Open-World Dynamic Prompt and Continual Visual Representation Learning

Youngeun Kim[1,*†], Jun Fang[2,*‡], Qin Zhang[2], Zhaowei Cai[3],
Yantao Shen[2], Rahul Duggal[2], Dripta S. Raychaudhuri[2],
Zhuowen Tu[2], Yifan Xing[2], Onkar Dabeer[2]

[1] Yale University
[2] AWS AI Labs
[3] Amazon AGI

`youngeun.kim@yale.edu {junfa, qzaamz, zhaoweic, yantaos, dugrahul,`
`driptarc, ztu, yifax, onkardab}@amazon.com`

Abstract. The open world is inherently dynamic, characterized by ever-evolving concepts and distributions. Continual learning (CL) in this dynamic open-world environment presents a significant challenge in effectively generalizing to unseen test-time classes. To address this challenge, we introduce a new practical CL setting tailored for open-world visual representation learning. In this setting, subsequent data streams systematically introduce novel classes that are disjoint from those seen in previous training phases, while also remaining distinct from the unseen test classes. In response, we present **D**ynamic **P**rompt **a**nd **R**epresentation **L**earner (**DPaRL**), a simple yet effective Prompt-based CL (PCL) method. Our DPaRL learns to generate *dynamic* prompts for inference, as opposed to relying on a *static* prompt pool in previous PCL methods. In addition, DPaRL jointly learns dynamic prompt generation and discriminative representation at each training stage whereas prior PCL methods only refine the prompt learning throughout the process. Our experimental results demonstrate the superiority of our approach, surpassing state-of-the-art methods on well-established open-world image retrieval benchmarks by an average of 4.7% improvement in Recall@1 performance.

Keywords: Dynamic Prompt Generation · Continual Learning · Open-World Visual Representation Learning

1 Introduction

Continual learning (CL) without catastrophic forgetting is a challenging task in practice as retraining with old data is prohibitively expensive when data volume grows [10, 14, 23]. It is particularly challenging for real-world applications in the *open-world* [2, 5], where new concepts keep emerging and distribution shifts are

---

* Equal contribution.
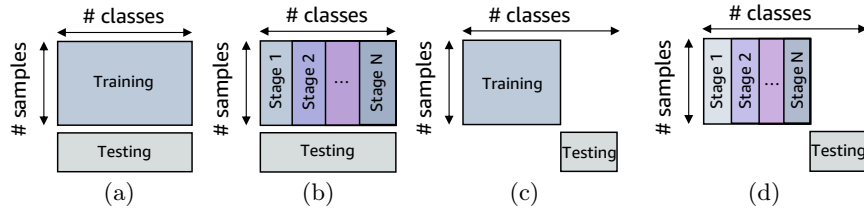† Work conducted during an internship at Amazon.
‡ Corresponding author.

Fig. 1: Illustration of the distinctions between our problem setting and traditional settings. In our work, we aim to address the problem where training splits have no class overlaps. (a) Closed-world setting: both training and testing classes are identical. (b) Continual learning (CL) in closed-world setting: training classes are split into multiple divisions and are introduced through various CL stages. (c) Open-world setting: training and testing classes remain separate. The aim is to learn a robust representation from training classes that generalize to unseen classes. (d) **Open-World Continual Representation Learning (Our Problem Setting)**: continual learning tailored for an open-world scenario. We sequentially introduce training classes over multiple CL stages, ensuring they remain distinct from test-time unseen classes.

inevitable. Recent studies on Prompt-based Continual Learning (PCL) [32, 38, 39] demonstrated success in mitigating catastrophic forgetting without access to any past data. However, they heavily focus on the *closed-world* setting where test-time categories come from those encountered in various training stages.

Whereas in practice, the world is dynamic and concepts evolve over time and new categories are often introduced after a model is trained. Consider, for instance, an image retrieval system for clothing brands in online e-commerce. During training, one would gather images of all existing clothing brands available to train a model. However, after deployment, the system encounters new brands. The new data collected for system upgrade, constituting a new training stage, will likely include brands that are disjoint to those in older stages. Furthermore, the system will constantly face new brands at test-time that are disjoint to any of those encountered in the previous training stages and is expected to generalize well to those new categories.

Thus, to comprehensively evaluate the ability of a model in retaining semantic information from continuous streams of data for generalization to new concepts at test-time for practical applications, we propose a new continual *open-world* representation learning setting, as illustrated in Figure 1(d). In this setting, different data streams comprise distinct classes, and test-time classes are disjoint to those encountered in all training phases. Our setting is notably more challenging than the previous closed-world CL scenario. It closely mirrors real-world demands on recognition systems, where, during any evaluation stage, the test set consists of unseen entities to the training process up to that point. This setup reflects the need for practical recognition systems to encode general semantic information from training-time classes and apply this knowledge to new, unseen concepts without access to older training data during system updates.

We start with evaluating state-of-the-art PCL methods [32, 38, 39] in our proposed setting and find that they fail to generalize with the absence of past data. For example, in our experiments, the gap between the paragon of standard training with all data in one phase and the best PCL baseline (Coda) shows up to 20% gap in Recall@1 (Table 1), significantly larger than the 5% gap in the closed-world setting, which suggests that prior PCL methods encounter difficulties and are less effective in the open-world setting.

