MULTI-SYNAPTIC COOPERATION: A BIO-INSPIRED FRAMEWORK FOR ROBUST AND SCALABLE CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning aims to acquire new knowledge incrementally while retaining prior information, with catastrophic forgetting (CF) being a central challenge. Existing methods can mitigate CF to some extent but are constrained by limited capacity, which often requires dynamic expansion for long task sequences and makes performance sensitive to task order. Inspired by the richness and plasticity of synaptic connections in biological nervous systems, we propose the Multi-Synaptic Cooperation Network (MSCN), a generalized framework that models cooperative interactions among multiple synapses through multi-synaptic connections modulated by local synaptic activity. This design enhances model representational capacity and enables task-adaptive plasticity by means of multi-synaptic cooperation, providing a new avenue for expanding model capacity while improving robustness to task order. During learning, our MSCN dynamically activates task-relevant synapses while suppressing irrelevant ones, enabling targeted retrieval and minimizing interference. Extensive experiments across four benchmark datasets, involving both spiking and non-spiking neural networks, demonstrate that our method consistently outperforms state-of-the-art continual learning methods with significantly improved robustness to task-order variation. Furthermore, our analysis reveals an optimal trade-off between synaptic richness and learning efficiency, where excessive connectivity can impair circuit performance. These findings highlight the importance of the multi-synaptic cooperation mechanism for achieving efficient continual learning and provide new insights into biologically inspired, robust, and scalable continual learning.

1 INTRODUCTION

Continual learning aims to develop models capable of acquiring and retaining knowledge from a sequence of tasks or data distributions, thereby mimicking the human ability to learn progressively over time. This approach is also known as lifelong learning or incremental learning Thrun (1994) and holds promise for building adaptive and efficient systems in dynamic environments. A core challenge in continual learning is catastrophic forgetting McCloskey & Cohen (1989)—a phenomenon where the model's performance on previously acquired tasks degrades significantly when updated with new data De Lange et al. (2022); Parisi et al. (2019); Masana et al. (2023).

Recently, various approaches have been proposed to mitigate catastrophic forgetting Bonicelli et al. (2022); Tong et al. (2023); Qiao et al. (2024); Li et al. (2024a). These approaches can be broadly categorized into three primary types Masana et al. (2023): *Rehearsal-based* methods Lopez-Paz & Ranzato (2017); Bang et al. (2021); Van De Ven et al. (2020); Hayes et al. (2020), *Regularization-based* methods Kirkpatrick et al. (2017); Wołczyk et al. (2022); Schwarz et al. (2018); Li et al. (2024b), and *Architecture-based* methods Yoon et al. (2018); Zhou et al. (2022b); Li et al. (2019); Wang et al. (2023); Serra et al. (2018); Hung et al. (2019b); Kang et al. (2022). Among these three types, *Architecture-based* methods are particularly notable for the ability to dynamically adjust the network structure to accommodate new tasks. They primarily rely on two strategies: network expansion and pruning. Expansion strategies Hung et al. (2019b); Li et al. (2019); Yoon et al. (2018) start with a small model and dynamically expand the network to mitigate forgetting. In contrast, pruning strategies Mallya et al. (2018); Wang et al. (2022a); Kang et al. (2022) assign a sub-network

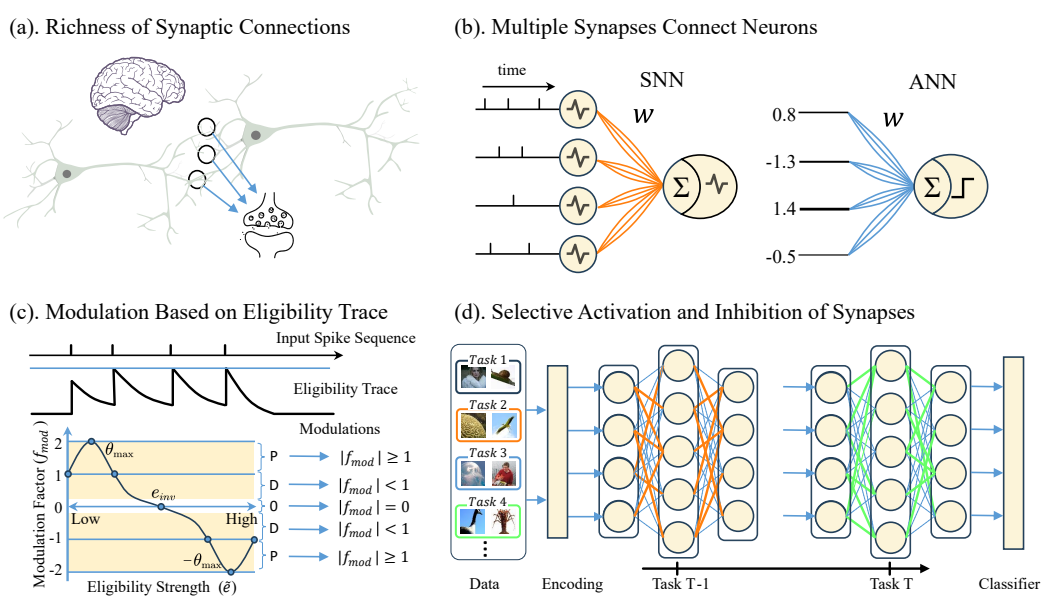

Figure 1: Overview of the MSCN framework. (a) The phenomenon of multi-synaptic connections in biological neurons. (b) Modeling of neurons with multiple synaptic connections in both SNNs and ANNs. (c) Local synaptic activity plasticity based on eligibility traces. Depending on the eligibility trace, synapses undergo potentiation (P), depression (D), or remain unchanged (O), reflecting the modulation's strength and direction. (d) The process of selectively activating task-relevant synapses and inhibiting irrelevant ones.

to each old task. These sub-networks are pruned from a pre-allocated dense model. Further training is restricted to the unpruned parameters only. While these strategies offer several advantages, including performance and the potential for forget-free learning in some cases, they still face notable challenges: (i) Dynamic expansion requires continuous network growth to accommodate new tasks, which makes it hardware-unfriendly; (ii) Both expansion and pruning methods are often sensitive to the task order, leading to significant performance variation depending on the sequence of tasks.

Remarkably, the brain achieves continual learning without suffering from dynamic structural growth Rasch & Born (2007); Joseph & Gu (2021), highlighting the potential of biologically inspired mechanisms as significant alternatives to artificial network expansion. Among these, multi-synaptic connectivity is believed to be a key factor in supporting continual learning Shi et al. (2025); Wu & Mel (2009); Ko et al. (2011), as its diverse synaptic architecture enhances the information representation capacity of neural circuits. Meanwhile, *Three-factor* learning rules Frémaux & Gerstner (2016); Gerstner et al. (2018) have been widely studied as biologically plausible models of synaptic plasticity, in which synaptic changes are not only modulated by global neuromodulatory signals but also depend on the local synaptic activity. Inspired by this, we propose MSCN, a novel framework that enhances representational flexibility and robustness by employing the multi-synaptic cooperation mechanism, rather than increasing network depth, width, or dynamically expanding the architecture, as illustrated in Fig. 1. Our framework consists of two components: the multi-synapse connectivity structure that augments the model's representational capacity within a fixed network architecture, and the synaptic plasticity modulation mechanism based on local synaptic activity. Such local synaptic activity is integrated via eligibility traces, which serve as modulatory signals to synaptic weight updates. During learning, task-relevant synapses are dynamically selected, while irrelevant ones are suppressed, thereby effectively minimizing interference across tasks.

Our contributions are as follows:

- We propose MSCN, the first continual learning framework that explicitly leverages the multi-synaptic cooperation mechanism, providing a biologically inspired and capacity-efficient solution. By maximally harnessing synaptic resources, MSCN unlocks the potential

of fixed-capacity networks and substantially boosts the scalability of continual learning systems.

- We design a modulatory mechanism based on local synaptic activities that modulates synaptic plasticity through eligibility traces, enabling precise, activity-dependent modulation at the synaptic level. This modulation significantly strengthens the robustness of continual learning models to task order variations, ensuring stable performance even under highly dynamic and unpredictable training sequences.
- Extensive experiments on four benchmark datasets across both spiking and non-spiking architectures demonstrate that MSCN consistently outperforms state-of-the-art continual learning methods in terms of accuracy, forgetting mitigation, and robustness to task order, while also exhibiting competitive computational efficiency.

2 Related work

Continual Learning methods are roughly divided into three categories: *Rehearsal-based* methods store past experiences in memory to mitigate forgetting. Some works Rebuffi et al. (2017); Tiwari et al. (2022); Zhou et al. (2022c); Jeeveswaran et al. (2023) design sampling strategies to allocate a limited memory budget, while others Lin et al. (2022); Rolnick et al. (2019); Sun et al. (2023) build special subspace of old tasks as the memory. *Regularization-based* methods aim to consolidate previous knowledge by adding extra regularization terms to the loss function. Some works Li & Hoiem (2017); Kirkpatrick et al. (2017); Cha et al. (2020) constrain important weights in the parameter space Akyürek et al. (2021); Rudner et al. (2022); Kim et al. (2023), feature representations Gao et al. (2022); Jeeveswaran et al. (2023), or output logits Li & Hoiem (2017); Oh et al. (2022) to remain close to those of the old model. *Architecture-based* methods dedicate different incremental model structures towards each task to minimize forgetting Zhou et al. (2022a); Lu et al. (2024); Kang et al. (2022). Some works Serra et al. (2018); Yoon et al. (2019); Hu et al. (2023) adopt modular architectures by dynamically expanding additional components Yan et al. (2021); Zhu et al. (2022), or freeze subsets of parameters Abati et al. (2020); Liu et al. (2021) to overcome forgetting. In this work, to better investigate the capacity efficiency and robustness of our method, we implement our multi-synaptic cooperation mechanism based on the architecture-based methods without dynamically expanding the network.

