CONTINUAL PRETRAINER IS AN INCREMENTAL MODEL GENERALIZER

Anonymous authors
Paper under double-blind review

ABSTRACT

With the necessity of lifelong-learnable machines over continuously changing real-world problems in practice, there has been rapid progress in continual learning these days. However, most recent works on continual learning focuses on alleviating catastrophic forgetting of a model trained over a sequence of vision tasks, considering only the performance on the tasks themselves rather than the representation transferability. In this paper, we tackle a novel problem of Continual Pre-training, which aims to increment the generalization of model representations, encouraging positive transfer for future problems. An initial empirical study shows a rather surprising finding that the transfer quality of the pre-trained model representation with both supervised and unsupervised task sequences does not show noticeable performance degradation even with full-finetuning. Furthermore, we propose a simple yet efficient Continual Pre-training method with GLobal Attention Discretization (GLAD) which introduces a new constraint to increment the global transferability of the backbone while projecting model weights to adapt to target problems via additional weight vectors. Our continual pre-training method breaks the barriers between pre-training and fine-tuning steps and leads to an integrated design that combines continual representation learning with continual learning of the task-specific learners.

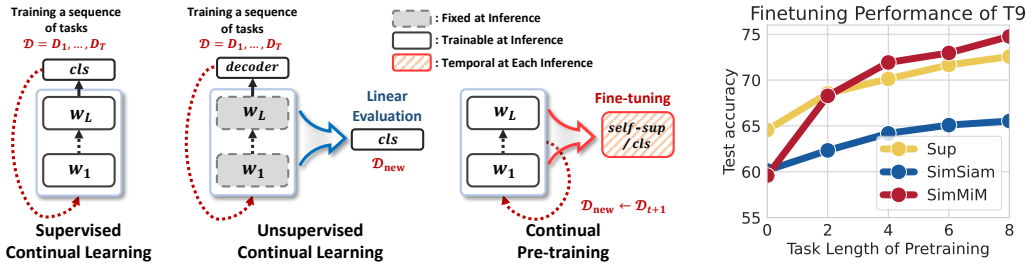
1 INTRODUCTION

Unsupervised Representation Learning (URL) (Radford et al., 2015; Gidaris et al., 2018; Grill et al., 2020; Xie et al., 2021) is a pertinent branch of machine learning where a model exploits data in a self-supervised manner without human-generated signals to extract the generic representations. Yet, the standard scenario resorts to an IID (Independent and Identically Distributed) dataset, consisting of the complete unlabeled instances before starting the model train. So far as the world is incomplete, the deep learning models in real-world need to update continuously, reflecting on ever-changing environments throughout a lifetime.

It carries the lifelong learnability of the representation model and aims to satisfy the community's thirst to achieve long-term preservation of learned knowledge from a sequence of unlabeled tasks. As motivated by the continual learning (CL) field (Thrun, 1995; Silver & Mercer, 2002; Kumar & Daume III, 2012; Li & Hoiem, 2016), Unsupervised Continual Learning (UCL) (Rao et al., 2019; Madaan et al., 2022; Fini et al., 2022), where continual learner trains on a sequence of unsupervised tasks, has recently been explored to address the limitations of unsupervised representation learning and provide comprehensive analyses and potential for unsupervised continual learning regarding representation quality and forgetting.

However, the recently proposed UCL framework and its interpretations remain clear limitations from the point of view of model generalization. Though maximizing the transferability of the learned representations on target problems is essential for general-purpose ed models, prior works confined the validations and analyses of UCL to linear evaluation on fixed representation model backbones. It is suitable for measuring direct differences in model drift (i.e., catastrophic forgetting) during a sequential pre-training, yet, it cannot disclose the effect of knowledge transfer and adaptation of the model to the unseen target problems.

Beyond limited understandings of UCL from prior works, we provide comprehensive analyses on the transferability of continual (representation) learning models to the in- and out-of-distribution



(a) Illustration of continual training schemes (b) Generalization of SCL/UCL

Figure 1: **(a) Continual Pre-training is concerned with maximizing model generalization for full-finetuning on future tasks.** It performs pre-training on sequential tasks regardless of label supervision, while UCL framework prevents the model from updating backbone weights at evaluation to measure representational drift when it proceeds to the next task. **(b) Continual pre-training increments model generalization.** Sequential pre-training on more tasks further increases fine-tuning performance on the unseen task (T9).

tasks using high-resolution visual images for searching the incremental model generalization, which is one of the foremost desires in representation learning area. We conduct experiments utilizing three different sequential training frameworks, illustrated in Figure 1 (a). Interestingly, when the model trains on a sequence of earlier nine tasks from ImageNet-split containing ten tasks in total, both supervised and unsupervised continual learning frameworks gain steady increases in their transferability to unseen tasks (T9), nullifying catastrophic forgetting (Please see Figure 1 (b)).

We explain these phenomena because the transferability hinges upon rich global features in pre-trained models (Xie et al., 2022a; Wei et al., 2022), while the model mostly loses local features during CL, particularly severe when they train in a supervised/contrastive manner. Therefore, we propose a new UCL framework based on Masked Image Modeling (Xie et al., 2022b; He et al., 2022) that outperforms existing supervised/unsupervised CL frameworks in the model generalization to effectively increment global representation. Lastly, leveraging our motivations, we suggest a new method for continual pre-training encouraging rich global features during supervised fine-tuning, named **GL**obal **A**ttention **D**iscretization (**GLAD**), We believe our observations and proposed approach lead to removing the barriers between the standard pretraining-finetuning scheme and continual learning towards incremental model generalization via never-ending fine-tuning.

We succinctly summarize the main contributions of the paper threefold:

- We unveil the behavior of representational transferability and forgetting of global and local features under multiple supervised/unsupervised continual learning frameworks at scale, with Vision Transformer backbones.
- We suggest a new learning/evaluation paradigm of the popular pretraining-finetuning scheme amalgamating to continual learning that aims to continuously increase the generalization of the pre-training backbone during the endless sequential fine-tuning phases.
- We further suggest a simple yet efficient remedy to increment global feature expressiveness throughout continual pre-training, dubbed *GLAD*, which enables the model rapidly adapts to the target problem while preserving high transfer affinity to unseen future tasks.

2 Related Work

Continual Learning SI (Zenke et al., 2017) introduces an additional surrogate loss that reduces the weight shift during continual learning by maintaining the training trajectory according to the weight importance of previous tasks. DEN (Yoon et al., 2018) adaptively controls the network capacity by adding/pruning parameters when new tasks arrive. DER (Buzzega et al., 2020) stores a few training instances of previous tasks as well as their predicted logits and minimize the similarity to produce similar logit predictions on past tasks. BiC (Wu et al., 2019) adds a new layer at the top of the backbone to correct classification bias on new tasks. Similarly, WA (Zhao et al., 2020) corrects the prediction bias by rescaling the FC layer with averaged weights normalization on past tasks. DyToX (Douillard et al., 2022) adopts ViT and performs ensembled prediction with task-specific

classifiers leveraging additional task-specific tokens. However, dominant research resorts to the sophisticated human annotation of inputs during training a sequence of tasks.

CURL (Rao et al., 2019) learns unsupervised representation on task sequences with a generative model adopting task-specific inference. However, the proposed method is validated for only MNIST-scale datasets due to their limited scalability by design. Madaan et al. (2022) suggest a new unsupervised continual learning framework in a contrastive manner using Siamese structures. They demonstrate the scalability of the proposed framework through comprehensive analyses of learned representations. CaSSLe (Fini et al., 2022) utilizes a similar contrastive self-supervised framework for unsupervised continual learning, yet provides further extensive validations including diverse self-supervised learning backbones over ImageNet-100.

