PANDA - Patch And Distribution-Aware Augmentation for Long-Tailed Exemplar-Free Continual Learning

Siddeshwar Raghavan[1], **Jiangpeng He**[2,*], **Fengqing Zhu**[1]

[1]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, USA
[2]Department of Computer Science, Indiana University, Bloomington, USA
raghav12@purdue.edu, jhe2@iu.edu, zhu0@purdue.edu

Abstract

Exemplar-Free Continual Learning (EFCL) restricts the storage of previous task data and is highly susceptible to catastrophic forgetting. While pre-trained models (PTMs) are increasingly leveraged for EFCL, existing methods often overlook the inherent imbalance of real-world data distributions. We discovered that real-world data streams commonly exhibit dual-level imbalances, dataset-level distributions combined with extreme or reversed skews within individual tasks, creating both intra-task and inter-task disparities that hinder effective learning and generalization. To address these challenges, we propose PANDA, a Patch-and-Distribution-Aware Augmentation framework that integrates seamlessly with existing PTM-based EFCL methods. PANDA amplifies low-frequency classes by using a CLIP encoder to identify representative regions and transplanting those into frequent-class samples within each task. Furthermore, PANDA incorporates an adaptive balancing strategy that leverages prior task distributions to smooth inter-task imbalances, reducing the overall gap between average samples across tasks and enabling fairer learning with frozen PTMs. Extensive experiments and ablation studies demonstrate PANDA's capability to work with existing PTM-based CL methods, improving accuracy and reducing catastrophic forgetting.

Code — https://gitlab.com/viper-purdue/panda
Extended version — https://arxiv.org/abs/2511.09791

Introduction

Continual Learning systems have made remarkable strides in overcoming catastrophic forgetting (Kirkpatrick et al. 2017a; Li et al. 2019; Kemker et al. 2018) and improving learning with growing streams of data. Despite these advancements, a large number of methods still assume perfectly balanced tasks with uniform class distributions (Zhou et al. 2024f; Wang et al. 2024; De Lange et al. 2022). In reality, data streams are both sensitive and unconstrained, and they often follow long-tailed distributions at two levels. Globally, certain classes dominate while others are rare; within a task, this skew can be more extreme or temporarily reversed. This type of two level imbalance remains underexplored in continual learning, resulting in a gap between benchmark studies and real-world applications. For instance, camera traps may log thousands of deer and rabbits, but only a few predators. During a migration window, deer can dominate a week's footage. In medical imaging, pneumonia is the most common condition, while pneumoconiosis appears sporadically and can occasionally overtake pneumonia as the most common case.

Data sensitivity and the high cost of storage make exemplar-based continual learning methods impractical in many real-world settings. Historically, exemplar-free (Li and Hoiem 2016; Kirkpatrick et al. 2017b; Smith et al. 2023b) approaches could not match the performance of exemplar-replay (Rolnick et al. 2018) techniques. However, this has changed with the recent advancement of Pre-Trained Models (PTM) trained on extensive datasets, which has inspired a growing number of continual learning techniques that benefit from their rich feature representations (He, Duan, and Zhu 2025; Goswami et al. 2023; Zhang et al. 2023; Zhou et al. 2024d,a, 2023; Wang et al. 2022b; Smith et al. 2023a; Wang et al. 2022a; Zhou et al. 2024b,c; Sun et al. 2025). Despite these advances, Exemplar-Free Continual Learning (EFCL) is still in its early stages when it comes to handling two-level imbalances. Prior work on imbalanced Long-Tailed-EFCL(LT-EFCL) (Qi et al. 2025; Hong et al. 2024) addresses only what we term single-level imbalance (SLI), designed to tackle a global long-tailed distribution.

To address this, we first **formalize the Dual Level Imbalance (DLI)** setting by introducing task-specific imbalances that deviate from the dataset level distribution. Then, to handle both single-level imbalance (SLI) and our proposed DLI in PTM-based EFCL frameworks, we present PANDA, a Patch-and-Distribution-Aware oversampling module that integrates into any existing EFCL method. PANDA combines two complementary mechanisms: (1)**Intra-task balancing** (within task), using a frozen CLIP encoder to selectively transfer semantic-rich patches from tail class samples to head class samples to equalize distribution. (2) For **Inter-task smoothening** we blend the previous task's minima and maxima with the current task via a learnable β to calibrate classifier thresholds.

We evaluate PANDA across both single and dual level imbalance settings, showcasing the improvement in accuracy and reduced forgetting. Our contributions are as follows:

1. We formalize the Dual Level Imbalance (DLI) setting

*Corresponding Author, Project lead

to inject task specific imbalance that deviates from the overall dataset distribution.

2. We propose PANDA, an integratable patch and distribution aware oversampling module that (a) transfers CLIP-identified patches from rare to common classes to balance intra-task distributions and (b) smoothening inter-task distribution shifts via a learnable β, reducing classifier bias without storing past data.
3. We integrate PANDA into existing PTM EFCL frameworks and demonstrate significant improvements in accuracy and mitigate catastrophic forgetting, highlighting its broad applicability and impact.

Related Work

Continual Learning Categories

Traditionally, continual learning has primarily focused on deep learning models trained from scratch (Zhou et al. 2024f; De Lange et al. 2022; He et al. 2020), categorized into replay-based methods, regularization techniques, and parameter-isolation approaches. **Replay-based** methods (Rebuffi et al. 2016; Rolnick et al. 2018; Isele and Cosgun 2018; Chaudhry et al. 2019; Raghavan, He, and Zhu 2024b) utilize a memory buffer to retain a subset of past data samples, enabling knowledge rehearsal to mitigate forgetting of previously learned classes. **Regularization** methods (Li and Hoiem 2016; Kirkpatrick et al. 2017b; Lee et al. 2017; Jung et al. 2016) incorporate additional penalty terms into the loss function to preserve prior task information while effectively learning new data. In contrast, **parameter-isolation** techniques (Serrà et al. 2018; Mallya and Lazebnik 2018; Mallya, Davis, and Lazebnik 2018; Rusu et al. 2016) assign separate model parameters to different tasks, preventing interference and mitigating forgetting. While most of these approaches focus on balanced settings, recent research has begun exploring **imbalanced** and **long-tailed** scenarios with replay strategies (He et al. 2023; Liu et al. 2022; Bang et al. 2021; Raghavan, He, and Zhu 2024a; Chrysakis and Moens 2020; He 2024).

