MATRYOSHKA CONTINUAL LEARNING: NESTED REPRESENTATION MITIGATES FORGETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learners are prone to catastrophic forgetting, where learning new tasks overwrites the representations that earlier tasks depend on. To address this, we propose Matryoshka Continual Learning (MCL), which structures a single embedding as a sequence of nested sub-embeddings like Matryoshka dolls. The first few dimensions constitute a compact representation that encodes coarse features and is trained to support the classification task on its own. Each wider prefix contains the preceding one and extends it with progressively finer, constructive detail, yielding a coarse-to-fine hierarchy within a single vector. Embedding is learned through a teacher–student setup, where the teacher has access to privileged knowledge, forcing a student with the same architecture to follow its lead without access to that knowledge. Because MCL changes the representation rather than the learning rule, it plugs into existing CL methods. We evaluate MCL against 9 baselines on Seq-CIFAR-100, Seq-Tiny-ImageNet, and the Robotic Tactile Internet (RTI) datasets under Task-Incremental Learning (Task-IL) and Class-Incremental Learning (Class-IL) settings. In Task-IL, MCL attains 87.29% accuracy on Seq-CIFAR-100 and 77.59% on Seq-Tiny-ImageNet, 0.32 and 1.97 points above the second-best method, respectively. In Class-IL, MCL ranks the second-best method in accuracy on both benchmarks, while achieving the highest plasticity–stability score.

1 INTRODUCTION

Continual learning requires a model to learn a sequence of tasks after the earlier training data are no longer available (Wang et al., 2024). Each update must improve performance on the current task without erasing the decision structure learned from previous ones. This stability–plasticity conflict is difficult even in Task-Incremental Learning (Task-IL), where the task identity is known at test time, and becomes more severe in Class-Incremental Learning (Class-IL), where one classifier must separate every class seen so far without knowing the task identity (Vahedifar et al., 2026b).

Most continual learning methods protect old knowledge through an auxiliary mechanism. Regularization-based methods estimate which parameters mattered for earlier tasks and penalize changes to them (Kirkpatrick et al., 2017; Zenke et al., 2017; Li & Hoiem, 2018; Vahedifar & Zhang, 2026b). They are computationally efficient but scale poorly. Memory-based methods introduce an exemplar buffer from previous tasks which deviates from the strict continual learning setting (Rebuffi et al., 2017; Buzzega et al., 2020; Douillard et al., 2022). They enforce a strict rule against accessing prior data in a continual learning setup, introducing privacy issues into the system. On the other hand, Buffer-free methods focus mostly on networks and architectural growth, but they are expensive to tune and train (Kang et al., 2022; Zajac et al., 2024; Vahedifar et al., 2026a; Rusu et al., 2022; Sokar et al., 2021). To overcome the shortcomings of earlier methods, we introduce Matryoshka Continual Learning (MCL), which takes a different approach by changing the representation itself. Inspired by Kusupati et al. (2022), MCL trains a single embedding whose nested prefixes encode features from coarse to fine. It therefore learns continually with fixed backbone across tasks and without storing data from earlier tasks. Figure 1 illustrates how the nested prefixes support classification.

MCL trains every prefix on the current task and distills the old-class predictions of a frozen checkpoint from the previous task at every width. The distillation weight follows each prefix's reach. The leading coordinates feed every classifier in the nest, so narrow prefixes are anchored most strongly. The outer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

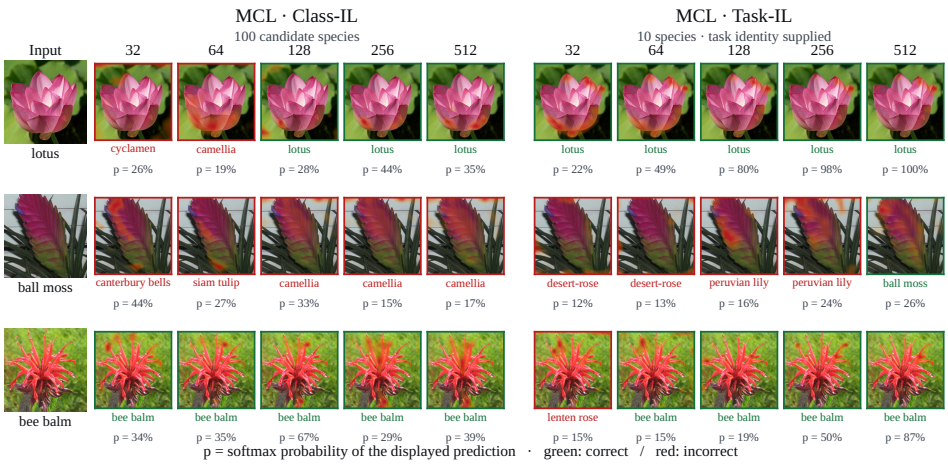

Figure 1: Qualitative figures for nested effects in Task-IL and Class-IL.

coordinates feed only into the widest classifiers, so they are more weakly constrained and absorb task-specific changes. We call this scheme Self-Distillation (SD). Classical knowledge distillation (Hinton et al., 2015) distills knowledge from a larger teacher to a smaller student. However, in SD, the teacher and the student are the same network at different points in training. The teacher has privileged knowledge of earlier tasks, and the student must imitate the teacher's predictions without access to that knowledge. Our contributions can be summarized:

- We introduce a buffer-free, plug-and-play method that enables nesting embeddings to overcome catastrophic forgetting in continual learning.
- We introduce width-dependent self-distillation for continual learning. A frozen teacher with privileged knowledge of earlier tasks anchors each prefix in proportion to its reach, and the student follows it without access to that knowledge.
- We compared MCL with 9 buffer-free continual learning methods across both Task-IL and Class-IL scenarios. In addition, we tested its real-world practicality on the Robotic Tactile Internet (RTI) dataset, which was created to simulate a Physical Artificial Intelligence (PAI) scenario.

2 Related Work

Continual learning methods are commonly distinguished by how they balance stability on previous tasks with plasticity for new ones. Regularization-based approaches retain a fixed architecture and constrain parameter updates using an estimate of parameter importance. Elastic Weight Consolidation (EWC) penalizes changes to parameters based on a diagonal approximation of the Fisher information matrix (Kirkpatrick et al., 2017), whereas Synaptic Intelligence (SI) estimates the importance of each parameter based on its contribution along the optimization trajectory (Zenke et al., 2017). Learning without Forgetting (LwF) instead regularizes the model at the functional level by distilling predictions from the previous model using current-task inputs (Li & Hoiem, 2018). These approaches are simple and buffer-free, but their importance estimates or distillation targets can become inaccurate over long task sequences, producing either insufficient protection of earlier knowledge or excessive constraints on learning new tasks.

A second family mitigates interference by assigning different parameters or computational paths to different tasks. Progressive Neural Networks (PNNs) add a frozen network column for every task and transfer features through lateral connections (Rusu et al., 2022); Hard Attention to the Task (HAT) learns task-dependent attention masks (Serrà et al., 2018). Winning Sub- Network (WSN) similarly learns binary task-adaptive subnetworks whose selected parameters are protected from subsequent updates (Kang et al., 2022). Fixed-capacity sparse methods such as SpaceNet (Sokar et al., 2021) and Neuro-Inspired Stability-Plasticity Adaptation (NISPA) (Gurbuz & Dovrolis, 2022) preserve stable pathways while pruning, rewiring, or reallocating the remaining capacity. More recently,

Shapley Neuron Values (SNV) ranks neurons using approximate Shapley values and freezes those most important for each task (Vahedifar et al., 2026a), whereas Prediction Error-based Classification (PEC) addresses Class-IL through class-specific predictors whose prediction errors against a frozen random target network yield class scores (Zajac et al., 2024). Although these methods can strongly reduce forgetting, they may require task identities, stored masks (Kang et al., 2022), multiple heads or modules (Zajac et al., 2024), expensive importance estimation (Vahedifar et al., 2026a), or progressively consume the network's available capacity (Rusu et al., 2022).

Replay-based methods remain strong baselines for Class-IL: Incremental Classifier and Representation Learning (iCaRL) stores representative exemplars and class prototypes (Rebuffi et al., 2017), Gradient Episodic Memory (GEM) constrains gradients using episodic memories (Lopez-Paz & Ranzato, 2017), and Dark Experience Replay (DER) and its extension DER++ combine rehearsal with stored logits and knowledge distillation (Buzzega et al., 2020). Their effectiveness, however, depends on retaining previous samples or generated surrogates, introducing storage, privacy, and buffer-management requirements. A separate recent line exploits pretrained vision transformers (ViTs): Learning to Prompt (L2P) (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), and COntinual Decomposed Attention-based Prompting (CODA-Prompt) (Smith et al., 2023) learn prompt pools or decomposed prompts, while Random Projections and Pre-trained models for Continual learning (RanPAC) uses frozen pretrained features, random projections, and class prototypes (McDonnell et al., 2023), and Slow Learner with Classifier Alignment (SLCA) combines slow representation updates with post-hoc classifier alignment (Zhang et al., 2023). These methods demonstrate the value of strong pretraining but operate in a different regime from continual learning with a model trained from scratch.

3 Methodology: Matryoshka Continual Learning

Continual learning methods typically apply a single mechanism uniformly to the feature vector shared across all coordinates. MCL instead imposes an order on the coordinates, so that the degree of stability varies along the vector. Figure 2 shows the main idea.

Problem formulation. A learner observes tasks $t = 1, \dots, T$ in sequence. Task t provides a training set $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{n_t}$ drawn i.i.d. from a distribution P_t over $\mathcal{X} \times \mathcal{C}_t$, with pairwise disjoint label sets, $\mathcal{C}_t \cap \mathcal{C}_k = \emptyset$ for $t \neq k$. We write $\mathcal{O}_t = \bigcup_{k<t} \mathcal{C}_k$ for the classes observed before task t ($\mathcal{O}_1 = \emptyset$). The model is $h = g \circ f_\theta$, with backbone $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and a linear or cosine head g with weights $W \in \mathbb{R}^{|\mathcal{O}_{t+1}| \times d}$. While training on task t, the learner observes $\mathcal{D}_t$ and a state $\mathcal{S}_{t-1}$ carried over from the previous task, but no samples from $\mathcal{D}_k$ for $k < t$. MCL uses

$$\mathcal{S}_{t-1} = \big(\theta_{t-1},\, W_{t-1},\, \{\mu_c\}_{c \in \mathcal{O}_t}\big), \qquad (1)$$

the frozen parameters of the previous model. And for Class-IL only, one feature mean $\mu_c \in \mathbb{R}^d$ per observed class. The state contains no exemplars, no generative model, and no per-parameter importance estimates. After task t, the learner should minimize the average risk over all tasks seen so far,

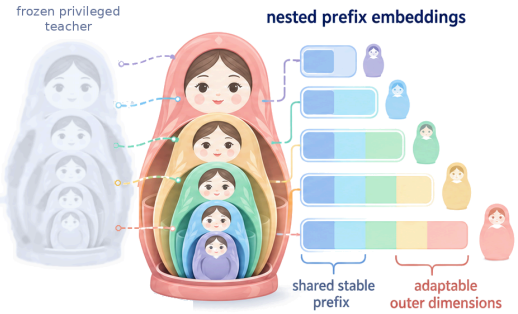

Figure 2: MCL nests coarse-to-fine prefixes in a single embedding. A frozen copy of the previous model acts as a privileged teacher and anchors each prefix, most strongly the narrow shared prefix, while the outer dimensions adapt to the new task.

$$\mathcal{R}_t(h) = \frac{1}{t} \sum_{k=1}^{t} \mathbb{E}_{(x,y)\sim P_k} \big[\mathbb{1}\{\hat{y}(x) \neq y\}\big], \qquad (2)$$

although only $\mathcal{D}_t$ is observed. Minimizing the empirical risk on $\mathcal{D}_t$ alone leaves f_θ unconstrained on the support of P_k for $k < t$; the resulting increase in the $k < t$ terms of Eq. 2 is catastrophic forgetting. A method must therefore trade *plasticity*, the decrease of the $k = t$ term, against *stability*, the preservation of the terms $k < t$. The two standard protocols differ only in the prediction rule $\hat{y}$.

Let $z_c(x)$ be the logit of class c and $k(x)$ the task of test input x:

$$\text{Task-IL: } \hat{y}(x) = \arg\max_{c \in \mathcal{C}_{k(x)}} z_c(x), \qquad \text{Class-IL: } \hat{y}(x) = \arg\max_{c \in \mathcal{O}_{t+1}} z_c(x). \tag{3}$$

In Task-IL the candidate set has size $|\mathcal{C}_{k(x)}|$ and is fixed; in Class-IL it grows with t, and old and new classes compete directly. Class-IL is therefore the harder protocol.

3.1 The Matryoshka Doll

A Matryoshka doll contains a smaller doll, which in turn contains a smaller one. Nested representation learning in this work applies this structure to a feature vector: several classifiers read nested prefixes of the vector, rather than a single classifier reading all d coordinates.

Let $\mathcal{M} = \{m_1 < \cdots < m_K\}$ with $m_K = d$ be the prefix widths; we call the prefix of width m a *doll* and use $\mathcal{M} = \{32, 64, 128, 256, 512\}$ ($K = 5$, $d = 512$). Writing $f_{1:m}(x)$ for the first m coordinates of $f_\theta(x)$. The doll of width m produces

$$z^{(m)}(x) \;=\; \begin{cases} W_{:,1:m}\, f_{1:m}(x) + b, & \text{linear head (Task-IL),} \\ s\, \dfrac{W_{:,1:m}}{\|W_{:,1:m}\|_{\text{row}}}\, \dfrac{f_{1:m}(x)}{\|f_{1:m}(x)\|}, & \text{cosine head, scale } s \text{ (Class-IL),} \end{cases} \tag{4}$$

where both norms are taken over the width-m prefix and the cosine head has no bias. All dolls share a single matrix W, so K classifiers share the same parameters. When task t arrives, W is extended by $|\mathcal{C}_t|$ rows; all rows remain trainable. Training all dolls jointly makes the coordinate order meaningful: the first m_1 coordinates must classify on their own, each wider doll must improve on the one it contains by including more embeddings into decision. Therefor, information is consequently arranged from coarse and shared at the front to fine and specific at the back.

The nested head does more than share parameters: it makes some coordinates harder to change than others. What decides this is how many dolls read each coordinate. A coordinate in the first 32 receives five contributions, one per doll; a coordinate beyond 256 receives one. Consequently, high-reach coordinates move mainly along directions they share. This favors a structure that transfers across tasks at the front of the vector, leaving the back free to absorb task-specific information. The ordering is a property of the head: it is not estimated from data.

Reach. Coordinate j is read by every doll of width at least j. We call the number of such dolls the *reach* of j,

$$r(j) = \big|\{m \in \mathcal{M} : m \geq j\}\big|. \tag{5}$$

For $\mathcal{M} = \{32, 64, 128, 256, 512\}$, reach decreases along the vector:

Coordinates j	1–32	33–64	65–128	129–256	257–512
Reach $r(j)$	5	4	3	2	1

Reach counts gradient contributions. Any loss built from per-doll terms sends coordinate j exactly $r(j)$ gradient contributions.