Self-supervised Learning Simsiam (Chen et al., 2020a) maximizes the similarity of input prediction upon two different augmentations using the Siamese network, learning input-invariant self-supervision. BarlowTwins (Zbontar et al., 2021) aims to remove cross-correlation across different feature vector embeddings from Siamese networks. DINO (Caron et al., 2021) distills teacher model predictions to the student by minimizing cross-entropy loss between their predictions, where the teacher model is updated through an exponential moving average from the student model. Unlike contrastive learning-based directions, Masked Image Modeling (MIM) has recently been developed inspired by masked language models for natural language understanding. SimMIM (Xie et al., 2022b) and MAE (He et al., 2022) adopt an encoder-decoder structure that zeroes out random spatial patches in each patchfied image and learns representations by predicting pixel values in masked patches. MSN (Assran et al., 2022) combines Siamese networks with masked modeling that maximize the prediction similarity between patchfied masked inputs and the augmented target views.

3 Recap: Pretraining-Finetuning and Continual Learning

3.1 Pretraining-Finetuning Scheme in Representation Learning

Given a neural network $f_{\boldsymbol{w}}$ parameterized by weights $\boldsymbol{w}$, recent works have addressed the broad machine learning problems described to $\mathcal{D}_{target}$ by optimizing learnable weights with respect to complex objective functions. Beyond statistical initialization of network weights (Glorot & Bengio, 2010; He et al., 2015), pre-training, where leveraging learned weights from scaled benchmark datasets (e.g., ImageNet (Deng et al., 2009)) as the initialization of $\boldsymbol{w}$, has been widely adopted to promote a rapid and stable convergence curve during training. Self-supervised learning (Chen et al., 2020a; He et al., 2020; Caron et al., 2020; 2021; Bardes et al., 2021; Xie et al., 2022a) has recently become prevalent for pre-training, demonstrating superior generalization performance compared to supervised counterparts by capturing task-agnostic input features. Let h and g be an encoder and a decoder parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively, the standard objective function is to minimize self-supervised loss given input data $\boldsymbol{d}$ without supervision as follows:

$$\boldsymbol{\theta}^*,\ \boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \ell\left(g_{\boldsymbol{\phi}} \circ h_{\boldsymbol{\theta}}\left(\boldsymbol{d}\right)\right), \tag{1}$$

where $\circ$ indicates function composition. The loss function is often designed in several formulations based on similarity, identity correlation, and contrastive loss. After the pre-training phase, the encoder transfers learned features to backbone neural networks for fine-tuning, $\boldsymbol{w} \leftarrow \boldsymbol{\theta}^*$.

3.2 Supervised and Unsupervised Continual Learning

Supervised Continual Learning (SCL) (Mallya & Lazebnik, 2018; Riemer et al., 2019; Aljundi et al., 2019; Chaudhry et al., 2019; 2020; Chrysakis & Moens, 2020; Titsias et al., 2020; Shen et al., 2020; Douillard et al., 2022; Yoon et al., 2020; 2022) is about achieving forward positive transfer on unlimited sequential future tasks while maintaining proficiency on previous tasks. Let $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_T\}$ be a sequence of T tasks, where the dataset $\mathcal{D}_t$ for the t-th task consists of n_t training instances $\mathcal{X}_t \in \mathbb{R}^{n_t \times C \times H \times W}$ and corresponding labels $\mathcal{Y}_t \in \mathbb{R}^{n_t}$, where $C, H,$ and W denotes a channel, height, and width of images, respectively. A continual learner $f_{\boldsymbol{w}}$, parameterized by a set of weights $\boldsymbol{w}$, aims to predict classes by minimizing the following optimization problem: $\text{minimize}_{\boldsymbol{w}}\ L\left(f\left(\mathcal{X}_{1:T}; \boldsymbol{w}\right), \mathcal{Y}_{1:T}\right)$, where L is a cross-entropy loss function. Yet, we assume that $f_{\boldsymbol{w}}$ can access each task in a specific timestep that loses the authorization to revisit the data instances

of previous tasks when the next task arrives. That is, the model solves the following non-stationary problem at task t throughout continuous task training:

$$\underset{\boldsymbol{w}}{\text{minimize}} \; L\left(f\left(\mathcal{X}_t; \boldsymbol{w}, \mathcal{D}_{1:t-1}\right), \mathcal{Y}_t\right). \tag{2}$$

Obtained models directly evaluate the performance of each task, categorizing task- and class-incremental learning setups according to the accessibility to task oracle during inference.

For Unsupervised Continual Learning (UCL), the t-th task consists of training instances $\mathcal{X}_t$ without any human-annotated labels and is formulated in representation learning frameworks on a sequence of unlabeled tasks, often referred to as continual self-supervised learning. A representation learner $f_{\boldsymbol{w}}$ aims to achieve the best solution that learns the informative representation from entire datasets by minimizing the following optimization problem: $\text{minimize}_{\boldsymbol{w}} \; L\left(f\left(\mathcal{X}_{1:T}; \boldsymbol{w}\right)\right)$, where L is an arbitrary loss function for representation learning (e.g., self-supervised losses (Chen et al., 2020b; Grill et al., 2020; Zbontar et al., 2021)). Similar to Section 3.2, at each timestep t, the model resorts to the accessible dataset $\mathcal{D}_t$ to solve the non-stationary problem as follows:

$$\boldsymbol{w}^* = arg\min_{\boldsymbol{w}} \; L\left(f\left(\mathcal{X}_t; \boldsymbol{w}, \mathcal{X}_{1:t-1}\right)\right) \approx arg\min_{\boldsymbol{w}} \; L\left(f\left(\mathcal{X}_{1:t}; \boldsymbol{w}\right)\right). \tag{3}$$

Since a direct comparison of the quality of representation models is intractable, recent representation learning literature validates obtained representation models by probing generic transferability on multiple downstream tasks. In a similar vein, prior UCL works (Madaan et al., 2022; Fini et al., 2022) adopt supervised prediction tools like the KNN classifier and linear evaluation while keeping the learned backbone fixed. However, we argue that such evaluation paradigms cannot appropriately measure the transferability of representation on unseen out-of-distribution problems.

4 Continual Pre-training for Incremental Model Generalization

4.1 Forgetting and Generalization in Continual Pre-training

Continual learning literature struggles to minimize catastrophic forgetting as a model continuously trains on unseen tasks over time. Prior works have demonstrated that a model with a standard CL setting suffers from forgetting due to loss of local features and attention to past tasks. But, we argue that they still lack understanding of the effect on the model generalization, and a model preserving the localized representation on pretraining tasks may not be beneficial for fine-tuning. I throw a question mark at this point:

"So, is the model generalization getting worse as it goes through training sequential tasks?"

Surprisingly, the answer is *NO*. We found that the generalization is consistently getting increased. To validate the model generalization during task sequential training, we formally define a scenario of continual pre-training, a general-purpose integrated learning framework in which a model adapts to an unlimited number of sequential tasks, encouraging incremental forward knowledge transfer.