To tackle this challenging yet practical open-world CL setting, we introduce a simple yet effective method called Dynamic Prompt and Representation Learner (DPaRL). Unlike previous prompt pool based PCL methods [32, 38, 39] that depend on a static prompt pool (Figure 2(a)) for test-time predictions. Notably, DPaRL trains an innovative dynamic prompt generation network jointly with the discriminative representation backbone model (Figure 2(b)), departing from the prompt-only learning paradigm in prior PCL methods.

Our key insight lies in that simply combining the fixed prompts learnt over training classes does not generalize well to unseen test-time classes in the practical open-world setting. On the other hand, the on-the-fly dynamically generated prompts in our method exhibit a stronger capability of capturing the diverse semantics spread across the numerous divisions of classes observed during the different training stages of continual learning, improving generalization to test-time unseen classes. This is achieved through the proposed *dynamic* prompt generation network, which possesses stronger representation power with a learnable mapping function compared to the traditional approach of combining multiple prompts in a *static* prompt pool. Additionally, unlike prior PCL methods, DPaRL jointly updates the dynamic prompt generation process and the discriminative representation backbone weights, facilitating maximized interaction and comparison between old concepts and newer ones to encapsulate diverse semantics that aid generalization to unseen test-time classes.

Consequently, over the four open-world image retrieval benchmarks on cars, online catalogs, products, and species images where the task is to search for the most similar images belonging to the same class as the query, via a distance measurement in an embedding space, DPaRL achieves significant accuracy boosts. It improves the average Recall@1 performance by absolute 4.7% and 9.1% compared to state-of-the-art continual learning methods, from 76.1% to 80.8% over 10 CL stages and from 68.0% to 77.1% over 100 CL stages, respectively.

In summary, our contributions are three folds:

1. We establish a new practical setting on continual visual representation learning in the open-world.
2. We introduce a simple yet powerful method, Dynamic Prompt and Representation Learner (DPaRL), which dynamically generates prompts while effectively updating the discriminative representation backbone. This enhancement improves generalization for unseen open-world classes at test time.
3. We outperform state-of-the-art continual learning methods in the proposed practical setting, both rehearsal-free and rehearsal-based.
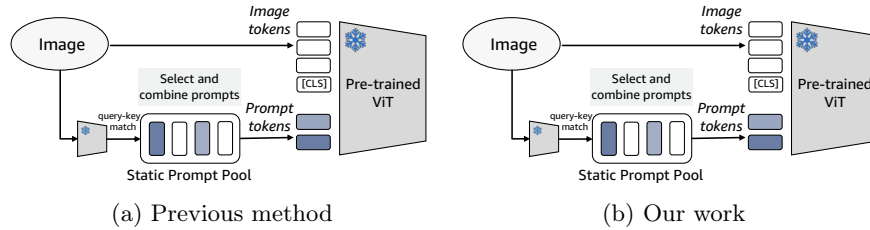
(a) Previous method   (b) Our work

Fig. 2: Prior PCL methods combine prompts from a *static* prompt pool trained on the training class distribution, leading to a loss of generalization capability when facing unseen classes during test time. Our work introduces a Dynamic Prompt Generation network that generates *dynamic* prompts on the fly by integrating a given image with stage tokens, followed by a specialized mapping function and adjustable discriminative representation backbone weights, providing generalizable prompts for unseen testing classes, distinguishing it from prior PCL methods.

2   Related Work

Continual learning. Continual learning (CL) has been recognized for decades as a crucial research domain, focusing on the ongoing adaptation of models over time. Early works primarily centered around *regularization-based* methods in an effort to address catastrophic forgetting [1, 21, 43] by introducing regularization between its current and previously learned parameters. However, they often underperform when faced with complex datasets.

A separate line of work, *architecture-based* approach, involves expanding the network architecture in response to new learning phases [20, 26, 29, 30, 42]. While these methods generally outperform regularization-based counterparts, they do so at the cost of introducing additional parameters. Recent advancements include *rehearsal-based* methods that incorporate a memory component for storing past data [4, 6, 7, 15, 28]. This archived data aids in training newer stages, typically yielding superior performance compared to other CL techniques. Yet, this data retention can introduce privacy concerns and significant memory overhead.

Given these constraints, *rehearsal-free* strategies are gaining attention in the community. These aim to mitigate catastrophic forgetting without relying on historical data. A prominent method within this category uses model inversion to produce rehearsal images [9, 13, 31, 41]. However, model inversion is computationally intensive and time-consuming. Additionally, these methods lag significantly in performance when compared to their rehearsal-based counterparts.

Recent developments have witnessed the rise of *prompt-based* approaches [27, 32, 34, 37–39, 45]. These strategies offer robust protection against catastrophic forgetting by learning a small set of parameters, or prompts, instead of directly training all parameters. For instance, L2P [39] employs a pool of prompts, selecting suitable ones for insertion based on input data clustering. Building on this concept, DualPrompt [38] presents general prompts designed to encapsulate shared knowledge across various learning stages. Advancing this further,

the state-of-the-art method CodaPrompt [32] facilitates end-to-end training by incorporating an attention mechanism for prompt combination within a pool. However, during inference, the above works rely on the combination of prompts from a static pool learned from training to feed into a frozen backbone model, which has limitations in cases where testing data has distribution shifts or disjoint classes from the training data. This motivates us to design a Dynamic Prompt and Representation Learner (DPaRL) to address these drawbacks.