Neural Network Dynamics in continual learning describe how internal representations and connectivity patterns evolve as new tasks are learned sequentially Márton et al. (2022). These dynamics manifest across multiple levels, including synaptic updates, activation trajectories, and parameter plasticity under task transitions Vyas et al. (2020). Recent works have begun to explicitly model neural dynamics in continual learning by introducing mechanisms such as synaptic trajectories, context-dependent modulation, and task-driven weight routing Li & Wang (2017); Li et al. (2024c); Xu et al. (2024). These studies highlight the importance of capturing temporal evolution in network parameters to support adaptive behavior over extended task sequences. Additional research explores dynamic mechanisms such as gating, masking, and sparsity-inducing priors to modulate parameter updates and isolate task-specific pathways Abati et al. (2020); Wang et al. (2022b); Yan et al. (2022). Recently, multi-synaptic (redundant) connections between neuron pairs have been shown to enhance computational capacity Zenke & Laborieux (2024); Hofmann et al. (2025). Nevertheless, the cooperative interactions among multiple synapses and their implications for continual learning remain largely underexplored. In contrast, our method introduces multi-synaptic dynamics within each connection and the modulation based on local synapse activity, enabling representational diversity and adaptive modulation of synaptic plasticity. This novel design does not require increasing network depth/width or dynamically expanding the network; instead, it increases capacity and enhances robustness through a multi-synaptic cooperation mechanism.

3 Method

3.1 Modeling Multi–Synaptic Spiking Neuron

Biological systems achieve continual learning without relying on architectural growth Song et al. (2024); Shi et al. (2025). A key neurobiological observation is the presence of multiple synaptic

contacts between the same axon–dendrite pair, providing redundancy and adaptability Trachtenberg et al. (2002); Yang et al. (2014). Since our design operates at the synaptic level, the proposed method can be applied to both ANN and SNN Maass (1997); Gütig & Sompolinsky (2006) architectures. To better capture biological principles, we first model the multi-synaptic cooperation mechanism in SNNs, which are more consistent with biological processes and offer event-driven, temporally sparse, and energy-efficient computation Gütig & Sompolinsky (2006). As a concrete instantiation, we adopt the leaky integrate-and-fire (LIF) neuron Lapicque (1907), a widely used model balancing biological plausibility and computational simplicity Shiu et al. (2024); Brand & Petruccione (2024). The membrane potential $V(t)$ evolves over continuous time as follows:

$$\tau_{\mathrm{m}} \frac{dV(t)}{dt} = -\big(V(t) - V_{\mathrm{rest}}\big) + I(t) \tag{1}$$

where τ_{m} is the membrane time constant, V_{rest} is the resting potential, and $I(t)$ denotes the total synaptic input current. This formulation captures leakage and current integration but assumes a single synapse per connection, limiting diversity. To address this, we generalize the neuron model by introducing $P \geq 1$ parallel synapses for each synaptic connection. Consider a neuron with N presynaptic neurons, each forming P distinct synaptic pathways to it. Therefore, in continuous time, the membrane potential of the postsynaptic neuron is given by:

$$V(t) = \sum_{i=1}^{N} \sum_{p=1}^{P} w_{ip} \mathrm{PSP}_{ip}(t) - \ \vartheta \sum_{j} e^{-\frac{t - t_{\mathrm{s}}^{j}}{\tau_{\mathrm{m}}}} \tag{2}$$

where w_{ip} denotes the synaptic weight of the p-th parallel synapse associated with the i-th presynaptic neuron, PSP_{ip} represents the postsynaptic potential of this synapse and ϑ denotes the firing threshold. To preserve synaptic heterogeneity Deng et al. (2025) and enable independent optimization, we introduce distinct decay time constants across parallel synapses. Accordingly, for the p-th synapse of presynaptic neuron i, the spike arrival times are denoted by t_{ip}^{f}, and PSP_{ip} is defined as:

$$\mathrm{PSP}_{ip}(t) = \sum_{f} K_{ip}\Big(t - t_{ip}^{f}\Big) \tag{3}$$

$$K_{ip}(t) = e^{-\frac{t}{\tau_{s_{ip}}}} \tag{4}$$

where $K_{ip}(t)$ denotes the kernel function of the p-th parallel synapse, and $\tau_{s_{ip}}$ represents its decay time constant, which is initialized to different values (non-trainable). This design allows multiple temporal and weighted channels to influence the synaptic plasticity, thereby enhancing the diversity of spatiotemporal representations. On this basis, modeling in the ANN architecture can be more readily formulated. Given space limitations, the corresponding implementation is presented in Appendix A.1.

3.2 Plasticity Modulation Based on Eligibility Traces

Building on multi-synaptic connections, we introduce a modulation mechanism of synaptic plasticity based on local synaptic activity. Specifically, we propose the eligibility trace as the basis for modulating local synaptic plasticity. We begin by formulating the modulation signal in continuous time as a cumulative sum of synaptic spike events. For a connection between two neurons, P parallel synapses share a common eligibility trace, which is defined as:

$$\frac{d\tilde{e}}{dt} = -\frac{\tilde{e}}{\tau} + \sum_{f} \delta\left(t - t^{f}\right) \tag{5}$$

where τ is the decay time constant, $\delta(\cdot)$ denotes the Dirac delta function. For practical implementation, we adopt a discrete-time formulation, and the dynamics of $\tilde{e}$ are updated as follows:

$$\tilde{e}[t+1] = \tilde{e}[t] - \frac{\tilde{e}[t]}{\tau} + S[t+1] \tag{6}$$

where $S[t+1] \in \{0,1\}$ indicates whether a spike occurred at time step $t+1$. This design captures recent spiking accumulation while allowing the eligibility trace to decay in the absence of input, enabling the eligibility trace to track local synaptic activity over time.

The effect of the modulation factor on synaptic plasticity is governed by a nonlinear function $f_{\text{mod}}(\tilde{e})$, which determines how strongly and in what direction the synapse should change in response to local synaptic activity signals. Following Zhang et al. (2023), we adopt a piecewise quadratic form for f_{mod}, which adjusts the strength and direction (potentiation or depression) of synaptic plasticity depending on the eligibility trace. In practice, $\tilde{e}$ is normalized over all model synapses to $[-1,1]$ to reflect the relative strength of local synaptic activity. The modulation function is defined as:

$$f_{\text{mod}}(\tilde{e}) = \begin{cases} 1 + (\theta_{\max} - 1)\left(\dfrac{|\tilde{e}|}{e_1}\right)^2, & 0 \le |\tilde{e}| \le e_1, \\ \theta_{\max}\left[1 - \left(\dfrac{|\tilde{e}| - e_1}{e_{\text{inv}} - e_1}\right)^2\right], & e_1 \le |\tilde{e}| \le e_{\text{inv}}, \\ -f_{\text{mod}}(2e_{\text{inv}} - |\tilde{e}|), & e_{\text{inv}} \le |\tilde{e}| \le 2e_{\text{inv}}, \\ 0, & otherwise \end{cases} \tag{7}$$

where $e_1 = 0.5\, e_{\text{inv}}$ sets the low–moderate boundary, $\theta_{\max}$ controls the maximum modulation strength, and e_{inv} marks the zero point of f_{mod}. (Fig. 1c). This function modulates plasticity based on local synaptic activity: potentiation occurs when $|f_{\text{mod}}| \ge 1$, depression arises when $0 < |f_{\text{mod}}| < 1$, and complete depression is observed when $f_{\text{mod}} = 0$. The final modulated synaptic weight change is computed as:

$$\Delta w = -\eta \cdot \left|f_{\text{mod}}(\tilde{e})\right| \cdot \frac{\partial \mathcal{L}}{\partial w} \tag{8}$$

where η is the learning rate and $\partial \mathcal{L}/\partial w$ is the gradient. The modulation function f_{mod} adjusts the size and direction of the synaptic weight update. By dynamically adjusting synaptic updates in response to recent local activity, this design mimics the biological mechanism of robust learning through activity-dependent modulation Wu et al. (2021); Wu & Maass (2025), thereby enabling the network to adapt to changing task demands and maintain stable performance.

3.3 Generalizing to Classic Architecture-Based Methods

In this section, we integrate our method into the *Architecture-based* setting, adopting the same setup as in Kang et al. (2023); Serra et al. (2018); Wortsman et al. (2020). Consider a standard supervised continual learning setting with T tasks presented sequentially. For each task j, the model receives a dataset $\mathcal{D}_j = \{(\mathbf{x}_{i,j}, y_{i,j})\}_{i=1}^{n_j}$ consisting of n_j labeled samples. A fixed-topology deep neural network $\mathcal{F}(\cdot;\boldsymbol{\theta})$, parameterized by model parameters $\boldsymbol{\theta}$, is employed. The objective at each step is to optimize the model for the current task j:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{minimize}} \frac{1}{n_j} \sum_{i=1}^{n_j} \mathcal{L}\left(\mathcal{F}(\mathbf{x}_{i,j};\boldsymbol{\theta}), y_{i,j}\right) \tag{9}$$

