

# HiDe-PET: Continual Learning via Hierarchical Decomposition of Parameter-Efficient Tuning

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Hang Su, *Member, IEEE*, Jun Zhu, *Fellow, IEEE*

**Abstract**—The deployment of pre-trained models (PTMs) has greatly advanced the field of continual learning (CL), enabling positive knowledge transfer and resilience to catastrophic forgetting. To sustain these advantages for sequentially arriving tasks, a promising direction involves keeping the pre-trained backbone frozen while employing parameter-efficient tuning (PET) techniques to instruct representation learning. Despite the popularity of Prompt-based PET for CL, its empirical design often leads to sub-optimal performance in our evaluation of different PTMs and target tasks. To this end, we propose a unified framework for CL with PTMs and PET that provides both theoretical and empirical advancements. We first perform an in-depth theoretical analysis of the CL objective in a pre-training context, decomposing it into hierarchical components namely within-task prediction, task-identity inference and task-adaptive prediction. We then present Hierarchical Decomposition PET (HiDe-PET), an innovative approach that explicitly optimizes the decomposed objective through incorporating task-specific and task-shared knowledge via mainstream PET techniques along with efficient recovery of pre-trained representations. Leveraging this framework, we delve into the distinct impacts of implementation strategy, PET technique and PET architecture, as well as adaptive knowledge accumulation amidst pronounced distribution changes. Finally, across various CL scenarios, our approach demonstrates remarkably superior performance over a broad spectrum of recent strong baselines. A preliminary version of our code is released at <https://github.com/thu-ml/HiDe-Prompt>.

**Index Terms**—Continual Learning, Incremental Learning, Pre-trained Models, Parameter-Efficient Tuning, Catastrophic Forgetting.

## 1 INTRODUCTION

The proficiency of artificial intelligence (AI) in accommodating real-world dynamics is largely contingent on its capability of continual learning (CL), which has benefited significantly in recent years from the deployment of pre-trained models (PTMs). In essence, PTMs provide CL with not only positive knowledge transfer but also resilience to catastrophic forgetting [1]–[4], which are critical to improve the performance of CL methods. Given that adapting PTMs to sequentially arriving tasks may compromise these advantages via progressive overwriting of pre-trained knowledge, numerous efforts have been devoted into implementing parameter-efficient tuning (PET) techniques for CL, i.e., keeping the pre-trained backbone frozen and introducing a few lightweight parameters to instruct representation learning. However, current advances predominantly center around empirical designs of Prompt-based PET [5]–[8], which struggle to adequately achieve the CL objective and therefore often exhibit sub-optimal performance across different PTMs and target tasks (see Sec. 6.2). In response to this critical challenge, there is an urgent need for a more profound understanding of CL with PTMs and PET, coupled with endeavors to enhance its effectiveness and generality.

In this work, we propose a unified framework for CL with PTMs and PET, seeking to advance this direction with both theoretical and empirical insights. We initiate our explorations with an in-depth theoretical analysis of the CL objective in a pre-training context. Considering the signifi-

cant impact of pre-trained knowledge on CL, this objective can be decomposed into three hierarchical components in response to sequentially arriving tasks, namely within-task prediction (WTP), task-identity inference (TII) and task-adaptive prediction (TAP). They prove to be sufficient and necessary to ensure low errors in common CL settings. Based on the theoretical analysis, we devise an innovative approach named Hierarchical Decomposition PET (HiDe-PET) to explicitly optimize WTP, TII and TAP.

The principal concept behind HiDe-PET leverages the unique strengths of PTMs for CL, with a particular focus on the effective instruction and efficient recovery of pre-trained representations. As a generic architecture applicable to mainstream PET techniques (e.g., Prompt [9], [10], Adapter [11], LoRA [12], etc.), we construct an ensemble of task-specific parameters that incorporates the knowledge of each task to optimize WTP, and a set of task-shared parameters that accumulates knowledge in a task-agnostic manner to improve TII. We further recover the distribution of uninstructed and instructed representations through preserving their statistical information, so as to optimize TII and TAP, respectively. In this way, our HiDe-PET is able to adeptly predict the identity of task-specific parameters from uninstructed representations and collect appropriate instructed representations for final predictions.

The proposed framework facilitates a thorough assessment of important factors emerged in CL with PTMs and PET, including the implementation strategy, PET technique and PET architecture. For example, we dissect representative strategies for stabilizing task-shared parameters and for preserving pre-trained representations, empirically analyzing their effectiveness. Moreover, we evaluate the behaviors of different PET techniques in achieving the CL objective,

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Hang Su, and Jun Zhu are with Dept. of Comp. Sci. & Tech., Institute for AI, BNRIst Center, THBI Lab, Tsinghua-Bosch Joint Center for ML, Tsinghua University, Beijing, China (email: wly19@tsinghua.org.cn; jingyi\_xie96@163.com; xxzhang1993@gmail.com; {suhangss, dcszj}@tsinghua.edu.cn). Corresponding authors: Jun Zhu and Hang Su.

where the Prompt-based PET tends to be less effective in WTP, consequently displaying lower sensitivity to TII and clearly lagging behind the LoRA/Adapter-based PET. Motivated by the inherent connections between TII and out-of-distribution (OOD) detection, we further devise a PET hierarchy tailored for adaptive knowledge accumulation, and unravel the relationship between task-specific and task-shared PET architectures in representation learning.

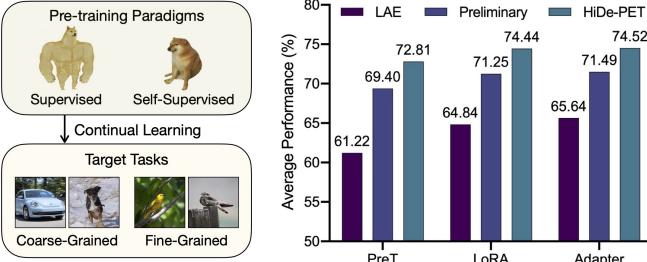


Fig. 1. Summary of CL performance with different PET techniques. We compare our HiDe-PET, our preliminary version [13] and LAE [14], and report the final average accuracy (FAA) over three pre-trained checkpoints and four CL benchmarks.

Upon extensive experiments, our HiDe-PET clearly outperforms a wide range of recent strong baselines, and demonstrates remarkable generality across a variety of PET techniques, pre-trained checkpoints and CL benchmarks (summarized in Fig. 1). We further provide empirical analysis to elucidate the respective contributions of the three hierarchical components. It is noteworthy that some results have been presented in our preliminary version [13], which mainly considered a specific implementation of task-specific parameters via Prompt-based PET. In contrast, the current version presents a unified framework for CL with PTMs and PET. It incorporates substantial extensions to the implementation strategy, PET technique and PET architecture, culminating in a comprehensive analysis and improved performance.

Overall, our main contributions are as follows:

- We present an in-depth theoretical analysis of the CL objective in a pre-training context, decomposing it into three hierarchical components for model design;
- We propose an innovative approach that exploits mainstream PET techniques and pre-trained representations to explicitly optimize the decomposed objective;
- We conduct extensive experiments to demonstrate the effectiveness and generality of our approach, coupled with a thorough assessment of important factors emerged in CL with PTMs and PET.

## 2 RELATED WORK

**Continual Learning (CL)** is characterized by learning sequentially arriving tasks and performing well on them. The primary challenge of CL stems from the dynamic nature of data distribution, which leads to catastrophic forgetting of old tasks while acquiring new ones [3], [15]. As summarized in a recent survey [3], representative methods include selective stabilization of important parameters for old tasks [16]–[18], approximation and recovery of old data distributions [19]–[21], exploitation of robust and well-distributed representations [22]–[24], manipulation of optimization programs

via gradient projection [25]–[27], construction of (relatively) dedicated parameters for each task [28]–[30], etc.

In the realm of CL, current efforts have predominantly centered around computer vision (CV), specifically for visual classification tasks. These efforts have progressively expanded their scope to include more complex visual tasks, as well as natural language processing (NLP), reinforcement learning (RL) and other related applications. Across various representative CL settings, especially those lacking the oracle of task identity during the testing phase, robust CL models often necessitate the storage and rehearsal of old training samples [21], [31], [32], which creates additional resource overheads and potential privacy concerns. These issues have been largely avoided through effective use of pre-trained knowledge in recent work, as discussed later.

**Pre-training and Fine-tuning:** Transfer learning with pre-trained models (PTMs) can significantly improve the performance of target tasks and has therefore become a prevalent paradigm in many areas of artificial intelligence (AI). Since the pre-trained knowledge is usually generalized and difficult to cover all specific domains, PTMs necessitate further fine-tuning for better adaptation. The most straightforward way is to update all model parameters, but involves keeping a separate copy of fine-tuned model parameters for each task. This leads to considerable resource overheads especially for advanced PTMs of increasing size.

In order to improve the efficiency of fine-tuning, some lightweight alternatives have been proposed that update only a few extra parameters with most of the model parameters frozen, collectively referred to as the parameter-efficient tuning (PET) techniques. These PET techniques were originally proposed for NLP and are now widely used for CV as well. Representative practices include *Prompt Tuning* (ProT) [9] and *Prefix Tuning* (PreT) [10] that prepend short prompt tokens into the original inputs or intermediate inputs, *Adapter* [11] that inserts small neural modules between backbone layers, *Low-Rank Adaptation* (LoRA) [12] that approximates the updates of model parameters with low rank matrices, etc. A recent work [33] unified these PET techniques in a similar form, corresponding to modifying specific hidden states of the PTMs.

**CL with PTMs and PET:** While much of the past work in CL has focused on training from scratch, a growing body of efforts have delved into the benefits of PTMs, which provide not only positive knowledge transfer but also resilience to catastrophic forgetting [1], [2]. Meanwhile, PTMs also need to improve the ability of CL to accommodate sequentially arriving tasks and to refine pre-trained knowledge from them. However, fine-tuning PTMs becomes much more difficult when considering CL, since repetitive adaptation of the same PTM may lead to progressive overwriting of pre-trained knowledge, whereas separate saving of the fine-tuned PTMs creates an additional linear increase in resource overhead on the time scale [4].

Therefore, applying PET techniques for CL has become an emerging direction in recent years, with Prompt-based PET being predominant. Many state-of-the-art methods have focused on designing appropriate prompt architectures for CL, such as task-shared prompts [6], [14], [34], task-specific prompts [6]–[8], [13] and their combinations [5], [6]. Since the frozen backbone with pre-trained knowledge

can provide robust and well-distributed representations, these methods have almost achieved the performance pinnacle of rehearsal-free CL under adequate supervised pre-training and general tasks. However, their sub-optimality in achieving the objective of CL has been clearly exposed under the more realistic self-supervised pre-training and fine-grained tasks [4], [13], in part due to the limited fitting ability of Prompt-based PET [35], [36]. LAE [14] is a recent work that assembles a pair of task-shared parameters to be implemented with mainstream PET techniques, but exhibits only moderate improvements in CL performance.

### 3 PRELIMINARIES

In this section, we first introduce the problem formulation of continual learning (CL) in a pre-training context. Then we describe representative parameter-efficient tuning (PET) techniques and PET-based CL methods.

#### 3.1 Problem Formulation

The CL problem can be generally defined as learning sequentially arriving tasks from their respective training sets  $\mathcal{D}_1, \dots, \mathcal{D}_t$  in order to perform well on their corresponding test sets. The training set and test set of each task are assumed to follow the same distribution. For supervised CL, each training set  $\mathcal{D}_t = \{(x_{t,n}, y_{t,n})\}_{n=1}^{N_t}$  comprises multiple sample-label pairs, where  $x_{t,n} \in \mathcal{X}_t$  and  $y_{t,n} \in \mathcal{Y}_t$  represent the sample and label elements, respectively, and  $N_t$  denotes the size of  $\mathcal{D}_t$ . Let's consider a neural network model composed of a backbone  $f_\theta$  with parameters  $\theta$ , and an output layer  $h_\psi$  with parameters  $\psi$ . This model aims to learn a projection from  $\mathcal{X} = \bigcup_{i=1}^t \mathcal{X}_i$  to  $\mathcal{Y} = \bigcup_{i=1}^t \mathcal{Y}_i$ , so that it can correctly predict the label  $\hat{y} = h_\psi(f_\theta(\mathbf{x})) \in \mathcal{Y}$  of an unseen test sample  $\mathbf{x}$  from observed tasks. Since  $\mathcal{D}_1, \dots, \mathcal{D}_t$  are usually limited in size and distinct in distribution, learning  $f_\theta$  from scratch can easily converge to an undesirable local minimum. In contrast, initialization of  $f_\theta$  with a substantial quantity of training samples external to  $\mathcal{D}_1, \dots, \mathcal{D}_t$ , i.e., applying adequate pre-training, helps  $\theta$  converge to a wide low-error region and thus can greatly improve the CL performance [1], [2].

Depending on the split of label space and the availability of task identity, there are three representative setups for CL [37], including task-incremental learning (TIL), domain-incremental learning (DIL), and class-incremental learning (CIL). Specifically, the label space  $\mathcal{Y}_i$  with  $i \in [t]$  is the same for DIL whereas disjoint for TIL and CIL. The task identity  $i \in [t]$  is provided at test time for TIL whereas not available for DIL and CIL. As a result, CIL is often considered more challenging and representative. Of note, although CIL is named from classification tasks, its definition can be generalized to other task types. To avoid additional resource overhead and potential privacy issues, the CL process is further restricted to be *rehearsal-free* [8], i.e., the sample elements of each  $\mathcal{D}_i$  are available only when learning task  $i$ , which particularly increases the challenge of CIL [3].

#### 3.2 PET Techniques

The backbone  $f_\theta$  of advanced pre-trained models (PTMs) often employs the transformer architecture [38] based on the

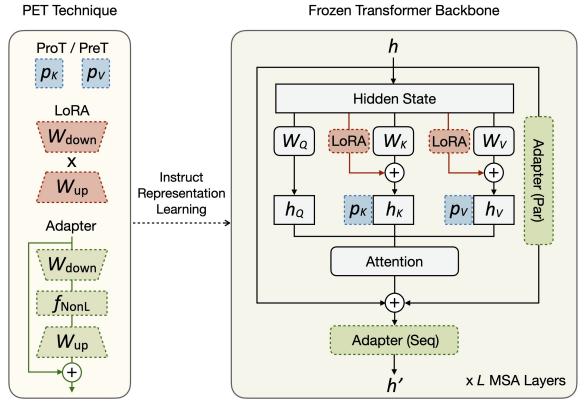


Fig. 2. Implementation of PET techniques for representation learning.

multi-head attention mechanism. For example, a pre-trained vision transformer (ViT) [39] consists of multiple consecutive multi-head self-attention (MSA) layers that transform an input sample into a sequence-like output representation  $\mathbf{r} \in \mathbb{R}^{d_r \times d}$  of sequence length  $d_r$  and embedding dimension  $d$ . Let's consider the  $l$ -th layer with input  $\mathbf{h}_{(l)}$  and output  $\mathbf{h}'_{(l)}$ , where  $\mathbf{h}'_{(L)}$  is equivalent to  $\mathbf{r}$  for total  $L$  layers. Since the output  $\mathbf{h}'_{(l)}$  then becomes the input  $\mathbf{h}_{(l+1)} \in \mathbb{R}^{d_{h_{(l+1)}} \times d}$  of the next layer, we omit the layer identity  $l$  for the sake of clarity. Then, the input  $\mathbf{h} \in \mathbb{R}^{d_h \times d}$  is further specified as the query  $\mathbf{h}_Q$ , key  $\mathbf{h}_K$  and value  $\mathbf{h}_V$ , and the output  $\mathbf{h}' \in \mathbb{R}^{d_h \times d}$  of the current layer is

$$\mathbf{h}' = \text{MSA}(\mathbf{h}_Q, \mathbf{h}_K, \mathbf{h}_V) = \text{Concat}(h_1, \dots, h_m) \mathbf{W}_O, \quad (1)$$

$$h_i = \text{Attn}(\mathbf{h}_Q \mathbf{W}_{Q,i}, \mathbf{h}_K \mathbf{W}_{K,i}, \mathbf{h}_V \mathbf{W}_{V,i}), i \in [m], \quad (2)$$

where  $\mathbf{W}_O, \mathbf{W}_{Q,i}, \mathbf{W}_{K,i}, \mathbf{W}_{V,i}$  are projection matrices,  $m$  is the number of heads, and  $\mathbf{h}_Q = \mathbf{h}_K = \mathbf{h}_V = \mathbf{h}$  in ViT. The concatenation (Concat) and attention (Attn) functions are specified in their original papers [38], [39].

