knowledge of previously learned classes.

Received 16 December 2024; revised 15 March 2025; accepted 16 March 2025. Date of publication 20 March 2025; date of current version 16 April 2025. This work was supported in part by Beijing Natural Science Foundation under Grant JQ23016, in part by the Science and Technology Development Fund of Macau Project under Grant 0070/2020/AMJ, Grant 0096/2023/RIA2, Grant 0044/2024/AGJ, and Grant 0123/2022/A3, and in part by Guangdong Provincial Key R&D Program under Grant 2019B010148001. The associate editor coordinating the review of this article and approving it for publication was Prof. Xuesong Wang. *(Corresponding author: Yanyan Liang.)*

Sunyuan Qiang, Xinxing Yu, Yanyan Liang, and Du Zhang are with the Macau University of Science and Technology, Macau 999078, China (e-mail: 3220004460@student.must.edu.mo; starsyu0206@gmail.com; yyliang@must.edu.mo; duzhang@must.edu.mo).

Jun Wan is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Macau University of Science and Technology, Macau 999078, China (e-mail: jun.wan@ia.ac.cn).

Digital Object Identifier 10.1109/LSP.2025.3553431

However, one key issue of CIL is catastrophic forgetting (CF) [6]. Upon training a model with newly coming classes, it tends to overwrite the knowledge related to the old ones, leading to a significant performance degradation on previously learned tasks. To alleviate this issue, many methods have been developed. Regularization methods [7], [8], [9] mitigate CF by adding penalty terms to limit abrupt variations in parameters. Replay methods [10], [11], [12] maintain knowledge by training jointly with old samples or generated data. Dynamic architecture methods [13], [14], [15], [16] allocate extra modules to adapt the new tasks while retaining the performance of the old tasks. Recently, a novel class of approach [17], [18], [19], [20] has surfaced that uses pre-trained knowledge with parameter-efficient fine-tuning (PEFT), significantly boosting the performance of CIL. Our proposed model also aligns with this promising direction of pre-trained models (PTMs), and further introduces collaborative multiple adapters.

Notably, despite PTMs based CIL methods benefit from the extensive generalization of pre-training, performing individually training within each PEFT module inevitably limits the potential for cooperation. Recent studies [16], [21] attempted in building multiple experts [22]. However, they are trained independently at each task or modules are selectively activated, resulting in insufficient knowledge sharing and limited performance. Some others [23], [24] employ low-quality networks, or even require task IDs and sample memory, which not only constrains model's performance but also limits the applications. Instead, we devise a multi-expert framework with PTMs for CIL by facilitating the sharing and integration of knowledge collaboratively. Meanwhile, our method avoids both the continual parameters increase and the preallocation of extensive inactive parameter spaces. The characteristics of adapter allow for the allocation of lightweight experts without incurring a substantial parameter overhead.

To this end, we introduce the class-incremental collaborative adapter experts (CICAE) model, which utilizes multiple experts to collaboratively and continuously update category knowledge for CIL. Specifically, each expert module employs adapter PEFT technique, which facilitates rapid adaptation to downstream tasks while conserving parameter usage. Subsequently, collaborative knowledge sharing (CKS) loss function is introduced among experts to enable collaborative progression for CIL. Finally, collaborative classifier alignment (CCA) is proposed to align classifiers with sampled category features collaboratively via CCA loss function.

The main contributions of this work include:

- We propose multi-expert collaborative adapters for CIL to concurrently learn multiple adapters.

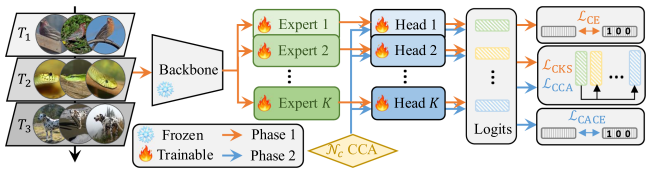

Fig. 1. Illustration of CICAE model. At each task, the model learns multiple adapters and heads with frozen backbone using $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{CKS}}$ in phase 1, thereby collaboratively facilitating the acquisition of incremental task knowledge. The CCA leverages the normal distribution $\mathcal{N}_c$ of features for all classes to collaboratively retrain the heads using $\mathcal{L}_{\text{CACE}}$ and $\mathcal{L}_{\text{CCA}}$ in phase 2, thereby achieving alignment between features and all-class categories.

- We present collaborative knowledge share (CKS), which enables each expert model to acquire collaborative information from others and advance together.
- We introduce collaborative classifier alignment (CCA), allowing each expert classifier head to further collaboratively align the prototype feature spaces.

