

Online-LoRA: Task-free Online Continual Learning via Low Rank Adaptation

Xiwen Wei

The University of Texas at Austin

xiwenwei@utexas.edu

Guihong Li

AMD

liguihong1995@gmail.com

Radu Marculescu

The University of Texas at Austin

radum@utexas.edu

Abstract

Catastrophic forgetting is a significant challenge in online continual learning (OCL), especially for non-stationary data streams that do not have well-defined task boundaries. This challenge is exacerbated by the memory constraints and privacy concerns inherent in rehearsal buffers. To tackle catastrophic forgetting, in this paper, we introduce Online-LoRA, a novel framework for task-free OCL. Online-LoRA allows to finetune pre-trained Vision Transformer (ViT) models in real-time to address the limitations of rehearsal buffers and leverage pre-trained models' performance benefits. As the main contribution, our approach features a novel online weight regularization strategy to identify and consolidate important model parameters. Moreover, Online-LoRA leverages the training dynamics of loss values to enable the automatic recognition of the data distribution shifts. Extensive experiments across many task-free OCL scenarios and benchmark datasets (including CIFAR-100, ImageNet-R, ImageNet-S, CUB-200 and CORe50) demonstrate that Online-LoRA can be robustly adapted to various ViT architectures, while achieving better performance compared to SOTA methods¹.

1. Introduction

Continual learning (CL) is pivotal in enabling machine learning systems to learn new concepts while preserving the previously learned knowledge. This ability is crucial for real-time applications like robotics, healthcare, and autonomous driving [45, 80]. However, a major hurdle in CL is *catastrophic forgetting*, where learning new information impairs performance on previously learned data.

Existing CL methods are typically classified along two dimensions: (1) task-based or task-free, depending on whether the boundaries between different tasks are known [30, 66, 99]; and (2) online or offline, based on whether the setting allows for multiple iterations over the data (offline) or requires a single pass through the data (online) [9, 28, 87].

In offline task-based CL, a series of tasks is presented sequentially. Typically, it is assumed that each task comprises

a dataset with samples drawn from a distinct independent and identically distributed (i.i.d.) distribution [22, 47]. The samples are i.i.d. within the same task, though across tasks the distributions may differ. Additionally, it is assumed that the probability distribution from which the data is drawn remains stationary between the training and inference phases for any given task [56, 98]. This assumption simplifies the learning problem by ensuring that the trained model can be effectively applied to new data from the same task without the need to account for distributional shifts [3].

While offline task-based CL has paved the way for understanding how models can sequentially learn, its assumptions are often misaligned with the intricacies of real-world data. For instance, the need of completed training before inference, known as the lack of anytime inference capability, does not hold for applications where decisions must be made on-the-fly based on data that is continuously changing. Moreover, the offline task-based CL assumes well-defined task boundaries, a condition seldom met outside tightly controlled experimental environments. In contrast, real-world data streams are inherently continuous and lack clear task boundaries, often exhibiting gradual transitions.

Motivated by the limitations of offline task-based setting, in this paper we focus on *task-free online CL*; this scenario is characterized by the constraints of seeing a stream of samples only once, and the absence of knowledge regarding task identities and task boundaries during both training and inference [19, 31, 46, 96].

Pre-trained Vision Transformers [20] have demonstrated superior performance on various vision tasks, hence integrating them into CL has attracted increasing interest [24, 89, 100]. Indeed, the extensive prior knowledge of pre-trained models enhances knowledge transfer [84], brings significant performance improvements compared to traditional SOTA methods trained from scratch, and provides robust generalizability and adaptability, especially valuable in data-scarce environments [26, 102]. Recent studies [17, 77, 90] have demonstrated the potential of using parameter efficient fine-tuning (PEFT) techniques like prompt tuning [48] and Low-Rank Adaptation (LoRA) [35] with pre-trained models for offline task-based CL.

Given the need for task-free OCL and the advantages

¹Code: <https://github.com/Christina200/Online-LoRA-official.git>

provided by pre-trained models and PEFT methods, we wonder, *whether task-free OCL can benefit from pre-trained models and PEFT as effectively as offline task-based CL*. To this end, we propose **Online-LoRA**, a new approach integrating pre-trained ViT and LoRA into the task-free OCL scenario. Online-LoRA learns incrementally with each new piece of information. More precisely, we propose an extensible architecture that expands the model with additional LoRA parameters where the loss surface plateaus [3]. Thus, by utilizing loss plateaus to recognize shifts in data distribution, our model remains robust in continuously changing environments. Furthermore, we propose a new online parameter regularization, aimed to mitigate forgetting and enhance memory efficiency. In our regularization, the importance weights are calculated on the LoRA parameters exclusively, rather than using the entire set of model parameters like in EWC [43]. This decreases the computational and memory requirements significantly, thus enabling online updates of the model parameter importance throughout the learning process.

We summarize our main contributions as follows:

- We propose Online-LoRA, an innovative approach that can efficiently learn from changing data in an online, task-free manner, thus enabling inference at any time. We achieve this through continual low rank adaptation and automatic detection of data distribution shifts based on loss plateaus.
- We present an online weight regularization mechanism that effectively mitigates forgetting by adapting the estimation of model parameter importance to the incoming data with minimal additional memory. We achieve this by using a Laplace approximation to estimate the uncertainty around the LoRA parameters.
- Our extensive evaluations with various ViT architectures across multiple task-free OCL benchmarks, under the settings of class and domain incremental learning, demonstrate that Online-LoRA consistently outperforms existing SOTA methods. Moreover, Online-LoRA exhibits robust performance across various task sequence lengths and ViT architectures, showcasing its effectiveness in diverse learning contexts.

2. Related work

2.1. Continual learning

Since many existing CL methods are offline task-based, their transition to the online, task-free setting is not trivial. Here, we discuss four categories of CL methods and their adaptability to task-free OCL.

Architecture-based methods in CL generate task-specific parameters by isolating subspaces or adding sub-networks

[21, 23, 37, 39, 65, 69, 70, 93, 94]. However, these approaches need task identity during training and inference, making them unsuitable for task-free settings; also, they typically involve significant additional parameters [40, 86, 97]. In contrast, [6] introduces virtual gradient updates from a virtual model, enabling ‘any-time inference’ for OCL.

Regularization-based methods selectively regularize the update of network parameters depending on their importance to the old tasks [10, 53, 92]. The importance of parameters can be determined using an approximation based on Fisher Information Matrix (FIM), as in EWC [43], Synaptic Intelligence [98] and MAS [1]. However, because EWC calculates the FIM at task transitions, it is not feasible in task-free OCL. On the other hand, EWC++ [12] combines the regularization approaches of EWC [43] and Synaptic Intelligence [98] and makes it suitable for online settings.

Rehearsal-based methods address catastrophic forgetting by combining old training examples from a memory buffer with current data [8, 11, 52, 56, 63, 68, 73, 82]. In principle, these methods are suitable for our task-free OCL setting, using strategies to retrieve and update the buffer [2, 4, 18, 25, 38, 41, 71, 72]. For instance, REMIND [27] enables efficient replay in OCL using compressed representations, while [7] integrates rehearsal with regularization techniques. However, their effectiveness decreases with smaller buffers [5] and they pose challenges in data-sensitive environments [74].

Prompt-based methods construct a pool of task-specific prompts, select and attach them to the pre-trained model [36, 67, 85]. Most of these methods assume explicit task boundaries and require information on these task boundaries for training [77, 88]; this is not feasible in task-free OCL. However, L2P [89] is suitable for task-free OCL as it employs an instance-wise prompt query. Similarly, MVP [61] is also suitable because it utilizes an instance-wise logit masking. In the class-incremental experiments within the original L2P codebase [89], a training trick is employed to mask out the classes not relevant to the current task. This trick contradicts the task-free OCL setting of having “no task identity information during training”. Thus, to ensure a fair comparison, we evaluate our Online-LoRA against L2P [89] and MVP [61] under the Stochastic Incremental Blurry task boundary (Si-blurry) scenario, a new scenario introduced in the MVP paper [61].

2.2. Parameter efficient fine-tuning

Parameter Efficient Fine-Tuning (PEFT) is an effective approach for transfer learning [33]. Instead of fine-tuning an entire pre-trained model, PEFT fine-tunes specific sub-modules within the network by adding a small amount of additional parameters. PEFT reduces computation, but achieves similar performance to full fine-tuning. PEFT has been successfully applied to vision transformer mod-

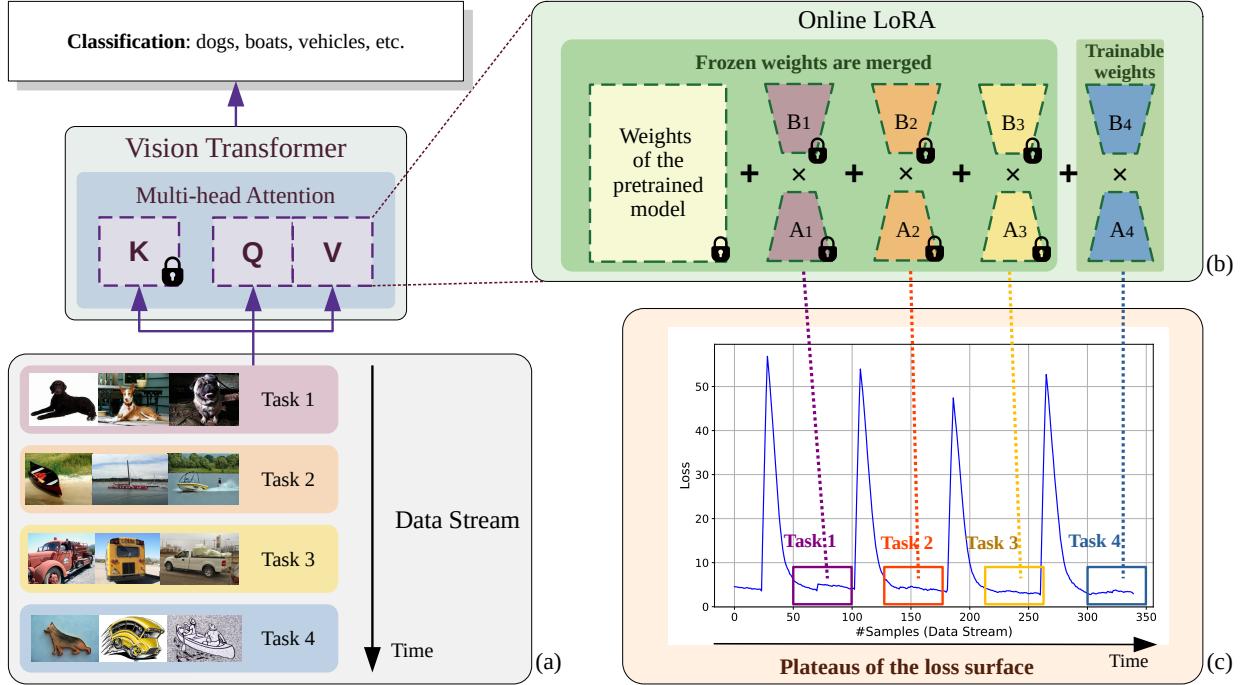


Figure 1. The overview of Online-LoRA. As the data is continuously streamed (a), a new pair of trainable LoRA parameters (A_4, B_4) is added (b) every time the loss surface encounters a plateau (c). Subsequently, the previous LoRA parameters ($A_1, B_1; A_2, B_2; A_3, B_3$) are frozen (the lock sign in (b)) and merged to the weights of the pre-trained ViT model.

els [16, 49], and one notable example is LoRA [35].

In LoRA, for a pre-trained weight matrix $W_{init} \in R^{d \times k}$, the update ΔW in $W \leftarrow W_{init} + \Delta W$ is reformulated as a low-rank decomposition: $\Delta W = BA$, where $A \in R^{r \times k}$ and $B \in R^{d \times r}$, and the rank $r \ll \min(d, k)$. W_{init} remains fixed during training and does not receive gradient updates, while A and B contain trainable parameters.

The application of PEFT in transformer-based models has gained popularity in CL research [17, 76, 90]. For example, SSIAT [79] incrementally tunes adapters [34] in pre-trained ViT. Additionally, C-LoRA [75] and InfLoRA [51] use separate LoRA sub-modules for each new task, and employ regularization to minimize interference between new and old tasks. However, these methods depend on the explicit knowledge of task boundaries, hence they are task-based offline CL approaches. To the best of our knowledge, our Online-LoRA is the first to extend LoRA to the task-free OCL scenario for transformer-based vision models.