To facilitate effective transfer of pre-trained knowledge while preventing its catastrophic forgetting, the backbone parameters  $\theta$  are usually frozen and additional lightweight parameters are introduced to instruct representation learning, referred to as the PET techniques [33]. Here we describe some representative ones (see Fig. 2):

**Prompt Tuning** (ProT) [9] and **Prefix Tuning** (PreT) [10] both prepend a few learnable parameters  $\mathbf{p} \in \mathbb{R}^{d_p \times d}$  of sequence length  $d_p$  and embedding dimension  $d$  to  $\mathbf{h}$ , collectively known as the *Prompt-based PET*. For ProT in ViT, an identical  $\mathbf{p}$  is prepended to  $\mathbf{h}_Q$ ,  $\mathbf{h}_K$  and  $\mathbf{h}_V$ :

$$\mathbf{h}' = \text{MSA}([\mathbf{p}; \mathbf{h}_Q], [\mathbf{p}; \mathbf{h}_K], [\mathbf{p}; \mathbf{h}_V]), \quad (3)$$

where  $[\cdot; \cdot]$  represents the concatenation operation along the dimension of sequence length. Since the output in  $\mathbb{R}^{(d_h+d_p) \times d}$  has increased dimensions, ProT is often used for only the last layer in CL [5], [7]. In contrast, PreT splits  $\mathbf{p}$  into  $\mathbf{p}_K \in \mathbb{R}^{d_p/2 \times d}$  for  $\mathbf{h}_K$  and  $\mathbf{p}_V \in \mathbb{R}^{d_p/2 \times d}$  for  $\mathbf{h}_V$ :

$$\mathbf{h}' = \text{MSA}(\mathbf{h}_Q, [\mathbf{p}_K; \mathbf{h}_K], [\mathbf{p}_V; \mathbf{h}_V]), \quad (4)$$

where the output has the same dimension as the input  $\mathbf{h} \in \mathbb{R}^{d_h \times d}$ , allowing PreT to be implemented in multiple layers. In particular, the output of PreT can be reframed as

$$\mathbf{h}' \leftarrow (1 - \lambda(\mathbf{h}))\mathbf{h}' + \lambda(\mathbf{h}) f_{\text{NonL}}(\mathbf{h} \mathbf{W}_Q \mathbf{p}_K^\top) \mathbf{p}_V, \quad (5)$$

where  $f_{\text{NonL}}$  is the nonlinear (NonL) softmax function, and  $\lambda(\mathbf{h})$  is a scalar that depends on the input [33].

**Adapter** [11] inserts lightweight neural modules between backbone layers, each usually composed of a down-projection matrix  $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times r}$  that reduces the dimension of  $\mathbf{h}$  with bottleneck  $r$ , a nonlinear (NonL) activation function  $f_{\text{NonL}}$  and an up-projection matrix  $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r \times d}$ . These modules are implemented with residual connections, acting on the output  $\mathbf{h}'$  in a *sequential* (Seq) manner, i.e.,

$$\mathbf{h}' \leftarrow \mathbf{h}' + f_{\text{NonL}}(\mathbf{h}' \mathbf{W}_{\text{down}}) \mathbf{W}_{\text{up}}, \quad (6)$$

as well as on the input  $\mathbf{h}$  in a *parallel* (Par) manner, i.e.,

$$\mathbf{h}' \leftarrow \mathbf{h}' + f_{\text{NonL}}(\mathbf{h} \mathbf{W}_{\text{down}}) \mathbf{W}_{\text{up}}. \quad (7)$$

**LoRA** [12] approximates the updates of pre-trained parameter matrix  $\mathbf{W} \in \mathbb{R}^{d \times k}$  with a low-rank decomposition  $\mathbf{W} + \Delta \mathbf{W} = \mathbf{W} + \mathbf{W}_{\text{down}} \mathbf{W}_{\text{up}}$ , where  $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times r}$ ,  $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r \times k}$  and  $r$  is the low-rank bottleneck. For ViT, LoRA is typically used to update the  $\mathbf{W}_Q$  and  $\mathbf{W}_V$  of a backbone layer. As a special case, when  $\mathbf{W}$  performs linear projection of the input  $\mathbf{h}$ , the output is modified as

$$\mathbf{h}' \leftarrow \mathbf{h}' + s \cdot \mathbf{h} \mathbf{W}_{\text{down}} \mathbf{W}_{\text{up}}, \quad (8)$$

where  $s \geq 1$  is a scalar hyperparameter [33].

As we can see, a common feature of these PET techniques is that they all amount to modulating the (intermediate) representations of  $f_\theta$ , though differing in their specific implementations.

### 3.3 PET-Based CL Methods

With the widespread use of PTMs in CL, there have been a variety of methods that incorporate PET techniques on a continual basis. Most of these methods have focused on designing appropriate PET architectures to formulate lightweight parameters tailored for CL, which can be conceptually summarized into task-specific parameters, task-shared parameters, and their explicit or implicit combinations. In particular, task-specific parameters require their identity to be predicted during the testing phase, while task-shared parameters need to mitigate catastrophic forgetting when learning each task. We briefly describe these methods here, with a comprehensive summary in Appendix Table 6 for further comparison.

**L2P** [5] constructs a prompt pool  $\mathbf{p}_1, \dots, \mathbf{p}_M$  of size  $M$  and instructs pre-trained representations in a ProT manner. Each prompt is associated with a learnable key, optimized by the cosine distance of the top- $N$  keys to a query function  $q(\mathbf{x}) = f_\theta(\mathbf{x})[0]$ . The most relevant prompt(s) can therefore be selected via uninstructed representations.

**DualPrompt** [6] employs the task-specific prompt  $\mathbf{p}_t'$  and task-shared prompt  $\mathbf{p}'$  to instruct respective layers in a PreT manner. The layer-wise  $\mathbf{p}_t'$  is associated with a task-specific key, optimized by its cosine distance to  $q(\mathbf{x})$ , and the best-matched key is selected via uninstructed representations.

**S-Prompt** [7] employs only the task-specific prompt  $\mathbf{p}_t$  and instructs pre-trained representations in a ProT manner. The task identity is inferred from preserved task centroids with  $k$ -Nearest Neighbors ( $k$ NNs>). S-Prompt also employs a multi-head output layer associated with the task identity.

**CODA-Prompt** [8] performs a weighted summation of the prompt pool, i.e.,  $\mathbf{p} = \sum_{i=1}^M \alpha_i \mathbf{p}_i$ , and instructs multiple layers in a PreT manner. Each weighting factor  $\alpha_i$  for  $i \in [M]$  is associated to a learnable key, optimized by its cosine distance to  $q(\mathbf{x})$ . The inference of  $\alpha_i$  can therefore construct an appropriate prompt for each input sample.

**LAE** [14] employs two kinds of task-shared parameters to incorporate knowledge from more recent tasks and more remote tasks, respectively, applicable to PreT, Adapter and LoRA for multiple layers. Their update speeds are regulated via combinatorial strategies such as temporary freezing and exponential moving average (EMA).

Besides, there are several relevant methods with different focuses. For example, SLCA [4] updates the entire backbone with a reduced learning rate, and further preserves pre-trained representations via dedicated covariance matrices. RanPAC [34] projects pre-trained representations to a high-dimension space and preserves them via a shared covariance matrix. These methods are often parameter-inefficient, i.e., the parameter cost is comparable to  $\theta$  due to a complexity much larger than  $O(d^2)$ , and therefore not prioritized in this work.<sup>1</sup>

Taken together, three notable limitations have surfaced in current progress of harnessing PET techniques for CL. First, the above methods often rely on empirical designs, making it difficult to ascertain their effectiveness in achieving the objective of CL in a pre-training context. In particular, their performance exhibits significant variations across different PTMs and target tasks, as demonstrated in Sec. 6.2. Second, these methods predominantly center around Prompt-based PET, which has been shown to be less effective under self-supervised pre-training [36] and fine-grained tasks [40], leaving underexplored the particular behaviors and potential benefits of other alternatives. Third, these methods share some analogous strategies, such as stabilizing task-shared parameters and recovering pre-trained representations, without a comprehensive analysis of different implementations and assimilation of their respective strengths. Therefore, there is an urgent need to establish a unified framework that incorporates both theoretical and empirical insights for CL with PTMs and PET, which constitutes our main motivation.

## 4 THEORETICAL ANALYSIS

In this section, we present an in-depth theoretical analysis of the CL objective in a pre-training context. We decompose it into three hierarchical components, which inspire our design of PET-based CL methods.

**Three Hierarchical Components:** For CL of sequentially arriving  $\mathcal{D}_i$ ,  $\mathcal{X}_i$  and  $\mathcal{Y}_i$  are the domain and label of task  $i \in [t]$ . Due to the space limit, here we take CIL where  $\mathcal{Y}_i \cap \mathcal{Y}_{i'} = \emptyset$  and  $\forall i \neq i'$  as a typical scenario for theoretical analysis, and leave the results of DIL and TIL to Appendix A. Let  $\mathcal{X}_i = \bigcup_j \mathcal{X}_{i,j}$  and  $\mathcal{Y}_i = \bigcup_j \mathcal{Y}_{i,j}$ , where  $j \in [|\mathcal{Y}_i|]$  indicates the  $j$ -th class in task  $i$ . Now assume we have a ground event denoted as  $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_t\}$  and a pre-trained backbone  $f_\theta$ . For any sample  $\mathbf{x} \in \bigcup_{i=1}^t \mathcal{X}_i$ , a general goal of the CIL problem is to learn  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathcal{D}, \theta)$ , where  $i \in [t]$  and  $j \in [|\mathcal{Y}_i|]$ . In alignment with a pioneering theoretical study for

1. Our preliminary version [13] also employed dedicated covariance matrices as the main implementation to acquire better performance.

CL from scratch [41], we first decompose our objective into two probabilities, including task-identity inference (TII) and within-task prediction (WTP), denoted as  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)$  and  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)$ , respectively:

$$P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathcal{D}, \theta) = P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta), \quad (9)$$

where the former predicts which task to perform and the latter performs that task. Let  $\bar{i} \in [t]$  and  $\bar{j} \in [\|\mathcal{Y}_i\|]$  be the ground truth of an  $\mathbf{x}$  w.r.t. the task identity and within-task index. Eq. (9) shows that if we can improve either the WTP performance  $P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)$ , the TII performance  $P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)$ , or both, then the CIL performance  $P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta)$  would be improved. However, such improvement is limited since it is upper-bounded by WTP or TII, both of which act on the output space of  $f_\theta$  that has not been adapted to target tasks.<sup>2</sup> On the top of WTP and TII, there exists a performance gap stemming from rectifying the final predictions over all observed tasks after potential adaptation of  $f_\theta$ , e.g., PET in Sec. 3.2. We refer to this hierarchical process as task-adaptive prediction (TAP) and denote it as  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$ , where  $y \in \mathcal{Y}_{\bar{i},\bar{j}}$  is the ground truth label of  $\mathbf{x}$ , and  $\mathcal{X}^y$  represents the domain of class  $y$  in all observed tasks. Then, the ultimate goal of CL is formulated as a multi-objective optimization problem, i.e.,

$$\max[P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta), P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)]. \quad (10)$$

Notice that the TII probability is a categorical distribution over all observed tasks up to  $t$ , while the TAP probability is over all observed classes  $\bigcup_{i=1}^t \mathcal{Y}_i$ . Interestingly,  $P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta)$  and  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$  are equivalent when training from scratch, whereas become different when considering the significant impact of pre-trained knowledge in  $\theta$ .

To resolve the above WTP, TII and TAP, we derive the sufficient and necessary conditions with the widely-used cross-entropy loss. Specifically, we define

$$H_{\text{WTP}}(\mathbf{x}) = \mathcal{H}(\mathbf{1}_j, \{P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)\}_j), \quad (11)$$

$$H_{\text{TII}}(\mathbf{x}) = \mathcal{H}(\mathbf{1}_i, \{P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)\}_i), \quad (12)$$

$$H_{\text{TAP}}(\mathbf{x}) = \mathcal{H}(\mathbf{1}_y, \{P(\mathbf{x} \in \mathcal{X}^c | \mathcal{D}, \theta)\}_c), \quad (13)$$

where  $H_{\text{WTP}}$ ,  $H_{\text{TII}}$ , and  $H_{\text{TAP}}$  are the cross-entropy values of WTP, TII, and TAP, respectively.  $c \in [\|\bigcup_{i=1}^t \mathcal{Y}_i\|]$  denotes the index of all observed classes.  $\mathcal{H}(p, q) \triangleq -\mathbb{E}_p[\log q] = -\sum_k p_k \log q_k$ .  $\mathbf{1}$  is a one-hot encoding function.

We now present the first theorem under the CIL setting (see Appendix A for the complete proof and the corresponding extensions to DIL and TIL settings):

**Theorem 1.** For continual learning (CL) in a pre-training context, if  $\mathbb{E}_{\mathbf{x}}[H_{\text{WTP}}(\mathbf{x})] \leq \delta$ ,  $\mathbb{E}_{\mathbf{x}}[H_{\text{TII}}(\mathbf{x})] \leq \epsilon$ , and  $\mathbb{E}_{\mathbf{x}}[H_{\text{TAP}}(\mathbf{x})] \leq \eta$ , we have the loss error  $\mathcal{L} \in [0, \max\{\delta + \epsilon, \eta\}]$ , regardless whether the WTP predictor, TII predictor and TAP predictor are trained together or separately.

With the use of cross-entropy, the CIL performance tends to be better as the bounds are tightened. In Theorem 1 we have shown that good performances of WTP, TII and TAP are sufficient to guarantee a good performance of CIL. For completeness, we now study the necessary conditions of a well-performed CIL model in Theorem 2.

2. As discussed in Sec. 3.3, the output space of  $f_\theta$  will usually be adapted by PET techniques or be employed as the query function.

**Theorem 2.** For CL in a pre-training context, if the loss error  $\mathcal{L} \leq \xi$ , then there always exist (1) a WTP predictor, s.t.  $H_{\text{WTP}} \leq \xi$ ; (2) a TII predictor, s.t.  $H_{\text{TII}} \leq \xi$ ; and (3) a TAP predictor, s.t.  $H_{\text{TAP}} \leq \xi$ .

Theorem 2 suggests that if a CIL model is well trained (i.e., with low loss error), then the WTP error, TII error and TAP error for sequentially arriving tasks are always implied to be small. Akin to the connection between  $P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta)$  and  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$ , Theorem 1 and Theorem 2 would degenerate into the main conclusion of the previous theoretical study [41] if the pre-trained knowledge carried by  $\theta$  is not considered, suggesting that the presented theorems are particularly directed to the impact of pre-training on CL.

**Connection of TII to OOD Detection:** In essence, the TII probability specifies the CL problem with task-wise input samples. Although the definition of “task” in literature can generalize to an incoming training batch of distinct distribution [3], it may not be pertinent in describing realistic data streams with apparent similarity and dissimilarity. Indeed, the CL problem is often associated with the out-of-distribution (OOD) detection [42], i.e., the ability of a model to detect an unseen input sample, which has been shown to behave similarly as task prediction when training from scratch [41]. Inspired by this, we further explore the necessary conditions to optimize TII/OOD for CL in a pre-training context, in order to facilitate adaptive knowledge accumulation from more pronounced distribution changes.

We again use cross-entropy to measure the performance of TII and OOD detection, so as to establish their connection in each task. We first define the OOD detector of the  $i$ -th task as  $P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)$ . Different from the TII probability, the OOD detection probability here is a Bernoulli distribution:

$$H_{\text{OOD},i}(\mathbf{x}) = \begin{cases} \mathcal{H}(1, P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)), & \mathbf{x} \in \mathcal{X}_i \\ \mathcal{H}(0, P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)), & \mathbf{x} \notin \mathcal{X}_i \end{cases}, \quad (14)$$

where  $H_{\text{OOD},i}$  is the cross-entropy value of an OOD detector of task  $i$ , and  $P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)$  can be predicted with an appropriate distance function. The TII probability can then be defined with the OOD detectors, i.e.,  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) = \frac{P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}$ .

Now we have the following theorem to explore the connection between TII and OOD detection in a pre-training context (see Appendix B for the complete proof):

**Theorem 3.** For CL in a pre-training context, if  $H_{\text{OOD},i}(\mathbf{x}) \leq \epsilon_i$  for  $i \in [t]$ , then we have  $H_{\text{TII}}(\mathbf{x}) \leq (\sum_i \mathbf{1}_{\mathbf{x} \in \mathcal{X}_i} e^{\epsilon_i})(\sum_i \mathbf{1}_{\mathbf{x} \notin \mathcal{X}_i} (1 - e^{-\epsilon_i}))$ . Likewise, if  $H_{\text{TII}}(\mathbf{x}) \leq \epsilon$ , then  $H_{\text{OOD},i}(\mathbf{x}) \leq \epsilon$  for  $i \in [t]$ .

Theorem 3 shows that the TII performance improves if the OOD detection performance improves, and vice versa. In particular,  $H_{\text{TII}}(\mathbf{x})$  converges to 0 as  $\epsilon_i$  converges to 0. We further derive the sufficient and necessary conditions of improving CL with WTP, OOD detection and TAP, as detailed in Appendix B.

