LCA: LOCAL CLASSIFIER ALIGNMENT FOR CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental requirement for intelligent systems is the ability to learn continuously under changing environments. However, models trained in this regime often suffer from catastrophic forgetting. Leveraging pre-trained models has recently emerged as a promising solution, since their generalized feature extractors enable faster and more robust adaptation. While some earlier works mitigate forgetting by fine-tuning only on the first task, this approach quickly deteriorates as the number of tasks grows and the data distributions diverge. More recent research instead seeks to consolidate task knowledge into a unified backbone, or adapting the backbone as new tasks arrive. However, such approaches may create a (potential) *mismatch* between task-specific classifiers and the adapted backbone. To address this issue, we propose a novel *Local Classifier Alignment* (LCA) loss to better align the classifier with backbone. Theoretically, we show that this LCA loss can enable the classifier to not only generalize well for all observed tasks, but also improve robustness. Furthermore, we develop a complete solution for continual learning, following the model merging approach and using LCA. Extensive experiments on several standard benchmarks demonstrate that our method often achieves leading performance, sometimes surpasses the state-of-the-art methods with a large margin.

1 INTRODUCTION

In real-world scenarios, data arrives continuously, requiring deployed models to adapt to evolving distributions while preserving previously learned knowledge. This challenge, known as the stability–plasticity dilemma in continual learning, captures the difficulty of balancing adaptation with retention. A widely adopted benchmark approximates this setting by partitioning a dataset into disjoint tasks and training a model sequentially without access to earlier data (Wang et al., 2024). At test time, the model must handle all tasks without being given the task identity. This setup is referred to as *Class-Incremental Learning* (CIL) (Van de Ven et al., 2022).

Pre-trained models (PTMs) have recently emerged as a strong foundation for this setting, in contrast to earlier methods that trained networks from scratch (Li & Hoiem, 2017; Kirkpatrick et al., 2017). PTMs can be obtained through supervised or self-supervised training, and include large multimodal models such as CLIP Radford et al. (2021) as well as vision backbones like the Vision Transformer (Dosovitskiy et al., 2020). Their broad generalization ability makes them effective feature extractors, requiring only lightweight adaptation and thereby reducing forgetting compared to traditional training.

Nevertheless, naive sequential adaptation still leads to degradation on past tasks. Restricting updates to the first task avoids forgetting but blocks useful transfer to later ones. Since each task introduces unique information, the final model must capture both shared and task-specific components. A natural approach is to incrementally adapt on each task and then merge the resulting models into a unified backbone. Insights from Linear Mode Connectivity and the Lottery Ticket Hypothesis (Frankle et al., 2020) suggest that task-specific solutions can be connected through low-loss paths and rely on sparse, critical parameters, making them amenable to combination. Advances in model merging (Ilharco et al., 2022; Yadav et al., 2023) further support this strategy, with both theoretical guarantees (Li et al., 2025) and empirical evidence on adaptation tasks (Akiba et al., 2025) showing that merged models can preserve and even enhance knowledge. Thus, it is reasonable to treat each

task-specific model as a standalone expert containing complementary knowledge, whose consolidation yields a stronger overall backbone. However, merging backbones across tasks introduces a new challenge: classifiers trained independently may no longer align with the integrated backbone. Because these classifiers cannot be retrained without access to past data, even small parameter shifts can lead to severe drops in performance on earlier tasks. Addressing this misalignment is the central focus of our work.

Our contributions in this work are as follows:

- We introduce a novel loss, *Local Classifier Alignment* (LCA), for training CIL classifiers. We provide theoretical analysis showing that LCA ensures both high generalization and robustness. Such a theory is crucial to support reliability of LCA and trustworthy CIL.
- We propose a new CIL framework in which LCA serves as the main alignment loss to reduce mismatches between classifiers and the backbone after learning new tasks. In our framework, each class is represented as a Gaussian in the latent space, and LCA jointly optimizes all classifiers.
- We conduct extensive experiments on seven benchmark datasets. Our results show that LCA consistently improves performance over baselines, enhances robustness under diverse scenarios, and can be integrated with other CIL methods to further boost their effectiveness. Figure 1 shows superiority of LCA on seven benchmark datasets.

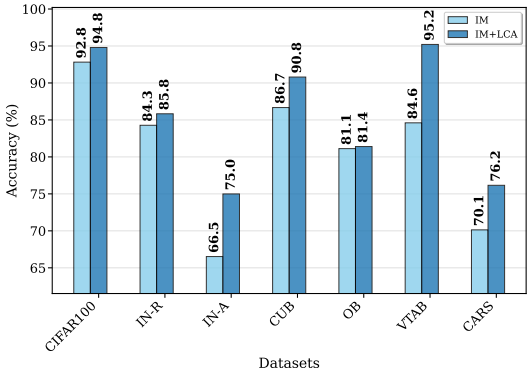


Figure 1: A comparison between IM and IM with LCA. IM is the result after only done the Incremental Merging step, while IM+LCA has Local Classifier Alignment as the last step.

2 Related Works

Representation-Based Methods. This line of research leverages the representational power of large pre-trained models for continual learning (Wang et al., 2024). Trained on massive and diverse datasets, these models provide strong transferability and inherent robustness against forgetting. One prominent direction adapts prompting techniques from natural language processing, where prompts are modeled as learnable parameters attached to the inputs (Wang et al., 2022b;a; Tran et al., 2025). These prompts act as additional instructions that guide model predictions during continual learning. Another direction adapts the pre-trained backbone only once during the first task by using parameter-efficient fine-tuning (PEFT) modules and then relies on class prototypes for inference (Panos et al., 2023; Zhou et al., 2025; McDonnell et al., 2023; Perez et al., 2018). Such prototype-based approaches classify via cosine similarity can avoid forgetting because accumulated prototypes across tasks are tasks order-invariant. However, empirical evidence shows that adaptation only in the first task is insufficient, since data distributions in real-world scenarios vary substantially across tasks, which limits this method's applicability in long training horizons.

