Revisiting Continual Ultra-fine-grained Visual Recognition with Pre-trained Models

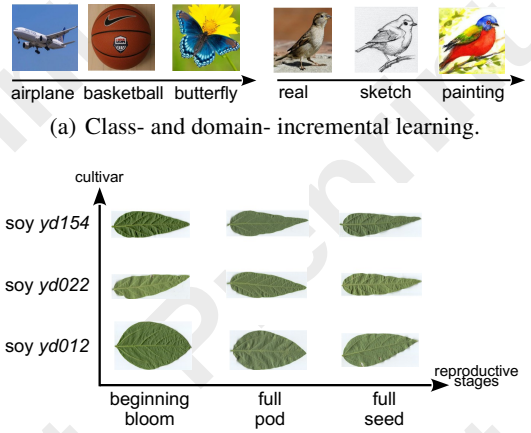
Pengcheng Zhang[1], **Xiaohan Yu**[2], **Meiying Gu**[1], **Yuchen Wu**[1], **Yongsheng Gao**[3] and **Xiao Bai**[*1]
[1]School of Computer Science and Engineering, State Key Laboratory of Complex & Critical Software Environment, Jiangxi Research Institute, Beihang University, China
[2]School of Computing, Macquarie University, Australia
[3]School of Engineering and Built Environment, Griffith University, Australia
pengchengz@buaa.edu.cn, xiaohan.yu@mq.edu.au, {gumeiying, wuyuchen}@buaa.edu.cn,
yongsheng.gao@griffith.edu.au, baixiao@buaa.edu.cn

Abstract

Continual ultra-fine-grained visual recognition (C-UFG) aims to continuously learn to categorize the increasing number of cultivates (VC-UFG) and consistently recognize crops across reproductive stages (HC-UFG), which is a fundamental goal of intelligent agriculture. Despite the progress made in general continual learning, C-UFG remains an underexplored issue. This work establishes the first comprehensive C-UFG benchmark using massive soy leaf data. By analyzing recent pre-trained model (PTM) based continual learning methods on the proposed benchmark, we propose two simple yet effective PTM-based methods to boost the performance of VC-UFG and HC-UFG, respectively. On top of those, we integrate the two methods into one unified framework and propose the first unified model, **Unic**, that is capable of tackling the C-UFG problem where VC-UFG and HC-UFG coexist in a single continual learning sequence. To understand the effectiveness of the proposed methods, we first evaluate the models on VC-UFG and HC-UFG challenges and then test the proposed Unic on a unified C-UFG challenge. Experimental results demonstrate the proposed methods achieve superior performance for C-UFG. The code is available at https://github.com/PatrickZad/unicufg.

1 Introduction

Recent years have witnessed the great success of deep learning for general [Deng *et al.*, 2009; Ridnik *et al.*, 2021] and fine-grained [Wah *et al.*, 2011; Van Horn *et al.*, 2018] visual recognition problems. Differently in agricultural production, crops of the same species are only identifiable at an ultra-fine granularity [Yu *et al.*, 2021b], which brings a more challenging goal, ultra-fine-grained visual recognition (UFG), to develop AI techniques for agriculture. Existing works [Yu *et al.*, 2023b; Wang *et al.*, 2021; Yu *et al.*, 2023a; Yu *et al.*, 2022] have largely promoted the development of UFG on a fixed set of categories. However, these techniques are not applicable to handle either the increasing number of new cultivars or the various reproductive stages of the same category.



(a) Class- and domain- incremental learning.

(b) Continual ultra-fine-grained recognition.

Figure 1: Illustration of (a) existing typical continual learning problems and (b) the proposed continual ultra-fine-grained problem.

To this end, we propose to formulate the continual ultra-fine-grained visual recognition (C-UFG) problem, a two-dimensional challenge as in Figure 1(b), which involves both continually learning to recognize new crop cultivars and consistently adapting to new reproductive stages of seen categories. Formally, we refer to continual learning of increasing cultivars as vertical C-UFG (VC-UFG), and continual learning of growing crops as horizontal C-UFG (HC-UFG). As discussed by [Wang *et al.*, 2024], the former can be viewed as a class-incremental learning (CIL) problem where the increasing tasks have disjoint label spaces, and the latter is an instance of domain-incremental learning (DIL) where all tasks share the same label space. While most previous continual learning methods focus only on one of the two scenarios (Figure 1(a)), a realistic C-UFG problem is a unified continual learning challenge that requires tackling both the increasing crop cultivars and reproductive stages of seen cultivars in a single learning sequence.

*Corresponding author.

While sequentially training the model on multiple tasks, both CIL and DIL assume that the task ID is unknown during inference. Thus the key challenge in continual learning is to effectively adapt to a new task (plasticity) without catastrophic forgetting of seen tasks (stability) [De Lange *et al.*, 2021]. To tackle this problem, early works mainly explored *Regularization-based* methods [Li and Hoiem, 2017; Kirkpatrick *et al.*, 2017] that introduce learning regularization to constrain the disruption of learned knowledge when training on new data, and *replay-based* methods [Lopez-Paz and Ranzato, 2017; Rolnick *et al.*, 2019] that memorize a subset of previous raw images or embedded features to jointly learn with new data. Despite their effectiveness on general continual learning problems, these methods struggle on ultra-fine-grained data as UFG requires substantial optimization of parameters to adapt to each task [Zhang *et al.*, 2023]. In contrast, recent continual learning methods with pre-trained models (PTMs) [Cha *et al.*, 2021; Wang *et al.*, 2022c; Wang *et al.*, 2022b; Smith *et al.*, 2023] are shown to be superior given the stable pre-trained features and highly plastic parameter-efficient tuning mechanism [Jia *et al.*, 2022; Li and Liang, 2021]. This motivates us to first conduct a comprehensive analysis of existing PTM-based continual learning methods and then explore more effective AI models for C-UFG with PTMs.