## 5 HIERARCHICAL DECOMPOSITION PET

Based on the above analysis, we now present an innovative approach named Hierarchical Decomposition PET (HiDe-PET) to explicitly optimize the three hierarchical components tailored for CL in a pre-training context (see Fig. 3).

**Algorithm 1** Training Algorithm of HiDe-PET

**Input:** Pre-trained transformer backbone  $f_\theta$ , training sets  $\mathcal{D}_1, \dots, \mathcal{D}_t$ , number of tasks  $t$ , number of epochs  $E$  and  $\bar{E}$ .  
**Output:** Parameters  $e_1, \dots, e_t, g, \omega$  and  $\psi$

```

1: Initialize  $g, \omega$  and  $\psi$ 
2: for  $i = 1, \dots, t$  do
3:   Initialize  $\hat{\psi}$  with  $\psi$ 
4:   Construct  $e_i$  with  $e_1, \dots, e_{i-1}$ 
5:   for  $epoch = 1, \dots, E$  do
6:     Update  $g$  and  $\hat{\psi}$  with  $\mathcal{L}_{\text{CE}}(\hat{\psi}, g)$ 
7:     Update  $e_i$  and  $\psi$  with  $\mathcal{L}_{\text{WTP}}(\psi, e_i)$  in Eq. (15)
8:   for  $c \in \mathcal{Y}_i$  do
9:     Obtain  $\hat{\mathcal{G}}_{i,c}$  from  $f_\theta, g$  and  $\mathcal{D}_i$ 
10:    Obtain  $\mathcal{G}_{i,c}$  from  $f_\theta, e_i$  and  $\mathcal{D}_i$ 
11:   for  $epoch = 1, \dots, \bar{E}$  do
12:     Optimize  $\omega$  with  $\mathcal{L}_{\text{TII}}(\omega)$  in Eq. (16)
13:     Optimize  $\psi$  with  $\mathcal{L}_{\text{TAP}}(\psi)$  in Eq. (17)
14:   return  $(e_1, \dots, e_t, g, \omega, \psi)$ 

```

Our HiDe-PET is applicable to mainstream PET techniques for learning sequentially arriving tasks, from which the pre-trained knowledge can also be evolved.

**5.1 Optimization of Decomposed Objective**

The principal concept behind HiDe-PET stems from two unique strengths of PTMs in CL: (1) the pre-trained representations can be effectively adapted to the distribution of target tasks through implementing PET techniques, and (2) the distribution of target tasks can be efficiently recovered through preserving statistical information of their pre-trained representations. The optimization of WTP, TII and TAP is described as below.

First, we improve **WTP** through incorporating task-specific knowledge from  $\mathcal{D}_i$  to capture the distribution of any task  $i \in [t]$ . Specifically, we employ multiple sets of task-specific parameters  $e_1, \dots, e_t$  to instruct pre-trained representations, implemented via mainstream PET techniques. Their concrete forms can be defined as the prompt parameters  $\{p\}$  for ProT and PreT, the projection matrices  $\{\mathbf{W}_{\text{down}}, \mathbf{W}_{\text{up}}\}$  for Adapter and LoRA, etc. When learning task  $t$ , we keep the previous parameters  $e_1, \dots, e_{t-1}$  frozen to avoid catastrophic forgetting. In order to inherit knowledge obtained from CL, we employ a PET ensemble strategy that initializes  $e_t$  with  $e_{t-1}$  and optimizes  $e_t$  with a weighted combination of all previous parameters  $e_t \leftarrow \alpha \sum_{i \in [t-1]} e_i + (1 - \alpha)e_t$ , where  $\alpha \in [0, 1]$  is a hyperparameter that controls the strength of obtained knowledge that facilitates  $e_t$  in learning task  $t$ . For  $H_{\text{WTP}}$ , we then optimize  $e_t$  and  $\psi$  via the cross-entropy (CE) loss:

$$\begin{aligned} \mathcal{L}_{\text{WTP}}(\psi, e_t) &= \mathcal{L}_{\text{CE}}(\psi, e_t) \\ &= \frac{1}{|\mathcal{D}_t|} \sum_{(\mathbf{x}, y) \in \mathcal{D}_t} -\log \frac{\exp(h_\psi(f_{\theta, e_t}(\mathbf{x})[y]))}{\sum_{y' \in \mathcal{Y}_t} \exp(h_\psi(f_{\theta, e_t}(\mathbf{x}))[y'])}. \end{aligned} \quad (15)$$

Second, we improve **TII** and **TAP** through recovering the distributions of pre-trained representations to achieve the optimum over all tasks. Specifically, after learning any task  $i \in [t]$ , we collect the uninstructed and instructed representations, i.e., the backbone projection of  $\mathcal{D}_i$  with  $f_\theta(\cdot)$  and  $f_{\theta, e_i}(\cdot)$ , respectively. We then approximate the distributions of these representations with their statistical

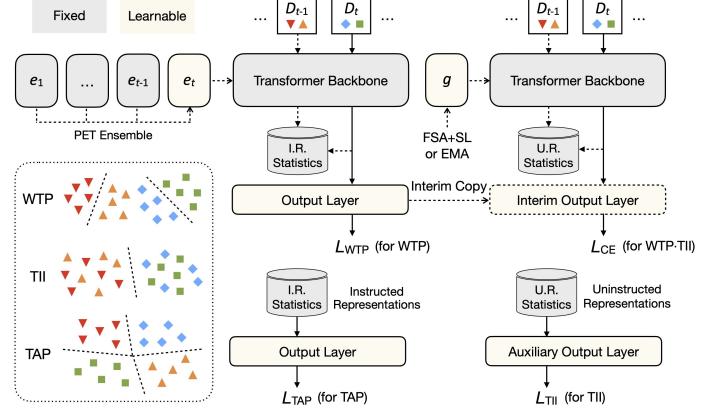


Fig. 3. Illustration of HiDe-PET. See Fig. 2 for detailed implementations of PET techniques and the frozen transformer backbone.

information for subsequent recovery. Taking classification tasks as an example, for each class  $c \in \mathcal{Y}_i$  and  $i \in [t]$ , we denote the approximated distributions of uninstructed and instructed representations as  $\hat{\mathcal{G}}_{i,c}$  and  $\mathcal{G}_{i,c}$ , respectively.

For  $H_{\text{TII}}$ , we build an auxiliary output layer  $\hat{h}_\omega(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^t$  with parameters  $\omega$ , learning the projection from uninstructed representations to the task identity:

$$\mathcal{L}_{\text{TII}}(\omega) = \sum_{i \in [t]} \sum_{c \in \mathcal{Y}_i} \sum_{\hat{\mathbf{r}} \in \hat{\mathcal{S}}_{i,c}} -\log \frac{\exp(\hat{h}_\omega(\hat{\mathbf{r}})[i])}{\sum_{j=1}^t \exp(\hat{h}_\omega(\hat{\mathbf{r}})[j])}, \quad (16)$$

where  $\hat{\mathcal{S}}_{i,c}$  is a collection of uninstructed representations sampled in a class-balanced manner from  $\hat{\mathcal{G}}_{i,c}$  for  $c \in \mathcal{Y}_i$  and  $i \in [t]$ . Therefore, we can determine the task identity via uninstructed representations and then obtain the corresponding instructed representations.

For  $H_{\text{TAP}}$ , we use a similar strategy to optimize the final output layer  $h_\psi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{|\mathcal{Y}|}$  with parameters  $\psi$ , learning the projection from instructed representations to all observed classes:

$$\mathcal{L}_{\text{TAP}}(\psi) = \sum_{i \in [t]} \sum_{c \in \mathcal{Y}_i} \sum_{\mathbf{r} \in \mathcal{S}_{i,c}} -\log \frac{\exp(h_\psi(\mathbf{r})[c])}{\sum_{j=1}^t \sum_{c' \in \mathcal{Y}_j} \exp(h_\psi(\mathbf{r})[c'])}, \quad (17)$$

where  $\mathcal{S}_{i,c}$  is a collection of instructed representations sampled in a class-balanced manner from  $\mathcal{G}_{i,c}$  for  $c \in \mathcal{Y}_i$  and  $i \in [t]$ .

**Improvement of Uninstructed Representations:** Since the TII process depends heavily on the uninstructed representations of  $\mathcal{D}_i$  collected from  $f_\theta(\cdot)$ , its effectiveness tends to be severely affected by the pre-trained checkpoints and target tasks. This issue becomes particularly pronounced when implementing more advanced PET techniques, which better incorporate task-specific knowledge for WTP and are thus more sensitive to TII. To address this issue, we further deploy a set of task-shared parameters  $g$  to improve TII in a task-agnostic manner.  $g$  is implemented via mainstream PET techniques analogous to  $e_i$ , while optimized with the cross-entropy  $\mathcal{L}_{\text{CE}}(\hat{\psi}, g)$  to accumulate task-shared knowledge from sequentially arriving  $\mathcal{D}_i$  for  $i \in [t]$ , where  $\hat{\psi}$  is an interim copy of  $\psi$  to avoid overwriting. We then use  $f_{\theta, g}(\cdot)$  instead of  $f_\theta(\cdot)$  to collect uninstructed representations<sup>3</sup>, approximate each  $\hat{\mathcal{G}}_{i,c}$  and optimize Eq. (16).

3. For naming consistency, we still use “uninstructed representations” to denote the projection of  $f_{\theta, g}(\cdot)$ .

Notably,  $\mathbf{g}$  needs to overcome its own catastrophic forgetting that leads to not only the loss of information in representation learning but also the representation shift in subsequent recovery. There are many CL methods attempting to address this challenge [4], [14], [34], but their strategies remain sub-optimal in balancing sequentially arriving tasks (see Table 3). Here we propose a simple yet effective strategy by taking advantages of first-session adaptation (FSA) [34], [43] and slow learner (SL) [4], [14]. Specifically, we learn  $\mathbf{g}$  in the first task with a larger learning rate that is adequately strong for representation learning, and then in subsequent tasks with a smaller learning rate for further fine-tuning. In this way, task-shared knowledge is effectively incorporated into  $\mathbf{g}$  and accumulates over time.

**Recovery of Task Distributions:** Since the pre-trained representations are well-distributed in general, there are many feasible strategies to approximate and preserve their distributions  $\hat{\mathcal{G}}_{i,c}$  and  $\mathcal{G}_{i,c}$ . The most straightforward option is to save randomly selected prototypes [44], yet not adequately employing the relationships between them. For classification tasks, the class-wise distribution is typically single-peaked and thus can be naturally approximated as a Gaussian with its mean and covariance [4], [13]. In order to reduce storage complexity, dedicated covariance matrices need to be further simplified for practical use, suffering from information loss to varying degrees [4], [13], [34]. Considering both storage efficiency and task-type generality, our default implementation is to obtain multiple representation centroids with  $k$ -Nearest Neighbors ( $k$ NNs) and add Gaussian noise to them. We also provide an empirical comparison of different implementations in Table 4.

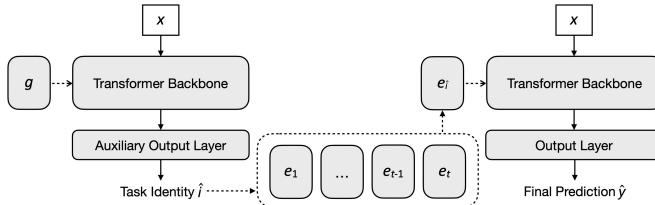


Fig. 4. Inference of HiDe-PET during the testing phase.

Overall, the entire HiDe-PET consists of two training stages (see Algorithm 1 and Fig. 3), corresponding to the pre-trained transformer backbone and the (auxiliary) output layer. At test time, HiDe-PET first predicts the task identity  $\hat{i} = \hat{h}_\omega(f_{\theta,g}(x))$  and then the overall class label  $\hat{y} = h_\psi(f_{\theta,e_i}(x))$  (see Fig. 4). Compared to the backbone parameters  $\theta$ , the trainable parameters  $e_t$ ,  $\mathbf{g}$ ,  $\omega$  and  $\psi$ , as well as the representation recovery are all lightweight, thus ensuring resource efficiency.

## 5.2 Adaptive Knowledge Accumulation

Within HiDe-PET, the parallel organization of  $e_1, \dots, e_t$  and  $\mathbf{g}$  facilitates the incorporation of task-specific and task-shared knowledge for many representative CL scenarios. In fact, the functions of  $e_1, \dots, e_t$  and  $\mathbf{g}$  are usually not exclusive, depending on the similarity and dissimilarity between task distributions. Motivated by the intrinsic connection between TII and OOD detection in our theoretical analysis, we unify the task-specific and task-shared PET architectures with a hierarchy of expandable parameter sets (see Fig. 5),

which may degenerate into either case of Sec. 5.1. We further explore a particular implementation of this hierarchy in order to accumulate knowledge adaptively from more pronounced distribution changes.

Let us assume the existence of multiple parameter sets that are implemented via mainstream PET techniques and are expanded or retrieved upon the input samples. For example, the sample elements of sequentially learning tasks  $i \in [t-1]$  have derived  $k$  parameter sets  $\mathbf{g}_1, \dots, \mathbf{g}_k$ . If the incoming  $x \in \mathcal{X}_t$  is identified as OOD from the previously observed distributions of  $\mathcal{X}_i$ , it learns an expanded set of parameters  $\mathbf{g}_{k+1}$  through the task-specific loss  $\mathcal{L}_{CE}(\hat{\psi}, \mathbf{g}_{k+1})$ , otherwise it retrieves and updates the most relevant one  $\mathbf{g}_j$  with  $j \in [k]$  through  $\mathcal{L}_{CE}(\hat{\psi}, \mathbf{g}_j)$ , where  $\hat{\psi}$  is an interim copy of the output layer parameters  $\psi$  to avoid overwriting.

**OOD Detection Strategy:** Given that the previous  $\mathcal{X}_i$  are not available in CL to describe their distributions, we take inspirations from recent metric-based OOD detectors [45] and formulate an effective criterion with their uninstructed representations:

$$P_i(x \in \mathcal{X}_i | \mathcal{D}, \theta) = \mathbf{1}(\text{Dis}(x, \hat{\mathcal{G}}_i) > \lambda_{OOD}), x \in \mathcal{X}_t, \quad (18)$$

where  $\hat{\mathcal{G}}_i$  for  $i \in [t-1]$  denotes the approximated distribution of uninstructed representations, which can be further specified as  $\hat{\mathcal{G}}_i = \bigcup_{c \in \mathcal{Y}_i} \hat{\mathcal{G}}_{i,c}$  for classification tasks.  $\lambda_{OOD}$  denotes the OOD detection threshold.  $\mathbf{1}(\cdot)$  is the indicator function.  $\text{Dis}(x, \hat{\mathcal{G}}_i)$  measures the distance of task-wise distributions, which can be implemented via the average Euclidean distance between  $f_{\theta,g}(x)$  and  $\hat{r} \sim \hat{\mathcal{G}}_i$ . Consequently, if  $x$  is identified as OOD for all tasks  $i \in [t-1]$ , then it will be associated with  $\mathbf{g}_{k+1}$ . Otherwise, it will retrieve the associated  $\mathbf{g}_j$  for  $j \in [k]$  corresponding to the majority of the most relevant task  $\hat{i} = \arg \min_{i \in [t-1]} \text{Dis}(x, \hat{\mathcal{G}}_i)$ . In particular, to overcome catastrophic forgetting when updating  $\mathbf{g}_j$ , we employ the same strategy as for learning the task-shared parameters  $\mathbf{g}$ , e.g., a combination of FSA [43] and SL [4]. As a special case, we have  $k=1$  if all input samples are identified as in-distribution, for which only  $\mathbf{g}_1$  exists and is equivalent to  $\mathbf{g}$ .

**Connection of PET Architectures:** We now extend the above discussion with the criterion of OOD detection in Eq. (18). Given a task sequence  $1, \dots, t$ , using a larger  $\lambda_{OOD}$  would make  $k \rightarrow 1$  and  $\{\mathbf{g}_1, \dots, \mathbf{g}_k\} \rightarrow \{\mathbf{g}\}$ , while using a smaller  $\lambda_{OOD}$  would make  $k \rightarrow t$  and  $\{\mathbf{g}_1, \dots, \mathbf{g}_k\} \rightarrow \{e_1, \dots, e_t\}$ . Therefore, the representation learning of HiDe-PET in Sec. 5.1 is equivalent to a parallel combination of these two extreme conditions for TII and WTP, respectively. This is a reasonable choice as most CL benchmarks employ randomly split classes of the same dataset as the task sequence, i.e., there is no actual task structure.