Incremental Backbone Evolution. A complementary line of research updates the backbone throughout the task sequence. For example, SLCA (Zhang et al., 2023) reduces forgetting by applying a smaller learning rate to the backbone than to the classifier, while methods such as MagMax (Marczak et al., 2024), EASE (Zhou et al., 2024), and MOS (Sun et al., 2025b) focus on integrating task-specific components into a unified backbone. In all cases, past classifiers are typically frozen to avoid bias toward the current task's data, which inevitably creates a mismatch between the evolving backbone and the fixed classifiers. To mitigate this issue, EASE reweights old classifiers using semantic similarity between new and old prototypes, while MOS dynamically selects suitable backbone adapters at inference time. Our approach differs in that we retrain all classifiers after the unification step using a novel loss that emphasizes local robustness, thereby directly reducing the mismatch between backbone and classifiers and improving the stability of the overall model.

Model Merging. Another related direction is model merging, which has recently gained attention for constructing a unified model from independently trained task-specific ones. Early methods such as FisherMerging (Matena & Raffel, 2022) use the Fisher Information Matrix to weight parameter importance, while Task Arithmetic (Ilharco et al., 2022) represents task updates as vectors, enabling explicit addition or subtraction of knowledge. More advanced approaches like TIES-Merge (Yadav et al., 2023) address task interference by resolving sign conflicts across task vectors. Although these methods show strong performance in out-of-domain generalization (Rame et al., 2022), applying them directly to continual learning often introduces a mismatch between the merged backbone and fixed classifiers, and incurs storage overhead since parameters from all past tasks must be retained. In contrast, our approach builds on the spirit of model merging by incrementally consolidating PEFT modules, which reduces memory requirements, and coupling this with a continuous training scheme that solidifies accumulated knowledge across tasks.

3 METHODOLOGY

This section presents our methodology for continual learning, which consists of two complementary components. The first focuses on incrementally merging task-specific backbones into a unified model, while maintaining proximity across tasks by initializing new training from the latest merged backbone. The second addresses the mismatch that arises when frozen task-specific classifiers interact with the consolidated backbone, introducing an alignment mechanism based on class-wise local regions. Together, these components enable the model to effectively retain past knowledge while adapting to new tasks.

3.1 PROBLEM FORMULATION

In class-incremental learning (CIL), the objective is to train a single model sequentially on a series of (potentially infinite number of) tasks. Each task i is associated with a dataset

$$\mathcal{D}_i = \{(x, y) \mid x \in \mathcal{X}_i,\ y \in \mathcal{Y}_i\},$$

where x denotes an input sample and y its corresponding label. Here, $\mathcal{X}_i$ represents the input space and $\mathcal{Y}_i$ the label space for task i. A defining characteristic of the CIL setting is that the label spaces of different tasks are strictly disjoint:

$$\forall i \neq j, \quad \mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset.$$

This assumption implies that each new task introduces a set of novel classes that the model has not encountered before. Consequently, the model must continuously expand its knowledge while preserving performance on previously learned classes, making the prevention of catastrophic forgetting a central challenge in CIL.

There are different strategies for constructing classifiers in continual learning, such as Nearest Class Mean (NCM) classifiers, which assign labels by comparing test features to stored class prototypes. In this work, however, we focus on the more general and widely used setup of progressively expanding Multi-Layer Perceptrons (MLPs) on top of the feature extractor. Instead of training a single unified classifier over all classes, a new MLP head is added for each task, with its output dimension matching the number of classes introduced by that task. This approach not only reduces storage but also mitigates forgetting, since earlier classifiers remain frozen and are not modified during subsequent training. After t tasks, the classifier consists of a set of task-specific heads $\{\theta_1^{\text{cls}}, \theta_2^{\text{cls}}, \ldots, \theta_t^{\text{cls}}\}$, and inference is performed by evaluating each head separately and concatenating their outputs:

$$h(x) = \text{concat}\big(h(x; \theta_1^{\text{cls}}),\ h(x; \theta_2^{\text{cls}}),\ \ldots,\ h(x; \theta_t^{\text{cls}})\big).$$

3.2 INCREMENTAL KNOWLEDGE CONSOLIDATION

Although pre-trained models possess strong generalization, they still lack the domain-specific knowledge needed to serve as effective feature extractors. As a result, a fine-tuning stage is required. To prevent forgetting during continuous fine-tuning, early methods often assume that task distributions remain close and restrict adaptation to the first task. However, as the number of tasks grows and their distributions diverge, performance on earlier tasks inevitably deteriorates. Inspired by research on model merging, we propose an incremental integration scheme that unifies task-specific backbones into a single consolidated backbone.

Algorithm 1 Incremental Merging (IM)

Input: Datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$, pretrained model $\theta_{\text{pretrained}}$, base PEFT θ_{peft_0}
Output: Merged PEFT module θ_{merged}

$\tau \leftarrow \mathbf{0}$ ▷ Task vector accumulator
for $i = 1$ to T **do** ▷ Sequentially process tasks
  $\theta_{\text{peft}_i} \leftarrow \texttt{finetune}(\theta_{\text{peft}_{i-1}}, \mathcal{D}_i)$ ▷ Task-i adaptation
  $\tau_{\text{curr}} \leftarrow \theta_{\text{peft}_i} - \theta_{\text{peft}_0}$ ▷ Task vector
  for $k = 1$ to d **do** ▷ d = number of parameters
    if $|\tau_{\text{curr}}^{(k)}| \geq |\tau^{(k)}|$ **then**
      $\tau^{(k)} \leftarrow \tau_{\text{curr}}^{(k)}$ ▷ Keep larger magnitude, preserve sign
    end if
  end for
  $\theta_{\text{merged}} \leftarrow \theta_{\text{peft}_0} + \alpha \cdot \tau$ ▷ Update merged PEFT
end for

To avoid excessive growth in stored parameters across tasks, we only keep the previously merged parameters and the current task's parameters at each step. We construct task vectors by subtracting the base PEFT parameters, then perform element-wise selection based on absolute magnitude. Specifically, we flatten both vectors and, for each coordinate, retain the value with the larger absolute magnitude while preserving its sign. If all task parameters were merged simultaneously, the procedure would resemble sign-based selection using the summed contributions, as in TIES-Merging (Yadav et al., 2023). The incremental merging scheme is summarized in Algorithm 1.