Proposition 3.1 (Gradient decomposition). *Let $\mathcal{L} = \sum_{m \in \mathcal{M}} a_m\, \mathcal{L}^{(m)}$, where each $\mathcal{L}^{(m)}$ depends on the features only through $z^{(m)}$. For either head in Eq. 4,*

$$\frac{\partial \mathcal{L}}{\partial f_j} \;=\; \sum_{m \in \mathcal{M},\, m \geq j} a_m\, \frac{\partial \mathcal{L}^{(m)}}{\partial f_j}. \tag{6}$$

Proof. $z^{(m)}$ depends on f_j if and only if $j \leq m$. For the cosine head, this still holds, because the prefix norm involves only $f_{1:m}$. □

3.2 Self-Distillation with Privileged Information

Self-distillation trains a model and then imitates its own earlier self-information. We freeze a copy of the network at the end of task $t-1$. Then we treat that copy as the teacher for task t. Subsequently, every image in the current batch goes through both networks; therefore, we penalize the divergence

between their predictions over the old classes $\mathcal{O}_t$. This idea works because the teacher does not need to see a previous class to provide useful information about it. For example, it can see a photo of a dog, and the old-class logits light up in a particular pattern. That pattern is a readout of where the old decision boundaries sit, so copying the pattern keeps the boundaries where they were. Constraining outputs also buys more freedom than constraining weights. The elegancy of self-distillation is that, unlike regularization, which forces weights not to change, distillation lets them change arbitrarily where they desire, as long as their answers on the old classes stay put.

Let $\mathcal{A}_t$ denote the classes scored by the cross-entropy. On task t we minimize

$$\mathcal{L} \;=\; \underbrace{\textstyle\sum_{m\in\mathcal{M}} c_m\, \ell\big(z^{(m)}|_{\mathcal{A}_t},\, y\big)}_{\text{(i) multi-doll cross-entropy}} + \lambda_{\mathrm{S}} \underbrace{\textstyle\sum_{m\in\mathcal{M}} w_m\, \tau^2\, \mathrm{KL}\big(\sigma(\tilde{z}^{(m)}|_{\mathcal{O}_t}/\tau)\,\big\|\,\sigma(z^{(m)}|_{\mathcal{O}_t}/\tau)\big)}_{\text{(ii) doll-weighted self-distillation}} \tag{7}$$

where $z|_{\mathcal{B}}$ restricts logits to a class subset $\mathcal{B}$, $\ell(\cdot, y)$ is the cross-entropy with y re-indexed into $\mathcal{B}$, σ is the softmax, and τ its temperature. In both protocols the softmax in (ii) is taken jointly over $\mathcal{O}_t$, and both terms are averaged over the batch. Each term carries a single scalar: c_m and w_m are fixed, and λ_{S} sets the stability–plasticity trade-off.

(i) Multi-doll cross-entropy. This term enforces the nesting. If only the widest doll is scored, nothing requires the first m_1 coordinates to be discriminative; scoring every doll requires each prefix to be a classifier in its own right, which produces the ordering of Sec. 3.1. We set $c_m = 1/K$, the mean rather than the sum, so that the classification gradient matches the scale received by a single-head baseline.

(ii) Doll-weighted self-distillation. The temperature τ raises the entropy of both distributions, so that non-maximal logits, which carry most of the information about the old decision boundaries, contribute to the loss. The weights w_m set the anchoring strength per doll:

$$w_m \;=\; \frac{(d/m)^{\beta}}{\sum_{m'\in\mathcal{M}} (d/m')^{\beta}}, \qquad \beta \geq 0. \tag{8}$$

For $\beta = 0$ all dolls are weighted equally ($w_m = 1/K$); increasing β shifts weight to narrow dolls. This follows Proposition 3.1: drift in the m_1-prefix propagates to all K classifiers, whereas the outermost coordinates affect one. For $\beta = 1$, $w_m \propto d/m$ and $w_{32}/w_{512} = 16$. The single scalar β thus converts the nesting order into a stability schedule.

4 Results and Discussion

Datasets and Architecture. We created a new dataset, the Robotic Tactile Internet(RTI) dataset, containing haptic tele-manipulation for real-world scenarios or PAI scenarios, with details explained in Appendix C. Additionally, we use Seq-CIFAR-100 (Krizhevsky & Hinton, 2009) and Seq-Tiny-ImageNet (Le & Yang, 2015). All methods train with the Adam optimizer (Kingma & Ba, 2015) without weight decay, a batch size of 64, and at most 200 epochs per task, with early stopping on validation loss (patience 15) and restoration of the best epoch. We use ResNet-18 (He et al., 2016) and evaluate on both Task-IL and Class-IL. We follow the Generalizable Two-phase Evaluation Protocol (GTEP) (Cha & Cho, 2025): hyperparameters are selected on half of the classes and then transferred, without further tuning, to a disjoint set of the remaining evaluation classes. As noted by GTEP tuning hyperparameters on validation data spanning the entire task sequence violates the sequential-access assumption of continual learning and leaks future-task information, yielding biased results.

Baselines and metrics. We compare MCL against Stochastic Gradient Descent (SGD) (lower bound) and joint training (upper bound). Additionally, EWC (Kirkpatrick et al., 2017), SI (Zenke et al., 2017), LwF (Li & Hoiem, 2018), WSN (Kang et al., 2022), PEC (Zajac et al., 2024), SpaceNet (Sokar et al., 2021), NISPA (Gurbuz & Dovrolis, 2022), Unified Continual Learning and Machine Unlearning (UniCLUN) (Chatterjee et al., 2024), and SNV(Vahedifar et al., 2026a). WSN is evaluated only in Task-IL, whereas PEC is evaluated only in Class-IL. We report final average accuracy (ACC), backward transfer (BWT), forward transfer (FWT), and the harmonic plasticity–stability score (PS) as the primary CL metrics; they are defined in Appendix D. Additionally, measurement costs such as GFLOP, GPU hours, Memory, parameters, Peak GPU time, and inference time are provided in Appendix E. Each experiment uses a single NVIDIA RTX PRO 6000A GPU, and all results are averaged over 3 runs.

Table 1: Continual-learning results. Best methods are **highlighted**; second-best are underlined.

Method	Task-IL ACC↑	Task-IL BWT↑	Task-IL FWT↑	Task-IL PS↑	Class-IL ACC↑	Class-IL BWT↑	Class-IL FWT↑	Class-IL PS↑
Seq-CIFAR-100								
Joint training (upper bound)	$87.89_{\pm 1.25}$	–	–	–	$65.43_{\pm 2.40}$	–	–	–
EWC	$70.88_{\pm 2.65}$	$-5.79_{\pm 1.66}$	$\mathbf{+1.19_{\pm 1.14}}$	$79.35_{\pm 1.45}$	$8.39_{\pm 0.66}$	$-84.46_{\pm 2.30}$	$-2.42_{\pm 0.54}$	$26.17_{\pm 3.21}$
SI	$67.34_{\pm 1.14}$	$-11.40_{\pm 1.55}$	$+0.29_{\pm 0.40}$	$79.10_{\pm 0.79}$	$8.56_{\pm 0.49}$	$-85.10_{\pm 1.69}$	$-2.38_{\pm 0.61}$	$25.34_{\pm 2.36}$
LwF	$77.81_{\pm 2.49}$	$-7.96_{\pm 1.57}$	$-1.53_{\pm 3.27}$	$86.42_{\pm 1.77}$	$10.06_{\pm 1.79}$	$-80.02_{\pm 1.20}$	$-2.42_{\pm 0.54}$	$32.15_{\pm 1.59}$
WSN	$84.70_{\pm 1.84}$	$\mathbf{+0.00_{\pm 0.00}}$	$-0.30_{\pm 1.65}$	$\mathbf{89.42_{\pm 1.62}}$	–	–	–	–
PEC	–	–	–	–	$32.84_{\pm 1.25}$	$\mathbf{-12.70_{\pm 0.55}}$	$\mathbf{+0.00_{\pm 0.00}}$	$56.43_{\pm 0.38}$
SpaceNet	$63.61_{\pm 1.32}$	$-1.62_{\pm 0.29}$	$-0.20_{\pm 0.69}$	$71.68_{\pm 1.10}$	$6.77_{\pm 0.43}$	$-65.46_{\pm 1.45}$	$-1.86_{\pm 0.19}$	$45.26_{\pm 0.86}$
NISPA	$80.02_{\pm 1.55}$	$\mathbf{+0.00_{\pm 0.00}}$	$-2.70_{\pm 1.88}$	$85.82_{\pm 1.44}$	$8.24_{\pm 0.54}$	$-78.89_{\pm 1.14}$	$-1.93_{\pm 0.26}$	$33.30_{\pm 1.30}$
SNV	$86.97_{\pm 0.73}$	$\mathbf{+0.00_{\pm 0.00}}$	$+0.49_{\pm 0.19}$	$86.50_{\pm 1.57}$	$\mathbf{36.18_{\pm 2.06}}$	$-18.62_{\pm 1.21}$	$\mathbf{+0.00_{\pm 0.00}}$	$53.84_{\pm 2.67}$
UniCLUN	$69.46_{\pm 2.70}$	$-21.36_{\pm 1.86}$	$-0.14_{\pm 0.44}$	$82.11_{\pm 1.46}$	$32.75_{\pm 4.04}$	$-58.84_{\pm 7.47}$	$-1.91_{\pm 0.66}$	$55.18_{\pm 6.17}$
MCL (ours)	$\mathbf{87.29_{\pm 5.70}}$	$-7.65_{\pm 4.53}$	$+0.64_{\pm 2.25}$	$86.51_{\pm 3.23}$	$33.22_{\pm 2.55}$	$-41.19_{\pm 5.33}$	$-0.23_{\pm 0.44}$	$\mathbf{60.14_{\pm 1.56}}$
SGD (lower bound)	$50.81_{\pm 2.99}$	$-34.44_{\pm 2.83}$	$+1.01_{\pm 2.51}$	$70.99_{\pm 2.21}$	$8.72_{\pm 0.70}$	$-85.85_{\pm 1.37}$	$-2.42_{\pm 0.54}$	$24.29_{\pm 2.00}$
Seq-Tiny-ImageNet								
Joint training (upper bound)	$81.36_{\pm 0.42}$	–	–	–	$57.24_{\pm 0.53}$	–	–	–
EWC	$58.90_{\pm 1.16}$	$-11.85_{\pm 2.13}$	$0.57_{\pm 0.41}$	$75.29_{\pm 0.94}$	$7.48_{\pm 0.26}$	$-74.82_{\pm 0.17}$	$-1.01_{\pm 0.27}$	$37.71_{\pm 0.25}$
SI	$63.85_{\pm 2.14}$	$-3.21_{\pm 1.30}$	$-0.82_{\pm 0.49}$	$76.23_{\pm 1.34}$	$7.48_{\pm 0.28}$	$-72.74_{\pm 0.10}$	$-1.01_{\pm 0.27}$	$39.70_{\pm 0.09}$
LwF	$67.39_{\pm 1.49}$	$-6.46_{\pm 1.22}$	$0.08_{\pm 0.58}$	$80.46_{\pm 1.05}$	$8.04_{\pm 0.20}$	$-71.42_{\pm 0.49}$	$-0.96_{\pm 0.27}$	$41.05_{\pm 0.33}$
WSN	$73.68_{\pm 0.10}$	$0.00_{\pm 0.00}$	$0.50_{\pm 0.70}$	$\mathbf{83.00_{\pm 0.50}}$	–	–	–	–
PEC	–	–	–	–	$16.18_{\pm 0.36}$	$-27.15_{\pm 1.88}$	$-1.42_{\pm 0.10}$	$36.72_{\pm 1.33}$
SpaceNet	$12.17_{\pm 0.46}$	$\mathbf{0.10_{\pm 0.20}}$	$0.00_{\pm 0.60}$	$1.01_{\pm 0.54}$	$1.01_{\pm 0.01}$	$\mathbf{-3.40_{\pm 0.60}}$	$-1.10_{\pm 0.10}$	$2.30_{\pm 0.10}$
NISPA	$50.50_{\pm 3.74}$	$-1.10_{\pm 1.50}$	$0.60_{\pm 0.90}$	$61.50_{\pm 4.90}$	$5.46_{\pm 0.29}$	$-49.70_{\pm 4.40}$	$-0.70_{\pm 0.10}$	$49.70_{\pm 0.20}$
SNV	$75.62_{\pm 0.51}$	$0.00_{\pm 0.00}$	$0.60_{\pm 0.10}$	$80.10_{\pm 0.30}$	$\mathbf{30.70_{\pm 1.73}}$	$-17.30_{\pm 2.40}$	$\mathbf{8.00_{\pm 1.60}}$	$52.10_{\pm 1.30}$
UniCLUN	$49.21_{\pm 9.01}$	$-24.60_{\pm 9.60}$	$-0.50_{\pm 0.60}$	$71.00_{\pm 4.80}$	$4.12_{\pm 2.66}$	$-39.80_{\pm 6.40}$	$-1.20_{\pm 0.00}$	$44.30_{\pm 4.40}$
MCL (ours)	$\mathbf{77.59_{\pm 1.28}}$	$-5.38_{\pm 0.99}$	$\mathbf{+0.70_{\pm 0.62}}$	$79.98_{\pm 0.86}$	$29.04_{\pm 1.75}$	$-21.53_{\pm 0.80}$	$-1.06_{\pm 0.22}$	$\mathbf{57.65_{\pm 1.01}}$
SGD (lower bound)	$40.24_{\pm 3.01}$	$-36.32_{\pm 1.33}$	$0.88_{\pm 0.94}$	$66.71_{\pm 1.38}$	$7.59_{\pm 0.34}$	$-74.35_{\pm 0.79}$	$-1.00_{\pm 0.27}$	$38.17_{\pm 0.76}$
RTI								
Joint training (upper bound)	$93.38_{\pm 1.06}$	–	–	–	$81.96_{\pm 0.75}$	–	–	–
EWC	$86.12_{\pm 4.63}$	$-5.79_{\pm 5.23}$	$-2.37_{\pm 1.65}$	$86.15_{\pm 3.00}$	$18.92_{\pm 0.71}$	$-93.00_{\pm 1.54}$	$-8.84_{\pm 8.09}$	$12.99_{\pm 2.65}$
SI	$85.64_{\pm 2.74}$	$-3.82_{\pm 1.08}$	$-2.60_{\pm 1.70}$	$84.02_{\pm 4.16}$	$18.87_{\pm 0.98}$	$-91.96_{\pm 1.07}$	$-7.19_{\pm 10.90}$	$14.77_{\pm 1.79}$
LwF	$91.35_{\pm 1.86}$	$-1.63_{\pm 1.01}$	$-3.70_{\pm 6.04}$	$91.16_{\pm 1.38}$	$19.02_{\pm 0.87}$	$-92.89_{\pm 1.65}$	$-7.90_{\pm 9.62}$	$13.17_{\pm 2.82}$
WSN	$93.49_{\pm 1.19}$	$0.00_{\pm 0.00}$	$-0.01_{\pm 1.81}$	$\mathbf{92.47_{\pm 1.23}}$	–	–	–	–
PEC	–	–	–	–	$45.94_{\pm 0.64}$	$\mathbf{-19.72_{\pm 1.58}}$	$0.00_{\pm 0.00}$	$\mathbf{65.83_{\pm 0.56}}$
SpaceNet	$83.94_{\pm 2.29}$	$\mathbf{+0.18_{\pm 0.15}}$	$-0.26_{\pm 5.31}$	$79.80_{\pm 2.77}$	$17.36_{\pm 1.47}$	$-83.27_{\pm 2.18}$	$-8.40_{\pm 1.08}$	$27.80_{\pm 2.90}$
NISPA	$91.39_{\pm 2.03}$	$0.00_{\pm 0.00}$	$\mathbf{+6.06_{\pm 4.67}}$	$88.22_{\pm 3.55}$	$18.39_{\pm 1.05}$	$-90.51_{\pm 1.03}$	$-11.49_{\pm 0.59}$	$17.15_{\pm 1.67}$
SNV	$\mathbf{94.60_{\pm 0.65}}$	$0.00_{\pm 0.00}$	$+0.00_{\pm 0.41}$	$90.10_{\pm 1.49}$	$\mathbf{53.18_{\pm 3.27}}$	$-24.20_{\pm 1.34}$	$\mathbf{+9.00_{\pm 1.74}}$	$46.30_{\pm 2.41}$
UniCLUN	$85.73_{\pm 1.63}$	$-9.95_{\pm 2.46}$	$+1.33_{\pm 1.64}$	$87.98_{\pm 0.96}$	$40.74_{\pm 2.52}$	$-63.24_{\pm 6.00}$	$-8.37_{\pm 6.82}$	$51.96_{\pm 5.55}$
MCL (ours)	$90.41_{\pm 0.98}$	$-3.45_{\pm 1.41}$	$-2.77_{\pm 3.78}$	$90.67_{\pm 0.82}$	$42.91_{\pm 0.89}$	$-36.13_{\pm 0.94}$	$-8.11_{\pm 8.93}$	$64.72_{\pm 0.72}$
SGD (lower bound)	$76.14_{\pm 4.63}$	$-19.78_{\pm 4.29}$	$-2.07_{\pm 0.54}$	$81.84_{\pm 3.40}$	$18.98_{\pm 0.82}$	$-93.44_{\pm 1.16}$	$-7.57_{\pm 9.73}$	$12.24_{\pm 2.03}$

