This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

BiLoRA: Almost-orthogonal Parameter Spaces for Continual Learning

Hao Zhu[1,*] Yifei Zhang[2,*] Junhao Dong[2] Piotr Koniusz[1,3,◇]
[1]Data61♥CSIRO [2]Nanyang Technological University [3]Australian National University
allenhaozhu@gmail.com, {yifei.zhang,junhao003}@ntu.edu.sg,
piotr.koniusz@data61.csiro.au

Abstract

Continual learning requires models to learn tasks sequentially while maintaining a delicate balance between stability (retaining knowledge of previous tasks) and plasticity (adapting to new tasks). A key challenge is preventing interference between tasks which degrades performance when learning new tasks over previously learned tasks. Recent approaches leverage parameter-efficient fine-tuning (PEFT) which adapts pre-trained models by injecting a small number of learnable parameters. However, existing PEFT-based continual learning methods such as InfLoRA face fundamental limitations, i.e., they rely on complex optimization procedures to learn orthogonal task-specific spaces which is increasingly difficult as tasks accumulate. Thus, we propose a novel bilinear reformulation that fundamentally reimagines the task separation through fixed orthogonal bases. Our key insight is that by expanding the parameter space quadratically through two fixed bases, we can achieve "almost orthogonal" task subspaces probabilistically, eliminating the need for explicit interference elimination procedures. We provide theoretical guarantees that this approach reduces the probability of task interference from $\mathcal{O}\big((k/d)^2\big)$ to $\mathcal{O}\big((k/d^2)^2\big)$, ensuring reliable task separation without complex optimization. Through extensive experiments on ImageNet-R, CIFAR-100, and DomainNet, we validate our theoretical bounds and demonstrate state-of-the-art performance with reduced parameter count. The code is available at: https://github.com/yifeiacc/BiLoRA.

1. Introduction

Continual learning, where models must learn tasks sequentially while preserving knowledge of previous tasks, remains a fundamental challenge in machine learning [26, 27, 32]. The key difficulty lies in maintaining a delicate balance between stability (retaining previous knowledge) and plasticity (adapting to new tasks). Traditional approaches often suffer from the so-called catastrophic forgetting [24, 29] where learning new tasks significantly degrades performance on previously learned tasks [1, 6, 7, 16, 19, 36, 40].

Recent advances in large-scale pre-trained models [4, 5, 11, 12] open new possibilities for addressing this challenge. However, adapting these foundational models efficiently while preventing interference between tasks remains difficult. Parameter-efficient fine-tuning (PEFT) emerged as a promising direction, particularly Low-Rank Adaptation (LoRA) [15] which introduces lightweight, learnable components alongside frozen pre-trained weights.

The application of PEFT methods to continual learning has shown initial success [38, 39], offering both parameter efficiency and continuous adaptation capabilities. However, current PEFT-based approaches face significant challenges in balancing stability and plasticity [37]. The fundamental issue lies in the shared parameter space where task-specific updates become entangled, making it impossible to modify one task's parameters without affecting other tasks [39].

Recently, InfLoRA [22] attempted to address such an interference by enforcing orthogonality between task-specific matrices. While this approach provides task separation guarantees, it faces two critical limitations: (1) The number of tasks is strictly bounded by the dimensionality of the parameter space - with model dimension d and rank r, only $\lfloor d/r \rfloor$ tasks can be supported. (2) Early tasks may claim key parameter subspaces, leaving later tasks with potentially suboptimal allocations that cannot be adjusted within a fixed-dimensional parameter space [8, 34]. These limitations stem from attempting to maintain perfect orthogonality within a fixed-dimensional parameter space.

We present a fundamentally different approach to task separation that leverages fixed orthogonal bases through a novel bilinear reformulation [35]. Our key insight is that while finding perfectly orthogonal subspaces becomes increasingly difficult with more tasks, we can achieve "almost orthogonal" spaces with high probability by expanding the parameter space quadratically (as shown in Figure 1). Specifically, by reformulating task adaptation through two fixed orthogonal bases, we can theoretically guarantee that

*Equal contribution. ◇Corresponding author.

the probability of significant interference between any two tasks becomes vanishingly small. This probabilistic guarantee of almost-orthogonal bases eliminates the need for explicit orthogonalization procedures while ensuring robust task separation. Moreover, our approach naturally maps to the frequency domain, where the Fourier basis provides both inherent structure for the task separation and efficient computation through FFT.

Our main contributions are listed below:

i. We introduce BiLoRA, a novel bilinear reformulation that quadratically expands the parameter space while supporting both perfect and almost-orthogonal task separation through fixed orthogonal bases.

ii. We provide theoretical guarantees for two complementary strategies: (1) perfect orthogonality supporting $\mathcal{O}\big(d^2/k\big)$ tasks via a without-replacement selection, and (2) almost-orthogonal spaces with interference probability $\mathcal{O}\big((k/d^2)^2\big)$ via a with-replacement selection – both significantly improving upon the $\mathcal{O}(d/r)$ limit of existing methods.

iii. We demonstrate that the Fourier basis provides an optimal implementation of our framework, offering both theoretical guarantees and practical efficiency while supporting both separation strategies.

iv. Through extensive experiments on ImageNet-R, CIFAR-100, and DomainNet, we validate both separation strategies and achieve the state-of-the-art performance in PEFT-based continual learning setting.

2. Related Work

Parameter-Efficient Model Adaptation. Recent advances in parameter-efficient fine-tuning (PEFT) have demonstrated promising results in adapting large pre-trained models [41, 46]. Additive approaches such as Adapter [14], AdapterFormer [3] insert trainable modules into pre-trained networks. While effective, these approaches increase model complexity and cannot maintain task separation. Input-based methods such as Prompt-tuning [21] and VPT [17] modify input representations via learnable tokens but their expressive power is limited to manipulating the input space. Another line of work focuses on weight decomposition, exemplified by LoRA [15], which decomposes weight updates into low-rank matrices. While achieving parameter efficiency, LoRA and its variants such as PACE [25] suffer from task interference in continual learning, as each sequential adaptation shares and modifies the same parameter space. In contrast, our method introduces a novel bilinear decomposition in the frequency domain and provides natural task separation without architectural modifications or input manipulations.