Instead, Eq. (18) applies to the apparent similarity and dissimilarity between task distributions, which is more realistic in applications and enables adaptive knowledge accumulation from CL for enhanced utilization. Here we leverage the unique property of LoRA-based PET to construct a specialized implementation, serving as a plug-in module for Algorithm 1. Unlike the commonly-used Prompt-based PET that updates only attached tokens, the LoRA-based PET specifies  $\mathbf{g}_1, \dots, \mathbf{g}_k$  as the approximated updates of  $\theta$ , where the most relevant  $\mathbf{g}_j$  is selected and temporarily added to  $\theta$  in CL. Therefore, the learning of subsequent tasks can

be significantly improved from the accumulated knowledge and further contribute to it (see Fig. 5). Moreover, this allows for the flexible evolution of pre-trained knowledge with target tasks in a lifelong manner, deviating from the conventional practice of fixing it at the initial checkpoint.

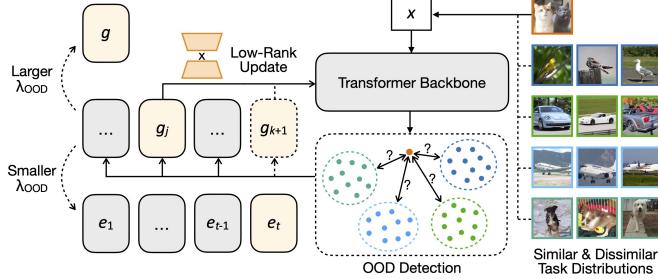


Fig. 5. Adaptive knowledge accumulation with HiDe-PET.

In brief, Sec. 4 and 5 serve as a unified framework for CL with PTMs and PET, applicable to explore the distinct impacts of implementation strategy, PET technique and PET architecture, as well as the adaptive knowledge accumulation for enhanced utilization.

## 6 EXPERIMENT

In this section, we perform extensive experiments to demonstrate the effectiveness and generality of our HiDe-PET. We first describe the experimental setups, and then present the experimental results with a comprehensive analysis.

### 6.1 Experimental Setup

To ensure the breadth and adequacy of the experiments, we consider a variety of CL benchmarks, pre-trained checkpoints, recent strong baselines, PET techniques and evaluation metrics. For the sake of comparison fairness, we follow the official implementations to reproduce all baselines.

**Benchmark:** We focus on four datasets that are widely used in previous work to evaluate CL [4]–[6], [8], [14], and split each into 10 target tasks of disjoint classes for CIL experiments. The first two are *general datasets*, such as CIFAR-100 [46] of 100-class small-scale images that are common objects in the real world and ImageNet-R [47] of 200-class large-scale images that are hard examples of ImageNet-21K [48] or newly collected examples of different styles. The latter two are *fine-grained datasets*, such as CUB-200 [49] of 200-class bird images and Cars-196 [50] of 196-type car images. We consider three pre-trained ViT-B/16 checkpoints that differ in paradigm and dataset, including Sup-21/1K, iBOT-21K and Sup-Weak. Specifically, Sup-21/1K [8] is essentially a supervised checkpoint that performs self-supervised learning on ImageNet-21K and then supervised fine-tuning on ImageNet-1K. iBOT-21K [51] is a self-supervised checkpoint that currently achieves the first-place classification performance for self-supervised learning on ImageNet-21K. Sup-Weak [52] is a supervised checkpoint on a subset of ImageNet-1K, in which 389 similar classes to subsequent CL are intentionally removed.

**Baseline:** We compare our HiDe-PET with a range of recent strong baselines as described in Sec. 3.3, including L2P [5], DualPrompt [6], S-Prompt [7], CODA-Prompt [8] and LAE [14]. In brief, these baselines cover different PET

architectures, but mainly target the Prompt-based PET. LAE [14] is the most recent baseline among them, and is also the most relevant one to ours as it applies to a variety of mainstream PET techniques. Similar to the previous work [13], we modify the implementation of S-Prompt [7] by inserting task-specific prompts into multiple MSA layers in a PreT manner and using a single-head output layer, in order to evaluate the impact of PET architecture and better adapt to the CIL experiments. The modified S-Prompt is referred to as S-Prompt++. Following LAE [14], we consider three kinds of mainstream PET techniques in our HiDe-PET and our preliminary version [13], including PreT [10], LoRA [12] and Adapter [11].

**Evaluation:** We use  $A_{i,i'}$  to denote the prediction accuracy on task  $i$  after learning task  $i'$  (with single-head evaluation for CIL), and define the average accuracy of all seen tasks as  $AA_{i'} = \frac{1}{i'} \sum_{i=1}^{i'} A_{i,i'}$ . After learning all tasks, we report both the final average accuracy  $FAA = AA_t$  that serves as the primary metric to evaluate CL performance, and the cumulative average accuracy  $CAA = \frac{1}{t} \sum_{i=1}^t AA_i$  that further reflects the historical performance. Moreover, we evaluate the behaviors of different PET techniques with the average learning accuracy  $ALA = \frac{1}{t-1} \sum_{i=2}^t A_{i,i-1}$  for learning plasticity and the final forgetting measure  $FFM = \frac{1}{t-1} \sum_{i=1}^{t-1} \max_{i' \in [t-1]} (A_{i,i'} - A_{i,t})$  for memory stability, as well as evaluate the TII performance in our HiDe-PET.

**Implementation:** We follow similar implementations as previous work.<sup>4</sup> Specifically, we employ a pre-trained ViT-B/16 backbone with an Adam optimizer ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ), a batch size of 128, and a cosine-decaying learning rate of 0.001. We train Split CIFAR-100 for 20 epochs and other benchmarks for 50 epochs to ensure convergence on each task. The image inputs are resized to  $224 \times 224$  and normalized to  $[0, 1]$ . The PET architecture of each baseline is also similar to its original paper, which has been shown to yield strong performance. Specifically, L2P [5] is implemented with the prompt pool size  $M = 30$ , the prompt length  $d_p = 5$  and the Top-5 keys. DualPrompt [6] is implemented with the prompt length  $d_p = 5$  of task-shared prompts inserted into layers 1-2 and the prompt length  $d_p = 20$  of task-specific prompts inserted into layers 3-5. S-Prompt++ [7] is implemented similarly to DualPrompt but replaces all task-shared prompts with task-specific prompts, inserted into layers 1-5 with the prompt length  $d_p = 20$ . CODA-Prompt [8] is implemented with the prompt pool size  $M = 100$  and the prompt length  $d_p = 8$ , inserted into layers 1-5. LAE [14] and our HiDe-PET are implemented with the prompt length  $d_p = 20$  for PreT, and the low-dimension bottleneck  $r = 10$  for Adapter and LoRA, inserted into layers 1-5. We insert the Adapter modules in both sequential and parallel manners, while employ LoRA to update both  $W_K$  and  $W_V$ . Therefore, the extra parameter costs of PreT, Adapter and LoRA are identical [14]. Note that, our HiDe-PET and our preliminary version [13] adopt a similar PET architecture as S-Prompt++, but replace the task-specific keys with an auxiliary output layer  $h_\omega$  to predict the task identity and further preserve

4. The training regime and supervised checkpoints are identical to those in CODA-Prompt [8], which are slightly different from those in our preliminary version [13] and lead to some performance differences.

TABLE 1

Overall performance of continual learning. PTM: pre-trained model. FAA (%): final average accuracy. CAA (%): cumulative average accuracy.

PTM	Method	Split CIFAR-100		Split ImageNet-R		Split CUB-200		Split Cars-196	
		FAA ( $\uparrow$ )	CAA ( $\uparrow$ )	FAA ( $\uparrow$ )	CAA ( $\uparrow$ )	FAA ( $\uparrow$ )	CAA ( $\uparrow$ )	FAA ( $\uparrow$ )	CAA ( $\uparrow$ )
Sup-21/1K	L2P [5]	84.25	88.84	71.34	76.87	70.90	76.70	41.06	46.47
	DualPrompt [6]	83.75	89.11	71.65	77.51	68.21	75.15	42.68	51.60
	S-Prompt++ [7]	82.41	87.68	71.15	77.16	68.01	75.04	39.62	47.85
	CODA-Prompt [8]	86.65	90.78	75.11	81.45	71.43	78.61	45.67	53.28
	LAE-PreT [14]	87.36	91.63	74.95	81.23	78.46	83.65	42.80	52.12
	LAE-LoRA [14]	88.38	92.45	76.27	82.99	80.02	84.47	50.90	58.38
	LAE-Adapter [14]	88.37	92.50	75.69	82.80	80.52	84.75	55.20	61.63
	HiDe-PreT	91.11	94.11	78.93	83.44	87.95	88.48	68.73	69.19
	HiDe-LoRA	91.21	93.99	<b>79.32</b>	<b>83.97</b>	<b>88.76</b>	<b>89.32</b>	69.65	69.36
	HiDe-Adapter	<b>91.23</b>	<b>94.26</b>	78.65	83.55	88.49	89.17	<b>70.98</b>	<b>71.31</b>
iBOT-21K	L2P [5]	79.32	85.13	61.31	70.05	45.93	56.02	45.25	45.75
	DualPrompt [6]	78.17	85.15	61.42	70.06	41.46	54.57	34.61	42.28
	S-Prompt++ [7]	79.85	85.89	60.84	69.01	39.88	53.71	36.46	43.34
	CODA-Prompt [8]	81.58	87.36	67.15	76.54	47.79	59.24	39.50	43.32
	LAE-PreT [14]	82.22	88.05	65.85	75.34	45.83	60.31	49.14	52.59
	LAE-LoRA [14]	84.63	90.24	70.49	79.06	56.16	68.38	58.66	62.59
	LAE-Adapter [14]	84.68	90.31	69.93	79.14	58.04	70.01	61.76	65.61
	HiDe-PreT	88.13	92.17	70.57	77.89	70.72	74.09	63.98	64.18
	HiDe-LoRA	<b>89.72</b>	<b>93.34</b>	<b>74.46</b>	<b>80.89</b>	<b>76.10</b>	<b>79.99</b>	67.73	68.64
	HiDe-Adapter	89.46	93.12	74.25	80.48	75.17	79.42	<b>69.62</b>	<b>70.11</b>
Sup-Weak	L2P [5]	67.73	78.84	47.95	56.51	43.99	58.85	33.25	38.97
	DualPrompt [6]	69.09	79.56	51.21	59.67	46.05	58.51	35.08	42.99
	S-Prompt++ [7]	71.05	81.34	47.87	56.62	42.91	57.70	36.20	43.35
	CODA-Prompt [8]	65.45	76.43	53.21	63.61	44.91	57.73	35.59	41.90
	LAE-PreT [14]	67.25	77.34	55.55	64.78	48.56	61.73	36.63	41.56
	LAE-LoRA [14]	68.43	78.57	57.40	66.84	48.99	61.50	35.35	39.93
	LAE-Adapter [14]	68.55	78.59	57.92	67.79	49.79	62.25	37.17	41.72
	HiDe-PreT	<b>77.65</b>	<b>85.14</b>	57.98	65.79	65.03	71.63	52.89	55.09
	HiDe-LoRA	77.46	84.89	<b>59.40</b>	<b>67.05</b>	<b>66.84</b>	<b>71.91</b>	52.61	54.78
	HiDe-Adapter	76.71	84.55	58.94	67.53	66.26	71.24	<b>54.38</b>	<b>56.23</b>

statistical information of pre-trained representations.<sup>5</sup>

## 6.2 Experimental Result

Now we present the results of our empirical investigation, including the overall performance of all methods, an ablation study of the three hierarchical components, the distinct impacts of implementation strategy, PET technique and PET architecture, as well as the adaptive knowledge accumulation over similar and dissimilar tasks.

**Overall Performance:** Table 1 summarizes the results of all methods across various pre-trained checkpoints and CL benchmarks. It can be seen that our HiDe-PET implemented with three mainstream PET techniques achieves consistently the highest performance, and the performance lead tends to be more pronounced under the more challenging scenarios. Specifically, the performance of all methods is affected to varying degrees when considering fine-grained tasks (i.e., CUB-200 and Cars-196) and weakened pre-training in terms of self-supervised paradigm (i.e., iBOT-21K) and reduced pre-training samples (i.e., Sup-Weak). Among these competitors, the sub-optimality of Prompt-based PET in CL

5. We consider a lightweight implementation in terms of the auxiliary output layer and representation recovery, which slightly compromise the performance but largely improve resource efficiency.

is clearly exposed, which underperforms LoRA/Adapter-based PET within and across methods. The LoRA/Adapter version of LAE [14] is the strongest competitor but still severely affected by the double challenges of pre-trained checkpoints and CL benchmarks. In contrast, our HiDe-PET adapts to them effectively with strong generality.

It is noteworthy that self-supervised pre-training is often considered more practical than supervised pre-training, owing to the expense of annotating massive pre-training samples [4], [13]. Meanwhile, Sup-Weak avoids potential overlap between PTMs and target tasks, providing a more restrictive scenario for CL [52]. Sup-Weak is also more analogous to the widely used setting of CIL experiments in literature, i.e., the model first learns half of the classes and then learns the other classes in multiple incremental phases, where the baselines of Prompt-based PET have been shown to perform poorly on it [53]. These considerations underscore more profound advantages of our HiDe-PET in CL. Moreover, our HiDe-PET clearly outperforms its preliminary version [13] under the same implementation (see Fig. 6, +1.98% / +2.56% / +2.66% / +5.63% FAA on the four CL benchmarks in average), which will be discussed latter.

**Ablation Study:** Table 2 presents an ablation study to validate the effectiveness of optimizing the three hierarchi-

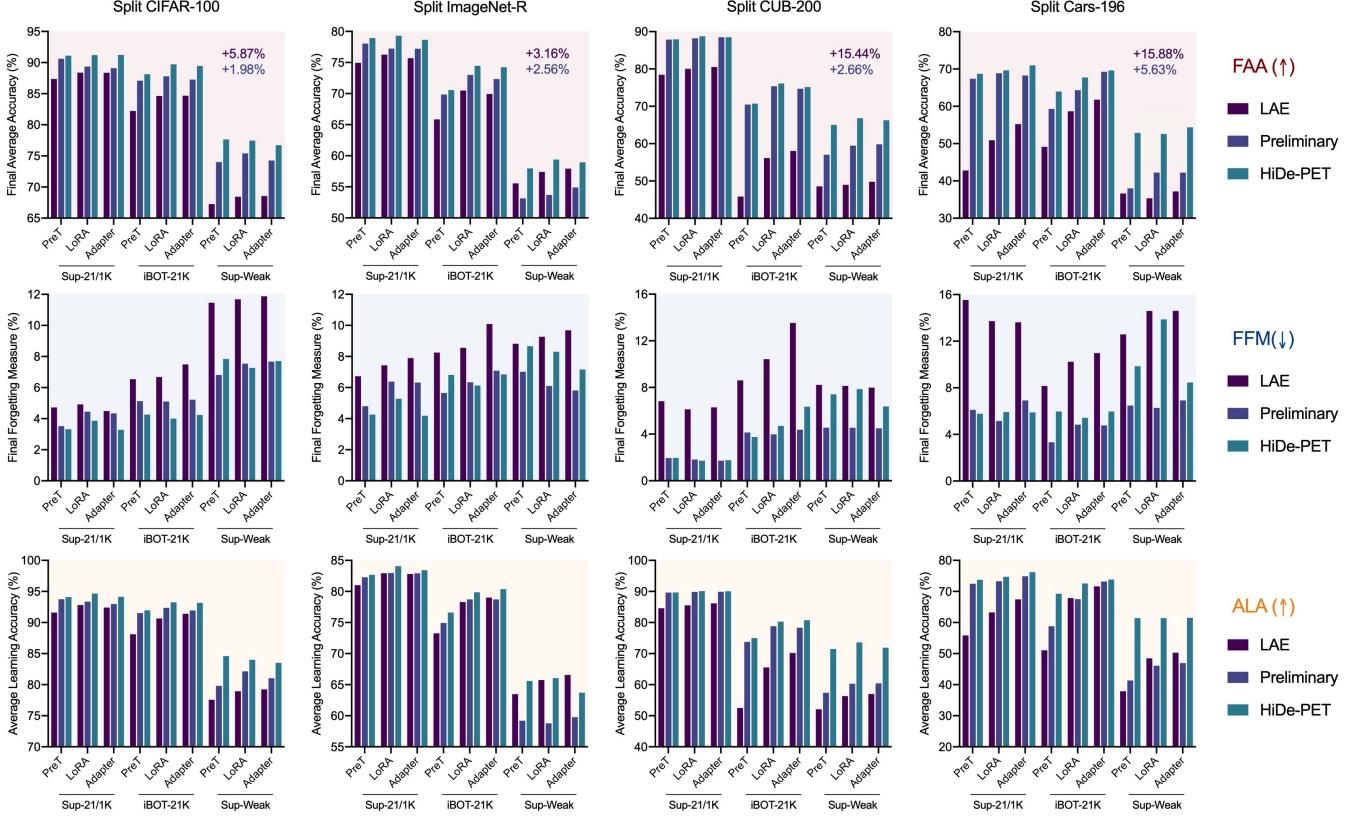


Fig. 6. Comparison of our HiDe-PET, our preliminary version [13] and LAE [14] implemented with different PET techniques. FAA (%): final average accuracy. FFM (%): final forgetting measure of old tasks. ALA (%): average learning accuracy of new tasks. We present in the first row the average FAA lead of our HiDe-PET over LAE (dark purple) and our preliminary version (dark blue).

cal components in HiDe-PET. Specifically, we progressively incorporate the designs of within-task prediction (WTP), task-identity inference (TII) and task-adaptive prediction (TAP) on the top of a naive architecture, which consists of only task-specific parameters  $e_1, \dots, e_t$  implemented via mainstream PET techniques. In general, the optimization of each component all contributes to the strong performance of HiDe-PET. Although their contributions are relatively comparable under supervised pre-training and general tasks, the improvement of TAP becomes more significant under self-supervised pre-training and fine-grained tasks, demonstrating the necessity of TAP within the CL objective. Besides, the improvement of TII often becomes more apparent with WTP+TAP rather than with WTP alone, suggesting that WTP, TII and TAP operate in concert rather than in isolation.