Continual Pre-training (CP) is a learning paradigm that a model pre-trained sequential source tasks solves the target task $\mathcal{T}_{target}$ problem in a supervised manner. Note that fine-tuned model on $\mathcal{T}_{target}$ is also used for fine-tuning future tasks. Given the dataset of target task $\mathcal{D}_{target} = \{\mathcal{X}, \mathcal{Y}\}$ and a classifier $\delta_{\boldsymbol{u}}$ parameterized by $\boldsymbol{u}$, we formulate the objective of continual pre-training as follows:

$$\boldsymbol{w}^*, \boldsymbol{u}^* = arg\min_{\boldsymbol{w}^{(t)}, \boldsymbol{u}} \ell\left(f\left(\mathcal{X}; \boldsymbol{w}^{(t)}, \boldsymbol{u}\right), \mathcal{Y}\right), \quad \text{where} \quad \boldsymbol{w}^{(t+1)} = \boldsymbol{w}^*. \tag{4}$$

Each fine-tuning step independently introduces its own classifier. The formulation is aligned with the continual learning problem described in Section 3.2, but the objective is different. While continual learning focuses on preserving local attention on past tasks to increment fine-grained problem-solving skills throughout training tasks sequentially, continual pre-training is about never-ending model generalization to achieve a better adaptation to the out-of-distribution task in the future.

However, the degree of effectiveness varies on the framework, as described in Figure 1 (b). As Xie et al. (2022a) discussed, both supervised (*Sup*) and contrastive self-supervised (*SimSiam* from Madaan et al. (2022)) frameworks train local attention at higher layers in neural networks, which is unsuitable for our motivation. To build a new UCL framework for a better generalizable representation

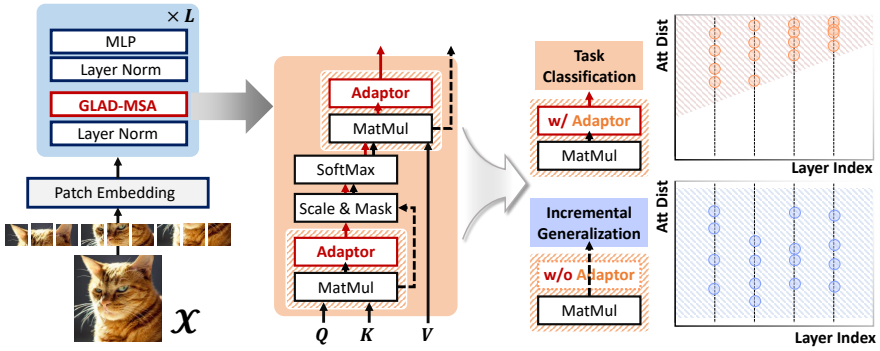


Figure 2: **Illustration of the proposed GLobal Attention Discretization (GLAD) for Continual Pre-training.** Our GLAD introduces a new Multi-head Self-attention operation named GLAD-MSA with a parametric adaptor. The pre-trained model with adaptors fine-tunes given problems with a constraint which encourages divergence of attention distance in each layer, leading to incremental positive transfer of backbone parameters for future tasks.

model, we focus on Masked Image modeling (MIM) (Pathak et al., 2016; He et al., 2022) that self-trains input representation by minimizing regression loss to predict RGB pixel values in randomly zeroed patches in patchfied input images. MIM captures global attention across all layers, allowing better generalization to unseen tasks at fine-tuning. Our proposed UCL framework (*SimMIM*) obtains superior model generalization ability over other supervised and unsupervised counterparts, demonstrating our motivation about the relationship between transferability and global attention.

4.2 Continual Self-supervised Learning with Masked Image Modeling

We formulate a representational learner $f_{\boldsymbol{w}}$, composed of a neural encoder $h_{\boldsymbol{\theta}}$ and a decoder $g_{\boldsymbol{\phi}}$. We build backbones using Vision Transformer variants (Dosovitskiy et al., 2020; Liu et al., 2021) due to their powerful generality and remarkable performance on high-resolution visual tasks. They are flexible to transfer the obtained representations to downstream tasks requiring various input image sizes in demands when existing UCL frameworks (Madaan et al., 2022; Rao et al., 2019) allow the fixed image size for representation learning and fine-tuning since the architectures are basically composed of multi-layer perceptrons and convolutional neural networks. At training t-th task with a training instance $\boldsymbol{x}_t \in \mathbb{R}^{C\times H\times W} \in \mathcal{X}_t$, a model segments $\boldsymbol{x}_t$ into smaller image patches where the width and height are $s < H, W$, and randomly zeros a fraction of image patches out with a fixed ratio τ. An encoder tokenizes masked patches to the embedding space and fed into multiple self-attention blocks to capture latent representation features. A decoder reconstructs encoded features to approximate the input image. The objective is to minimize the following loss function for continual representation learning ($\|\cdot\|_\mu$ denotes any norm, often $\mu \in \{1, 2\}$ is used):

$$\ell\left(\boldsymbol{x}_t; \boldsymbol{w}\right) = \left\|f(\boldsymbol{m} *_s \boldsymbol{x}_t; \boldsymbol{w}) - \boldsymbol{x}_t\right\|_\mu = \left\|g\left(h\left(\boldsymbol{m} *_s \boldsymbol{x}_t; \boldsymbol{\theta}\right); \boldsymbol{\phi}\right) - \boldsymbol{x}_t\right\|_\mu,$$
$$\text{where } \boldsymbol{m} = \{0,1\}^{\lfloor \frac{H}{s} \rfloor \cdot \lfloor \frac{W}{s} \rfloor} \sim B\left(\left\lfloor \frac{H}{s} \right\rfloor \cdot \left\lfloor \frac{W}{s} \right\rfloor, \rho\right). \tag{5}$$

With patch size s, $*_s$ denotes a patch-wise multiplication operation between a training instance $\boldsymbol{x}$ and a generated mask vector $\boldsymbol{m}$ drawn by the binary distribution B with sparsity ratio ρ. $B(i, \rho)$ is a i independent binary sampling with a ratio ρ to pick 1. The model updates a set of weights that predicts masked regions of input images, conditioning other available areas. We simply adopt ℓ_1 regularization to minimize the distance between predicted patches and the targets, followed by earlier reconstruction-based works (Xie et al., 2022b; He et al., 2022), and after completing a sequential training, the obtained encoder $h_{\boldsymbol{\theta}}$ can be utilized for many different downstream tasks. And we find that our proposed framework outperforms supervised and contrastive benchmarks in model transferability during continual pre-training, also described in Figure 1 (b).