Task Separation Mechanisms in Continual Learning. The prevention of catastrophic forgetting via effective task separation remains a central challenge in continual learning. Parameter allocation strategies, such as PackNet [23] and HAT [31], assign dedicated parameters to each task but face scalability issues with increasing tasks. Gradient projection approaches, including OWM [42] and GPM [30], project gradients to orthogonal spaces. InfLoRA [22] extends this concept to PEFT but requires complex gradient manipulation. Architecture-based methods such as DualPrompt [38] and L2P [39] employ specialized architectures for task separation, increasing design complexity. Our frequency-domain approach differs fundamentally by providing inherent task separation through orthogonal basis functions, avoiding the computational overhead of gradient projection while maintaining mathematical elegance.

Parameter-Efficient Continual Learning. Managing computational and memory resources efficiently remains a critical challenge in continual learning. Traditional approaches such as [2, 27, 32, 47] update all parameters, requiring substantial computational resources and storage. Progressive expansion methods, *e.g.*, CODA-P [34], add parameters as new tasks arrive, leading to linear growth in model size. Noteworthy are also efforts in open-world object detection, *e.g.*, OW-DETR [10] & OWOBJ [45]. Memory-based methods use experience replay or knowledge distillation and store additional data/model copies. In contrast, our frequency-domain bilinear decomposition achieves both parameter and computational efficiency via structured sparsity in the frequency domain. It enjoys constant memory footprint (regardless of task number) and fast computation via FFT.

3. Preliminaries

Parameter-efficient fine-tuning via low-rank adaptation (LoRA) is a popular approach for adapting pre-trained models to downstream tasks. In this section, we first review the foundations of LoRA and formalize the challenges in applying it to continual learning.

Low-Rank Adaptation. Given a pre-trained model with weight matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$, LoRA [15] introduces learnable parameters to adapt the model via:

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + \mathbf{A}\mathbf{B}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$ are low-rank matrices with rank $r \ll d$. This decomposition significantly reduces the number of trainable parameters from $\mathcal{O}(d^2)$ to $\mathcal{O}(rd)$.

Continual Learning with LoRA. In the continual learning setting, we encounter a sequence of tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$ that must be learned sequentially. When applying LoRA to continual learning, each task requires its own adaptation parameters for a linear layer subsuming $\{1, \ldots, t\}$ tasks to-date:

$$\mathbf{W}_t = \mathbf{W} + \sum_{i=1}^{t} \mathbf{A}_i\mathbf{B}_i, \tag{2}$$

where $\mathbf{A}_i$ and $\mathbf{B}_i$ are the task-specific parameters for task i.

Existing Solution. InfLoRA [22] addresses this challenge by ensuring orthogonality in the $\mathbf{B}$ matrices through gradient-guided subspace allocation:

$$\text{span}(\mathbf{B}_i) \perp \text{span}(\mathbf{B}_j), \quad \forall i \neq j. \tag{3}$$

This approach first designs orthogonal $\mathbf{B}$ matrices via the gradient space projection, then learns corresponding $\mathbf{A}$ matrices. However, InfLoRA faces two fundamental limitations. The sequential nature of subspace allocation leads to potentially suboptimal basis assignments for early tasks, with no mechanism to adjust these allocations based on the requirements of later tasks. Moreover, a strict dimensionality constraint emerges from the orthogonality requirement: with each task requiring r orthogonal dimensions in a d-dimensional parameter space, the maximum number of supported tasks is bounded by $T \leq \lfloor d/r \rfloor$. This creates an inherent tension between the number of tasks that can be accommodated, the expressiveness available to each task through rank r, and the underlying model dimension d.

4. Methodology

While InfLoRA successfully addresses catastrophic forgetting in continual learning by the orthogonal parameter allocation, our analysis reveals its two significant limitations. The first limitation is a strict capacity constraint - when using a model with dimension $d = 1024$ and rank $r = 32$, InfLoRA can only accommodate 32 tasks before exhausting its available orthogonal subspaces. The second limitation stems from InfLoRA's sequential allocation strategy: as tasks arrive sequentially, early tasks may allocate parameter subspaces that would have been more beneficial for later tasks. However, these allocations cannot be adjusted at a later time.

Our solution develops three key stages: (i) We introduce fixed orthogonal bases as a replacement for learned orthogonal projections, ensuring each task has equal access to the entire parameter space. Building on this foundation, (ii) we develop a bilinear reformulation that expands the available parameter space from d to d^2 dimensions, dramatically increasing the number of tasks we can support. (iii) We then implement this approach using Fourier transforms, which provides both theoretical guarantees for task separation and computational efficiency.

Our theoretical analysis demonstrates that this approach achieves near-perfect task separation while supporting quadratically more tasks than previous methods.

4.1. Bilinear-LoRA (BiLoRA)

Intuition and Motivation. Instead of learning orthogonal bases for each task, we propose to utilize existing orthogonal bases (*e.g.*, DFT, DCT) to partition the parameter space. In fixed orthogonal basis approaches, a task's weight update can be represented as:

$$\Delta\mathbf{W}_t = \mathbf{U}\mathbf{B}_t, \tag{4}$$

where $\mathbf{U} \in \mathbb{R}^{d\times d}$ is a fixed orthogonal basis and $\mathbf{B}_t \in \mathbb{R}^{d\times d}$ has only r non-zero rows, indicating which basis vectors are selected. This fixed basis approach addresses the sequential allocation issue in previous methods by providing a fair, predetermined subspace allocation strategy. Unlike learned orthogonal bases where early tasks might monopolize crucial subspaces, each task now has equal access to the entire basis set. However, the fundamental limitation remains: in a d-dimensional space, we can only support $\lfloor d/r \rfloor$ tasks without interference, regardless of the allocation strategy. This inherent capacity limit motivates our bilinear extension, which expands the available parameter space quadratically.