Open-World Visual Representation Learning. In open-world visual representation learning, the model learns discriminative representations that align distances between representations with their semantic similarities, and uses these knowledge for testing on unseen classes. Previous works in this domain have primarily focused on designing loss functions [11, 17, 24] for enhanced accuracy. Recently, several studies [2, 12] have demonstrated the effectiveness of utilizing pre-trained visual foundation models and fine-tuning them to achieve significant performance improvements. While open-world visual representation learning has been under investigation for an extended period, there has been limited exploration within the continual learning (CL) framework for this task. [44] investigated biometric datasets using a regularization-based CL approach. However, the investigation into state-of-the-art prompt-based CL methods remains missing. Furthermore, as their studies primarily centered on biometric data, there is a pressing need to explore established continual learning benchmarks specific to open-world visual representation learning in more natural settings.

3   Methodology

In this section, we introduce our open-world Continual Learning (CL) setting and motivate our Dynamic Prompt and Representation Learner (DPaRL).

3.1   Problem Setting

Two fundamental visual representation learning settings are the *closed-world* and *open-world* paradigms. In a *closed-world* setting (as depicted in Figure 1(a)), the training and testing data classes are entirely identical. Conversely, the *open-world* setting (illustrated in Figure 1(c)) presents a more practical scenario. Here, training and testing classes are entirely distinct, thus requiring the model to learn representations that generalize to unseen concepts.

When faced with continuous streams of data for visual recognition system updates, built upon the *closed-world* paradigm, Figure 1(b) showcases the well-known continual learning setting for image recognition. In this work, we introduce a new practical setting: *Open-World Continual Representation Learning* as shown in Figure 1(d). Instead of testing on seen classes of the various training stages thus far, the testing classes are entirely disjoint to those faced in different training stages, necessitating the model to not only learn continually but also generalize to new concepts, making it a more dynamic and challenging setting.
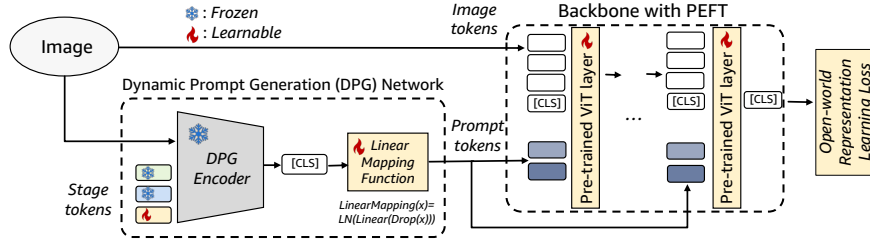
Fig. 3: Illustration of our proposed method: Dynamic Prompt and Representation Leaner (DPaRL). We dynamically generate prompt tokens on the fly from Dynamic Prompt Generation (DPG) networks by integrating information from stage tokens and image tokens. The [CLS] token from DPG is converted to prompt tokens via a linear mapping function. The generated prompt tokens are added to the backbone ViT and trained with a loss function. The learnable parameters include the current stage token, weights in mapping function, weights in backbone, and weights in loss function.

3.2   Prompt-based Continual Learning Paradigm

Prompt-based Continual Learning (PCL) methods [32, 38, 39] use a pre-trained Vision Transformer (ViT) as a discriminative backbone for closed-world image classification, as shown in the Figure 2(a). These methods create a prompt pool with multiple prompt tokens, updating only the learnable parameters in this pool during training. In inference, the learned prompt pool is *static*, and PCL methods select tokens from this pool to feed into multiple ViT backbone layers for predictions. Our method adopts this PCL paradigm but adapts to the open-world representation learning setting with a prototypical metric learning loss. Additionally, we introduce a *dynamic* prompt generation (DPG) network to replace the static prompt pool for effective joint learning with the discriminative representation backbone model, as detailed below.

3.3   Dynamic Prompt and Representation Leaner (DPaRL)

Overall Pipeline. In the open-world setting where training and testing classes are disjoint, existing prompt pool designs in PCL [32, 38, 39] exhibit limitations, i.e., limited separation between intra- and inter- testing class distance distributions as shown in Figure 4 and large performance gap from paragon in Table 1. To mitigate, we first propose dynamically generating prompts by blending the given image information and continual learning stage's information through a dynamic prompt generation (DPG) network. It aims to provide better representation power compared to the prompt pool design, which combines a fixed prompt pool trained from the training class distribution. Moreover, we enhance the capability of the discriminative representation backbone model to more effectively generalize to the open-world concepts through a joint dynamic prompt and representation learning paradigm. We refer to Figure 2, 3 for detailed design differences between our approach and prior prompt pool-based methods.

Dynamic Prompt Generation Network. In order to retain information from each CL stage of our Dynamic Prompt Generation (DPG) process, we introduce stage tokens, denoted as S. During training stage t, we train a stage token S_t while freezing the previous stage tokens $S_{t-(q-1)} \sim S_{t-1}$ in a first-in-first-out (FIFO) queue manner, where q is the max queue size. Guided by the ablation study detailed in Section 4.4, we typically set $q = 5$ using a FIFO order to aim for both high accuracy and scalability. This ensures that the knowledge from the previous stages remains unchanged, and the total number of stage tokens limits by the queue size without linearly scaling with stage numbers. Within the deep neural network such as ViT, the self-attention block facilitates the integration of information between the stage-wise tokens and the instance-wise image tokens.

As illustrated in Figure 3, the DPG produces a [CLS] token to dynamically acquire task-specific high-level information for prior knowledge of class-level features. However, it differs in size from the prompt tokens required by the discriminative backbone model (right side of Figure 3). To address this, we introduce a mapping function between the [CLS] token and the prompt tokens.