Exemplar Free Continual Learning With PTM

Continual Learning With PTMs (Zhou et al. 2024d) has demonstrated greater resilience to forgetting and improved performance compared to models trained from scratch, even when they're exemplar-free in nature, making them a promising direction for more efficient continual learning systems. Continual learning with pre-trained models can be broadly categorized into two main approaches: **Prompt-based** methods (Wang et al. 2022b; Smith et al. 2023a; Wang et al. 2022a; Hong et al. 2024) and **Representation-based** methods (Zhou et al. 2024b, 2023, 2024e; Qi et al. 2025; Marouf et al. 2024; Zhang et al. 2023; Goswami et al. 2023; Zhou et al. 2024c; Sun et al. 2025). Prompt-based methods utilize lightweight, trainable parameters (prompts) to guide the model in learning task-specific image samples. These prompts are attached to the input alongside image patches, helping the model adapt efficiently to new tasks while leveraging pre-trained knowledge. Representation-based methods leverage pre-trained knowledge by keeping the backbone completely frozen while replacing classification weights with prototypes, using a nearest mean classifier for classification or adapters or a mix of these strategies.

Data Augmentation Methods

In the context of long-tailed learning and imbalanced datasets, oversampling tail-end classes is a common technique, but focusing on the semantic information is necessary to aid in the augmentation. The Cutout (Devries and Taylor 2017) technique randomly removes regions from the image, while Mixup (Zhang et al. 2017) creates new samples by interpolating two images in the dataset. In contrast, CutMix (Yun et al. 2019) cuts a patch from one image and pastes it onto another, while also adjusting the labels proportionally. A recent data augmentation technique for long-tailed learning employs contrastive learning to generate semantically consistent data, thereby addressing the imbalance (Pan et al. 2024). However, none of these methods are inherently designed for continual learning, where the entire sample distribution is unknown. While balancing task distributions can help mitigate imbalance, the inherent distribution shifts across different tasks can still lead to a biased classifier. Our work, **PANDA**, aims to address this gap by introducing a more adaptive augmentation strategy tailored for continual learning settings.

Methodology

Preliminaries For Long-Tailed Continual Learning

We assume the problem of supervised classification in the context of continual learning (Zhou et al. 2024f), where we encounter a series of N tasks $T_1, T_2, ..., T_N$. Each task T_k where $k \in N$ contains a disjoint set of images from a dataset paired with its corresponding labels. The overall dataset is structured to follow a long-tailed distribution, with the ordering of head and tail classes shuffled to better represent real-world scenarios. Based on this dataset, we construct data streams that we treat as distinct tasks The distribution is characterized by an exponential decay (Cao et al. 2019), parameterized by ρ denotes the ratio between the most and least frequent classes. During training, we can only access the data from the current task T_k, where C^k denotes the number of classes in the task k and n_j^k denotes the number of samples in class j of task k. As we are focusing on the exemplar-free setting, we don't have any storage systems for rehearsal and focus on learning the samples from the current task. Training samples in any task k, denoted as $\mathscr{X}_k = \{x_j^k, y_j^k | j \in \{1, 2, ..., C^k\}\}$ are i.i.d. samples drawn from the current distribution D characterized by ρ, where x_j^k are the images in task k and y_j^k are the corresponding labels. The goal of the continual learner is to classify all classes learned up to task k denoted by $C^{1:k}$, which implies not only learning new task information, but also reducing catastrophic forgetting on previous tasks.

Dual Level Imbalance

Prior work on imbalanced continual learning (Liu et al. 2022; Chrysakis and Moens 2020; Raghavan, He, and Zhu 2024a; Qi et al. 2025; Hong et al. 2024) primarily focusing on the

exponential decay of class samples (controlled by parameter ρ) and do not explicitly control how imbalance impacts individual tasks, we call this the Single-Level Imbalance (SLI) setting. Although SLI setting introduce varying levels of imbalance across the entire dataset, they fail to address task-specific imbalance. To bridge this gap, we propose a DLI setup as shown in Figure 1 that incorporates both dataset-level (controlled by ρ) and task-level (controlled by ρ^*, $*$ denotes the task affected) imbalances. This approach amplifies inter-task imbalance (between different tasks) and can intensify intra-task imbalance, making the problem more realistic and challenging in the context of continual learning.

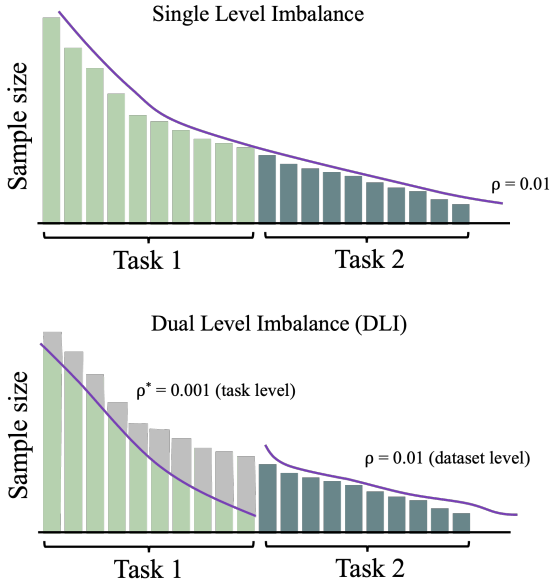

Figure 1: An illustrative figure for the Dual Level Imbalance (DLI) setting: the top image represents the conventional long-tailed CL scenario (SLI), while the bottom image demonstrates the DLI setup, where task-level imbalances are introduced in addition to the dataset-level imbalance. **NOTE: The order of tasks and classes in tasks are shuffled, and are not ordered from head to tail.**