II. Method

A. Notations

In CIL [10], at each task t, only current dataset $(\mathbf{x}, y) \sim \mathcal{D}_t$ is available, where $\mathbf{x}$ and y are the sample pair from data and target space $\mathcal{X}_t$ and $\mathcal{Y}_t$. The model is required to classify all previously learned tasks $\mathcal{D}_{1:t}$. Note that label spaces at different tasks are assumed to be non-overlapping, i.e., $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ for $i \neq j$. $|\mathcal{Y}_t|$ denotes the number of class at task t.

B. Class-Incremental Collaborative Adapter Experts Model

The overall architecture of our method is shown in Fig. 1. This section describes each component in detail and the entire training process is summarized in Section II-B5.

1) Individual Adapter Tuning: Individual learning by each expert is the pivotal foundation for synergistically progressing towards our final multi-expert model. We employ adapter [25], known for its stable performance in CIL [26], to fine-tune the model and learn task-specific knowledge at each task t.

$$\mathbf{x}'_\ell = \text{FFN}(\mathbf{x}_\ell) + s \cdot \text{ReLU}\left(\mathbf{x}_\ell \cdot \mathbf{W}_{\text{down}}\right) \cdot \mathbf{W}_{\text{up}}, \tag{1}$$

where $\mathbf{x}_\ell$ denotes feature at ℓ-th attention block given an input image $\mathbf{x}$. $\mathbf{W}_{\text{down}}$ and $\mathbf{W}_{\text{up}}$ are additional learnable parameters. Then, margin cosine cross entropy loss is used for training objective and enhance the clustering quality of the features.

$$\mathcal{L}_{\text{CE}} = -\log \frac{e^{s(\cos(\mathbf{w}_i, \mathbf{z}) - m)}}{e^{s(\cos(\mathbf{w}_i, \mathbf{z}) - m)} + \sum_{c \neq i}^{|\mathcal{Y}_t|} e^{s(\cos(\mathbf{w}_c, \mathbf{z}))}}, \tag{2}$$

where cos denotes the cosine similarity. s and m denote the scale and margin factor, respectively. $\mathbf{w}_i$ represents the weight of the i-th class in the classifier $\mathbf{W}_{\text{cls}}$ head. $\mathbf{z} = F_{\theta,\phi}(\mathbf{x})$ is the output feature from backbone model θ and adapter module ϕ.

2) Multiple Collaborative Adapters: To scale the learning of an individual adapter to a multi-expert collaborative learning process, we introduce multiple collaborative adapters module. Formally, give backbone model θ, we developed K adapters parameterized by $\{\phi_k\}_{k=1}^K$. In contrast to the expandable adapters [16], we avoid the issue of parameters increasing as tasks grow. Then, features and prediction logits are as follows.

$$\{\mathbf{z}_k\}_{k=1}^K = \{F_{\theta,\phi_k}(\mathbf{x})\}_{k=1}^K, \{\mathbf{l}_k\}_{k=1}^K = \{\mathbf{W}_{\text{cls}_k} \cdot \mathbf{z}_k\}_{k=1}^K. \tag{3}$$

However, owing to the inherently limited parameter count of adapter modules designed for PEFT, the diversity of knowledge within each expert is significantly constrained. Directly increasing experts and applying ensemble show no substantial gains in performance, as illustrated in Fig. 2(a). Instead, we propose to assign different learning rates (LRs) to the adapters, thereby enhancing the diversity among experts. Intuitively, this strategy allows experts with higher LRs to efficiently absorb new knowledge while those with lower LRs help stabilize and retain existing knowledge, as in Table VI. The objective function remains CE loss for each expert in (2), i.e., $\mathcal{L}_{\text{CE}}^k$.

3) Collaborative Knowledge Share: In last subsection, each expert is still learning in isolation. To enable cooperative learning among multiple adapter experts, collaborative knowledge share (CKS) is proposed, allowing each expert to assimilate additional knowledge from its counterparts. Based on the logits in (3), the probability and CKS loss can be expressed as

$$\mathbf{p}_{k,c} = \frac{e^{\mathbf{l}_{k,c}}}{\sum_j e^{\mathbf{l}_{k,j}}}, \quad \mathcal{L}_{\text{CKS}} = \sum_{k_1}^{K} \sum_{k_2, k_2 \neq k_1}^{K} \text{KL}\left(\mathbf{p}_{k_1} \| \mathbf{p}_{k_2}\right), \tag{4}$$

where c denotes the target class index. $\mathbf{l}_{k,j}$ denotes the j-th logits element of k-th expert. KL denotes the Kullback Leibler divergence, which is employed to ensure the consistency of the probability space handled by CE loss and to extract the latent knowledge from the logits. The knowledge of each adapter maintains a unified foundation based on the pre-trained space, while simultaneously exhibiting task-specific diversity. With the proposed CKS module, this knowledge is collaboratively integrated during the fine-tuning process, thereby facilitating the effective performance of CIL.