A straightforward approach is to employ a single linear layer for dimensionality transformation. However, such a method introduces an excessive number of additional parameters, and the resulting over-parameterization can lead to overfitting of the highly compressed [CLS] token information. To tackle this issue, we impose a constraint on the weight parameter $W = AB^T \in \mathbb{R}^{C_{in} \times C_{out}}$, where $A \in \mathbb{R}^{C_{in} \times R}$, $B \in \mathbb{R}^{C_{out} \times R}$, ensuring a maximum rank of $R < min(C_{in}, C_{out})$. This constraint is inspired by the success of the LoRA technique [16]. The experimental results in Section 4.4 illustrate the critical importance of the low-rank mapping function design for our DPG module. Moreover, we apply Dropout [33] and LayerNorm [3] to this low-rank linear mapping function

$$Mapping(x) = LayerNorm(LowRankLinear(DropOut(x))), \qquad (1)$$

which further helps to avoid overfitting and stabilize training.

With the help of this specialized mapping function along with a pre-trained neural network, the prompt P is dynamically obtained from the input image I and stage tokens $S_{t,q} := S_{t-(q-1)} \sim S_t$:

$$P = Mapping(Network([S_{t,q};\ I])). \qquad (2)$$

We reshape P to a size of $N_p \times C \times L$, where N_p is the number of prompts, C is the channel dimension, and L indicates the number of layers where prompts are applied to the backbone model. The generated prompts are inserted across the first L layers in the ViT backbone. In our DPG approach to prompting, we follow the technical foundations of prompting from the prior state-of-the-art PCL methods [32,38]. This is because we want to make fair comparison with the previous work, and *our primary focus is on the formation of prompt*. The prompt for the l-th layer, denoted as $P_l \in \mathbb{R}^{N \times C}$, is partitioned into $\{P_{l,k}, P_{l,v}\} \in \mathbb{R}^{\frac{N}{2} \times C}$ and is prefixed to the input token embedding for both *key* and *value* in the attention mechanism [36]. This can be formulated as:

$$h_i = Attention(X_l W_q^i,\ [P_k; X] W_k^i,\ [P_v; X] W_v^i). \qquad (3)$$

Here, we omit layer index l for brevity, and i is head index in multi-head self-attention of transformer architecture.

Joint Dynamic Prompt and Representation Learning. Differing from prior PCL methods [32,38,39] that freeze the backbone model (on the right side of the pipeline in Figure 2(a)) during training, we propose our Dynamic Prompt and Representation Learner (DPaRL), to jointly learn the *dynamic prompt generation* with the discriminative representation learning *backbone* (on the right side of Figure 2(b) and Figure 3). This approach aims to maximize the capability of entire pipeline to integrate old-stage concepts and new ones to encapsulate diverse semantics that aid generalization to open-world unseen classes and unknown domain shifts. Leveraging parameter-efficient fine-tuning techniques, our DPaRL successfully maximizes accuracy performance and minimizes catastrophic forgetting simultaneously. The overall framework is depicted in Figure 3. Note that the DPG *encoder* weights (on the left side of the pipeline) for prompt generation are *frozen* in our method, while the *learnable* parameters include stage tokens, weights in the mapping function and representation loss function, as well as the discriminative representation backbone weights.

4   Experiments

In this section, we empirically evaluate the effectiveness of our Dynamic Prompt and Representation Learner (DPaRL) in the established challenging continual open-world visual representation learning.

4.1   Datasets & Evaluation Metric

Datasets. We conducted our training and testing across 4 prominent open-world image retrieval benchmarks: Cars [19], In-Shop [22], SOP [25] and iNat2018 [35]. We evaluate baselines and our proposed method following our practical continual open-world visual representation learning setting. In this dynamic scenario, classes are introduced incrementally at each learning stage, and the testing phase encompasses classes that have not been encountered during training. Our experiments consist of 10 or 100 stages, with the total training classes approximately distributed uniformly across these stages. The detailed data information including number of classes and samples are provided in the *Supplementary Material*. For inference, we use the entire testing dataset across all CL stages since there is no overlap between training and testing classes. Thus, our objective is to learn class relations while maintaining generalization across CL stages.
Evaluation Metric. To evaluate model performance, we compute the Recall@1 metric using the [CLS] token at the last layer of the discriminative representation backbone network. Given our goal to optimize the model's adaptability throughout the CL phases, we report average Recall@1 across all CL stage. Let R_n be the Recall@1 performance at stage n, then we report $R_N = \frac{1}{N_s} \sum_{n=1}^{N_s} R_n$ where N_s

is the total number of stages. We also report the forgetting score F_N by following [32,38] to measure the performance drop on *the same unseen testing dataset* across different training stages, that is, $F_N = \frac{1}{N_s-1} \sum_{n=2}^{N_s} \max_{m \in \{1,...,n-1\}} (0, R_m - R_n)$. We stress that R_N holds greater importance as it captures both learning capacity and forgetting, while F_N merely offers supplementary context.

4.2   Baselines and Implementations

Baselines. We validate 3 types of representative baselines:

- Rehearsal-based: Experience Replay (ER) [7],
- Regularization-based: Learning without Forgetting (LwF) [21]
- Rehearsal-free prompt-based: Learning to Prompt (L2P) [39], DualPrompt [38], and CodaPrompt [32].