Following Gao et al. (2023); Wortsman et al. (2020), task identities are assumed to be available during both training and inference, under a multi-head setting in which each task is assigned a distinct output head. For each task j, a binary mask $\mathbf{m}_j^*$ is learned to activate the relevant synapses. The training objective is formulated as:

$$\boldsymbol{\theta}^*, \mathbf{m}_j^* = \underset{\boldsymbol{\theta}, \mathbf{m}_j}{\text{minimize}} \frac{1}{n_j} \sum_{i=1}^{n_j} \Big[\mathcal{L}\big(\mathcal{F}(\mathbf{x}_{i,j};\boldsymbol{\theta} \odot \mathbf{m}_j), y_{i,j}\big) - \mathcal{L}\big(\mathcal{F}(\mathbf{x}_{i,j};\boldsymbol{\theta}), y_{i,j}\big)\Big] \tag{10}$$

where $\odot$ denotes element-wise multiplication. A shared learnable relevance score $\mathbf{r}$ is maintained across tasks, with each entry corresponding to a synapse Kang et al. (2023). Trained jointly with

the network parameters, $\mathbf{r}$ enables the model to identify task-relevant connections. For task j, the subnetwork $\hat{\boldsymbol{\theta}}_j$ is formed by selecting the top $c\%$ of weights ranked by relevance, where c is the layerwise capacity ratio Wortsman et al. (2020). The selected weights are indicated by the binary mask $\boldsymbol{m}_j$, in which a value of 1 signifies that the corresponding weight is active during the forward pass, and 0 indicates it is deactivated. To preserve past knowledge, an accumulated mask $\mathbf{M}_{j-1} = \bigvee_{i=1}^{j-1} \mathbf{m}_i$ (with $\bigvee$ as logical OR) is applied when learning task j. The parameters $\boldsymbol{\theta}$ are updated as:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \boldsymbol{\Delta\theta} \odot (\mathbf{1} - \mathbf{M}_{j-1}) \tag{11}$$

where $\boldsymbol{\Delta\theta}$ denotes the gradient step, and the term $(\mathbf{1} - \mathbf{M}_{j-1})$ ensures that only unallocated synapses remain trainable, thereby preserving the stability of parameters from previously learned tasks.

4 Experiments

4.1 Experimental Setup

We conduct comprehensive experiments under diverse training configurations, input domains, datasets, and network architectures, and evaluate continual learning performance using two widely adopted metrics Kang et al. (2022); Konishi et al. (2023): *Average Accuracy (ACC)* and *Backward Transfer (BWT)*. ACC measures the model's average accuracy over all tasks after learning task T, reflecting its overall generalization ability. BWT measures the impact of new tasks on prior ones, with higher values better and 0 indicating no forgetting. Unless otherwise stated, we set the synapse count to $P = 3$ in all main experiments, ablation studies, and computational cost analyses. Implementation and hardware details are provided in Appendix A.2.

Table 1: Performance comparison on four datasets, evaluating performance under both SNN and ANN frameworks. We report the results across 5 independent runs with different random seeds under the same experimental setup. Table 5 in Appendix A.2 shows the standard deviations.

Network	Method	PMNIST		10-split CIFAR-100		TinyImageNet		5-Datasets	
		ACC (%) ↑	BWT (%) ↑	ACC (%) ↑	BWT (%) ↑	ACC (%) ↑	BWT (%) ↑	ACC (%) ↑	BWT (%) ↑
SNN	*MTL*	96.52	/	79.83	/	79.24	/	89.93	/
	EWC[PNAS] Kirkpatrick et al. (2017)	91.45	-3.20	73.75	-4.89	60.29	-25.47	57.06	-44.55
	HAT[ICML] Serra et al. (2018)	93.25	-2.07	73.67	-0.13	62.18	-8.51	72.72	-22.90
	GPM[ICLR] Saha et al. (2021)	94.80	-1.62	77.48	-1.37	70.07	-2.92	79.70	-15.52
	HLOP[ICLR] Xiao et al. (2024)	95.15	-1.30	78.58	-0.26	71.40	-0.52	88.65	-3.71
	MSCN	**96.34**	**0.0**	**79.54**	**0.0**	**73.22**	**0.0**	**88.84**	**0.0**
ANN	EWC[PNAS] Kirkpatrick et al. (2017)	92.01	-0.03	72.77	-3.59	64.51	-0.04	88.64	-0.04
	GPM[ICLR] Saha et al. (2021)	94.96	-0.02	73.18	-1.17	67.39	1.45	91.22	-0.01
	PackNet[CVPR] Mallya et al. (2018)	96.37	0.0	72.39	0.0	55.46	0.0	92.81	0.0
	SupSup[NeurIPS] Wortsman et al. (2020)	96.31	0.0	75.47	0.0	59.60	0.0	93.28	0.0
	WSN[ICML] Kang et al. (2022)	96.41	0.0	76.38	0.0	71.96	0.0	93.41	0.0
	TAMiL[ICLR] Bhat et al. (2023)	96.87	-3.15	76.73	-3.47	72.55	-3.02	93.47	-4.72
	SPG[ICML] Konishi et al. (2023)	96.35	–	74.82	–	73.26	–	93.32	–
	DFGP[ICCV] Yang et al. (2023)	94.64	-0.01	74.59	0.0	–	–	92.09	-0.01
	Bayesian[ICML] Thapa & Li (2024)	96.74	–	75.57	–	73.93	–	93.36	–
	MSCN	**97.53**	**0.0**	**77.37**	**0.0**	**75.03**	**0.0**	**93.69**	**0.0**

4.2 Comparison to the state-of-the-art methods

We first conduct a comprehensive evaluation of MSCN in a multi-head task-incremental learning scenario, using four widely used benchmark datasets and employing both SNNs and ANNs. As summarized in Table 1, where the dataset complexity roughly increases from left to right, the results consistently demonstrate the strength and reliability of MSCN across diverse datasets and architectures. Notably, on TinyImageNet with ANN, MSCN achieves an accuracy that exceeds the second-best method, Bayesian Thapa & Li (2024), by 1.10%. Although several approaches, such as WSN Kang et al. (2022), SupSup Wortsman et al. (2020), and PackNet Mallya et al. (2018), achieve zero backward transfer, they do not match the overall accuracy of MSCN. These findings indicate that multi-synaptic connectivity is an effective and scalable design for continual learning. Moreover, Fig. 2 shows that MSCN attains superior per-task performance on most tasks, further validating its representational strength. Since most existing works on continual learning are based on ANN

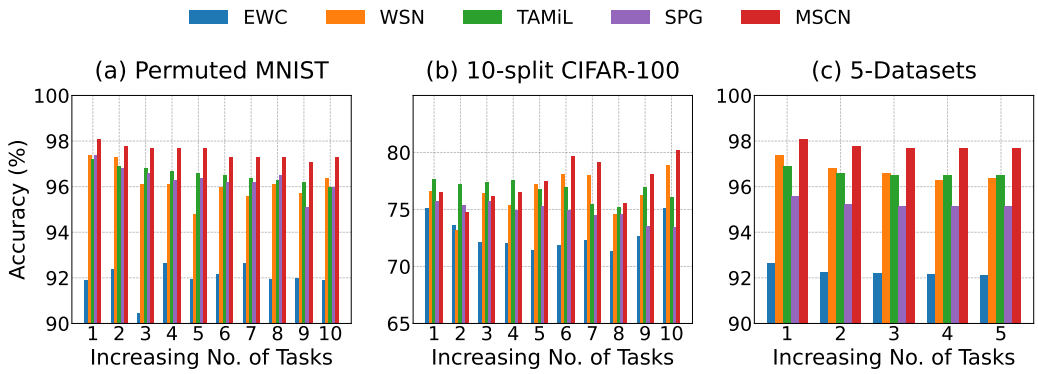

Figure 2: Per-task accuracy across the entire incremental learning.

architectures, we investigate our method extensively within the same framework in the following sections to ensure fairness while highlighting its strengths.

4.3 Robustness to Task Order Permutations

We evaluate order robustness by training on multiple CIFAR-100 Split permutations and measuring per-task accuracy variation over three task orders. As illustrated in Fig. 3d, we observe that EWC Kirkpatrick et al. (2017) and GPM Saha et al. (2021) display large fluctuations across task sequences, highlighting their strong sensitivity to order. WSN Kang et al. (2022) performs competitively with EWC but shows a tendency to overfit to particular task orders (Fig. 3a–b). In contrast, MSCN achieves stable accuracy across all tasks and permutations, with only minimal variation (Fig. 3c). This consistency indicates that MSCN can flexibly adapt to new tasks while mitigating interference and preserving prior knowledge, highlighting the important role of local activity-dependent synaptic modulation as a foundational mechanism for building scalable continual learning systems. Additional experiments on five task orders are reported in Fig. 10 of the Appendix A.3.4.

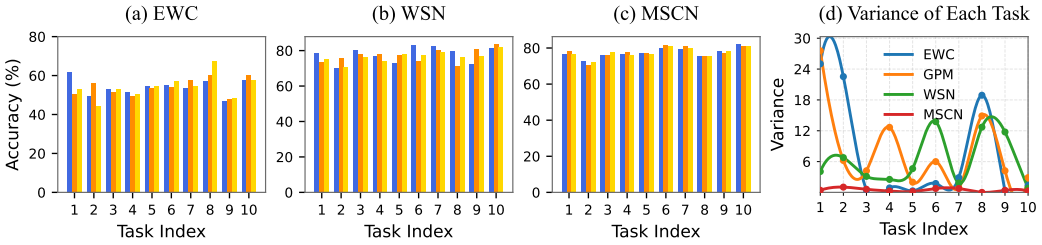

Figure 3: Task order robustness comparison on CIFAR-100 Split. Bar plots in (a), (b), and (c) show per-task accuracy under three different task sequences. Specifically, (a) corresponds to the EWC method, (b) to the WSN method, and (c) to our MSCN method. (d) shows the standard deviation of the accuracy for each task across different task sequences.