4.3 Continual Pre-training via GLobal Attention Discretization

Motivated by our findings in Section 4.1 and Section 4.2, we propose a new method for continual pre-training, named *Global Attention Discretization (GLAD)*. Our proposed method promotes diverse

attention distance to preserve transferable attention for future problems while minimizing current task loss. We first build a backbone with GLAD-MSA illustrated in Figure 2, a multi-head self-attention operation with adaptor vector $\boldsymbol{v} \in \mathbb{R}^{d_{\text{out}}}$, where d_{out} is an output dimension of MSA operation. We project the MSA features by multiplying a diagonal form of an adaptor $\text{diag}(\boldsymbol{v})$ to maintain flexibility to address the current target task while constraining the multi-head attention to preserve locality inductive bias. Let $\boldsymbol{a}^{l,i}$ be attention passed over adaptor operation (*dark dashed arrow*) from i-th head at layer l, the objective function of our GLAD is as follows:

$$arg\min_{\boldsymbol{w},\boldsymbol{v}} \sum_{n=1}^{N} \ell\left(f\left(\boldsymbol{x}^{(n)};\boldsymbol{w},\boldsymbol{v}\right),\boldsymbol{y}^{(n)}\right) \quad s.t., \sum_{l=1}^{L} \log\left(\sqrt{E\left[\left(\boldsymbol{a}^{l,i}-\overline{\boldsymbol{a}}^{l}\right)^{2}\right]}+\epsilon\right) \geq \tau, \tag{6}$$

E indicates the expectation and $\overline{\boldsymbol{a}}^{l} = \frac{1}{H^{l}}\sum_{i}^{H^{l}} \boldsymbol{a}^{l,i}$ at layer l. ϵ is a small constant value. We constrain the log variance of attention to guarantee sufficient divergence of attention heads at each layer with a specific degree $\tau > 0$ by jointly minimizing the task loss with an additional regularizer formulated to the negative sum of layerwise log attention variance. Note that our proposed method is robust to utilize any kind of multi-head self-attention modules, we demonstrate the efficacy in vanilla Vision Transformer (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021). The learned backbone weights $\boldsymbol{w}$ excepting classifier and GLAD-adaptors can be reused for fine-tuning future tasks. We describe the overall continual pre-training procedure with GLAD in Algorithm 1.

Algorithm 1 Continal Pre-training with GLobal Attention Discretization (GLAD)

input A sequence of tasks $\{\mathcal{D}_1, \mathcal{D}_2, \cdots\}$, backbone network f, learning rate $\eta \in \mathbb{R}^{+}$, hyperparameter λ, small constant ϵ, initialization $\boldsymbol{w}_{\text{init}}, \boldsymbol{v}_{\text{init}}$

for all task $\mathcal{T}_t = \mathcal{T}_1, \mathcal{T}_2, \ldots$ **do**
  Build a model $f_{\boldsymbol{w},\boldsymbol{v}}(\cdot)$ with GLAD-MSA ▷ Figure 2
  Initialize $\boldsymbol{w} \leftarrow \boldsymbol{w}^{*}$ excluding classifier **if** $f_{\boldsymbol{w}^{*}}$ **exists, otherwise** $\boldsymbol{w} \leftarrow \boldsymbol{w}_{\text{init}}$
  Initialize $\boldsymbol{v} \leftarrow \boldsymbol{v}_{\text{init}} := (1, \ldots, 1) \in \mathbb{R}^{d_{\text{out}}}$
  for batch $\boldsymbol{x}_n, \boldsymbol{y}_n \sim \mathcal{D}_t$ **do**
    $\mathcal{L} = \ell\left(f\left(\boldsymbol{x}_n;\boldsymbol{w},\boldsymbol{v}\right),\boldsymbol{y}_n\right) + \lambda\sum_{l=1}^{L} - \log\left(\sqrt{E\left[\left(\boldsymbol{a}^{l,i}-\overline{\boldsymbol{a}}^{l}\right)^{2}\right]}+\epsilon\right)$ ▷ Equation 6
    $\boldsymbol{w} \leftarrow \nabla_{\boldsymbol{w}}\mathcal{L}$
    $\boldsymbol{v} \leftarrow \nabla_{\boldsymbol{v}}\mathcal{L}$
  end for
  $\boldsymbol{w}^{*} \leftarrow \boldsymbol{w}$
end for

5 Experiments

5.1 Baselines and Datasets

Backbone architectures and baselines For all experiments, we use ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021) as backbone architectures. We follow Siamese networks by Madaan et al. (2022) and implement a masked image modeling-based continual self-supervised learning framework under SimMIM (Xie et al., 2022b) and MAE (He et al., 2022) for UCL. CURL (Rao et al., 2019) is one of the pioneer works on unsupervised continual learning literature, but it is not scalable for high-resolution visual images by design. We utilize several continual learning methods: Synaptic Intelligence (SI) (Zenke et al., 2017), Dark Experience Replay (DER) (Buzzega et al., 2020), and Lifelong Unsupervised Mixup (LUMP) (Madaan et al., 2022). We further describe details on architectures, CL methods, and hyperparameter setups in Appendix A.

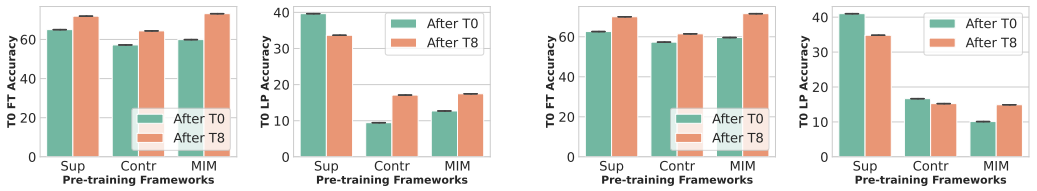
Datasets and validation tasks We use ImageNet-1K dataset (Deng et al., 2009) by splitting it into ten tasks where each task contains 100 classes. We construct a sequential dataset with nine earlier tasks and assign the last task as a validation set for non-iid task evaluation.

5.2 Experimental Results

Forward transfer through continual pre-training We validate the first task (T0)'s evaluation performance according to Continual Pre-training in Table 1, to measure the change of evaluation

Table 1: **Finetuning and linear evaluation performance with their forward transfer** of the first task on ImageNet 1K Split after supervised/unsupervised continual learning. We report the Top-1/Top-5 performance for all individual experiments. Higher is better for both metrics and the best results are highlighted in **bold**.

	(Continual) Pre-training	Supervised		Contrastive (Chen & He, 2021)		Masked Model (Xie et al., 2022b)	
		Final Acc	Fwd Transfer	Final Acc	Fwd Transfer	Final Acc	Fwd Transfer
Finetuning	1K Pretrained	87.48 / 98.08	–	–	–	–	–
	22K Pretrained	87.76 / 98.48	–	–	–	–	–
	Base Model	71.90 / 90.64	6.88 / 3.70	64.38 / 86.18	7.18 / 4.56	73.18 / 91.76	13.26 / 7.46
	SI (Zenke et al., 2017)	70.00 / 90.38	7.40 / 4.52	61.46 / 84.82	4.15 / 2.42	71.54 / 90.76	11.92 / 6.58
	DER (Buzzega et al., 2020)	70.57 / 90.12	8.94 / 4.86	62.37 / 85.46	9.28 / 6.08	70.10 / 90.10	19.55 / 12.24
	LUMP (Madaan et al., 2022)	N/A	N/A	64.01 / 86.42	6.24 / 4.32	**75.11 / 92.38**	**21.28 / 12.42**
Linear Probe	1K Pretrained	87.48 / 97.98	–	–	–	–	–
	22K Pretrained	86.53 / 98.06	–	–	–	–	–
	Base Model	33.66 / 62.20	-5.98 / -4.30	17.10 / 40.60	**7.64 / 13.94**	17.46 / 40.60	4.76 / 6.36
	SI (Zenke et al., 2017)	**34.82 / 63.18**	-6.18 / -5.56	15.24 / 36.76	-1.42 / -2.38	14.92 / 37.80	4.82 / 8.26
	DER (Buzzega et al., 2020)	34.59 / 62.29	-5.86 / -6.16	14.84 / 36.13	3.68 / 5.22	6.22 / 21.52	-0.84 / -1.04
	LUMP (Madaan et al., 2022)	N/A	N/A	18.50 / 42.05	7.54 / 11.38	19.26 / 43.21	7.38 / 11.22