4) Collaborative Classifier Alignment: We extend CA [26], [27] into a multi-expert learning framework with collaboration for all-class alignment, enabling classifiers to mutually learn from each another and improve overall model performance. Specifically, given class prototypes $\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} \mathbf{z}_n$ and covariance $\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} (\mathbf{z}_n - \boldsymbol{\mu}_c)(\mathbf{z}_n - \boldsymbol{\mu}_c)^\top$ derived from the feature data $\mathbf{z}$ of the current task after incremental training, multidimensional feature distributions $\mathcal{N}_{\mathcal{Y}_t} = \{\mathcal{N}_c(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\}_{c \in \mathcal{Y}_t}$ can be formulated for each class $c \in \mathcal{Y}_t$ in current task, which are subsequently integrated with the distributions $\mathcal{N}_{\mathcal{Y}_{1:t-1}}$ from previous tasks, thereby enabling the extraction of all-class features $\hat{\mathbf{z}}$ to mitigate the unavailability of old data in CIL and facilitate classifier alignment (both old and new classes), with the training objective defined as

$$\mathcal{L}_{\text{CACE}} = -\log \frac{e^{\cos(\mathbf{w}_i, \hat{\mathbf{z}})/\tau}}{\sum_{c=1}^{|\mathcal{Y}_t|} e^{\cos(\mathbf{w}_c, \hat{\mathbf{z}})/\tau}}. \tag{5}$$

Note that $\mathcal{L}_{\text{CACE}}$ is consistently utilized for each expert, i.e., $\mathcal{L}_{\text{CACE}}^k$. Then, similar to (3) and (4), we foster collaborative progression and mutual knowledge integration among different classifiers, and the probability of k-th classifier and collaborative classifier alignment (CCA) loss can be expressed as

$$\hat{\mathbf{p}}_{k,c} = \frac{e^{\cos(\mathbf{w}_{k,c}, \hat{\mathbf{z}}_k)}}{\sum_j e^{\cos(\mathbf{w}_{k,j}, \hat{\mathbf{z}}_k)}}, \mathcal{L}_{\text{CCA}} = \sum_{k_1}^{K} \sum_{k_2, k_2 \neq k_1}^{K} \text{KL}\left(\hat{\mathbf{p}}_{k_1} \| \hat{\mathbf{p}}_{k_2}\right), \tag{6}$$

where $\mathbf{w}_{k,j}$ denotes the j-th class weight of k-th expert. $\hat{\mathbf{z}}_k$ are sampled feature data from k-th expert's normal distribution.

5) Summary: The overall framework of our model is shown in Fig. 1, which mainly consists two phases. In the first phase, collaborative fine-tuning of multiple experts is performed. We

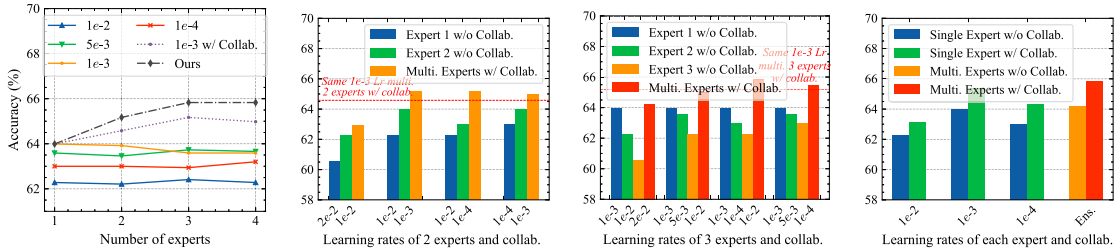

Fig. 2. Ablations results of the last accuracy $\mathcal{A}_{Last}$ on ImageNet-A. (a) Same LRs ensemble and our model with different number of experts. (b) Different LRs for each expert with 2 experts. (c) Different LRs for each expert with 3 experts. (d) Performance of single and multi. experts on collaboration.