4.1 Discussion

Overall Performance. As shown in Table 1, MCL ranks first or second in 13 of the 24 columns, whereas SNV ranks first or second in 17. SNV freezes selected neurons for each task using Shapley Neuron Valuation. It achieves high ACC in the Task-IL scenario. This shows that subnetwork-based methods can achieve high ACC while delivering 0 BWT. However, because they freeze subnetworks, they are not scalable. MCL tolerates some forgetting in exchange for plasticity: on Seq-CIFAR-100 and Seq-Tiny-ImageNet, its Task-IL FWT is positive, and its Task-IL accuracy is the highest, and it has the highest PS in two of the six settings. Class-IL separates the methods more sharply. The regularization baselines (EWC, SI, LwF) remain below 20% accuracy and incur more than 70 points of forgetting on every benchmark. MCL reaches $29.04 - 42.91\%$ accuracy with BWT between -21.53 and -41.19. SNV leads Class-IL accuracy on all three benchmarks, by 2.96 points on Seq-CIFAR-100, 1.66 on Seq-Tiny-ImageNet, and 7.79 on RTI, although MCL scores a higher PS on each.

Accuracy Evaluation Across Tasks. Figure 3.a,c, track accuracy and training cost as tasks accumulate. Regularization and sparsity methods (EWC, SI, LwF, NISPA, SpaceNet) fall onto the SGD curve from the second task onward and end at only 6.8–10.1%. While, MCL is ahead of SNV after tasks 2 and 3 (70.8 vs. 68.7%, 60.6 vs. 58.7%), but SNV maintains the highest continual accuracy from task 5 onward and finishes at 36.2%. MCL finishes second at 33.2%, narrowly ahead of PEC (32.8%). The gap to Joint (65.4% at task 10) shows how much Class-IL forgetting remains and indispensable need for more capable continual learning methods. In Task-IL (Figure 3.c), SNV stays above the Joint upper bound at every step and finishes at 86.98%. MCL matches or exceeds SNV after tasks 2, 3, 4 and 6 and stays above Joint for five of the first six tasks, but declines over the last four tasks.

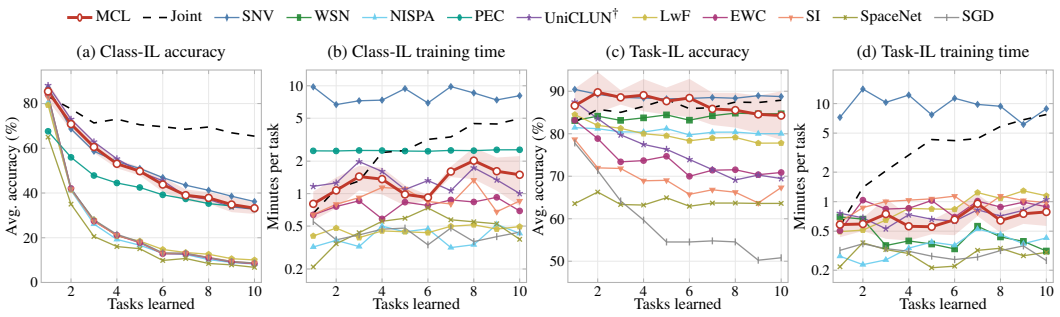

Figure 3: Accuracy and training cost as tasks accumulate for CIFAR-100. (a) Class-IL and (c) Task-IL average accuracy over the tasks seen so far. (b) Class-IL and (d) Task-IL training minutes per task on a log scale. Shading is MCL ± 1 SD; means over three seeds.

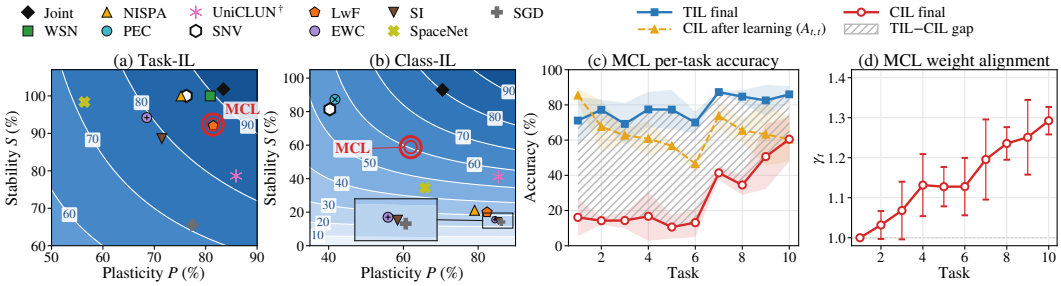

Figure 4: (a) Task-IL and (b) Class-IL stability–plasticity plane for CIFAR-100. (c) MCL accuracy on each task after all ten tasks; hatching marks the Task-IL−Class-IL gap. (d) Weight-alignment factor applied to each new Class-IL head. Ribbons and bars: ± 1 SD over three seeds.

Training cost. Figure 3.b,d, shows training times for all methods. SNV is the most expensive method in both settings, at 6.7–9.8 minutes per task in Class-IL and 6.1–14.1 in Task-IL. MCL's cost grows far more slowly than Joint's and stays about six times below SNV's in Class-IL and fourteen times below in Task-IL. MCL's Class-IL runs take roughly twice as long as its Task-IL runs; about a quarter of that difference comes from training for longer before early stopping (375 versus 297 epochs per run on average), and the rest from a higher per-epoch cost.

Stability and plasticity. Figure 4.a,b, place each method by how well it learns a new task (P) and how much of each learned task it retains (S). In Task-IL, MCL reaches PS $= 86.5\%$, essentially the same operating point as LwF 86.4%, and ahead of EWC and SI. Parameter-isolation methods reach $S = 100\%$ by construction; WSN converts this into the highest continual PS (89.4%), while UniCLUN has the highest plasticity (85.9%) but forgets more than any method except SGD ($S = 78.6\%$). MCL is the only continual method with both P and S above 50% (61.9%, 58.8%), and it has the highest continual PS in Class-IL.

Where class-incremental accuracy is lost. Figure 4.c, resolves MCL's final accuracy by task. When the task identity is given, MCL retains 69–87% on every task after all ten tasks, so the old tasks remain solvable. Without it, the six oldest tasks end at only 11–17%, whereas the four most recent keep 35–61%. Every task was learned reasonably well when it arrived ($A_{t,t} = 47$–85%), so the Class-IL deficit is not a failure to learn: it is concentrated in older tasks and appears when their classes must compete with those of later tasks. Figure 4.d, reports the nested weight-alignment factor $\gamma_t = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \overline{\|w^{\text{old}}_{:m}\|} / \overline{\|w^{\text{new}}_{:m}\|}$ by which each new head is rescaled. It rises from 1.00 to 1.29 over the sequence: new heads are increasingly weaker than the accumulated old heads, and the correction MCL applies grows accordingly.

Plate measures. Figure 5 evaluates ten continual learners on a single protocol: 100 Flowers-102 (Nilsback & Zisserman, 2008) species presented as ten disjoint tasks to an ImageNet-pretrained ResNet-18, scored on the official held-out split after the final task. The intent is not another accuracy table but a decomposition of *how* each method fails. Panels (a, b, d, e) separate a method's accuracy from the confidence it attaches to it and from how that accuracy is distributed across species; panels (c, f) ask which part of the representation the final decision rests on. Read together, they show

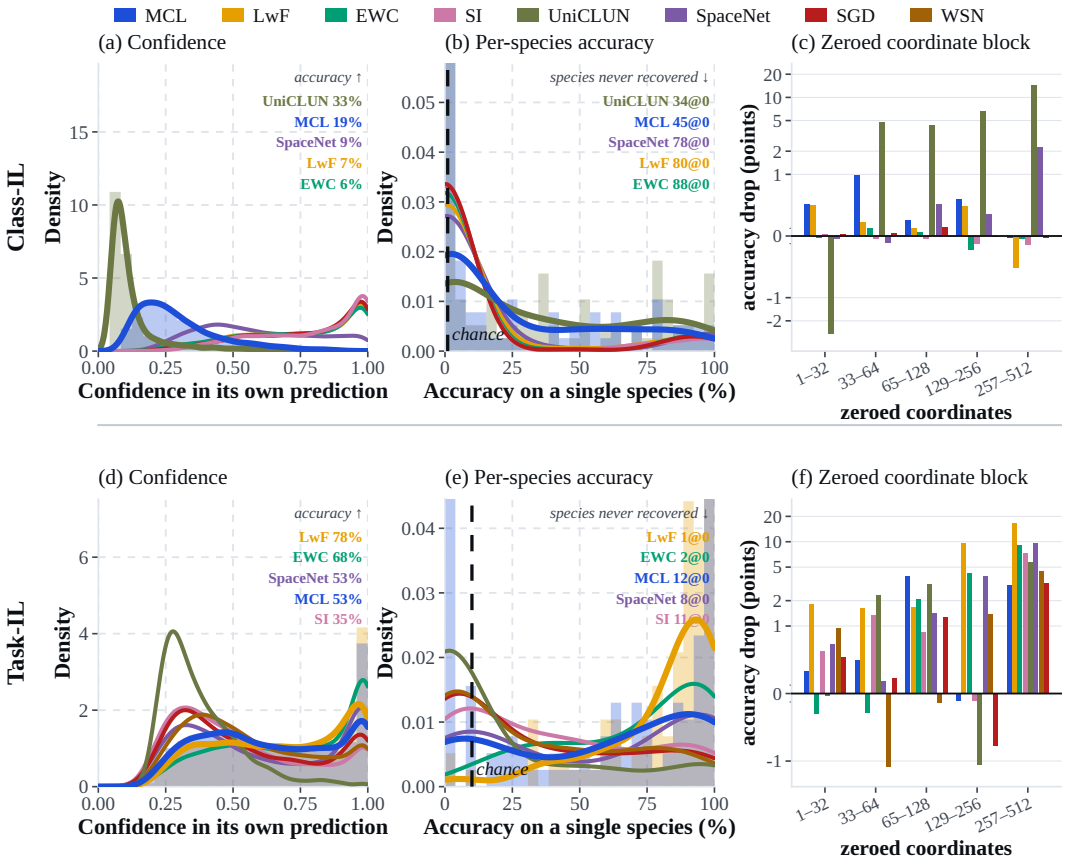

Figure 5: Eight continual learners after ten tasks of Flowers-10. **(a, d)** Confidence each method places in its own prediction, and how often that prediction is correct. **(b, e)** The same accuracy resolved per species, with the count of species never recovered; the density axis is scaled to the curves, so the histogram spike at zero is cropped. **(c, f)** Accuracy lost when one disjoint block of feature coordinates is zeroed, on a symmetric-log axis.

that the buffer-free regularization baselines and the nested models fail in qualitatively different ways, and that the difference is invisible in a mean. Under Class-IL (a), SGD, SI, EWC and LwF place the bulk of their probability mass above 0.9 while answering correctly only 5–7% of the time, a gap between mean confidence and accuracy of $+0.69$ to $+0.74$. MCL sits at $+0.10$ and the UniCLUN at -0.21. Panel (b) shows the same collapse from the other side: SI and SGD never again classify 90 of the 100 species correctly, EWC 88 and LwF 80, against 45 for MCL and 34 for UniCLUN. The failure is concrete. SI returns giant white arum lily at $p = 1.00$ for a frangipani, a bird of paradise, and a camellia alike. A deployed system cannot abstain on such predictions, which is the practical cost of forgetting that an accuracy column hides. This calibration gap, rather than the accuracy ranking, is where the nested models clearly outperform the regularization baselines.

Training dynamics. Figure 6 shows that early stopping, not the epoch cap, ends every task: no task reached 200 epochs, the restored best epoch lies between 8 and 33, and each task trains for 23–48 epochs.

Qualitative Results. Fig. 7 shows how visually similar classes stay mixed at every width.

5 Conclusion

In this work, we propose Matryoshka Continual Learning (MCL) a nesting framework that uses different embedding widths to mitigate the catastrophic forgetting problem in continual learning. The main idea is that nested embeddings can be guided by a privileged teacher with access to prior

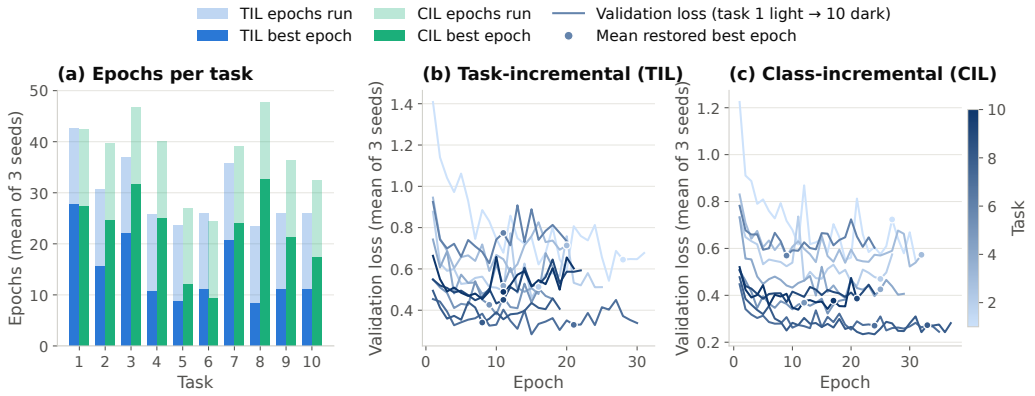

Figure 6: MCL training dynamics for CIFAR-100. (a) Epochs trained (light) and restored best epoch (dark) per task. (b, c) Validation loss per task (task 1 light to task 10 dark), mean of three seeds up to the shortest run; dots: mean best epoch. Early stopping with patience 15 and a 200-epoch cap.