Implementations Details. Consistently across all methods, we adhere to the training setups by following [32]. Specifically, we utilize the Adam optimizer [18] with a batch size of 128 to train for 20 epochs to ensure comprehensive model convergence. The learning rate is cosine decaying with an initial value of 0.001.

We utilize ArcFace [11] as the training loss function as it is a prevalent choice in open-world representation learning to achieve state-of-the-art performance [2]. For new-stage training, the prototype weights in the loss function of old training stages are appended to the new training stage as initialization. All prototype weights in the loss function, including those corresponding to the classes for old and new stages, are trainable during each training stage.

For established PCL techniques, we maintain the prompt lengths and locations for L2P [39], DualPrompt [38], and CodaPrompt [32] as suggested in these works. For our DPaRL method, in each CL stage, we incorporate a learnable stage token with the first-in-first-out (FIFO) order and a max size of 5 tokens based on insights from Section 4.4. We apply our specialized mapping function in the DPG module to avoid over-fitting and stabilize training. We adopt LoRA [16] as the PEFT method to update backbone weights as guided by Section 4.5.

4.3   Performance Comparison with Prior Arts

Table 1 shows a comprehensive performance comparison with various CL methods across all four open-world image retrieval tasks. We adopt the official codebase of CodaPrompt [32] for the implementation of PCL baselines including L2P [39], DualPrompt [38], and CodaPrompt [32]. Additionally, we report the performance of representative rehearsal-based and regularization-based approaches, *i.e.*, ER [7] and LWF [21], respectively, for a more comprehensive comparison. Note that, for a fair comparison, all methods use the pre-trained ViT-Base on ImageNet-21k as the backbone model, as shown on the right side of Figure 3.

From the results in the Table 1, we observe that the performance trend of PCL-based methods (L2P, DualPrompt, CodaPrompt, our DPaRL) is similar to what was observed in the closed-world image recognition task [32], where PCL

Table 1: Performance on 4 open-world image retrieval benchmarks with 10 CL stages. We report the averaged *Recall@1* (R_N) and *Forgetting Score* (F_N) across all continual learning stages, the results are averaged over 3 runs. We present the *LowerBound* with the zero-shot performance of ImageNet-21k pre-trained ViT-B/16 backbone. *Upper-Bound* denotes the upper bound performance when trained on the entire dataset at a single stage. All PCL methods, including L2P, DualPrompt, CodaPrompt, and our DPaRL, utilize the same pre-trained encoder, ViT-B/16 on ImageNet-21k, to generate the prompt tokens. The numbers in brackets show the *Recall@1 improvement* of our DPaRL over the previous best baselines, with the previous best baseline underlined.

Dataset	Cars [19]		In-Shop [22]		SOP [25]		iNat2018 [35]		Average	
Metric	$R_N(\uparrow)$	$F_N(\downarrow)$	$R_N(\uparrow)$	$F_N(\downarrow)$	$R_N(\uparrow)$	$F_N(\downarrow)$	$R_N(\uparrow)$	$F_N(\downarrow)$	$R_N(\uparrow)$	$F_N(\downarrow)$
LowerBound	53.86	–	47.16	–	62.85	–	71.74	–	58.90	–
UpperBound	85.96	–	89.28	–	85.21	–	82.68	–	85.78	–
ER [7]	65.38	0.49	82.99	2.48	65.38	0.50	36.94	18.58	62.67	5.51
LwF [21]	56.53	2.91	81.52	3.39	66.18	13.20	39.69	33.81	60.98	13.33
L2P [39]	56.27	0.00	67.85	0.00	76.88	0.40	78.95	0.00	69.99	0.10
DualPrompt [38]	55.54	0.00	65.56	0.00	77.04	0.66	78.71	0.03	69.21	0.17
CodaPrompt [32]	65.23	0.36	78.61	0.00	81.62	0.01	78.97	0.01	76.11	0.10
DPaRL (Ours)	73.22(↑7.8)	0.03	86.28(↑3.3)	0.02	83.69(↑2.1)	0.07	80.02(↑1.0)	0.03	80.80(↑4.7)	0.09

methods outperform other continual learning approaches by a significant margin on both recall@1 metric R_N and forgetting score F_N[4], aligning with findings from previous research [32, 38, 39]. This suggests that PCL methods are strong baselines for continual learning on both closed-world image recognition tasks and open-world visual representation learning tasks. Notably, our DPaRL approach consistently outperforms other methods across all datasets, resulting in an average Recall@1 boost of 4.7% over the leading previous method, CodaPrompt [32]. This supports our hypothesis about the limitations of the conventional PCL design in open-world scenarios.

In the following, we provide step-by-step guidance on the design choices of our DPaRL. Additional analysis and results are presented in the supplementary material, covering the superior performance of DPaRL in few-shot settings, effectiveness of DPaRL in closed-world benchmarks, dynamic changes in the accuracy curve across CL stages, and robustness of DPaRL against various DPG encoders. Due to space limitations, we refer to the supplementary material for a more comprehensive exploration of these aspects.

4.4   Effectiveness of Dynamic Prompt Generation

To solely investigate the effectiveness of our Dynamic Prompt Generation design, we freeze the backbone weights similar to prior PCL methods [32, 38, 39] in our DPaRL and denoted as the DPG method.