PANDA Framework

We introduce PANDA, a training-free debiasing augmentation framework. The core objective of PANDA is to increase the effective number of training samples for tail classes (least frequent) by leveraging the rich contextual diversity found in head-class images (most frequent). Following the mechanism of vision transformers, each image is divided into $N \times N$ non-overlapping regions, referred to as patches. PANDA synthesizes new images by transferring the most semantically relevant patches, which are identified using a frozen CLIP (Radford et al. 2021) encoder, from tail-class samples to the head-class samples. Images typically consist of an object of interest with background content that does not affect the classification context. By using this knowledge during patch swapping from head to tail classes, PANDA enriches the diversity of tail-class samples and improves the class distribution and reduces bias towards head class data.

At task T_k with sample distribution $D_{T=k}$, the training set is

$$\mathscr{X}_k = \{(x_j^k, y_j^k) \mid j = 1, \ldots, C^k\}.$$

We partition $\mathscr{X}_k$ into head and tail classes according to sample frequency,

$$\mathscr{X}_k = \{\mathscr{X}_k^h,\ \mathscr{X}_k^t\}, \tag{1}$$

$$\text{with distributions} = \ [D_{T=k}^h,\ D_{T=k}^t]. \tag{2}$$

Let each head-class image be denoted as $x^h \in \mathbb{R}^{H\times W\times C}$ with label y^h and each tail-class image as $x^t \in \mathbb{R}^{H\times W\times C}$ with label y^t. Our goal is to compose a new sample (x', y') that balances the intra-task distribution and leverages prior-task distribution statistics to mitigate inter-task distribution shifts.

Using a frozen pretrained CLIP encoder, we partition each image x^h and x^t into N semantic patches, each annotated with positional encoding P_e. To exploit CLIP's joint language–vision alignment, we then convert the class label into a pseudo-sentence of the form:

$$t = \textbf{Image of a } \{\textbf{label}\} \tag{3}$$

We then compute text and image embeddings using the frozen CLIP encoders (Eqn 4).

$$\begin{aligned} z_t &= g(t) \quad \text{(text features)} \\ z_i &= f(P_i) \quad \text{(image patch features)} \end{aligned} \tag{4}$$

The cosine similarity between z_t and each patch feature z_i is computed as

$$S_i = \frac{z_i \cdot z_t}{\|z_i\|\|z_t\|}, \quad i = 1, \ldots, N. \tag{5}$$

For each image, we select the top N/2 patches whose similarity scores exceed a confidence threshold of 0.45 (determined by experiments, overall range 0 to 1), thereby preventing cross-class contamination and eliminating bad patch choices. We denote their indices by

$$i_h^* = \arg\top_{k=N/2}\{S_i^h\}, \quad i_t^* = \arg\top_{k=N/2}\{S_i^t\}.$$

We define binary masks $M^h, M^t \in \{0,1\}^{H\times W}$ so that

$$M(u,v) = \begin{cases} 1, & (u,v) \in \text{patches } i^*, \\ 0, & \text{otherwise.} \end{cases}$$

Let $(M^h)'$ be the inverse of M^h. We compose the new image as

$$x' = (M^h)' \odot x^h \ + \ M^t \odot x^t, \tag{6}$$

$$y' = y^t, \tag{7}$$

where $\odot$ denotes element wise mask multiplication. This procedure grafts the tail object's most semantic patches into the head image while preserving its original context.

To prevent overfitting to these synthetic compositions, we apply standard image augmentations including flip crop, color jitter and Gaussian blur, before passing x' to the learner.