Figure 7: Qualitative figures for nested prefixes in Class-IL. For each query (left; one held-out image drawn at random from the first class of each task), the five most similar held-out images under the final MCL model, using only the first $m \in \{32, 128, 512\}$ embedding dimensions.

knowledge, which then guides students to follow it without having access to that prior knowledge.

Limitation. As with existing baselines, MCL still exhibits a substantial gap to the joint-training upper bound in the class-incremental (Class-IL) setting. This gap suggests that Class-IL remains a challenging regime for current continual learning methods and warrants further investigation.

Future work. MCL construction assumes an encoder whose coordinates can be ordered and truncated, which holds for the vision backbones studied here but needs rethinking for language models, where the natural units are tokens rather than embedding dimensions. We leave this limitation for future work, where nesting needs to be well defined for large language models.

AI USE STATEMENT

In this work, we used generative AI tools to polish the text, fix typos, improve wording, and check grammar. The method, the datasets, and the references were developed and compiled by the authors without generative assistance. LLM-generated code was verified and tested for correctness and it is runnable in uploaded supplementary material. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Anonymous GitHub: https://anonymous.4open.science/r/Matryoshka-Continual-Learning-21F9

We provided a complete version of the code to reproduce every result of the paper at the anonymous GitHub link and uploaded it to the **supplementary materials in Open Review** on the submission page. The code and the RTI dataset will be published upon acceptance of the paper.

REFERENCES

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *NeurIPS*, volume 33, pp. 15920–15930, 2020.

Sungmin Cha and Kyunghyun Cho. Hyperparameters in continual learning: A reality check. *Transactions on Machine Learning Research*, 2025.

Romit Chatterjee, Vikram Chundawat, Ayush Tarun, Ankur Mali, and Murari Mandal. A unified framework for continual learning and machine unlearning, 2024. URL https://arxiv.org/abs/2408.11374.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *CVPR*, 2022.

Mustafa Burak Gurbuz and Constantine Dovrolis. NISPA: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks. In *ICML*, volume 162, pp. 8157–8174, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, pp. 831–839, 2019. doi: 10.1109/CVPR.2019.00092.

Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D. Yoo. Forget-free continual learning with winning subnetworks. In *ICML*, volume 162, pp. 10734–10750, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

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*, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *NeurIPS*, volume 35, 2022.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Cs231n course project report, Stanford University, 2015. URL https://vision.stanford.edu/teaching/cs231n/reports/2015/pdfs/yle_project.pdf.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, volume 30, 2017.

Mark D. McDonnell, Dong Gong, Amin Parveneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *NeurIPS*, volume 36, pp. 12022–12053, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *CVPR*, pp. 2001–2010, 2017.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks, 2022. URL https://arxiv.org/abs/1606.04671.

Joan Serrà, Dídac Surís, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, volume 80, pp. 4548–4557, 2018.

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 *CVPR*, pp. 11909–11919, 2023.

Ghada Sokar, Decebal Constantin Mocanu, and Mykola Pechenizkiy. SpaceNet: Make free space for continual learning. *Neurocomputing*, 439:1–11, 2021. doi: 10.1016/j.neucom.2021.01.078.

Mohammad Ali Vahedifar and Qi Zhang. Signal prediction for loss mitigation in tactile internet: A leader-follower game-theoretic approach. In *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2025. doi: 10.1109/MLSP62443.2025.11204284. URL https://doi.org/10.1109/MLSP62443.2025.11204284.

Mohammad Ali Vahedifar and Qi Zhang. Shapley features for robust signal prediction in tactile internet. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026a.

Mohammad Ali Vahedifar and Qi Zhang. No forgetting learning: Buffer-free continual learning classification, 2026b. URL https://arxiv.org/abs/2503.04638.

Mohammad Ali Vahedifar, Abhisek Ray, and Qi Zhang. Shapley neuron values for continual learning: Which neurons matter most?, 2026a. URL https://arxiv.org/abs/2605.15877.

Mohammad Ali Vahedifar, Qi Zhang, and Alexandros Iosifidis. Towards lifelong deep learning: A review of continual learning and unlearning methods, 2026b. URL https://doi.org/10.5281/zenodo.18774099. Preprint, version 2.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *TPAMI*, 46(8):5362–5383, 2024.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pp. 631–648, 2022a.

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 *CVPR*, pp. 139–149, 2022b.

Michal Zajac, Tinne Tuytelaars, and Gido M van de Ven. Prediction error-based classification for class-incremental learning. In *ICLR*, 2024.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In *ICCV*, 2023.

A APPENDIX

The appendix primarily contains details of Class-IL Components B, the RTI dataset details and the Physical Artificial Intelligence (PAI) scenario C, Continual Learning metrics D, and additional experiments E that cannot be included in the main body due to page limits.

B CLASS-IL COMPONENTS

In Class-IL, all observed classes compete under a single prediction rule (Eq. 3), which introduces biases that Task-IL does not expose. MCL addresses them with two training-time choices and two end-of-task operations; none of them changes the Task-IL model beyond the weight rescaling noted below.

Local cross-entropy and cosine head. A batch from P_t contains no class in $\mathcal{O}_t$, so scoring old logits on it can only lower the scores. We therefore set $\mathcal{A}_t = \mathcal{C}_t$ and leave the old classes to the distillation term. With a linear head, old classes are further disadvantaged by weight-norm imbalance: old rows stop growing while newer rows continue to grow (Hou et al., 2019). The cosine head of Eq. 4 removes row norms from the decision. Under Task-IL, $\mathcal{A}_t = \mathcal{C}_t$ holds by definition of the prediction rule, and the linear head is kept.

Nested weight aligning. At the end of each task $t > 1$, we rescale the new-class rows to the mean norm of the old rows. Because each doll reads only a prefix of a row, the ratio is computed per doll and averaged:

$$\gamma_t = \frac{1}{K} \sum_{m \in \mathcal{M}} \frac{\text{mean}_{c \in \mathcal{O}_t} \|W_{c,1:m}\|}{\text{mean}_{c \in \mathcal{C}_t} \|W_{c,1:m}\|}, \qquad W_{\mathcal{C}_t,:} \leftarrow \gamma_t \, W_{\mathcal{C}_t,:}. \tag{9}$$

For $K = 1$ this reduces to standard weight aligning. Under the cosine head the rescaling leaves Class-IL logits invariant; it affects only subsequent optimization, since the gradient with respect to a normalized row scales with the inverse of its norm. Under Task-IL, the linear head's weights, but not its bias, are rescaled.

Exemplar-free class-mean readout. At the end of task t, in evaluation mode, we store for each $c \in \mathcal{C}_t$ the class mean of a power-normalized full-width feature,

$$\mu_c = \text{mean}_{(x,y) \in \mathcal{D}_t,\, y=c} \; \phi(x), \qquad \phi(x) = \frac{\text{sign}(f_\theta(x)) \odot \sqrt{|f_\theta(x)|}}{\|\text{sign}(f_\theta(x)) \odot \sqrt{|f_\theta(x)|}\|}. \tag{10}$$

Means of earlier classes are not up to date. At test time, the Class-IL score of class c combines the widest-doll logit with a nearest-class-mean score,

$$g_c(x) = (1-\alpha)\,\text{std}\big(z^{(d)}(x)\big)_c + \alpha\,\text{std}\big(-\|\phi(x) - \mu_\cdot\|^2\big)_c, \tag{11}$$

where $\alpha \in [0,1]$ and $\text{std}(\cdot)$ standardizes a score vector over $\mathcal{O}_{t+1}$ to zero mean and unit variance per input, placing both terms on a common scale. The readout adds no parameters or gradients and modifies only the Class-IL prediction rule.

Prediction rule. Under Class-IL the model predicts with the fused score of Eq. equation 11, not with the head logit alone:

$$\hat{y}(x) = \underset{c \,\in\, \mathcal{O}_{t+1}}{\arg\max}\, g_c(x), \qquad g_c(x) = (1-\alpha)\,\text{std}\big(z^{(d)}(x)\big)_c + \alpha\,\text{std}\big(-\|\phi(x) - \mu_\cdot\|^2\big)_c. \tag{12}$$

Setting $\alpha = 0$ recovers the widest-doll logit and $\alpha = 1$ the nearest-class-mean score; both are limiting cases of the same rule rather than separate methods. Under Task-IL the prediction rule is unchanged and α is inactive, so Eq. equation 12 is the sole place the readout enters any reported number.

B.1 CONSTANTS AND HOW THEY WERE SET

MCL has three constants beyond the optimizer: the fusion weight α of Eq. equation 12, the cosine scale s of Eq. equation 4, and the width exponent β of Eq. equation 8. They were fixed in three different ways, and we state each explicitly rather than reporting a single tuned configuration.

Symbol	Role	Value	How fixed
α	fusion weight, Eq. equation 12	0.5	searched, $\{0.25, 0.5, 1.0\}$
s	cosine scale, Eq. equation 4	16	fixed a priori, not searched
β	width exponent, Eq. equation 8	0	defines the arm; $\beta = 1$ is MCL-G

Table 2: The three constants of MCL. Only α is selected on data.

α is searched jointly with the learning rate, the distillation weight λ_S and the temperature τ over $\alpha \in \{0.25, 0.5, 1.0\}$, $\lambda_S \sim \log\mathrm{U}(0.1, 10)$ and $\tau \in \{1, 2, 4\}$, under the same held-out tuning protocol as every baseline. The selected configurations are $(\alpha, \lambda_S, \tau, \mathrm{lr}) = (0.5,\ 0.425,\ 4.0,\ 7.16\times10^{-4})$ for Class-IL and $(0.5,\ 0.486,\ 2.0,\ 9.78\times10^{-4})$ for Task-IL.

C Real-World Physical AI Scenario

Physical AI over the Tactile Internet. PAI integrates learning-enabled models with embodied systems that perceive, reason, and interact with the physical world. A representative example is robotic teleoperation, where a human operator remotely controls a robot while receiving haptic feedback. The Tactile Internet supports such bidirectional interaction by transmitting control commands and touch sensations over long distances, enabling mission-critical applications such as remote surgery and robotic manipulation (Vahedifar & Zhang, 2025). Because haptic devices typically operate at sampling rates of 1 kHz or higher, the real-time control and inference path must satisfy an extremely low latency budget, commonly targeted at approximately 1 ms (Vahedifar & Zhang, 2026a). A deployed PAI system is not trained only once. It continually receives new interaction data as operators, environments, and procedures change.

Robotic Tactile Internet dataset. To study this setting, we collected a Robotic Tactile Internet (RTI) dataset containing haptic tele-manipulation signals from a virtual client–server environment implemented in CHAI3D. The experimental platform follows the IEEE 1918.1.1 architecture and uses a Novint Falcon haptic device operating at a 1 kHz control loop. Participants controlled a virtual tool while performing five actions: no movement, pressing, tapping, free-air movement, and dragging. Each action was performed for 10 seconds, and every participant repeated the complete five-action sequence four times. This produced 20 action segments per participant. The dataset contains recordings from 20 anonymized participants aged 20 to 33. We used this dataset for teleoperation classification in a continual learning setup.

The RTI dataset and its accompanying documentation will be made publicly available following completion of the required legal, ethical, and institutional review and confirmation that its release complies with the applicable participant-consent and data-protection requirements.

D Continual-Learning Metrics

Let K denote the number of tasks. After completing every task T_i, we evaluate the model on every task T_j, including tasks not yet encountered. This produces the accuracy matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$, where

$$A_{i,j} = \text{accuracy on task } T_j \text{ after completing task } T_i. \tag{13}$$

Accuracies are represented as proportions in the following definitions.

Average Accuracy (ACC). ACC measures final performance across all tasks:

$$\mathrm{ACC} = \frac{1}{K}\sum_{j=1}^{K} A_{K,j}. \tag{14}$$

Higher ACC indicates better overall performance at the end of the task sequence.

Backward Transfer (BWT). BWT measures how subsequent learning changes performance on previously learned tasks:

$$\mathrm{BWT} = \frac{1}{K-1}\sum_{j=1}^{K-1} \left(A_{K,j} - A_{j,j}\right). \tag{15}$$

Negative BWT indicates forgetting, zero indicates no net change, and positive BWT indicates that later tasks improve earlier performance.

Forward Transfer (FWT). FWT measures whether learning previous tasks improves performance on a new task before that task is trained:

$$\mathrm{FWT} = \frac{1}{K-1} \sum_{j=2}^{K} \left(A_{j-1,j} - b_j \right), \tag{16}$$

where b_j is the accuracy of a randomly initialized model on task T_j.

Plasticity. Plasticity measures the normalized improvement produced by learning each new task:

$$P = \frac{1}{K-1} \sum_{j=2}^{K} \frac{A_{j,j} - A_{j-1,j}}{1 - A_{j-1,j} + \varepsilon}, \tag{17}$$

where $\varepsilon > 0$ prevents division by zero. Higher P indicates a greater ability to acquire new knowledge.

Stability. Stability measures how well performance on earlier tasks is preserved:

$$S = 1 - \frac{1}{K-1} \sum_{j=1}^{K-1} \left(A_{j,j} - A_{K,j} \right). \tag{18}$$

Under the definition of BWT above, $S = 1 + \mathrm{BWT}$. Higher values indicate less forgetting.

Plasticity–Stability Score (PS). We summarize the balance between plasticity and stability using their harmonic mean:

$$\mathrm{PS} = \frac{2\,P\,S}{P + S + \varepsilon}. \tag{19}$$

Because the symbol PS appears on both sides if written this way, we denote the reported score by PS and the two constituent quantities by P and S. A high score requires both successful acquisition and retention (Vahedifar & Zhang, 2026b).

E Additional Results of Continual Learning Job

Table 3: Training cost comparison.