Bilinear Reformulation. This single-sided projection in Eq. (4) limits the expressiveness and capacity for task separation. To address this, we propose a bilinear decomposition:

$$\Delta\mathbf{W}_t = \mathbf{U}\mathbf{B}_t\mathbf{V}^T \tag{5}$$

where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d\times d}$ are fixed orthogonal bases. This bilinear extension provides two key advantages. While the linear model operates in a d-dimensional space, our bilinear model naturally expands the parameter space to d^2 dimensions due to the interaction of the two basis transformations. For unit-norm inputs $\boldsymbol{x}_i$ and $\boldsymbol{x}_t$, their interaction is:

$$\begin{aligned}
\langle \boldsymbol{y}_i, \boldsymbol{y}_t \rangle &= \langle \boldsymbol{U}\boldsymbol{B}_i\boldsymbol{V}^T\boldsymbol{x}_i, \boldsymbol{U}\boldsymbol{B}_t\boldsymbol{V}^T\boldsymbol{x}_t \rangle && (6)\\
&= \boldsymbol{x}_i^T\boldsymbol{V}\boldsymbol{B}_i^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{B}_t\boldsymbol{V}^T\boldsymbol{x}_t && (7)\\
&= \boldsymbol{x}_i^T\boldsymbol{V}\boldsymbol{B}_i^T\boldsymbol{B}_t\boldsymbol{V}^T\boldsymbol{x}_t \ \ (\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}) && (8)\\
&\leq \|\boldsymbol{B}_i^T\boldsymbol{B}_t\|_F \text{ (by Cauchy-Schwarz).} && (9)
\end{aligned}$$

This shows how all three conditions work together to maintain task independence: orthonormal $\boldsymbol{U}$ preserves output separation, orthonormal $\boldsymbol{V}$ preserves input transformation, and orthogonal $\boldsymbol{B}$ matrices ensure task parameter separation. When $\boldsymbol{B}_i$ and $\boldsymbol{B}_t$ are orthogonal, the Frobenius norm $\|\boldsymbol{B}_i^T\boldsymbol{B}_t\|_F$ becomes zero, guaranteeing task separation regardless of inputs. This quadratic expansion dramatically reduces the interference probability from $\mathcal{O}\big((k/d)^2\big)$ for the linear case to $\mathcal{O}\big((k/d^2)^2\big)$ for our bilinear case that supports significantly more tasks and maintains their independence.

4.2. Task Separation with DFT

Building upon the bilinear reformulation, we show that parameter-efficient finetuning with Discrete Fourier Transform (DFT) [9] naturally fits into our bilinear framework while providing enhanced task separation via the frequency domain orthogonality. Rather than learning the basis matrices $\mathbf{U}$ and $\mathbf{V}$, which would substantially increase the parameter count, we employ fixed orthogonal bases ($\mathbf{F}$) through DFT:

$$\Delta\mathbf{W}_t = \mathbf{F}\mathbf{B}_t\mathbf{F}^H, \tag{10}$$

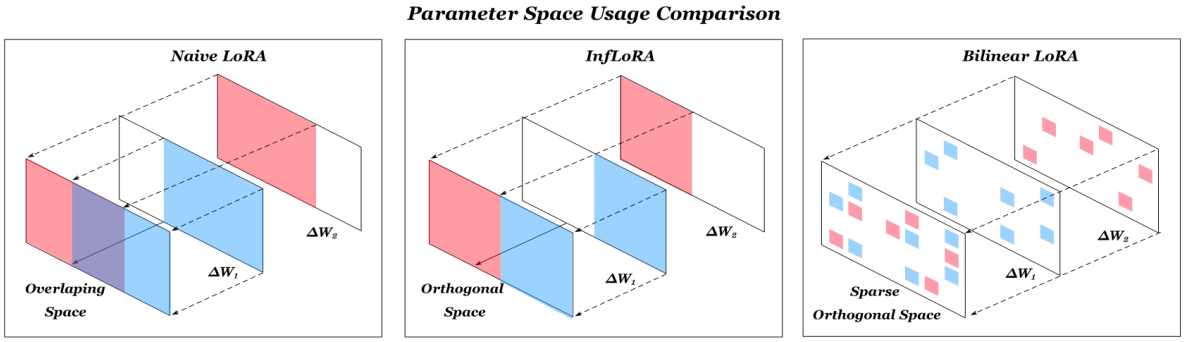


Figure 1. Comparisons of the parameter space utilization across different LoRA variants. Traditional LoRA allows task parameters to overlap. InfLoRA enforces a strict orthogonality through fixed subspace allocation. Our proposed Bilinear-LoRA leverages task separation while maintaining independence.

where $\mathbf{F} \in \mathbb{C}^{d\times d}$ is the 1D DFT matrix, $\mathbf{F}^H$ is its Hermitian transpose, and $\mathbf{B}_t \in \mathbb{C}^{d\times d}$ contains the learnable frequency components for task t. This formulation is a special case of our bilinear framework with $\mathbf{U} = \mathbf{F}, \mathbf{V} = \mathbf{F}$, with the additional advantage that $\mathbf{F}^H = \mathbf{V}^H = \mathbf{V}^T$ due to the unitary nature of the DFT matrix. The 2D FFT of a matrix $\Delta\mathbf{W}$ can be computed by applying 1D FFT to its rows followed by 1D FFT on the columns, which is precisely what our formulation $\mathbf{F}\mathbf{B}_t\mathbf{F}^H$ achieves, enabling efficient implementation through fast Fourier transform algorithms.

Structured Parameter Space. Working in the frequency domain offers great flexibility for task separation. To ensure our design meets both theoretical and practical requirements, we constrain the task-specific matrices $\mathbf{B}_t$ with two key properties:

(i) We enforce sparsity in the frequency domain:

$$\|\mathbf{B}_t\|_0 = k \ll d^2, \tag{11}$$

where only k frequency components are non-zero. This sparse representation enables two theoretically distinct allocation strategies:

- Without replacement: Enforcing $\text{supp}(\mathbf{B}_t) \cap \text{supp}(\mathbf{B}_i) = \emptyset$ for $i \neq t$ supports $\lfloor d^2/k \rfloor$ tasks with guaranteed zero interference.
- With replacement: Relaxing the separation constraint enables unlimited tasks with collision probability reduced to $\mathcal{O}\big((k/d^2)^2\big)$.