To further limit drift between tasks, training proceeds sequentially by using the latest merged backbone as the initialization for each new task, instead of reinitializing from the base parameters. This practice keeps successive solutions close in parameter space, a property emphasized in model-merging theory (Li et al., 2025) as important for stable and effective merging.

3.3 Local Classifier Alignment

In contrast with a generalized feature extractor, task-specific classifiers should remain well separated to achieve strong classification performance. Because datasets from previous tasks are unavailable, updating earlier classifiers during new training would move them away from their previously optimized values. The standard pipeline therefore trains and merges the backbone while freezing the old classifiers, which introduces a mismatch between the unified backbone and those fixed heads. We address this discrepancy with an alignment step that bridges the two components.

Consider the CIL approach where we use a pretrained model such as VIT to produce high-quality embeddings of the input samples, a Gaussian distribution (or prototype) to represent a class, and a classifier. This approach has been investigated heavily and often high-performing CIL methods. At each time step t, one can use a learning method to train a classifier h_t based on the classifier h_{t-1} which already fitted for prior tasks and a dataset $\boldsymbol{D}_t$ for the current task. The overall performance of h_t depends heavily on the employed learning algorithm. A good algorithm can well align the prototypes and classifier. However, this might not be always the case, and hence can degrade the overall CIL performance.

We propose a simple finetuning step, called *Local Classifier Alignment (LCA)*, to better align the classifier and prototypes. Specifically, LCA minimizes the following loss

$$L(\boldsymbol{D}, h_t) = \frac{1}{C_t} \sum_{i=1}^{C_t} L_i \tag{1}$$

$$L_i = \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{D}_i} \left[\ell(h_t, \boldsymbol{x})\right] + \lambda \mathbb{E}_{\boldsymbol{x}, \boldsymbol{x}' \sim \boldsymbol{D}_i} \left[\left|\ell(h_t, \boldsymbol{x}) - \ell(h_t, \boldsymbol{x}')\right|\right] \tag{2}$$

where $\boldsymbol{D}_i$ contains i.i.d. samples from distribution $\mathcal{N}_i$, and $\boldsymbol{D} = \{\boldsymbol{D}_1, ..., \boldsymbol{D}_t\}$.

Basically, LCA tries to simultaneously minimize the loss for each class and keep the loss less sensitive to a small change in the input samples around the class prototypes. The class loss can be seen from the first term $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{D}_i}[\ell(h_t, \boldsymbol{x})]$, while the sensitivity comes from the second term in each L_i.

Such a term can be seen as a regularizer to penalize the classifier for unstable predictions. A larger value for λ suggests a stronger penalty for sensitivity of the loss.

It is worth noting that the novelty of LCA comes from the regularization term. It not only helps improve robustness of the classifier, but also can reduce overlapping between classes. Indeed, when using the first term as the main objective for training the classifier, some samples randomly generated by $\mathcal{N}_i$ of one class can lie far from the i-th class prototype and hence can be closer to some other class prototypes. This can harm the training for the classifier. The second term can help us reduce the negative effect from those potentially harmful samples, under a suitable choice of the regularization constant.

3.4 Theoretical analysis

We next analyze the generalization ability of the classifier h_t. Until time step t, the classifier h_t already has learned from C_t classes. We want to estimate its test error for those learned tasks. To this end, the classical tradition is to estimate $L(P, \boldsymbol{h}) = \frac{1}{C_t}\sum_{i=1}^{C_t} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_i}\left[\ell(h_t, \boldsymbol{x})\right]$, which represents the overall expected error for the learned tasks.

Let $\bigcup_{i=1}^{t} \mathcal{Z}_i$ be the decomposition of the data space into non-overlapping local areas, with $\mathcal{Z}_i$ as the local area with centroid μ_i, where μ_i is the mean of the Gaussian distribution $\mathcal{N}_i$, for each index i. We have the following bound for the expected error of the overall classifier up to time step t, whose proof appears in Appendix A.

Theorem 3.1. *Consider a model h_t learned from a dataset $\boldsymbol{D} = \{\boldsymbol{D}_1, ..., \boldsymbol{D}_t\}$, where $\boldsymbol{D}_i$ contains n_i i.i.d. samples from distribution $\mathcal{N}_i$ for each $i \le C_t$, and a bounded loss ℓ. Denote $P = \frac{1}{C_t}\sum_{i=1}^{C_t}\mathcal{N}_i$ as the overall distribution, $n = \sum_{i=1}^{C_t} n_i$, $\ell_{\max} = \sup_{\boldsymbol{z}} \ell(\boldsymbol{h}, \boldsymbol{z})$, and $\bar{\epsilon}_i(h_t) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}_i, \boldsymbol{s}\sim\boldsymbol{D}_i}[|\ell(h_t, \boldsymbol{z}) - \ell(h_t, \boldsymbol{s})| : \boldsymbol{z}, \boldsymbol{s} \in \mathcal{Z}_i]$ for each index i. For any $\delta > 0$, the following holds with probability at least $1 - \delta$:*

$$L(P, h_t) \le L(\boldsymbol{D}, h_t) + \sum_{i=1}^{C_t} \frac{n_i}{n}\bar{\epsilon}_i(h_t) + \ell_{\max}\sqrt{\frac{C_t \ln 4 + 2\ln(1/\delta)}{n}} \tag{3}$$

This result shows that the test error of the classifier h_t can be controlled by both the training error and the robustness term $\bar{\epsilon} = \sum_{i=1}^{C_t} \frac{n_i}{n}\bar{\epsilon}_i(h_t)$, which tells how robust is the loss w.r.t. to a small change in the input of the samples around the class prototypes. A stronger robustness (i.e., smaller $\bar{\epsilon}$) can lead to a tighter bound on the test error, suggeting a better model. When both $L(\boldsymbol{D}, h_t)$ and $\bar{\epsilon}$ are small, the model must have small test error and hence generalize well on unseen data. On the other hand, a bad model will exhibit a large training error $L(\boldsymbol{D}, h_t)$ or large $\bar{\epsilon}$. Therefore, Theorem 3.1 plays as the theoretical foundation for our LCA loss (1). Training by this loss would arguably improve both performance and robustness of the classifier, which is crucial for real-world CIL tasks.