---

[4] PCL methods consistently achieve forgetting scores close to zero, indicating that performance continually improves. This trend is also evident in the dynamic change of accuracy curve in the supplementary material, highlighting that PCL methods serve as robust baselines for addressing catastrophic forgetting in open-world tasks.
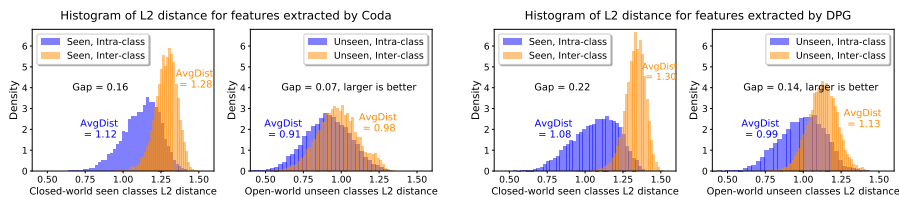
Fig. 4: Histogram of L2 distance between a pair of samples with embedding features extracted by Coda [32] (SOTA PCL method) and our DPG (the naive version of DPaRL with freezing backbone weights). Left and right figures are distributions of the seen and unseen classes in training and testing Cars [19] dataset, respectively. Our DPG exhibits enhanced separation between inter- and intra-classes, particularly on open-world unseen test classes.

We visualize the distributions of L2 distance between pairwise samples within intra- and inter-classes of CodaPrompt (Coda) [32] and our DPG in Figure 4 to assess generalizability. A good model should *exhibit distinguishable distributions between intra- and inter-classes for both closed-world seen classes and open-world unseen classes.* Coda, employing a static prompt pool design, achieves a notable separation on closed-world intra- and inter-classes but exhibits a close distribution for open-world intra- and inter-classes, impeding effective generalization to unseen test classes. In contrast, our DPG shows a more distinct separation, with an average gap of 0.14 (twice as large as Coda's gap) between inter- and intra-class distances in testing data. This indicates better generalizability with a dynamic prompt generation mechanism for unseen classes.

Moreover, accuracy results with Recall@1 in Table 2 and Table 4 reveal that our DPG significantly outperforms the best baseline (Coda), with an average of 9.1% boost in 100 CL stages and 3.2% boost in 10 CL stages, respectively. This highlights the superiority of Dynamic Prompting Generation design, particularly in practical scenarios involving a large number of stages. In the next Section 4.4.1 and Section 4.4.2, we investigate the importance of our design components of stage tokens and mapping function in the DPG network.

4.4.1 Effectiveness of Stage Tokens

In our DPaRL design with Dynamic Prompt Generation (DPG), we introduce stage tokens to retain information on old training stages, and we impose a max limit on the number of stage tokens ($M_{stg\_tok} = 5$) to control storage and computation costs as the number of stages scales up, e.g., $N_{stg} = 100$. Hence, when the stage number goes beyond the max limit M_{stg_tok}, we replace some of previous stage tokens. *Our hypothesis is that the earlier stage information learned from previous stage tokens can propagate into the later ones.* In this section, we conduct ablations in a practical CL setting with a large number of stages (100 stages) to study the impact of stage tokens.

Table 2: Impact of stage tokens with a large number of CL stages, $N_{stg} = 100$. Recall@1 results on In-Shop, SOP, and iNat2018 datasets are reported. It's worth noting that the Cars dataset is relatively small, preventing its division into 100 stages. DPG is a naive version of our DPaRL method by freezing the backbone weights.

Method	# Stage Token	Order	In-Shop	SOP	iNat	Average
L2P [39]	N/A	N/A	48.39	70.39	76.82	65.20
Dual [38]	N/A	N/A	52.50	73.76	76.91	67.72
Coda [32]	N/A	N/A	52.96	73.92	76.98	67.95
DPG	None	N/A	71.18	80.04	77.57	76.26 (↑8.31)
DPG	5	**FIFO (default)**	72.05	80.81	78.29	77.05 (↑9.10)
DPG	5	FILO	71.76	80.60	78.12	76.83 (↑8.88)
DPG	5	Random	72.00	80.83	78.20	77.01 (↑9.06)
DPG	10	FIFO	72.15	80.81	78.28	77.08 (↑9.13)
DPG	100	full-size baseline	72.20	80.84	78.37	77.14 (↑9.19)

With and without stage token. From Table 2, DPG with 5 stage tokens achieves a noticeable accuracy improvement over DPG without stage tokens, 77.05 vs. 76.26 on average, indicating the importance of stage token for retaining stage-wise class information in old training stages.

Impact of the order. We conduct ablations on three different orders: first-in-first-out (*FIFO*), first-in-last-out (*FILO*), and randomly replacing one stage token (*Random*). Table 2 reveals that *FIFO* performs the best on average. *FILO* achieves slightly lower accuracy. Nevertheless, the *Random* method achieves performance very close to the best, indicating the robustness of our DPG design to different orders of stage tokens. We hypothesize this due to stage information learned from previous ones effectively propagating to the new stage tokens, regardless of the order or the location where the new stage token is inserted.

Selection of maximum limit M_{stg_tok}. When the stage number is large ($N_{stg} = 100$), keeping a maximum token number of 5 achieves comparable performances (77.05 vs. 77.14 on average) to that of the full token number baseline with $M_{stg\_tok} = 100$. This signifies a highly effective selection with a constant number of stage tokens to achieve high accuracy, as well as good scalability with respect to the stage number. Hence, we use $M_{stg\_tok} = 5$ throughout this paper.