In practice, both strategies achieve nearly identical performance due to the quadratic expansion of the parameter space, with sparsity simultaneously reducing memory requirements and decreasing the probability of task interference.

(ii) To ensure our weight updates remain real-valued while operating in the complex frequency domain, we enforce the Hermitian symmetry:

$$\mathbf{B}_t[u, v] = \mathbf{B}_t[-u, -v]^*, \tag{12}$$

where $u \in \big\{-\frac{d}{2}+1, \ldots, \frac{d}{2}\big\}$ and $v \in \big\{-\frac{d}{2}+1, \ldots, \frac{d}{2}\big\}$ represent the frequency coordinates in the 2D Fourier domain. This symmetry constraint guarantees that after the inverse Fourier transform, the resulting weight update $\Delta\mathbf{W}_t$ will be real-valued, as required for neural network computations.

Learning Procedure. To efficiently implement the above constraints, we represent the sparse frequency matrix $\mathbf{B}_t$ by a binary mask $\mathbf{S}_t$ and corresponding values $\mathbf{\Theta}_t$:

$$\mathbf{B}_t = \mathbf{S}_t \odot \mathbf{\Theta}_t, \tag{13}$$

where $\mathbf{S}_t$ indicates the support (non-zero locations) and $\mathbf{\Theta}_t$ contains the complex values at those locations. This representation essentially treats $\mathbf{B}_t$ as a sparse matrix where $\mathbf{S}_t$ and $\mathbf{\Theta}_t$ serve as index and value tensors, respectively.

Algorithm 1 details the optimization process for each task. The procedure first initializes $\mathbf{\Theta}_t$ to zeros and generates a frequency mask $\mathbf{S}_t$ that satisfies both sparsity $\big(\|\mathbf{S}_t\|_0 = k\big)$ and symmetry $\big(\mathbf{S}_t[u, v] = \mathbf{S}_t[-u, -v]\big)$ constraints. For the "without replacement" strategy, we additionally enforce $\text{supp}(\mathbf{S}_t) \cap \text{supp}(\mathbf{S}_i) = \emptyset, \forall i < t$.

This approach ensures that task learning remains confined to designated frequency components while maintaining the constraints necessary for real-valued weight updates. By operating directly in the frequency domain and exploiting sparsity, we achieve both computational efficiency and effective task separation.

4.3. Theoretical Analysis

Below, we provide theoretical guarantees for our frequency domain task separation model. We first establish probabilistic bounds on task interference through frequency separation.

Theorem 1 (Frequency Separation)**.** *For tasks i and j, let $\mathbf{S}_i$ and $\mathbf{S}_j$ be independently and uniformly sampled frequency selection masks with k components each from a total of d^2 components. For $k \ll d$, the probability of perfect separa-*

Algorithm 1: Fourier Domain Task Learning.

Require: Base model weights $\mathbf{W}$, task dataset $\mathcal{D}_t$, number of frequency components k.
Ensure: Task-specific parameters $\mathbf{\Theta}_t$.

Initialize $\mathbf{\Theta}_t \leftarrow \mathbf{0}^{d\times d}$
Generate frequency mask $\mathbf{S}_t$ satisfying:
$|\mathbf{S}_t|_0 = k$ ▷ Sparsity constraint
$\mathbf{S}_t[u, v] = \mathbf{S}_t[-u, -v]$ ▷ Symmetry constraint
For each training batch $(x, y) \in \mathcal{D}_t$
  $\Delta\mathbf{W}_t \leftarrow \mathbf{F}(\mathbf{S}_t \odot \mathbf{\Theta}_t)\mathbf{F}^H$
  $\hat{y} \leftarrow (\mathbf{W} + \Delta\mathbf{W}_t)x$ ▷ Forward pass
  $\mathcal{L} \leftarrow \text{Loss}(\hat{y}, y)$
  Compute gradients: $\frac{\partial\mathcal{L}}{\partial\mathbf{\Theta}_t}$
  Update only selected components:
  $\mathbf{\Theta}_t \leftarrow \mathbf{\Theta}_t - \eta(\mathbf{S}_t \odot \frac{\partial\mathcal{L}}{\partial\mathbf{\Theta}_t})$
  Enforce Hermitian symmetry:
    $\mathbf{\Theta}_t[u, v] \leftarrow \mathbf{\Theta}_t[-u, -v]^*$
End
return $\mathbf{\Theta}_t$

tion (no frequency collision) is at least:

$$p\big(\mathit{supp}(\mathbf{S}_i) \cap \mathit{supp}(\mathbf{S}_j) = \emptyset\big) \geq 1 - \frac{k^2}{d^2}. \tag{14}$$

The above theorem reveals a fundamental advantage of our bilinear approach: by expanding to a $d \times d$ frequency grid, tasks naturally achieve separation with high probability. Consider a practical example with model dimension $d = 1024$ and $k = 32$ components per task: the probability of perfect separation is at least 99.9%. This is analogous to randomly placing k items in a $d \times d$ grid versus a 1D array of size d – the quadratic expansion in space dramatically reduces the collision probability from $\mathcal{O}(k^2/d)$ to $\mathcal{O}(k^2/d^2)$. This demonstrates that our bilinear framework can support significant scaling in both model capacity and number of tasks while preserving near-perfect separation. Moreover, these theoretical bounds are conservative, *i.e.*, in practice, with our Fourier-based implementation, we observe even better separation due to the structured nature of frequency components. This leads to our main result on the task interference:

Theorem 2 (Task Interference Bound)**.** *For tasks i and j with updates $\Delta\mathbf{W}_i$ and $\Delta\mathbf{W}_j$, with probability at least $1-\delta$:*

$$\left\|\Delta\mathbf{W}_i^T \Delta\mathbf{W}_j\right\|_F \leq \varepsilon, \tag{15}$$

when the number of frequency components k satisfies:

$$k \leq c\frac{d^2}{T}\log\left(\frac{1}{\delta}\right), \tag{16}$$

where c is a universal constant and T is the total number of tasks.