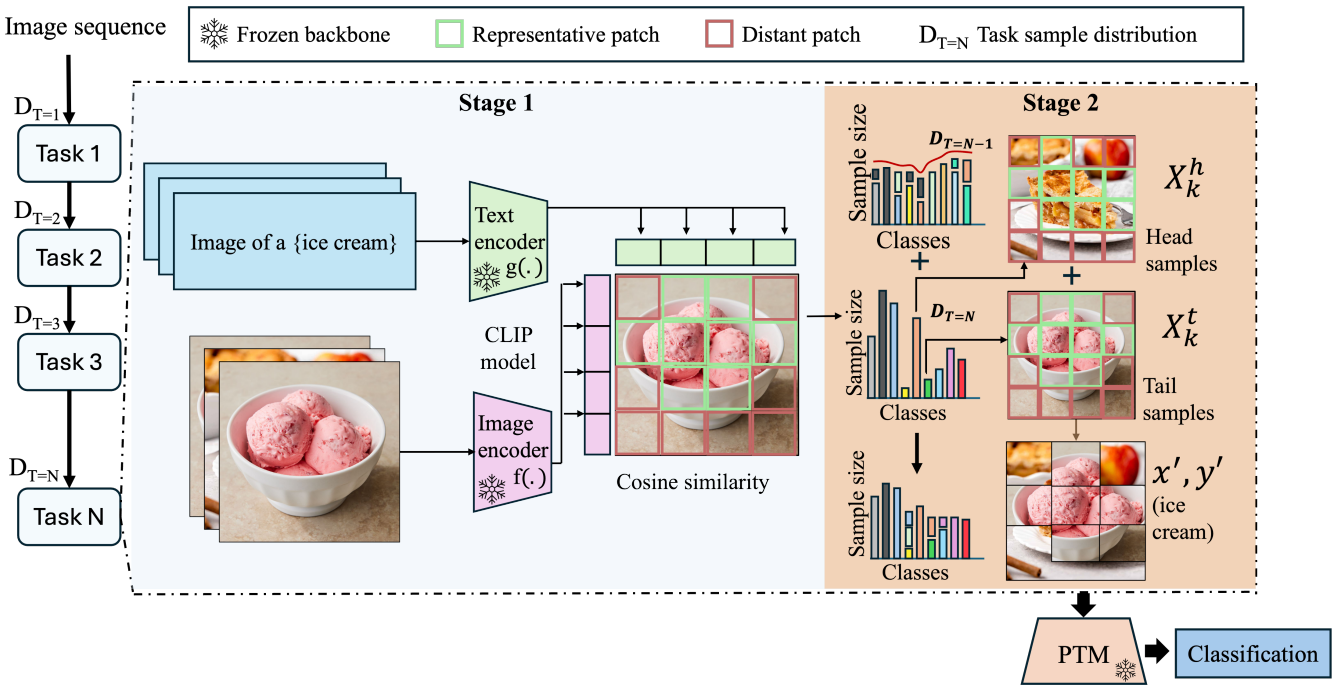

Figure 2: An overview of the PANDA framework, designed to improve Long-Tail Continual Learning through a training-free approach that contextually enriches tail-end samples by utilizing the distributions of head-class samples task-wise. We employ the frozen CLIP (Radford et al. 2021) model to identify and extract the most relevant patches from head-class samples, transferring them to tail-class samples. Additionally, we incorporate prior task distributions to balance the current task, to mitigate inter-class and inter-task imbalances. The augmented samples are then fed into the Continual Learning pipeline.

We iterate this augmentation until the average class counts for head and tail differ by at most q samples, achieving

$$\mathscr{X}_k^h = \mathscr{X}_k^h - M, \tag{8}$$

$$\mathscr{X}_k^t = \mathscr{X}_k^t + M, \tag{9}$$

$$\left|\text{Avg}[D_{T=k}^h] - \text{Avg}[D_{T=k}^t]\right| \leq q. \tag{10}$$

Adaptive Distribution Smoothening To align distributions across tasks, we maintain vectors of prior task maxima and minima and compute

$$\text{adjusted}_m = \beta\,\text{prior}_m + (1-\beta)\,\text{current}_m, \quad m \in \{\min, \max\}. \tag{11}$$

The coefficient β modulates how strongly the previous task's performance influences the current one, smoothing the transition from $(D_{T=k-1})$ to $(D_{T=k})$. When the performance of current task drops compared to the previous one, β is lowered to enable rapid adaptation. An improvement in performance leads to an increase in β to reinforce stability, and if performance is roughly unchanged, β remains the same.

Experiments

In this section, we provide a detailed overview of our experiments and the datasets utilized. We evaluate the performance of existing exemplar free continual learning methods with a PTM backbone on long-tailed datasets (Single and Dual level imbalances) both with and without the integration of our PANDA framework. Finally, we present ablation studies to demonstrate the contribution of each component in our proposed approach.

Datasets

We utilize two widely used publicly available datasets, CIFAR-100 (Krizhevsky 2009) and a 100-class subset of iNaturalist (Van Horn et al. 2018). To introduce class imbalance, we generate long-tailed versions of CIFAR-100 using an exponential decay factor of $\rho = 0.01$, where $0 < \rho \leq 1$, which controls the ratio between the most and least frequent classes (Raghavan, He, and Zhu 2024a) at the dataset level. Task wise imbalance for DLI is controlled by ρ^* where $*$ denotes the affected task. We ensure that the least frequent classes contain a minimum of three samples. The iNaturalist dataset inherently follows a long-tailed distribution representative of real-world scenarios, and we randomly select 100 classes using a fixed seed of 1993. All images were resized to an image resolution of 224×224

Implementation Details

We adapt the publicly available PyTorch implementation of the PTM based CL algorithms from previous work (Sun et al. 2023; Qi et al. 2025; Hong et al. 2024) for our experiments. In our PANDA framework, we utilize a frozen CLIP (Radford et al. 2021) backbone. To ensure consistency, we split