4.4.2 Effectiveness of Mapping Function

In our DPaRL framework (Figure 3), as elaborated in Section 3.3, we introduce a mapping function to match the size of [CLS] token and prompt tokens. We apply rank-R constraint to avoid overfitting and reduce the memory consumption from the fully-connected layer inside the mapping function. To measure the sensitivity of parameter R, we conduct experiments with varying R from 64 to 768.

The results for both In-Shop and iNat in Table 3 indicate the existence of an optimal rank with $R_{opt} = 256$ that strikes a balance between accuracy and the number of parameters, saving $2/3$ of the total parameters compared to the full rank case. This default rank $R_{opt} = 256$ is subsequently used for all experiments.

Note that when the rank reaches 768, representing the full rank of the mapping function essentially serving as a fully-connected layer, a significant per-

Table 3: Impact of the rank of weight matrix inside the mapping function on our DPG design. Recall@1 results on In-Shop and iNat2018 with 10 CL stages are reported.

Rank	64	128	256	384	512	768
In-Shop	82.33	83.17	**84.09**	83.86	82.98	71.40
iNat2018	79.21	79.85	**80.02**	79.74	79.63	78.98
#Parameters	2.0M	3.9M	7.9M	12.0M	16.1M	23.5M

formance drop of 12.69% is observed for the In-Shop dataset in Table 3. This suggests that an overparameterized expanding layer results in a severe accuracy degradation for a highly compressed [CLS] token, indicating the importance of our design. Moreover, our design has a robust rank range from 128 to 384, which closely matches the accuracy performance of the peak result.

4.5 Effectiveness of Dynamic Prompt and Representation Learner

Previous PCL methods [32,38,39] freeze backbone weights of the discriminative model to reduce catastrophic forgetting. However, this approach could limit accuracy performance as it doesn't provide the degree of freedom to update the backbone, particularly in open-world setting where data distribution and domain shifts naturally exist. In this section, we explore effective ways to update backbone weights jointly with our dynamic prompt learning. We explore 3 approaches here: 1) fine-tuning (FT) the entire backbone weights, 2) parameter-efficient fine-tuning (PEFT) with AdaptFromer [8] (*Adapter*) or Low-Rank Adaptation [16] weights on the backbone (*LoRA*), and 3) joint prompt generation and representation learning on the backbone weights (*Coda* and *DPaRL, w/ X*).

Table 4 shows that vanilla FT or fine-tuning with PCL methods not only leads to lower accuracy but also suffers from higher catastrophic forgetting compared to *DPG*. While, *LoRA* gets higher accuracy (+10.4% on average) and reduces forgetting (-2.3% on average) compared to vanilla FT, as well as outperforms *Adapter*, indicating a more effective way to update the backbone. Still, *LoRA* is less effective than *DPG*, with an average accuracy gap of 1.7% and a forgetting gap of 0.6%, indicating the superiority of our DPG design for open-world CL problems. Lastly, our joint Dynamic Prompt and Representation Learner (DPaRL) with LoRA improves average accuracy by 3.2% and reduces forgetting by 0.5% compared to LoRA-only baseline. It also boosts accuracy by 1.4% on average compare to DPG method without updating backbone weights at a negligible cost of 0.09 forgetting raises, indicating the effectiveness of our DPaRL design to maximizing accuracy without causing catastrophic forgetting issues. Therefore, unless stated otherwise, we adopt joint learning with dynamic prompt generation with LoRA as our default DPaRL method in this paper.

4.6 Performance on Closed-World Setting in Other Domains

In this section, we assess the effectiveness of our DPaRL method in domains other than the established image retrieval benchmarks such as medical image

Table 4: Accuracy performance (R_N,↑) and forgetting scores (F_N,↓) in 10 CL stages with different fine-tuning methods: fine-tuning (FT) the entire backbone weights, PEFT with prompt tuning (Coda and our DPG), Adapter [8], and LoRA [16].

Dataset	Cars		In-Shop		SOP		iNat2018		Average	
Metric	R_N(↑)	F_N(↓)	R_N(↑)	F_N(↓)	R_N(↑)	F_N(↓)	R_N(↑)	F_N(↓)	R_N(↑)	F_N(↓)
Coda [32], freeze(frz) backbone	65.23	0.36	78.61	0.00	81.62	0.01	78.97	0.01	76.11	0.10
DPG (Ours), frz backbone	70.62	**0.00**	84.09	**0.00**	82.69	**0.00**	80.02	**0.01**	79.36	**0.00**
FineTune [32]	57.57	5.04	82.23	0.54	73.45	2.23	55.69	3.84	67.24	2.91
Adapter [8]	64.59	2.47	83.94	0.00	82.18	0.06	77.27	0.31	76.99	0.71
LoRA [16]	67.77	0.99	83.29	0.40	81.97	0.27	77.47	0.79	77.63	0.61
Coda, w/ FT	50.03	12.79	82.29	0.45	74.03	3.29	57.42	4.39	65.94	5.23
Coda, w/ Adapter	64.39	0.86	83.56	0.45	81.98	0.05	78.56	0.06	77.12	0.36
Coda, w/ LoRA	69.06	1.23	82.79	0.33	82.51	0.25	79.12	0.34	78.37	0.54
DPaRL, w/ FT	65.00	4.61	82.36	0.33	74.24	2.71	56.73	4.12	69.58	2.94
DPaRL, w/ Adapter	68.47	1.00	85.72	0.01	82.57	0.01	78.04	0.03	78.70	0.26
DPaRL, w/ LoRA (default)	**73.22**	0.03	**86.28**	0.02	**83.69**	0.07	**80.02**	0.22	**80.80**	0.09

domain. However, there are no available open-world benchmarks so we follow the standard *closed-world* setting in Figure 1 (b). We conduct experiments on MedMNIST [40], which consists of approximately 700K biomedical images. It includes 7 datasets, each with 7∼11 classes. We evenly split the classes into 3 CL stages and report the average top-1 accuracy over each stage. Results in Table 5 demonstrate that DPaRL generalizes well to the medical domain, outperforming other methods with an average improvement of 6.50%. For additional tasks in other domains, please refer to the supplementary material.