TABLE I
PERFORMANCE COMPARISON ON IMAGENET-R, IMAGENET-A, CUB200, AND CIFAR100 DATASETS

Method	Venue	Params	ImageNet-R		ImageNet-A		CUB200		CIFAR100	
			$\mathcal{A}_{Last}$(%)	$\mathcal{A}_{Avg}$(%)	$\mathcal{A}_{Last}$(%)	$\mathcal{A}_{Avg}$(%)	$\mathcal{A}_{Last}$(%)	$\mathcal{A}_{Avg}$(%)	$\mathcal{A}_{Last}$(%)	$\mathcal{A}_{Avg}$(%)
Joint	-	86M	$81.72_{\pm 0.35}$	-	$50.56_{\pm 1.75}$	-	$88.17_{\pm 0.32}$	-	$89.71_{\pm 0.07}$	-
Fine-Tune	-	86M	$20.93_{\pm 0.86}$	$40.35_{\pm 0.74}$	$06.03_{\pm 4.74}$	$16.57_{\pm 5.80}$	$22.05_{\pm 1.69}$	$45.67_{\pm 2.04}$	$22.17_{\pm 1.09}$	$41.83_{\pm 1.60}$
SLCA [27]	ICCV'23	86M	$79.35_{\pm 0.28}$	$83.29_{\pm 0.46}$	$61.05_{\pm 0.63}$	$68.88_{\pm 2.31}$	$84.68_{\pm 0.09}$	$90.77_{\pm 0.79}$	$91.26_{\pm 0.37}$	$94.29_{\pm 0.92}$
ADAM-Adapter [18]	IJCV'24	1.19M	$65.79_{\pm 0.98}$	$72.42_{\pm 1.41}$	$48.81_{\pm 0.08}$	$58.84_{\pm 1.37}$	$85.84_{\pm 0.08}$	$91.33_{\pm 0.49}$	$87.29_{\pm 0.27}$	$91.21_{\pm 1.33}$
ADAM-SSF [18]	IJCV'24	0.20M	$66.61_{\pm 0.09}$	$74.36_{\pm 1.00}$	$48.94_{\pm 0.14}$	$58.79_{\pm 2.82}$	$85.67_{\pm 0.15}$	$90.99_{\pm 0.76}$	$85.27_{\pm 0.21}$	$89.90_{\pm 0.98}$
ADAM-VPT [18]	IJCV'24	0.04M	$65.29_{\pm 1.52}$	$72.97_{\pm 0.56}$	$29.29_{\pm 7.42}$	$39.14_{\pm 7.59}$	$85.28_{\pm 0.47}$	$90.89_{\pm 0.86}$	$85.04_{\pm 1.04}$	$89.49_{\pm 0.58}$
LAE [28]	ICCV'23	0.19M	$72.29_{\pm 0.14}$	$77.99_{\pm 0.46}$	$47.18_{\pm 1.17}$	$58.15_{\pm 0.73}$	$80.97_{\pm 0.51}$	$87.22_{\pm 1.21}$	$85.25_{\pm 0.43}$	$89.80_{\pm 1.20}$
L2P [17]	CVPR'22	0.04M	$72.34_{\pm 0.17}$	$77.36_{\pm 0.64}$	$44.04_{\pm 0.93}$	$51.24_{\pm 2.26}$	$67.02_{\pm 1.90}$	$79.62_{\pm 1.60}$	$84.06_{\pm 0.88}$	$88.26_{\pm 1.34}$
DualPrompt [29]	ECCV'22	0.25M	$69.10_{\pm 0.62}$	$74.28_{\pm 0.66}$	$53.19_{\pm 0.74}$	$64.59_{\pm 0.08}$	$68.48_{\pm 0.47}$	$80.59_{\pm 1.50}$	$86.93_{\pm 0.24}$	$91.13_{\pm 0.32}$
CODAPrompt [30]	CVPR'23	3.84M	$73.31_{\pm 0.50}$	$78.47_{\pm 0.53}$	$52.08_{\pm 0.12}$	$63.92_{\pm 0.12}$	$77.23_{\pm 1.12}$	$81.90_{\pm 0.85}$	$83.21_{\pm 3.39}$	$87.71_{\pm 3.17}$
RanPAC [31]	NeurIPS'23	1.19M	$77.83_{\pm 0.10}$	$82.73_{\pm 0.36}$	$63.08_{\pm 0.48}$	$70.99_{\pm 1.20}$	$89.73_{\pm 0.12}$	$93.41_{\pm 0.27}$	$91.50_{\pm 0.11}$	$\underline{94.42}_{\pm 0.69}$
EASE [16]	CVPR'24	11M	$75.99_{\pm 0.16}$	$81.29_{\pm 0.42}$	$57.36_{\pm 0.48}$	$66.47_{\pm 1.19}$	$85.07_{\pm 1.35}$	$90.93_{\pm 1.08}$	$88.27_{\pm 0.44}$	$92.04_{\pm 0.81}$
SSIAT [26]	CVPR'24	1.19M	$79.38_{\pm 0.59}$	$83.63_{\pm 0.43}$	$62.43_{\pm 1.63}$	$70.83_{\pm 1.63}$	$88.75_{\pm 0.38}$	$93.00_{\pm 0.90}$	$91.35_{\pm 0.26}$	$94.35_{\pm 0.60}$
CPrompt [32]	CVPR'24	0.26M	$77.14_{\pm 0.11}$	-	-	-	-	-	$87.82_{\pm 0.21}$	-
InfLoRA [33]	CVPR'24	-	$75.65_{\pm 0.14}$	-	-	-	-	-	$87.06_{\pm 0.25}$	-
VPT-NSP^2 [34]	NeurIPS'24	-	$78.88_{\pm 0.50}$	-	-	-	-	-	$91.74_{\pm 0.63}$	-
VQPrompt [35]	NeurIPS'24	0.50M	$75.70_{\pm 0.13}$	$79.75_{\pm 0.35}$	$55.67_{\pm 0.14}$	$62.23_{\pm 2.08}$	$86.27_{\pm 0.27}$	$91.26_{\pm 0.55}$	$90.23_{\pm 0.21}$	$93.15_{\pm 0.88}$
CICAE (1 expert)	Ours	3.57M	$\underline{79.87}_{\pm 0.35}$	$\underline{84.04}_{\pm 0.37}$	$\underline{65.08}_{\pm 0.70}$	$\underline{72.16}_{\pm 1.72}$	$\underline{90.56}_{\pm 0.07}$	$\underline{94.16}_{\pm 0.49}$	$\underline{91.77}_{\pm 0.17}$	$94.23_{\pm 0.81}$
CICAE (3 expert)	Ours	3.57M	$\mathbf{80.39}_{\pm 0.34}$	$\mathbf{84.26}_{\pm 0.34}$	$\mathbf{65.57}_{\pm 0.57}$	$\mathbf{72.43}_{\pm 1.59}$	$\mathbf{90.72}_{\pm 0.10}$	$\mathbf{94.21}_{\pm 0.48}$	$\mathbf{91.94}_{\pm 0.18}$	$\mathbf{94.44}_{\pm 0.67}$