Corollary 1. *Given the notations and assumptions as in Theorem 3.1, if $n_i = m$ for all i, then the following holds with probability at least $1 - \delta$:*

$$L(P, h_t) \le L(\boldsymbol{D}, h_t) + \frac{1}{C_t}\sum_{i=1}^{C_t} \bar{\epsilon}_i(h_t) + \ell_{\max}\sqrt{\frac{\ln 4}{m} + \frac{2\ln(1/\delta)}{mC_t}} \tag{4}$$

This is a direct consequence of Theorem 3.1. It suggests that, to assure high generalization, the number m of samples for each class should not be too small when doing alignment. If one use a small m, the uncertainty part in (4) will be large, meaning some randomness (by noise) can harm the alignment step.

Remark 1. Although the LCA loss (1) is introduced to the CIL context, the loss is general enough to be employed in many other contexts. Indeed, one can use LCA loss to train a CIL classifier. Also, one can easily plug LCA to do a finetuning step for the existing CIL methods. In those cases, the theoretical benefits of LCA shown in Theorem 3.1 may remain valid.

It is worth noting that the result in Theorem 3.1 holds when the class distributions (or prototypes) are fixed. This means Theorem 3.1 applies to the cases that learning a new task does not change the latent features of the examples from old classes. For the cases of prototype changes, our theoretical results do not directly apply.

Table 1: Final average accuracy comparison on seven datasets with ViT-B/16-IN1K as the pretrained backbone. The best result is highlighted in bold, while the second best is highlighted in italic.

Method	CIFAR100	IN-R	IN-A	CUB	OB	VTAB	CARS	Overall
APER+Adapter	90.8 ± 0.5	78.8 ± 0.6	58.9 ± 1.3	89.7 ± 1.3	80.3 ± 0.4	90.7 ± 0.6	50.6 ± 1.1	77.1
APER+Finetune	81.7 ± 0.9	72.1 ± 0.8	58.7 ± 3.7	89.5 ± 1.4	77.8 ± 1.2	91.8 ± 1.4	53.2 ± 1.4	75.0
APER+SSF	89.5 ± 1.0	78.1 ± 1.1	61.6 ± 0.5	89.6 ± 1.1	80.3 ± 0.6	91.8 ± 1.5	51.3 ± 1.1	77.5
APER+VPT-Deep	89.0 ± 0.9	78.8 ± 0.7	57.0 ± 0.4	89.0 ± 1.2	79.8 ± 0.4	91.9 ± 1.4	50.6 ± 3.0	76.6
APER+VPT-Shallow	88.1 ± 0.9	67.3 ± 3.5	56.9 ± 1.4	89.5 ± 1.4	79.7 ± 0.9	91.5 ± 0.8	50.9 ± 0.8	74.8
CODA-Prompt	91.0 ± 0.2	78.2 ± 0.4	48.1 ± 0.9	75.6 ± 1.2	71.0 ± 0.1	65.6 ± 2.6	26.3 ± 0.6	65.1
DualPrompt	86.7 ± 0.6	74.6 ± 0.5	55.3 ± 1.5	78.9 ± 1.0	74.4 ± 1.2	84.0 ± 5.9	49.4 ± 2.1	71.9
EASE	91.7 ± 0.3	82.4 ± 0.5	*67.8 ± 1.8*	89.5 ± 1.2	80.8 ± 0.2	*93.3 ± 0.1*	48.1 ± 1.2	79.1
L2P	87.7 ± 1.4	77.3 ± 0.6	52.6 ± 1.7	75.8 ± 1.8	73.8 ± 1.2	82.4 ± 2.8	53.4 ± 1.2	71.9
MOS	*94.3 ± 0.3*	83.3 ± 0.6	67.6 ± 2.0	**92.3 ± 0.6**	**86.1 ± 0.7**	92.4 ± 0.5	71.4 ± 19.6	*83.9*
SLCA	93.7 ± 0.3	*85.1 ± 0.3*	45.1 ± 19.8	90.2 ± 0.9	*82.7 ± 0.6*	91.1 ± 3.4	*74.6 ± 2.2*	80.4
IM	92.8 ± 0.1	84.3 ± 1.0	66.5 ± 1.1	86.7 ± 0.8	81.1 ± 0.8	84.6 ± 4.9	70.1 ± 1.5	80.9
IM+LCA	**94.8 ± 0.3**	**85.8 ± 0.2**	**75.0 ± 0.5**	*90.8 ± 0.3*	81.4 ± 0.5	**95.2 ± 1.1**	**76.2 ± 1.4**	**85.6**

4 EXPERIMENTS

This section presents the details of our experiments on seven benchmark datasets, along with an ablation study to examine some design aspects of our method.

4.1 EXPERIMENT SETUP

Dataset. We follow the setup in (McDonnell et al., 2023) to select datasets and define the number of classes in each incremental task. Specifically, we evaluate on CIFAR100 (Krizhevsky & Hinton, 2009), ImageNet-R (IN-R) (Hendrycks et al., 2021a), ImageNet-A (IN-A) (Hendrycks et al., 2021b), CUB-200 (CUB) (Welinder et al., 2010), OmniBenchmark (OB) (Zhang et al., 2022), VTAB (Zhai et al., 2019), and StanfordCars (CARS) (Krause et al., 2013). All datasets are split into 10 tasks, except VTAB which is divided into 5.

Baselines. the following methods are used for comparison:

- **Ours:** We consider two variants: **IM** (Incremental Merging) and **IM+LCA**, where the latter incorporates LCA in the alignment phase.
- Pre-trained based CIL methods: *CODA-Prompt* (Smith et al., 2023), *DualPrompt* (Wang et al., 2022a), *L2P* (Wang et al., 2022b), *EASE* (Zhou et al., 2024), *MOS* (Sun et al., 2025b), *SLCA* (Zhang et al., 2023), *APER* (Zhou et al., 2025).