Table 5: Top-1 (%) on medical data MedMNIST [40]. Our DPaRL surpasses other PCL methods by 6.50% on average.

Methods	Blood	Derma	OrganA	OrganC	OrganS	Path	Tissue	Average
L2P	56.03	34.00	61.12	52.87	46.25	56.39	41.29	49.71
Dual	66.28	33.16	65.42	57.79	48.22	60.69	40.34	53.13
Coda	57.37	33.72	65.41	65.98	46.82	79.85	32.30	<u>54.49</u>
DPaRL (Ours)	**70.49**	**38.71**	**70.39**	**71.29**	**51.23**	**81.44**	**43.38**	**60.99(↑6.50)**

5   Conclusion

In this paper, we introduce a practical setting for continual visual representation learning in the open-world. Existing continual learning methods fail to generalize well in this proposed setting. To mitigate this challenge, we propose Dynamic Prompt and Representation Learner (DPaRL), to enhance the adaptability and performance in continual learning for open-world visual recognition. Our method consistently demonstrates superior performance across various open-world image retrieval datasets and settings, by dynamically generating prompts from a deep neural network coupled with a dedicated mapping function, and effectively updating the discriminative representation backbone. Notably, it overcomes the intrinsic constraints found in static prompt pool designs, thereby expanding the limits of model generalization in open-world scenarios.

References

1. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Proceedings of the European conference on computer vision (ECCV). pp. 139–154 (2018)
2. An, X., Deng, J., Yang, K., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Unicom: Universal and compact representation learning for image retrieval. arXiv preprint arXiv:2304.05884 (2023)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
4. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems **33**, 15920–15930 (2020)
5. Cao, T., Wang, Y., Xing, Y., Xiao, T., He, T., Zhang, Z., Zhou, H., Tighe, J.: Pss: Progressive sample selection for open-world visual representation learning. In: European Conference on Computer Vision. pp. 278–294. Springer (2022)
6. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. arXiv preprint arXiv:1812.00420 (2018)
7. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H., Ranzato, M.: On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019)
8. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. arXiv preprint arXiv:2205.13535 (2022)
9. Choi, Y., El-Khamy, M., Lee, J.: Dual-teacher class-incremental learning with data-free generative replay. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3543–3552 (2021)
10. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3366–3385 (2021)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Gao, Q., Zhao, C., Ghanem, B., Zhang, J.: R-dfcil: Relation-guided representation learning for data-free class incremental learning. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
14. Hadsell, R., Rao, D., Rusu, A.A., Pascanu, R.: Embracing change: Continual learning in deep neural networks. Trends in cognitive sciences **24**(12), 1028–1040 (2020)
15. Hayes, T.L., Cahill, N.D., Kanan, C.: Memory efficient experience replay for streaming learning. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 9769–9776. IEEE (2019)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

17. Kim, S., Kim, D., Cho, M., Kwak, S.: Proxy anchor loss for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3238–3247 (2020)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
20. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: International Conference on Machine Learning. pp. 3925–3934. PMLR (2019)
21. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
22. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016)
23. Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., Sanner, S.: Online continual learning in image classification: An empirical survey. Neurocomputing **469**, 28–51 (2022)
24. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE international conference on computer vision. pp. 360–368 (2017)
25. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4004–4012 (2016)
26. Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y.W., Hadsell, R.: Continual unsupervised representation learning. Advances in neural information processing systems **32** (2019)
27. Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., Almahairi, A.: Progressive prompts: Continual learning for language models. arXiv preprint arXiv:2301.12314 (2023)
28. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
29. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
30. Serra, J., Suris, D., Miron, M., Karatzoglou, A.: Overcoming catastrophic forgetting with hard attention to the task. In: International conference on machine learning. pp. 4548–4557. PMLR (2018)
31. Smith, J., Hsu, Y.C., Balloch, J., Shen, Y., Jin, H., Kira, Z.: Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9374–9384 (2021)
32. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: 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)
33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)

34. Tang, Y.M., Peng, Y.X., Zheng, W.S.: When prompt-based incremental learning does not meet strong pretraining. In: ICCV. pp. 1706–1716 (2023)
35. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Wang, L., Xie, J., Zhang, X., Huang, M., Su, H., Zhu, J.: Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. NeurIPS (2023)
38. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European Conference on Computer Vision. pp. 631–648. Springer (2022)
39. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)
40. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
41. Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8715–8724 (2020)
42. Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. arXiv preprint arXiv:1708.01547 (2017)
43. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International conference on machine learning. pp. 3987–3995. PMLR (2017)
44. Zhao, B., Tang, S., Chen, D., Bilen, H., Zhao, R.: Continual representation learning for biometric identification. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1198–1208 (2021)
45. Zhou, D.W., Ye, H.J., Zhan, D.C., Liu, Z.: Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. arXiv preprint arXiv:2303.07338 (2023)