The best results are in bold and the second best results are underlined. We report the results for CICAE model with 1 and 3 expert inference, respectively.

freeze the backbone model and optimize the parameters of the adapters $\{\phi_k\}_{k=1}^{K}$ and the classifier weights $\{\mathbf{W}_{\text{cls}_k}\}_{k=1}^{K}$ to minimize the loss function $\mathcal{L}_1$. In the second phase, with the semantic shift [16], [26], [36] and category distributions, we exclusively optimize the weights $\{\mathbf{W}_{\text{cls}_k}\}_{k=1}^{K}$ for collaborative classifier alignment using the $\mathcal{L}_2$.

$$\mathcal{L}_1 = \sum_{k=1}^{K} \mathcal{L}_{\text{CE}}^{k} + \lambda_1 \cdot \mathcal{L}_{\text{CKS}}, \quad \mathcal{L}_2 = \sum_{k=1}^{K} \mathcal{L}_{\text{CACE}}^{k} + \lambda_2 \cdot \mathcal{L}_{\text{CCA}}, \tag{7}$$

where λ_1 and λ_2 are hyperparameters controlling the collaborative knowledge sharing. During the inference, predictions are made by directly summing the logits across multiple experts, $\hat{y} = \arg\max_{c \in \{1,2,\ldots,C\}} \sum_{k=1}^{K} \mathbf{l}_{k,c}$, where $\mathbf{l}_{k,c}$ is the logit value at the position of c-th class index and k-th adapter.

TABLE II
PERFORMANCE COMPARISON ON OBJECTNET AND OMNIBENCHMARK DATASETS

Method	ObjectNet B0 Inc10 (%)		Omnibenchmark B0 Inc30 (%)	
	$\mathcal{A}_{Avg}$	$\mathcal{A}_{Last}$	$\mathcal{A}_{Avg}$	$\mathcal{A}_{Last}$
Fine-tune	19.14	8.73	23.61	10.57
Fine-tune Adapter	50.22	35.95	62.32	50.53
LwF [7]	33.01	20.65	47.14	33.95
SDC [36]	39.04	29.06	60.94	50.28
L2P [17]	63.78	52.19	73.36	64.69
DualPrompt [29]	59.27	49.33	73.92	65.52
CODAPrompt [30]	66.07	53.29	77.03	68.09
SimpleCIL [18]	65.45	53.59	79.34	73.15
ADAM-FT [18]	61.41	48.34	73.02	65.03
ADAM-VPTS [18]	64.54	52.53	79.63	73.68
ADAM-VPTD [18]	67.83	54.65	81.05	74.47
ADAM-SSF [18]	69.15	56.64	80.53	74.00
ADAM-Adapter [18]	67.18	55.24	80.75	74.37
SLCA [27]	72.55	61.30	82.80	74.10
RanPAC [31]	76.73	65.98	85.44	79.13
EASE [16]	70.84	57.86	81.11	74.85
SSIAT [26]	76.46	65.93	84.41	77.63
CICAE (1 expert)	<u>76.89</u>	<u>66.29</u>	<u>85.50</u>	<u>79.72</u>
CICAE (3 expert)	**77.19**	**66.55**	**85.54**	**79.80**

III. EXPERIMENTAL RESULTS

Following previous works [26], [34], [35], the experiments are conducted on six CIL benchmark datasets, and a cross-domain CIL dataset, i.e., ImageNet-R [37], ImageNet-A [38], CUB200 [39], CIFAR100 [40], ObjectNet [41], OmniBenchmark [42], and VTAB [43]. The number of experts is set to 3. Basically, the LRs for each expert are set to $1e$-2, $2e$-2, and $1e$-3. The epoch is set to 20 and 10 in the first and the incremental tasks. The average accuracy $\mathcal{A}_{Avg}$ and the last accuracy $\mathcal{A}_{Last}$ are reported as quantitative metrics.