4.4 Ablation Study

To better understand the individual contributions of the key components in MSCN, we perform an ablation study by selectively removing the multi-synaptic connectivity structure and the neuromodulatory mechanism. The results across four benchmark datasets are summarized in Table 2. The full MSCN model, with both components enabled, consistently achieves the highest accuracy across all datasets. Disabling the neuromodulatory mechanism results in a noticeable performance drop, particularly on TinyImageNet and PMNIST. Conversely, removing the multi-synaptic structure also results in performance degradation, particularly on CIFAR-100 and 5-Datasets. We observe that when both components are ablated, performance drops further across all datasets, confirming that

the two mechanisms act cooperatively and that their interaction is essential for MSCN's ability to achieve robust and scalable continual learning.

Table 2: An ablation study of MSCN on ACC. ✓indicates that the component is included, while ✗indicates that it is excluded.

Multi-synapse	Modulation	PMNIST	CIFAR-100	TinyImageNet	5-Datasets
✓	✓	97.53 (± 0.19)	77.37 (± 0.23)	75.03 (± 0.27)	93.69 (± 0.21)
✗	✓	96.79 (± 0.20)	77.03 (± 0.22)	73.81 (± 0.26)	93.47 (± 0.24)
✓	✗	96.53 (± 0.21)	76.81 (± 0.25)	73.78 (± 0.29)	93.51 (± 0.22)
✗	✗	96.34 (± 0.22)	76.34 (± 0.24)	72.59 (± 0.31)	93.32 (± 0.25)

4.5 Effect of Synapse Count

To probe the role of multi-synaptic connectivity, we vary synapse and neuron counts and measure performance across tasks. On CIFAR-100 Split (Fig. 4), increasing synapses per connection consistently raises per-task accuracy as the number of tasks grows. Jointly scaling synapses and neurons (Fig. 5) reveals a saturation regime: accuracy improves with capacity but plateaus once model capacity exceeds task complexity. Fig. 6 summarizes four benchmarks; rows (top→bottom) are PMNIST, CIFAR-100, TinyImageNet, and 5-Datasets, reflecting increasing complexity. We observe that, although the optimal number of synapses varies with task difficulty, it generally stabilizes once a certain threshold is exceeded. Interestingly, the observed performance trend happens to mirror how synaptic counts are distributed in the brain—typically confined to a limited but effective rangeToni et al. (1999); Watson et al. (2025). Additional experimental results are provided in Appendix A.3.3.

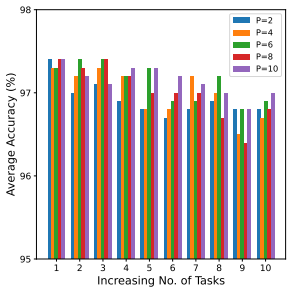

Figure 4: Per-task accuracy under different synapse counts on CIFAR-100 Split.

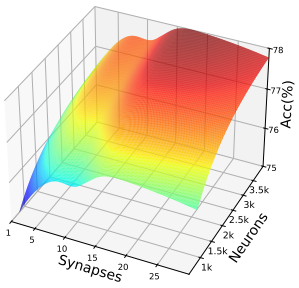

Figure 5: Accuracy under different synapse and neuron counts on CIFAR-100 Split.

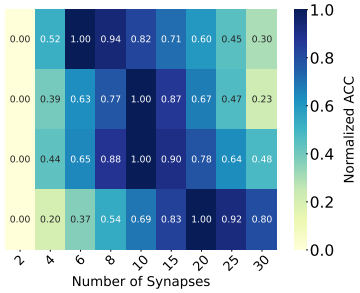

Figure 6: Accuracy changes across four datasets under varying numbers of synapses

4.6 Capacity Analysis

To evaluate the capacity efficiency of MSCN, we adopt the commonly used metric CAP Kang et al. (2022); Wortsman et al. (2020) (defined in Appendix A.2.2) and compare it with baseline approaches, where lower CAP values indicate better capacity utilization. Fig. 7 shows the relationship between accuracy and total capacity usage across four benchmark datasets. The results demonstrate that MSCN consistently achieves the highest accuracy with significantly lower capacity overhead. On Permuted MNIST and TinyImageNet (Fig. 7a,c), MSCN outperforms all baselines while requiring substantially fewer resources. Similar trends are observed on CIFAR-100 and the 5-Datasets benchmark (Fig. 7b,d), where methods such as PackNet Mallya et al. (2018) and SupSup Wortsman et al. (2020) consume much more capacity but yield lower or comparable performance. These results indicate that the multi-synaptic cooperation mechanism enables more effective knowledge storage and reuse across tasks.

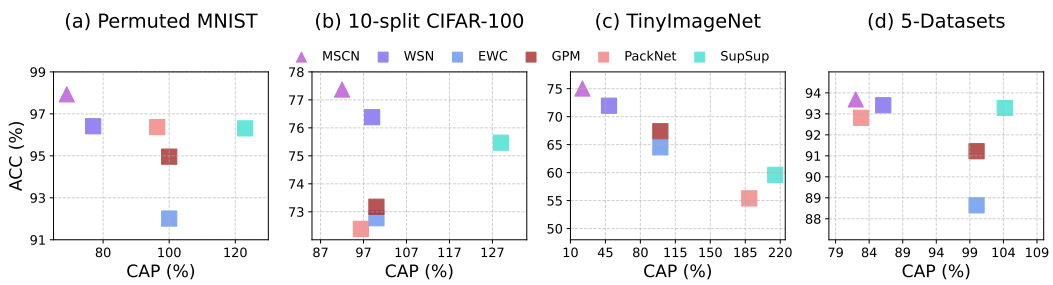

Figure 7: Accuracy over CAP (%) across four benchmarks.

4.7 Computational Efficiency

Since multi-synaptic connections inevitably increase the parameter count, we conducted experiments under the same parameter budget as the baselines to fairly evaluate computational efficiency. We further evaluated robustness under five task orders using AOPD (where lower values indicate stronger robustness; see Appendix A.2.2 for details). As shown in Table 3, although reducing the parameter count leads to a drop in MSCN's accuracy, our method consistently achieves the best efficiency and robustness while maintaining competitive accuracy. These results demonstrate that the proposed multi-synaptic cooperation mechanism achieves high computational efficiency and establishes the foundation for robust and scalable continual learning.

Table 3: Computational efficiency under the same parameter budget (Training time in hours).

Method	PMNIST			10-split CIFAR-100		
	Training Time ↓	ACC (%) ↑	AOPD (%) ↓	Training Time ↓	ACC (%) ↑	AOPD (%) ↓
PackNet	0.59 (± 0.15)	96.43 (± 0.18)	2.23	1.13 (± 0.10)	72.45 (± 0.20)	5.36
SupSup	0.53 (± 0.12)	96.36 (± 0.22)	1.27	0.87 (± 0.08)	75.54 (± 0.17)	3.81
WSN	0.38 (± 0.05)	96.49 (± 0.13)	0.29	0.78 (± 0.06)	76.47 (± 0.34)	2.59
MSCN	**0.33** (± 0.11)	**96.89** (± 0.19)	**0.23**	**0.65** (± 0.04)	**76.40** (± 0.14)	**2.41**

Method	TinyImageNet			5-Datasets		
	Training Time ↓	ACC (%) ↑	AOPD (%) ↓	Training Time ↓	ACC (%) ↑	AOPD (%) ↓
PackNet	1.45 (± 0.12)	55.51 (± 0.25)	6.51	3.45 (± 0.08)	92.89 (± 0.12)	4.37
SupSup	0.97 (± 0.07)	59.65 (± 0.24)	6.94	3.26 (± 0.10)	93.31 (± 0.16)	2.83
WSN	0.92 (± 0.04)	72.03 (± 0.41)	4.98	3.05 (± 0.08)	93.50 (± 0.13)	1.37
MSCN	**0.75** (± 0.03)	**74.04** (± 0.21)	**4.73**	**2.82** (± 0.09)	**93.33** (± 0.09)	**1.26**

5 Conclusion

In this paper, we propose MSCN, the first continual learning framework that explicitly models multi-synaptic cooperation. By equipping each connection with multiple plastic synapses and employing local synaptic activity–based modulation, MSCN achieves effective knowledge retention and adaptability within a fixed architecture. Extensive evaluations across diverse datasets and architectures demonstrate that our approach consistently outperforms state-of-the-art baselines in terms of computational efficiency and task order robustness. These findings highlight that the synergistic interplay between multi-synaptic connectivity and localized plasticity modulation substantially enhances the network's representational capacity, providing new insights for robust and scalable continual learning. In addition, the similarity between our model's synapse counts and those found in biological systems further supports the plausibility of our design. Future work will explore allocating different numbers of synapses across neuron connections to further optimize synaptic resource utilization.

REFERENCES

Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3931–3940, 2020.

Afra Feyza Akyürek, Ekin Akyürek, Derry Tanti Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. *arXiv preprint arXiv:2110.07059*, 2021.

J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi. Rainbow memory: Continual learning with a memory of diverse samples. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8214–8223, 2021. doi: 10.1109/CVPR46437.2021.00812.

Prashant Bhat, Bahram Zonooz, and Elahe Arani. Task-aware information routing from common representation space in lifelong learning. *arXiv preprint arXiv:2302.11346*, 2023.

L. Bonicelli, M. Boschini, A. Porrello, C. Spampinato, and S. Calderara. On the effectiveness of lipschitz-driven rehearsal in continual learning. In *Advances in Neural Information Processing Systems*, volume 35 of *Advances in Neural Information Processing Systems*, 2022.

Dean Brand and Francesco Petruccione. A quantum leaky integrate-and-fire spiking neuron and network. *npj Quantum Information*, 10(1):125, 2024.

Alexander R Callan, Martin Heß, Felix Felmy, and Christian Leibold. Arrangement of excitatory synaptic inputs on dendrites of the medial superior olive. *Journal of Neuroscience*, 41(2):269–283, 2021.