The above theorem translates our collision probability into the practical guidance for implementing our method. The bound in Eq. (16) tells us exactly how many frequency components we can allocate per task while keeping interference under control. This enables us to realize a budget – as one adds more tasks (T increases), one needs to be more conservative with the per-task allocations (k decreases). The $\log(1/\delta)$ term provides a safety margin, *i.e.*, by making it larger, one can be more confident about avoiding the interference at the cost of slightly reduced capacity. Most importantly, the above result shows we can handle $\mathcal{O}(d^2)$ tasks – quadratically more than traditional approaches – while still maintaining the task independence with high probability.

5. Experiments

In this section, we conduct extensive experiments to evaluate our method. Firstly, we describe the experimental settings, then compare our method with state-of-the-art baselines on various benchmarks. Finally, we provide detailed ablation studies and analysis.

Continual Learning Setting. We evaluate our approach in a class-incremental learning setting where the model must classify all previously seen classes without providing the task identity during inference. For all datasets, we follow the protocol of Gao *et al*. [8] where classes are partitioned into disjoint subsets representing sequential tasks. For the classification layer, we employ a dynamic expansion approach, adding new output nodes for each task's classes while freezing weights corresponding to previously learned classes. This prevents the new task from directly interfering with classification boundaries of previous tasks.

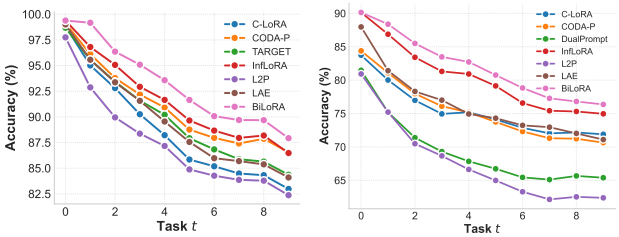
Datasets. We evaluate our method on ImageNet-R [13], CIFAR-100 [20], and DomainNet [28]. Following the protocols established in [8], we maintain consistent incremental task settings for fair comparison with baseline continual learning (CL) methods. All experiments are conducted 5 times with different random seeds, and we report both the final task performance and averaged results across all tasks.

Baselines. We compare our approach against several PEFT-based CL methods: L2P [39], DualPrompt [38], and CODA-P [34]. We also evaluate against recent advances including C-LoRA [33] and LAE [8]. Additionally, we establish performance bounds using two non-CL approaches: a *sequential* baseline (lower bound) that employs naive sequential training in a class-incremental setting, and a *joint* training approach (upper bound) that trains all classes simultaneously. For fair comparison, all methods are implemented using the same pre-trained model and fine-tuned using LoRA.

Implementation Details. Our model employs ViT-B/16 [5] as the backbone, which is pre-trained using self-supervision and widely adopted in class-incremental learning (CIL). Following standard practice, we insert adapters only in the query

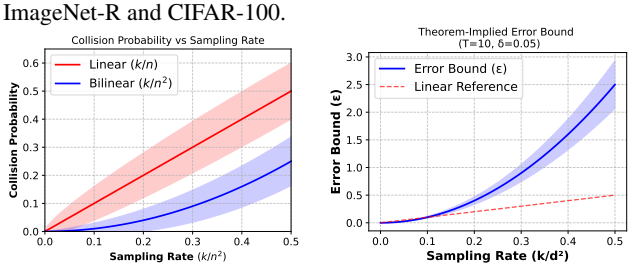
Table 1. Results (%) on ImageNet-R, CIFAR-100 and DomainNet. We report results over 5 trials.

Tasks	ImageNet-R (10 Task)		CIFAR-100 (10 Task)		DomainNet (10 Task)	
Method	**Last** (↑)	Avg. (↑)	**Last** (↑)	Avg. (↑)	**Last** (↑)	**Avg.** (↑)
Joint Training	81.14 ± 0.34	–	91.920.05	-	77.720.04	-
Sequential	62.54 ± 0.24	67.98 ± 0.27	62.18 ± 3.59	80.42 ± 0.23	54.34 ± 1.24	70.35 ± 0.21
L2P [39]	65.41 ± 0.52	69.39 ± 0.43	82.48 ± 0.20	87.64 ± 0.25	71.56 ± 0.06	76.49 ± 0.02
DualPrompt [38]	71.47 ± 0.35	75.82 ± 0.29	84.42 ± 0.30	90.06 ± 0.07	74.64 ± 0.06	78.57 ± 0.03
CODA-P [34]	71.70 ± 0.39	76.71 ± 0.10	86.62 ± 0.11	91.08 ± 0.28	74.83 ± 0.15	79.83 ± 0.07
C-LoRA [33]	71.89 ± 0.45	75.33 ± 0.28	82.97 ± 0.47	88.81 ± 0.34	70.34 ± 0.15	76.26 ± 0.15
LAE [8]	71.70 ± 0.39	76.71 ± 0.10	84.15 ± 0.10	89.84 ± 0.03	67.23 ± 0.42	76.76 ± 0.17
InfLoRA [22]	75.65 ± 0.19	80.82 ± 0.12	86.51 ± 0.73	91.70 ± 0.32	75.45 ± 0.22	80.57 ± 0.57
`Ours`	**77.95** ± 0.14	**81.52** ± 0.26	**87.46** ± 0.76	**92.50** ± 0.62	**76.56** ± 0.41	**81.65** ± 0.62



(a) CIFAR-100 (b) ImageNet-R

Figure 2. The performance of different methods during learning on ImageNet-R and CIFAR-100.