A. Comparisons With Prior Arts

1) Main Results: We compare our method with prior arts all on ImageNet-21 k pre-trained ViT model for a fair evaluation. Tables I and II show $\mathcal{A}_{Avg}$ and $\mathcal{A}_{Last}$ on six CIL benchmarks. Evidently, for metric $\mathcal{A}_{Last}$, our model exhibits state-of-the-art performance across all datasets, regardless of whether a single expert or an ensemble of three experts is used. Benefiting from the shared knowledge of 3 experts, our single expert inference already reaches a competitive performance. Notably, for ImageNet-A, we achieve a last accuracy of 65.57 and an average accuracy of 72.43, exceeding the previous best result of 63.08 (+2.49) and 70.99 (+1.44). Moreover, the training parameters of 3 experts remain acceptable (about 3.6 M), while offering noticeable performance enhancements.

2) Cross-Domain CIL Results: The accuracy of each task on the cross-domain VTAB dataset is shown in Table III. Our model achieves the superior performance via 3 experts, with an average accuracy of 95.02. Moreover, the average accuracy of single expert also exceed other methods, reaching 94.76.

TABLE III
PERFORMANCE COMPARISON ON CROSS-DOMAIN VTAB CIL DATASET

Method	$T_1(\%)$	$T_2(\%)$	$T_3(\%)$	$T_4(\%)$	$T_5(\%)$	Avg ↑
ADAM-Adapter [18]	87.60	86.07	89.14	82.72	84.35	85.97
ADAM-SSF [18]	89.60	88.21	89.94	80.50	82.38	86.13
ADAM-VPT [18]	90.20	87.57	89.69	80.39	82.18	86.01
SLCA [27]	94.80	92.43	93.54	93.98	94.33	93.82
LAE [28]	97.99	85.26	79.68	78.78	74.36	83.21
RanPAC [31]	96.30	93.07	93.84	92.36	91.97	93.50
EASE [16]	96.20	92.50	92.96	92.82	93.55	93.60
SSIAT [26]	96.10	92.71	94.09	93.68	94.50	94.21
CICAE (1 expert)	96.20	93.93	94.34	94.43	<u>94.93</u>	<u>94.76</u>
CICAE (3 expert)	96.20	94.43	94.38	94.71	**95.41**	**95.02**

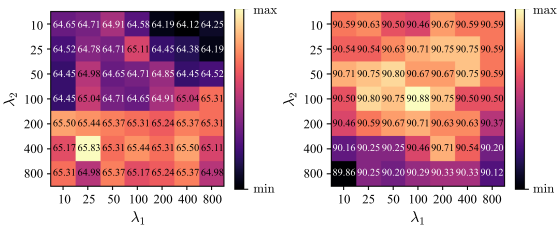

Fig. 3. Parameter analysis of λ_1, λ_2 on ImageNet-A and CUB200 datasets.

B. Components Analysis

1) Effect of Multi. Experts (W/o Collab. v.s. W/ Collab.): We summarize our proposed two KL collaborative loss functions as Collab. in figures. As in Fig. 2(a), although the multiple experts use the same $1e$-3 LR, employing the collaboration ($1e$ -3 w/ Collab), results in a further performance boost. Moreover, Fig. 2(d) shows that, with different LRs, the ensemble predictions from the multiple experts also demonstrate performance gains. Compared with individually trained ensembles without collaboration, our model facilitates knowledge sharing among multiple experts, resulting in performance improvement.

2) Effect of Single Expert (W/o Collab. v.s. W/ Collab.): As shown in Fig. 2(d), we also evaluate the performance gain of individual experts. Across all three single expert models, there is a noticeable and steady performance enhancement following the utilization of collaboration.

3) Same LRs v.s. Different LRs and Number of Experts: As illustrated in Fig. 2(a), employing the same LR leads to similar performance between single and multiple experts without collaboration. This phenomenon could potentially be attributed to a diminished diversity due to the same LRs, alongside the lack of knowledge transfer. Moreover, our model consistently outperforms baselines across varying expert numbers, with stable performance improvements as adapters increase from 1 to 3, demonstrating effective leveraging of multiple adapters. At 4 adapters, performance plateaus, indicating sufficient diversity in captured collaborative knowledge, and we directly set up 3 adapters to balance performance and parameter efficiency. Specifically, as shown in Fig. 2(b) and (c), whether utilizing two experts or three, an ensemble with appropriately different LRs can typically outperforms the optimal results of ensembles sharing the same LR (red dotted line).