Method	Task-IL				Class-IL			
	Train min/task↓	GPU h↓	Energy (Wh)↓	Peak device (MB)↓	Train min/task↓	GPU h↓	Energy (Wh)↓	Peak device (MB)↓
Seq-CIFAR-100								
Joint training (upper bound)	$40.33_{\pm 2.27}$	$0.678_{\pm 0.038}$	$84.53_{\pm 5.13}$	$2001.0_{\pm 2.1}$	$28.37_{\pm 0.68}$	$0.479_{\pm 0.011}$	$67.12_{\pm 3.43}$	$2094.0_{\pm 3.2}$
EWC	$8.73_{\pm 0.75}$	$0.151_{\pm 0.012}$	$17.97_{\pm 1.29}$	$2148.5_{\pm 55.7}$	$7.77_{\pm 1.26}$	$0.136_{\pm 0.021}$	$15.56_{\pm 2.00}$	$2276.4_{\pm 151.9}$
SI	$9.67_{\pm 0.45}$	$0.168_{\pm 0.008}$	$20.97_{\pm 1.03}$	$3297.1_{\pm 773.7}$	$9.11_{\pm 0.44}$	$0.158_{\pm 0.007}$	$20.74_{\pm 1.20}$	$3542.4_{\pm 134.5}$
LwF	$8.95_{\pm 0.68}$	$0.154_{\pm 0.011}$	$37.11_{\pm 2.49}$	$4793.7_{\pm 126.4}$	$4.58_{\pm 0.41}$	$0.082_{\pm 0.007}$	$12.02_{\pm 0.63}$	$2144.3_{\pm 52.1}$
WSN	$4.53_{\pm 0.23}$	$0.082_{\pm 0.004}$	$14.21_{\pm 0.77}$	$2391.1_{\pm 50.9}$	–	–	–	–
PEC	–	–	–	–	$25.07_{\pm 0.16}$	$0.426_{\pm 0.003}$	$96.16_{\pm 1.58}$	$4473.6_{\pm 0.0}$
SpaceNet	**$2.89_{\pm 0.03}$**	**$0.052_{\pm 0.001}$**	**$3.79_{\pm 0.07}$**	**$1334.8_{\pm 12.1}$**	$4.89_{\pm 0.55}$	$0.085_{\pm 0.009}$	**$6.19_{\pm 0.65}$**	**$1335.5_{\pm 11.6}$**
NISPA	$3.63_{\pm 0.74}$	$0.066_{\pm 0.012}$	$7.58_{\pm 1.43}$	$1991.9_{\pm 135.5}$	**$4.02_{\pm 0.01}$**	**$0.072_{\pm 0.000}$**	$8.75_{\pm 0.08}$	$1985.6_{\pm 49.8}$
SNV	$97.00_{\pm 23.73}$	$1.624_{\pm 0.395}$	$426.21_{\pm 112.91}$	$2428.9_{\pm 0.0}$	$81.31_{\pm 8.34}$	$1.400_{\pm 0.139}$	$319.94_{\pm 42.47}$	$2428.1_{\pm 1.2}$
UniCLUN	$7.42_{\pm 0.60}$	$0.128_{\pm 0.010}$	$25.77_{\pm 2.35}$	$2514.8_{\pm 0.0}$	$13.58_{\pm 1.79}$	$0.233_{\pm 0.030}$	$39.95_{\pm 5.95}$	$2514.8_{\pm 0.0}$
MCL (ours)	$6.83_{\pm 1.03}$	$0.120_{\pm 0.017}$	$14.97_{\pm 2.18}$	$2142.2_{\pm 55.7}$	$13.31_{\pm 1.49}$	$0.228_{\pm 0.025}$	$24.22_{\pm 3.10}$	$2110.8_{\pm 96.5}$
SGD (lower bound)	$3.06_{\pm 0.19}$	$0.057_{\pm 0.003}$	$8.86_{\pm 0.49}$	$1915.0_{\pm 0.0}$	$4.30_{\pm 0.08}$	$0.077_{\pm 0.001}$	$12.05_{\pm 0.18}$	$1916.4_{\pm 1.2}$
Seq-Tiny-ImageNet								
Joint training (upper bound)	$17.21_{\pm 1.26}$	$2.87_{\pm 0.21}$	$424.0_{\pm 16.4}$	$3405.4_{\pm 1.2}$	$15.32_{\pm 1.63}$	$2.55_{\pm 0.27}$	$400.2_{\pm 24.2}$	$3404.0_{\pm 0.0}$
EWC	**$1.09_{\pm 0.06}$**	**$0.182_{\pm 0.010}$**	$48.3_{\pm 2.6}$	$3179_{\pm 0}$	$0.99_{\pm 0.03}$	$0.164_{\pm 0.004}$	**$43.1_{\pm 1.1}$**	$3179_{\pm 0}$
SI	$1.54_{\pm 0.14}$	$0.257_{\pm 0.024}$	$66.7_{\pm 5.7}$	$3539_{\pm 0}$	$1.12_{\pm 0.09}$	$0.186_{\pm 0.015}$	$47.7_{\pm 3.6}$	$3539_{\pm 1}$
LwF	$5.02_{\pm 0.30}$	$0.836_{\pm 0.050}$	$222.3_{\pm 13.3}$	$13462_{\pm 68}$	**$0.95_{\pm 0.05}$**	**$0.158_{\pm 0.008}$**	$43.5_{\pm 2.1}$	$3377_{\pm 0}$
WSN	$4.97_{\pm 0.15}$	$0.83_{\pm 0.03}$	$89.0_{\pm 3.7}$	$3555.7_{\pm 1.2}$	–	–	–	–
PEC	–	–	–	–	$35.81_{\pm 7.45}$	$5.97_{\pm 1.24}$	$819.8_{\pm 48.3}$	$13992.5_{\pm 0.0}$
SpaceNet	$1.14_{\pm 0.12}$	$0.19_{\pm 0.02}$	**$13.1_{\pm 1.2}$**	**$1596.3_{\pm 0.0}$**	$18.29_{\pm 1.30}$	$3.05_{\pm 0.22}$	$202.4_{\pm 13.1}$	**$1596.3_{\pm 0.0}$**
NISPA	$1.41_{\pm 0.34}$	$0.23_{\pm 0.06}$	$34.9_{\pm 8.4}$	$4705.6_{\pm 553.8}$	$2.99_{\pm 0.18}$	$0.50_{\pm 0.03}$	$73.5_{\pm 3.9}$	$5659.8_{\pm 334.7}$
SNV	$46.96_{\pm 7.83}$	$7.83_{\pm 1.31}$	$2051.0_{\pm 369.3}$	$6263.1_{\pm 1.2}$	$45.05_{\pm 8.18}$	$7.51_{\pm 1.36}$	$1888.3_{\pm 406.0}$	$6265.9_{\pm 1.2}$
UniCLUN	$4.40_{\pm 0.32}$	$0.73_{\pm 0.05}$	$157.2_{\pm 11.5}$	$5759.1_{\pm 40.0}$	$4.62_{\pm 0.75}$	$0.77_{\pm 0.12}$	$171.4_{\pm 27.0}$	$5759.1_{\pm 40.0}$
MCL (ours)	$1.30_{\pm 0.06}$	$0.217_{\pm 0.010}$	$57.0_{\pm 2.5}$	$3377_{\pm 0}$	$3.01_{\pm 0.13}$	$0.502_{\pm 0.022}$	$112.5_{\pm 4.6}$	$3379_{\pm 0}$
SGD (lower bound)	$0.74_{\pm 0.03}$	$0.123_{\pm 0.005}$	$33.7_{\pm 1.2}$	$3169_{\pm 0}$	$0.81_{\pm 0.03}$	$0.135_{\pm 0.005}$	$36.4_{\pm 1.3}$	$3169_{\pm 0}$
RTI								
Joint training (upper bound)	$6.36_{\pm 0.36}$	$0.1073_{\pm 0.0060}$	$25.70_{\pm 1.19}$	$2347_{\pm 29}$	$5.22_{\pm 0.15}$	$0.0882_{\pm 0.0025}$	$23.08_{\pm 0.40}$	$2340_{\pm 25}$
EWC	$2.24_{\pm 0.07}$	$0.0387_{\pm 0.0011}$	$9.10_{\pm 0.25}$	$2226_{\pm 16}$	$2.18_{\pm 0.08}$	$0.0377_{\pm 0.0014}$	$8.93_{\pm 0.14}$	$2273_{\pm 46}$
SI	$2.63_{\pm 0.09}$	$0.0453_{\pm 0.0015}$	$12.30_{\pm 0.31}$	$2759_{\pm 19}$	$2.27_{\pm 0.13}$	$0.0394_{\pm 0.0021}$	$10.49_{\pm 0.53}$	$2745_{\pm 17}$
LwF	$4.01_{\pm 0.18}$	$0.0682_{\pm 0.0030}$	$20.08_{\pm 0.90}$	$3333_{\pm 20}$	$1.89_{\pm 0.14}$	$0.0328_{\pm 0.0024}$	$9.26_{\pm 0.67}$	$2165_{\pm 22}$
WSN	$3.07_{\pm 0.19}$	$0.0530_{\pm 0.0031}$	$13.34_{\pm 0.91}$	$2476_{\pm 1}$	–	–	–	–
PEC	–	–	–	–	$12.42_{\pm 0.01}$	$0.2098_{\pm 0.0001}$	$62.82_{\pm 0.05}$	$4643_{\pm 1}$
SpaceNet	$1.77_{\pm 0.21}$	**$0.0304_{\pm 0.0034}$**	**$2.56_{\pm 0.39}$**	**$1541_{\pm 12}$**	**$1.49_{\pm 0.08}$**	**$0.0257_{\pm 0.0014}$**	**$2.14_{\pm 0.11}$**	**1534**
NISPA	$2.17_{\pm 0.50}$	$0.0373_{\pm 0.0083}$	$5.40_{\pm 1.21}$	$2073_{\pm 40}$	$2.01_{\pm 0.17}$	$0.0346_{\pm 0.0029}$	$5.05_{\pm 0.45}$	$2136_{\pm 42}$
SNV	$19.56_{\pm 8.28}$	$0.5785_{\pm 0.1408}$	$186.12_{\pm 49.31}$	$2619_{\pm 1}$	$23.71_{\pm 2.92}$	$0.4886_{\pm 0.0501}$	$139.45_{\pm 18.52}$	$2621_{\pm 1}$
UniCLUN	$4.69_{\pm 0.38}$	$0.0795_{\pm 0.0063}$	$20.31_{\pm 1.49}$	2874	$6.61_{\pm 0.29}$	$0.1115_{\pm 0.0049}$	$29.81_{\pm 1.28}$	2874
MCL (ours)	**$1.75_{\pm 0.21}$**	$0.0305_{\pm 0.0035}$	$7.71_{\pm 0.81}$	$2190_{\pm 37}$	$2.44_{\pm 0.42}$	$0.0420_{\pm 0.0070}$	$9.50_{\pm 1.68}$	$2197_{\pm 20}$
SGD (lower bound)	$1.12_{\pm 0.13}$	$0.0198_{\pm 0.0021}$	$5.44_{\pm 0.57}$	2071	$1.28_{\pm 0.03}$	$0.0225_{\pm 0.0005}$	$6.17_{\pm 0.14}$	2071

Table 4: Memory and storage measurements for Task-IL.

Method	Peak alloc. (MB)↓	Peak reserved (MB)↓	Peak device (MB)↓	Optimizer (MB)↓	Model params (MB)↓	Resident storage (MB)↓	Checkpoint (MB)↓
Seq-CIFAR-100							
Joint training (upper bound)	$636.2_{\pm 2.4}$	$784.3_{\pm 2.1}$	$2001.0_{\pm 2.1}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$44.9_{\pm 0.0}$
EWC	$801.7_{\pm 52.5}$	$931.8_{\pm 55.7}$	$2148.5_{\pm 55.7}$	$89.6_{\pm 0.0}$	$44.8_{\pm 0.0}$	$134.4_{\pm 0.0}$	$134.5_{\pm 0.0}$
SI	$1984.8_{\pm 766.0}$	$2080.4_{\pm 773.7}$	$3297.1_{\pm 773.7}$	$89.6_{\pm 0.0}$	$44.8_{\pm 0.0}$	$313.4_{\pm 0.0}$	$313.6_{\pm 0.0}$
LwF	$3404.3_{\pm 122.6}$	$3577.0_{\pm 126.4}$	$4793.7_{\pm 126.4}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.8_{\pm 0.0}$
WSN	$945.6_{\pm 1.8}$	$1174.4_{\pm 50.9}$	$2391.1_{\pm 50.9}$	$178.6_{\pm 0.0}$	$89.4_{\pm 0.0}$	$213.0_{\pm 0.0}$	$213.4_{\pm 0.0}$
PEC	–	–	–	–	–	–	–
SpaceNet	**$105.2_{\pm 0.8}$**	**$126.5_{\pm 12.1}$**	**$1334.8_{\pm 12.1}$**	**$11.3_{\pm 0.0}$**	**$5.6_{\pm 0.0}$**	**$26.8_{\pm 0.0}$**	**$14.1_{\pm 0.0}$**
NISPA	$503.4_{\pm 75.4}$	$775.2_{\pm 135.5}$	$1991.9_{\pm 135.5}$	$70.0_{\pm 0.0}$	$35.0_{\pm 0.0}$	$78.8_{\pm 0.0}$	$78.8_{\pm 0.0}$
SNV	$770.6_{\pm 4.5}$	$1210.1_{\pm 0.0}$	$2428.9_{\pm 0.0}$	$89.4_{\pm 0.0}$	$44.8_{\pm 0.0}$	$49.9_{\pm 0.0}$	$46.8_{\pm 0.0}$
UniCLUN	$1074.3_{\pm 2.4}$	$1296.0_{\pm 0.0}$	$2514.8_{\pm 0.0}$	$92.2_{\pm 0.0}$	$92.2_{\pm 0.0}$	$104.6_{\pm 0.0}$	$104.7_{\pm 0.0}$
MCL (ours)	$780.7_{\pm 46.1}$	$925.5_{\pm 55.7}$	$2142.2_{\pm 55.7}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.7_{\pm 0.0}$	$89.8_{\pm 0.0}$
SGD (lower bound)	$543.5_{\pm 0.0}$	$698.4_{\pm 0.0}$	$1915.0_{\pm 0.0}$	$89.4_{\pm 0.0}$	$44.8_{\pm 0.0}$	$44.8_{\pm 0.0}$	$44.9_{\pm 0.0}$
Seq-Tiny-ImageNet							
Joint training (upper bound)	$1548.2_{\pm 23.7}$	$2188.7_{\pm 1.2}$	$3405.4_{\pm 1.2}$	$89.8_{\pm 0.0}$	$89.8_{\pm 0.0}$	$89.8_{\pm 0.0}$	$45.0_{\pm 0.0}$
EWC	$1564.8_{\pm 1.2}$	$1814.0_{\pm 0.0}$	$3178.8_{\pm 0.0}$	$85.3_{\pm 0.0}$	$42.8_{\pm 0.0}$	$128.0_{\pm 0.0}$	$128.2_{\pm 0.0}$
SI	$1910.7_{\pm 1.0}$	$2174.0_{\pm 0.0}$	$3538.8_{\pm 0.0}$	$85.3_{\pm 0.0}$	$42.8_{\pm 0.0}$	$298.5_{\pm 0.0}$	$298.7_{\pm 0.0}$
LwF	$11773.7_{\pm 0.9}$	$12097.3_{\pm 68.1}$	$13462.1_{\pm 68.1}$	$85.6_{\pm 0.0}$	$85.6_{\pm 0.0}$	$85.7_{\pm 0.0}$	$85.8_{\pm 0.0}$
WSN	$1870.6_{\pm 3.6}$	$2339.0_{\pm 1.2}$	$3555.7_{\pm 1.2}$	$178.7_{\pm 0.0}$	$89.5_{\pm 0.0}$	$213.1_{\pm 0.0}$	$213.5_{\pm 0.0}$
PEC	–	–	–	–	–	–	–
SpaceNet	**$340.8_{\pm 0.0}$**	**$385.9_{\pm 0.0}$**	**$1596.3_{\pm 0.0}$**	**$40.9_{\pm 0.0}$**	**$20.5_{\pm 0.0}$**	$97.3_{\pm 0.0}$	$51.2_{\pm 0.0}$
NISPA	$1905.1_{\pm 80.4}$	$3489.0_{\pm 553.8}$	$4705.6_{\pm 553.8}$	$272.2_{\pm 0.0}$	$136.1_{\pm 0.0}$	$306.2_{\pm 0.0}$	$306.2_{\pm 0.0}$
SNV	$3232.3_{\pm 3.4}$	$5044.3_{\pm 1.2}$	$6263.1_{\pm 1.2}$	$89.4_{\pm 0.0}$	$44.9_{\pm 0.0}$	**$73.4_{\pm 0.0}$**	**$46.9_{\pm 0.0}$**
UniCLUN	$3388.5_{\pm 1.2}$	$4540.3_{\pm 40.0}$	$5759.1_{\pm 40.0}$	$92.4_{\pm 0.0}$	$92.4_{\pm 0.0}$	$141.6_{\pm 0.0}$	$141.8_{\pm 0.0}$
MCL (ours)	$1816.8_{\pm 0.4}$	$2012.0_{\pm 0.0}$	$3376.8_{\pm 0.0}$	$85.6_{\pm 0.0}$	$85.6_{\pm 0.0}$	$85.8_{\pm 0.0}$	$85.9_{\pm 0.0}$
SGD (lower bound)	$1430.9_{\pm 0.3}$	$1804.0_{\pm 0.0}$	$3168.8_{\pm 0.0}$	$85.3_{\pm 0.0}$	$42.8_{\pm 0.0}$	$42.8_{\pm 0.0}$	$42.9_{\pm 0.0}$
RTI							
Joint training (upper bound)	$643_{\pm 25}$	$916_{\pm 29}$	$2347_{\pm 29}$	89.4	89.4	89.5	$44.8_{\pm 0.0}$
EWC	$696_{\pm 3}$	$795_{\pm 16}$	$2226_{\pm 16}$	89.4	44.7	134.1	$134.2_{\pm 0.0}$
SI	$1246_{\pm 7}$	$1328_{\pm 19}$	$2759_{\pm 19}$	89.4	44.7	312.9	$313.1_{\pm 0.0}$
LwF	$1849_{\pm 29}$	$1902_{\pm 20}$	$3333_{\pm 20}$	89.4	89.4	89.5	$89.6_{\pm 0.0}$
WSN	$912_{\pm 0}$	$1045_{\pm 1}$	$2476_{\pm 1}$	178.6	89.3	156.7	$157.0_{\pm 0.0}$
PEC	–	–	–	–	–	–	–
SpaceNet	**$105_{\pm 0}$**	**$127_{\pm 12}$**	**$1541_{\pm 12}$**	**11.1**	**5.6**	**26.5**	**$14.0_{\pm 0.0}$**
NISPA	$395_{\pm 5}$	$642_{\pm 40}$	$2073_{\pm 40}$	69.4	34.7	78.0	$78.1_{\pm 0.0}$
SNV	$785_{\pm 5}$	$1188_{\pm 1}$	$2619_{\pm 1}$	89.4	44.7	49.8	$46.8_{\pm 0.0}$
UniCLUN	$1209_{\pm 1}$	1441	2874	92.0	92.0	104.4	$104.5_{\pm 0.0}$
MCL (ours)	$672_{\pm 27}$	$759_{\pm 37}$	$2190_{\pm 37}$	89.4	89.4	89.5	$89.6_{\pm 0.0}$
SGD (lower bound)	566	640	2071	89.4	44.7	44.7	$44.8_{\pm 0.0}$