Specifically, we establish a benchmark of C-UFG consisting of challenging VC-UFG, HC-UFG and unified C-UFG tasks. By evaluating existing open-sourced PTM-based continual learning methods on the VC-UFG and HC-UFG tasks, we observe significant performance decay of the methods compared to their results on general objects while two co-exist designs, task-aware prompt learning and selective prompt reuse, still benefit C-UFG. We further discover that the key challenges that limit the final performance are the inter-task prompts in VC-UFG and the classifier bias in HC-UFG. Based on the preliminary analysis, we first propose two simple yet effective PTM-based methods for VC-UFG and HC-UFG, respectively. For VC-UFG, we employ deep and dense keys to reduce the use of inter-task prompts. Adaptive Classifier Adaptation is proposed to improve the robustness of classifiers to inter-task prompts. For HC-UFG, we propose to learn task-shared momentum prompts to alleviate the bias of classifiers on more recent domains. On top of those, we propose a unified C-UFG model, **Unic**, to integrate the two methods into a unified framework to tackle realistic hybrid C-UFG problems where VC-UFG and HC-UFG tasks exist in a single continual learning sequence.

In summary, this work makes the following contributions:

- We formulate a realistic continual ultra-fine-grained recognition problem for agriculture production. Comprehensive C-UFG datasets are established to facilitate future development of advanced techniques.
- By analyzing existing PTM-based continual learning methods on the proposed C-UFG datasets, we design two simple yet effective PTM-based methods for VC-UFG and HC-UFG, respectively. On top of those, we propose a unified model, Unic, to effectively tackle the more realistic hybrid C-UFG problems.
- We evaluate the proposed method on VC-UFG, HC-UFG and unified C-UFG datasets. The evaluation results demonstrate that our proposed method boosts the PTM-based continual learning performance on the three challenging problems.

2 Related Work

Ultra-fine-grained Visual Recogntion. With the advancement of deep learning, fine-grained visual recognition (FG) [Wah *et al.*, 2011; Van Horn *et al.*, 2018; Liu *et al.*, 2022] has greatly developed in the last decade. However, identifying objects at a very fine granularity, *i.e.* UFG [Yu *et al.*, 2021b], is still a challenging task. Compared with FG, UFG relies on genetic source banks rather than human experts to obtain accurate data labels. It promotes the classification granularity from the species level to a subordinate level where even human experts can hardly identify the visual difference between two categories. For UFG,, [Yu *et al.*, 2021b] established the first comprehensive benchmark of ultra-fine-grained cultivar leaf data and verified the performance of modern neural networks. A key challenge in UFG is the over-fitting of training data. For this, MaskCov [Yu *et al.*, 2021a] proposed a random mask covariance network for representation learning. Spare [Yu *et al.*, 2022] proposed part representation learning and erasing in a self-supervised framework. Benefiting from the superiority of vision transformers [Dosovitskiy *et al.*, 2021], transformer-based models for UFG [Yu *et al.*, 2023b; Yu *et al.*, 2023a] further boost the overall performance.

Continual Learning with PTMs. PTMs enable effective replay-free continual learning. L2P [Wang *et al.*, 2022c] for the first time proposed to introduce a prompt pool with paired prompts and keys with adaptive selection of the prompts for inference. DualPrompt [Wang *et al.*, 2022b] further improved the prompt learning and introduced task-shared prompts. CODA-P [Smith *et al.*, 2023] designed an attention mechanism to learn weights for merging the task-specific prompts. S-Prompts [Wang *et al.*, 2022a] proposed a simple yet effective clustering-based prompt select mechanism for domain-incremental scenarios. LAE [Gao *et al.*, 2023] proposes a unified PTM-based continual learning frameworks for different parameter-efficient fine-tuning [Jia *et al.*, 2022; Houlsby *et al.*, 2019; Hu *et al.*, 2022] methods. HiDE [Wang *et al.*, 2023] mainly analyzes the effect of supervised and self-supervised PTMs and decompose the PTM-base continual learning problem to improve the continual learning performance of self-supervised PTMs. VQ-Prompt [Jiao *et al.*, 2024] proposes a vector-quantization framework for prompt learning to further enhance the continual learning capability of discrete task-specific prompts.

We also note that a previous work [Zhang *et al.*, 2023] made an attempt in C-UFG. However, the previous C-UFG challenge takes cultivates at different reproductive stages as different categories and regards that a VC-UFG problem, which neither considers the inherent invariance between the reproductive stages nor the existence of HC-UFG problems. Moreover, the more realistic unified C-UFG problem is also ignored. In contrast, this work presents a more comprehensive C-UFG benchmark that establishes VC-UFG, HC-UFG

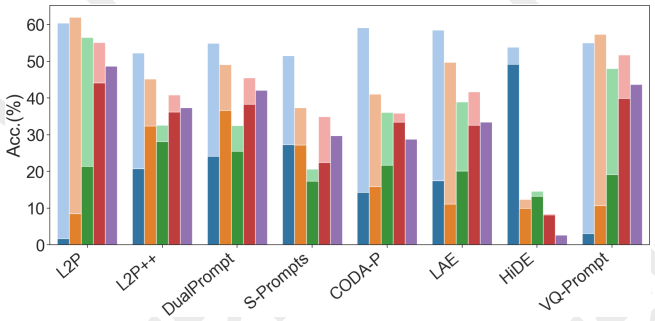


Figure 2: Model classification accuracies on SoyGene-C.