Method	CIFAR100-LT ($\rho = 1$)		CIFAR100-LT ($\rho = 0.01$)		iNaturalist (100 cls)	
	Avg Acc(↑)	Avg For(↓)	Avg Acc(↑)	Avg For(↓)	Avg Acc(↑)	Avg For(↓)
Prompt Methods						
L2P (Wang et al. 2022b)	89.23	6.41	73.34	7.87	78.41	4.72
CodaPrompt (Smith et al. 2023a)	91.30	5.26	76.52	7.55	83.85	4.58
DualPrompt (Wang et al. 2022a)	87.36	10.38	74.24	8.14	81.39	10.69
DAP (Hong et al. 2024)	71.95	18.56	62.98	15.13	66.38	13.67
L2P + PANDA	–	–	**81.32** (↑ 7.98)	**6.08** (↓ 1.79)	**85.47** (↑ 7.06)	**3.37** (↓ 1.35)
CodaPrompt + PANDA	–	–	**87.49** (↑ 2.94)	**4.61** (↓ 2.94)	**90.45** (↑ 6.60)	**3.30** (↓ 1.28)
DualPrompt + PANDA	–	–	**81.00** (↑ 6.76)	**7.38** (↓ 0.76)	**85.44** (↑ 4.05)	**9.44** (↓ 1.25)
DAP + PANDA	–	–	**67.17** (↑ 4.19)	**12.63** (↓ 2.50)	**69.15** (↑ 2.77)	**10.88** (↓ 2.79)
Other Methods						
SimpleCIL (Zhou et al. 2024b)	82.40	7.33	79.01	8.14	89.90	4.48
Adam w/ SSF (Zhou et al. 2024b)	89.05	4.94	86.55	4.63	91.05	2.91
RanPAC (Zhou et al. 2023)	94.89	3.95	90.35	5.22	94.35	2.38
EASE (Zhou et al. 2024e)	92.88	6.65	89.94	6.76	86.91	5.27
CoFiMA (Marouf et al. 2024)	94.29	4.68	93.05	5.57	**94.55** (↑ 0.99)	3.88
SLCA (Zhang et al. 2023)	93.86	7.01	91.73	6.73	92.54	7.22
FeCAM (Goswami et al. 2023)	91.15	4.57	82.99	7.06	87.87	**3.33** (↓ 1.11)
APART (Qi et al. 2025)	86.78	10.86	81.94	13.42	83.47	12.66
APER (Zhou et al. 2024c)	90.93	5.24	87.66	5.66	92.22	3.17
MOS (Sun et al. 2025)	94.26	3.53	91.60	4.69	95.49	2.77
SimpleCIL + PANDA	–	–	**80.20** (↑ 1.19)	**7.98** (↓ 0.26)	**91.92** (↑ 2.02)	**4.44** (↓ 0.04)
Adam w/ SSF + PANDA	–	–	**88.08** (↑ 1.53)	**4.32** (↓ 0.31)	**92.61** (↑ 1.56)	**2.38** (↓ 0.53)
RanPAC + PANDA	–	–	**91.91** (↑ 1.56)	**4.38** (↓ 0.84)	**95.70** (↑ 1.35)	**1.97** (↓ 0.41)
EASE + PANDA	–	–	**91.97** (↑ 2.03)	**6.65** (↓ 0.11)	**92.45** (↑ 5.54)	**5.25** (↓ 0.02)
CoFiMA + PANDA	–	–	**93.83** (↑ 0.78)	**4.91** (↓ 0.66)	93.56	**2.98** (↓ 0.90)
SLCA + PANDA	–	–	**92.05** (↑ 0.32)	**6.23** (↓ 0.50)	**93.27** (↑ 0.73)	**4.58** (↓ 2.64)
FeCAM + PANDA	–	–	**86.48** (↑ 5.95)	**6.68** (↓ 0.38)	**92.42** (↑ 4.55)	4.44
APART + PANDA	–	–	**83.39** (↑ 1.45)	**11.48** (↓ 1.94)	**85.91** (↑ 2.44)	**10.02** (↓ 2.64)
APER + PANDA	–	–	**88.94** (↑ 1.28)	**5.26** (↓ 0.40)	**92.42** (↑ 0.20)	**3.19** (↓ 0.02)
MOS + PANDA	–	–	**92.04** (↑ 0.80)	**4.48** (↓ 0.21)	**95.85** (↑ 0.36)	**2.63** (↓ 0.14)

Table 1: Average top-1 accuracy in the **long-tailed** scenario (single-level imbalance) on CIFAR-100-LT (10 tasks) and iNaturalist (10 tasks) Best accuracy highlighted in **boldface**. NOTE: $\rho = 1$ is the balanced case and PANDA framework has no added effect.

the datasets, construct long-tailed distributions, and set the random seed to 1993. The training batch size is set to 48, while the test batch size is 128. Optimizer and scheduler settings are inherited from each method individually, following the implementations in (Sun et al. 2023; Qi et al. 2025; Hong et al. 2024). Each experiment is repeated 10 times on a single NVIDIA A40 GPU, and we report the average accuracy and average forgetting in Table 1 and the statical confidence is included in the supplementary material. In our PANDA framework, we select the top $N/2$ patches as the most representative based on the cosine similarity scores derived from comparing text and image embeddings.

Evaluation Metrics

In this paper, we use **Average Accuracy** and **Average Forgetting** to evaluate and compare the performance of different continual learning methods with PTMs. Average Accuracy metric quantifies overall performance on a balanced test set from previously encountered tasks. Given a total of k tasks, let $a_{m,n}$ represent the performance on a held-out test set for task n after being trained sequentially from task 1 to m.

$$\text{Average Accuracy } (A_k) = \frac{1}{k}\sum_{i=1}^{k} a_{k,i} \tag{12}$$

Average Forgetting for a task k is calculated as the difference between its maximum performance obtained in the past and the current performance (Wang et al. 2024)

$$F_{m,k} = \max_{i\in\{1,..,k-1\}} \left(a_{i,m} - a_{k,m}\right), \forall j < k \tag{13}$$

Results And Analysis

Conventional Long-Tailed setting In this section, we compare and discuss the performance of a wide range of PTM-based exemplar-free continual learning baselines against their counterparts integrated with our PANDA framework in the single and dual level long-tailed setting. The evaluation is carried out on CIFAR-100 (Krizhevsky 2009), and iNaturalist (Van Horn et al. 2018) with a Single Level Imbalance of $\rho = 0.01$ and summarized in Table 1. It showcases that the performance of existing EFCL methods in the SLI long-tailed setting drops compared to the balanced setting ($\rho = 1$), even on methods designed to tackle imbalanced data streams due to severe head class bias.