**Implementation Strategy:** Now we evaluate the implementation strategy of task-shared parameters and representation recovery, which are both important for the optimization of our HiDe-PET and potentially shared by many recent methods. In contrast to task-specific parameters discussed above, task-shared parameters  $\mathbf{g}$  aim to improve pre-trained representations in a task-agnostic manner, demanding effective strategies to mitigate catastrophic forgetting. Various strategies have been employed in previous work, including (1) fix-and-tuning (F&T) [14], which updates the output layer with frozen  $\mathbf{g}$  in earlier epochs and then updates  $\mathbf{g}$  for representation learning in later epochs; (2) first-session adaptation (FSA) [43], which updates  $\mathbf{g}$  for representation learning exclusively from the first task and then fixes  $\mathbf{g}$  in subsequent tasks; (3) slow learner (SL) [4], which reduces

the learning rate of  $\mathbf{g}$  in all tasks; and (4) exponential moving average (EMA) [14], which employs an interim copy of  $\mathbf{g}$  to learn each task and then updates  $\mathbf{g}$  with a small momentum.

However, most of these strategies have their own limitations. Both F&T and SL impose restrictions on the extent of updates, sacrificing the effectiveness of representation learning and suffering from potential representation shifts in subsequent recovery. FSA adeptly integrates knowledge from the first task and completely avoids representation shifts, but is unable to perform subsequent representation learning. Considering their complementary properties, we propose a simple but effective strategy that employs FSA for learning the first task and SL for learning subsequent tasks, which clearly outperforms other strategies (see Table 3). Notably, EMA can be seen as a coarse implementation of FSA+SL and indeed achieves the second-highest performance. Therefore, the task-shared parameters in HiDe-PET may also be implemented with EMA by updating  $\mathbf{g}$  from each  $e_i$  for  $i \in [t]$ , offering a slight compromise in performance but reducing a half of the training cost.

On the other hand, we evaluate effective strategies for representation recovery. For classification tasks, the pre-trained representations of each class tend to be single-peaked and therefore can be modeled as a Gaussian with dedicated mean and covariance. Although the covariance achieves considerable performance as shown in Table 4, it requires a storage complexity  $O(d^2)$  for embedding dimension  $d$ , which is comparable to the MSA layer of the backbone. There are three alternatives that reduces the storage complexity to  $O(d)$ , such as simplifying the co-

TABLE 2

Ablation study of the three hierarchical components in HiDe-PET.  
Naive: a naive baseline of only task-specific parameters.

Setup	Method	Split ImageNet-R		Split Cars-196	
		FAA ( $\uparrow$ )	CAA ( $\uparrow$ )	FAA ( $\uparrow$ )	CAA ( $\uparrow$ )
Sup-21/1K PreT	Naive	69.77	76.36	46.08	53.59
	WTP	73.01	78.20	48.23	54.11
	WTP+TII	75.68	80.95	52.89	54.18
	WTP+TAP	78.06	82.69	61.38	64.46
	WTP+TII+TAP	<b>78.93</b>	<b>83.44</b>	<b>68.73</b>	<b>69.19</b>
Sup-21/1K LoRA	Naive	74.54	80.66	45.39	54.28
	WTP	75.59	81.31	48.01	53.85
	WTP+TII	76.03	81.69	49.62	57.29
	WTP+TAP	78.33	83.68	63.28	65.87
	WTP+TII+TAP	<b>79.32</b>	<b>83.97</b>	<b>69.65</b>	<b>69.36</b>
Sup-21/1K Adapter	Naive	75.17	81.46	47.20	55.69
	WTP	76.10	82.23	53.12	59.35
	WTP+TII	76.80	82.60	55.93	60.89
	WTP+TAP	76.98	82.73	64.65	67.38
	WTP+TII+TAP	<b>78.65</b>	<b>83.55</b>	<b>70.98</b>	<b>71.31</b>
iBOT-21K PreT	Naive	63.78	73.47	41.54	47.96
	WTP	64.98	73.54	52.99	56.31
	WTP+TII	66.33	74.98	53.89	57.01
	WTP+TAP	69.84	77.02	59.75	61.28
	WTP+TII+TAP	<b>70.57</b>	<b>77.89</b>	<b>63.98</b>	<b>64.18</b>
iBOT-21K LoRA	Naive	67.07	77.07	53.13	59.07
	WTP	68.06	77.39	56.03	61.18
	WTP+TII	68.54	77.60	59.48	63.36
	WTP+TAP	72.60	79.95	61.50	64.88
	WTP+TII+TAP	<b>74.46</b>	<b>80.89</b>	<b>67.73</b>	<b>68.64</b>
iBOT-21K Adapter	Naive	68.17	77.57	53.58	59.69
	WTP	69.11	77.29	57.71	62.21
	WTP+TII	69.65	77.90	62.12	65.66
	WTP+TAP	71.32	79.17	62.78	65.59
	WTP+TII+TAP	<b>74.25</b>	<b>80.48</b>	<b>69.62</b>	<b>70.11</b>

variance to variance, preserving randomly selected prototypes, and obtaining multiple representation centroids with  $k$ NNs. Among them, the multi-centroid demonstrates superior performance and is applicable to different task types, which therefore becomes our default implementation. Interestingly, the variance achieves comparable performance as the covariance and the multi-centroid under general tasks (i.e., Split ImageNet-R) and supervised pre-training (i.e., Sup-21/1K) while requires negligible parameter costs. This further strengthens the advantages of our HiDe-PET in such scenarios targeted by previous work.

**PET Technique:** While mainstream PET techniques universally amount to modulating specific hidden states of the PTMs [33], their potential differences in CL are noteworthy. As mentioned above, Prompt-based PET usually lags behind LoRA/Adapter-based PET for both LAE and HiDe-PET (see Table 1 and Fig. 6). The performance gap tends to be more pronounced under the more challenging scenarios of pre-trained checkpoints and CL benchmarks. Such differences in CL could stem from the limited capacity of Prompt-based PET in handling self-supervised pre-training [36] and fine-grained tasks [40], as validated by our results for both task-specific parameters (see WTP in Table 2) and task-shared parameters (see FSA+SL in Table 3). This reinforces the necessity of embracing more advanced PET techniques, as our HiDe-PET does.

Beyond the overall performance, the choice of PET

TABLE 3  
Comparison of different strategies for learning task-shared parameters.  
TII (%): performance of task identity inference. FAA-U (%): final average accuracy of learning all classes from uninstructed representations. F&T: fix-and-tuning. FSA: first-session adaptation. SL: slow learner. EMA: exponential moving average.

Setup	Method	Split ImageNet-R		Split Cars-196	
		TII ( $\uparrow$ )	FAA-U ( $\uparrow$ )	TII ( $\uparrow$ )	FAA-U ( $\uparrow$ )
Sup-21/1K PreT	F&T [14]	76.45	74.50	59.46	52.94
	FSA [43]	75.85	73.76	68.42	58.85
	SL [4]	77.06	74.68	66.13	56.67
	EMA [14]	76.17	73.93	68.35	58.99
	FSA+SL	<b>77.15</b>	<b>75.02</b>	<b>69.43</b>	<b>59.32</b>
Sup-21/1K LoRA	F&T [14]	71.90	69.85	64.65	57.68
	FSA [43]	77.74	75.72	70.90	61.37
	SL [4]	77.26	75.41	68.68	59.33
	EMA [14]	78.33	<b>76.51</b>	71.20	61.87
	FSA+SL	<b>78.43</b>	76.35	<b>71.92</b>	<b>62.89</b>
Sup-21/1K Adapter	F&T [14]	75.45	73.88	57.16	52.11
	FSA [43]	78.15	76.30	72.75	63.51
	SL [4]	77.52	75.98	63.20	56.16
	EMA [14]	78.29	76.30	73.59	64.46
	FSA+SL	<b>80.09</b>	<b>78.52</b>	<b>73.71</b>	<b>64.93</b>

TABLE 4  
Comparison of different strategies for representation recovery. We set  $k = 10$  for  $k$ NNs to obtain Multi-Centroid. #Param: average parameter costs per class, where  $d = 768$  in this case.

Setup	Method	Split ImageNet-R		Split Cars-196	
		FAA ( $\uparrow$ )	#Param ( $\downarrow$ )	FAA ( $\uparrow$ )	#Param ( $\downarrow$ )
Sup-21/1K PreT	No Recovery	75.68	0	52.89	0
	Prototype	76.88	10d	62.65	10d
	Variance	77.54	1d	57.04	1d
	Covariance	77.58	$d^2$	<b>73.14</b>	$d^2$
	Multi-Centroid	<b>78.93</b>	9.5d	68.73	8.0d
iBOT-21K PreT	No Recovery	58.88	0	41.89	0
	Prototype	67.08	10d	46.34	10d
	Variance	70.55	1d	48.27	1d
	Covariance	68.85	$d^2$	<b>66.42</b>	$d^2$
	Multi-Centroid	<b>70.57</b>	9.5d	63.98	8.7d
Sup-Weak PreT	No Recovery	54.63	0	44.35	0
	Prototype	55.06	10d	46.91	10d
	Variance	55.49	1d	47.77	1d
	Covariance	57.46	$d^2$	<b>56.06</b>	$d^2$
	Multi-Centroid	<b>57.98</b>	9.1d	52.89	8.7d

technique also exerts distinct influences on the three hierarchical components. Compared to Prompt-based PET, LoRA/Adapter-based PET excels in WTP performance through more effectively incorporating task-specific knowledge, but reveals a heightened sensitivity to TII performance, manifested in the errors of predicting an inappropriate set of task-specific parameters. This effect is further compensated by the TAP performance. As shown in Table 2, the effectiveness of TII is remarkably pronounced when coupled with WTP+TAP for LoRA/Adapter-based PET, whereas diminishes for either Prompt-based PET or WTP alone. Moreover, our HiDe-PET outperforms its preliminary version [13] especially in LoRA/Adapter-based PET and the more challenging scenarios (see Fig. 6), thanks to the improved TII performance with task-shared parameters.

**PET Architecture:** The generality of our HiDe-PET is also reflected in its PET architecture, which strategically exploits both task-specific and task-shared parameters for representation learning. These two kinds of parameters are

TABLE 5

Evaluation of adaptive knowledge accumulation (AKA) with LoRA-based PET. Full (%): average accuracy of learning subsequent tasks with all training samples. Few (%): average accuracy of learning subsequent tasks with a few training samples (5 per class).

PTM	Method	Split Dogs-120		Split CUB-200		Split Cars-196		Split Aircraft-102		CL of Mixture	
		Full ( $\uparrow$ )	Few ( $\uparrow$ )	Full ( $\uparrow$ )	Few ( $\uparrow$ )	Full ( $\uparrow$ )	Few ( $\uparrow$ )	Full ( $\uparrow$ )	Few ( $\uparrow$ )	FAA ( $\uparrow$ )	CAA ( $\uparrow$ )
Sup-21/1K	w/o AKA	92.32	88.36	89.51	81.26	83.77	57.04	77.98	55.56	77.32	81.14
	w/ AKA	92.50	88.92	91.39	84.91	89.46	64.30	83.38	62.37	83.27	86.78
iBOT-21K	w/o AKA	81.70	54.08	74.66	45.79	79.27	38.15	79.24	53.97	68.38	74.57
	w/ AKA	84.41	65.10	82.95	62.30	88.32	56.12	85.06	62.76	74.99	81.55
Sup-Weak	w/o AKA	88.14	80.88	71.10	47.46	59.03	36.51	62.68	35.78	53.23	58.54
	w/ AKA	88.17	81.38	77.50	56.92	77.42	50.37	72.64	46.32	65.48	70.41

used to acquire knowledge with different levels of differentiation and need to overcome their respective challenges as described in Sec. 3.3. Within HiDe-PET, they both contribute to the outstanding performance in Table 1 and complement each other (see WTP in Table 2 and FSA+SL in Table 3). In contrast, our preliminary version [13] and LAE [14] exclusively engage either task-specific or task-shared parameters, missing out on fully harnessing the benefits of PTMs and PET. We further present an extensive comparison of our HiDe-PET, our preliminary version and LAE in terms of the overall performance, memory stability and learning plasticity, so as to better demonstrate the respective contributions of different PET architectures (see Fig. 6).

It is noteworthy that previous work such as L2P [5] and DualPrompt [6] also explicitly or implicitly exploit both task-specific and task-shared prompts, but in a *sequential* manner to instruct representation learning of each task (corresponding to WTP in our framework). In contrast, our HiDe-PET optimizes these two kinds of parameters in a *parallel* manner to improve the three hierarchical components, allowing for a more adequate differentiation of the acquired knowledge. Interestingly, using only task-shared parameters coupled with representation recovery within HiDe-PET (i.e., FSA+SL in Table 3) has already achieved better performance than these methods (see Sec. 7 for a conceptual explanation), serving as a strong baseline to evaluate current progress. The inherent connections of task-specific and task-shared parameters will be further explored below with a PET hierarchy inspired by OOD detection.

**Adaptive Knowledge Accumulation:** As analyzed in Sec. 5.2, the use of  $e_1, \dots, e_t$  and  $g$  can be seen as a special case tailored for target tasks randomly split from the same dataset, i.e., there is no actual task structure. When considering more realistic CL scenarios with apparent similarity and dissimilarity between task distributions, we devise a hierarchy of expandable parameter sets  $g_1, \dots, g_k$  upon OOD detection to achieve adaptive knowledge accumulation (AKA), and focus on the LoRA-based PET to examine if the pre-trained knowledge can evolve flexibly with target tasks in CL. Here we construct such scenario with the two fine-grained datasets (i.e., CUB-200 [49] and Cars-196 [50]) and another two (i.e., Dogs-120 [54] and Aircraft-102 [55]), which cover both natural and artificial objects. Each dataset is randomly split into 10 tasks. We collect 5 tasks per dataset and mix them as a task sequence (20 tasks in total) for CL, while leaving the rest 5 tasks per dataset for validation.

In CL, the OOD detection threshold  $\lambda_{OOD}$  determines the expansion of parameter sets. As shown in Fig. 7, using a larger  $\lambda_{OOD}$  tends to expand fewer parameter sets, and

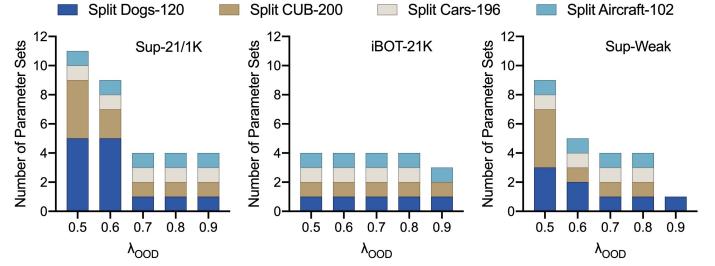


Fig. 7. OOD detection threshold  $\lambda_{OOD}$  and the number of expanded parameter sets  $g_1, \dots, g_k$  under different pre-trained checkpoints.

vice versa. In particular, the choice of  $\lambda_{OOD}$  is relatively insensitive and can consistently construct one parameter set for each dataset (with  $\lambda_{OOD} = 0.7$  or 0.8) under different pre-trained checkpoints. Then we validate the effectiveness of adaptive knowledge accumulation in Table 5. Inspired by a recent work [56], we evaluate the improvements of pre-trained knowledge through the average accuracy of learning the validation tasks under large-shot or few-shot setting. Through selecting the most relevant  $g_j$  and adding it to  $\theta$ , the pre-trained backbone  $f_\theta$  is able to learn each task more effectively. With the improved  $f_\theta$ , the overall performance of CL (i.e., FAA and CAA) for the mixed task sequence is also significantly enhanced.