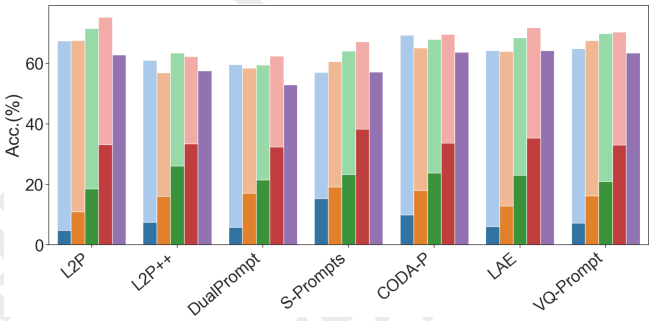


Figure 3: Model classification accuracies on SoyAgeing-C.

and unified C-UFG challenges to better mitigate the gap between research and real-world agriculture demands.

3 Preliminary Analysis

3.1 Problem Definition

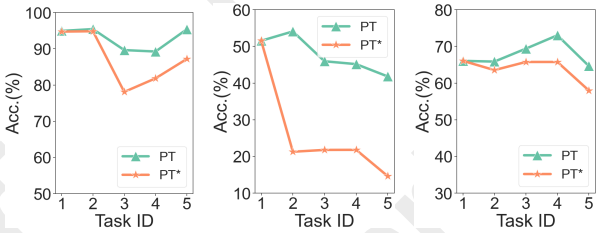
Formally, we denote by T the total number of UFG tasks that sequentially arrive in a C-UFG learning sequence. For each task $t \in \{1, 2, \ldots, T\}$, we refer to $\mathcal{D}_t = \{\mathcal{X}_t, \mathcal{Y}_t\}$ as its data where $\mathcal{X}_t$ stands for the images and $\mathcal{Y}_t$ is the corresponding labels, respectively. During continual learning, a C-UFG model learns on the training sets of the T tasks one by one. Afterward, the model is evaluated on all T test sets. In VC-UFG, the labels spaces of any two different tasks are disjoint, *i.e.* $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ given $i \neq j$. While in HC-UFG, all tasks share the same label space, *i.e.* $\mathcal{Y}_i = \mathcal{Y}_j$ for $i, j \in \{1, 2, \ldots, T\}$.

Similar to previous continual learning datasets [De Lange *et al.*, 2021; Wang *et al.*, 2024], we construct the following C-UFG datasets the data released by [Yu *et al.*, 2021b]:

- **SoyGene-C** where we evenly separate the original $1,110$-class SoyGene dataset into five 222-class UFG tasks. This formulates a typical VC-UFG challenge.
- **SoyAgeing-C** where we treat the data at five reproductive stages in the original SoyAgeing dataset as an HC-UFG challenge.
- **SoyGlobal-C** where we separate the original $1,938$ classes of SoyGlobal into eight 216-class sets and one 210-class set. This presents a more difficult VC-UFG challenge with a long task sequence and limited training samples.
- **UniUFG-C** where we alternately arrange the first three tasks in SoyGene-C and SoyAgeing-C to form a continual learning sequence containing both VC-UFG and HC-UFG scenarios. This presents a more realistic and challenging unified C-UFG problem.

3.2 Evaluation and Discussion

For analysis, we include recent PTM-based continual learning methods L2P [Wang *et al.*, 2022c], DualPrompt [Wang *et al.*, 2022b], S-Prompts [Wang *et al.*, 2022a], CODA-P [Smith *et al.*, 2023], LAE [Gao *et al.*, 2023], HiDE [Wang *et al.*, 2023], VQ-Prompt [Jiao *et al.*, 2024], for comparisons. For L2P, we empirically found that introducing deep and task-aware prompt tuning as in DualPrompt, which we refer to as L2P++, largely improves the results. Although S-Prompts is designed for DIL, it is also suitable for VC-UFG. While the CIL models are also applicable in HC-UFG when simply treating the same cultivar at different reproductive stages as different categories, they ignore the intrinsic invariance of crops during growing and introduce extra computation complexity. For this, we replace the expandable classifiers with fix-sized ones in those methods in HC-UFG. HiDE is excluded for HC-UFG as its task identity prediction requires expandable classifiers.



(a) CIFAR-100-5. (b) SoyGene-C. (c) SoyAgeing-C.

Figure 4: Prompt tuning on tasks in SoyGene-C, CIFAR-100-5 and SoyAgeing-C.

We mainly test the methods on SoyGene-C and SoyAgeing-C as in Figure 2 and Figure 3, respectively. The immediate (**light colors**) and final (**dark colors**) classification accuracies on each task are grouped together for illustration. The former is the best accuracy obtained during continual learning and the latter gives the model performance after training.