Method	$\rho^* = 0.05, * = 2, \rho = 0.01$		$\rho^* = 0.05, * = 3, \rho = 0.01$		$\rho^* = 0.05, * = 4, \rho = 0.01$	
	Avg Acc(%)	Avg For(%)	Avg Acc(%)	Avg For(%)	Avg Acc(%)	Avg For(%)
	CIFAR100-LT (10 tasks)					
CodaPrompt	84.27	4.67	82.88	5.03	82.29	4.80
RanPAC	89.74	4.10	90.21	**2.91**(↓ 0.54)	88.39	4.26
MOS	**93.54** (↑ 1.32)	3.14	92.10	3.34	91.69	4.13
CoFiMA	93.97	4.02	92.18	4.66	90.39	5.17
CodaPrompt + PANDA	**89.44** (↑ 5.17)	**3.62** (↓ 1.05)	**87.77** (↑ 4.89)	**3.46** (↓ 1.56)	**85.35** (↑ 3.06)	**4.62** (↓ 0.18)
RanPAC + PANDA	**90.89** (↑ 1.15)	**3.84** (↓ 0.26)	**92.65** (↑ 2.44)	3.45	**90.05** (↑ 1.66)	**3.93** (↓ 0.33)
MOS + PANDA	92.22	**2.61** (↓ 0.53)	**93.21** (↑ 1.11)	**2.80** (↓ 0.54)	**92.82** (↑ 1.13)	**3.18** (↓ 0.95)
CoFiMA + PANDA	**94.38** (↑ 0.41)	**3.27** (↓ 0.75)	**93.25** (↑ 1.07)	**3.86** (↓ 0.80)	**92.05** (↑ 1.66)	**4.82** (↓ 0.35)

Table 2: Average Accuracy (%) and Forgetting (%) on CIFAR100-LT with dual-level imbalance. ρ indicates dataset level imbalance, ρ^* indicates task level imbalance and $*$ indicates the task. The best results are in **boldface**

For prompt-based methods (L2P (Wang et al. 2022b), CodaPrompt (Smith et al. 2023a), DualPrompt (Wang et al. 2022a), DAP (Hong et al. 2024)), we attribute the performance drop to head-class bias. Prompt tuning modifies only a small set of parameters rather than full model weights, and in highly imbalanced settings this limited adaptability fails to capture tail-class diversity. As a result, tail feature representations remain weak and classification suffers. Even with DAP's dual-anchor design for imbalance, head samples still dominate both adaptation and retention as tail gradients are too small to meaningfully adjust the general prompt, and the stabilizing anchor effectively ignores tail samples.

For **representation-based methods** (SimpleCIL, ADAM, RanPAC, EASE, MOS, APER, APART), except for EASE, the classifier is trained only on the first task, and in subsequent tasks, only the fully connected layer is updated using prototypes from the nearest mean classifier. In contrast, EASE introduces a new adapter for each task while keeping the PTM backbone frozen, allowing it to adapt to task-specific nuances during training. One reason for the lower performance in the long-tailed setting is the assumption that class distributions remain stable when updating the classifier using nearest mean prototypes. However, in continual learning with class imbalance, these distributions shift over time, making the prototypes less reliable and increasing the likelihood of misclassifications.

Lastly, specialized approaches such as FeCAM (Goswami et al. 2023), SLCA (Zhang et al. 2023), and CoFiMA (Marouf et al. 2024) address continual learning from different perspectives. FeCAM leverages prototype-based classification using class centroids, CoFiMA ensembles model weights across tasks, and SLCA adapts learning rates and aligns classifiers to mitigate overfitting. Although effective on balanced data, these methods overfit frequent classes in biased streams, resulting in weak tail-class representations and reduced overall performance. On iNaturalist, we note that CoFiMA without PANDA yields slightly higher accuracy, though PANDA integration markedly reduces forgetting. For FeCAM, the opposite trend holds. We attribute this to the stability–plasticity trade-off, since these methods are already near their performance limits, gains in forgetting reduction often come at the cost of accuracy, and vice versa.

Dual-Level Imbalance (DLI) Setting Next, we compare the results in the DLI setting proposed in our paper and present the performance in Table 2 for the top four methods selected from Table 1. We notice that the existing methods struggle when the both, dataset-level and task-level imbalance is severe. We present 3 different cases with varying levels of imbalance (lower the ρ, more severe the imbalance). ρ represents dataset-level imbalance and ρ^* represents the task-level imbalance. By incorporating PANDA, we demonstrate that leveraging prior task distribution information alongside current task imbalances to stabilize the present task distribution and thereby reduce distribution shifts from previous tasks leading to performance gains.

Ablation Studies

Comparison Against Long-Tailed Learning Augmentation Methods In this section, first, we evaluate PANDA by comparing it with four widely used long-tailed augmentation methods, namely, CutMix (Yun et al. 2019), Mixup (Zhang et al. 2017), Remix (Chou et al. 2020), and Contrastive CutMix (Pan et al. 2024). From Table 3 we can see that the existing methods like CutMix and Mixup do not even reach the baseline performance as the algorithm doesn't take the distributions into the balancing of classes. Comparatively, Remix performs better than the baseline because it is designed to adjust the mixing ratio and region based on the content of the images. This adaptive behavior allows the model to see a more informative and balanced mix of features, leading to better performance. Under single-level imbalance, Contrastive CutMix further improves performance by aligning augmented representations through contrastive supervision. However, under Dual-Level Imbalance (DLI), these methods are unable to cope with large cross-task distribution shifts, while PANDA explicitly addresses this challenge, delivering consistently superior accuracy and minimal forgetting.

Comparison Against Other Masking Methods In PANDA, we use a frozen CLIP encoder to select the most semantically relevant patches for synthesizing tail-class augmentations. To assess this CLIP-based selection, we compare it with Attention Affinity Masking. Following attention-masking techniques from DinoV2 (Oquab et al. 2023), we extract multi-head self-attention maps from a frozen ViT, av-

Augmentation technique	Single level imbalance $\rho = 0.02$		Dual level imbalance $\rho^* = 0.05, * = 3, \rho = 0.02$	
	Avg Accuracy (%)	Avg Forgetting(%)	Avg Accuracy (%)	Avg Forgetting(%)
RanPAC	84.39	5.82	85.07	5.97
RanPAC + CutMix	85.43	7.97	84.03	6.77
RanPAC + Mixup	81.33	8.03	77.29	7.06
RanPAC + Remix	86.50	7.55	86.51	5.73
RanPAC + Con-CutMix	87.27	6.48	84.19	6.01
RanPAC + PANDA (Ours)	**90.31**	**5.03**	**90.08**	**4.52**