For fair comparison and reproducibility, we adopt the implementation from (Sun et al., 2025a) for all baseline methods and use the same pre-trained backbone, ViT-B/16 trained on ImageNet-1K (ViT-B/16-IN1K) (Dosovitskiy et al., 2020).

Training. We apply LoRA (Hu et al., 2022) and ensure a consistent computational budget by fixing the low-rank dimension to 64. All experiments are carried out on a single NVIDIA RTX 4090 GPU running Ubuntu 22.04. Detail on hyperparameters is mentioned in Appendix B.

Evaluation Metrics. Following standard practice, we report the final average accuracy, defined as $\text{AA} = \frac{1}{T}\sum_{i=1}^{T} A_{T,i}$, where $A_{T,i}$ denotes the accuracy on task i after training on the final task T.

4.2 EXPERIMENT RESULTS

4.2.1 OVERALL BENCHMARK

Table 1 reports the mean and standard deviation of accuracy across three random seeds (1993, 1994, 1995). With LCA, the final model achieves the highest performance on five out of seven benchmark datasets, yielding an overall improvement of nearly 2%. Unlike methods such as EASE (Zhou et al., 2024) and MOS (Sun et al., 2025b), which either expand the backbone to integrate new tasks or rely on complex inference procedures, our approach requires no additional mechanism. The backbone is incrementally merged without storing past parameters or samples, and the only extra storage is the

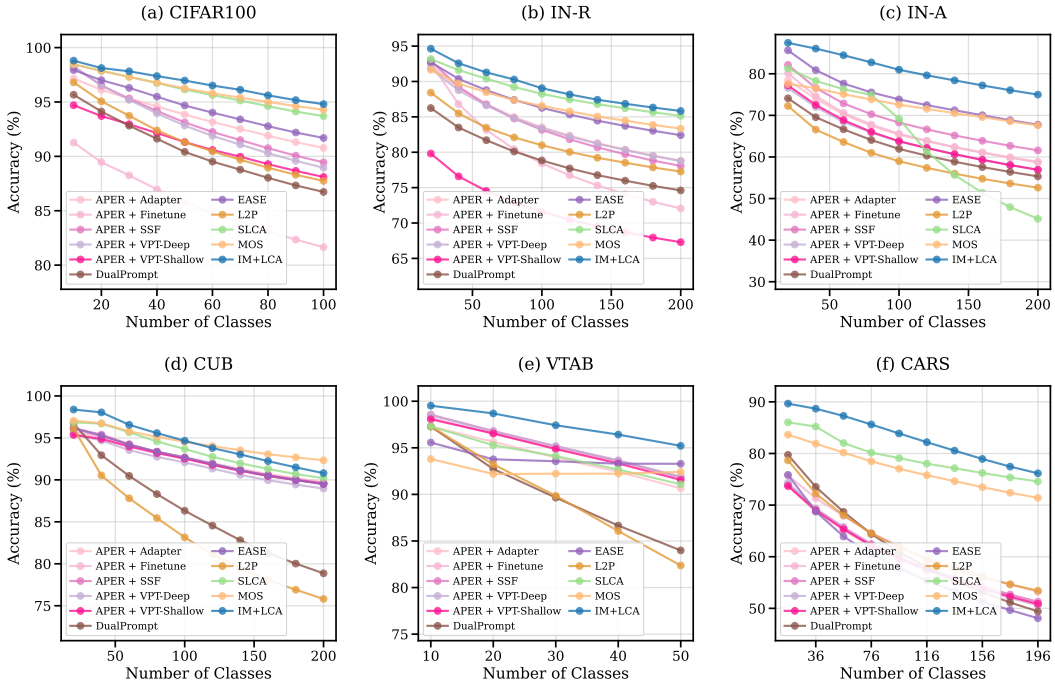


Figure 2: Performance curves of different methods across all tasks and datasets. All methods use ViT-B/16-IN1K as the pre-trained backbone without any additional exemplars.

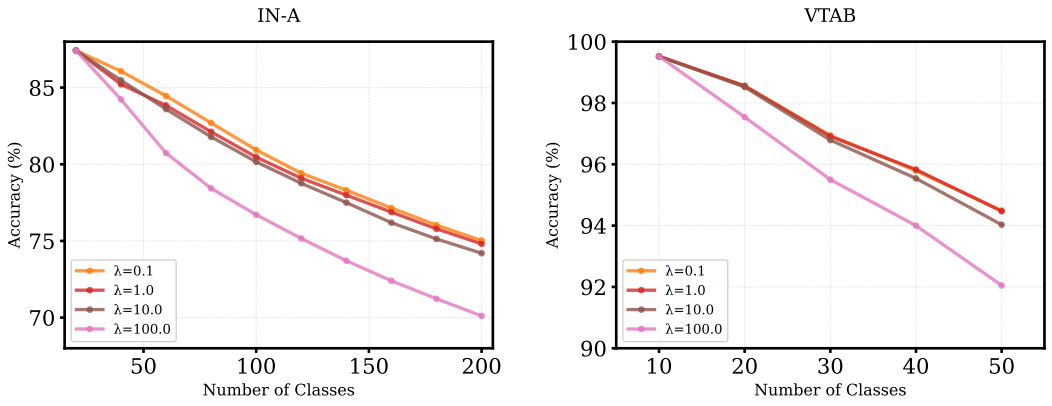
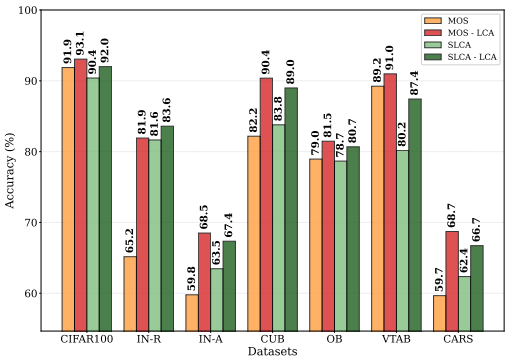


Figure 3: Effect of λ on the accuracy on two datasets.