On SoyGene-C, while most methods show similar immediate results, L2P++ and DualPrompt achieve superior overall final accuracy. By comparing the two methods with the others, we observe two co-exist key designs shared by L2P++ and DualPrompt that make them different from their counterparts, *i.e.* **task-aware prompt learning** and **selective prompt reuse**. Specifically, L2P++ and DualPrompt assign nonoverlapping task-oriented prompts for each task and optimize only the correlated prompts when learning a specific task. During inference, a subset of learned prompts is adaptively selected for reuse. S-Prompts also shares the two de-

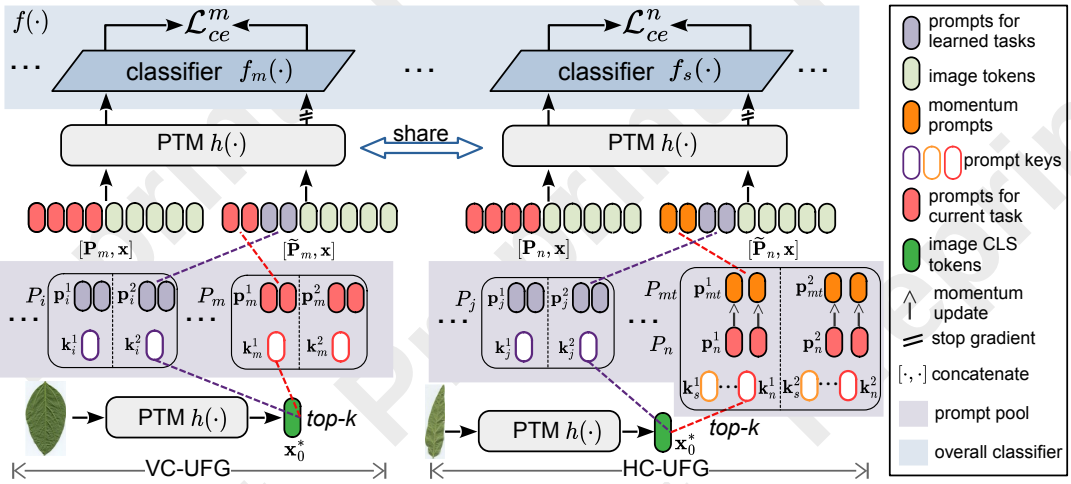


Figure 5: The overall architecture of the proposed Unic for unified C-UFG. For the convenience of illustration, we show only a single prompted layer and the associated prompts. We show a unified C-UFG learning sequence where task m is a VC-UFG task and task n is a HC-UFG task. The left part illustrates how Unic learns on a VC-UFG task and the right part shows that on an HC-UFG task.

signs except that its prompt selection is less compatible with UFG tasks, which explains its overall performance. Differently, L2P, VQ-Prompt optimize the prompts in a task-agnostic manner, and CODA-P, LAE, HiDE design different ways to fuse the learned prompts for reuse.

For VC-UFG, **task-aware prompt learning** guarantees disjoint optimization of prompts across tasks while the task-agnostic mechanism may change the already learned prompts in a new task as the semantic boundary between UFG categories is extremely ambiguous. For **selective prompt reuse**, we hypothesize that it alleviates the impact of involving inter-task prompts for inference compared with the prompt fusing methods. To verify the impact of inter-task prompts, we conduct a comparison between two prompt tuning models, PT* and PT, on SoyGene-C and a 5-split version of CIFAR-100 pacman-key –populate archlinux[Krizhevsky *et al.*, 2009]. PT is trained independently on the tasks to show the effectiveness of inner-task prompts, while PT* learns the prompts only on the first task and reuses them as inter-task prompts in the subsequent tasks where only the classifiers are learnable. As in Figure 4(a), inter-task prompts do not largely hinder the model performance on general data, while Figure 4(b) suggests that inter-task prompts cause significant decay of recognizing UFG objects. This is mainly due to that learning to recognize UFG categories requires substantially fitting the prompts in the data and limits the generalizability. As fusing learned prompts always injects inter-task prompts for inference, it limits the continual learning performance.

Differently on SoyAgeing-C, we observe that the compared methods show similar overall performance where the early learned knowledge is largely forgotten. By further evaluating PT and PT* on SoyAgeing-C tasks as in Figure 4(c), we surprisingly find that using inter-task prompts is clearly less harmful than on SoyGene-C tasks. This suggests that learning a single generalizable group of prompts for HC-UFG tasks is possible, and a key challenge in HC-UFG is to tackle the bias of the fix-sized classifiers on recently seen data.

Despite the effectiveness of existing methods on VC-UFG or HC-UFG, a more realistic yet underexplored C-UFG problem is a mixture of VC-UFG and HC-UFG where a new UFG task can either introduce new cultivates or present a new reproductive stage of seen cultivates. This motivates us to further develop a unified model that tackles the two C-UFG challenges in a single continual learning procedure.

4 Method

Based on the discussion in Section 3.2, we propose a unified C-UFG model, **Unic**, that is capable of tackling VC-UFG and HC-UFG in a single continual learning procedure. In this section, we first present our effective solutions for VC-UFG and HC-UFG in Section 4.1 and Section 4.2, respectively. We then explain how the two methods can be integrated into a unified model for C-UFG in Section 4.3.