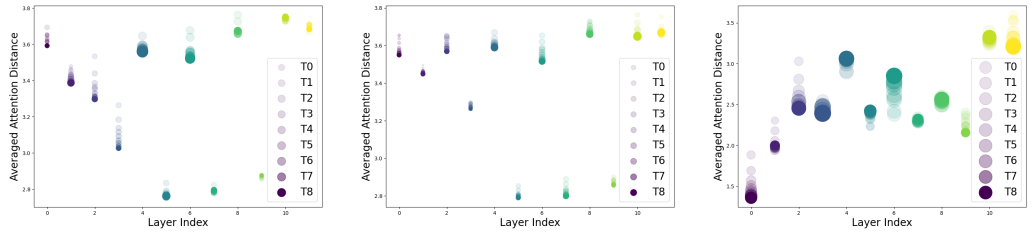
(a) FInetuning - Base (b) Linear Probe - Base (c) FInetuning - SI (d) Linear Probe - SI

Figure 3: **(a-b): Visualization of the model transferability by comparing the accuracy on the Out-of-distribution task (*T9*)** after pre-training the first task (*After T0*) with after pre-training on nine sequential tasks (*After T8*). We perform pre-training under Supervised (Sup), Contrastive Self-supervised Learning (Contr), and Masked Image Modeling (MIM). **(c-d): Same visualization with a CL method**, *SI*, during pre-training.

performance on the in-distribution task. The evaluation of T0 from full pre-trained models over the Imagenet-1K and -22K datasets obtains high validation accuracy on fine-tuning and linear evaluation as they do not perform continual pre-training. Fine-tuning the base continual pre-training models, which perform a simple CL strategy without additional methods during pre-training task sequences, achieves performance increases in T0 as they pre-trained longer task sequences, obtaining positive values in Forward Transfer. The results are similar to all supervised and unsupervised continual pre-training frameworks, including Contrastive Self-supervised Learning (Madaan et al., 2022) and Masked Image Modeling (Section 4.2). We also performed the continual pre-training with multiple continual learning methods. We found that continual learning methods follow consistent tendencies according to the continual pre-training frameworks when LUMP with Masked Image Modeling gains the highest accuracy on T0 with the strong forward transfer. On the other hand, the model degrades the linear evaluation performance in supervised continual pre-training, which had to do with catastrophic forgetting reported in conventional continual learning scenarios, and demonstrates increased performance by combining with continual learning methods.

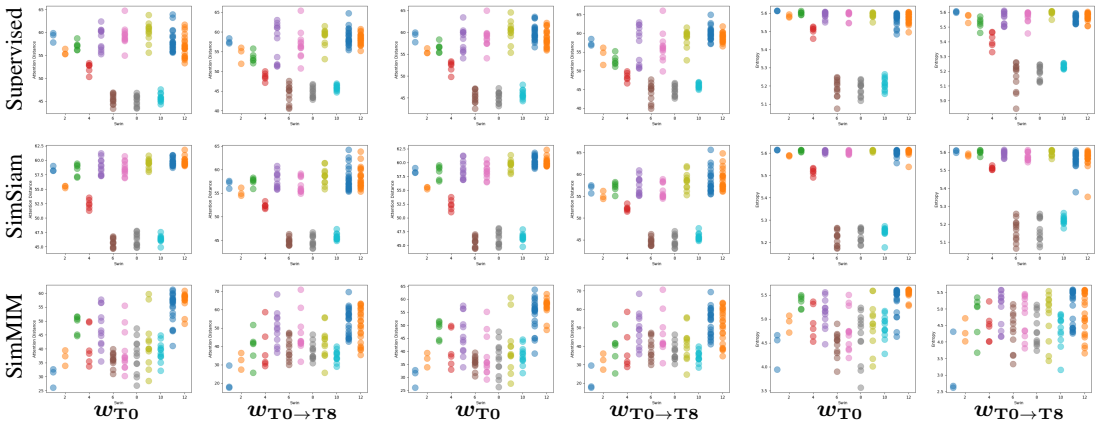
Analyses for an Out-Of-Distribution (OOD) task We denote the last task (T9) as an Out-Of-Distribution problem, excluding it from the pre-training task sequence. In Figure 3, we visualize the top-1 validation accuracy on an OOD task over three continual pre-training frameworks w/ and w/o SI. Similar to in-distribution evaluation, MIM Continual Pre-training achieves higher fine-tuning performance both on the base model and SI. The linear probe performance of Supervised CP surpasses unsupervised counterparts, and we expect that representation from supervised learning contains directly helpful features to classify the high-resolution and complex task problems even without the re-update of backbone weights. In contrast, MIM remarkably underperforms on linear evaluations due to its characteristic property; Masked Modeling focuses on capturing global attention rather than local attention, which provides a better generalization to adapt to unseen tasks but is inadequate for solving the problem.

To understand how continual learning frameworks exhibit incremental model generalization and fine-tuning performance, we analyze the behavior of layer attention during continual pre-training using the Swin-T backbone. In Figure 4, We visualize the layer-by-layer changes in aggregated attention



Supervised SimSiam SimMIM

Figure 4: **Visualization of aggregated attention distance** on an OOD task (T9) at each layer at the end of each continual pre-training task phase (T0→T8). The radius of marker indicates the standard deviation over attention heads in the corresponding layer.



(a) Attention Distance of **T0** (b) Attention Distance of **T9** (c) Attention Entropy of **T9**

Figure 5: **(a-b): Visualization of the attention distance** of an in-distribution (T0) and out-of-distribution task (T9) with respect to three continual pre-trained frameworks right after the completion of the first ($\boldsymbol{w}_{\mathbf{T0}}$) and last task ($\boldsymbol{w}_{\mathbf{T0\to T8}}$). **(c): Visualization of the entropy of each attention head's distribution** of unseen task T9.

distance for T9 while the model pre-trains ImageNet-1K-Split sequentially until the penultimate task (T0→T8). Interestingly, aggregated attention distance significantly decreases the scale and increases the diversity across attention heads. This demonstrates that the continual pre-trainer performs incremental model generalization by getting more tending to memorize richer global attention rather than focusing on local attention. In Supervised and Contrastive Continual Pre-training frameworks, lower layers tend to drastically change toward capturing global attention, and it is also coincident with well-known observations that lower layers in neural networks are more concerned with global features. Also, Masked Image Modeling (SimMIM) results demonstrate the salient effectiveness of capturing global attention compared with the other two frameworks.

Understanding the role of global and local attention We explicate behaviors of multi-head attention in transformer backbones during continual pre-training with different supervised and unsupervised frameworks. We investigate the degree of locality inductive bias in the continual pre-training models by measuring the distance of attention heads at each layer followed by Xie et al. (2022a), visualized in different dots. As shown in Figure 5 (a) and (b), supervised continual pre-training (*Supervised*) captures the locality at lower layers and tends to focus on global attention at upper layers. And we find that the model continuously promotes the locality preference during continual pre-training. Also, the attention distance primarily depends on the pre-training models rather than fine-tuning tasks, whether in- or out-of-distribution. And the results of contrastive self-supervised learning (*SimSiam*) are similar to *Supervised*. Interestingly, Masked Image Modeling Continual Pre-training (*SimMIM*) behaves very differently from the other two frameworks. Like the lower layers, the deeper layers also have a diverse focus on locality and this tendency becomes stronger as they continue to pre-train more tasks.