4) Effect of Loss Weight: The hyper-parameters λ_1 and λ_2 are responsible for controlling the knowledge sharing among adapters and classifiers experts. From the results in Fig. 3, we determine that λ_1, λ_2 to 25, 400 and 100, 100 for ImageNet-A and CUB datasets, indicating a balance between individual model learning and knowledge transfer among experts.

5) Ablation Study of All Components: The ablation results for each component are provided in Table IV. Our baseline is directly constructed using 3 Adapters ensemble with default CA [26]. We report accuracy results of both 3 single experts and the multiple experts ensemble. The performance of components is consistently improved, validating the effectiveness of our proposed method. Particularly, we achieve last accuracy from 62.41 to 65.83 (+3.42) with multiple experts.

TABLE IV
ABLATIONS RESULTS OF ALL COMPONENTS ON IMAGENET-A

C.Adp.	CKS	CCA	Single $\mathcal{A}_{Avg}(\%)$	Single $\mathcal{A}_{Last}(\%)$	Multi. $\mathcal{A}_{Last}(\%)$
Vanilla Multi.-Adp.			$72.61_{\pm 0.14}$	$62.41_{\pm 0.10}$	62.41
✓			$72.69_{\pm 0.35}$	$63.09_{\pm 0.70}$	64.19
✓	✓		$72.90_{\pm 0.34}$	$63.28_{\pm 0.44}$	64.65
✓		✓	$73.60_{\pm 0.64}$	$64.21_{\pm 0.83}$	65.24
✓	✓	✓	$73.69_{\pm 0.49}$	$64.27_{\pm 0.91}$	**65.83**

TABLE V
PERFORMANCE COMPARISON ON DIFFERENT EXPERTS ARCHITECTURES

Method	$\mathcal{A}_{Avg}$	$\mathcal{A}_{Last}$
Incr Experts	68.15	57.60
Entropy Gate	73.29	63.86
CICAE (Ours)	74.53	65.83

TABLE VI
PERFORMANCE COMPARISON OF NEW AND OLD TASKS ON ADAPTERS

Adapter-idx	LR	ImageNet-A New	ImageNet-A Old
0	1e-2	72.58	62.29
1	1e-3	70.97	64.87
2	1e-4	68.55	63.94
Ens.	-	73.39	65.16

6) Expert Architecture: Table V shows the results of different expert architectures, including incremental experts and entropy gated MoE. The former continuously allocates modules as tasks increases, while the latter employs entropy to gate and select expert for prediction. It can be clearly observed that they remain limited in performance due to insufficient knowledge sharing, whereas our method attains superior results.

7) Adapters Preferences: Table VI reveals that adapters with different LRs inherently induce task-specific preferences: higher LR prioritizes plasticity, lower LR ensures stability, and minimal LR can be considered to stabilize the pre-trained knowledge. Their ensemble further amplifies mutual strengths, leading to enhanced overall performance.

IV. CONCLUSION

In this study, we propose the collaborative adapter experts (CICAE) model. Our multiple adapter-based experts model incorporates the collaborative knowledge sharing (CKS) that interlinks various experts, thereby enhancing mutual assistance and fostering a collective improvement. Collaborative classifier alignment (CCA) is further utilized to align classifier with the features cooperatively. Experimental results on CIL benchmark datasets validate the effectiveness of our model.

Future work: While our model demonstrates promising performance, several limitations remain. First, the current experiments are conducted on standard CIL benchmarks with relatively balanced task distributions. Future work should explore more complex CIL settings with highly imbalanced data or domain shifts. Second, the inference overhead introduced by multiple adapter experts may hinder deployment. Hence, developing lightweight variants of the collaborative model is the direction of our future efforts.

REFERENCES

[1] H. Wang, X. Liu, Z. Qiao, and H. Tao, "Inducing causal meta-knowledge from virtual domain: Causal meta-generalization for hyperspectral domain generalization," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5538416.

[2] H. Wang, X. Liu, Z. Qiao, G. Wang, and H. Chen, "Multimodal remote sensing data classification based on Gaussian mixture variational dynamic fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5621214.

[3] Y. Wu et al., "Large scale incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 374–382.

[4] B. Zheng, D. Zhou, H. Ye, and D. Zhan, "Multi-layer rehearsal feature augmentation for class-incremental learning," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 61649–61663.

[5] D. Zhou, Z. Cai, H. Ye, L. Zhang, and D. Zhan, "Dual consolidation for pre-trained model-based domain-incremental learning," 2024, arXiv:2410.00911.

[6] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Front. Psychol.*, vol. 4, 2013, Art. no. 504.

[7] Z. Li and D. Hoiem, "Learning without forgetting," in *Proc. Eur. Conf. Comput. Vis.*, 2016, vol. 9908, pp. 614–629.

[8] J. Zhu, G. Luo, B. Duan, and Y. Zhu, "Class incremental learning with deep contrastive learning and attention distillation," *IEEE Signal Process. Lett.*, vol. 31, pp. 1224–1228, 2024.