4.1 Continual Learning for VC-UFG

Baseline. For VC-UFG, we employ L2P++ as the baseline model for its simple yet effective design. When learning on the m-th VC-UFG task, L2P++ introduces a set of task-aware prompts $P_m = \{\mathbf{p}_m^l\}_{l=1}^L$, where $\mathbf{p}_m^l \in \mathbb{R}^{M\times N\times D}$, M is the number of prompted layers. By performing deep visual prompt tuning [Jia *et al.*, 2022] with a cross-entropy loss $\mathcal{L}_{ce}^m$, a local classifier f_m and the prompts and jointly learned for the m-th task. Meanwhile, $\mathbf{p}_m^l$ is associated with a learnable key $\mathbf{k}_m^l \in \mathbb{R}^D$. For a training image I, the pre-trained ViT [Dosovitskiy *et al.*, 2021] is employed to extract a global pre-trained feature $\mathbf{x}_0^* \in \mathbb{R}^D$. A matching loss

$$\mathcal{L}_{match}^m = \frac{1}{L}\sum_{l=1}^{L}\left(1 - \frac{\mathbf{k}_m^l \cdot \mathbf{x}_0^*}{|\mathbf{k}_m^l|\,|\mathbf{x}_0^*|}\right) \quad (1)$$

is then introduced to optimize the cosine similarity between $\mathbf{k}_m^l$ and $\mathbf{x}_0^*$. For inference, the model concatenates all learned local classifiers as the global classifier. By measuring the similarity between $\mathbf{x}_0^*$ of a test image and all learned keys, the

prompts associated with the *top-k* similar keys, where $k = L$, are adaptively selected for reuse.

Deep and dense keys. As in Figure 4(b), reusing the inter-task prompt can largely hinder the VC-UFG results. To reduce the inter-task prompts for inference, we introduce deep and dense keys for adaptive selection. Different from that in L2P++, we assign a learnable key $\mathbf{k}_m^l \in \mathbb{R}^D$ for each task-aware prompt $\mathbf{p}_m^l \in \mathbb{R}^{N\times D}$ at each prompted layer, resulting in deep keys similar to deep prompts. We then proportionally increase the number of prompts L and decrease the prompt length N. This makes the keys more dense without changing the total prompt length LN.

Adaptive Classifier Adaptation. Without knowing the task identity, inter-task prompts are inevitably involved during L2P++ inference. For this, we propose to learn a training sample with both task-agnostic and task-aware prompts, where the task-agnostic prompts are selected as in inference. Learning with task-agnostic prompts involves two learnable modules, *i.e.* the selected prompts and the local classifier. However, optimizing the task-agnostic prompts limits the final VC-UFG performance as discussed in Section 3.2. This inspires us to perform Adaptive Classifier Adaptation (**ACA**) where only the classifier is optimized to adapt to the task-agnostic prompts.

As in the left part of Figure 5, we show only a single prompted layer and the prompts for illustration purpose. Learning on the m-th task, which is a VC-UFG task, is supervised by a classification loss

$$\mathcal{L}_{ce}^m = \mathcal{L}(f_m(h([\mathbf{P}_m, \mathbf{x}])), y) + \lambda_v \mathcal{L}(f_m(h([\widetilde{\mathbf{P}}_m, \mathbf{x}])), y) \tag{2}$$

where $\mathcal{L}$ is the softmax cross-entropy loss and y is the label of the training image. $\lambda_v \in (0, 1])$ is a constant factor of ACA. $\mathbf{P}_m, \widetilde{\mathbf{P}}_m \in \mathbb{R}^{LN\times D}$ are the concatenated task-aware and task-agnostic prompts. The gradient propagated to $\widetilde{\mathbf{P}}_m$ is stopped to avoid changing the prompts. Meanwhile, the prompt keys are optimized as in L2P++. Thus the overall VC-UFG training loss is given by

$$\mathcal{L}_V^m = \mathcal{L}_{ce}^m + \beta_v \mathcal{L}_{match}^m \tag{3}$$

where β is the weight of the matching loss.

4.2 Continual Learning for HC-UFG

Based on the observation in Figure 3, we assume that a single shared group of prompts is sufficient to tackle a series of HC-UFG tasks that share the same label space. This also helps to alleviate the bias of the shared classifier to more recent data. For this, we introduce the momentum prompts to encode the common HC-UFG knowledge during continual learning. We also make the classifier adapt to the momentum prompts similar to ACA.

Specifically, we start with standard deep prompt tuning on the first HC-UFG task, resulting in learned prompts $P_1 = \{\mathbf{p}_1^l\}_{l=1}^L$ for each prompt layer and the classifier f_1. We also initialize the momentum prompts $P_{mt} = \{\mathbf{p}_{mt}^l\}_{l=1}^L$ using P_1. Then for the n-th task ($n > 1$), we initialize the prompts $P_n = \{\mathbf{p}_n^l\}_{l=1}^L$ by P_{n-1} and perform deep prompt tuning to update P_n. After each update of P_n, we also update P_{mt} as

$$\mathbf{p}_{mt}^l = \mu \mathbf{p}_{mt}^l + (1-\mu)\mathbf{p}_n^l, l = 1, \cdots, L \tag{4}$$

where $\mu \in (0, 1)$ is the momentum factor. The overall training loss for HC-UFG is thus given by

$$\mathcal{L}_{ce}^n = \mathcal{L}(f_1(h([\mathbf{P}_n, \mathbf{x}])), y) + \lambda_h \mathcal{L}(f_1(h([\mathbf{P}_{mt}, \mathbf{x}])), y) \tag{5}$$

where $\mathcal{L}$ is the softmax cross-entropy loss and y is the label of the training image. $\lambda_v \in (0, 1])$ is a constant factor. We also stop the gradient propagated to $\mathbf{P}_{mt}$. For inference, the momentum prompts $\mathbf{P}_{mt}$ are used for recognizing any ultra-fine-grained object.