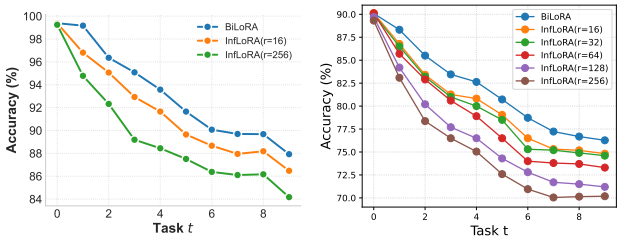
(a) Collision Probability (b) Task Interference

Figure 3. Analysis of task separation in parameter space. (a) Comparison between linear (k/d) and bilinear (k/d^2) collision probabilities as function of sampling rate, with $\pm 2\sigma$ confidence intervals showing the consistent advantage of bilinear allocation. (b) The resulting interference bound $\|\Delta \mathbf{W}_i^T \Delta \mathbf{W}_j\|_F$ (95% conf. interval), validating Theorem 2's prediction that maintaining $k \leq c(d^2/T)\log(1/\delta)$ ensures bounded interference between tasks.

and key linear layers of the transformer's attention blocks. For our method (`Ours`), we set the frequency parameter $k = 3000$, and maintain a LoRA rank of $r = 16$ across all LoRA-based methods. The models are optimized using Adam [18] with a batch size of 64. Following Zhang *et al.* [44], we employ different learning rates, *i.e.*, $1e^{-3}$ for the classification layer and $1e^{-5}$ for adapter parameters. We additionally implement cosine annealing during training. The hyperparameters are set as follows: $\delta = 1$, $\lambda = 0.001$, $\gamma = 0.5$, and $\eta = 0.2$.

5.1. Performance Evaluation

Overall Performance. Our method achieves state-of-the-art results across all three benchmark datasets (Table 1). On ImageNet-R, we achieve 77.95% final accuracy and 81.52% average accuracy, outperforming previous best results by 2.3%. The improvement is particularly notable on CIFAR-100, where our method approaches the joint training performance (91.92%) with 87.46% final accuracy while maintaining the ability to learn sequentially. For DomainNet, which tests adaptation across diverse visual domains, we observe consistent improvements with 76.56% final accuracy and 81.65% average accuracy. These results demonstrate that our frequency-based task separation effectively balances plasticity and stability in continual learning.



(a) CIFAR-100 (b) ImageNet-R

Figure 4. Investigation of different r for InfLoRA in terms of Accuracy score on CIFAR-100 and ImageNet-R.

Scaling to More Tasks. A key advantage of our approach is its ability to handle an increasing number of tasks effectively (Table 2). With 5 tasks, we maintain performance close to the joint training, achieving 78.32% final accuracy. More importantly, as we scale to 10 and 20 tasks, our method enjoys significantly less performance degradation compared to baselines. Even with 20 tasks, we maintain 72.41% final accuracy and 79.28% average accuracy, demonstrating that our bilinear formulation successfully addresses the task capacity limitations of previous approaches.

Anti-catastrophic Forgetting Analysis. Figure 2 demonstrates how different methods maintain performance on previously learned tasks as new tasks are introduced. On both ImageNet-R and CIFAR-100, our method exhibits significantly better resistance to catastrophic forgetting. While baseline methods show noticeable accuracy drops after learning each new task, our approach maintains a more stable performance trajectory throughout the learning process. This

Table 2. Results (%) for 5 tasks, 10 tasks, and 20 tasks on ImageNet-R. We report results averaged over 5 trials.

Tasks	5 Task		10 Task		20 Task	
Method	**Last** (↑)	**Average** (↑)	**Last** (↑)	**Average** (↑)	**Last**(↑)	**Average** (↑)
Joint Training	81.14 ± 0.34	-	81.14 ± 0.34	-	81.14 ± 0.34	-
Sequential	58.74 ± 1.28	72.91 ± 0.28	62.54 ± 0.24	67.98 ± 0.27	34.62 ± 0.85	51.15 ± 1.50
L2P [39]	64.13 ± 0.78	68.66 ± 0.41	65.41 ± 0.52	69.39 ± 0.43	57.92 ± 0.28	64.57 ± 0.29
DualPrompt [38]	67.88 ± 0.17	71.16 ± 0.31	71.47 ± 0.35	75.82 ± 0.29	61.00 ± 0.72	65.80 ± 0.67
CODA-P [34]	73.09 ± 0.21	76.91 ± 0.21	71.70 ± 0.39	76.71 ± 0.10	67.28 ± 0.30	72.34 ± 0.17
C-LoRA [33]	75.85 ± 0.31	78.85 ± 0.34	71.89 ± 0.45	75.33 ± 0.28	65.71 ± 0.60	70.63 ± 0.85
LAE [8]	73.84 ± 0.14	77.29 ± 0.45	71.70 ± 0.39	76.71 ± 0.10	66.98 ± 0.35	73.72 ± 0.05
InfLoRA [22]	77.52 ± 0.37	82.01 ± 0.12	75.65 ± 0.19	80.82 ± 0.12	71.01 ± 0.45	77.28 ± 0.45
`Ours`	**78.32** ± 0.37	**83.31** ± 0.52	**77.95** ± 0.14	**81.52** ± 0.26	**72.41** ± 0.65	**79.28** ± 0.76

stability can be attributed to our frequency-domain task separation mechanism: by allocating distinct frequency components to each task, we effectively prevent new task updates from interfering with previously learned features. The consistent performance gap between our method and existing approaches grows wider as more tasks are learned, highlighting the long-term benefits of our bilinear formulation in preserving task-specific knowledge.

Understanding Capacity Limitations of InfLoRA. Figure 4 provides a compelling evidence for why previous orthogonal subspace approaches struggle with scaling. For CIFAR-100, we compare InfLoRA's performance using standard rank ($r = 16$) versus high rank ($r = 256$) settings. For ImageNet-R, we conduct a more extensive comparison across multiple ranks. While a higher rank theoretically provides more expressivity per task [43], it significantly reduces the number of available orthogonal subspaces $\lfloor \frac{d}{r} \rfloor$. The results clearly show a trade-off: as the rank increases, the performance degrades more rapidly as the tasks accumulate, despite having more parameters per task. This empirically validates our theoretical argument about the fundamental limitations of the orthogonal subspace allocation, demonstrating why our bilinear frequency-domain approach is necessary.