Sungmin Cha, Hsiang Hsu, Taebaek Hwang, Flavio P Calmon, and Taesup Moon. Cpr: Classifier-projection regularization for continual learning. *arXiv preprint arXiv:2006.07326*, 2020.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Tiny episodic memories in continual learning. In *International Conference on Machine Learning*, pp. 1286–1295. PMLR, 2019.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Bohan Deng, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Modular continual learning in a unified visual environment. In *European Conference on Computer Vision (ECCV)*, 2021.

Zhichao Deng, Zhikun Liu, Junxue Wang, Shengqian Chen, Xiang Wei, and Qiang Yu. Hetsyn: Versatile timescale integration in spiking neural networks via heterogeneous synapses. *arXiv preprint arXiv:2508.11644*, 2025.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

Nicolas Frémaux and Wulfram Gerstner. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in neural circuits*, 9:85, 2016.

Qiang Gao, Xiaojun Shan, Yuchen Zhang, and Fan Zhou. Enhancing knowledge transfer for task incremental learning with data-free subnetwork. *Advances in Neural Information Processing Systems*, 36:68471–68484, 2023.

Qiankun Gao, Chen Zhao, Bernard Ghanem, and Jian Zhang. R-dfcil: Relation-guided representation learning for data-free class incremental learning. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Frontiers in neural circuits*, 12:53, 2018.

Aayush Gupta, Vinay Verma, and Piyush Rai. Bias mitigation in continual learning using adaptive task balancing. In *International Conference on Learning Representations*, 2020a.

G. Gupta, K. Yadav, and L. Paull. La-MAML: Look-ahead meta learning for continual learning. In *Advances in Neural Information Processing Systems*, volume 2020-December of *Advances in Neural Information Processing Systems*, 2020b.

Robert Gütig and Haim Sompolinsky. The tempotron: a neuron that learns spike timing–based decisions. *Nature neuroscience*, 9(3):420–428, 2006.

T.L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan. REMIND your neural network to prevent catastrophic forgetting. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12353 LNCS:466–483, 2020. doi: 10.1007/978-3-030-58598-3_28.

Martin Hofmann, Moritz Franz Peter Becker, Christian Tetzlaff, and Patrick Mäder. Concept transfer of synaptic diversity from biological to artificial neural networks. *Nature communications*, 16(1): 5112, 2025.

Zhiyuan Hu, Yunsheng Li, Jiancheng Lyu, Dashan Gao, and Nuno Vasconcelos. Dense network expansion for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11858–11867, 2023.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in neural information processing systems*, 32, 2019a.

S.C.Y. Hung, C.-H. Tu, C.-E. Wu, C.-H. Chen, Y.-M. Chan, and C.-S. Chen. Compacting, picking and growing for unforgetting continual learning. In *Advances in Neural Information Processing Systems*, volume 32 of *Advances in Neural Information Processing Systems*, 2019b.

Kishaan Jeeveswaran, Prashant Bhat, Bahram Zonooz, and Elahe Arani. Birt: Bio-inspired replay in vision transformers for continual learning. *arXiv preprint arXiv:2305.04769*, 2023.

J. Joseph and A. Gu. Reproducibility report: La-maml: Look-ahead meta learning for continual learning. *Corr*, 2021.

H. Kang, R.J.L. Mina, S.R.H. Madjid, J. Yoon, M. Hasegawa-Johnson, S.J. Hwang, and C.D. Yoo. Forget-free Continual Learning with Winning Subnetworks. In *Proceedings of Machine Learning Research*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10734–10750, 2022.

H. Kang, J. Yoon, S.R. Madjid, S.J. Hwang, and C.D. Yoo. On the soft-subnetwork for few-shot class incremental learning. 11th International Conference on Learning Representations, ICLR 2023, 2023.

Do-Yeon Kim, Dong-Jun Han, Jun Seo, and Jaekyun Moon. Warping the space: Weight space rotation for class-incremental few-shot learning. In *The Eleventh International Conference on Learning Representations*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114.

Ho Ko, Sonja B Hofer, Bruno Pichler, Katherine A Buchanan, P Jesper Sjöström, and Thomas D Mrsic-Flogel. Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91, 2011.

Tatsuya Konishi, Mori Kurokawa, Chihiro Ono, Zixuan Ke, Gyuhak Kim, and Bing Liu. Parameter-level soft-masking for continual learning. In *International Conference on Machine Learning*, pp. 17492–17505. PMLR, 2023.

Louis Édouard Lapicque. Louis lapicque. *J. physiol*, 9:620–635, 1907.

D. Li, T. Wang, J. Chen, K. Kawaguchi, C. Lian, and Z. Zeng. Multi-view class incremental learning. *Information Fusion*, 102, 2024a. doi: 10.1016/j.inffus.2023.102021.

D. Li, T. Wang, J. Chen, Q. Ren, K. Kawaguchi, and Z. Zeng. Towards continual learning desiderata via HSIC-bottleneck orthogonalization and equiangular embedding. volume 38 of *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13464–13473, 2024b. doi: 10.1609/aaai.v38i12.29249.

Depeng Li, Tianqi Wang, Junwei Chen, Wei Dai, and Zhigang Zeng. Harnessing neural unit dynamics for effective and scalable class-incremental learning. *arXiv preprint arXiv:2406.02428*, 2024c.

M. Li and D. Wang. Insights into randomized algorithms for neural networks: Practical issues and common pitfalls. *Information Sciences*, 382–383:170–178, 2017. doi: 10.1016/j.ins.2016.12.007.

X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. *International Conference on Machine Learning*, pp. 3925–3934, 2019.

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

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning. *arXiv preprint arXiv:2202.02931*, 2022.

Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 2544–2553, 2021.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Aojun Lu, Tao Feng, Hangjie Yuan, Xiaotian Song, and Yanan Sun. Revisiting neural networks for continual learning: An architectural perspective. *arXiv preprint arXiv:2404.14829*, 2024.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

A. Mallya, D. Davis, and S. Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 67–82, 2018.

C.D. Márton, S. Zhou, and K. Rajan. Linking task structure and neural network dynamics. *Nature Neuroscience*, 25(6):679–681, 2022. doi: 10.1038/s41593-022-01090-w.

Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost Van De Weijer. Class-Incremental Learning: Survey and Performance Evaluation on Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, May 2023. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2022.3213473.

M. McCloskey and N.J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24 (C):109–165, 1989. doi: 10.1016/S0079-7421(08)60536-8.

Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *International conference on machine learning*, pp. 15699–15717. PMLR, 2022.

Youngmin Oh, Donghyeon Baek, and Bumsub Ham. Alife: Adaptive logit regularizer and feature replay for incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 35:14516–14528, 2022.

G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012.

Z. Qiao, Q. Pham, Z. Cao, H. H. Le, P. N. Suganthan, X. Jiang, and R. Savitha. *Class-incremental learning for time series: Benchmark and evaluation*, 2024.

B. Rasch and J. Born. Maintaining memories by reactivation. *Current Opinion in Neurobiology*, 17 (6):698–703, 2007. doi: 10.1016/j.conb.2007.11.007.

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*, pp. 2001–2010, 2017.

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

Tim GJ Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual learning via sequential function-space variational inference. In *International Conference on Machine Learning*, pp. 18871–18887. PMLR, 2022.

G. Saha, I. Garg, and K. Roy. Gradient projection memory for continual learning. ICLR 2021 - 9th International Conference on Learning Representations, 2021.

J. Schwarz, J. Luketina, W. M. Czarnecki, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell. Progress & compress: A scalable framework for continual learning. *Progress & Compress: A Scalable Framework for Continual Learning*, pp. 4528–4537, 2018.

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

Qianqian Shi, Faqiang Liu, Hongyi Li, Guangyu Li, Luping Shi, and Rong Zhao. Hybrid neural networks for continual learning inspired by corticohippocampal circuits. *Nature Communications*, 16(1):1272, 2025.

Philip K Shiu, Gabriella R Sterne, Nico Spiller, Romain Franconville, Andrea Sandoval, Joie Zhou, Neha Simha, Chan Hyuk Kang, Seongbong Yu, Jinseop S Kim, et al. A drosophila computational brain model reveals sensorimotor processing. *Nature*, 634(8032):210–219, 2024.

Yuhang Song, Beren Millidge, Tommaso Salvatori, Thomas Lukasiewicz, Zhenghua Xu, and Rafal Bogacz. Inferring neural activity before plasticity as a foundation for learning beyond backpropagation. *Nature neuroscience*, 27(2):348–358, 2024.

Wenju Sun, Qingyong Li, Jing Zhang, Wen Wang, and Yangli-ao Geng. Decoupling learning and remembering: A bilevel memory framework with knowledge projection for task-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20186–20195, 2023.

Jeevan Thapa and Rui Li. Bayesian adaptation of network depth and width for continual learning. In *Forty-first International Conference on Machine Learning*, 2024.

S. Thrun. A lifelong learning perspective for mobile robot control. *Intelligent Robots and Systems*, pp. 23–30, 1994.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 99–108, 2022.

S. Tong, X. Dai, Z. Wu, M. Li, B. Yi, and Y. Ma. Incremental learning of structured memory via closed-loop transcription. 11th International Conference on Learning Representations, ICLR 2023, 2023.

Nicolas Toni, P-A Buchs, Irina Nikonenko, CR Bron, and Dominique Muller. Ltp promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature*, 402(6760): 421–425, 1999.

Joshua T Trachtenberg, Brian E Chen, Graham W Knott, Guoping Feng, Joshua R Sanes, Egbert Welker, and Karel Svoboda. Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. *Nature*, 420(6917):788–794, 2002.

Marta Turegano-Lopez, Felix de Las Pozas, Andrea Santuy, Jose-Rodrigo Rodriguez, Javier DeFelipe, and Angel Merchan-Perez. Tracing nerve fibers with volume electron microscopy to quantitatively analyze brain connectivity. *Communications Biology*, 7(1):796, 2024.