4.3 A Unified Model for C-UFG

Based on the continual learning methods for VC-UFG and HC-UFG introduced in Section 4.1 and Section 4.2, we further propose a unified model **Unic** to tackle a more realistic C-UFG problem where VC-UFG and HC-UFG co-exist. As illustrated in Figure 5, suppose task m in the C-UFG learning sequence introduces a set of new categories and task n shares the same label space $\mathcal{Y}_m$ with task s where $s < n$ and $\mathcal{Y}_s \cap \mathcal{Y}_i = \emptyset$ for $i < s$. Learning on task m is thus a VC-UFG problem and task n introduces an HC-UFG problem.

For task m, Unic keeps the design of keys and prompts as in Section 4.1. For task n, while we employ the shared momentum prompts in Unic as in Section 4.2, we also introduce learnable keys similar to that in VC-UFG to adaptively select learned prompts during inference. Note that this makes each momentum prompt $\mathbf{p}_{mt}^l$ associated with the keys learned on all HC-UFG tasks sharing the same label space, *i.e.* $\{\mathbf{k}_c^l | s \le c \le n, \mathcal{Y}_c = \mathcal{Y}_s = \mathcal{Y}_n\}$. To avoid duplicate selection of the same momentum prompt, we calculate the matching score of $\mathbf{p}_{mt}^l$ for the *top-k* selection as

$$\max_{c\in[s,n], \mathcal{Y}_c=\mathcal{Y}_s=\mathcal{Y}_n} \frac{\mathbf{k}_c^l \cdot \mathbf{x}_0^*}{|\mathbf{k}_c^l|\,|\mathbf{x}_0^*|} \tag{6}$$

where $\mathbf{x}_0^*$ is the global feature of the test image given by the pre-trained model.

We also employ ACA in Unic. This differs from that in vanilla VC-UFG and HC-UFG as the momentum prompts learned on HC-UFG tasks can be selected to adapt classifiers of VC-UFG tasks and the prompts for VC-UFG tasks are possible to help adapt the HC-UFG classifier. By doing so, we integrate our effective designs for VC-UFG and HC-UFG into a unified framework that is capable of learning continuously on a sequence of hybrid VC-UFG and HC-UFG tasks. On an arbitrary C-UFG task t, the overall training objective is given by

$$\mathcal{L}^t = \mathcal{L}_{ce}^t + \beta \mathcal{L}_{match}^t \tag{7}$$

where $\mathcal{L}_{match}^t$ shares the same fromula with Equation 1. $\mathcal{L}_{ce}^t$ is similar to Equation 2 except that f_t is replaced by f_s if a previous task s shares the same label space with task t and $\mathcal{Y}_s \cap \mathcal{Y}_i = \emptyset$ for $i < s$.

5 Experiment

Similar to [Jiao *et al.*, 2024], the overall performance of a model is assessed by two metrics, *i.e.* Final Average Accuracy (*FAA*) and Cumulative Average Accuracy (*CAA*). The

Method	SoyGene-C		SoyGlobal-C		SoyAgeing-C		UniUFG-C	
	FAA(↑)	*CAA*(↑)	*FAA*(↑)	*CAA*(↑)	*FAA*(↑)	*CAA*(↑)	*FAA*(↑)	*CAA*(↑)
Upper bound	58.58	-	36.65	-	69.15	-	64.61	-
L2P [Wang *et al.*, 2022c]	24.04	37.37	20.01	30.05	26.04	41.06	19.87	31.25
L2P++ [Wang *et al.*, 2022c]	33.00	38.57	22.24	33.07	28.14	36.70	36.96	42.11
DualPrompt [Wang *et al.*, 2022b]	31.97	38.57	20.83	29.38	31.00	41.70	24.42	34.94
S-Prompts [Wang *et al.*, 2022a]	24.82	33.78	15.84	25.95	30.67	42.88	36.76	44.33
CODA-P [Smith *et al.*, 2023]	25.60	38.97	14.63	26.03	29.80	43.51	24.25	35.60
LAE [Gao *et al.*, 2023]	20.02	34.40	11.87	18.03	28.32	39.78	19.01	25.42
HiDE [Wang *et al.*, 2023]	15.67	28.74	10.65	21.06	-	-	-	-
VQ-Prompt [Jiao *et al.*, 2024]	23.32	36.04	10.85	16.95	26.18	39.79	18.89	26.82
Unic-V / Unic-H / Unic	**37.02**	**43.45**	**25.53**	**35.77**	**32.67**	**44.69**	**39.16**	**46.35**

Table 1: C-UFG performance comparison between the proposed methods and previous continual learning methods.

former refers to the final average accuracy after learning all the tasks. The latter measures the average of historical *FAA* values after learning each task. We present experimental on the three C-UFG problems, *i.e.* VC-UFG, HC-UFG and unified C-UFG, to verify the effectiveness of the proposed methods. The model proposed for VC-UFG (Section 4.1) is denoted as Unic-V and that for HC-UFG (Section 4.2) is denoted as Unic-H.