3. Online-LoRA

3.1. Problem formulation

We define a data stream of unknown distributions $D = \{D_1, \dots, D_N\}$ over $X \times Y$, where X and Y are input and output space respectively [58]. At each time step s , the system receives a batch of non i.i.d samples x_k^t, y_k^t from the current distribution D_t of task t ; the system sees this batch

only once. Moreover, at any moment s , the distribution D_t can itself experience sudden or gradual changes from D_t to D_{t+1} . The system is unaware of when and how these distribution changes happen.

For simplicity, we assume that D_t is the data distribution of task t , and any shift from D_t to D_{t+1} is sudden. Of note, this remains a task-free setting, since gradual transitions from D_t to D_{t+1} can still be modeled by adding intermediate tasks and making these distributions increasingly similar, thus effectively blurring the explicit boundaries between tasks. Our Online-LoRA does not assume any task boundaries at any time.

3.2. Loss-guided model adaptation

In existing LoRA-based CL methods [75, 90], new LoRA parameters are added for each new task t' , resulting in a set of LoRA parameters denoted as $\{A_{t'}, B_{t'}\}$, where $A_{t'} \in R^{d \times r}, B_{t'} \in R^{r \times k}, d$ and k are the input and output dimensions of the attention layer, and rank $r \ll \min(d, k)$. When learning task t , if the initial ViT weights are denoted as W_{init} , then for an input sample X , the model output Y becomes:

$$Y = (W_{init} + \sum_{t'=0}^t B_{t'} A_{t'}) X \quad (1)$$

This *incremental model* effectively mitigates the catastrophic forgetting by minimizing the interference between

old and new tasks (see Figure 1). As shown, LoRA is applied only to the query and value projection matrices in all the attention layers. Since data from previous tasks is not available when training on future tasks, the LoRA parameters of old tasks are frozen and merged with the pre-trained weights to reduce the memory overhead. However, the existing LoRA-based methods rely on the knowledge of task boundaries during training, as a new pair of LoRA parameters is initialized at the beginning of each new task. In task-free OCL, data flows continuously without clear task boundaries, and there is no information about the start or the end of a task. There brings the need for a mechanism to determine when to initialize the new LoRA parameters.

To this end, we consider the idea of *loss surface* [3]. As learning progresses, a decreasing loss indicates effective learning from current samples. Conversely, an increasing loss suggests a shift in data distribution, hindering effective learning. We assume that the model converges before the distribution shifts. Then between these phases, *plateaus of the loss surface* occurs, signaling that the model has reached a stable state by fitting well to the current data distribution (see Appendix C for more details). At these plateaus, it is best to consolidate the learned knowledge by freezing the current LoRA weights and initializing a pair of new, trainable LoRA parameters. To prevent the accumulation of additional LoRA parameters, the frozen LoRA weights are merged into the pre-trained attention weights.

3.3. Online parameter importance estimation

Many studies have demonstrated the efficacy of weight regularization in reducing catastrophic forgetting [1, 12, 43]; this technique relies on estimating the importance of each parameter. However, in an online scenario where data distribution shifts constantly, parameter importance also varies continually. Therefore, a static estimation of parameter importance is not applicable. Furthermore, Online-LoRA utilizes pre-trained Vision Transformer (ViT) models, which have a *substantial number of parameters*. Techniques such as calculating the Fisher information matrix at each task-switch for parameter importance estimation are computationally inefficient in this context [43].

However, it is still useful to consider the model training from a Bayesian perspective as EWC does [43]. In Bayesian machine learning, model parameters are treated as random variables, and the prior knowledge about these parameters is updated via Bayes' rule. More precisely, given data D :

$$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) - \log p(D) \quad (2)$$

Assume D is split into two independent parts: current sample x and data observed at the last time step D_{prev} . We can rewrite the posterior probability of the parameters and the equation (2) becomes:

$$\log p(\theta|D) = \log p(x|\theta) + \log p(\theta|D_{prev}) - \log p(x) \quad (3)$$

Since calculating the posterior probability is usually intractable, following the work on the Laplace approximation [57], we approximate this posterior as a Gaussian centered at the maximum a-posteriori (MAP) solution θ_{MAP} with covariance given by the inverse of the Hessian. In our work, the empirical Fisher information matrix is used to approximate the covariance of the approximated posterior. More specifically, the LoRA adapter $\sum_{t'} B_{t'} A_{t'} X$ is treated as two separate linear layers with weights $\sum_{t'} A_{t'} \in R^{d \times r}$ and $\sum_{t'} B_{t'} \in R^{r \times k}$, respectively, rather than as a single linear layer with a low rank weight matrix [95]; this division enhances memory efficiency. In EWC [43], the size of the importance weight matrix equals to the number of parameters squared. For instance, to employ EWC in ViT-B/16, the model needs to store and update a $86.6M \times 86.6M$ matrix, representing a significant memory and computational overhead. By handling the LoRA adapter as two distinct layers, our Online-LoRA approach employs two smaller importance weight matrices, $\Omega^{A,l} \in R^{d \times r}$ and $\Omega^{B,l} \in R^{r \times k}$, for each attention layer. The combined size of these matrices is proportional to the total number of LoRA parameters, calculated as follows: #attention heads \times 2 (for Q and V projection matrices) \times input size \times rank \times 2. For a ViT-B/16 model with a LoRA rank of 4, this equates to a total of: 12 heads \times 2 \times 768 input size \times 4 rank \times 2 = 147,456. This additional memory footprint is negligible ($\sim 0.17\%$ of the total parameters of the ViT-B/16 model), which enables the *online* updates of the importance weights.

In offline CL, the parameter importance is computed based on the entire training dataset of the current task. This is not applicable to online CL because each training sample can only be seen once. We employ a small *hard buffer* containing samples with the highest loss (computed with the current model), selected from both the current sample and the existing buffer. The hard buffer is continually updated to replace any samples whose loss decreases significantly as the model trains, ensuring that it contains genuinely challenging examples. Due to concerns about memory constraints and privacy, the hard buffer is minimal (holds only 4 samples), yet vital to parameter importance estimation.

Therefore, we propose a memory and computationally efficient estimation of the parameter importance, focusing on the sensitivity of loss relative to LoRA parameters. The *importance weight matrices* $\Omega^{A,l} \in R^{d \times r}$ and $\Omega^{B,l} \in R^{r \times k}$ match the dimensions of LoRA parameters:

$$\Omega_{ij}^{A,l} = \frac{1}{N} \sum_{k=1}^N \nabla_{W_{ij}^{A,l}} \log p(x_k|\theta) \circ \nabla_{W_{ij}^{A,l}} \log p(x_k|\theta) \quad (4)$$

$$\Omega_{ij}^{B,l} = \frac{1}{N} \sum_{k=1}^N \nabla_{W_{ij}^{B,l}} \log p(x_k|\theta) \circ \nabla_{W_{ij}^{B,l}} \log p(x_k|\theta) \quad (5)$$

with parameters $W^{A,l}$ and $W^{B,l}$ of the new and trainable LoRA modules added to the l^{th} attention layer. x_k are samples from the hard buffer described above. The parameters denoted by θ encompass the entire model, which includes the pre-trained ViT, the frozen LoRA parameters, and the new trainable LoRA parameters $W^{A,l}$ and $W^{B,l}$ for all attention layers l . Finally, \circ is the Hadamard product (i.e. the element-wise product of two matrices).

After updating the importance weights, the model continues the learning process while penalizing changes to parameters that have been identified as important so far. Because $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(y, D; \theta)$, θ_{MAP} is given by model weights at the last loss surface plateau. As such, our final learning objective is:

$$\min_{W^A, W^B} \mathcal{L}(F(X; \theta), Y) + \mathcal{L}(F(X_B; \theta), Y_B) + L_{\text{LoRA}}(W^A, W^B) \quad (6)$$

$$L_{\text{LoRA}}(W^A, W^B) = \frac{\lambda}{2} \sum_{l \in |Attn|} ((\Omega^{A,l} \circ (W^{A,l}) \circ (W^{A,l})) \\ + (\Omega^{B,l} \circ (W^{B,l}) \circ (W^{B,l}))) \quad (7)$$

where $Attn$ is the set of attention layers in the model, $\mathcal{L}(F(X; \theta), Y)$ is the loss of current samples, $\mathcal{L}(F(X_B; \theta), Y_B)$ is the loss of hard buffer samples. \circ is the element-wise product of two matrices.

4. Experiments

4.1. Evaluation benchmarks

We evaluate our approach under three different scenarios: disjoint class-incremental, Si-Blurry class-incremental, and domain-incremental.

Disjoint class-incremental setting is when the datasets are split into disjoint tasks, each consisting of a unique set of classes. We conduct experiments with three datasets under this setting: Split-CIFAR-100 splits the CIFAR-100 dataset [44] into 10 tasks with 10 classes per task. Split-ImageNet-R splits the ImageNet-R dataset [32] into 10 tasks with 20 classes per task. Split-ImageNet-S splits the ImageNet-Sketch dataset [83] randomly into 10 tasks with 100 classes per task or into 20 tasks with 50 classes per task. Split-CUB-200 splits the CUB-200-2011 dataset [81] into 5 tasks with 40 classes per task.

Stochastic incremental-Blurry (Si-Blurry) [62] class-incremental setting is when the class distributions change in a stochastic manner, with classes overlapping across tasks and the task boundaries being dynamic and not clearly defined. We randomly select 50% of the entire classes to be "disjoint classes" (newly encountered classes that never appeared before), and 10% to be "blurry classes" (classes that do not belong to a fixed task and may appear in multiple learning tasks over time).

Domain-incremental setting is when the input distribution shifts over time, but the classes remain consistent. We use the CORo50 dataset [55] for this setting; it has 11 distinct domains (8 for training, 3 for testing). The samples from the training domains arrive sequentially.

4.2. Experimental details

Baselines We compare Online-LoRA against SOTA task-free OCL methods. The Upper-bound (*UB*) baseline refers to supervised fine-tuning on the entire dataset of i.i.d. data, representing the optimal performance. The SOTA methods selected for comparison include AGEM [13], ER [14], EWC++ [12], MIR [2], GDumb [64], DER++ [8], PCR [52], LODE (with DER++ [8]) [50], EMA (with DER++ [8] and with RAR [101]) [78], L2P [89] and MVP [61].

Evaluation metrics To evaluate the OCL performance, we choose three metrics, A_{AUC} , A_{Final} , and *Forgetting*. The A_{AUC} [91] evaluates the model accuracy throughout training, measuring the performance of anytime inference. The final accuracy A_{Final} [15, 59] measures the performance after the training is finished. *Forgetting* [89] measures the average difference between the final performance obtained for each task compared to the best performance on each task. Higher A_{AUC} and A_{Final} are better, while lower *Forgetting* is better. See Appendix A for the detailed definitions.

Implementation details We employ a ViT-B/16 (86.6M parameters) and a ViT-S/16 (48.6M parameters) [20] pre-trained on ImageNet as our backbone. For each setup, we evaluate all methods, including ours and other SOTA methods, using the same pre-trained models (see Appendix F.2).

We use the Adam optimizer [42] to train our Online-LoRA, with a 0.0002 learning rate for ViT-B/16 and 0.0005 for ViT-S/16. We set the size of the minimal hard buffer to 4, regularization factor λ to 2000 for all settings. See Appendix B for experimental details of Online-LoRA. For the other approaches, we refer to their original codebases for implementation and hyperparameter selection for a fair comparison (details in Appendix F). The buffer size of the replay-based methods is 500 (results for other buffer sizes in Appendix G). Given our focus on online CL, the training epoch is set to 1 for all experiments.

4.3. Main results

Results on disjoint class-incremental setting. Table 1 summarizes the results on the disjoint class-incremental benchmarks Split CIFAR-100, Split ImageNet-R, Split ImageNet-S, and Split CUB-200. Our Online-LoRA, outperforms all other compared methods consistently across the ViT-B/16 and ViT-S/16. On Split ImageNet-S, Online-LoRA exhibits standout performance, significantly outperforming all other methods, and notably reducing the gap to the upper bound.