Gido M. Van De Ven, Hava T. Siegelmann, and Andreas S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1):4069, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17866-2.

S. Vyas, M.D. Golub, D. Sussillo, and K.V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020. doi: 10.1146/annurev-neuro-092619-094115.

F.-Y. Wang, D.-W. Zhou, L. Liu, H.-J. Ye, Y. Bian, D.-C. Zhan, and P. Zhao. Beef: Bi-compatible class-incremental learning via energy-based expansion and fusion. 11th International Conference on Learning Representations, ICLR 2023, 2023.

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

Zifeng Wang, Zheng Zhan, Yifan Gong, Geng Yuan, Wei Niu, Tong Jian, Bin Ren, Stratis Ioannidis, Yanzhi Wang, and Jennifer Dy. Sparcl: Sparse continual learning on the edge. *Advances in Neural Information Processing Systems*, 35:20366–20380, 2022b.

Jake F Watson, Victor Vargas-Barroso, Rebecca J Morse-Mora, Andrea Navas-Olive, Mojtaba R Tavakoli, Johann G Danzl, Matthias Tomschik, Karl Rössler, and Peter Jonas. Human hippocampal ca3 uses specific functional connectivity rules for efficient associative memory. *Cell*, 188(2): 501–514, 2025.

M. Wołczyk, K.J. Piczak, B. Wójcik, Ł. Pustelnik, P. Morawiecki, J. Tabor, T. Trzciński, and P. Spurek. Continual learning with guarantees via weight interval constraints. volume 162 of *Proceedings of Machine Learning Research*, pp. 23897–23911, 2022.

M. Wortsman, V. Ramanujan, R. Liu, A. Kembhavi, M. Rastegari, J. Yosinski, and A. Farhadi. Supermasks in superposition. volume 2020-December of *Advances in Neural Information Processing Systems*, 2020.

Chi-Hong Wu, Raul Ramos, Donald B Katz, and Gina G Turrigiano. Homeostatic synaptic scaling establishes the specificity of an associative memory. *Current biology*, 31(11):2274–2285, 2021.

Xundong E Wu and Bartlett W Mel. Capacity-enhancing synaptic learning rules in a medial temporal lobe online learning model. *Neuron*, 62(1):31–41, 2009.

Yujie Wu and Wolfgang Maass. A simple model for behavioral time scale synaptic plasticity (btsp) provides content addressable memory with binary synapses and one-shot learning. *Nature communications*, 16(1):342, 2025.

Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Di He, and Zhouchen Lin. Hebbian learning based orthogonal projection for continual learning of spiking neural networks. *arXiv preprint arXiv:2402.11984*, 2024.

Mingkun Xu, Faqiang Liu, Yifan Hu, Hongyi Li, Yuanyuan Wei, Shuai Zhong, Jing Pei, and Lei Deng. Adaptive synaptic scaling in spiking networks for continual learning and enhanced robustness. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Qingsen Yan, Dong Gong, Yuhang Liu, Anton Van Den Hengel, and Javen Qinfeng Shi. Learning bayesian sparse networks with full experience replay for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 109–118, 2022.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3014–3023, 2021.

Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, and Xingwei Wang. Data augmented flatness-aware gradient projection for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5630–5639, 2023.

Guang Yang, Cora Sau Wan Lai, Joseph Cichon, Lei Ma, Wei Li, and Wen-Biao Gan. Sleep promotes branch-specific formation of dendritic spines after learning. *Science*, 344(6188):1173–1178, 2014.

J. Yoon, E. Yang, J. Lee, and S.J. Hwang. Lifelong learning with dynamically expandable networks. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018.

Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. *arXiv preprint arXiv:1902.09432*, 2019.

Friedemann Zenke and Axel Laborieux. Theories of synaptic memory consolidation and intelligent plasticity for continual learning. *arXiv preprint arXiv:2405.16922*, 2024.

Tielin Zhang, Xiang Cheng, Shuncheng Jia, Chengyu T Li, Mu-ming Poo, and Bo Xu. A brain-inspired algorithm that mitigates catastrophic forgetting of artificial and spiking neural networks with low computational cost. *Science Advances*, 9(34):eadi2947, 2023.

Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. *arXiv preprint arXiv:2205.13218*, 2022a.

J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. Ibot: Image bert pre-training with online tokenizer. ICLR 2022 - 10th International Conference on Learning Representations, 2022b.

Xiao Zhou, Renjie Pi, Weizhong Zhang, Yong Lin, Zonghao Chen, and Tong Zhang. Probabilistic bilevel coreset selection. In *International conference on machine learning*, pp. 27287–27302. PMLR, 2022c.

Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9296–9305, 2022.

A Appendix

A.1 Multi-Synaptic Cooperation Network in ANNs

A.1.1 Modeling Multi-Synaptic Neuron

In biological neurons, a single axon can establish multiple synaptic contacts with the same dendritic branch, enabling the signal to influence the target neuron through parallel pathways Turegano-Lopez et al. (2024); Callan et al. (2021). To simulate this structure in artificial neural networks, we represent each inter-neuronal connection not by a single weight but by a vector of parallel synaptic weights. Each element in this vector is trained independently, and their combined effect determines the connection strength. This design remains compatible with existing network architectures, supports standard gradient-based learning, and naturally extends plasticity rules developed for spiking neural networks.

To model multi-synaptic connections in ANNs, we first associate each connection (i, j) with a column vector of synaptic weights $\mathbf{w}_{ij}$:

$$\mathbf{w}_{ij} = \left[w_{ij}^1, \ldots, w_{ij}^P\right]^\top \tag{12}$$

where P is a hyperparameter representing the number of synaptic connections between each pair of neurons. To preserve synaptic diversity and enable distinct optimization of parallel synapses, we

assign different activation functions to different parallel synapses. The input to postsynaptic neuron j from presynaptic neuron i is then defined as:

$$g_{ij}(x_i) \;=\; \sum_{p=1}^{P} \sigma_p\big(w_{ij}^p x_i\big) \tag{13}$$

where σ_p denotes the activation function specific to the p-th synapse. When an input vector $\mathbf{x}$ is given to a fully connected layer, the input to neuron j is computed as

$$z_j \;=\; \sum_{i=1}^{N} g_{ij}(x_i) \;=\; \sum_{i=1}^{N}\sum_{p=1}^{P} \sigma_p\big(w_{ij}^p x_i\big) \tag{14}$$

where N is the number of input neurons in the layer. The value z_j is then passed through a ReLU activation function resulting in the output $y_j = \max(0, z_j)$. Similarly, in convolutional layers, the additional synaptic dimension is incorporated into each filter. As a result, the summation in Eq. (14) also spans spatial locations and input channels.

For local synaptic plasticity modulation, we associate the local activity of each neuron with its output value. The eligibility trace is defined as follows:

$$\tilde{e}[t+1] = \tilde{e}[t] - \frac{\tilde{e}[t]}{\tau} + z_j \tag{15}$$

where τ is a decay time constant. Each synaptic weight is updated using the following learning rule:

$$\Delta w_{ij}^p = -\,\eta \;\cdot\; \big|f_{\text{mod}}\big(\tilde{e}_{ij}\big)\big| \;\cdot\; \frac{\partial \mathcal{L}}{\partial w_{ij}^p} \tag{16}$$

where η is the learning rate. The eligibility trace $\tilde{e}_{ij}$ keeps track of local synaptic activity. The function f_{mod} uses this information to smoothly adjust the synaptic strength—potentiation or depression—similar to how synaptic plasticity is modulated in spiking neural models. When $P = 1$, Eq. (12)–Eq. (16) become the same as in standard ANNs, so this method remains fully compatible with existing implementations.

A.2 Implementation Details

A.2.1 Datasets.

The datasets used in our experiments are summarized in Table 4. We evaluate our approach on four standard continual learning benchmarks: Permuted MNIST (PMNIST), CIFAR-100, TinyImageNet, and 5-Datasets, which are presented in roughly increasing order of dataset complexity.

PMNIST is a variant of the original MNIST dataset consisting of 28×28 grayscale images of handwritten digits. Each task applies a fixed but unique random permutation to the pixel positions, making it a widely adopted benchmark for evaluating robustness in task-incremental learning. For PMNIST, we assign each task 60,000 training and 10,000 testing samples to increase the challenge.

CIFAR-100. CIFAR-100 is an object recognition dataset with 100 natural image classes. Following the protocol in Rebuffi et al. (2017), we partition the dataset into 10 tasks, each comprising 10 disjoint classes with their corresponding images.

TinyImageNet contains 100,000 64×64 color images across 200 classes. We construct 40 sequential tasks by splitting the dataset into 5-way classification problems. For fair comparison, we randomly sample a subset of the original dataset and align the test set with the training set as in Serra et al. (2018).

5-Datasets combines tasks from five diverse datasets: MNIST, SVHN, Fashion-MNIST, CIFAR-10, and NotMNIST, each treated as an independent task. This setting evaluates the model's generalizability under distribution shift and cross-domain learning.

For CIFAR-100 and TinyImageNet, we follow standard settings with 500 training and 100/50 test images per class, respectively.

Table 4: Dataset statistics

Dataset	PMNIST	CIFAR-100	TinyImageNet	5-Datasets
Tasks	10	10	40	5
Classes	10	100	200	/
Training Samples	60,000	50,000	100,000	/
Test Samples	10,000	10,000	10,000	/

A.2.2 Evaluation Metrics

Following Kang et al. (2022); Wortsman et al. (2020); Mallya et al. (2018), we evaluate all methods based on the following metrics:

Accuracy (ACC) measures the average of the final classification accuracy on all tasks:

$$\text{ACC} = \frac{1}{T}\sum_{i=1}^{T} acc_{T,i} \tag{17}$$

where $acc_{T,i}$ is the test accuracy for task i after training on task T.