Interestingly, the idea of updating the pre-trained backbone with a mixture of LoRA experts [57] has been shown effective to accumulate knowledge from multi-task learning, which is consistent with our results. In contrast, the design of the OOD detection, FSA+SL, and representation recovery enables our HiDe-PET to achieve this purpose in a lifelong manner. Besides, our HiDe-PET can also adapt to task-agnostic CL [3] through expanding  $e_1, \dots, e_t$  upon the OOD detection. We leave it as a further work.

## 7 DISCUSSION AND CONCLUSION

In this work, we present a unified framework for CL with PTMs and PET, in order to advance this direction with improved effectiveness and generality. Our framework features a profound integration of theoretical and empirical insights, a broad coverage of relevant techniques, as well as a robust adaptation to different scenarios. Considering the particular impact of pre-trained knowledge on CL, we decompose the CL objective into three hierarchical components, i.e., WTP, TII and TAP, and devise an innovative approach to explicitly optimize them with mainstream PET techniques. During the optimization process, pre-trained representations are effectively instructed via task-specific

and task-shared PET architectures, and are efficiently recovered through preserving their statistical information.

Our framework allows for a comprehensive evaluation of various technical elements inherent in CL with PTMs and PET. Through an extensive empirical investigation, we demonstrate the better performance of LoRA/Adapter-based PET over Prompt-based PET within both task-specific and task-shared PET architectures, which tends to be more evident under the more challenging scenarios in terms of pre-trained checkpoints and CL benchmarks. We also unravel the distinct behaviors of different PET techniques in response to the three hierarchical components, as well as the respective challenges and complementary effects of different PET architectures. These technical elements are potentially shared by many recent methods, making it possible to dissect their specific implementations and incorporate the most appropriate ones. Owing to the above extensive explorations, our approach achieves remarkably superior performance across various CL scenarios over a wide range of recent strong baselines.

Intriguingly, the correspondence of our approach to the three hierarchical components suggests a more profound connection between existing methods. As discussed in Sec. 3.3, the use of task-specific parameters [5]–[8], [13] necessitates learning to predict their identities, equivalent to optimizing the decomposed WTP performance  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)$  and TII performance  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)$ . In contrast, the use of task-shared parameters [4]–[6], [14], [34] needs to overcome catastrophic forgetting, equivalent to optimizing the pre-decomposed performance  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathcal{D}, \theta)$ . On the top of representation learning, the use of representation recovery [4], [34], [44] to rectify the output layer further improves the TAP performance  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$ . This connection is summarized by the multi-objective optimization problem in Eq. (10), and also demonstrates why our approach clearly outperforms other baselines and why the use of only task-shared parameters and representation recovery (i.e., FSA+SL in Table 3) is powerful enough. Subsequent efforts in CL with PTMs and PET could employ this as a theoretical reference to develop more advanced methods.

Moreover, the hierarchical decomposition along with the design of our approach showcase a close relationship with the mechanisms of robust biological CL. In the mammalian brain, the memory of an experience is consolidated with the interplay of hippocampus and neocortex, known as the complementary learning system theory [58], [59] that has been widely used to inspire CL in artificial intelligence (AI). The hippocampus-dependend and neocortex-dependend memories tend to be more specific and more generalized, respectively [60], [61], and the retrieval of these two memory paths is adaptively switched from the concrete scenarios [62]. Within hippocampus, the activation of distinct populations of memory cells also undergoes adaptive switching [63], and the neural representations of previous experiences are frequently recovered [64]. The entire process is consistent with the parallel organization of task-specific and task-shared parameters, the exclusive selection of the former and the representation recovery of task distributions.

In the era of large-scale PTMs, we would emphasize the pressing need for these adaptive algorithms that are

designed with machine learning fundamentals and real-world considerations. By leveraging the power of PTMs and the adaptability of CL, we can customize solutions to address the unique challenges posed by specific domains, and envision extending our approach to numerous areas such as healthcare, robotics and industrial manufacturing. Such an elevated goal requires extending the target of CL from homogeneous to heterogeneous tasks, which also provides novel opportunities to explore generalizable knowledge behind them. Taken together, we expect this work to not only facilitate direct applications but also set the stage for the robustness, adaptability and reliability of future AI systems, as a general purpose of CL research.

## ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106302), NSFC Projects (Nos. 62061136001, 92248303, 62106123, 61972224), BNRist (BNR2022RC01006), Tsinghua Institute for Guo Qiang, and the High Performance Computing Center, Tsinghua University. L.W. is also supported by the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20230350 and the Shuimu Tsinghua Scholar. J.Z. is also supported by the XPlorer Prize.

## REFERENCES

- [1] V. V. Ramasesh, A. Lewkowycz, and E. Dyer, "Effect of scale on catastrophic forgetting in neural networks," in *ICLR*, 2021.
- [2] S. V. Mehta *et al.*, "An empirical investigation of the role of pre-training in lifelong learning," *arXiv preprint arXiv:2112.09153*, 2021.
- [3] L. Wang *et al.*, "A comprehensive survey of continual learning: Theory, method and application," *IEEE TPAMI*, 2024.
- [4] G. Zhang *et al.*, "Slca: Slow learner with classifier alignment for continual learning on a pre-trained model," in *ICCV*, 2023.
- [5] Z. Wang *et al.*, "Learning to prompt for continual learning," in *CVPR*, 2022.
- [6] Z. Wang *et al.*, "Dualprompt: Complementary prompting for rehearsal-free continual learning," in *ECCV*, 2022.
- [7] Y. Wang, Z. Huang, and X. Hong, "S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning," *NeurIPS*, 2022.
- [8] J. S. Smith *et al.*, "Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning," in *CVPR*, 2023.
- [9] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *EMNLP*, 2021.
- [10] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL-IJCNLP*, 2021.
- [11] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *NeurIPS*, 2017.
- [12] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [13] L. Wang *et al.*, "Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality," *NeurIPS*, 2023.
- [14] Q. Gao *et al.*, "A unified continual learning framework with general parameter-efficient tuning," in *ICCV*, 2023.
- [15] G. I. Parisi *et al.*, "Continual lifelong learning with neural networks: A review," *Neur. Netw.*, 2019.
- [16] J. Kirkpatrick *et al.*, "Overcoming catastrophic forgetting in neural networks," *PNAS*, 2017.
- [17] L. Wang *et al.*, "Afec: Active forgetting of negative transfer in continual learning," *NeurIPS*, 2021.
- [18] R. Aljundi *et al.*, "Memory aware synapses: Learning what (not) to forget," in *ECCV*, 2018.
- [19] S.-A. Rebuffi *et al.*, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017.
- [20] H. Shin *et al.*, "Continual learning with deep generative replay," *NeurIPS*, 2017.

- [21] L. Wang *et al.*, "Memory replay with data compression for continual learning," in *ICLR*, 2021.
- [22] Q. Pham, C. Liu, and S. Hoi, "Dualnet: Continual learning, fast and slow," *NeurIPS*, 2021.
- [23] H. Cha, J. Lee, and J. Shin, "Co2l: Contrastive continual learning," in *ICCV*, 2021.
- [24] O. Ostapenko *et al.*, "Foundational models for continual learning: An empirical study of latent replay," in *CoLLAs*, 2022.
- [25] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *NeurIPS*, 2017.
- [26] S. Wang *et al.*, "Training networks in null space of feature covariance for continual learning," in *CVPR*, 2021.
- [27] G. Saha, I. Garg, and K. Roy, "Gradient projection memory for continual learning," in *ICLR*, 2020.
- [28] J. Serra *et al.*, "Overcoming catastrophic forgetting with hard attention to the task," in *ICML*, 2018.
- [29] L. Wang *et al.*, "Coscl: Cooperation of small continual learners is stronger than a big one," in *ECCV*, 2022.
- [30] L. Wang *et al.*, "Incorporating neuro-inspired adaptability for continual learning in artificial intelligence," *Nat. Mach. Intell.*, 2023.
- [31] Y. Wu *et al.*, "Large scale incremental learning," in *CVPR*, 2019.
- [32] J. Knoblauch, H. Husain, and T. Diethé, "Optimal continual learning has perfect memory and is np-hard," in *ICML*, 2020.
- [33] J. He *et al.*, "Towards a unified view of parameter-efficient transfer learning," in *ICLR*, 2021.
- [34] M. D. McDonnell *et al.*, "Ranpac: Random projections and pre-trained models for continual learning," *NeurIPS*, 2023.
- [35] M. Jia *et al.*, "Visual prompt tuning," in *ECCV*, 2022.
- [36] S. Yoo *et al.*, "Improving visual prompt tuning for self-supervised vision transformers," in *ICML*, 2023.
- [37] G. M. Van de Ven and A. S. Tolias, "Three scenarios for continual learning," *arXiv preprint arXiv:1904.07734*, 2019.
- [38] A. Vaswani *et al.*, "Attention is all you need," *NeurIPS*, 2017.
- [39] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.
- [40] X. Ma *et al.*, "When visual prompt tuning meets source-free domain adaptive semantic segmentation," *NeurIPS*, 2023.
- [41] G. Kim *et al.*, "A theoretical study on solving continual learning," *NeurIPS*, 2022.
- [42] J. Yang *et al.*, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021.
- [43] A. Panos *et al.*, "First session adaptation: A strong replay-free baseline for class-incremental learning," *arXiv preprint arXiv:2303.13199*, 2023.
- [44] Q. Tran *et al.*, "Koppa: Improving prompt-based continual learning with key-query orthogonal projection and prototype-based one-versus-all," *arXiv preprint arXiv:2311.15414*, 2023.
- [45] Y. Sun *et al.*, "Out-of-distribution detection with deep nearest neighbors," in *ICML*, 2022.
- [46] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Tech. Rep.*, 2009.
- [47] D. Hendrycks, S. Basart *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *ICCV*, 2021.
- [48] T. Ridnik *et al.*, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.
- [49] C. Wah *et al.*, "The caltech-ucsd birds-200-2011 dataset," 2011.
- [50] J. Krause *et al.*, "3d object representations for fine-grained categorization," in *ICCVW*, 2013.
- [51] J. Zhou *et al.*, "Image bert pre-training with online tokenizer," in *ICLR*, 2021.
- [52] G. Kim, B. Liu, and Z. Ke, "A multi-head model for continual learning via out-of-distribution replay," in *CoLLAs*, 2022.
- [53] Y.-M. Tang, Y.-X. Peng, and W.-S. Zheng, "When prompt-based incremental learning does not meet strong pretraining," in *ICCV*, 2023.
- [54] A. Khosla *et al.*, "Novel dataset for fine-grained image categorization: Stanford dogs," in *CVPRW*, 2011.
- [55] S. Maji *et al.*, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
- [56] W. Liao *et al.*, "Does continual learning meet compositionality? new benchmarks and an evaluation framework," *NeurIPS*, 2023.
- [57] X. Wu, S. Huang, and F. Wei, "Mole: Mixture of lora experts," in *ICLR*, 2023.
- [58] D. Kumaran, D. Hassabis, and J. L. McClelland, "What learning systems do intelligent agents need? complementary learning systems theory updated," *Trends Cogn. Sci.*, 2016.
- [59] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory," *Psychol. Rev.*, 1995.
- [60] P. W. Frankland and B. Bontempi, "The organization of recent and remote memories," *Nature Rev. Neurosci.*, 2005.
- [61] L. Wang *et al.*, "Triple-memory networks: A brain-inspired method for continual learning," *IEEE TNNLS*, 2021.
- [62] I. Goshen *et al.*, "Dynamics of retrieval strategies for remote memories," *Cell*, 2011.
- [63] B. Lei *et al.*, "Social experiences switch states of memory engrams through regulating hippocampal rac1 activity," *PNAS*, 2022.
- [64] D. Kudithipudi *et al.*, "Biological underpinnings for lifelong learning machines," *Nat. Mach. Intell.*, 2022.
- [65] Y. Zuo *et al.*, "Hierarchical prompts for rehearsal-free continual learning," *arXiv preprint arXiv:2401.11544*, 2024.



TNNLS, NeurIPS, ICLR, CVPR, ICCV, ECCV, etc.

**Liyuan Wang** is currently a postdoctoral researcher in Tsinghua University, working with Prof. Jun Zhu at the Department of Computer Science and Technology. Before that, he received the B.S. and Ph.D. degrees from Tsinghua University. His research interests include continual learning, incremental learning, lifelong learning and brain-inspired AI. His work in continual learning has been published in major conferences and journals in related fields, such as Nature Machine Intelligence, IEEE TPAMI, IEEE



**Jingyi Xie** is currently a researcher engineer in Prof. Jun Zhu's group. She received the B.Sc. and M.Sc. degree with the Department of Mathematics and Statistics, Wuhan University, Wuhan, China. Her current research interests include representation learning, continual learning, and deep learning.



**Xingxing Zhang** received the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University in 2020 and B.E. degree in 2015. She was also a visiting student with the Department of Computer Science, University of Rochester, USA, from 2018 to 2019. She was a postdoc in the Department of Computer Science, Tsinghua University, from 2020 to 2022. Her research interests include continual learning and zero/few-shot learning. She has received the excellent Ph.D. thesis award from the Chinese Institute of Electronics in 2020.



Award" in ICME2018.



**Hang Su**, IEEE member, is an associated professor in the department of computer science and technology at Tsinghua University. His research interests lie in the adversarial machine learning and robust computer vision, based on which he has published more than 50 papers including CVPR, ECCV, TMI, etc. He has served as area chair in NeurIPS and the workshop co-chair in AAAI22. he received "Young Investigator Award" from MICCAI2012, the "Best Paper Award" in AVSS2012, and "Platinum Best Paper

**Jun Zhu** received his B.S. and Ph.D. degrees from the Department of Computer Science and Technology in Tsinghua University, where he is currently a Bosch AI professor. He was an adjunct faculty and postdoctoral fellow in the Machine Learning Department, Carnegie Mellon University. His research interest is primarily on developing machine learning methods to understand scientific and engineering data arising from various fields. He regularly serves as senior Area Chairs and Area Chairs at prestigious conferences, including ICML, NeurIPS, ICLR, IJCAI and AAAI. He was selected as "AI's 10 to Watch" by IEEE Intelligent Systems. He is a Fellow of the IEEE and an associate editor-in-chief of IEEE TPAMI.

## APPENDIX A

### THEORETICAL FOUNDATION I

In this section, we present the complete proof of our hierarchical decomposition under different CL scenarios.

#### A.1 Class-Incremental Learning (CIL)

##### Proof of Theorem 1

*Proof.* For class-incremental learning (CIL) with pre-training, assume  $\mathbb{E}_{\mathbf{x}}[H_{WTP}(\mathbf{x})] \leq \delta$ ,  $\mathbb{E}_{\mathbf{x}}[H_{TII}(\mathbf{x})] \leq \epsilon$ , and  $\mathbb{E}_{\mathbf{x}}[H_{TAP}(\mathbf{x})] \leq \eta$ . Let  $y \in \mathcal{Y}_{\bar{i}, \bar{j}}$  be the ground truth of an  $\mathbf{x}$ , where  $\bar{i} \in \{1, \dots, t\}$  and  $\bar{j} \in \{1, \dots, |\mathcal{Y}_{\bar{i}}|\}$  denote the task identity and within-task index, respectively.

As we defined,

$$\begin{aligned} H_{WTP}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, j} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)\}_j) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta), \end{aligned} \quad (19)$$

$$\begin{aligned} H_{TII}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_{\bar{i}}, \{P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)\}_i) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta), \end{aligned} \quad (20)$$

$$\begin{aligned} H_{TAP}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_y, \{P(\mathbf{x} \in \mathcal{X}^c | \mathcal{D}, \theta)\}_c) \\ &= -\log P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta). \end{aligned} \quad (21)$$

Then, we have

$$\begin{aligned} &\mathcal{H}(\mathbf{1}_{\bar{i}, \bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, j} | \mathcal{D}, \theta)\}_{i,j}) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathcal{D}, \theta) \\ &= -\log(P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta) - \log P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta) \\ &= H_{WTP}(\mathbf{x}) + H_{TII}(\mathbf{x}). \end{aligned} \quad (22)$$

Taking expectations on Eq. (21), we have

$$\mathcal{L}_1 = \mathbb{E}_{\mathbf{x}}[H_{TAP}(\mathbf{x})] \leq \eta. \quad (23)$$

Taking expectations on both sides of Eq. (22), we have

$$\begin{aligned} \mathcal{L}_2 &= \mathbb{E}_{\mathbf{x}}[\mathcal{H}(\mathbf{1}_{\bar{i}, \bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, j} | \mathcal{D}, \theta)\}_{i,j})] \\ &= \mathbb{E}_{\mathbf{x}}[H_{WTP}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}}[H_{TII}(\mathbf{x})] \\ &\leq \delta + \epsilon. \end{aligned} \quad (24)$$

Considering the multi-objective optimization problem  $\max[P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathcal{D}, \theta), P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)]$  in Eq. (10), we have the loss error

$$\begin{aligned} \mathcal{L} &= \max\{\mathbb{E}_{\mathbf{x}}[\mathcal{H}(\mathbf{1}_{\bar{i}, \bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, j} | \mathcal{D}, \theta)\}_{i,j})], \mathbb{E}_{\mathbf{x}}[H_{TAP}(\mathbf{x})]\} \\ &= \max\{\mathcal{L}_2, \mathcal{L}_1\} \\ &= \max\{\delta + \epsilon, \eta\}. \end{aligned} \quad (25)$$

This finishes the proof.