5.1 Implementation Details

For the PTM, we use ViT-Base [Dosovitskiy *et al.*, 2021] pretrained on ImageNet-21K [Deng *et al.*, 2009] as in previous works. The pre-trained position embedding is resized using bicubic interpolation to match the input size. For the training on any UFG task, we resize the image to 440×440 and randomly crop a 384×384 patch followed by random horizontal flip for training. The model is trained for 160 epochs with an Adam optimizer. The learning rate is initialized as 0.03 and gradually decayed to 0.0003 using a cosine scheduler. For inference, we directly resize the images to 384×384. Following previous works, the total number of prompts at each prompted layer is 20 and the prompted layers are the first 5 transformer layers. The prompt length N is set to 1 and the number of prompts L is 20. The weights of ACA are set to $\lambda_v = 0.5$ and $\lambda_h = 0.3$. The weight of the matching loss is set to 0.5 for both β_v and β. And the momentum μ is set to 0.9999 for Unic-H.

5.2 Comparison with SOTA

VC-UFG. As in Table 1, we evaluate Unic-V and previous continual learning methods on both SoyGene-C and SoyGlobal-C datasets to verify their performance for VC-UFG. On SoyGene-C, it can be observed that both L2P++ and DualPrompt clearly surpass the other compared methods as discussed in Section 3.2. Compared with L2P++ and DualPrompt, the proposed Unic-V further boost the *FAA* and *CAA* by a large margin. On SoyGlobal-C, while L2P++ consistently outperforms the other compared methods, DualPrompt does not obtain a comparable performance. We hypothesize that the insufficient training samples limit the generalizability of the shared prompts of DualPrompt. Meanwhile, the proposed Unic-V consistently shows a superior performance. We also conduct prompt tuning on SoyGene and SoyGlobal to obtain the respective upper-bound models as in the first row of Table 1. Compared with the upper-bound model, the continual learning methods still fall behind by a large margin. This suggests there still exists a large room to enhance the VC-UFG performance.

HC-UFG To verify the effectiveness of the proposed Unic-H, we test Unic-H and previous methods on SoyAgeing-C as in Table 1. Compared with the continual learning methods, S-Prompts shows a superior result as it suits more for domain-incremental learning. CODA-P also obtains a comparable performance, in which we believe the attention-based prompt fusing plays a major role. Compared with those methods, our proposed Unic-H consistently improves the results by a clear margin, demonstrating the effectiveness of the momentum prompts. We also observe that the performance gap between the upper bound and the continual learning methods becomes larger on HC-UFG. While the upper bound naturally learns an unbiased classifier, the continual learning methods are significantly limited by the classifier bias.

Unified C-UFG To compare the effectiveness between the proposed Unic and previous continual learning methods, we slightly modify the compared methods to share the same classifier between HC-UFG tasks and apply the methods to UniUFG-C. As in Table 1, both L2P++ and S-Promtps obtain better results than the other compared methods. Compared with these two methods, the proposed Unic consistently shows superior final average accuracy. The cumulative average accuracy is only slightly inferior to S-Prompts. Compared with the upper-bound model, we also observe a large margin of the C-UFG performance. While respectively improving the VC-UFG and HC-UFG models can benefit the unified C-UFG model, the unified framework is also an important factor to further improve the overall performance.

5.3 Analytical Study

The effect of deep keys. To understand the effect of using deep keys, we test the proposed Unic-V on SoyGene-C as in Table 2. It can be observed that introducing deep keys improves both the final average accuracy and cumulative average accuracy. We also test to vary the configuration of deep

Model	Deep keys	SoyGene-C	
		FAA(↑)	*CAA(↑)*
Unic-V w/o ACA	-	33.95	40.56
Unic-V w/o ACA	✓	35.90	42.24
Unic-V	-	34.86	41.68
Unic-V	✓	**37.02**	**43.45**

Table 2: Comparison between Unic-V with and without the deep keys. We also test to remove ACA to make a more comprehensive understanding.

keys in Unic-V without ACA. The results of using deep keys are consistently shown to be superior on SoyGene-C. These overall experiments suggest that introducing deep keys effectively improves the model performance for VC-UFG.

Model	Dense keys	SoyGene-C	
		FAA(↑)	*CAA(↑)*
Unic-V w/o ACA	-	33.80	40.47
Unic-V w/o ACA	✓	35.90	42.24
Unic-V	-	35.50	42.53
Unic-V	✓	**37.02**	**43.45**

Table 3: Comparison between Unic-V with and without the dense keys. We also test to remove ACA to make a more comprehensive understanding.

The effect of dense keys. To understand the effect of dense keys, we conduct the experiments as in Table 3. For models without dense keys, we set the number of prompts L to 4 and the length of each prompt N to 5 which results in the same final length of prompts as in Unic-V with dense keys for fair comparisons. The results in Table 3 suggest that using dense keys clearly improves the performance of Unic-V for HC-UFG. Moreover, we observe that the model with dense keys consistently improves the results when removing ACA in model training, demonstrating the effectiveness of dense keys.

ACA	SoyGene-C		SoyAgeing-C	
	FAA(↑)	*CAA(↑)*	*FAA(↑)*	*CAA(↑)*
-	35.90	42.24	30.33	43.85
✓	**37.02**	**43.45**	**32.67**	**44.69**

Table 4: Comparison between models with and without ACA. We respectively test Unic-V and Unic-H on SoyGene-C and SoyAgeing-C datasets.

The effect of ACA. ACA is proposed to adapt the classifier to task-agnostic prompts for robust C-UFG. To understand the effect of ACA, we first test Unic-V and Unic-H with and without ACA on SoyGene-C and SoyAgeing, respectively. As in Table 4, the proposed ACA consistently improves the continual learning performances for VC-UFG and HC-UFG. In addition, we also test adaptive prompt adaptation (APA) and adaptive joint adaptation (AJA). The former fixes the classifier to optimize only the task-agnostic prompts, and the latter jointly optimizes the classifier and task-agnostic prompts. As in Table 5, neither AJA nor APA improves the C-UFG performance compared with the model without ACA (Table 4). Moreover, APA largely hinders overall performance as it changes the learned prompts.