Next, we visualize the entropy conditioned solely on the distribution of each attention head in Figure 5 (c) by computing $-\sum_i \boldsymbol{a}_i \log(\boldsymbol{a}_i)$ for each attention head $\boldsymbol{a}$. Similar to the analyses of attention distance, we observe that the attention of SimMIM focuses on much broader over many tokens while the other two frameworks concentrate on a few tokens, resulting in narrow attention.

Pre-training T0→T4	Continual Pre-training & Fine-tuning T5→T8			
	T5	T6	T7	T8
Supervised Learning	64.56 / 86.02	69.06 / 89.98	75.06 / 91.44	77.80 / 93.86
+ GLAD (Ours)	**65.78 / 86.84**	**69.84 / 90.52**	**76.12 / 91.76**	**79.04 / 94.54**
SimMIM (Xie et al., 2022b)	**68.34** / 88.24	72.22 / 91.78	77.38 / 93.22	80.12 / 94.74
+ GLAD (Ours)	68.24 / **88.46**	**73.24 / 92.10**	**78.94 / 93.46**	**81.62 / 95.22**
MAE (He et al., 2022)	**42.19 / 70.01**	52.07 / 78.56	63.07 / 85.38	71.19 / 90.94
+ GLAD (Ours)	41.75 / 69.46	**54.74 / 80.84**	**68.01 / 87.59**	**74.30 / 92.19**

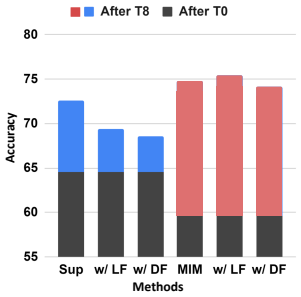


Figure 6: **Left: Per-task fine-tuning (top-1 / top-5) performance on T5 to T8 during continual pre-training** with base frameworks and our proposed method. **Right: Analysis for freezing the weights on a few lower/deeper layers during continual pre-training** after pre-training first task. We use supervised (*Sup*) and masked image modeling (*MIM*). *LF* and *DF* denote *freezing lower layers* and *freezing deeper layers*, respectively.

Freezing the partial layer weights during continual pre-training We further analyze the effect of the layers for incremental generalization during continual pre-training in Figure 6 right. After supervised/unsupervised training of the first task (*After T0*), we freeze the two lowest or two deepest layer weights during the successive continual pre-training up to the final task (*After T8*). For the MIM-based UCL framework, we use MAE with a ViT-B backbone. In supervised learning, both partial gradient update policies reduce the degree of incremental generalization during continual pre-training. It significantly reduces the representation model's fine-tuning performance compared to the fully-trained model. However, interestingly, prohibiting the update of layer weights at both ends less affects MIM-based continual pre-training. We expect that this property comes from its flexibility in learning diverse attention across all layers. For further analyses, please see Appendix B.3.

Incremental model generalization via Global Attention Discretization We now validate our proposed method, *Global Attention Discretization (GLAD)*, which encourages incremental model generalization during supervised continual pre-training. As discussed earlier, supervised training tends to focus on global attention at deeper layers. That is, the model is hard to stray far from the weight space of the limited locality inductive bias, which is evident in the slower movement of averaged attention distance from supervised continual pre-training compared to the SimMIM-based (Please see Figure 4) and results in suboptimal adaptation to arriving tasks. In Figure 6 left, we report the fine-tuning performance during continual pre-training. Note that to see the effect of MIM-based continual pre-training, we first perform continual pre-training over earlier five tasks from ImageNet-1K Split under supervised and MIM-based unsupervised continual learning. Next, the pre-trained models fine-tune the remaining tasks in a sequential manner. We adopt SimMIM and Masked AutoEncoder (MAE) to better understand general behaviors in Masked Image Modeling during unsupervised continual learning. Our proposed GLAD achieves significant gain in the performance of each task during sequential full-finetuning over different pre-trained initialization from supervised and masked image modeling.

6 Conclusion

As powerful representation models have great versability to solve various downstream tasks, exploring incremental pre-training strategy on a number of sequential tasks can be a practical and important approach. This paper delves into how supervised and unsupervised continual learning affects model generalization from various perspectives. To our surprise, continual learning models preserve or even increment their transferability on in- and out-of-distribution tasks, increasing fine-tuning performance as pre-training more tasks. We scrutinize the behavior of representations in continual learning frameworks in the pre-training, including masked image modeling-based unsupervised continual learning, and find that the continual learner tends to forget class-discriminative features while progressively accumulating transferable features. Motivated by our observations, we propose a new method for continual pre-training to help backbone weights gain transferability during fine-tuning by introducing a new MSA module with parametric adaptors. We believe the exploration of continuous learnability of the representation model would contribute to developing eco-friendly and resource-efficient training regimes for broad research/industry fields.

REPRODUCIBILITY STATEMENT

We will release our code and the implementation details to be publicly available for reproducibility. In addition, we provide more details in Appendix for reproducibility, including architecture and baselines setup with hyperparameters.

REFERENCES

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. Continual learning with tiny episodic memories. *arXiv preprint arXiv:1902.10486*, 2019.

Arslan Chaudhry, Naeemullah Khan, Puneet K Dokania, and Philip HS Torr. Continual learning in low-rank orthogonal subspaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Enrico Fini, Victor G Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Rethinking the representational continuity: Towards unsupervised continual learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Hrka5PA7LW.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Gehui Shen, Song Zhang, Xiang Chen, and Zhi-Hong Deng. Generative feature replay with orthogonal weight modification for continual learning. *arXiv preprint arXiv:2005.03490*, 2020.

Daniel L Silver and Robert E Mercer. The task rehearsal method of life-long learning: Overcoming impoverished data. In *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2002.

Sebastian Thrun. *A Lifelong Learning Perspective for Mobile Robot Control*. Elsevier, 1995.

Michalis K Titsias, Jonathan Schwarz, Alexander G de G Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022a.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sk7KsfW0-.

Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gdj2EKPB.

Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. Online coreset selection for rehearsal-based continual learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=f9D-5WNG4Nv.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

A Details for Problem Setups

Datasets For all experiments and analyses, we use ImageNet (Deng et al., 2009) dataset, containing 1000 classes of high-resolution object images with their corresponding labels. We split them into 10 tasks, where each task consists of 100 different classes. We use only 10% of training instances in each task for pre-training, and use the full set for the fine-tuning and linear probe. Accuracy is measured by the validation dataset for each task. Note that Figure 6 Left use full training set for sequential fine-tuning procedure from T5 to T8.

Architectures and baselines We follow Madaan et al. (2022) for an unsupervised continual learning framework with contrastive self-supervised learning using SimSiam (Chen et al., 2020a). For masked image modeling, we follow the setting of SimMIM (Xie et al., 2022b) and MAE (He et al., 2022) using their official code repositories[1][2] where the masking ratio is 0.6 and 0.75, respectively. We use Vision Transformer (Dosovitskiy et al., 2020) (ViT-B) and Swin Transformer (Liu et al., 2021) (Swin-T) for backbone architectures. In ViT-B, the embedding dimension is 768, the layer depth is 12, the number of heads is 12, and the patch size is 16. In Swin-T, the embedding dimension is 96, the layer depth at each block is $[2, 2, 6, 2]$ (in total 12), the number of heads at each block is $[3, 6, 12, 24]$, the patch size is 4, and the sliding window size is 7. We set the input image size to 224 for all experiments but 192 for SimMIM pre-training. For continual learning methods, we use SI (Zenke et al., 2017), DER (Buzzega et al., 2020), and LUMP (Madaan et al., 2022). The implementation is built upon an official code of LUMP[3].