##### Proof of Theorem 2

*Proof.* For CIL with pre-training, its loss error  $\mathcal{L} \leq \xi$ . Assume  $\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} \subseteq \mathcal{X}_{\bar{i}}$ . According to the proof of Theorem 1, we have

$$\begin{aligned} H_{WTP}(\mathbf{x}) &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta) \\ &= -\log \frac{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathcal{D}, \theta)}{P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)} \\ &\leq -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathcal{D}, \theta) \\ &= \mathcal{H}(\mathbf{1}_{\bar{i}, \bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, j} | \mathcal{D}, \theta)\}_{i,j}) \\ &= \mathcal{L}_2 \leq \xi. \end{aligned} \quad (26)$$

Likewise, we have

$$\begin{aligned} H_{TII}(\mathbf{x}) &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta) \\ &= -\log \frac{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathcal{D}, \theta)}{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)} \\ &\leq -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathcal{D}, \theta) \\ &= \mathcal{H}(\mathbf{1}_{\bar{i}, \bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{\bar{i}, j} | \mathcal{D}, \theta)\}_{i,j}) \\ &= \mathcal{L}_2 \leq \xi. \end{aligned} \quad (27)$$

Considering the multi-objective optimization problem  $\max[P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathcal{D}, \theta), P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)]$  in Eq. (10), each component must guarantee the loss error less than  $\xi$ , i.e.,

$$\begin{aligned} H_{TAP}(\mathbf{x}) &= -\log P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta) \\ &= \mathcal{L}_1 \leq \xi. \end{aligned} \quad (28)$$

This finishes the proof.

#### A.2 Domain-Incremental Learning (DIL)

For domain-incremental learning (DIL), Let  $\mathcal{X}_i = \bigcup_j \mathcal{X}_{i,j}$  and  $\mathcal{Y}_i = \bigcup_j \mathcal{Y}_{i,j}$ , where  $j \in \{1, \dots, |\mathcal{Y}_i|\}$  denotes the  $j$ -th class in task  $i$  and  $\mathcal{Y}_i = \mathcal{Y}_{i'}$  for  $\forall i \neq i'$  in  $\{1, \dots, t\}$ . Now assume we have a ground event denoted as  $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_t\}$  and a pre-trained model  $f_\theta$ . For any sample  $\mathbf{x} \in \bigcup_{i=1}^t \mathcal{X}_i$ , a general goal of the DIL problem is to learn  $P(\mathbf{x} \in \mathcal{X}_{*,j} | \mathcal{D}, \theta)$ , where  $\mathcal{X}_{*,j}$  represents the  $j$ -th class domain in any task. This can be decomposed into two probabilities, including task-identity inference (TII) and within-task prediction (WTP), denoted as  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)$  and  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)$ , respectively. Based on Bayes' theorem, we have

$$P(\mathbf{x} \in \mathcal{X}_{*,j} | \mathcal{D}, \theta) = \sum_i P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta), \quad (29)$$

where  $\{*, j\}$  represents the  $j$ -th class in each domain.

Then we have the following theorems in terms of the sufficient and necessary conditions for improving DIL with pre-training.

**Theorem 4.** For DIL with pre-training, if  $\mathbb{E}_{\mathbf{x}}[H_{WTP}(\mathbf{x})] \leq \delta$ ,  $\mathbb{E}_{\mathbf{x}}[H_{TII}(\mathbf{x})] \leq \epsilon$ , and  $\mathbb{E}_{\mathbf{x}}[H_{TAP}(\mathbf{x})] \leq \eta$ , we have the loss error  $\mathcal{L} \in [0, \max\{\delta + \epsilon + \log t, \eta\}]$ , regardless whether the WTP predictor, TII predictor and TAP predictor are trained together or separately.

##### Proof of Theorem 4

*Proof.* Let  $\bar{j} \in \{1, \dots, |\mathcal{Y}_t|\}$  and  $y \in \mathcal{Y}_t$  be the ground truth of an  $\mathbf{x}$  w.r.t. the within-task index and class label. Eq. (29) suggests that if we can improve either the WTP performance  $P(\mathbf{x} \in \mathcal{X}_{\bar{i}, \bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)$ , the TII performance  $P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)$ , or both, then the DIL performance  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$  would be improved. However, such an improvement is limited since it is upper-bounded by WTP or TII. To further improve the DIL performance, we propose a hierarchical decomposition of the objective. That is, besides the improvement of  $P(\mathbf{x} \in \mathcal{X}_{*, \bar{j}} | \mathcal{D}, \theta)$ , we also need to directly improve the performance of task-adaptive prediction (TAP), denoted as  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$ , where  $y \in \{1, \dots, |\mathcal{Y}_t|\}$ ,  $\mathcal{X}^y$  represents the domain of class  $y$  in all observed domains, and  $\mathcal{X}^y = \bigcup_i \mathcal{X}_{i, \bar{j}}$ . Then the final goal of DIL is formulated as a multi-objective optimization problem, i.e.,  $\max[P(\mathbf{x} \in \mathcal{X}_{*, \bar{j}} | \mathcal{D}, \theta), P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)]$ . Notice that the TII probability is a categorical distribution over all observed domains  $1, \dots, t$ , while the TAP probability is over all observed classes  $\bigcup_{i=1}^t \mathcal{Y}_i$ .

As similarly defined in CIL, here

$$\begin{aligned} H_{\text{WTP}}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)\}_j) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta), \end{aligned} \quad (30)$$

$$\begin{aligned} H_{\text{TII}}(\mathbf{x}) &= \mathcal{H}(\gamma, \{P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)\}_i) \\ &= -\gamma_i \log P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta), \end{aligned} \quad (31)$$

$$\begin{aligned} H_{\text{TAP}}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_y, \{P(\mathbf{x} \in \mathcal{X}^c | \mathcal{D}, \theta)\}_c) \\ &= -\log P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta), \end{aligned} \quad (32)$$

where  $\gamma = \{\gamma_i\}_{i=1}^t$  represents the possibility of  $\mathbf{x}$  belonging to each observed domain,  $\gamma_i \in [0, 1]$  and  $\sum_i \gamma_i = 1$ .

Then, for any simplex  $\gamma$ , we have

$$\begin{aligned} &\mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{*,j} | \mathcal{D}, \theta)\}_j) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{*,\bar{j}} | \mathcal{D}, \theta) \\ &= -\log \left( \sum_i P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta) P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \right) \\ &\leq -\sum_i \gamma_i \log \left( \frac{P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta) P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\gamma_i} \right) \\ &= -\sum_i \gamma_i \log P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta) \\ &\quad - \sum_i \gamma_i \log P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) + \sum_i \gamma_i \log(\gamma_i) \\ &= \sum_i \gamma_i H_{\text{WTP}} + H_{\text{TII}} + \mathcal{H}(\gamma). \end{aligned} \quad (33)$$

Taking expectations on Eq. (32), we have

$$\mathcal{L}_1 = \mathbb{E}_{\mathbf{x}}[H_{\text{TAP}}(\mathbf{x})] \leq \eta. \quad (34)$$

Taking expectations on both sides of Eq. (33), we have

$$\begin{aligned} \mathcal{L}_2 &= \mathbb{E}_{\mathbf{x}}[\mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{*,j} | \mathcal{D}, \theta)\}_j)] \\ &\leq \sum_i \gamma_i \mathbb{E}_{\mathbf{x}}[H_{\text{WTP}}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}}[H_{\text{TII}}(\mathbf{x})] + \mathcal{H}(\gamma) \\ &\leq \delta + \epsilon + \log t. \end{aligned} \quad (35)$$

Considering the multi-objective optimization problem  $\max[P(\mathbf{x} \in \mathcal{X}_{*,\bar{j}} | \mathcal{D}, \theta), P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)]$ , we have the loss error

$$\begin{aligned} \mathcal{L} &= \max\{\mathbb{E}_{\mathbf{x}}[\mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{*,j} | \mathcal{D}, \theta)\}_j)], \mathbb{E}_{\mathbf{x}}[H_{\text{TAP}}(\mathbf{x})]\} \\ &= \max\{\mathcal{L}_2, \mathcal{L}_1\} \\ &= \max\{\delta + \epsilon + \log t, \eta\}. \end{aligned} \quad (36)$$

This finishes the proof.

**Theorem 5.** For DIL with pre-training, if the loss error  $\mathcal{L} \leq \xi$ , then there always exist (1) a WTP predictor, s.t.  $H_{\text{WTP}} \leq \xi$ ; (2) a TII predictor, s.t.  $H_{\text{TII}} \leq \xi$ ; and (3) a TAP predictor, s.t.  $H_{\text{TAP}} \leq \xi$ .

**Proof of Theorem 5** For DIL with pre-training, its loss error  $\mathcal{L} = \max[\mathcal{L}_1, \mathcal{L}_2] \leq \xi$ . Assume  $\mathbf{x} \in \mathcal{X}_{*,\bar{j}} \subseteq \mathcal{X}^y$ . According to the proof of Theorem 4, if we define  $P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathcal{D}, \theta) = P(\mathbf{x} \in \mathcal{X}_{*,\bar{j}} | \mathcal{D}, \theta)$ , we have

$$\begin{aligned} H_{\text{WTP}}(\mathbf{x}) &= -\log P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta) \\ &= -\log \frac{P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathcal{D}, \theta)}{P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)} \\ &\leq -\log P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathcal{D}, \theta) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{*,\bar{j}} | \mathcal{D}, \theta) \\ &= \mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{*,j} | \mathcal{D}, \theta)\}_j) \\ &= \mathcal{L}_2 \leq \xi. \end{aligned} \quad (37)$$

Likewise, if we define  $\gamma_i = 1$  and  $\gamma_{i'} = 0$  for  $\forall i \neq i'$  in  $\{1, \dots, t\}$ , we have

$$\begin{aligned} H_{\text{TII}}(\mathbf{x}) &= -\sum_i \gamma_i \log P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \\ &= -\log \frac{P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)}{P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)} \\ &\leq -\log(P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)) \\ &= -\log(P(\mathbf{x} \in \mathcal{X}_{*,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)) \\ &= \mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{*,j} | \mathcal{D}, \theta)\}_j) \\ &= \mathcal{L}_2 \leq \xi. \end{aligned} \quad (38)$$

Considering the multi-objective optimization problem  $\max[P(\mathbf{x} \in \mathcal{X}_{*,\bar{j}} | \mathcal{D}, \theta), P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)]$ , each component must guarantee the loss error less than  $\xi$ , i.e.,

$$\begin{aligned} H_{\text{TAP}}(\mathbf{x}) &= -\log P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta) \\ &= \mathcal{L}_1 \leq \xi. \end{aligned} \quad (39)$$

This finishes the proof.

### A.3 Task-Incremental Learning (TIL)

For task-incremental learning (TIL), let  $\mathcal{Y}_i = \bigcup_j \mathcal{Y}_{i,j}$  and  $\mathcal{Y}_i = \bigcup_j \mathcal{Y}_{i,j}$ , where  $j \in \{1, \dots, |\mathcal{Y}_i|\}$  indicates the  $j$ -th class in task  $i$ . Unlike CIL and DIL, TIL has the task identity provided during the testing phase and has  $\mathcal{Y}_i \cap \mathcal{Y}_{i'} = \emptyset$  for  $\forall i \neq i'$  in  $\{1, \dots, t\}$ , respectively. Now assume we have a ground event denoted as  $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_t\}$  and a pre-trained model  $f_\theta$ . For any sample  $\mathbf{x} \in \bigcup_{i=1}^t \mathcal{X}_i$ , a general goal of the TIL problem is to learn  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)$ , where  $i \in \{1, \dots, t\}$  and  $j \in \{1, \dots, |\mathcal{Y}_i|\}$ . In fact, this is equivalent to WTP alone. Then we have the following theorems in terms of the sufficient and necessary conditions for improving TIL with pre-training.

**Theorem 6.** For TIL with pre-training,  $\mathbb{E}_{\mathbf{x}}[H_{\text{TII}}(\mathbf{x})] = 0$ , and TAP is degraded into WTP. If  $\mathbb{E}_{\mathbf{x}}[H_{\text{WTP}}(\mathbf{x})] \leq \delta$ , we have the loss error  $\mathcal{L} \in [0, \delta]$ .

#### Proof of Theorem 6

*Proof.* For TIL with pre-training, assume  $\mathbb{E}_{\mathbf{x}}[H_{\text{WTP}}(\mathbf{x})] \leq \delta$ . Given an  $\mathbf{x}$  with the task identity  $i \in \{1, \dots, t\}$ , let  $j \in \{1, \dots, |\mathcal{Y}_i|\}$  be the ground truth of  $\mathbf{x}$  w.r.t. the within-task index, and  $y \in \mathcal{Y}_{i,j}$  be the ground truth label of  $\mathbf{x}$ .

As similarly defined in CIL, here

$$\begin{aligned} H_{\text{WTP}}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)\}_j) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{i,\bar{j}} | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta), \end{aligned} \quad (40)$$

$$\begin{aligned} H_{\text{TII}}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_{\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)\}_i) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \\ &= -\log 1 \\ &= 0, \end{aligned} \quad (41)$$

$$\begin{aligned} H_{\text{TAP}}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_y, \{P(\mathbf{x} \in \mathcal{X}^c | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta)\}_c) \\ &= -\log P(\mathbf{x} \in \mathcal{X}^y | \mathbf{x} \in \mathcal{X}_i, \mathcal{D}, \theta) \\ &= H_{\text{WTP}}(\mathbf{x}). \end{aligned} \quad (42)$$

Then, we have

$$\begin{aligned} & \mathcal{H}(\mathbf{1}_{\bar{i},\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathcal{D}, \theta)\}_{i,j}) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta) \\ &= -\log(P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta) - \log P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta) \quad (43) \\ &= H_{\text{WTP}}(\mathbf{x}) + H_{\text{TII}}(\mathbf{x}) \\ &= H_{\text{WTP}}(\mathbf{x}). \end{aligned}$$

Taking expectations on both sides of Eq. (43), we have

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{\mathbf{x}}[\mathcal{H}(\mathbf{1}_{\bar{i},\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathcal{D}, \theta)\}_{i,j})] \\ &= \mathbb{E}_{\mathbf{x}}[H_{\text{WTP}}(\mathbf{x})] \\ &\leq \delta. \end{aligned} \quad (44)$$

Considering the TIL objective  $P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)$ , we have the loss error  $\mathcal{L} \leq \delta$ . This finishes the proof.

**Theorem 7.** For TIL with pre-training, if the loss error  $\mathcal{L} \leq \xi$ , then there always exists a WTP predictor, s.t.  $H_{\text{WTP}} \leq \xi$ .

### Proof of Theorem 7

For TIL with pre-training, its loss error  $\mathcal{L} \leq \xi$ . Assume  $\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} \subseteq \mathcal{X}_{\bar{i}}$ . According to the proof of Theorem 6, we have

$$\begin{aligned} H_{\text{WTP}}(\mathbf{x}) &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta) \\ &= -\log \frac{P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta)}{P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)} \\ &\leq -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathcal{D}, \theta) \\ &= \mathcal{H}(\mathbf{1}_{\bar{i},\bar{j}}, \{P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathcal{D}, \theta)\}_{i,j}) \\ &\leq \xi. \end{aligned} \quad (45)$$

This finishes the proof.

## A.4 Impact of Pre-Training on CL

In this work, our theoretical contribution lies in the impact of *pre-training* on CL, where we demonstrate the sufficient and necessary conditions to achieve good CL performance. This is clearly different from the previous work on CL from scratch [41], as analyzed below.

First, the condition is different. We formulate the hierarchical components as  $\theta$ -conditional probabilities, i.e.,  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta)$ ,  $P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)$  and  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$  for WTP, TII and TAP, respectively, where  $\theta$  captures the pre-trained knowledge in initialization. When training from scratch, since the randomly-initialized parameters carry no information and is greatly different from the optimal solution, it needs be substantially changed in CL and should not be considered in objectives. In contrast, the pre-trained parameters carry beneficial knowledge for target tasks. If the pre-training is adequately strong,  $\theta$  is already close to the optimal solution and only requires appropriate fine-tuning. Therefore,  $\theta$  needs to be stabilized (usually frozen) in CL and should be considered in its objective. Then, the due to the incorporation of  $\theta$ , the sufficient and necessary conditions to achieve good CL performance become different, which should further improve TAP on the top of WTP and TII.