Table 3: Average Accuracy (%) and Average Forgetting (%) on **CIFAR100-LT** with single and Dual-Level Imbalance compared with other LT augmentation techniques. ρ indicates dataset level imbalance, ρ^* indicates task level imbalance and $*$ indicates the task. The best results are in **boldface**

erage the heads, normalize, and keep the top N/2 patches by attention score. As shown in Table 4, PANDA's CLIP-based masks consistently outperform affinity-based masking. We attribute this to CLIP's language-guided semantic alignment, which isolates meaningful regions more effectively. Affinity masks often include irrelevant background, and naive cropping lacks semantic focus. Qualitative examples in Figure 3 show that Attention Affinity patching blends head and tail features, causing confusion, whereas PANDA cleanly transfers representative tail regions into the head image, preserving class identity.

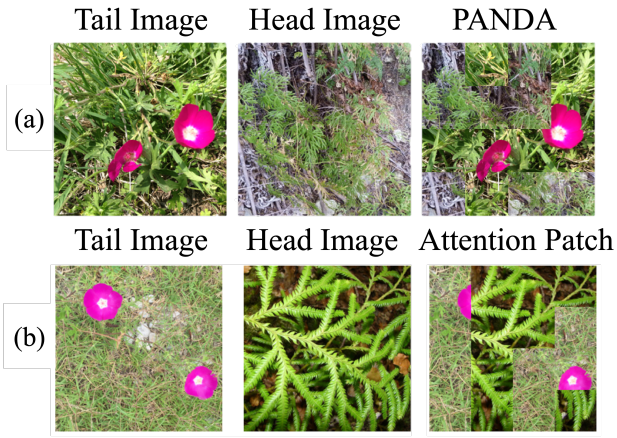

Figure 3: The figure illustrates the effect of (a) **TOP:** PANDA patching and (b) **BOTTOM:** Attention based patching. Attention patching removes critical sections from Tail image while PANDA preserves them.

APART + Masking	Avg Acc (%) (↑)	Avg For (%) (↓)
Baseline	81.94	13.42
Attention Affinity	80.87 (↓ 1.07)	14.63 (↑ 1.21)
PANDA (*ours*)	**83.39** (↑ 1.45)	**11.48** (↓ 1.94)

Table 4: Performance comparison on Attention-Affinity Mask vs. Low Level Feature Transfer vs CLIP-Similarity Mask on 10 tasks, $\rho = 0.01$ with CIFAR100-LT on the APART (Qi et al. 2025). Decreased and improved performance is highlighted.

Resource Usage Additionally, we detail the computational resources required including GPU memory usage and runtime for existing algorithms in Table 5. We compare these metrics to those observed with the integration of PANDA. A modest resource increase substantially boosts continual learning performance by reducing bias, achieved by balancing distributions both within and across tasks.

iNaturalist (100 classes) - 10 tasks		
	GPU Memory (MB)	Run time (Hr)
L2P	2994	1.31
CodaPrompt	19234	1.16
RanPAC	5052	0.33
ADAM w/ SSF	9050	0.61
CoFiMa	16488	1.20
SLCA	12220	0.94
L2P + PANDA	3282 (+288 MB ↑)	1.59 (+0.28 ↑)
CodaPrompt + PANDA	19368 (+134 MB ↑)	1.33 (+0.17 ↑)
RanPAC + PANDA	5517 (+465 MB ↑)	0.42 (+0.09 ↑)
ADAM w/ SSF + PANDA	9384 (+334 MB ↑)	1.07 (+0.46 ↑)
CoFiMa+PANDA	17066 (+ 578 MB ↑)	1.43 (+0.23 ↑)
SLCA+PANDA	12612 (+392 MB ↑)	1.08 (+0.14 ↑)

Table 5: The running time and memory usage for existing methods compared with addition of our PANDA augmentation framework on **iNaturalist subset** dataset (100 classes). We highlighted the increase in resources with our framework

Limitations

While we provide a comprehensive analysis, experimentation, and ablation of our proposed framework, PANDA depends on a pre-trained CLIP model for feature alignment and patch selection, making its performance inherently tied to the efficiency of the model.

Conclusion

This paper introduces Patch and Distribution Aware Augmentation (PANDA) for LT-EFCL, leveraging pre-trained models to address imbalances. PANDA is a training-free method that uses CLIP to extract tail-class patches and integrate them into head-class samples, which balances intra-task imbalances, smoothens inter-task distribution shifts, and reduces bias. We also formalize a Dual-Level Imbalance (DLI) setting for task-level imbalances. Extensive experiments show our method surpasses baselines to improve accuracy and mitigate forgetting.

References

Bang, J.; Kim, H.; Yoo, Y.; Ha, J.-W.; and Choi, J. 2021. Rainbow Memory: Continual Learning with a Memory of Diverse Samples. arXiv:2103.17230.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H. S.; and Ranzato, M. 2019. Continual Learning with Tiny Episodic Memories. *ArXiv*, abs/1902.10486.

Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; and Juan, D.-C. 2020. Remix: Rebalanced Mixup. *ArXiv*, abs/2007.03943.

Chrysakis, A.; and Moens, M.-F. 2020. Online Continual Learning from Imbalanced Data. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1952–1961. PMLR.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2022. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385.

Devries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *ArXiv*, abs/1708.04552.

Goswami, D.; Liu, Y.; Twardowski, B.; and van de Weijer, J. 2023. FeCAM: Exploiting the Heterogeneity of Class Distributions in Exemplar-Free Continual Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

He, J. 2024. Gradient Reweighting: Towards Imbalanced Class-Incremental Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16668–16677.

He, J.; Duan, Z.; and Zhu, F. 2025. CL-LoRA: Continual Low-Rank Adaptation for Rehearsal-Free Class-Incremental Learning. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30534–30544.