	Split-CIFAR-100		Split-ImageNet-R		Split-ImageNet-S		Split-CUB-200	
	$A_{\text{Final}} (\uparrow)$	Forgetting (\downarrow)	$A_{\text{Final}} (\uparrow)$	Forgetting (\downarrow)	$A_{\text{Final}} (\uparrow)$	Forgetting (\downarrow)	$A_{\text{Final}} (\uparrow)$	Forgetting (\downarrow)
ViT-B/16	AGEM [13]	12.67 \pm 1.87	82.51 \pm 2.27	5.60 \pm 2.74	53.97 \pm 1.97	0.16 \pm 0.04	9.42 \pm 0.17	10.84 \pm 1.57
	ER [14]	44.85 \pm 1.83	44.67 \pm 4.29	40.99 \pm 3.96	32.38 \pm 0.89	30.21 \pm 0.70	37.14 \pm 1.83	31.66 \pm 0.83
	EWC++ [12]	10.61 \pm 0.74	84.10 \pm 1.11	3.86 \pm 2.02	56.95 \pm 1.46	0.32 \pm 0.28	22.46 \pm 4.69	26.14 \pm 3.46
	MIR [2]	48.36 \pm 3.11	43.41 \pm 1.02	41.51 \pm 2.99	31.32 \pm 5.17	30.33 \pm 3.81	35.92 \pm 1.75	31.64 \pm 2.97
	GDumb [64]	41.00 \pm 19.97	-	8.87 \pm 1.36	-	1.65 \pm 0.22	-	9.09 \pm 1.03
	PCR [52]	48.48 \pm 0.15	46.23 \pm 1.29	46.11 \pm 3.03	25.50 \pm 0.41	38.75 \pm 0.22	35.01 \pm 2.12	41.11 \pm 1.43
	DER++ [8]	36.64 \pm 6.11	56.94 \pm 7.55	30.90 \pm 8.04	24.26 \pm 4.14	6.47 \pm 0.06	15.34 \pm 0.15	26.61 \pm 1.27
	LODE (DER++) [50]	44.29 \pm 1.48	45.54 \pm 3.32	42.20 \pm 6.46	31.83 \pm 1.05	9.97 \pm 2.29	8.48\pm1.24	39.20 \pm 4.25
	EMA (DER++) [78]	42.28 \pm 4.36	55.59 \pm 1.48	41.75 \pm 1.98	32.65 \pm 1.55	16.88 \pm 2.23	36.28 \pm 1.09	35.26 \pm 3.31
	EMA (RAR) [78]	47.10 \pm 0.82	50.01 \pm 0.35	30.04 \pm 0.33	39.36 \pm 0.04	14.06 \pm 0.37	36.28 \pm 1.09	33.34 \pm 1.11
ViT-S/16	Ours	49.40\pm1.36	41.74\pm2.58	48.18\pm0.63	23.85\pm1.48	47.06\pm0.24	28.09 \pm 3.25	41.46\pm0.31
	<i>UB</i>	89.50 \pm 0.04	-	76.78 \pm 0.44	-	63.82 \pm 0.02	-	82.81 \pm 1.07
	AGEM [13]	7.43 \pm 2.15	82.45 \pm 5.46	2.35 \pm 0.87	48.01 \pm 0.05	2.75 \pm 2.86	18.81\pm0.44	1.40 \pm 0.17
	ER [14]	31.91 \pm 2.06	52.21 \pm 6.41	32.73 \pm 0.20	45.37 \pm 1.72	19.53 \pm 1.44	45.10 \pm 0.48	21.81 \pm 3.02
	EWC++ [12]	6.80 \pm 2.13	81.59 \pm 7.43	1.32 \pm 0.83	53.54 \pm 0.19	4.08 \pm 3.24	21.28 \pm 0.46	2.07 \pm 0.54
	MIR [2]	29.08 \pm 1.14	39.42 \pm 1.60	34.73\pm0.29	48.66 \pm 0.69	13.96 \pm 1.95	42.61 \pm 0.08	22.95 \pm 1.12
	GDumb [64]	10.87 \pm 4.94	-	5.33 \pm 1.09	-	2.09 \pm 0.32	-	3.28 \pm 0.99
	PCR [52]	32.89 \pm 1.47	39.90 \pm 2.51	21.96 \pm 0.27	45.12 \pm 0.08	14.37 \pm 0.95	43.96 \pm 0.48	22.28 \pm 2.73
	DER++ [8]	17.67 \pm 4.04	51.65 \pm 3.67	22.17 \pm 4.27	54.79 \pm 0.89	18.15 \pm 0.66	46.22 \pm 0.95	29.53 \pm 2.37
	LODE (DER++) [50]	28.65 \pm 3.06	40.42 \pm 1.58	31.65 \pm 0.72	43.72 \pm 0.09	17.59 \pm 0.84	47.85 \pm 0.23	26.81 \pm 0.89
ViT-S/16	EMA (DER++) [78]	12.76 \pm 0.65	41.17 \pm 1.75	20.89 \pm 3.05	48.03 \pm 1.79	12.93 \pm 0.13	22.59 \pm 0.16	35.79 \pm 5.27
	EMA (RAR) [78]	19.21 \pm 2.16	41.99 \pm 1.73	16.11 \pm 0.35	50.58 \pm 0.83	14.50 \pm 2.71	23.79 \pm 2.91	34.53 \pm 1.04
	Ours	32.16\pm0.24	38.64\pm0.65	33.21 \pm 0.50	42.76\pm0.18	22.45\pm0.43	44.56 \pm 0.24	37.41\pm0.16
	<i>UB</i>	86.55 \pm 0.01	-	69.94 \pm 0.34	-	59.28 \pm 0.11	-	73.91 \pm 1.16

Table 1. Results of disjoint class-incremental learning. ‘ \uparrow ’ means higher is better and ‘ \downarrow ’ means lower is better. Regularization-based methods (EWC++, AGEM, and LODE) yield low accuracy and low forgetting on Split ImageNet-S. This is because their overly rigid constraints on model updates hinder effective learning. The best results are noted by **bold**. *UB* is the upper-bound performance.

	CIFAR-100 [44]		ImageNet-R [32]		ImageNet-S [83]	
	$A_{\text{AUC}} (\uparrow)$	$A_{\text{Final}} (\uparrow)$	$A_{\text{AUC}} (\uparrow)$	$A_{\text{Final}} (\uparrow)$	$A_{\text{AUC}} (\uparrow)$	$A_{\text{Final}} (\uparrow)$
ViT-B/16	L2P	43.01 \pm 9.37	39.86 \pm 2.28	22.71 \pm 1.86	27.08 \pm 2.49	10.02 \pm 0.42
	MVP	47.52 \pm 9.74	44.49 \pm 0.93	27.79 \pm 0.62	31.64 \pm 1.77	10.68 \pm 0.45
	Ours	60.12\pm5.79	61.70\pm6.29	45.05\pm1.59	48.00\pm6.01	30.81\pm2.09
	<i>UB</i>	89.50 \pm 0.04	-	76.78 \pm 0.44	-	63.82 \pm 0.02
ViT-S/16	L2P	37.82 \pm 12.19	30.88 \pm 1.39	24.31 \pm 1.83	21.83 \pm 2.13	2.00 \pm 0.12
	MVP	40.31 \pm 9.52	35.55 \pm 2.11	27.04 \pm 1.09	26.67 \pm 3.70	2.27 \pm 0.14
	Ours	52.84\pm7.97	58.72\pm1.44	39.47\pm1.93	36.61\pm4.63	15.35\pm0.92
	<i>UB</i>	86.55 \pm 0.01	-	69.94 \pm 0.34	-	59.28 \pm 0.11

Table 2. Results of Si-blurry class-incremental learning. ‘ \uparrow ’ means higher is better and ‘ \downarrow ’ means lower is better. All datasets are split into 5 blurry tasks. To ensure a fair comparison with L2P [89] and MVP [61], we exclude the loss from hard buffer samples in Online-LoRA. The best results are noted by **bold**.

As shown in Table 1, Online-LoRA maintains a consistent and strong performance across various dataset sizes. In comparison, GDumb [64] exhibits unstable performance on the smaller dataset, Split CIFAR-100, and performs poorly on larger datasets such as Split-ImageNet-R and Split ImageNet-S. The main issue with GDumb is its exclusive reliance on a replay buffer to retrain the model. With larger datasets, a small buffer size tends to cause class imbalance, as it cannot represent the dataset diversity adequately. Online-LoRA, on the other hand, does not face this issue because it utilizes a small but highly targeted ‘hard buffer’ consisting of samples that the current model finds most challenging, as indicated by their high loss values.

This selective buffering approach is not only effective, as shown in Section 4.5, but it also sidesteps the drawbacks of a large memory buffer by not overly relying on it.

Results on Si-blurry class-incremental setting.

Table 2 summarizes the results on the Si-blurry class-incremental benchmarks with datasets CIFAR-100, ImageNet-R, and ImageNet-S. In the Si-blurry scenario, Online-LoRA consistently outperforms all the considered methods by significant margins across both metrics, A_{AUC} and A_{Final} . The superior performance in anytime inference can be largely attributed to Online-LoRA strategic utilization of loss surface plateaus, which consolidates the knowledge precisely when needed. Online-LoRA is

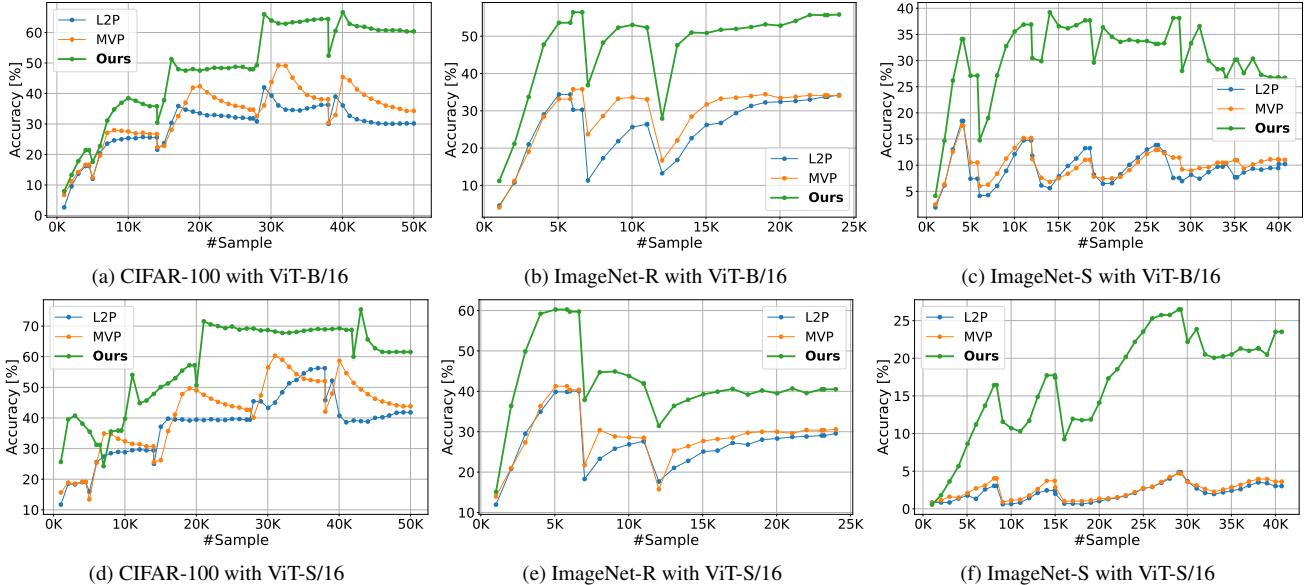


Figure 2. Average accuracy versus number of samples for Si-Blurry CIFAR-100, ImageNet-R, and ImageNet-S scenarios. As shown, the Online-LoRA consistently outperforms competing methods, maintaining high accuracy throughout.

also more flexible than EWC [43] which does so only at specific discrete moments; Online-LoRA also avoids the excessive frequency of updates that introduce noise as seen in EWC++ [12].

Figure 2 displays the trend of accuracy as more samples are provided, highlighting the consistent performance of Online-LoRA across two different ViT architectures. Compared to other methods, Online-LoRA effectively learns new knowledge from incoming samples, which leads to an increase in accuracy.