Method	SoyGene-C		SoyAging-C	
	FAA(↑)	*CAA(↑)*	*FAA(↑)*	*CAA(↑)*
AJA	31.85	42.67	30.46	43.77
APA	28.67	39.01	26.51	42.74
ACA	**37.02**	**43.45**	**32.67**	**44.69**

Table 5: Comparison between models with different adaptive adaptation methods.

The value of momentum. We also test to vary the momentum factor in Unic-H and Unic to find the optimal value. As in Table 6, we test three momentum values on both SoyAGeing-C and UniUFG-C datasets. It can be observed that setting the momentum factor to 0.9999 consistently gives superior performance on the two challenges. We thus employ $\mu = 0.9999$ by default.

Momentum	SoyAgeing-C		UniUFG-C	
	FAA(↑)	*CAA(↑)*	*FAA(↑)*	*CAA(↑)*
0.999	30.08	43.74	36.65	45.07
0.9999	**32.67**	**44.69**	**39.16**	**46.35**
0.99999	31.49	43.94	37.83	44.73

Table 6: Model performance under different momentum factors. We respectively test Unic-H and Unic on SoyAgeing-C and UniUFG-C datasets.

6 Conclusion

To facilitate future development of AI techniques for agriculture in ultra-fine-grained cultivates recognition, this work establishes the first comprehensive C-UFG benchmark, which consists of three realistic C-UFG challenges, VC-UFG, HC-UFG and unified C-UFG. By analyzing existing PTM-based continual learning methods on the proposed C-UFG challenges, we discover their beneficial designs and the main issues in improving the C-UFG performance. Based on the preliminary analysis, we first propose two simple yet effective methods for tackling VC-UFG and HC-UFG problems, respectively. On top of those, we propose a unified C-UFG model Unic that integrates the two methods into a unified framework to tackle C-UFG problems where VC-UFG and HC-UFG co-exist in a single continual learning sequence. By quantitatively evaluating the proposed methods and existing continual learning methods, we demonstrate our proposed methods effectively boost the C-UFG performance on the proposed challenges.

Acknowledgements

This work is supported by the National Natural Science Foundation of China 62276016 and 62372029.

References

[Cha *et al.*, 2021] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 9516–9525, 2021.

[De Lange *et al.*, 2021] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Gao *et al.*, 2023] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11483–11493, 2023.

[He *et al.*, 2022] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 852–860, 2022.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[Hu *et al.*, 2022] 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.

[Jia *et al.*, 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*, pages 709–727. Springer, 2022.

[Jiao *et al.*, 2024] Li Jiao, Qiuxia Lai, YU LI, and Qiang Xu. Vector quantization prompting for continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Kirkpatrick *et al.*, 2017] 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.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.

[Liu *et al.*, 2022] Yang Liu, Lei Zhou, Pengcheng Zhang, Xiao Bai, Lin Gu, Xiaohan Yu, Jun Zhou, and Edwin R Hancock. Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In *European Conference on Computer Vision*, pages 57–73. Springer, 2022.

[Lopez-Paz and Ranzato, 2017] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[Ridnik *et al.*, 2021] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[Rolnick *et al.*, 2019] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

[Smith *et al.*, 2023] 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*, pages 11909–11919, 2023.

[Van Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pages 8769–8778, 2018.

[Wah *et al.*, 2011] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. In *California Institute of Technology*, 2011.

[Wang *et al.*, 2021] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. In *32nd British Machine Vision Conference 2021*, 2021.

[Wang *et al.*, 2022a] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022.

[Wang *et al.*, 2022b] 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*, pages 631–648. Springer, 2022.

[Wang *et al.*, 2022c] 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*, pages 139–149, 2022.

[Wang *et al.*, 2023] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, pages 69054–69076, 2023.

[Wang *et al.*, 2024] 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*, 2024.

[Yu *et al.*, 2020] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Patchy image structure classification using multi-orientation region transform. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12741–12748, 2020.

[Yu *et al.*, 2021a] Xiaohan Yu, Yang Zhao, Yongsheng Gao, and Shengwu Xiong. Maskcov: A random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition*, 119:108067, 2021.

[Yu *et al.*, 2021b] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Xiaohui Yuan, and Shengwu Xiong. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10285–10295, 2021.

[Yu *et al.*, 2022] Xiaohan Yu, Yang Zhao, and Yongsheng Gao. Spare: Self-supervised part erasing for ultra-fine-grained visual categorization. *Pattern Recognition*, 128:108691, 2022.

[Yu *et al.*, 2023a] Xiaohan Yu, Jun Wang, and Yongsheng Gao. Cle-vit: Contrastive learning encoded transformer for ultra-fine-grained visual categorization. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4531–4539, 2023.

[Yu *et al.*, 2023b] Xiaohan Yu, Jun Wang, Yang Zhao, and Yongsheng Gao. Mix-vit: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognition*, 135:109131, 2023.

[Zhang *et al.*, 2023] Pengcheng Zhang, Xiaohan Yu, Xiao Bai, Jin Zheng, Xiaoyu Wu, and Yongsheng Gao. Diving into continual ultra-fine-grained visual categorization. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 113–120. IEEE, 2023.