Table 5: Memory and storage measurements for Class-IL.

Method	Peak alloc. (MB)↓	Peak reserved (MB)↓	Peak device (MB)↓	Optimizer (MB)↓	Model params (MB)↓	Resident storage (MB)↓	Checkpoint (MB)↓
Seq-CIFAR-100							
Joint training (upper bound)	$694.9_{\pm 27.7}$	$877.3_{\pm 3.2}$	$2094.0_{\pm 3.2}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$44.9_{\pm 0.0}$
EWC	$897.3_{\pm 191.4}$	$1059.8_{\pm 151.9}$	$2276.4_{\pm 151.9}$	$89.6_{\pm 0.0}$	$44.8_{\pm 0.0}$	$134.4_{\pm 0.0}$	$134.5_{\pm 0.0}$
SI	$2219.6_{\pm 160.6}$	$2325.7_{\pm 134.5}$	$3542.4_{\pm 134.5}$	$89.6_{\pm 0.0}$	$44.8_{\pm 0.0}$	$313.5_{\pm 0.0}$	$313.7_{\pm 0.0}$
LwF	$810.1_{\pm 23.2}$	$927.6_{\pm 52.1}$	$2144.3_{\pm 52.1}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.8_{\pm 0.0}$
WSN	–	–	–	–	–	–	–
PEC	$2215.7_{\pm 0.0}$	$3259.0_{\pm 0.0}$	$4473.6_{\pm 0.0}$	**$0.9_{\pm 0.0}$**	$66.8_{\pm 0.0}$	$66.8_{\pm 0.0}$	$67.0_{\pm 0.0}$
SpaceNet	**$105.9_{\pm 0.1}$**	**$127.2_{\pm 11.6}$**	**$1335.5_{\pm 11.6}$**	$11.3_{\pm 0.0}$	**$5.6_{\pm 0.0}$**	**$26.8_{\pm 0.0}$**	**$14.2_{\pm 0.0}$**
NISPA	$543.8_{\pm 17.0}$	$769.0_{\pm 49.8}$	$1985.6_{\pm 49.8}$	$70.0_{\pm 0.0}$	$35.0_{\pm 0.0}$	$78.8_{\pm 0.0}$	$78.8_{\pm 0.0}$
SNV	$811.3_{\pm 2.4}$	$1209.4_{\pm 1.2}$	$2428.1_{\pm 1.2}$	$89.4_{\pm 0.0}$	$44.9_{\pm 0.0}$	$50.0_{\pm 0.0}$	$46.9_{\pm 0.0}$
UniCLUN	$1074.3_{\pm 2.4}$	$1296.0_{\pm 0.0}$	$2514.8_{\pm 0.0}$	$92.2_{\pm 0.0}$	$92.2_{\pm 0.0}$	$104.6_{\pm 0.0}$	$104.7_{\pm 0.0}$
MCL (ours)	$765.8_{\pm 72.7}$	$894.1_{\pm 96.5}$	$2110.8_{\pm 96.5}$	$89.6_{\pm 0.0}$	$89.6_{\pm 0.0}$	$89.9_{\pm 0.0}$	$89.8_{\pm 0.0}$
SGD (lower bound)	$543.8_{\pm 0.0}$	$699.7_{\pm 1.2}$	$1916.4_{\pm 1.2}$	$89.6_{\pm 0.0}$	$44.8_{\pm 0.0}$	$44.8_{\pm 0.0}$	$44.9_{\pm 0.0}$
Seq-Tiny-ImageNet							
Joint training (upper bound)	$1562.6_{\pm 2.4}$	$2187.3_{\pm 0.0}$	$3404.0_{\pm 0.0}$	$89.8_{\pm 0.0}$	$89.8_{\pm 0.0}$	$89.8_{\pm 0.0}$	$45.0_{\pm 0.0}$
EWC	$1566.1_{\pm 1.8}$	$1814.0_{\pm 0.0}$	$3178.8_{\pm 0.0}$	$85.6_{\pm 0.0}$	$42.8_{\pm 0.0}$	$128.0_{\pm 0.0}$	$128.2_{\pm 0.0}$
SI	$1912.5_{\pm 1.8}$	$2174.7_{\pm 1.2}$	$3539.4_{\pm 1.2}$	$85.6_{\pm 0.0}$	$42.8_{\pm 0.0}$	$298.5_{\pm 0.0}$	$298.7_{\pm 0.0}$
LwF	$1817.4_{\pm 1.6}$	$2012.0_{\pm 0.0}$	$3376.8_{\pm 0.0}$	$85.6_{\pm 0.0}$	$85.6_{\pm 0.0}$	$85.7_{\pm 0.0}$	$85.8_{\pm 0.0}$
WSN	–	–	–	–	–	–	–
PEC	$8560.9_{\pm 0.0}$	$12777.9_{\pm 0.0}$	$13992.5_{\pm 0.0}$	**$0.9_{\pm 0.0}$**	$89.1_{\pm 0.0}$	$89.1_{\pm 0.0}$	$89.4_{\pm 0.0}$
SpaceNet	**$343.3_{\pm 0.0}$**	**$385.9_{\pm 0.0}$**	**$1596.3_{\pm 0.0}$**	**$40.9_{\pm 0.0}$**	**$20.5_{\pm 0.0}$**	$97.3_{\pm 0.0}$	$51.2_{\pm 0.0}$
NISPA	$2022.2_{\pm 57.8}$	$4443.2_{\pm 334.7}$	$5659.8_{\pm 334.7}$	$272.2_{\pm 0.0}$	$136.1_{\pm 0.0}$	$306.2_{\pm 0.0}$	$306.2_{\pm 0.0}$
SNV	$3233.3_{\pm 1.6}$	$5047.1_{\pm 1.2}$	$6265.9_{\pm 1.2}$	$89.4_{\pm 0.0}$	$45.0_{\pm 0.0}$	**$73.5_{\pm 0.0}$**	**$47.0_{\pm 0.0}$**
UniCLUN	$3387.9_{\pm 1.4}$	$4540.3_{\pm 40.0}$	$5759.1_{\pm 40.0}$	$92.4_{\pm 0.0}$	$92.4_{\pm 0.0}$	$141.6_{\pm 0.0}$	$141.8_{\pm 0.0}$
MCL (ours)	$1819.0_{\pm 0.7}$	$2014.0_{\pm 0.0}$	$3378.8_{\pm 0.0}$	$85.6_{\pm 0.0}$	$85.6_{\pm 0.0}$	$86.1_{\pm 0.0}$	$86.3_{\pm 0.0}$
SGD (lower bound)	$1431.2_{\pm 0.1}$	$1804.0_{\pm 0.0}$	$3168.8_{\pm 0.0}$	$85.6_{\pm 0.0}$	$42.8_{\pm 0.0}$	$42.8_{\pm 0.0}$	$42.9_{\pm 0.0}$
RTI							
Joint training (upper bound)	$628_{\pm 26}$	$909_{\pm 25}$	$2340_{\pm 25}$	89.4	89.4	89.5	$44.8_{\pm 0.0}$
EWC	$755_{\pm 52}$	$842_{\pm 46}$	$2273_{\pm 46}$	89.4	44.7	134.1	$134.2_{\pm 0.0}$
SI	$1244_{\pm 2}$	$1314_{\pm 17}$	$2745_{\pm 17}$	89.4	44.7	312.9	$313.1_{\pm 0.0}$
LwF	$659_{\pm 1}$	$734_{\pm 22}$	$2165_{\pm 22}$	89.4	89.4	89.5	$89.6_{\pm 0.0}$
WSN	–	–	–	–	–	–	–
PEC	2176	$3212_{\pm 1}$	$4643_{\pm 1}$	**0.9**	49.0	49.0	49.0
SpaceNet	**105$_{\pm 0}$**	**120**	**1534**	11.1	**5.6**	**26.5**	**$14.0_{\pm 0.0}$**
NISPA	$395_{\pm 5}$	$705_{\pm 42}$	$2136_{\pm 42}$	69.4	34.7	78.0	$78.1_{\pm 0.0}$
SNV	$826_{\pm 3}$	$1190_{\pm 1}$	$2621_{\pm 1}$	89.4	44.7	49.9	$46.8_{\pm 0.0}$
UniCLUN	$1210_{\pm 1}$	1441	2874	92.0	92.0	104.4	$104.5_{\pm 0.0}$
MCL (ours)	$674_{\pm 29}$	$766_{\pm 20}$	$2197_{\pm 20}$	89.4	89.4	89.6	$89.6_{\pm 0.0}$
SGD (lower bound)	566	640	2071	89.4	44.7	44.7	$44.8_{\pm 0.0}$

Table 6: Inference cost comparison.