Results on domain-incremental setting. Table 3 summarizes the results on the domain-incremental setting. Our proposed method, Online-LoRA, not only significantly outperforms other SOTA methods, but also closes a substantial portion of the gap with the upper-bound (UB) performance.

To summarize, the Online-LoRA consistently achieves superior performance under various setups. These results indicate its robustness and adaptability, not only in different ViT setups, but also for dynamically evolving data. In addition to effectively mitigating forgetting, Online-LoRA shows good plasticity.

4.4. Exploration with length of task sequence

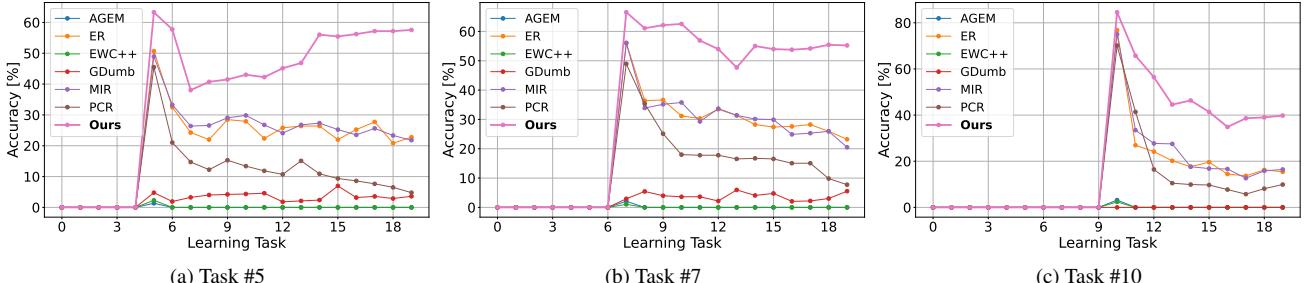
Table 4 summarizes the results on Split ImageNet-S dataset across varying task sequence lengths; Table 5 summarizes the results on Si-blurry ImageNet-S. As the task sequence is longer, all methods experience a decline in performance. However, Online-LoRA exhibits the smallest reduction in performance, showcasing its robustness against longer task sequences. This can be attributed to its utilization of loss surface plateaus, which effectively captures and adapts to shifts in data distribution at instance level.

	ViT-B/16		ViT-S/16	
	$A_{\text{Final}} (\uparrow)$	Forgetting (\downarrow)	$A_{\text{Final}} (\uparrow)$	Forgetting (\downarrow)
AGEM [13]	80.15 \pm 2.97	2.23 \pm 0.81	78.22 \pm 3.51	3.19 \pm 0.09
ER [14]	85.85 \pm 1.35	0.72 \pm 0.03	78.99 \pm 3.85	5.04 \pm 0.10
EWC++ [12]	78.65 \pm 6.51	2.31 \pm 0.17	79.03 \pm 4.54	4.80 \pm 0.69
MIR [2]	74.35 \pm 4.07	11.01 \pm 1.05	86.49 \pm 0.81	2.53 \pm 0.84
GDumb [64]	77.20 \pm 3.49	-	75.64 \pm 2.92	-
PCR [52]	87.16 \pm 0.73	0.78 \pm 0.03	75.20 \pm 1.48	0.61 \pm 0.02
DER++ [8]	81.88 \pm 7.06	10.13 \pm 7.00	89.33 \pm 0.62	0.42 \pm 0.57
LODE (DER++) [50]	77.02 \pm 2.22	17.30 \pm 2.82	83.48 \pm 5.84	24.54 \pm 0.94
L2P [89]	87.97 \pm 0.37	0.00\pm0.00	86.47 \pm 0.23	0.00\pm0.00
MVP [61]	84.82 \pm 0.54	0.00\pm0.00	79.85 \pm 0.33	3.55 \pm 0.39
Ours	93.71\pm0.01	0.00\pm0.00	90.96\pm0.02	0.00\pm0.00
Upper Bound (UB)	95.6 \pm 0.01	-	93.56 \pm 0.01	-

Table 3. Results of domain-incremental learning on CORoMe [55]. ‘ \uparrow ’ means higher is better and ‘ \downarrow ’ means lower is better. Online-LoRA not only achieves the highest final accuracy but also demonstrates the lowest forgetting.

In contrast, for prompt-based learning methods such as L2P, longer task sequences challenge the capacity of prompt pool as more task-specific information needs to be encoded. Similarly, for replay-based methods, the strategy of selecting informative samples from the buffer is prone to biases in longer task sequences. This bias may result in an inadequate representation of earlier tasks or an overemphasis on more recent tasks, hurting the methods overall performance.

Furthermore, Figure 3 shows the accuracy on the validation set for four tasks at the time they are first encountered and after each subsequent task is learned (see Appendix I for results of other tasks). As shown in Figure 3, Online-LoRA consistently outperforms the other SOTA methods in terms of preserving the performance of previously learned tasks, which underscores the effectiveness of our online parameter regularization in mitigating catastrophic forgetting.



(a) Task #5

(b) Task #7

(c) Task #10

Figure 3. Test accuracy of three tasks versus the number of learning tasks. ViT-B/16 model is used on Split ImageNet-S with 20 tasks. The accuracy for each task prior to the model being trained on it is recorded as zero, since no measurements are taken at that stage, as the model has not yet been exposed to the corresponding task.

Method	10 tasks		20 tasks	
	$A_{\text{Final}} (\uparrow)$	$Forgetting (\downarrow)$	$A_{\text{Final}} (\uparrow)$	$Forgetting (\downarrow)$
AGEM [13]	0.16±0.04	9.42±0.17	0.11±0.05	7.96±0.10
ER [14]	30.21±0.70	37.14±1.83	22.81±0.30	43.61±0.16
EWC++ [12]	0.32±0.28	22.46±4.69	0.11±0.05	5.26±0.45
MIR [2]	30.33±3.81	35.92±1.75	22.04±0.41	39.17±0.13
GDumb [64]	1.65±0.22	-	1.97±0.79	-
PCR [52]	38.75±0.22	35.01±2.12	17.87±2.18	45.46±0.07
DER++ [8]	6.47±0.06	15.34±0.15	2.29±0.23	23.14±0.06
LODE (DER++) [50]	9.97±2.29	8.48±1.24	13.47±0.66	35.89±1.63
EMA (DER++) [78]	16.88±2.23	36.28±1.09	11.55±0.66	38.56±0.22
EMA (RAR) [78]	14.06±0.37	36.28±1.09	9.05±0.60	29.77±1.70
Ours	47.06±0.24	28.09±3.25	44.19±2.09	28.48±0.24
Upper Bound (<i>UB</i>)		63.82±0.02		

Table 4. Comparison with other methods on Split ImageNet-S for different lengths of task sequences. ‘ \uparrow ’ means higher is better and ‘ \downarrow ’ means lower is better. ViT-B/16 model is used.

Method	Task sequence	ImageNet-S	
		$A_{\text{AUC}} (\uparrow)$	$A_{\text{Final}} (\uparrow)$
L2P [89]		10.02±0.42	13.58±4.04
MVP [61]	5 tasks	10.68±0.45	13.99±1.73
Ours		30.81±2.09	30.22±4.36
L2P [89]		9.06±0.43	12.49±3.39
MVP [61]	10 tasks	9.50±0.29	12.24±2.16
Ours		30.69±0.59	31.44±4.39
L2P [89]		6.57±0.54	7.13±0.89
MVP [61]	20 tasks	7.87±0.24	8.98±1.49
Ours		26.91±0.25	25.73±6.15

Table 5. Comparison with prompt-based methods on Si-blurry ImageNet-S at different length of task sequence. ViT-B/16 is used.

4.5. Ablation study

Table 6 shows the ablation study on the effectiveness of each component (“incremental LoRA” and “hard loss”) of Online-LoRA on Split ImageNet-R (10 tasks). The results demonstrate the crucial role of each component of Online-LoRA in overall performance. More results in Appendix E.

Simply fine-tuning a single set of LoRA parameters (i.e. without incorporating any components of Online-LoRA) results in significantly worse performance compared to our approach, with a 20% drop in accuracy (from 48.23% to 28.68%). Additionally, excluding the loss from hard buffer

Incremental LoRA	Hard loss	$A_{\text{Final}} (\uparrow)$	$Forgetting (\downarrow)$
-	-	28.68±0.13	53.45±0.04
✓	-	34.74±0.31	34.37±1.15
-	✓	36.08±0.19	35.75±0.33
✓	✓	48.23±0.74	23.85±1.08

Table 6. Ablation results of ViT-B/16 model on Split ImageNet-R dataset. ‘ \uparrow ’ means higher is better and ‘ \downarrow ’ means lower is better. “Incremental LoRA”: introducing new, trainable LoRA at each loss plateau with the model parameter regularization in Equation 7. . “Hard loss”: including $\mathcal{L}(F(X_B; \theta), Y_B)$ (the loss from hard buffer samples) in the final learning objective in Equation 6. ✓ indicates the presence of the component, – indicates its absence.

samples within the Online-LoRA framework leads to a substantial performance decline from 48.23% to 34.74% (a 13.5% decrease). This emphasizes the crucial role of maintaining a minimal buffer with only the four most challenging samples in mitigating forgetting.

Furthermore, the absence of new LoRA initialized at plateaus of the loss surface and model parameter regularization results in a significant performance decline of 12%, from 48.23% to 36.08%. This highlights the importance of continuously adding new LoRA parameters to minimize task interference and implementing online weight regularization to prevent catastrophic forgetting.

5. Conclusion

In this paper, we have presented Online-LoRA, a novel method for task-free online CL. Online-LoRA dynamically analyzes the loss surface to adapt the model to changing data distributions and uses online weight regularization to prevent catastrophic forgetting.

We have also provided empirical evidence to show the effectiveness of Online-LoRA across various scenarios. Notably, Online-LoRA shows substantial performance advantage over other state-of-the-art methods in scenarios involving long task sequences. Furthermore, Online-LoRA’s performance closely approaches the upper bound in domain-incremental settings. Given the widespread adoption of pre-trained models in CL, Online-LoRA offers a strong foundation for practical task-free online CL systems.

References

- [1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.
- [2] Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32, 2019.
- [3] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019.
- [4] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.
- [5] Benedikt Bagus and Alexander Gepperth. An investigation of replay-based approaches for continual learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.
- [6] Soumya Banerjee, Vinay K Verma, Avideep Mukherjee, Deepak Gupta, Vinay P Namboodiri, and Piyush Rai. Verse: Virtual-gradient aware streaming lifelong learning with anytime inference. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 493–500. IEEE, 2024.
- [7] Soumya Banerjee, Vinay Kumar Verma, and Vinay P Namboodiri. Streaming lifelong learning with any-time inference. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9486–9492. IEEE, 2023.
- [8] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.
- [9] Zhipeng Cai, Ozan Sener, and Vladlen Koltun. Online continual learning with natural distribution shifts: An empirical study with visual data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8281–8290, 2021.
- [10] Francisco Manuel Castro, Manuel J. Marín-Jiménez, Nicolás Gui Mata, Cordelia Schmid, and Alahari Karteek. End-to-end incremental learning. In *European Conference on Computer Vision*, 2018.
- [11] 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.
- [12] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018.
- [13] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.
- [14] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019.
- [15] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc’Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.
- [16] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.
- [17] Rajas Chitale, Ankit Vaidya, Aditya Kane, and Archana Ghotkar. Task arithmetic with lora for continual learning. *arXiv preprint arXiv:2311.02428*, 2023.
- [18] Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. In *International Conference on Machine Learning*, pages 1952–1961. PMLR, 2020.
- [19] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8250–8259, October 2021.
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [21] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. In *European conference on computer vision*, pages 386–402. Springer, 2020.
- [22] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.
- [23] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.
- [24] 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.
- [25] Yanan Gu, Xu Yang, Kun Wei, and Cheng Deng. Not just selection, but exploration: Online class-incremental continual learning via dual view consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7442–7451, 2022.