Backward Transfer (BWT) measures the influence of learning new tasks on the performance of previously learned ones. A negative BWT indicates forgetting, whereas a positive BWT suggests that learning later tasks improved the performance of earlier ones. BWT is computed as:

$$\text{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} (acc_{T,i} - acc_{i,i}) \tag{18}$$

where $acc_{i,i}$ denotes the accuracy of task i immediately after it is learned. A BWT close to zero implies stability, while highly negative values indicate catastrophic forgetting.

Capacity (CAP) measures the amount of network capacity used under each parameter pruning method Kang et al. (2022). It accounts for both the proportion of trainable parameters and the efficiency of binary encoding. The CAP metric is defined as:

$$\text{CAP} = (1 - C) + \frac{(1-\alpha)N}{32} \tag{19}$$

where α is the average mask compression rate (α=0.78), N is the number of tasks, and C is the percentage of non-fixed parameters. A smaller CAP value indicates higher effective network capacity.

Average Order-normalized Performance Disparity (AOPD) measures the robustness of these algorithms under different task orders. Following the protocol of Yoon et al. (2019), we assessed the task order robustness with the Order-normalized Performance Disparity (OPD) metric, which is computed as the disparity between the performance $\bar{A}_t$ of task t on R different task orders: $\text{OPD}_t \triangleq \max\{\bar{A}_t^1, \ldots, \bar{A}_t^R\} - \min\{\bar{A}_t^1, \ldots, \bar{A}_t^R\}$. The average OPD (AOPD) is defined by

$$\text{AOPD} \triangleq \frac{1}{T}\sum_{t=0}^{T-1} \text{OPD}_t \tag{20}$$

A.2.3 Experiment Settings

All experiments were conducted on a Linux server equipped with an Intel Xeon Gold 5220 (2.20 GHz) CPU and two NVIDIA Tesla V100-SXM2 GPUs (32 GB each, driver 535.129.03). Following Kang et al. (2022); Wortsman et al. (2020); Mallya et al. (2018), we use a two-layered MLP with 100 neurons per layer for PMNIST and use a modified version of AlexNet for the CIFAR-100 Split dataset and a reduced ResNet-18 Chaudhry et al. (2019); Saha et al. (2021) for 5-Datasets. For TinyImageNet, we also use the same network architecture Gupta et al. (2020a); Deng et al. (2021), which consists of 4 Conv layers and 3 fully connected layers. For a fair comparison, we follow the experimental setting in Kang et al. (2022); Thapa & Li (2024), and all methods are evaluated under the same multi-head setting with known task labels. The hyperparameter settings are presented in Table 7.

Table 5: Performance deviations of the proposed method and baselines on four datasets.

(a) PMNIST and 10-split CIFAR-100

Network	Method	PMNIST		10-split CIFAR-100	
		ACC (%)	BWT (%)	ACC (%)	BWT (%)
SNN	MTL	0.12	/	0.31	/
	EWCKirkpatrick et al. (2017)	0.51	0.42	0.63	0.57
	HAT Serra et al. (2018)	0.24	0.0	0.37	0.19
	GPM Saha et al. (2021)	0.43	0.34	0.41	0.38
	HLOP Xiao et al. (2024)	0.39	0.21	0.35	0.18
	MSCN	**0.22**	**0.0**	**0.25**	**0.0**
ANN	MTL	0.14	/	0.21	/
	EWCKirkpatrick et al. (2017)	0.56	0.01	0.57	0.49
	GPM Saha et al. (2021)	0.07	0.01	0.48	0.39
	PackNet Mallya et al. (2018)	0.04	0.0	0.41	0.0
	SupSup Wortsman et al. (2020)	0.09	0.0	0.32	0.0
	WSN Kang et al. (2022)	0.07	0.0	0.29	0.0
	TAMiL Bhat et al. (2023)	0.17	0.04	0.36	0.49
	MSCN	**0.19**	**0.0**	**0.23**	**0.0**

(b) TinyImageNet and 5-Datasets

Network	Method	TinyImageNet		5-Datasets	
		ACC (%)	BWT (%)	ACC (%)	BWT (%)
SNN	MTL	0.29	/	0.26	/
	EWCKirkpatrick et al. (2017)	0.59	0.72	0.48	0.65
	HAT Serra et al. (2018)	0.61	0.44	0.33	0.51
	GPM Saha et al. (2021)	0.46	0.29	0.36	0.42
	HLOP Xiao et al. (2024)	0.41	0.23	0.30	0.25
	MSCN	**0.31**	**0.0**	**0.29**	**0.0**
ANN	MTL	0.33	/	0.27	/
	EWCKirkpatrick et al. (2017)	0.44	0.03	0.26	0.02
	GPM Saha et al. (2021)	0.42	0.31	0.20	0.01
	PackNet Mallya et al. (2018)	0.35	0.0	0.12	0.0
	SupSup Wortsman et al. (2020)	0.40	0.0	0.21	0.0
	WSN Kang et al. (2022)	0.34	0.0	0.13	0.0
	TAMiL Bhat et al. (2023)	0.31	0.13	0.22	0.08
	MSCN	**0.27**	**0.0**	**0.21**	**0.0**

A.2.4 ARCHITECTURAL DETAILS

Two-layered MLP: In conducting the PMNIST experiments, we are following the exact setup as denoted by Saha et al. (2021) fully-connected network with two hidden layers of 100 neurons Lopez-Paz & Ranzato (2017).

Modified AlexNet: For the split CIFAR-100 dataset, we use a modified version of AlexNet similar to Gupta et al. (2020b); Saha et al. (2021).

Table 7: Experiment settings and hyperparameter configurations for different datasets

Dataset	PMNIST	10-split CIFAR-100	TinyImageNet	5-Datasets
learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
dropout rate	0.2	0.2	0.5	0.2
epochs	5	200	10	100
batch size	10	64	10	64
warmup ratio	0.05	0.05	0.05	0.05
optimizer	AdamW	AdamW	AdamW	AdamW
weight decay	1×10^{-2}	1×10^{-2}	1×10^{-2}	1×10^{-2}
architecture	Two-layered MLP	Modified AlexNet	4 Conv layers and 3 Fully connected layers	Reduced ResNet18

4 Conv layers and 3 Fully connected layers: For TinyImageNet, we use the same network architecture as Gupta et al. (2020b); Deng et al. (2021).

Reduced ResNet18: In conducting the 5-Dataset experiments, we use a smaller version of ResNet18 with three times fewer feature maps across all layers as denoted by Lopez-Paz & Ranzato (2017).

A.2.5 List of main notations

In Table 8, we list the main notations used in this paper, together with brief explanations, enabling quick reference to the meaning of each symbol.

Table 8: List of main notations

Notation	Description
m_j	Binary mask selecting active synapses for task j
M_{j-1}	Accumulated mask of all previous tasks up to $j-1$
r	Learnable relevance score for each synapse
c	Layer-wise capacity ratio for subnet selection
P	Number of parallel synapses in each connection (synapse count)
w_{ip}	Weight of the p-th synapse from presynaptic neuron i
N	Number of presynaptic neurons
$V(t)$	Membrane potential of a spiking neuron at time t
τ_m	Membrane time constant in LIF neurons
V_{rest}	Resting potential of the spiking neuron
$I(t)$	Total synaptic input current at time t
$\text{PSP}_{ip}(t)$	Postsynaptic potential from synapse p of neuron i
$\tau_{s,ip}$	Decay constant of the p-th parallel synapse of neuron i
$K_{ip}(t)$	Synaptic kernel of the p-th synapse of neuron i
t_{ip}^f	Arrival time of the f-th spike at synapse (i,p)
$\tilde{e}$	Eligibility trace representing local synaptic activity
τ	Decay time constant of the eligibility trace
$f_{\text{mod}}(\tilde{e})$	Modulation function for $\tilde{e}$

A.3 Extra Experiments

A.3.1 Integration with different types of methods

To further evaluate the applicability of our MSCN, we integrated it with different types of methods and conducted additional experiments on both 10-split CIFAR-100 and TinyImageNet under the same parameter budget. As shown in Table 9, integrating MSCN into regularization-based and replay-based methods consistently improves BWT. Notably, combining MSCN with ER on TinyImageNet improves BWT by 53.8% (an absolute decrease of 10.69). The observed BWT improvements are attributed to the high capacity efficiency of MSCN, which arises from the independent optimization of

Table 9: Integration with different types of methods.

Type	Method	10-split CIFAR-100		TinyImageNet	
		ACC (%) ↑	BWT (%) ↑	ACC (%) ↑	BWT (%) ↑
regularization-based	EWC	72.77 (± 0.57)	-3.59 (± 0.49)	64.51 (± 0.44)	-0.04 (± 0.03)
	EWC+MSCN	**73.26** (± **0.66**)	**-2.78** (± **0.19**)	**64.98** (± **0.54**)	**-0.03** (± **0.01**)
replay-based	ER	70.07 (± 0.73)	-7.70 (± 0.59)	48.32 (± 0.91)	-19.86 (± 0.70)
	ER+MSCN	**71.13** (± **0.62**)	**-5.24** (± **0.51**)	**49.26** (± **0.84**)	**-9.17** (± **0.55**)
architecture-based	Bayesian	75.57 (± 0.38)	0.00 (± 0.00)	73.93 (± 0.36)	0.00 (± 0.00)
	Bayesian+MSCN	**76.48** (± **0.34**)	**0.00** (± **0.00**)	**74.56** (± **0.33**)	**0.00** (± **0.00**)

multiple parallel synapses, as demonstrated in Fig. 7. Such higher capacity efficiency has been shown to reduce catastrophic forgetting Hung et al. (2019a); Mirzadeh et al. (2022); Farajtabar et al. (2020). Meanwhile, the modulation mechanism further enhances this property by depressing the effect of noisy samples and strengthening learning on clean ones. In contrast, for architecture-based methods, BWT remains zero because the weights of past tasks are frozen, which is exactly as expected. At the same time, when our MSCN is incorporated, all three types of methods achieve improved accuracy. These additional experiments further highlight the robustness of our MSCN.