[9] S. Kolouri, A. Abbasi, S. A. Koohpayegani, P. Nooralinejad, and H. Pirsiavash, "Multi-agent lifelong implicit neural learning," *IEEE Signal Process. Lett.*, vol. 30, pp. 1812–1816, 2023.

[10] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5533–5542.

[11] S. Qiang, J. Hou, J. Wan, Y. Liang, Z. Lei, and D. Zhang, "Mixture uniform distribution modeling and asymmetric mix distillation for class incremental learning," in *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, pp. 9498–9506.

[12] W. Ding, H. Sun, C. Pei, D. Jia, and L. Huang, "Multi-organ registration with continual learning," *IEEE Signal Process. Lett.*, vol. 31, pp. 1204–1208, 2024.

[13] S. Yan, J. Xie, and X. He, "DER: Dynamically expandable representation for class incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3014–3023.

[14] F. Wang, D. Zhou, H. Ye, and D. Zhan, "FOSTER: Feature boosting and compression for class-incremental learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, vol. 13685, pp. 398–414.

[15] S. Qiang, Y. Liang, J. Wan, and D. Zhang, "Dynamic feature learning and matching for class-incremental learning," 2024, arXiv:2405.08533.

[16] D. Zhou, H. Sun, H. Ye, and D. Zhan, "Expandable subspace ensemble for pre-trained model-based class-incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 23554–23564.

[17] Z. Wang et al., "Learning to prompt for continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 139–149.

[18] D.-W. Zhou, Z.-W. Cai, H.-J. Ye, D.-C. Zhan, and Z. Liu, "Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need," *Int. J. Comput. Vis.*, vol. 133, pp. 1012–1032, 2025.

[19] D. Park, Y. Yu, D. Katabi, and H. K. Kim, "Adversarial continual learning to transfer self-supervised speech representations for voice pathology detection," *IEEE Signal Process. Lett.*, vol. 30, pp. 932–936, 2023.

[20] H. Sun, D. Zhou, H. Zhao, L. Gan, D. Zhan, and H. Ye, "MOS: Model surgery for pre-trained model-based class-incremental learning," 2024, arXiv:2412.09441.

[21] J. Yu et al., "Boosting continual learning of vision-language models via mixture-of-experts adapters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 23219–23230.

[22] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," in *Proc. 2nd Int. Conf. Learn. Representations, Workshop*, 2014.

[23] L. Wang, X. Zhang, Q. Li, J. Zhu, and Y. Zhong, "COSCL: Cooperation of small continual learners is stronger than a big one," in *Proc. Eur. Conf. Comput. Vis.*, 2022, vol. 13686, pp. 254–271.

[24] T. Doan, S. Mirzadeh, and M. Farajtabar, "Continual learning beyond a single model," in *Proc. Conf. Lifelong Learn. Agents*, 2023, vol. 232, pp. 961–991.

[25] S. Chen et al., "AdaptFormer: Adapting vision transformers for scalable visual recognition," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, pp. 16664–16678.

[26] Y. Tan, Q. Zhou, X. Xiang, K. Wang, Y. Wu, and Y. Li, "Semantically-shifted incremental adapter-tuning is a continual vitransformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 23252–23262.

[27] G. Zhang, L. Wang, G. Kang, L. Chen, and Y. Wei, "SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19091–19101.

[28] Q. Gao et al., "A unified continual learning framework with general parameter-efficient tuning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11449–11459.

[29] Z. Wang et al., "DualPrompt: Complementary prompting for rehearsal-free continual learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, vol. 13686, pp. 631–648.

[30] J. S. Smith et al., "Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11909–11919.

[31] M. D. McDonnell, D. Gong, A. Parvaneh, E. Abbasnejad, and A. van den Hengel, "RanPAC: Random projections and pre-trained models for continual learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2023, pp. 12022–12053.

[32] Z. Gao, J. Cen, and X. Chang, "Consistent prompting for rehearsal-free continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 28463–28473.

[33] Y. Liang and W. Li, "Inflora: Interference-free low-rank adaptation for continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 23638–23647.

[34] Y. Lu et al., "Visual prompt tuning in null space for continual learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024.

[35] L. Jiao, Q. Lai, Y. Li, and Q. Xu, "Vector quantization prompting for continual learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024.

[36] L. Yu et al., "Semantic drift compensation for class-incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6980–6989.

[37] D. Hendrycks et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8320–8329.

[38] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15262–15271.

[39] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," 2011.

[40] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[41] A. Barbu et al., "ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9448–9458.

[42] Y. Zhang, Z. Yin, J. Shao, and Z. Liu, "Benchmarking Omni-vision representation through the lens of visual realms," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 594–611.

[43] X. Zhai et al., "A large-scale study of representation learning with the visual task adaptation benchmark," 2019, arXiv:1910.04867.