- [26] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.
- [27] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European conference on computer vision*, pages 466–483. Springer, 2020.
- [28] Jiangpeng He and Fengjing Zhu. Online continual learning for visual food classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2337–2346, 2021.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [30] Xu He, Jakub Sygnowski, Alexandre Galashov, Andrei A Rusu, Yee Whye Teh, and Razvan Pascanu. Task agnostic continual learning via meta learning. *arXiv preprint arXiv:1906.05201*, 2019.
- [31] Yuhang He, YingJie Chen, Yuhang Jin, Songlin Dong, Xing Wei, and Yihong Gong. Dyson: Dynamic feature space self-organization for online task-free class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.
- [32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.
- [33] 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.
- [34] 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.
- [35] 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. *arXiv preprint arXiv:2106.09685*, 2021.
- [36] 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.
- [37] Hyundong Jin and Eunwoo Kim. Helpful or harmful: Inter-task association in continual learning. In *European Conference on Computer Vision*, pages 519–535. Springer, 2022.
- [38] Xisen Jin, Junyi Du, and Xiang Ren. Gradient based memory editing for task-free continual learning. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020.
- [39] Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pages 10734–10750. PMLR, 2022.
- [40] Zixuan Ke, Bing Liu, and Xingchang Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in neural information processing systems*, 33:18493–18504, 2020.
- [41] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *European conference on computer vision*, volume 16, pages 411–428, 2020.
- [42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [43] 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.
- [44] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.
- [45] Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.
- [46] Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. *arXiv preprint arXiv:2001.00689*, 2020.
- [47] Timothée Lesort, Massimo Caccia, and Irina Rish. Understanding continual learning settings with data distribution drift analysis. *arXiv preprint arXiv:2104.01678*, 2021.
- [48] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [49] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [50] Yan-Shuo Liang and Wu-Jun Li. Loss decoupling for task-agnostic continual learning. *Advances in Neural Information Processing Systems*, 36, 2023.
- [51] Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23638–23647, 2024.
- [52] Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In *CVPR*, pages 24246–24255, 2023.
- [53] Yu Liu, Sarah Parisot, Gregory Slabaugh, Xu Jia, Ales Leonardis, and Tinne Tuytelaars. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 699–716. Springer, 2020.

- [54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.
- [55] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, pages 17–26, 2017.
- [56] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.
- [57] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.
- [58] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.
- [59] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3589–3599, 2021.
- [60] Microsoft. DeepSpeed: A deep learning optimization library. <https://github.com/microsoft/DeepSpeed>, 2024. Accessed: 2024-09-05.
- [61] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [62] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *ICCV*, 2023.
- [63] Julien Pourcel, Ngoc-Son Vu, and Robert M French. Online task-free continual learning with dynamic sparse distributed memory. In *European Conference on Computer Vision*, pages 739–756, 2022.
- [64] Ameya Prabhu, Philip Torr, and Puneet Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *The European Conference on Computer Vision (ECCV)*, August 2020.
- [65] Jathushan Rajasegaran, Munawar Hayat, Salman H Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [66] Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. *Advances in neural information processing systems*, 32, 2019.
- [67] Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madiam Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. *arXiv preprint arXiv:2301.12314*, 2023.
- [68] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [69] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *ArXiv*, abs/1606.04671, 2016.
- [70] Murray Shanahan, Christos Kapelanis, and Jovana Mitrović. Encoders and ensembles for task-free continual learning. *arXiv preprint arXiv:2105.13327*, 2021.
- [71] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021.
- [72] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021.
- [73] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.
- [74] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.
- [75] James Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *ArXiv*, abs/2304.06027, 2023.
- [76] James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14994–15004, 2023.
- [77] 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.
- [78] Albin Soutif-Cormerais, Antonio Carta, and Joost Van de Weijer. Improving online continual learning performance and stability with temporal ensembles. In *Conference on Lifelong Learning Agents*, pages 828–845. PMLR, 2023.
- [79] Yuwen Tan, Qinhao Zhou, Xiang Xiang, Ke Wang, Yuchuan Wu, and Yongbin Li. Semantically-shifted incremental adapter-tuning is a continual vittransformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23252–23262, 2024.
- [80] Eli Verwimp, Shai Ben-David, Matthias Bethge, Andrea Cossu, Alexander Gepperth, Tyler L Hayes, Eyke

- Hüllermeier, Christopher Kanan, Dhireesha Kudithipudi, Christoph H Lampert, et al. Continual learning: Applications and the road forward. *arXiv preprint arXiv:2311.11908*, 2023.
- [81] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds 200. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
 - [82] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pages 398–414. Springer, 2022.
 - [83] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.
 - [84] 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.
 - [85] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An ocam’s razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022.
 - [86] Zifeng Wang, Tong Jian, Kaushik Chowdhury, Yanzhi Wang, Jennifer Dy, and Stratis Ioannidis. Learn-prune-share for lifelong learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 641–650. IEEE, 2020.
 - [87] Zhen Wang, Liu Liu, Yajing Kong, Jiaxian Guo, and Dacheng Tao. Online continual learning with contrastive vision transformer. In *European Conference on Computer Vision*, pages 631–650. Springer, 2022.
 - [88] 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. *European Conference on Computer Vision*, 2022.
 - [89] 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.
 - [90] Martin Wistuba, Prabhu Teja Sivaprasad, Lukas Balles, and Giovanni Zappella. Continual learning with low rank adaptation. In *NeurIPS 2023 Workshop on Distribution Shifts (DistShifts)*, 2023.
 - [91] Hyun woo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. *ArXiv*, abs/2110.10031, 2021.
 - [92] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019.
 - [93] Ziyang Wu, Christina Baek, Chong You, and Yi Ma. Incremental learning via rate reduction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1133, 2020.
 - [94] Mengqi Xue, Haofei Zhang, Jie Song, and Mingli Song. Meta-attention for vit-backed continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 150–159, June 2022.
 - [95] Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *ArXiv*, abs/2308.13111, 2023.
 - [96] Fei Ye and Adrian G Bors. Online task-free continual generative and discriminative learning via dynamic cluster memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26202–26212, 2024.
 - [97] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.
 - [98] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.
 - [99] Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. Task agnostic continual learning using online variational bayes. *arXiv preprint arXiv:1803.10123*, 2018.
 - [100] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023.
 - [101] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. In *Advances in Neural Information Processing Systems*, volume 35, pages 14771–14783, 2022.
 - [102] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023.
 - [103] Yitao Zhu, Zhenrong Shen, Zihao Zhao, Sheng Wang, Xin Wang, Xiangyu Zhao, Dinggang Shen, and Qian Wang. Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis, 2023.

In this supplementary materials, a unique labeling with an "S" prefix (e.g., S1, S2, etc.) is used, distinguishing them from the main paper references.

A. Evaluation Metrics

In this section, we present the definitions of the three evaluation metrics we used in our experiments, supplementing Section 4.2 in the main paper.

Let $a_{i,j}$ be the testing accuracy on the i^{th} task after training on j^{th} task. The total number of tasks is denoted by T .

Final Accuracy The final accuracy A_{Final} is calculated as the average accuracy across all tasks after training on the final task:

$$A_{\text{Final}} = \frac{1}{T} \sum_{i=1}^T a_{i,T} \quad (8)$$

Area Under the Curve of Accuracy The A_{AUC} (Area Under the Curve of Accuracy) is defined as the area under the curve (AUC) of the accuracy-to-# of samples curve [91]. To construct the curve, the accuracy is measured after each sample is observed. A_{AUC} measures the any time inference accuracy of the model:

$$A_{\text{AUC}} = \sum_{i=1}^k f(i \cdot \Delta n) \cdot \Delta n, \quad (9)$$

where the step size Δn is defined as $\Delta n = 1$, representing the number of samples observed between inference queries, and $f(\cdot)$ denotes the curve in the accuracy-to-{number of samples} plot. A high A_{AUC} indicates that the method consistently maintains high accuracy throughout training.

Forgetting Forgetting is defined as the averaged differences between the historical maximum accuracy of task k and the accuracy of task k after all tasks finish training:

$$\text{forgetting} = \frac{1}{T-1} \sum_{k=1}^{T-1} \max_{t=1,2,\dots,T-1} (a_{k,t} - a_{k,T}) \quad (10)$$

The last task T is excluded because the forgetting of the last task is always 0.

B. Experimental Details

In this section, we provide details of the experiments we reported in the paper, supplementing Section 4 in the main paper.

Data preprocessing Because we focus on the ViT architectures ViT-B/16 and ViT-S/16, all input images are resized to 224×224 and normalized to $[0, 1]$.

Hyperparameters For tuning the threshold values for each dataset (CIFAR-100 [44], ImageNet-R [32], ImageNet-S [83], CUB-200 [81], and CORe50 [55]), we conducted a grid search following the protocol in [58]. The threshold grid is shown in Table 7. Table 8 shows the threshold values we used in our experiments. For CIFAR-100, ImageNet-R, and ImageNet-S, these threshold values remain consistent in both disjoint and Si-blurry class-incremental scenarios.

We set the regularization factor $\lambda=2000.0$ (see Equation 7 in the main paper) for all experiments.

C. Loss Surface

Figure 4 shows more qualitative examples of how the loss surface recognizes data distribution shifts, supplementing Section 3.2 in the main paper. MAS [3] introduces the *loss surface* to derive information about incoming streaming data in the task-free scenario. As shown in Figure 4, the peaks on the loss surface indicate shifts in the input data distribution. And the stable regions, namely plateaus, signal the convergence of the model. For instance, the Split CIFAR-100 dataset has 10 distinct tasks, with the data distribution remaining constant within each task. As a result, during the learning process of Split CIFAR-100, there are 9 shifts in data distribution, corresponding to 9 peaks in the loss surface, as illustrated in Figure 4.

To identify plateaus on the loss surface, we employ a *loss window*, which is a sliding window that moves across consecutive training losses. Within this window, we closely observe both the mean and variance of the losses. A plateau is identified when both metrics fall below a predefined threshold (see Appendix B for details). Upon detecting a plateau, we proceed to introduce new LoRA parameters and update the estimation of the model parameter importance. Our goal in identifying plateaus is to mark periods of stable prediction following shifts in data distribution. Therefore, we only classify a phase as a plateau if it follows a peak. A peak is recognized when the loss window's mean increases by an amount exceeding the standard deviation of the window within a single batch.

D. Results of Swin Transformer

In this section, we present the results for the disjoint class-incremental and domain-incremental settings (for details on these settings, see Section 4.1 in the main paper) using the Swin Transformer architecture [54]. For a fair comparison, the hyperparameters for the baseline methods are set according to the descriptions in Appendix F.3. For

Threshold	CIFAR-100	ImageNet-R	ImageNet-S	CUB-200	CORe50
Mean	[2.2, 2.6, 2.8, 3.0]		[5.2, 5.4, 5.6, 5.8, 6.0]		[18.0, 24.0, 30.0]
Variance		[0.02, 0.03, 0.04, 0.06, 0.08, 0.1]			[0.6, 0.8, 1.0, 1.2]

Table 7. Hyperparameter grid for the mean and variance threshold values of the loss window in our Online-LoRA.

Threshold	CIFAR-100	ImageNet-R	ImageNet-S	CORe50	CUB-200
Mean	2.6	5.2	5.6	6.0	24.0
Variance	0.03	0.02	0.06	0.1	1.0

Table 8. Mean and variance thresholds of the loss window for different datasets.

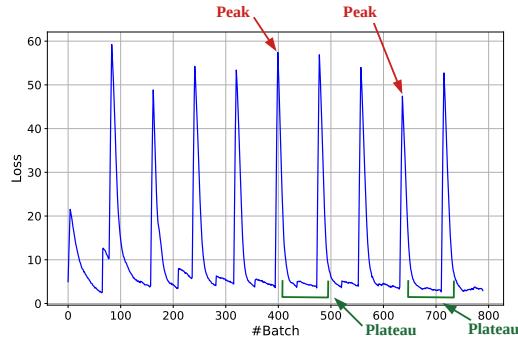


Figure 4. Loss surface of Online-LoRA on Split CIFAR-100 using ViT-B/16 model. Note that other peaks and plateaus exist but are not marked.

our method, we use a learning rate of 0.0003 for the Swin Transformer.

As shown in Table 9, our approach consistently outperforms other baseline methods in both disjoint class-incremental and domain-incremental learning settings. This demonstrates that our method remains effective across various ViT architectures, extending beyond the ViT-B/16 and ViT-S/16 models reported in Section 4.3 of the main paper.