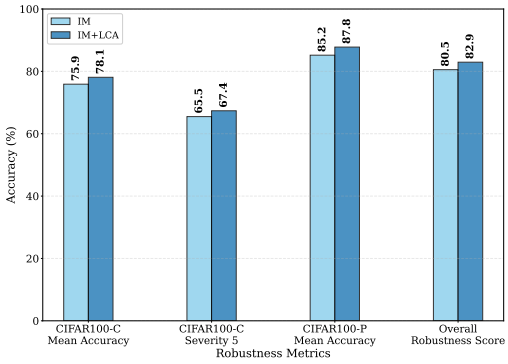
mean and covariance of each encountered class, which scales as $\mathcal{O}(n)$ with the number of classes. Figure 2 further shows that LCA consistently outperforms other methods across most datasets, with a notable improvement of 8% on ImageNet-A compared to the runner-up.

4.2.2 Robustness Measurement

Following the standard (Hendrycks & Dietterich, 2019; Taori et al., 2020), we measure the performance of IM and IM+LCA on two robustness benchmarks, CIFAR100-C and CIFAR100-P. CIFAR100-C is constructed by applying 19 types of common corruptions (e.g., noise, blur, weather, and digital distortions) at 5 severity levels to the original CIFAR-100 test set, thereby evaluating

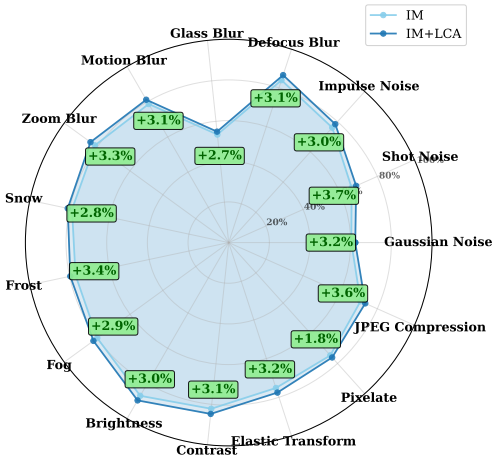


(a) Performance for variants of MOS and SLCA. The postfix LCA means that method uses our alignment loss.

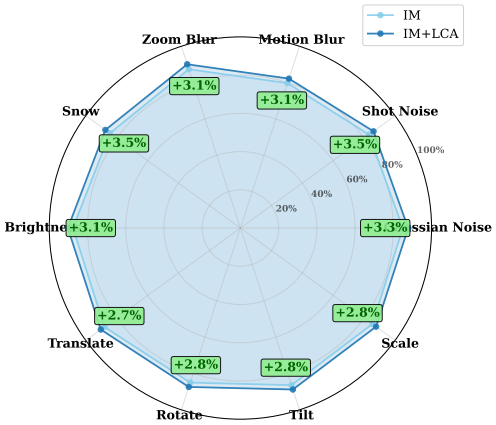


(b) Robustness comparison between IM and IM+LCA on CIFAR100-C and CIFAR100-P. Metrics include mean accuracy, accuracy at severity level 5, and overall robustness score.

Figure 4: (a) Complementary evaluation of LCA when using LCA for MOS and SLCA. (b) Robustness performance of IM and IM+LCA on corruption and perturbation benchmarks.



(a) Performance on CIFAR100-C.



(b) Performance on CIFAR100-P.

Figure 5: Accuracy performance of IM and IM+LCA under different corruption and perturbation types. The relative difference between IM and IM+LCA is highlighted.

model robustness under distribution shift. The final corruption robustness accuracy is computed as

$$\text{Acc}_\text{C} = \frac{1}{19 \times 5} \sum_{c=1}^{19} \sum_{s=1}^{5} A_{c,s},$$

where $A_{c,s}$ is the accuracy under corruption type c with severity s.

CIFAR100-P, in contrast, focuses on prediction stability under perturbations. Each image is perturbed by transformations such as translations, rotations, or noise, and the model's consistency is measured across perturbed versions. The perturbation robustness accuracy is defined as

$$\text{Acc}_\text{P} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{1}{N_p} \sum_{i=1}^{N_p} A_{p,i},$$

where $\mathcal{P}$ denotes the set of perturbation types, N_p is the number of perturbed samples for type p, and $A_{p,i}$ is the accuracy on the i-th perturbed sample.

Finally, to summarize robustness across both benchmarks, we report an overall robustness score:

$$\text{Robustness} = \tfrac{1}{2}\left(\text{Acc}_{\text{C}} + \text{Acc}_{\text{P}}\right).$$

Together, these benchmarks evaluate both accuracy under distributional corruption and resilience of predictions under small perturbations.

Figures 4b and 5 summarize the robustness results. Figure 4b shows that when being trained with LCA, the model obtains a clear improvement on robustness, with more than +2% gain in mean accuracy on CIFAR100-C and a +2.5% gain on CIFAR100-P. The radar plots in Figure 5 further reveal that LCA consistently improves accuracy across all corruption and perturbation types. These results confirm that LCA strengthens robustness both on average and across diverse perturbations.

4.2.3 Ablation Study

LCA as a complementary component. We further evaluate the applicability of LCA on SLCA (Zhang et al., 2023) and MOS (Sun et al., 2025b), two methods that also update the backbone progressively. To asses the impact of our proposed method, we construct two baselines by omitting the final step of these methods, and retaining only the update backbone part. From these, we build their counterparts, SLCA-LCA and MOS-LCA, which include the aligment step as the final step. We measure the average accuracy and report in Figure 4a. We do not find the optimal hyperparameters but fixing the value of λ at 0.1, yet the variants with LCA show improvements on all scenerios, notably in IN-A, CUB, VTAB, and CARS, in some cases even matching the reported results of the original methods (SLCA-LCA achieves 89.0% in CUB compare to 90.0% of the original, MOS-LCA achieves 93.1% in CIFAR100 compare to 94.3%).