	Task-IL				Class-IL			
Method	B1 ms/sample↓	GFLOPs/sample↓	B64 ms/sample↓	B64 sample/s↑	B1 ms/sample↓	GFLOPs/sample↓	B64 ms/sample↓	B64 sample/s↑
Seq-CIFAR-100								
Joint training (upper bound)	$1.837_{\pm 0.229}$	$1.111_{\pm 0.000}$	$0.0266_{\pm 0.0012}$	$38216.0_{\pm 1351.0}$	$1.844_{\pm 0.122}$	$1.111_{\pm 0.000}$	$0.0271_{\pm 0.0002}$	$37238.2_{\pm 120.2}$
EWC	$1.624_{\pm 0.112}$	$1.111_{\pm 0.000}$	$0.0247_{\pm 0.0006}$	$40610.6_{\pm 967.4}$	$1.755_{\pm 0.053}$	$1.111_{\pm 0.000}$	$0.0260_{\pm 0.0004}$	$38555.8_{\pm 612.4}$
SI	$1.756_{\pm 0.100}$	$1.111_{\pm 0.000}$	$0.0258_{\pm 0.0003}$	$39074.8_{\pm 502.2}$	$1.785_{\pm 0.148}$	$1.111_{\pm 0.000}$	$0.0270_{\pm 0.0010}$	$37464.4_{\pm 1043.3}$
LwF	$1.549_{\pm 0.039}$	$1.111_{\pm 0.000}$	$0.0250_{\pm 0.0003}$	$40058.7_{\pm 361.4}$	$1.861_{\pm 0.081}$	$1.111_{\pm 0.000}$	$0.0267_{\pm 0.0003}$	$37556.1_{\pm 373.1}$
WSN	$8.109_{\pm 0.404}$	$1.111_{\pm 0.000}$	$0.1233_{\pm 0.0038}$	$8556.2_{\pm 192.1}$	–	–	–	–
PEC	–	–	–	–	$6.270_{\pm 0.671}$	$0.365_{\pm 0.000}$	$0.2209_{\pm 0.0004}$	$4526.5_{\pm 7.3}$
SpaceNet	$0.377_{\pm 0.142}$	**$0.003_{\pm 0.000}$**	**$0.0055_{\pm 0.0012}$**	**$219715.3_{\pm 48405.0}$**	$0.271_{\pm 0.113}$	**$0.003_{\pm 0.000}$**	**$0.0041_{\pm 0.0006}$**	**$281369.5_{\pm 37115.0}$**
NISPA	$0.761_{\pm 0.212}$	$0.209_{\pm 0.000}$	$0.0121_{\pm 0.0028}$	$95640.6_{\pm 16886.8}$	$0.611_{\pm 0.131}$	$0.209_{\pm 0.000}$	$0.0104_{\pm 0.0007}$	$100432.2_{\pm 5334.5}$
SNV	$8.086_{\pm 0.280}$	$1.111_{\pm 0.000}$	$0.1237_{\pm 0.0071}$	$8500.0_{\pm 394.3}$	$121.286_{\pm 4.129}$	$44.434_{\pm 0.000}$	$1.9836_{\pm 0.0468}$	$505.1_{\pm 11.1}$
UniCLUN	$1.258_{\pm 0.095}$	$1.111_{\pm 0.000}$	$0.0249_{\pm 0.0009}$	$40523.6_{\pm 988.8}$	$1.806_{\pm 0.227}$	$1.111_{\pm 0.000}$	$0.0264_{\pm 0.0014}$	$38236.6_{\pm 1506.7}$
MCL (ours)	$1.818_{\pm 0.292}$	$1.111_{\pm 0.000}$	$0.0337_{\pm 0.0148}$	$38440.6_{\pm 2575.9}$	$2.848_{\pm 0.361}$	$1.111_{\pm 0.000}$	$0.0350_{\pm 0.0018}$	$29993.5_{\pm 1003.4}$
SGD (lower bound)	$1.561_{\pm 0.266}$	$1.111_{\pm 0.000}$	$0.0254_{\pm 0.0006}$	$39679.7_{\pm 726.6}$	$1.708_{\pm 0.125}$	$1.111_{\pm 0.000}$	$0.0265_{\pm 0.0005}$	$37980.6_{\pm 529.0}$
Seq-Tiny-ImageNet								
Joint training (upper bound)	$2.34_{\pm 0.18}$	$4.443_{\pm 0.000}$	$0.093_{\pm 0.002}$	$10783_{\pm 197}$	$2.64_{\pm 0.33}$	$4.443_{\pm 0.000}$	$0.093_{\pm 0.001}$	$10787_{\pm 119}$
EWC	$0.971_{\pm 0.368}$	$4.443_{\pm 0.000}$	$0.0969_{\pm 0.0025}$	$10323_{\pm 271}$	$0.941_{\pm 0.065}$	$4.443_{\pm 0.000}$	$0.0974_{\pm 0.0017}$	$10273_{\pm 181}$
SI	$0.880_{\pm 0.047}$	$4.443_{\pm 0.000}$	$0.0971_{\pm 0.0009}$	$10299_{\pm 92}$	$0.899_{\pm 0.032}$	$4.443_{\pm 0.000}$	$0.0974_{\pm 0.0007}$	$10269_{\pm 69}$
LwF	$0.820_{\pm 0.070}$	$4.443_{\pm 0.000}$	$0.0976_{\pm 0.0018}$	$10244_{\pm 185}$	$0.974_{\pm 0.029}$	$4.443_{\pm 0.000}$	$0.0980_{\pm 0.0003}$	$10209_{\pm 33}$
WSN	$7.15_{\pm 0.22}$	$4.44_{\pm 0.00}$	$0.153_{\pm 0.004}$	$6575_{\pm 151}$	–	–	–	–
PEC	–	–	–	–	$12.45_{\pm 0.48}$	$1.806_{\pm 0.000}$	$0.919_{\pm 0.005}$	$1088_{\pm 6}$
SpaceNet	$0.48_{\pm 0.16}$	$0.01_{\pm 0.00}$	**$0.008_{\pm 0.001}$**	**$171746_{\pm 10934}$**	$0.50_{\pm 0.12}$	$0.01_{\pm 0.00}$	**$0.009_{\pm 0.001}$**	**$182344_{\pm 30154}$**
NISPA	$1.41_{\pm 0.12}$	$0.84_{\pm 0.00}$	$0.033_{\pm 0.001}$	$30396_{\pm 655}$	$1.51_{\pm 0.14}$	$0.84_{\pm 0.00}$	$0.033_{\pm 0.001}$	$30604_{\pm 829}$
SNV	$7.71_{\pm 0.07}$	$4.44_{\pm 0.00}$	$0.162_{\pm 0.000}$	$6192_{\pm 3}$	$128.04_{\pm 3.00}$	$177.73_{\pm 0.00}$	$4.669_{\pm 0.038}$	$214_{\pm 2}$
UniCLUN	$2.28_{\pm 0.28}$	$4.44_{\pm 0.00}$	$0.094_{\pm 0.001}$	$10638_{\pm 111}$	$2.28_{\pm 0.42}$	$4.44_{\pm 0.00}$	$0.094_{\pm 0.001}$	$10654_{\pm 76}$
MCL (ours)	$0.840_{\pm 0.049}$	$4.443_{\pm 0.000}$	$0.0956_{\pm 0.0008}$	$10466_{\pm 89}$	$1.495_{\pm 0.102}$	$4.443_{\pm 0.000}$	$0.0961_{\pm 0.0002}$	$10406_{\pm 16}$
SGD (lower bound)	$0.843_{\pm 0.129}$	$4.443_{\pm 0.000}$	$0.0951_{\pm 0.0014}$	$10521_{\pm 153}$	$1.044_{\pm 0.375}$	$4.443_{\pm 0.000}$	$0.0972_{\pm 0.0012}$	$10288_{\pm 129}$
RTI								
Joint training (upper bound)	$0.649_{\pm 0.018}$	1.111	$0.0215_{\pm 0.0000}$	$46415_{\pm 54}$	$0.660_{\pm 0.030}$	1.111	$0.0228_{\pm 0.0001}$	$43929_{\pm 124}$
EWC	$0.692_{\pm 0.033}$	1.111	$0.0216_{\pm 0.0000}$	$46200_{\pm 45}$	$0.704_{\pm 0.023}$	1.111	$0.0228_{\pm 0.0001}$	$43917_{\pm 152}$
SI	$0.662_{\pm 0.001}$	1.111	$0.0217_{\pm 0.0001}$	$45998_{\pm 107}$	$0.659_{\pm 0.035}$	1.111	$0.0229_{\pm 0.0000}$	$43665_{\pm 28}$
LwF	$0.643_{\pm 0.029}$	1.111	$0.0218_{\pm 0.0001}$	$45896_{\pm 118}$	$0.658_{\pm 0.031}$	1.111	$0.0229_{\pm 0.0000}$	$43751_{\pm 27}$
WSN	$3.260_{\pm 0.171}$	1.111	$0.0529_{\pm 0.0001}$	$18894_{\pm 45}$	–	–	–	–
PEC	–	–	–	–	$0.758_{\pm 0.076}$	0.2675	$0.1388_{\pm 0.0003}$	$7208_{\pm 16}$
SpaceNet	$0.074_{\pm 0.017}$	**0.002786**	**$0.0013_{\pm 0.0001}$**	**$772264_{\pm 43835}$**	$0.075_{\pm 0.009}$	**0.002786**	**$0.0014_{\pm 0.0000}$**	**$748739_{\pm 3971}$**
NISPA	$0.152_{\pm 0.013}$	0.2091	$0.0063_{\pm 0.0000}$	$157682_{\pm 279}$	$0.147_{\pm 0.007}$	0.2091	$0.0063_{\pm 0.0000}$	$157484_{\pm 455}$
SNV	$3.247_{\pm 0.112}$	1.111	$0.0532_{\pm 0.0009}$	$18797_{\pm 318}$	$44.12_{\pm 1.51}$	44.434	$1.6901_{\pm 0.0398}$	$592_{\pm 14}$
UniCLUN	$0.624_{\pm 0.029}$	1.111	$0.0215_{\pm 0.0000}$	$46581_{\pm 25}$	$0.635_{\pm 0.028}$	1.111	$0.0227_{\pm 0.0000}$	$44123_{\pm 70}$
MCL (ours)	$0.644_{\pm 0.028}$	1.111	$0.0216_{\pm 0.0001}$	$46193_{\pm 170}$	$0.936_{\pm 0.008}$	1.111	$0.0241_{\pm 0.0000}$	$41553_{\pm 7}$
SGD (lower bound)	$0.631_{\pm 0.021}$	1.111	$0.0217_{\pm 0.0000}$	$46180_{\pm 57}$	$0.623_{\pm 0.010}$	1.111	$0.0227_{\pm 0.0000}$	$43970_{\pm 35}$

E.1 Qualitative Results

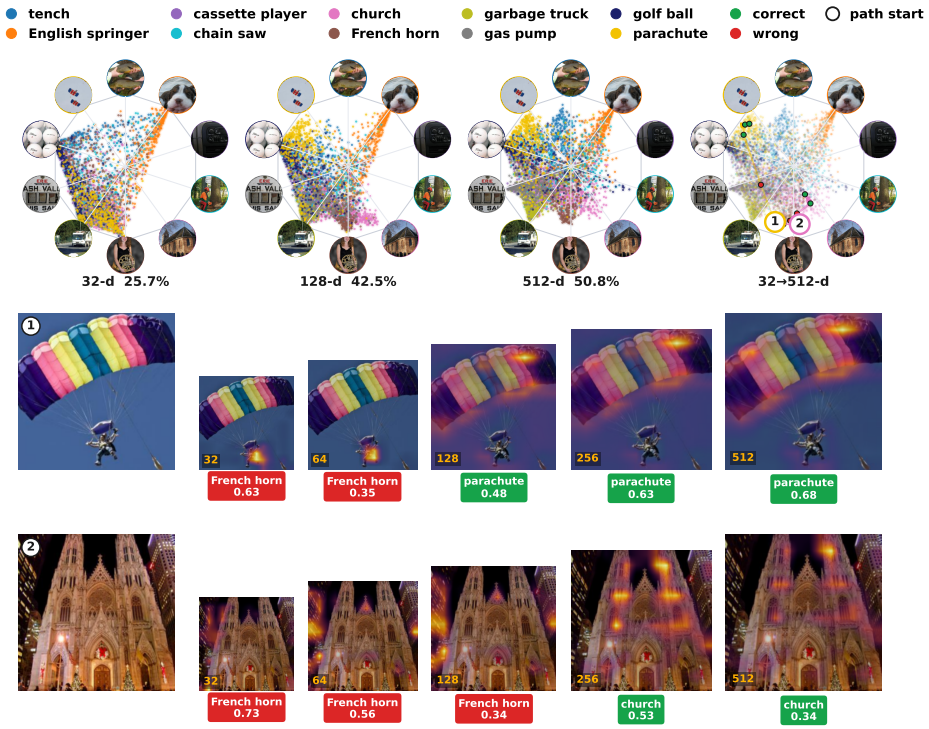

Figure 8: Top: class polygon at 32, 128, and 512 dimensions and the two images' paths; bottom: Grad-CAM and predictions at each width.

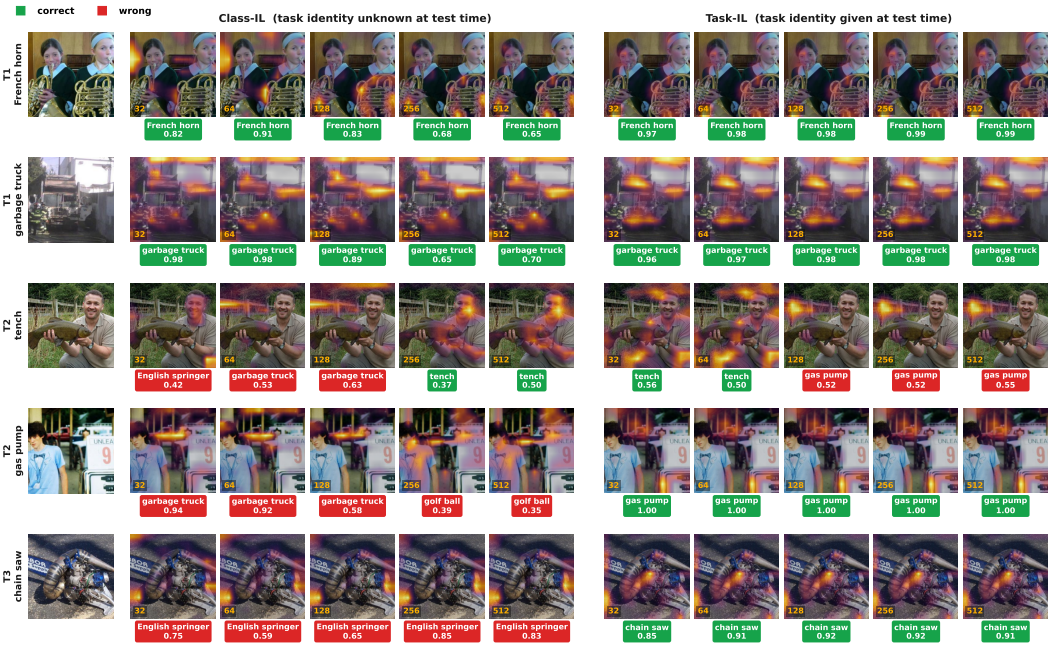

Figure 9: Nesting effect in both Task-IL and Class-IL scenarios.

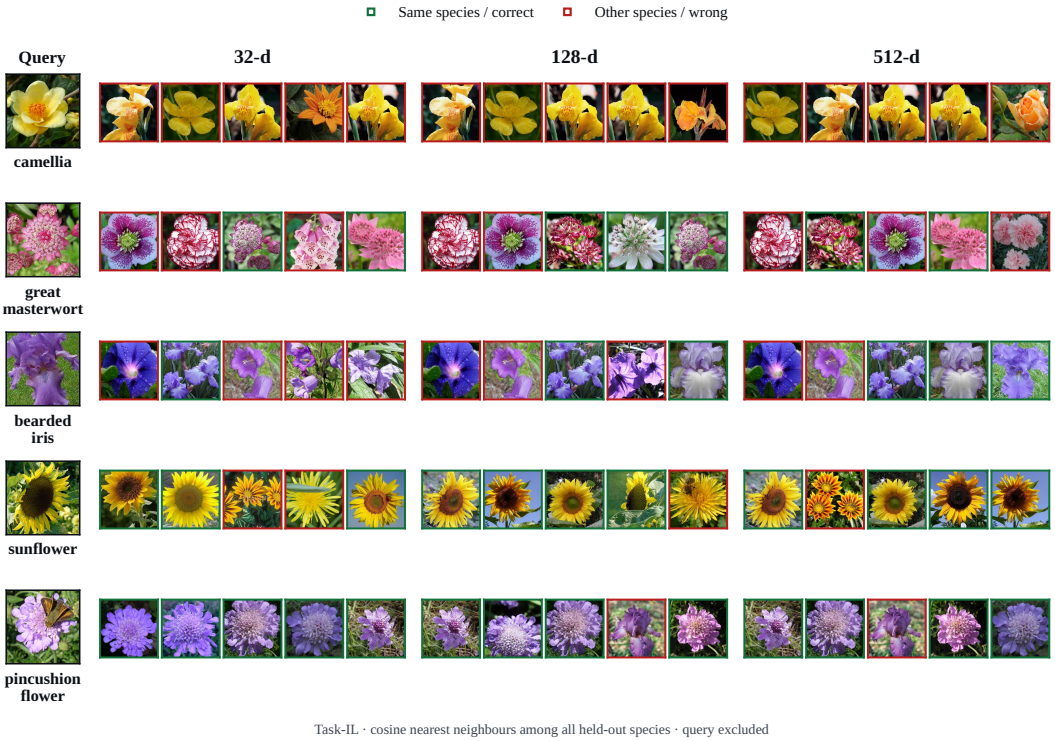

Figure 10: **What the nested prefixes see (task-incremental).** Same queries and layout as Fig. 7, for the final task-incremental MCL model.