A.3.2 Layer-Wise Capacity Analysis

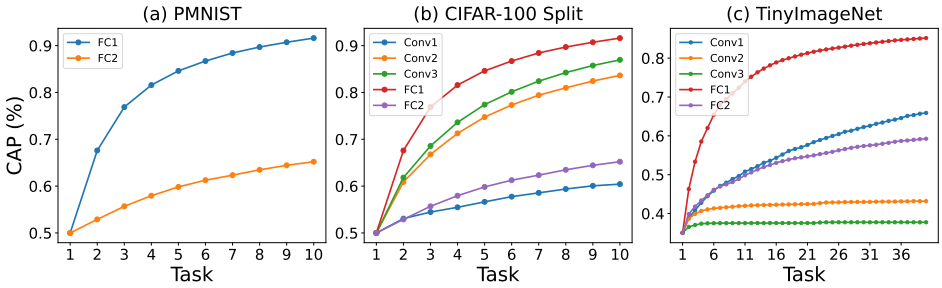

Figure 8: Synaptic capacity usage as the number of tasks increases on three benchmarks. Each curve shows the percentage of active synapses per layer as tasks are incrementally introduced.

As shown in Fig. 8, we analyze the synaptic capacity usage across three benchmarks: PMNIST, CIFAR-100 Split, and TinyImageNet. For each dataset, we measure the percentage of utilized synapses in each layer as tasks are incrementally learned. Across all three datasets, we observe a consistent pattern: capacity usage increases rapidly during the initial tasks, then gradually slows down as more tasks are introduced. This effect is particularly pronounced in the fully connected layers, such as FC1, which tend to accumulate more synaptic updates compared to early convolutional layers. The underlying reason is that the model needs to allocate new synaptic resources to encode novel task-specific features at the beginning. However, as training progresses, many new tasks can be handled by reusing synapses that represent similar features, reducing the need for additional capacity. This confirms the model's ability to reuse past representations more effectively as it acquires more knowledge, leading to a slower growth in capacity usage over time.

A.3.3 Analysis of Synapse Count

Fig. 9 illustrates the relationship between synapse count (denoted as P) and average accuracy as the number of tasks increases, evaluated on PMNIST and CIFAR-100 Split. We vary the number of synapses per connection across five settings (P=2, 4, 6, 8, 10) and track model performance throughout the incremental learning process. We observe that on PMNIST, accuracy remains high across all configurations; however, larger synapse counts (e.g., P=8, 10) tend to deliver more stable performance over multiple tasks. On CIFAR-100 Split, the benefits of increased synaptic capacity become more evident: higher P values consistently result in better average accuracy, particularly as

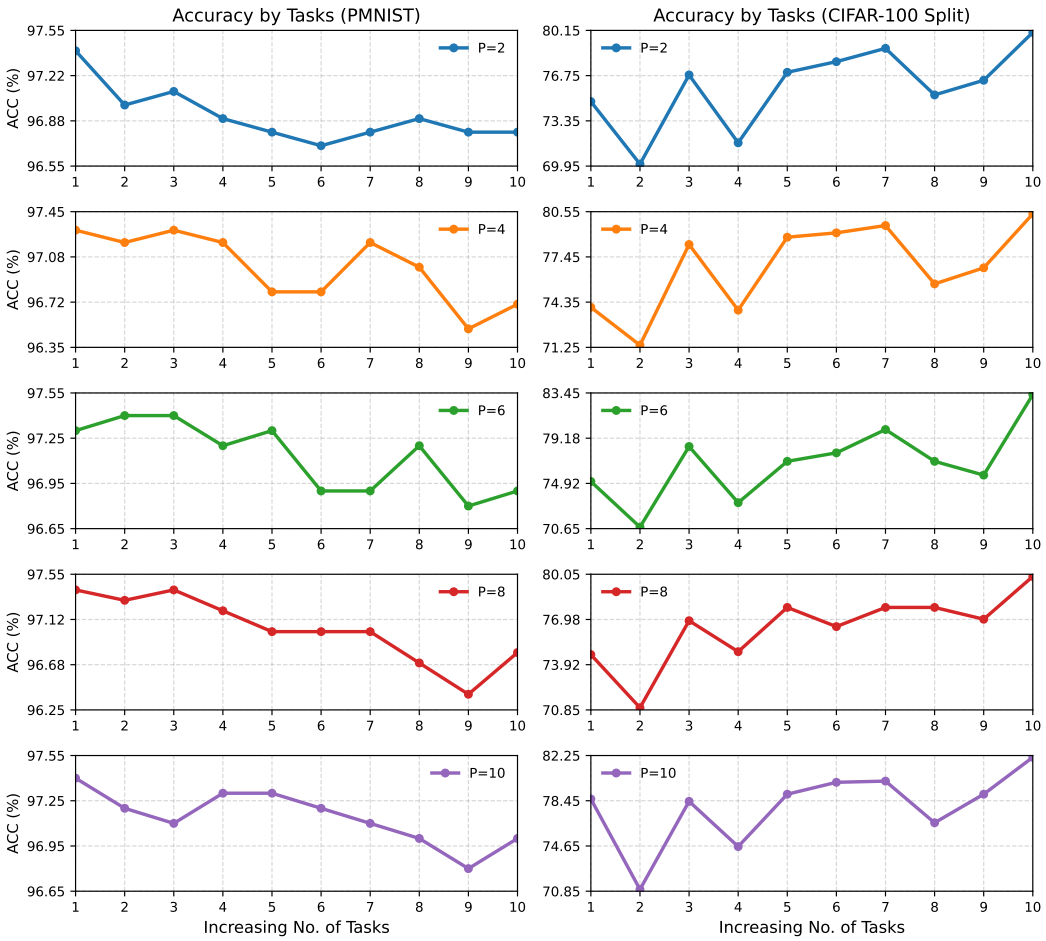

Figure 9: Average accuracy with increasing number of tasks under different synapse counts.

the number of tasks grows. These results confirm that synaptic multiplicity enhances the model's ability to retain knowledge and generalize over longer task sequences.

A.3.4 Order robustness analysis

To further investigate the robustness of our method to task permutations, we conduct additional experiments on the CIFAR-100 Split benchmark using five randomly shuffled task orders. Fig. 10 presents the per-task accuracy across all 10 tasks for three representative baselines—EWC Kirkpatrick et al. (2017), GPM Saha et al. (2021), and WSN Kang et al. (2022)—alongside our proposed MSCN. We observe that EWC (Fig. 10a) and GPM (Fig. 10b) are highly sensitive to task order, exhibiting considerable variance in accuracy for the same task index across different permutations. In contrast, WSN (Fig. 10c) achieves more stable performance, though moderate fluctuations persist, particularly on later tasks. Notably, our method, MSCN (Fig. 10d), maintains consistently high accuracy across all permutations and task indices, with significantly reduced inter-order variance. These results show that MSCN is robust to task order, ensuring stability in dynamic environments.

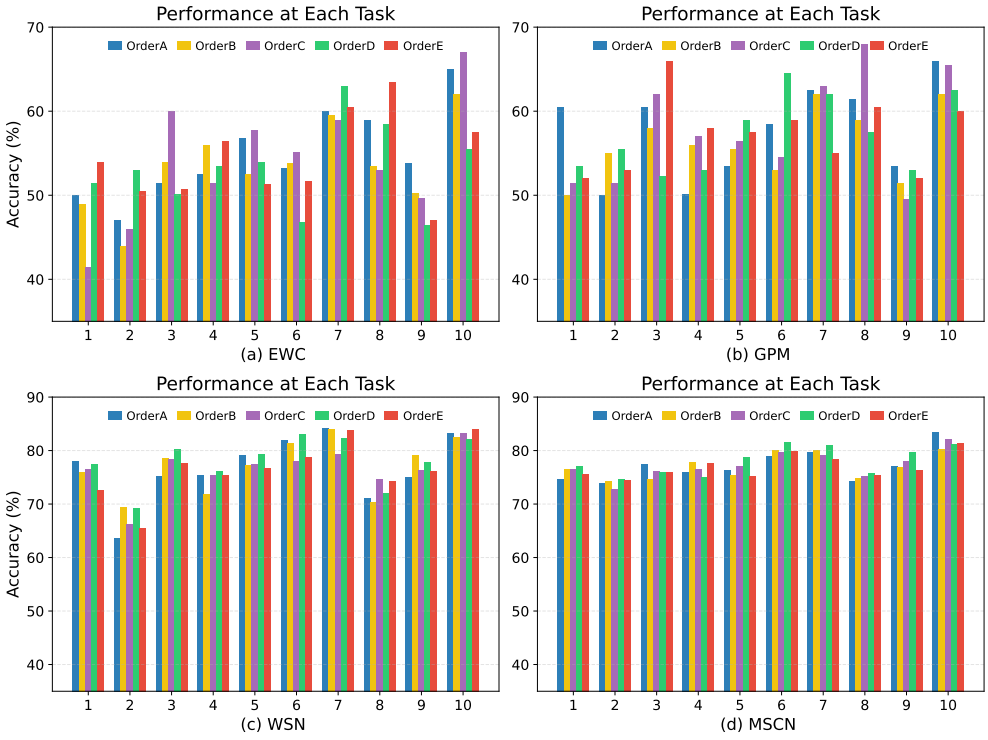

Figure 10: Task order robustness comparison on CIFAR-100 Split. Bar plots show per-task accuracy under five different task sequences.