Hyperparameter sensitivity. In our method, λ controls the strength of the robustness penalty in the loss function. While an appropriate choice of λ is important for achieving strong performance as Figure 3 shows that overly strong regularization can degrade results. In practice, we find that $\lambda = 0.1$ provides stable and reliable performance across all datasets.

5 Conclusion

This paper investigates the mismatch that may arise between a continuously updated backbone and the task-specific classifiers in continual learning. We propose a theoretically grounded loss function, with a provable bound on classification error, which enables the training of all classifiers using features generated from a Gaussian distribution. Extensive experiments on seven benchmark datasets confirm the effectiveness of our method, and additional robustness evaluations under various noisy conditions demonstrate consistent improvements. While this work primarily focuses on addressing the alignment phase, future research will explore integrating the proposed loss into the end-to-end training pipeline. Such an approach has the potential to further enhance the robustness of the backbone itself and the performance of the overall CIL method.

Despite having significant contributions, our work still remains some limitations. For example, we have not investigated the proposed LCA loss in other contexts. Furthermore, although providing a theoretical foundation and novel insights, the developed theory does not work with the cases when the whole backbone changes over time. Addressing those limitations can open interesting directions for future research.

Reproducibility Statement

We have taken several steps to ensure the reproducibility of our results. The hyperparameters required for both training and inference are explicitly reported in Appendix B. The complete source code is provided as supplementary material, where users can download and unzip the package, follow the installation instructions, and run the main script to reproduce our experiments. All datasets used in this work are publicly available, implementation details and evaluation protocols are described in Section 4.1. Together, these resources are intended to make it straightforward for others to replicate and build upon our work.

REFERENCES

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2):195–204, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2021b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. Fine-grained categorization and dataset bootstrapping using deep learning with humans in the loop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2013. doi: 10.1109/CVPR.2013.83. URL https://ai.stanford.edu/~jkrause/cars/car_dataset.html.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Hongkang Li, Yihua Zhang, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. *arXiv preprint arXiv:2504.10957*, 2025.

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

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123, 2022.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pp. 379–395. Springer, 2024.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

Aristeidis Panos, Yuriko Kobe, Daniel Olmeda Reino, Rahaf Aljundi, and Richard E Turner. First session adaptation: A strong replay-free baseline for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18820–18830, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11909–11919, 2023.

Hai-Long Sun, Da-Wei Zhou, De-Chuan Zhan, and Han-Jia Ye. Pilot: A pre-trained model-based continual learning toolbox. *SCIENCE CHINA Information Sciences*, 68(4):147101, 2025a. doi: https://doi.org/10.1007/s11432-024-4276-4.

Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. Mos: Model surgery for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20699–20707, 2025b.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Quyen Tran, Tung Lam Tran, Khanh Doan, Toan Tran, Dinh Phung, Khoat Than, and Trung Le. Boosting multiple views for pretrained-based continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical report, California Institute of Technology, 2010.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19148–19158, 2023.

Yuanhan Zhang, Zhenfei Yin, Jing Shao, and Ziwei Liu. Benchmarking omni-vision representation through the lens of visual realms. In *European Conference on Computer Vision*, pp. 594–611. Springer, 2022.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23554–23564, 2024.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

A PROOF OF THEOREM 3.1

Proof of Theorem 3.1. Let $p_i = P(\mathcal{Z}_i)$ be the probability measure of area $\mathcal{Z}_i$, and note that $\sum_{i=1}^{t} p_i = 1$.

We first observe that

$$L(P, h_t) \quad = \quad L(P, h_t) - \sum_{i=1}^{C_t} \frac{n_i}{n} L(\mathcal{N}_i, h_t) + \sum_{i=1}^{C_t} \frac{n_i}{n} L(\mathcal{N}_i, h_t) - L(\boldsymbol{D}, h_t) + L(\boldsymbol{D}, h_t) \quad (5)$$

Since $L(P, h_t) = \sum_{i=1}^{C_t} p_i L(\mathcal{N}_i, h_t)$ and $L(\mathcal{N}_i, h_t) \leq \ell_{\max}$, we observe that

$$\begin{aligned} L(P, h_t) - \sum_{i=1}^{C_t} \frac{n_i}{n} L(\mathcal{N}_i, h_t) &= \sum_{i=1}^{C_t} p_i L(\mathcal{N}_i, h_t) - \sum_{i=1}^{C_t} \frac{n_i}{n} L(\mathcal{N}_i, h_t) & (6) \\ &= \sum_{i=1}^{C_t} \left(p_i - \frac{n_i}{n} \right) L(\mathcal{N}_i, h_t) & (7) \\ &\leq \ell_{\max} \sum_{i=1}^{C_t} \left| p_i - \frac{n_i}{n} \right| & (8) \end{aligned}$$

Note that $(n_1, ..., n_{C_t})$ is a multinomial random variable with parameters n and $(p_1, ..., p_{C_t})$. For any $\epsilon > 0$, Bretagnolle-Huber-Carol inequality shows $\Pr\left(\sum_{i=1}^{C_t} \left| p_i - \frac{n_i}{n} \right| \geq 2\epsilon \right) \leq 2^{C_t} \exp(-2n\epsilon^2)$. In other words, for any $\delta > 0$, taking $\epsilon = \sqrt{\frac{C_t \ln 2 - \ln \delta}{2n}}$, the following holds true with probability at least $1 - \delta$:

$$L(P, h_t) - \sum_{i=1}^{C_t} \frac{n_i}{n} L(\mathcal{N}_i, h_t) \quad \leq \quad C \sqrt{\frac{C_t \ln 4 - 2 \ln \delta}{n}} \quad (9)$$