E. Supplementary Ablation Study

E.1. Ablation Study on Imagenet-S Dataset

In addition to the ablation results on Split Imagenet-R presented in Section 4.5 of the main paper, this section provides further ablation results on the Split Imagenet-Sketch dataset with varying task lengths. As shown in Table 10, our Online-LoRA consistently outperforms other variants that lack certain components. These results demonstrate that both the hard buffer loss and incremental LoRA, along with online parameter regularization, are crucial for the performance of our approach.

The baseline involves continuous fine-tuning of a single set of LoRA parameters. In contrast, Online-LoRA intro-

duces an incremental LoRA architecture coupled with parameter importance-based regularization, and preserves a hard buffer along with its loss computations. Individually, each component improves performance and reduces forgetting. However, integrating both components into the baseline achieves the optimal performance, demonstrating the efficacy of our complete approach.

E.2. Impact of Pre-trained Weights

In this section, we demonstrate that our experimental settings do not provide any unfair advantage to our Online-LoRA approach through the use of pre-trained ViT models.

First, it is important to note that all baseline methods in our experiments utilize the same pre-trained ViT models as their backbones, just like Online-LoRA. Consequently, all methods benefit from the pre-training to varying extents, particularly those originally implemented with ResNet18 backbones (Table 12). For detailed information on the backbones used by each baseline, please refer to Appendix F.2.

Second, we show that simply using pre-trained models without applying any CL methods or strategies fails to yield competitive performance. To illustrate this, we introduce three simple baselines:

- **Frozen FT:** This baseline freezes the pre-trained backbone (feature extractor). Only the classification head (the final layer used for classification) is continuously fine-tuned on the data stream. Given that the model is pre-trained on the ImageNet-21K dataset, if any unfair advantage exists due to data leakage or other factors, it should be evident here by showing strong performance.
- **Continual FT:** This baseline fully fine-tunes the pre-trained model, including both the backbone and the classification head, on each new data batch. This is consistent with our OCL setting where the model encounters each data batch only once. If the pre-trained weights alone brings any unfair advantage, this baseline should perform competitively, similar to methods specifically designed for CL.
- **Random Head:** This baseline uses the pre-trained model’s backbone with a newly initialized classifier head and performs only inference without any fine-tuning. Since the classifier head is randomly initialized, it should provide a clear lower bound for per-

Method	Split-ImageNet-S		CORe50	
	$A_{\text{Final}} (\uparrow)$	$Forgetting (\downarrow)$	$A_{\text{Final}} (\uparrow)$	$Forgetting (\downarrow)$
AGEM [13]	31.67±0.96	50.12±0.27	90.15±1.31	1.16±0.05
ER [14]	42.60±0.75	38.68±0.26	88.93±2.99	4.16±0.09
EWC++ [12]	29.57±1.57	51.87±0.04	90.91±1.28	0.04±0.02
MIR [2]	42.90±0.19	38.49±0.15	87.47±0.65	5.67±0.14
GDumb [64]	14.76±1.13	-	79.52±3.00	-
Ours	53.75±0.29	32.86±0.89	95.29±0.06	0.00±0.00
<i>UB</i>	71.98±0.23	-	97.56±0.02	-

Table 9. Results of disjoint class-incremental learning and domain-incremental learning using Swin Transformer. ‘ \uparrow ’ means higher is better and ‘ \downarrow ’ means lower is better. The best results are noted by **bold**. *UB* is the upper-bound performance. With Swin Transformer, our Online-LoRA method consistently outperforms other baseline methods across various settings, demonstrating its adaptability and effectiveness across different ViT architectures.

Incremental LoRA	Hard loss	10 tasks		20 tasks	
		$A_{\text{Final}} (\uparrow)$	$Forgetting (\downarrow)$	$A_{\text{Final}} (\uparrow)$	$Forgetting (\downarrow)$
-	-	30.66±0.25	38.70±0.40	24.49±2.61	39.29±2.57
✓	-	31.11±2.60	34.62±2.98	32.47±0.29	33.14±1.39
-	✓	36.26±0.12	39.29±2.57	35.43±4.99	32.56±2.72
✓	✓	47.06±0.24	28.09±3.25	44.19±2.09	28.48±0.24

Table 10. Ablation results of ViT-B/16 model on Split ImageNet-Sketch dataset. ‘ \uparrow ’ means higher is better and ‘ \downarrow ’ means lower is better. “Incremental LoRA”: introducing new, trainable LoRA at each loss plateau with the model parameter regularization in Equation 7 in paper. “Hard loss”: including $\mathcal{L}(F(X_B; \theta), Y_B)$ (the loss from hard buffer samples) in the final learning objective in Equation 6 in paper. A check mark (✓) indicates the presence of the component, while a dash (—) indicates its absence.

formance, demonstrating that without any adaptation or learning, the model’s performance is essentially at chance level.

As shown in Table 11, **Random Head** baseline achieves near-zero accuracy, confirming that merely using pre-trained weights without adaptation to the test dataset does not have an advantage. Although the **Frozen FT** and **Continual FT** baselines outperform some CL methods (which also use the same pre-trained models), they still suffer from severe forgetting and exhibit a significant performance gap compared to other methods, particularly our Online-LoRA, with nearly a 20% difference in final accuracy and a 30% difference in forgetting.

These results demonstrate that the performance advantages of our Online-LoRA method over the baseline CL methods are not simply due to the use of pre-trained models. Instead, they arise from the effectiveness of our approach. The pre-trained weights provide a common foundation for all methods, but it is our approach that leads to superior performance.

F. Baseline Settings

In this section, we provide the experimental settings for the baseline methods used in our experiments².

F.1. Overview of Baselines

- **AGEM** [13]: Averaged Gradient Episodic Memory, utilizes samples in the memory buffer to constrain parameter updates.
- **ER** [14]: Experience replay, a rehearsal-based method with random sampling in memory retrieval and reservoir sampling in memory update.
- **EWC++** [12]: An online version of EWC [43], a regularization method that limits the update of parameters crucial to past tasks.
- **MIR** [2]: Maximally Interfered Retrieval, a rehearsal-based method that retrieves memory samples with loss

²Codebases used: https://github.com/AlbinSou/online_ema.git, <https://github.com/liangyanshuo/Loss-Decoupling-for-Task-Agnostic-Continual-Learning.git>, <https://github.com/FelixHuiweiLin/PCR.git>, <https://github.com/RaptorMai/online-continual-learning.git>

Method	Accuracy (\uparrow)	Forgetting (\downarrow)
Random Head	0.08 \pm 0.00	-
Frozen FT	27.98 \pm 0.29	55.12 \pm 0.43
Continual FT	28.49 \pm 0.21	53.49 \pm 0.07
AGEM [13]	5.60 \pm 2.74	53.97 \pm 1.97
ER [14]	40.99 \pm 3.96	32.38 \pm 0.89
EWC++ [12]	3.86 \pm 2.02	56.95 \pm 1.46
MIR [2]	41.51 \pm 2.99	31.32 \pm 5.17
GDumb [64]	1.65 \pm 0.22	-
PCR [52]	46.11 \pm 3.03	25.50 \pm 0.41
DER++ [8]	30.90 \pm 8.04	24.26 \pm 4.14
LODE (DER++) [50]	42.20 \pm 6.46	31.83 \pm 1.05
EMA (DER++) [78]	41.75 \pm 1.98	32.65 \pm 1.55
EMA (RAR) [78]	30.04 \pm 0.33	39.36 \pm 0.04
Online-LoRA (ours)	48.18\pm0.63	23.85\pm1.48
UB	63.82 \pm 0.02	-

Table 11. Performance comparison between pre-trained models without CL strategies and pre-trained models with CL strategies on Split ImageNet-R (online class-incremental learning setting). ViT-B/16 backbone is used. While some methods do not outperform simple fine-tuning on a continuous data stream, other CL methods provide significant performance improvements to the pre-trained model. This demonstrates that the advantages of CL methods, including Online-LoRA, are not solely due to the use of pre-trained weights but also stem from the effectiveness of the methods themselves. UB is the upper-bound baseline trained on the i.i.d. data of the datasets. The best results are noted by **bold**.

increases given the estimated parameter update based on the current batch.

- **GDumb** [64]: Greedy Sampler and Dumb Learner, a strong baseline that greedily updates the memory buffer from the data stream with the constraint to keep a balanced class distribution.
- **PCR** [52]: Proxy-based contrastive replay, a rehearsal-based method that replaces the samples for anchor with proxies in a contrastive-based loss.
- **DER++** [8]: Dark Experience Replay++, a rehearsal-based method using knowledge distillation from past experiences.
- **LODE** [50]: Loss Decoupling, a rehearsal-based method that decouples the learning objectives of old and new tasks to minimize interference.
- **EMA** [78]: Exponential Moving Average, a model ensemble method that combines models from various training tasks.

- **L2P** [89]: Learning to Prompt, a prompt-based method that prepends learnable prompts selected from a prompt pool to the embeddings of a pre-trained transformer.
- **MVP** [61]: Mask and Visual Prompt tuning, a prompt-based method that uses instance-wise feature space masking.

F.2. Backbone

Among the baseline methods we compare, L2P [89] and MVP [61] originally reported results using a ViT-B/16 model [20] pre-trained on ImageNet21k, while the other baselines (AGEM [13], ER [14], EWC++ [12], MIR [2], GDumb [64], DER++ [8], PCR [52], LODE [50], EMA [78]) reported results using a ResNet18 [29] architecture.

To ensure a fair comparison, we standardized our experimental setup by evaluating all baselines using the same pre-trained ViT model (ViT-B/16 and ViT-S/16). For methods originally implemented with ResNet18, we reimplemented them with ViT to match the experimental conditions of L2P and MVP. As shown in Table 12, all methods perform better with the pre-trained ViT-B/16 than with ResNet18, supporting our argument that using a pre-trained ViT provides a more consistent and stronger baseline for performance comparisons.

F.3. Training Settings

The following settings are shared by the baseline methods (and our Online-LoRA) in the experiments:

- Buffer Size: 500. Methods using a buffer include AGEM [13], ER [14], MIR [2], GDumb [64], PCR [52], DER++ [8], LODE [50], and EMA [78].
- Optimizer: Adam.
- Batch Size: 64.

In Table 13, we summarize the hyperparameters used for all baseline methods in our experiments. To ensure a fair comparison, we adopted these hyperparameters from their original codebases. However, because the baseline methods used different backbones and batch sizes in their original experiments, we adjusted the learning rates for some baselines to standardize the comparison across all methods. For tuning the learning rates, we followed the protocol outlined in [58] and conducted a grid search on a small cross-validation set. The hyperparameter grid for the baselines is detailed in Table 14.

G. Exploration with Buffer Size

Table 15 we show more results of the impact of buffer sizes on the performance of replay-based methods (AGEM [13], ER [14], GDumb [64], MIR [2]).

Method	Acc. w/ ResNet18	Acc. w/ ViT-B/16	Performance Gain (%)
AGEM [13]	5.4±0.6	12.67±1.87	134.63
ER [14]	14.5±0.8	44.85±1.83	209.31
EWC++ [12]	4.8±0.2	10.61±0.74	121.04
MIR [2]	14.8±0.7	48.36±3.11	226.76
GDumb [64]	24.8±0.7	41.00±19.97	65.32
PCR [52]	21.8±0.9	48.48±0.15	122.39
DER++ [8]	15.5±1.0	36.64±6.11	136.39
LODE (DER++) [50]	37.8±1.1	44.29±1.48	17.17
EMA (DER++) [78]	23.2±1.2	42.28±4.36	82.24
EMA (RAR) [78]	35.4±1.2	47.10±0.82	33.05

Table 12. Performance comparison on CIFAR-100 between ResNet18 and pre-trained ViT-B/16 in an online class-incremental learning scenario. Acc. stands for Accuracy. All rehearsal-based methods use a buffer size of 500 for fair comparison. The results demonstrate that there is no unfair comparison in our experiments, as all methods benefit from the pre-trained ViT-B/16 model. The performance gain is computed as the percentage increase from the ResNet18 accuracy to the ViT-B/16 accuracy for each method.