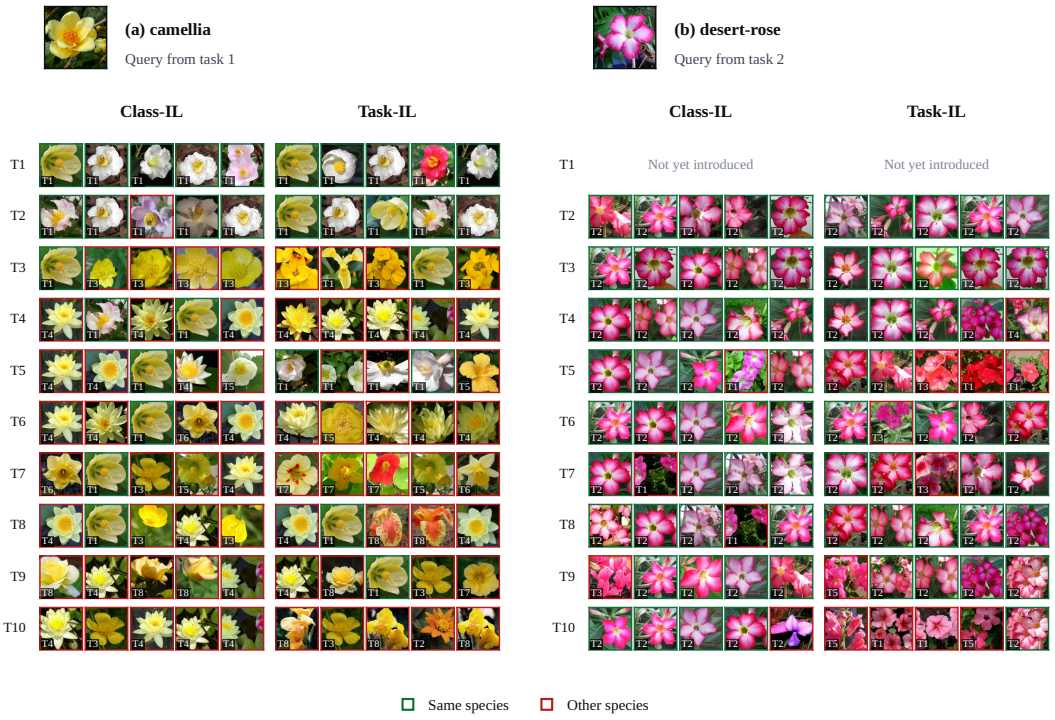

Figure 11: **Evolution of flower retrieval using the 32-dimensional MCL prefix.** The two panels follow fixed held-out queries from task 1 (camellia) and task 2 (desert-rose). Each row shows the five nearest neighbors, ordered by decreasing cosine similarity, after the indicated training task for the Class-IL and Task-IL models. Retrieval uses normalized 32-dimensional features and searches held-out images from all species seen so far in both scenarios, excluding the query itself. Green and red borders denote matching and different species, respectively; T_k tags identify each neighbor's training task.



Figure 12: **How learning new tasks reshapes an old class's neighborhood (512 dimensions).** A held-out French horn image (task 1) and a tench image (task 2), along with their five nearest held-out neighbors among the classes learned so far, after each task; class-incremental (left) and task-incremental (right) MCL. The corner tag gives the neighbor's task; green borders mark the query's class. In the class-incremental model, learning task 3 pulls chain-saw images (T3) into the French-horn neighborhood (3 of 5 neighbors), which recovers by the end of the sequence (5/5 after task 5). The task-incremental neighborhood contains at most one other-class image after any task. The tench neighborhood stays intact in the class-incremental model throughout.

E.2 HYPERPARAMETERS

Table 7: Hyperparameter search space and selected configurations on Seq-CIFAR-100.

		Task-IL Scenario		Class-IL Scenario	
Method	Search space	Trial	Selected configuration	Trial	Selected configuration
EWC	`lr` $\sim$ LogUniform$(10^{-4}, 10^{-2})$ `ewc_gamma` $\sim$ Uniform$(0.9, 1.0)$ `ewc_lambda` $\sim$ LogUniform$(1, 10^4)$	4	`lr=0.000706`, `ewc_gamma=0.912`, `ewc_lambda=2029.6`	23	`lr=0.00213`, `ewc_gamma=0.970`, `ewc_lambda=1.749`
SI	`lr` $\sim$ LogUniform$(10^{-4}, 10^{-2})$ `si_c` $\sim$ LogUniform$(10^{-2}, 10)$ `si_xi` $\sim$ LogUniform$(10^{-4}, 10^{-2})$	2	`lr=0.000131`, `si_c=0.333`, `si_xi=0.000119`	0	`lr=0.000444`, `si_c=0.0284`, `si_xi=0.00200`
LwF	Candidate list (30 configs) `lr` $\in$ [1.1e-4, 8.4e-3] `lwf_lambda` $\in [0.14, 9.13]$ `temperature` $\in \{1, 2, 3, 4\}$	13	`lr=0.00173`, `lwf_lambda=0.984`, `temperature=4`	27	`lr=0.000111`, `lwf_lambda=0.838`, `temperature=2`
WSN	Candidate list `lr` $\in$ [1.2e-4, 7.9e-3] `wsn_density` $\in \{0.1, 0.3, 0.5, 0.7\}$	15	`lr=0.00132`, `wsn_density=0.3`	–	–
PEC	`lr` $\sim$ LogUniform$(10^{-4}, 10^{-2})$	–	–	3	`lr=0.000140`
SpaceNet	`lr` $\sim$ LogUniform$(10^{-4}, 10^{-2})$ `density_factor` $\in \{0.5, 1.0, 1.5\}$ `rewire_fraction` $\in \{0.1, 0.2, 0.3\}$	8	`lr=0.00147`, `density_factor=0.5`, `rewire_fraction=0.3`	11	`lr=0.000689`, `density_factor=1.5`, `rewire_fraction=0.1`
NISPA	`lr` $\sim$ LogUniform$(10^{-4}, 10^{-2})$ `nispa_prune_perc` $\sim$ Uniform$(70, 95)$ `nispa_recovery_perc` $\sim$ Uniform$(1, 5)$	10	`lr=0.00190`, `nispa_prune_perc=79.3`, `nispa_recovery_perc=3.19`	6	`lr=0.00143`, `nispa_prune_perc=79.9`, `nispa_recovery_perc=4.91`
SNV	Candidate list (23 of the 30 draws of `lr` $\sim$ LogUniform$(10^{-4}, 10^{-2})$, keeping `lr` $\geq$ 3e-4) `truncation` $\in \{0.05, 0.1, 0.2\}$	14	`lr=0.00130`, `truncation=0.05`, `rot_aux=0`	14	`lr=0.00130`, `truncation=0.05`, `rot_aux=1.0`, `routing=rot+energy`
UniCLUN	`lr` $\sim$ LogUniform$(10^{-4}, 10^{-2})$ `alpha1, alpha2, alpha3` $\sim$ LogUniform$(0.1, 10)$	8	`lr=0.00125`, `alpha1=0.134`, `alpha2=0.132`, `alpha3=0.258`	10	`lr=0.000806`, `alpha1=0.398`, `alpha2=3.879`, `alpha3=2.500`
MCL (ours)	Candidate list (30 configs) `lr` $\in$ [1.1e$-$4, 8.4e$-$3] `lwf_lambda` $\in [0.14, 9.13]$ `temperature` $\in \{1, 2, 3, 4\}$ Fixed: `density_alpha=0.25`	9	`lr=0.000689`, `lwf_lambda=1.206`, `temperature=3`	9	`lr=0.000689`, `lwf_lambda=1.206`, `temperature=3`

Table 8: Hyperparameter search space and selected configurations on Seq-Tiny-ImageNet.

		Task-IL Scenario		Class-IL Scenario	
Method	Search space	Trial	Selected configuration	Trial	Selected configuration
EWC	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{ewc_lambda} \sim \text{LogUniform}(1, 10^{4})$ $\texttt{ewc_gamma} \sim \text{Uniform}(0.9, 1.0)$	30	lr=0.00196905, ewc_lambda=9383.91, ewc_gamma=0.982192	30	lr=0.000444289, ewc_lambda=4.01233, ewc_gamma=0.965093
SI	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{si_c} \sim \text{LogUniform}(0.01, 10)$ $\texttt{si_xi} \sim \text{LogUniform}(10^{-4}, 10^{-2})$	30	lr=0.00553151, si_c=0.0174479, si_xi=0.000791361	30	lr=0.000444289, si_c=0.0283496, si_xi=0.00200387
LwF	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{lwf_lambda} \sim \text{LogUniform}(0.1, 10)$ $\texttt{temperature} \in \{1, 2, 3, 4\}$	30	lr=0.00112303, lwf_lambda=5.62698, temperature=4	30	lr=0.00011095, lwf_lambda=0.838283, temperature=2
WSN	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{wsn_density} \in \{0.1, 0.3, 0.5, 0.7\}$	$15^{\dagger}$	lr=0.0013198195, wsn_density=0.3	–	–
PEC	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$	–	–	$3^{\dagger}$	lr=0.000140
SpaceNet	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{density_factor} \in \{0.5, 1.0, 1.5\}$ $\texttt{rewire_fraction} \in \{0.1, 0.2, 0.3\}$	$8^{\dagger}$	lr=0.0014655269, density_factor=0.5, rewire_fraction=0.3	$11^{\dagger}$	lr=0.0006890934, density_factor=1.5, rewire_fraction=0.1
NISPA	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{nispa_prune_perc} \sim \text{Uniform}(70, 95)$ $\texttt{nispa_recovery_perc} \sim \text{Uniform}(1, 5)$	$10^{\dagger}$	lr=0.0018959503, nispa_prune_perc=79.310, nispa_recovery_perc=3.191	$6^{\dagger}$	lr=0.0014262836, nispa_prune_perc=79.917, nispa_recovery_perc=4.905
SNV	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{truncation} \in \{0.05, 0.1, 0.2\}$	$14^{\dagger}$	lr=0.0012981647, truncation=0.05	$14^{\dagger}$	lr=0.0012981647, truncation=0.05
UniCLUN	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{alpha1, alpha2, alpha3} \sim \text{LogUniform}(0.1, 10)$	$8^{\dagger}$	lr=0.0012459165, alpha1=0.1335, alpha2=0.1316, alpha3=0.2582	$10^{\dagger}$	lr=0.0008060626, alpha1=0.3977, alpha2=3.8793, alpha3=2.5003
MCL (ours)	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$ $\texttt{lwf_lambda} \sim \text{LogUniform}(0.1, 10)$ $\texttt{temperature} \in \{1, 2, 4\}$ $\texttt{mcl_density_alpha} \in \{0.25, 0.5, 1.0\}$	30	lr=0.00196905, lwf_lambda=9.68706, temperature=2, mcl_density_alpha=0.5	30	lr=0.00196905, lwf_lambda=9.68706, temperature=2, mcl_density_alpha=0.5
SGD (lower bound)	$\texttt{lr} \sim \text{LogUniform}(10^{-4}, 10^{-2})$	30	lr=0.00785992	30	lr=0.000413965

Table 9: Hyperparameter search space and selected configurations on RTI.

		Task-IL Scenario		Class-IL Scenario	
Method	Search space	Trial	Selected configuration	Trial	Selected configuration
Joint (upper bound)	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$	7	`lr=0.00103`	9	`lr=0.000737`
EWC	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$ `ewc_gamma` $\sim \mathrm{Uniform}(0.9, 1.0)$ `ewc_lambda` $\sim \mathrm{LogUniform}(1, 10^{4})$	4	`lr=0.000706`, `ewc_gamma=0.912`, `ewc_lambda=2030`	23	`lr=0.00213`, `ewc_gamma=0.970`, `ewc_lambda=1.75`
SI	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$ `si_c` $\sim \mathrm{LogUniform}(10^{-2}, 10)$ `si_xi` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$	2	`lr=0.000131`, `si_c=0.333`, `si_xi=0.000119`	27	`lr=0.000171`, `si_c=0.015`, `si_xi=0.00344`
LwF	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$ `lwf_lambda` $\sim \mathrm{LogUniform}(0.1, 10)$ `temperature` $\in \{1, 2, 3, 4\}$	11	`lr=0.00146`, `lwf_lambda=1.90`, `temperature=3`	23	`lr=0.00144`, `lwf_lambda=0.817`, `temperature=1`
WSN	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$ `wsn_density` $\in \{0.1, 0.3, 0.5, 0.7\}$	25	`lr=0.00112`, `wsn_density=0.5`	–	–
PEC	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$	–	–	11	`lr=0.000152`
SpaceNet	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$ `density_factor` $\in \{0.5, 1.0, 1.5\}$ `rewire_fraction` $\in \{0.1, 0.2, 0.3\}$	8	`lr=0.00147`, `density_factor=0.5`, `rewire_fraction=0.3`	14	`lr=0.00190`, `density_factor=1.0`, `rewire_fraction=0.1`

NISPA	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$
	`nispa_prune_perc` $\sim \mathrm{Uniform}(70, 95)$
	`nispa_recovery_perc` $\sim \mathrm{Uniform}(1, 5)$
		3	`lr=0.000737`,
			`nispa_prune_perc=71.7`,
			`nispa_recovery_perc`
			`=1.36`
				3	`lr=0.000737`,
					`nispa_prune_perc=71.7`,
					`nispa_recovery_perc`
					`=1.36`

UniCLUN	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$ `alpha1, alpha2, alpha3` $\sim \mathrm{LogUniform}(0.1, 10)$ each	26	`lr=0.0015074642`, `alpha1=0.335346084`, `alpha2=0.101903056`, `alpha3=0.688482653`	27	`lr=0.000548`, `alpha1=1.36`, `alpha2=8.06`, `alpha3=2.40`
MCL (ours)	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$ `lwf_lambda` $\sim \mathrm{LogUniform}(0.1, 10)$ `temperature` $\in \{1, 2, 4\}$ `mcl_density_alpha` $\in \{0.25, 0.5, 1.0\}$	15	`lr=0.000977761019`, `lwf_lambda=0.486352755`, `temperature=2`, `mcl_density_alpha=0.5`	12	`lr=0.000716447896`, `lwf_lambda=0.424907445`, `temperature=4`, `mcl_density_alpha=0.5`
SGD (lower bound)	`lr` $\sim \mathrm{LogUniform}(10^{-4}, 10^{-2})$	10	`lr=0.000138`	9	`lr=0.000737`