He, J.; Lin, L.; Ma, J.; Eicher-Miller, H. A.; and Zhu, F. 2023. Long-tailed continual learning for visual food recognition. *arXiv preprint arXiv:2307.00183*.

He, J.; Mao, R.; Shao, Z.; and Zhu, F. 2020. Incremental Learning in Online Scenario . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13923–13932. Los Alamitos, CA, USA: IEEE Computer Society.

Hong, C.; Jin, Y.; Kang, Z.; Chen, Y.; Li, M.; Lu, Y.; and Wang, H. 2024. Dynamically Anchored Prompting for Task-Imbalanced Continual Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. Also available as arXiv:2404.14721.

Isele, D.; and Cosgun, A. 2018. Selective Experience Replay for Lifelong Learning. *ArXiv*, abs/1802.10269.

Jung, H.; Ju, J.; Jung, M.; and Kim, J. 2016. Less-forgetting Learning in Deep Neural Networks. *ArXiv*, abs/1607.00122.

Kemker, R.; McClure, M.; Abitino, A.; Hayes, T.; and Kanan, C. 2018. Measuring catastrophic forgetting in neural networks. *Proc. Conf. AAAI Artif. Intell.*, 32(1).

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017a. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U. S. A.*, 114(13): 3521–3526.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017b. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report.

Lee, S.-W.; Kim, J.-H.; Jun, J.; Ha, J.-W.; and Zhang, B.-T. 2017. Overcoming catastrophic forgetting by incremental moment matching. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4655–4665. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to Grow: A Continual Structure Learning Framework for Overcoming Catastrophic Forgetting. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3925–3934. PMLR.

Li, Z.; and Hoiem, D. 2016. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40: 2935–2947.

Liu, X.; Hu, Y.-S.; Cao, X.-S.; Bagdanov, A. D.; Li, K.; and Cheng, M.-M. 2022. Long-Tailed Class Incremental Learning. In *European Conference on Computer Vision*, 495–512. Springer.

Mallya, A.; Davis, D.; and Lazebnik, S. 2018. Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. In *European Conference on Computer Vision*.

Mallya, A.; and Lazebnik, S. 2018. PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7765–7773.

Marouf, I. E.; Roy, S.; Tartaglione, E.; and Lathuilière, S. 2024. Weighted Ensemble Models Are Strong Continual Learners.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Pan, H.; Guo, Y.; Yu, M.; and Chen, J. 2024. Enhanced Long-Tailed Recognition With Contrastive CutMix Augmentation. *IEEE Transactions on Image Processing*, 33: 4215–4230.

Qi, Z.-H.; Zhou, D.-W.; Yao, Y.; Ye, H.-J.; and Zhan, D.-C. 2025. Adaptive adapter routing for long-tailed class-incremental learning. *Mach. Learn.*, 114(3).

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Raghavan, S.; He, J.; and Zhu, F. 2024a. DELTA: Decoupling Long-Tailed Online Continual Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Raghavan, S.; He, J.; and Zhu, F. 2024b. Online Class-Incremental Learning For Real-World Food Image Classification. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8180–8189. Los Alamitos, CA, USA: IEEE Computer Society.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2016. iCaRL: Incremental Classifier and Representation Learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5533–5542.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T. P.; and Wayne, G. 2018. Experience Replay for Continual Learning. In *Neural Information Processing Systems*.

Rusu, A. A.; Rabinowitz, N. C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; and Hadsell, R. 2016. Progressive Neural Networks. *ArXiv*, abs/1606.04671.

Serrà, J.; Surís, D.; Miron, M.; and Karatzoglou, A. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023a. 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 (CVPR)*, 11909–11919.

Smith, J. S.; Tian, J.; Halbe, S.; Hsu, Y.-C.; and Kira, Z. 2023b. A Closer Look at Rehearsal-Free Continual Learning. arXiv:2203.17269.

Sun, H.-L.; Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2023. PILOT: A Pre-Trained Model-Based Continual Learning Toolbox. *arXiv preprint arXiv:2309.07117*.

Sun, H.-L.; Zhou, D.-W.; Zhao, H.; Gan, L.; Zhan, D.-C.; and Ye, H.-J. 2025. MOS: Model Surgery for Pre-Trained Model-Based Class-Incremental Learning. In *AAAI*, 20699–20707.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The iNaturalist Species Classification and Detection Dataset . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8769–8778. Los Alamitos, CA, USA: IEEE Computer Society.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5362–5383.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022a. DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning. *European Conference on Computer Vision*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6022–6031.

Zhang, G.; Wang, L.; Kang, G.; Chen, L.; and Wei, Y. 2023. 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*.

Zhang, H.; Cissé, M.; Dauphin, Y.; and Lopez-Paz, D. 2017. mixup: Beyond Empirical Risk Minimization. *ArXiv*, abs/1710.09412.

Zhou, D.-W.; Cai, Z.-W.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024a. Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. arXiv:2303.07338.

Zhou, D.-W.; Cai, Z.-W.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024b. Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. *International Journal of Computer Vision*.

Zhou, D.-W.; Cai, Z.-W.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024c. Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. *International Journal of Computer Vision*.

Zhou, D.-W.; Sun, H.-L.; Ning, J.; Ye, H.-J.; and Zhan, D.-C. 2024d. Continual learning with pre-trained models: A survey. In *IJCAI*, 8363–8371.

Zhou, D.-W.; Sun, H.-L.; Ye, H.-J.; and Zhan, D.-C. 2024e. Expandable Subspace Ensemble for Pre-Trained Model-Based Class-Incremental Learning. In *CVPR*, 23554–23564.

Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024f. Class-Incremental Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9851–9873.

Zhou, D.-W.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2023. Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. *Proceedings of the Neural Information Processing Systems*.