Interestingly, without considering the significant impact of pre-training,  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$  becomes equivalent to  $P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta) \cdot P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)$  due to classifying the same input in the same representation space, i.e.,  $\frac{\exp(h_{\psi}(f_{\theta}(\mathbf{x}))[y])}{\sum_{i=1}^t \sum_{y' \in \mathcal{Y}_i} \exp(h_{\psi}(f_{\theta}(\mathbf{x}))[y'])} = \frac{\exp(\hat{h}_{\omega}(f_{\theta}(\mathbf{x})[\bar{i}])}{\sum_{i=1}^t \exp(\hat{h}_{\omega}(f_{\theta}(\mathbf{x})[\bar{i}]))} \cdot$

$\frac{\exp(h_{\psi}(f_{\theta}(\mathbf{x}))[\bar{j}])}{\sum_{j \in \mathcal{Y}_i} \exp(h_{\psi}(f_{\theta}(\mathbf{x}))[\bar{j}])}$ . When considering the significant impact of pre-training, let  $\Delta$  denote the operation of PET, such as prompt parameters  $\{p\}$  for ProT and PreT, the projection matrices  $\{\mathbf{W}_{\text{down}}, \mathbf{W}_{\text{up}}\}$  for Adapter and LoRA, etc.  $P(\mathbf{x} \in \mathcal{X}^y | \mathcal{D}, \theta)$  becomes different from  $P(\mathbf{x} \in \mathcal{X}_{\bar{i},\bar{j}} | \mathbf{x} \in \mathcal{X}_{\bar{i}}, \mathcal{D}, \theta) \cdot P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)$ , because the latter is to calculate  $\frac{\exp(\hat{h}_{\omega}(f_{\theta}(\mathbf{x})[\bar{i}])}{\sum_{i=1}^t \exp(\hat{h}_{\omega}(f_{\theta}(\mathbf{x})[\bar{i}]))} \cdot \frac{\exp(h_{\psi}(f_{\theta,\Delta}(\mathbf{x}))[\bar{j}])}{\sum_{j \in \mathcal{Y}_i} \exp(h_{\psi}(f_{\theta,\Delta}(\mathbf{x})))}$  where TII is performed on unstructured representations. Whereas, TAP is to calculate  $\frac{\exp(h_{\psi}(f_{\theta,\Delta}(\mathbf{x}))[y])}{\sum_{i=1}^t \sum_{y' \in \mathcal{Y}_i} \exp(h_{\psi}(f_{\theta,\Delta}(\mathbf{x}))[y'])}$ , classifying the instructed representations.

## APPENDIX B THEORETICAL FOUNDATION II

In this section, we first present the complete proof of connecting TII to OOD detection, and then derive the sufficient and necessary conditions of improving CL with WTP, OOD detection and TAP.

### B.1 TII to OOD Detection

#### Proof of Theorem 3

For CL in a pre-training context, define the TII probability as  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) = \frac{P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}$ . If  $H_{\text{OOD},i}(\mathbf{x}) \leq \epsilon_i$  for  $i \in \{1, \dots, t\}$ , then we have

$$\begin{aligned} H_{\text{OOD},i}(\mathbf{x}) &= \\ & \begin{cases} \mathcal{H}(1, P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)) = -\log P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \leq \epsilon_i, & \mathbf{x} \in \mathcal{X}_i \\ \mathcal{H}(0, P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)) = -\log P_i(\mathbf{x} \notin \mathcal{X}_i | \mathcal{D}, \theta) \leq \epsilon_i, & \mathbf{x} \notin \mathcal{X}_i \end{cases}. \end{aligned} \quad (46)$$

This means  $P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \geq e^{-\epsilon_i}$  for  $\mathbf{x} \in \mathcal{X}_i$ , and  $P_i(\mathbf{x} \notin \mathcal{X}_i | \mathcal{D}, \theta) \geq e^{-\epsilon_i}$  (i.e.,  $P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \leq 1 - e^{-\epsilon_i}$ ) for  $\mathbf{x} \notin \mathcal{X}_i$ .

Let  $\bar{i} \in \{1, \dots, t\}$  denote the task identity of  $\mathbf{x}$ , then we have

$$\begin{aligned} H_{\text{TII}}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_{\bar{i}}, \{P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)\}_i) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta) \\ &= -\log \frac{P_i(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)} \\ &= \log[1 + \frac{\sum_{j \neq \bar{i}} P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}{P_i(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta)}] \\ &\leq \log[1 + \frac{\sum_{j \neq \bar{i}} 1 - e^{-\epsilon_j}}{e^{-\epsilon_{\bar{i}}}}] \quad (47) \\ &= \log[1 + e^{\epsilon_{\bar{i}}} \sum_{j \neq \bar{i}} 1 - e^{-\epsilon_j}] \\ &\leq e^{\epsilon_{\bar{i}}} \sum_{j \neq \bar{i}} 1 - e^{-\epsilon_j} \\ &= (\sum_i \mathbf{1}_{\mathbf{x} \in \mathcal{X}_i} e^{\epsilon_i})(\sum_i \mathbf{1}_{\mathbf{x} \notin \mathcal{X}_i} (1 - e^{-\epsilon_i})). \end{aligned}$$

The last inequation holds due to  $\log(1 + z) \leq z$  for  $z \geq 0$ .

Now, let us move on the proof of adequacy for TII and OOD detection. If  $H_{\text{TII}}(\mathbf{x}) \leq \epsilon$ , then we have

$$\begin{aligned} H_{\text{TII}}(\mathbf{x}) &= \mathcal{H}(\mathbf{1}_{\bar{i}}, \{P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)\}_i) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_{\bar{i}} | \mathcal{D}, \theta) \\ &\leq \epsilon. \end{aligned} \quad (48)$$

Further, for  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) = \frac{P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}$ , we have

$$\begin{aligned} H_{\text{OOD},i}(\mathbf{x}) &= \mathcal{H}(1, P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)) \\ &= -\log P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \\ &= -\log(P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)) \\ &= -\log P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) - \log \sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta) \quad (49) \\ &= H_{\text{TII}}(\mathbf{x}) - \log \sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta) \\ &\leq \epsilon \end{aligned}$$

The last inequation holds due to  $\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta) \geq 1$ .

Likewise, for  $\mathbf{x} \notin \mathcal{X}_i$ , we have

$$\begin{aligned} H_{\text{OOD},i}(\mathbf{x}) &= \mathcal{H}(0, P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)) \\ &= -\log P_i(\mathbf{x} \notin \mathcal{X}_i | \mathcal{D}, \theta) \\ &= -\log(1 - P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)) \\ &= -\log(1 - P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)) \quad (50) \\ &\leq -\log P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) \\ &\leq \epsilon \end{aligned}$$

This finishes the proof.

## B.2 Sufficient and Necessary Conditions

Now we discuss the upper bound of CIL in relation to WTP, OOD detection and TAP.

**Theorem 8.** For CIL with pre-training (i.e.,  $\theta$ ), define the TII probability as  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) = \frac{P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}$ . If  $H_{\text{WTP}}(\mathbf{x}) \leq \delta$ ,  $H_{\text{TAP}}(\mathbf{x}) \leq \eta$ , and  $H_{\text{OOD},i}(\mathbf{x}) \leq \epsilon_i$  for  $i \in \{1, \dots, t\}$ , then we have the loss error

$$\mathcal{L} \in [0, \max\{\delta + (\sum_i \mathbf{1}_{\mathbf{x} \in \mathcal{X}_i} e^{\epsilon_i})(\sum_i \mathbf{1}_{\mathbf{x} \notin \mathcal{X}_i} (1 - e^{-\epsilon_i})), \eta\}].$$

As shown in Theorem 8, the good performance of WTP, TAP, and OOD detection are sufficient to guarantee a good CIL performance. Now we further study the necessary conditions of a well-performed CIL model.

**Theorem 9.** For CIL with pre-training (i.e.,  $\theta$ ), define the TII probability as  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) = \frac{P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}$ . If the loss error  $\mathcal{L} \leq \xi$ , then there always exist (1) a WTP predictor, s.t.  $H_{\text{WTP}} \leq \xi$ ; (2) a TII predictor, s.t.  $H_{\text{TII}} \leq \xi$ ; (3) a TAP predictor, s.t.  $H_{\text{TAP}} \leq \xi$ ; and (4) an OOD detector for each task, s.t.  $H_{\text{OOD},i} \leq \xi$  for  $i \in \{1, \dots, t\}$ .

This theorem shows that if a good CIL model is trained, then a good WTP, a good TII, a good TAP, and a good OOD detector for each task are always implied.

**Proof of Theorem 8** For CIL with pre-training (i.e.,  $\theta$ ), define TII as  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) = \frac{P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}$ . If  $H_{\text{WTP}}(\mathbf{x}) \leq \delta$ ,  $H_{\text{TAP}}(\mathbf{x}) \leq \eta$ , and  $H_{\text{OOD},i}(\mathbf{x}) \leq \epsilon_i$  for  $i \in \{1, \dots, t\}$ , then using Theorem 8 we have  $H_{\text{TII}}(\mathbf{x}) \leq (\sum_i \mathbf{1}_{\mathbf{x} \in \mathcal{X}_i} e^{\epsilon_i})(\sum_i \mathbf{1}_{\mathbf{x} \notin \mathcal{X}_i} (1 - e^{-\epsilon_i}))$ . Further, using Theorem 1 we have the loss error

$$\begin{aligned} \mathcal{L} &= \max\{\mathbb{E}_{\mathbf{x}}[\mathcal{H}(\mathbf{1}_{\bar{i}, j}, \{P(\mathbf{x} \in \mathcal{X}_{i,j} | \mathcal{D}, \theta)\}_{i,j})], \mathbb{E}_{\mathbf{x}}[H_{\text{TAP}}(\mathbf{x})]\} \\ &= \max\{\mathbb{E}_{\mathbf{x}}[H_{\text{WTP}}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}}[H_{\text{TII}}(\mathbf{x})], \mathcal{L}_1\} \\ &= \max\{\delta + (\sum_i \mathbf{1}_{\mathbf{x} \in \mathcal{X}_i} e^{\epsilon_i})(\sum_i \mathbf{1}_{\mathbf{x} \notin \mathcal{X}_i} (1 - e^{-\epsilon_i})), \eta\}. \quad (51) \end{aligned}$$

This finishes the proof.

**Proof of Theorem 9** For CIL with pre-training (i.e.,  $\theta$ ), define the TII probability as  $P(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta) = \frac{P_i(\mathbf{x} \in \mathcal{X}_i | \mathcal{D}, \theta)}{\sum_j P_j(\mathbf{x} \in \mathcal{X}_j | \mathcal{D}, \theta)}$ . If the loss error  $\mathcal{L} \leq \xi$ , then using Theorem 2 there always exist (1) a WTP predictor, s.t.  $H_{\text{WTP}} \leq \xi$ ; (2) a TII predictor, s.t.  $H_{\text{TII}} \leq \xi$ ; and (3) a TAP predictor, s.t.  $H_{\text{TAP}} \leq \xi$ . Furthermore, if  $H_{\text{TII}} \leq \xi$ , using Theorem 3, we always have an OOD detector for each task, s.t.  $H_{\text{OOD},i} \leq \xi$  for  $i \in \{1, \dots, t\}$ . This finishes the proof.

## APPENDIX C IMPLEMENTATION DETAILS

Here we describe the supplementary implementation details of the empirical investigation.

**Comparison with Preliminary Version:** The major technical difference between our HiDe-PET and our preliminary version [13] lies in the use of task-shared parameters  $\mathbf{g}$  to improve TII, which is critical for LoRA/Adapter-based PET that is sensitive to the TII errors. To mitigate catastrophic forgetting in  $\mathbf{g}$ , we set a cosine-decaying learning rate of 0.01 for FSA, a cosine-decaying learning rate of 0.001 for SL, and a momentum of 0.1 for EMA. The PET ensemble strategy sets  $\alpha = 0.1$  in all cases. To ensure generality and resource efficiency, the specific implementations are slightly modified in three aspects. First, our preliminary version [13] followed the implementation of L2P [5] and DualPrompt [6], which employed a constant learning rate of 0.005 and a supervised checkpoint of ImageNet-21K (i.e., Sup-21K). We notice that many recent methods followed the implementation of CODA-Prompt [8], which employed a cosine-decaying learning rate of 0.001, a self-supervised/supervised checkpoint on ImageNet-21/1K (i.e., Sup-21/1K) and a different split of ImageNet-R. Considering that a smaller learning rate with cosine decay has been more commonly used for fine-tuning large-scale PTMs, we reproduce all baselines with the implementation of CODA-Prompt [8] in the current manuscript. This consideration further ensures the generality of our HiDe-PET in adapting to different experimental settings. Second, our preliminary version [13] devised a contrastive regularization (CR) term to balance the instructed representations for WTP and TAP, which brings some benefits to the performance of Prompt-based PET. In subsequent explorations, we observe that the CR term cannot improve the performance of LoRA/Adapter-based PET, and therefore remove it in the current manuscript to ensure generality in adapting to different PET techniques. Third, our preliminary version [13] employed dedicated covariance matrices (additional  $d^2$  parameters for each class) in representation recovery and a two-layer MLP (additional  $d^2$  parameters for the first layer) in  $\hat{h}_{\omega}$ , in order to acquire better performance. In contrast, the current manuscript employs multi-centroid (additional  $< 10d$  parameters for each class) in representation recovery and a one-layer MLP in  $\hat{h}_{\omega}$ , which slightly compromise the performance but largely improve resource efficiency.

**Adaptive Knowledge Accumulation:** In Sec. 5.2, we devise a PET hierarchy  $\mathbf{g}_1, \dots, \mathbf{g}_k$  inspired by OOD detection to demonstrate the connections between task-specific and task-shared PET architecture. We further consider a specialized

TABLE 6

Comparison of recent CL methods relevant to PTMs and PET.  $t$  is the total number of tasks.  $d$  is the embedding dimension.  $s$  is the expansion rate of embedding dimension. Full: fine-tuning of full backbone parameters. General: applicable to mainstream PET techniques, such as ProT, PreT, LoRA, Adapter, etc. Note that  $t \ll d$  in general, e.g.,  $t = 10$  and  $d = 768$  for all cases in this work.  $s = 100$  in [34].

Method	Year	Avenue	PET Technique	Task-Specific Parameters	Task-Shared Parameters	Representation Recovery
L2P [5]	2022	CVPR	ProT	✓	✓	N/A
DualPrompt [6]	2022	ECCV	PreT	✓	✓	N/A
S-Prompt [7]	2022	NeurIPS	ProT	✓	N/A	N/A
CODA-Prompt [8]	2023	CVPR	PreT	✓	N/A	N/A
SLCA [4]	2023	ICCV	Full	N/A	✓	$O(td^2)$
FSA [43]	2023	ICCV	Full	N/A	✓	$O(d^2)$
LAE [14]	2023	ICCV	General	N/A	✓	N/A
RanPAC [34]	2023	NeurIPS	PreT	N/A	✓	$O(s^2d^2)$
KOPPA [44]	2023	arXiv	PreT	✓	N/A	$O(td)$
H-Prompt [65]	2024	arXiv	PreT	✓	✓	$O(td^2)$
HiDe-Prompt [13]	2023	NeurIPS	PreT	✓	N/A	$O(td^2)$
HiDe-PET	2024	Current	General	✓	✓	$O(td)$

implementation of  $\mathbf{g}_1, \dots, \mathbf{g}_k$  for HiDe-PET in Algorithm 1, serving as a plug-in module to achieve adaptive knowledge accumulation from pronounced distribution changes. As described in Sec. 5.2,  $\mathbf{g}_1, \dots, \mathbf{g}_k$  are adaptively expanded or retrieved upon the distance  $\text{Dis}(\mathbf{x}, \hat{\mathcal{G}}_i)$  to each previous task  $i \in [t]$ . The learning of each  $\mathbf{g}_j$  for  $j \in [k]$  is identical to the learning of  $\mathbf{g}$ , i.e., a combination of FSA and SL. As for the exploitation of  $\mathbf{g}_1, \dots, \mathbf{g}_k$ , before learning each task  $i$ , the most relevant  $\mathbf{g}_j$  is first retrieved upon the current training samples  $\mathcal{D}_i$  and then temporarily added to the backbone parameters  $\theta$  to better incorporate task-specific knowledge (the improved  $\theta$  is denoted as  $\theta'$ ). The improved backbone  $f_{\theta'}$  is used to obtain  $\mathcal{G}_{i,c}$ , i.e., Step 10 in Algorithm 1. Whereas, the original backbone  $f_{\theta}$  is still used to obtain  $\hat{\mathcal{G}}_{i,c}$ , i.e., Step 9 in Algorithm 1. At the testing phase, the most relevant  $\mathbf{g}_j$  is retrieved upon the current testing samples and temporarily added to  $\theta$ .