Training setups and hyperparameters We use AdamW optimizer (Loshchilov & Hutter, 2017) with cosine learning rate decay and the warmup for all experiments. For the pre-training phase at each task, we train the model 60 epochs on supervised learning and 100 epochs on unsupervised learning models as self-supervised learning methods without label supervision may require more iterations to converge. For fine-tuning, we basically perform 30 epochs training. For fine-tuning from the model pre-trained Imagenet 1K & 22K in Table 1, we set the number of training epochs to 10 as they rapidly converge within a few iterations. We set the hyperparameter for balancing the degree of regularization term $\lambda = 100$ for SI, $\lambda = 0.1$ for DER, and $\alpha = 0.1$ for LUMP. And the buffer size is 200 for rehearsal-based continual learning methods like DER and LUMP. We set the batch size to 64 for SimSiam pre-training, otherwise 128. Table 2 summarizes the learning rate and training epochs for experiments and we linearly scale the learning rate with $batch\_size/512$ in practice to reflect the input variance, followed by Goyal et al. (2017).

Table 2: **Basic configurations** for three continual learning frameworks during pre-training and fine-tuning, where $\eta = 2e^{-4}$. We report the best combinations of the learning rate for pre-training and fine-tuning in the range of $[0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0] \times \eta$ and $[0.1, 0.5, 1.0, 5.0] \times \eta$, respectively.

Method	Supervised	Contrastive	Masked Modeling
Pre-training	*lr*: 1.0 η, *epochs*: 60	*lr*: 0.2 η, *epochs*: 100	*lr*: 5.0 η, *epochs*: 100
Fine-tuning	*lr*: 0.5 η, *epochs*: 30	*lr*: 1.0 η, *epochs*: 30	*lr*: 5.0 η, *epochs*: 30

B Additional Analyses

B.1 Attention Distance and Entropy from Different Transformer Backbones

sup:subsec:analyses-vitswin We plot the attention distance and distribution of attention heads per layer for ViT-B in Figure 7 and Figure 8, respectively. And also, we plot the attention distance and entropy of the distribution of attention heads per layer for Swin-T in Figure 9 and Figure 10. Both self-attention-based architectures similarly behave according to the learning frameworks, i.e., *Supervised*, *SimSiam*, and *SimMIM*. Note that two consecutive layers in Swin-T repeat relatively high and low values for both metrics since two successive swin transformer blocks (*S-MSA* and *SW-MSA* in their original paper) aggregate locality in different ranges.

[1] https://github.com/microsoft/SimMIM
[2] https://github.com/facebookresearch/mae
[3] https://github.com/divyam3897/UCL

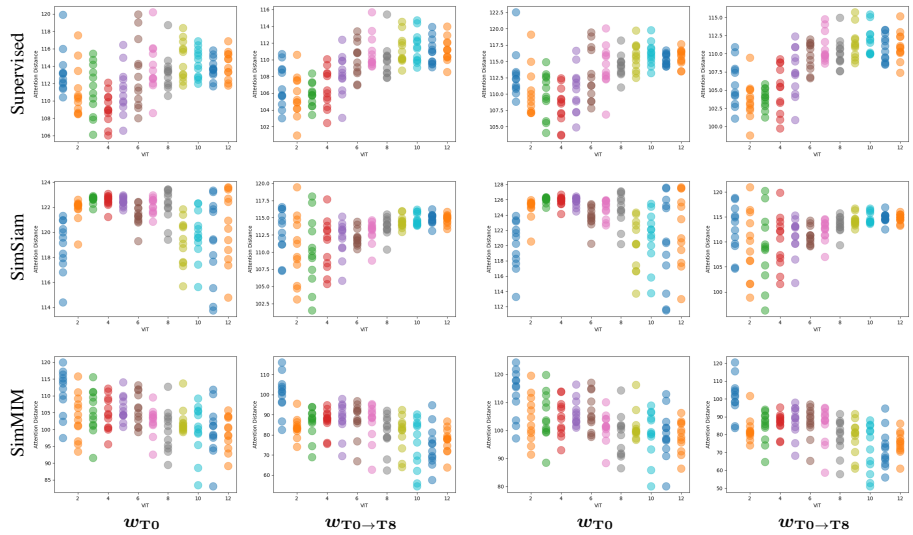


Figure 7: **ViT-B Attention distance** of the first (T0) and the last task (T8) in a task sequence with respect to three continual pre-trained frameworks right after the completion of the first ($\boldsymbol{w}_{\mathbf{T0}}$) and last task ($\boldsymbol{w}_{\mathbf{T0 \to T8}}$).

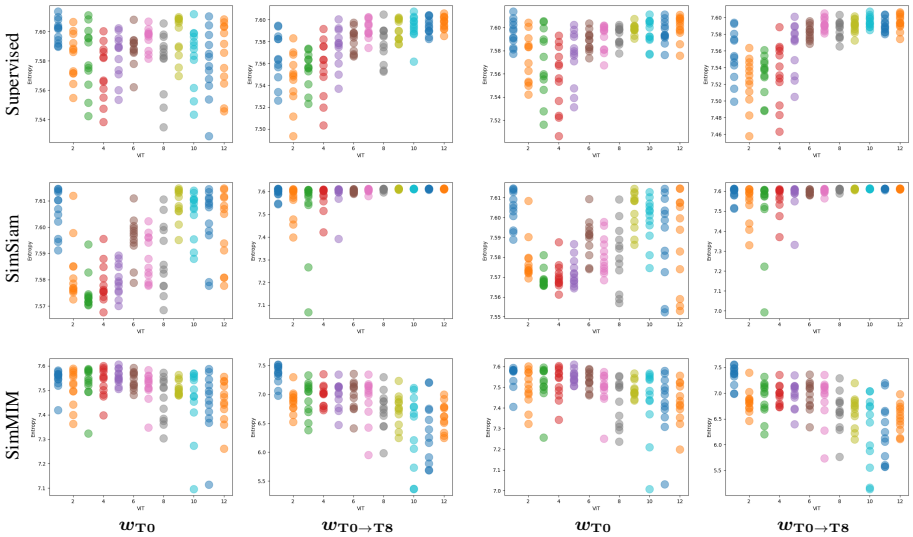


Figure 8: **ViT-B Attention entrtopy** of the first (T0) and the last task (T8) in a task sequence with respect to three continual pre-trained frameworks right after the completion of the first ($\boldsymbol{w}_{\mathbf{T0}}$) and last task ($\boldsymbol{w}_{\mathbf{T0 \to T8}}$).