5.2. Empirical Analysis

To validate our theoretical analysis, we conduct simulation experiments examining both the collision probability in the frequency space and the resulting task interference. These experiments provide empirical evidence for the advantages of our bilinear formulation and validate the theoretical bounds.

Collision Analysis. We first analyze the probability of frequency collisions between tasks as we vary the sampling rate k/n. Figure 3a compares the collision probability between the traditional linear allocation (k/n) and our bilinear approach (k/n^2). The results show that our bilinear strategy achieves quadratically lower collision rates, *e.g.*, at a sampling rate of 0.2, the linear model shows a collision probability of 0.2, while our bilinear model reduces this to 0.04. This significant reduction supports more tasks while maintaining independence. The confidence intervals, computed through binomial statistics, confirm the statistical significance of this improvement across different sampling rates.

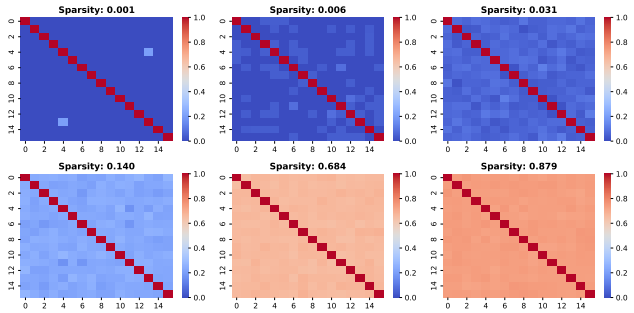


Figure 5. Visualization of cross-correlation between 16 different parameter updates $\Delta\mathbf{W}^t$ at varying sparsity levels. Each matrix element represents the inner product $\langle \Delta\mathbf{W}^t, \Delta\mathbf{W}^{t'} \rangle$ between pairs of updates. Lower values (darker colors) indicate a "stronger" orthogonality between updates $\Delta\mathbf{W}^t$, demonstrating how sparsity in frequency assignment maintains orthogonal subspaces for different tasks.

Task Interference. To verify Theorem 2, we measure the actual interference between tasks under different sampling rates. Figure 3b shows the Frobenius norm of the interference $\|\Delta\mathbf{W}_i^T \Delta\mathbf{W}_j\|_F$ as the sampling rate k/d^2 increases. The experimental results closely match our theoretical predictions, with the interference growing quadratically with the sampling rate. The shaded confidence intervals (95% confidence level) demonstrate the consistency of this relationship. Notably, when we maintain $k \leq c(d/T)\log(1/\delta)$ as specified by Theorem 2, the interference remains bounded by ϵ, validating our theoretical guarantees. To directly visualize the effectiveness of our frequency-domain task separation, Figure 5 shows the cross-correlation matrix between parameter updates ($\Delta\mathbf{W}^t$) across 15 different tasks. The predominantly dark off-diagonal elements indicate the minimal interference between different tasks' updates, validating that our frequency allocation strategy successfully maintains near-orthogonal subspaces.

5.3. Ablation Analysis

How Many Frequency Components Are Needed? Table 5 reveals an interesting finding: our method achieves the

optimal performance using only 3.2‰ of the available frequency components. Increasing the number of components beyond this point provides diminishing returns, suggesting that an efficient sweet spot for parameter allocation has been found. This result is consistent across all datasets, indicating that the frequency-domain task separation can be highly parameter-efficient while maintaining strong performance.

Robustness to Pre-trained Models. An important practical question is whether our method depends on specific pre-training approaches. Table 3 addresses this by comparing performance across different pre-trained models (DINO-1k and iBOT-1k). The results show that our method maintains its advantages regardless of the pre-training method, with consistent improvements over baselines. This robustness is particularly important for practical applications where different pre-trained models might be available.

Impact of Frequency Allocation Strategy. Below we ask "How should we allocate frequency components to tasks, and how does this affect scalability?" Table 4 compares three strategies: *shared allocation*, *random sampling with replacement*, and *random sampling without replacement*. While random sampling without replacement guarantees no collisions, it limits the maximum number of tasks to $\lfloor d^2/k \rfloor$. In contrast, random sampling with replacement can theoretically support unlimited tasks while achieving nearly identical performance – a crucial advantage for real-world applications. This negligible performance difference is explained by our bilinear formulation: with $k \ll d^2$, the collision probability $\left(\mathcal{O}\left((k/d^2)^2\right)\right)$ becomes so small that occasional frequency overlaps hardly impact performance. These results show that sampling with replacement offers the best of both worlds, *i.e.*, unlimited task scalability and strong performance – making it the preferred choice for practical applications. This finding distinguishes our approach from prior methods such as InfLoRA, which face strict limits on the task capacity due to the orthogonality constraints.

These results collectively demonstrate that our frequency-domain approach not only provides theoretical advantages but also translates to practical improvements in continual learning. The method is robust across different settings, efficient in parameter usage, and effectively addresses the limitations of previous approaches.

6. Conclusions

This work addresses a fundamental question in parameter-efficient continual learning: how to achieve a reliable task separation without the burden of learning orthogonal spaces. Our analysis reveals that the traditional pursuit of perfect orthogonality, while mathematically elegant, becomes increasingly impractical as task numbers grow. Instead, we demonstrate that "almost orthogonality" through bilinear extension provides a more scalable solution with provable guarantees. Our findings have profound implications for both theoretical understanding and practical applications of continual learning. The probabilistic analysis reveals that expanding the parameter space can be more effective than enforcing strict constraints, challenging the conventional wisdom in task separation. This insight, combined with the natural structure of fixed bases such as the Fourier transform, replaces the complex challenge of learning orthogonal spaces with an intuitive frequency allocation scheme. Looking forward, this work suggests a paradigm shift from the pursuit of perfect preservation to probabilistic guarantees in continual learning. This new perspective not only simplifies the implementation of continual learning systems but also provides clearer paths for adaptation to specific domain requirements. While our current focus has been on LoRA-based architectures, the principles of "almost orthogonality" and the bilinear extension could revolutionize how we approach parameter efficiency in deep learning more broadly.