Method	CIFAR-100	ImageNet-R	ImageNet-S	CUB-200	CORe50
AGEM [13]		LR=0.0001, WD=0.0001			
ER [14]		LR=0.0001, WD=0.0001, Episode memory per batch=10			
EWC++ [12]		LR=0.0001, WD=0.0001, $\lambda=100$, $\alpha=0.9$ Number of training batches after which the Fisher information will be updated: 50			
MIR [2]		LR=0.0001, WD=0.0001, Number of subsample=50			
GDumb [64]		LR=0.001, WD=0.0001, Minimal learning rate: 0.0005, Gradient clipping=10, Epochs to train for memory=30			
PCR [52]		LR=0.0001, WD=0.0001, Episode memory per batch=10, Temperature=0.09, Warmup of buffer before retrieve=4			
DER++ [8]		LR=0.0003, $\alpha=0.2$, $\beta=0.5$			
LODE [50]		LR=0.0003, $C=1.0$, $\rho=0.1$			
EMA [78]		LR=0.0002, λ for warm-up: 0.9, $\lambda=0.99$			
L2P [89]	LR=0.003, Size of the prompt pool=10, Length of a single prompt=10, Number of prepended prompt=4				
MVP [61]		LR=0.005, $\gamma=2.0$, $m=0.5$, $\alpha=0.5$			

Table 13. Hyperparameters for the baseline methods on ViT-B/16. LR: learning rate. WD: weight decay.

As shown in Table 15, when the buffer size increases, all replay-based methods see improvements in their performance across the benchmarks. Notably, when the buffer size hits 5000 (a large capacity; 20% of the ImageNet-R training set, 12.5% of the ImageNet-S training set), the difference in performance between GDumb and other replay-based methods narrows. This suggests that the sophisticated memory retrieval strategies employed by these other methods do not significantly outperform GDumb’s simple approach of training directly on the buffered data. Moreover, the performance of rehearsal-based methods drops when the buffer size shrinks. This highlights the efficiency of our

Online-LoRA, which achieves high performance using just a minimal buffer size of 4.

H. Computation Analysis

In this section, we present the model parameter size, training FLOPs, and training time for our Online-LoRA and the baseline methods.

As shown in Table H, our Online-LoRA model introduces approximately 0.6M additional parameters due to the inclusion of LoRA parameters, which represents a negligible increase (0.69%) compared to the original size of the ViT-B/16 model. Notably, our memory buffer con-

Method	CIFAR-100	ImageNet-R	ImageNet-S	CUB-200	CORe50
AGEM [13]		LR: [0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1] WD: [0.0001, 0.001, 0.01, 0.1]			
ER [14]		LR: [0.0001, 0.0003, 0.001, 0.003] WD: [0.0001, 0.001, 0.01, 0.1]			
EWC++ [12]		LR: [0.0001, 0.001, 0.01, 0.1] WD: [0.0001, 0.001]			
MIR [2]		LR: [0.0001, 0.001, 0.01, 0.1] WD: [0.0001, 0.001]			
GDumb [64]		LR: [0.001, 0.01, 0.1] WD: [0.0001, 0.000001]			
PCR [52]		LR: [0.0001, 0.001, 0.01, 0.1] WD: [0.0001, 0.001]			
DER++ [8]		LR: [0.0003, 0.003, 0.03]			
LODE [50]		LR: [0.0003, 0.003, 0.03]			
EMA [78]		LR: [0.0001, 0.0002, 0.0003, 0.0004, 0.0005]			

Table 14. Hyperparameter grid for the baseline methods using the ViT-B/16 backbone. LR: learning rate; WD: weight decay. Since L2P [89] and MVP [61] use the same backbone and batch size as in our experiments, their learning rates were not adjusted.

	Buffer size	Method	Split-ImageNet-R	Split-ImageNet-S	Core50	
500	500	AGEM [13]	5.60 \pm 2.74	0.16 \pm 0.04	80.15 \pm 2.97	
		ER [14]	40.99 \pm 3.96	30.21 \pm 0.70	85.85 \pm 1.35	
		MIR [2]	41.51 \pm 2.99	30.33 \pm 3.81	74.35 \pm 4.07	
		GDumb [64]	8.87 \pm 1.36	1.65 \pm 0.22	77.20 \pm 3.49	
1000	1000	AGEM [13]	7.16 \pm 1.56	0.23 \pm 0.04	78.73 \pm 3.87	
		ER [14]	44.71 \pm 2.63	34.32 \pm 0.53	84.27 \pm 4.11	
		MIR [2]	46.65 \pm 5.63	33.99 \pm 1.72	82.64 \pm 1.12	
		GDumb [64]	19.19 \pm 1.36	2.71 \pm 0.12	78.09 \pm 3.75	
5000	5000	AGEM [13]	7.21 \pm 0.34	0.12 \pm 0.02	77.57 \pm 3.56	
		ER [14]	47.23 \pm 2.71	37.65 \pm 0.23	81.32 \pm 2.19	
		MIR [2]	49.33\pm3.49	35.90 \pm 2.35	81.18 \pm 3.20	
		GDumb [64]	46.08 \pm 0.64	9.68 \pm 0.28	69.42 \pm 1.06	
Ours		48.18 \pm 0.63	47.06\pm0.24	93.71\pm0.01		
<i>UB</i>		76.78 \pm 0.44	63.82 \pm 0.02	95.60 \pm 0.01		

Table 15. Results of replay-based methods with different buffer size. A_{Final} metric and ViT-B/16 model is used. Each dataset has 10 disjoint tasks. *UB* is the upper-bound baseline trained on the i.i.d. data of the datasets. The best results are noted by **bold**.

tains only 4 data samples, whereas other baselines (except EWC++) require at least 500 samples in their buffers to achieve comparable performance (see Appendix G for more details). Regarding computational consumption measured by FLOPs during training, Online-LoRA demonstrates advantages over EWC++ [12], thanks to our efficient computation of the importance weight matrix, as explained in

Section 3.3 of the main paper. The extremely low FLOPs of GDumb [64] can be attributed to its design, which involves greedily updating the memory buffer without employing additional strategies. However, its training time is relatively high because retraining is triggered frequently to maintain a balanced memory buffer, which adds overhead despite the low FLOPs.

Method	#params (M)	FLOPs ($\times 10^{15}$)	Training time (s)
AGEM [13]	85.88	140.52	828.39
ER [14]	85.88	140.05	849.43
EWC++ [12]	85.88	214.36	1076.53
GDumb [64]	85.88	18.44	2078.59
MIR [2]	85.88	161.04	1069.29
Ours	86.47	151.20	864.60

Table 16. Computational statistics for Online-LoRA and baseline methods on CIFAR-100 in the online class-incremental learning scenario using the ViT-B/16 backbone. FLOPs are measured as ‘forward FLOPs per GPU’ using the DeepSpeed FLOPS Profiler [60]. All experiments are conducted on a single NVIDIA A100 GPU.

I. Task Accuracy

In this section, Figure 5 and Figure 6 show task accuracy as a function of the number of learning tasks as described in Section 4.4 in the main paper. The ViT-B/16 model is employed on the Split ImageNet-S dataset with 20 tasks. These results demonstrate that our Online-LoRA consistently outperforms the other methods in mitigating the forgetting of previously learned tasks.

Figure 5a shows that AGEM [13] begins with an initial accuracy of $\sim 10\%$. However, this accuracy drastically decreases for subsequent tasks, eventually dropping to zero. Given that the Split ImageNet-S dataset consists of 20 tasks with 500 classes per task, AGEM’s performance is no better than that of a random model, which would have an expected accuracy of 0.2%. This dramatic decline is primarily due to the increasingly restrictive constraints placed on gradient updates as the number of tasks increases. Such constraints significantly hurt the model’s ability to learn from new tasks, showing a fundamental weakness of AGEM in handling long sequences of diverse tasks. A similar issue was observed with EWC++ [12], another regularization-based approach.

In contrast, our Online-LoRA model does not encounter this problem even though an online weight regularization is used. This is because our model is continuously expanded by adding new LoRA parameters (see Section 3.2 in the main paper). This strategy allows the model to adapt to new information more flexibly, bypassing the learning limitations encountered by traditional regularization methods like AGEM and EWC++.

J. Code

Our code will be publicly available at: <https://github.com/Christina200/Online-LoRA-official.git>. Our implementation of LoRA is based on the codebase of MeLo [103].

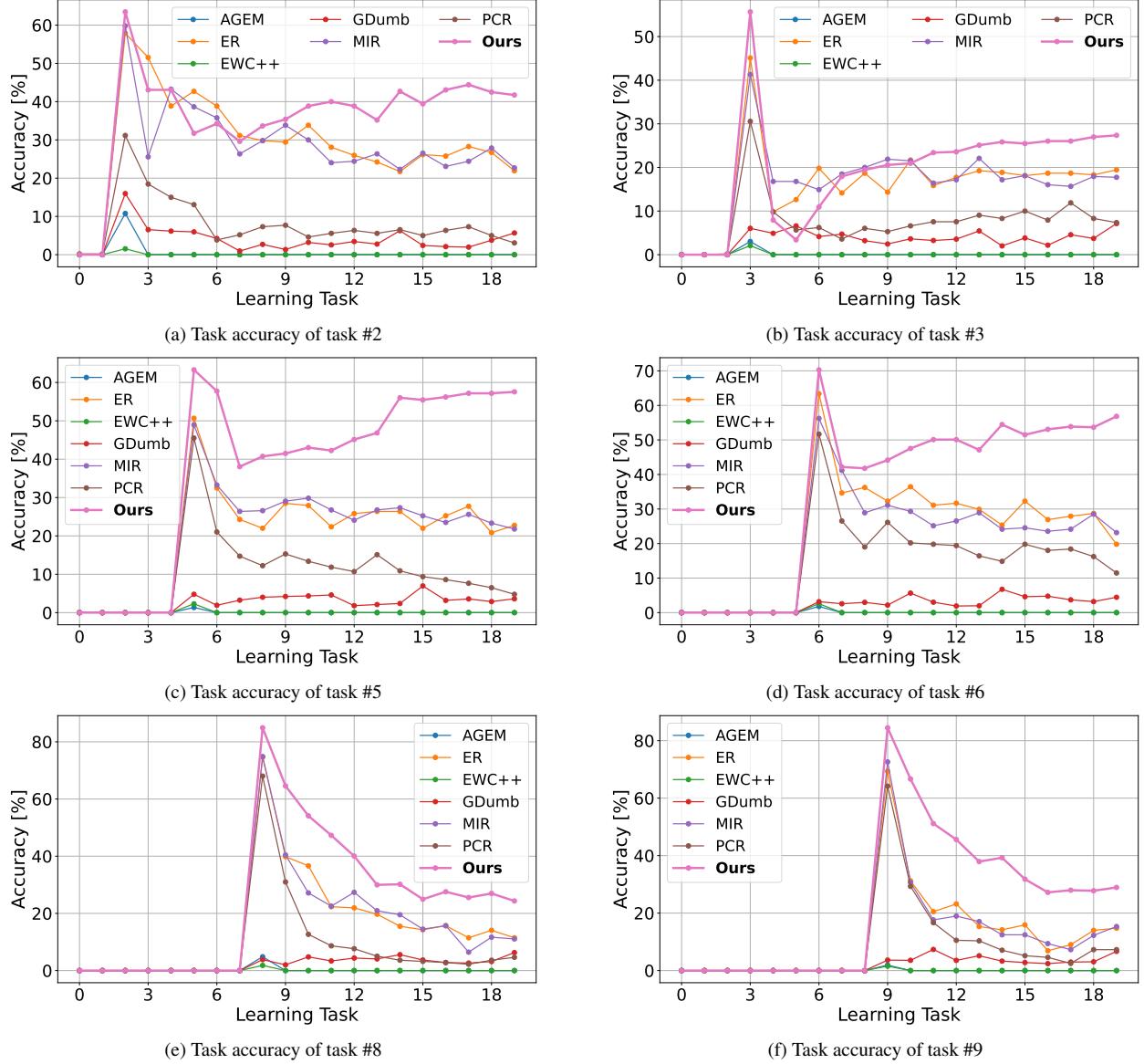


Figure 5. Task accuracy versus the number of learning tasks of task #2 to task #9. Our Online-LoRA consistently outperforms all the other methods in maintaining accuracy on previously learned tasks. Note that the recorded accuracy for initial tasks is zero, not due to poor model performance, but because our evaluation prioritizes mitigating forgetting in tasks the model has already encountered.