Next we observe the second term:

$$\begin{aligned} \sum_{i=1}^{C_t} \frac{n_i}{n} L(\mathcal{N}_i, h_t) - L(\boldsymbol{D}, h_t) &= \sum_{i=1}^{C_t} \frac{n_i}{n} L(\mathcal{N}_i, h_t) - \sum_{i=1}^{C_t} \frac{n_i}{n} L(\boldsymbol{D}_i, h_t) & (10) \\ &= \sum_{i=1}^{C_t} \frac{n_i}{n} \left[L(\mathcal{N}_i, h_t) - L(\boldsymbol{D}_i, h_t) \right] & (11) \\ &= \sum_{i=1}^{C_t} \frac{n_i}{n} \left(\mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}_i} [\ell(h_t, \boldsymbol{z}) - L(\boldsymbol{D}_i, h_t) : \boldsymbol{z} \in \mathcal{Z}_i] \right) & (12) \\ &= \sum_{i=1}^{C_t} \frac{n_i}{n} \left(\mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}_i, \boldsymbol{s} \sim \boldsymbol{D}_i} [\ell(h_t, \boldsymbol{z}) - \ell(h_t, \boldsymbol{s}) : \boldsymbol{z}, \boldsymbol{s} \in \mathcal{Z}_i] \right) & (13) \\ &\leq \sum_{i=1}^{C_t} \frac{n_i}{n} \left(\mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}_i, \boldsymbol{s} \sim \boldsymbol{D}_i} [|\ell(h_t, \boldsymbol{z}) - \ell(h_t, \boldsymbol{s})| : \boldsymbol{z}, \boldsymbol{s} \in \mathcal{Z}_i] \right) & (14) \\ &= \sum_{i=1}^{C_t} \frac{n_i}{n} \bar{\epsilon}_i(h_t) & (15) \end{aligned}$$

Combining (5) and (9) and (15) completes the proof. □

B TRAINING DETAILS

Table 2: Training and merging hyperparameters used in all experiments of LCA.

Hyperparameter	**Value**
Training epochs	10
Batch size	64
Base learning rate	1×10^{-2}
Weight decay	5×10^{-4}
Optimizer	SGD (momentum = 0.9)
Learning rate scheduler	CosineAnnealing (eta-min = 1×10^{-6})
Merging coefficient α	1.0
Alignment classifier epochs	10
Number of samples per class	512
Alignment batch size	128
LoRA rank	64
LoRA alpha	128
LoRA dropout	0.0
LoRA initialization	Gaussian

C DETAILS OF EXAMPLE GENERATION

In this section, we provide details on how Gaussian distributions are used to align the classifiers during the incremental training process. Pre-trained models typically produce well-structured representations, which allows us to approximate each class distribution using its empirical mean and covariance.

Storing class statistics. For each class c at training stage t, we extract features using the backbone parameters θ_t. The empirical mean $\mu_c \in \mathbb{R}^d$ and covariance $\Sigma_c \in \mathbb{R}^{d\times d}$ are computed as

$$\mu_c = \frac{1}{K}\sum_{i=1}^{|\mathcal{D}_t|} \mathbf{1}(y_i = c)\,\phi(x_i;\theta_t), \tag{16}$$

$$\Sigma_c = \frac{1}{K}\sum_{i=1}^{|\mathcal{D}_t|} \big(\phi(x_i;\theta_t) - \mu_c\big)\big(\phi(x_i;\theta_t) - \mu_c\big)^\top, \tag{17}$$

where $\phi(\cdot;\theta_t)$ denotes the feature extractor with backbone parameters θ_t, and $K = \sum_{i=1}^{|\mathcal{D}_t|} \mathbf{1}(y_i = c)$.

Feature replay via Gaussian sampling. Before each alignment step, we regenerate features for every class $c \in \mathcal{Y}_t$ by sampling from a multivariate Gaussian distribution:

$$\hat{\boldsymbol{z}}_c \sim \mathcal{N}(\mu_c, \Sigma_c). \tag{18}$$

We generate approximately five times the batch size of synthetic features per class, which provides sufficient diversity for classifier alignment.

D FURTHER ANALYSIS

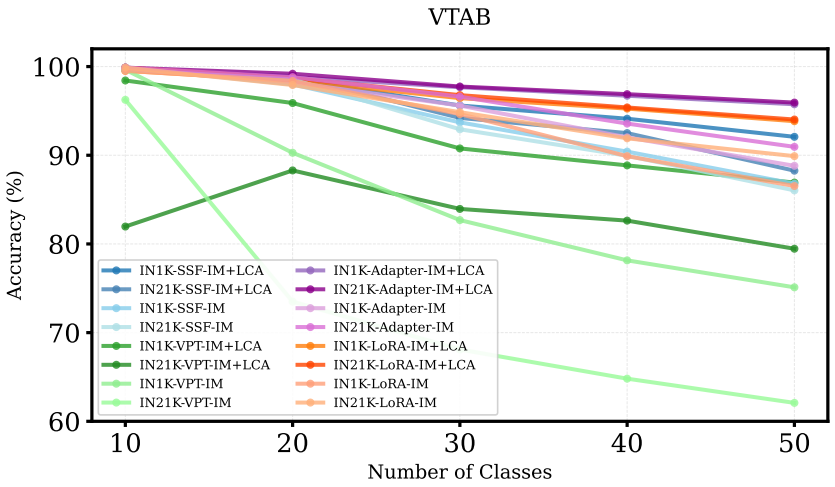


Figure 6: Ablation study on different PEFT strategies with two backbones, ViT-B/16-1K and ViT-B/16-21K.

Results with different fine-tuning strategies. We further examine the adaptability of incremental merging and LCA with different parameter-efficient fine-tuning (PEFT) strategies on the VTAB dataset. The strategies include SSF (Lian et al., 2022), Adapters (Rebuffi et al., 2017), VPT (Jia et al., 2022), and LoRA (Hu et al., 2022), evaluated with ViT-B/16 pretrained on both ImageNet-1K and ImageNet-21K.

Figure 6 shows that LCA consistently improves performance across all PEFT methods. In particular, VPT without LCA suffers a sharp performance drop, suggesting that incremental merging alone is insufficient to prevent forgetting in some cases. Adding LCA, however, proves beneficial across all methods, demonstrating its adaptability. Notably, despite having relatively few trainable parameters, Adapters emerge as a strong candidate within our pipeline, achieving the highest accuracy when combining with LCA.