Table 3. Results (%) of different methods on ImageNet-R (10 tasks) using various self-supervised pre-trained models. DINO-1k and iBOT-1k indicate that the ViT was pre-trained on ImageNet-1k using these respective methods.

	Method	**Last** (↑)	**Average** (↑)
DINO-1k	L2P [39]	56.71 ± 0.12	63.59 ± 0.21
	DualPrompt [38]	60.23 ± 0.42	66.57 ± 0.25
	CODA-P [34]	64.02 ± 0.68	71.50 ± 0.42
	C-LoRA [33]	63.07 ± 0.36	68.09 ± 0.41
	LAE [8]	61.03 ± 0.27	69.89 ± 0.15
	InfLoRA [22]	68.31 ± 0.28	76.15 ± 0.05
	InfLoRA-b5 [22]	66.16 ± 0.14	73.01 ± 0.17
iBOT-1k	L2P [39]	60.80 ± 0.35	66.58 ± 0.28
	DualPrompt [38]	63.78 ± 0.38	68.88 ± 0.16
	CODA-P [34]	68.02 ± 0.48	74.28 ± 0.47
	C-LoRA [33]	68.60 ± 0.07	73.47 ± 0.28
	LAE [8]	64.14 ± 0.29	72.59 ± 0.22
	InfLoRA [22]	71.84 ± 0.09	78.29 ± 0.09
	`Ours`	**73.72** ± 0.44	**79.91** ± 0.13

Table 4. Shared *vs.* non-shared frequency coordinates by sampling with or without replacement. We report the **the Last and Average Accuracy**.

Methods	**CIFAR-100**		**ImageNet-R**		**DomainNet**	
	Last (↑)	**Avg.** (↑)	**Last** (↑)	**Avg.** (↑)	**Last** (↑)	**Avg.** (↑)
`Ours` + Shared	74.42	77.52	84.46	89.15	74.15	78.97
`Ours` + Random	75.65	81.52	87.46	92.50	76.56	81.65
`Ours` + Random W/O R	75.72	81.65	87.48	92.60	76.74	81.94

Table 5. Performance comparison under different frequency sparsity levels.

Indicators	**Average** (↑)					
# Freq. k	1000	3000	6000	12000	24000	48000
Sparsity k/d^2	1.6‰	3.2‰	4.8‰	5.4‰	8.0‰	9.6‰
Euqal rank r	1	2	4	8	16	32
CIFAR-100	85.2	87.46	87.30	87.30	87.30	87.30
ImageNet-R	79.7	81.52	81.48	81.48	81.48	81.48
DomainNet	75.1	76.56	76.23	76.23	76.23	76.23

Acknowledgment

This work is funded by the CSIRO Science Digital (SD) and Advanced Engineering Biology Future Science Platforms (AEB-FSP).

References

[1] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *ICCV*, pages 844–853, 2021.

[2] Matteo Boschini, Lorenzo Bonicelli, Angelo Porrello, Giovanni Bellitto, Matteo Pennisi, Simone Palazzo, Concetto Spampinato, and Simone Calderara. Transfer without forgetting. In *Proceedings of the European Conference on Computer Vision*, pages 692–709, 2022.

[3] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, pages 16664–16678, 2022.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[5] 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. In *International Conference on Learning Representations*, 2021.

[6] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, pages 86–102, 2020.

[7] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, pages 128–135, 1999.

[8] 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 11449–11459, 2023.

[9] Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete fourier transform. In *Forty-first International Conference on Machine Learning*, 2024.

[10] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *CVPR*, 2022.

[11] Maryam Haghighat, Peyman Moghadam, Shaheer Mohamed, and Piotr Koniusz. Pre-training with random orthogonal projection image modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

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

[14] 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 *Proceedings of the International Conference on Machine Learning*, pages 2790–2799, 2019.

[15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[16] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *CVPR*, 2021.

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, pages 709–727, 2022.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] 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*, pages 3521–3526, 2017.

[20] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[22] 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.

[23] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

[24] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989.

[25] Yao Ni, Shan Zhang, and Piotr Koniusz. PACE: marrying the generalization of PArameter-efficient fine-tuning with consistency regularization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[26] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with

neural networks: A review. *Neural Networks*, pages 54–71, 2019.

[27] Can Peng, Piotr Koniusz, Kaiyu Guo, Brian C. Lovell, and Peyman Moghadam. Multivariate prototype representation for domain-generalized incremental learning. *Computer Vision and Image Understanding*, 249:104215, 2024.

[28] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

[29] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

[30] Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

[31] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR, 2018.

[32] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1591–1600, 2021.

[33] James Seale 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. *CoRR*, 2023.

[34] 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.

[35] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

[36] Lei Wang, Jun Liu, Liang Zheng, Tom Gedeon, and Piotr Koniusz. Meet JEANIE: A similarity measure for 3d skeleton sequences via temporal-viewpoint alignment. *Int. J. Comput. Vis.*, 132(9):4091–4122, 2024.

[37] 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.

[38] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proceedings of the European Conference on Computer Vision*, pages 631–648, 2022.

[39] 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.

[40] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, pages 374–382, 2019.

[41] Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, and Rex Ying. Low-Rank Adaptation for Foundation Models: A Comprehensive Review, 2024.

[42] Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

[43] Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] 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. *arXiv preprint arXiv:2303.05118*, 2023.

[45] Shan Zhang, Yao Ni, Jinhao Du, Yuan Xue, Philip Torr, Piotr Koniusz, and Anton van den Hengel. Open-world objectness modeling unifies novel object detection. In *CVPR*, 2025.

[46] Yifei Zhang, Hao Zhu, Aiwei Liu, Han Yu, Piotr Koniusz, and Irwin King. Less is More: Extreme Gradient Boost Rank-1 Adaption for Efficient Finetuning of LLMs, 2024.

[47] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19068–19079, 2023.