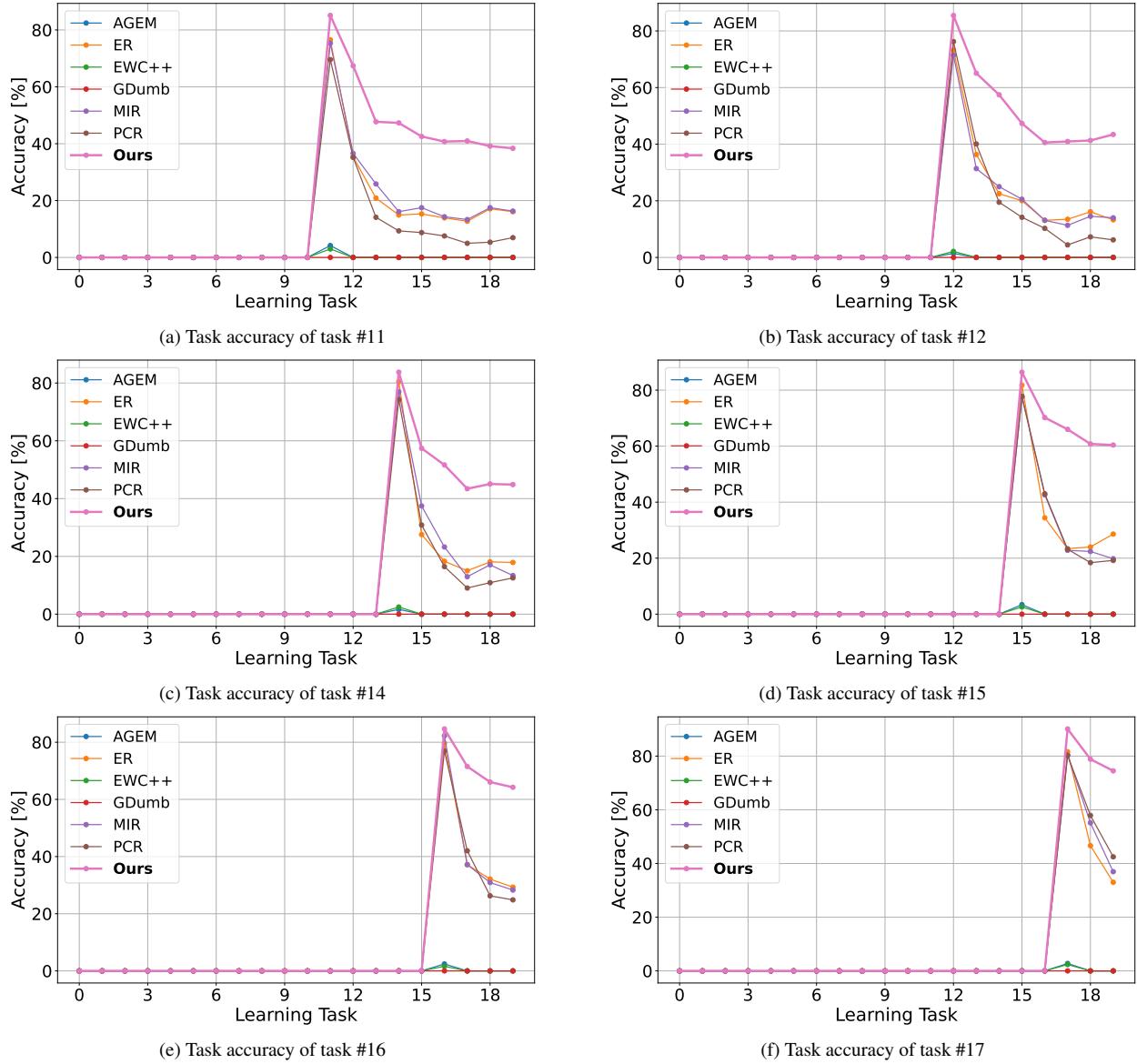


Figure 6. Task accuracy versus the number of learning tasks of task #11 to task #17. Compared to the results of task #2 to task #9 in Figure 5, our Online-LoRA has greater advantages over the other methods for these newer tasks #11 to task #17. Zero accuracy for initial tasks results from not measuring them at the time the specific task had not been learned yet.

References

- [1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.
- [2] Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32, 2019.
- [3] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019.
- [4] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.
- [5] Benedikt Bagus and Alexander Gepperth. An investigation of replay-based approaches for continual learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.
- [6] Soumya Banerjee, Vinay K Verma, Avideep Mukherjee, Deepak Gupta, Vinay P Namboodiri, and Piyush Rai. Verse: Virtual-gradient aware streaming lifelong learning with anytime inference. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 493–500. IEEE, 2024.
- [7] Soumya Banerjee, Vinay Kumar Verma, and Vinay P Namboodiri. Streaming lifelong learning with any-time inference. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9486–9492. IEEE, 2023.
- [8] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.
- [9] Zhipeng Cai, Ozan Sener, and Vladlen Koltun. Online continual learning with natural distribution shifts: An empirical study with visual data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8281–8290, 2021.
- [10] Francisco Manuel Castro, Manuel J. Marín-Jiménez, Nicolás Guil Mata, Cordelia Schmid, and Alahari Karteek. End-to-end incremental learning. In *European Conference on Computer Vision*, 2018.
- [11] 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.
- [12] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018.
- [13] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.
- [14] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019.
- [15] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc’Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.
- [16] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.
- [17] Rajas Chitale, Ankit Vaidya, Aditya Kane, and Archana Ghotkar. Task arithmetic with lora for continual learning. *arXiv preprint arXiv:2311.02428*, 2023.
- [18] Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. In *International Conference on Machine Learning*, pages 1952–1961. PMLR, 2020.
- [19] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8250–8259, October 2021.
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [21] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. In *European conference on computer vision*, pages 386–402. Springer, 2020.
- [22] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.
- [23] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.
- [24] 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.
- [25] Yanan Gu, Xu Yang, Kun Wei, and Cheng Deng. Not just selection, but exploration: Online class-incremental continual learning via dual view consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7442–7451, 2022.

- [26] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.
- [27] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European conference on computer vision*, pages 466–483. Springer, 2020.
- [28] Jiangpeng He and Fengjing Zhu. Online continual learning for visual food classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2337–2346, 2021.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [30] Xu He, Jakub Sygnowski, Alexandre Galashov, Andrei A Rusu, Yee Whye Teh, and Razvan Pascanu. Task agnostic continual learning via meta learning. *arXiv preprint arXiv:1906.05201*, 2019.
- [31] Yuhang He, YingJie Chen, Yuhang Jin, Songlin Dong, Xing Wei, and Yihong Gong. Dyson: Dynamic feature space self-organization for online task-free class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.
- [32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.
- [33] 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.
- [34] 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.
- [35] 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. *arXiv preprint arXiv:2106.09685*, 2021.
- [36] 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.
- [37] Hyundong Jin and Eunwoo Kim. Helpful or harmful: Inter-task association in continual learning. In *European Conference on Computer Vision*, pages 519–535. Springer, 2022.
- [38] Xisen Jin, Junyi Du, and Xiang Ren. Gradient based memory editing for task-free continual learning. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020.
- [39] Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pages 10734–10750. PMLR, 2022.
- [40] Zixuan Ke, Bing Liu, and Xingchang Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in neural information processing systems*, 33:18493–18504, 2020.
- [41] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *European conference on computer vision*, volume 16, pages 411–428, 2020.
- [42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [43] 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.
- [44] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.
- [45] Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.
- [46] Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. *arXiv preprint arXiv:2001.00689*, 2020.
- [47] Timothée Lesort, Massimo Caccia, and Irina Rish. Understanding continual learning settings with data distribution drift analysis. *arXiv preprint arXiv:2104.01678*, 2021.
- [48] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [49] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [50] Yan-Shuo Liang and Wu-Jun Li. Loss decoupling for task-agnostic continual learning. *Advances in Neural Information Processing Systems*, 36, 2023.
- [51] Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23638–23647, 2024.
- [52] Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In *CVPR*, pages 24246–24255, 2023.
- [53] Yu Liu, Sarah Parisot, Gregory Slabaugh, Xu Jia, Ales Leonardis, and Tinne Tuytelaars. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 699–716. Springer, 2020.

- [54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.
- [55] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, pages 17–26, 2017.
- [56] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.
- [57] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.
- [58] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.
- [59] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3589–3599, 2021.
- [60] Microsoft. DeepSpeed: A deep learning optimization library. <https://github.com/microsoft/DeepSpeed>, 2024. Accessed: 2024-09-05.
- [61] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [62] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *ICCV*, 2023.
- [63] Julien Pourcel, Ngoc-Son Vu, and Robert M French. Online task-free continual learning with dynamic sparse distributed memory. In *European Conference on Computer Vision*, pages 739–756, 2022.
- [64] Ameya Prabhu, Philip Torr, and Puneet Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *The European Conference on Computer Vision (ECCV)*, August 2020.
- [65] Jathushan Rajasegaran, Munawar Hayat, Salman H Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [66] Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. *Advances in neural information processing systems*, 32, 2019.
- [67] Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madiam Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. *arXiv preprint arXiv:2301.12314*, 2023.
- [68] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [69] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *ArXiv*, abs/1606.04671, 2016.
- [70] Murray Shanahan, Christos Kapelanis, and Jovana Mitrović. Encoders and ensembles for task-free continual learning. *arXiv preprint arXiv:2105.13327*, 2021.
- [71] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021.
- [72] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021.
- [73] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.
- [74] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.
- [75] James Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *ArXiv*, abs/2304.06027, 2023.
- [76] James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14994–15004, 2023.
- [77] 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.
- [78] Albin Soutif-Cormerais, Antonio Carta, and Joost Van de Weijer. Improving online continual learning performance and stability with temporal ensembles. In *Conference on Lifelong Learning Agents*, pages 828–845. PMLR, 2023.
- [79] Yuwen Tan, Qinhao Zhou, Xiang Xiang, Ke Wang, Yuchuan Wu, and Yongbin Li. Semantically-shifted incremental adapter-tuning is a continual vittransformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23252–23262, 2024.
- [80] Eli Verwimp, Shai Ben-David, Matthias Bethge, Andrea Cossu, Alexander Gepperth, Tyler L Hayes, Eyke

- Hüllermeier, Christopher Kanan, Dhireesha Kudithipudi, Christoph H Lampert, et al. Continual learning: Applications and the road forward. *arXiv preprint arXiv:2311.11908*, 2023.
- [81] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds 200. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
 - [82] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pages 398–414. Springer, 2022.
 - [83] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.
 - [84] 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.
 - [85] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An ocam’s razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022.
 - [86] Zifeng Wang, Tong Jian, Kaushik Chowdhury, Yanzhi Wang, Jennifer Dy, and Stratis Ioannidis. Learn-prune-share for lifelong learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 641–650. IEEE, 2020.
 - [87] Zhen Wang, Liu Liu, Yajing Kong, Jiaxian Guo, and Dacheng Tao. Online continual learning with contrastive vision transformer. In *European Conference on Computer Vision*, pages 631–650. Springer, 2022.
 - [88] 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. *European Conference on Computer Vision*, 2022.
 - [89] 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.
 - [90] Martin Wistuba, Prabhu Teja Sivaprasad, Lukas Balles, and Giovanni Zappella. Continual learning with low rank adaptation. In *NeurIPS 2023 Workshop on Distribution Shifts (DistShifts)*, 2023.
 - [91] Hyun woo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. *ArXiv*, abs/2110.10031, 2021.
 - [92] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019.
 - [93] Ziyang Wu, Christina Baek, Chong You, and Yi Ma. Incremental learning via rate reduction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1133, 2020.
 - [94] Mengqi Xue, Haofei Zhang, Jie Song, and Mingli Song. Meta-attention for vit-backed continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 150–159, June 2022.
 - [95] Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *ArXiv*, abs/2308.13111, 2023.
 - [96] Fei Ye and Adrian G Bors. Online task-free continual generative and discriminative learning via dynamic cluster memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26202–26212, 2024.
 - [97] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.
 - [98] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.
 - [99] Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. Task agnostic continual learning using online variational bayes. *arXiv preprint arXiv:1803.10123*, 2018.
 - [100] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023.
 - [101] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. In *Advances in Neural Information Processing Systems*, volume 35, pages 14771–14783, 2022.
 - [102] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023.
 - [103] Yitao Zhu, Zhenrong Shen, Zihao Zhao, Sheng Wang, Xin Wang, Xiangyu Zhao, Dinggang Shen, and Qian Wang. Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis, 2023.