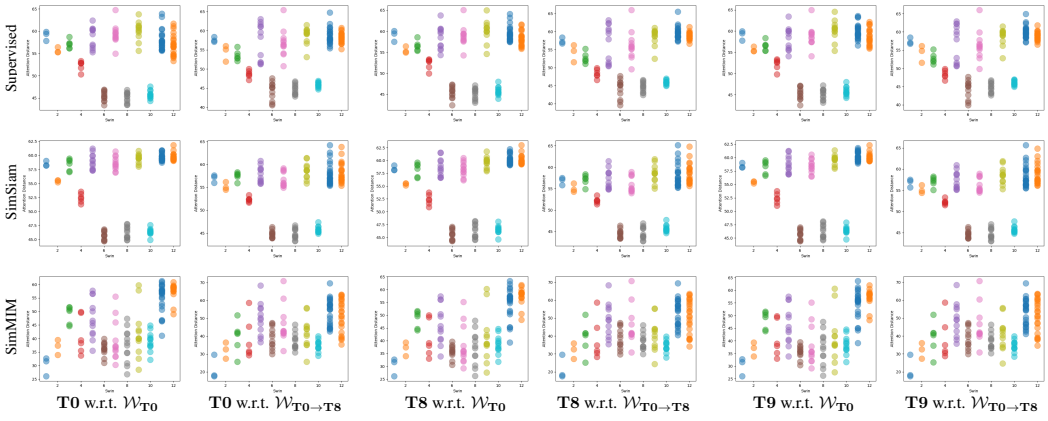


Figure 9: **Swin-T attention distance** of an in-distribution (T0) and ood task (T9) with respect to three continual pre-trained frameworks right after the completion of the first ($\boldsymbol{w}_{\mathbf{T0}}$) and last task ($\boldsymbol{w}_{\mathbf{T0 \to T8}}$).

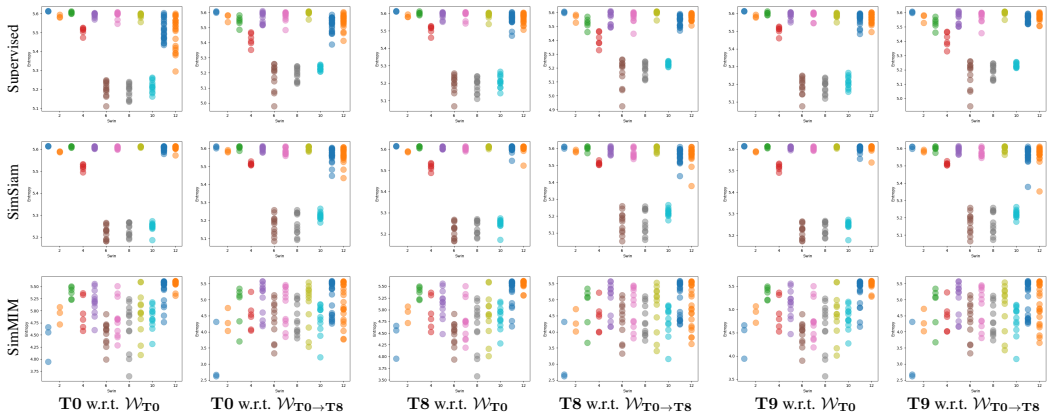


Figure 10: **Swin-T attention entropy** of an in-distribution (T0) and ood task (T9) with respect to three continual pre-trained frameworks right after the completion of the first ($\boldsymbol{w}_{\mathbf{T0}}$) and last task ($\boldsymbol{w}_{\mathbf{T0}\to\mathbf{T8}}$).

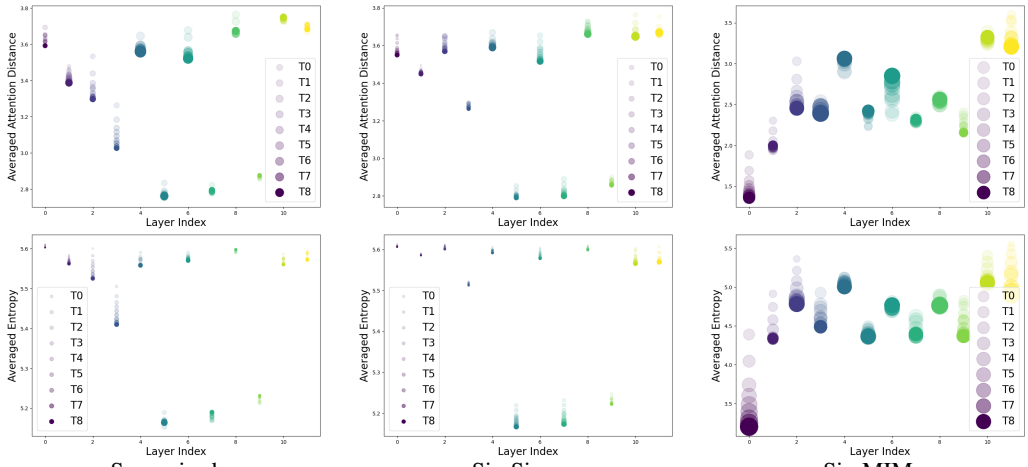


Figure 11: **Top row: Visualization of aggregated attention distance** on an OOD task (T9) at each layer at the end of each continual pre-training task phase (T0→T8). The radius of marker indicates the standard deviation over attention heads in the corresponding layer. **Bottom row: Aggregated attention entropy** on T9.

B.2 Change of Aggregated Attention Distance and Entropy during Continual Pre-training

sup:subsec:analyses-entr We additionally visualize the movement of aggregated entropy of the distribution from each attention head in Figure 11. We visualize the plot for attention distance in Figure 4 again for a better comparison between them. Similar to observations in attention distance visualization, aggregated attention entropy gradually decreases as proceeding to pre-train more tasks, encouraging incremental model generalization. And aggregated attention entropy for supervised and contrastive learning-based continual pre-training frameworks suffer from a small diversity with a high average amount of information for all attention heads, compared to SimMIM.

B.3 Aggregated Attention Distance and Entropy while Freezing Partial Layers

We further visualize the movement of aggregated attention distance and entropy when freezing the two lowest and deepest layers in Figure 12 and Figure 13, respectively. These experiments are exactly from Figure 6 Right. Interestingly, if SimMIM freezes a few layers during continual pre-training, the remaining trainable layers tend to decrease their attention distance and entropy more actively. We believe that this is a reason that the SimMIM does not find noticeable performance degeneration in fine-tuning, even freezing a few layers during continual pre-training. However, we didn't observe a

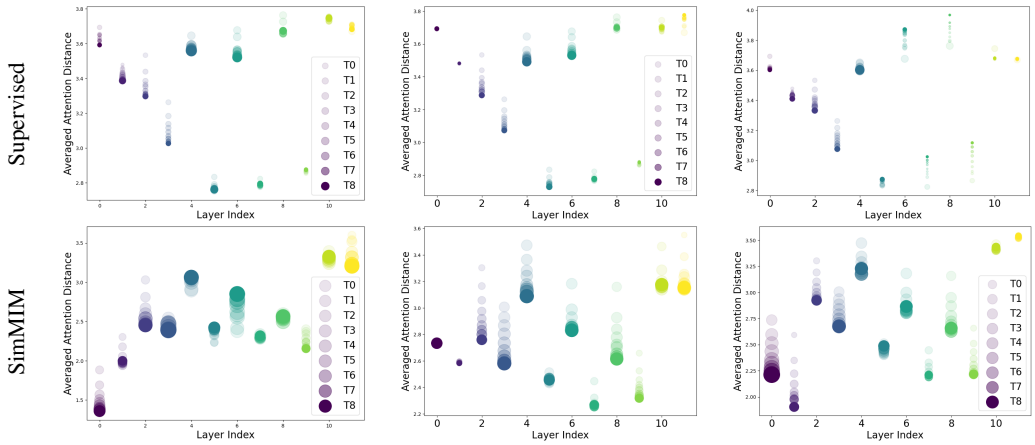


Figure 12: **Visualization of aggregated attention distance** on an OOD task (T9) at each layer at the end of each continual pre-training task phase (T0→T8). We freeze the two lowest or deepest layers after pre-training the first task (T0). The radius of the marker indicates the standard deviation over attention heads in the corresponding layer.

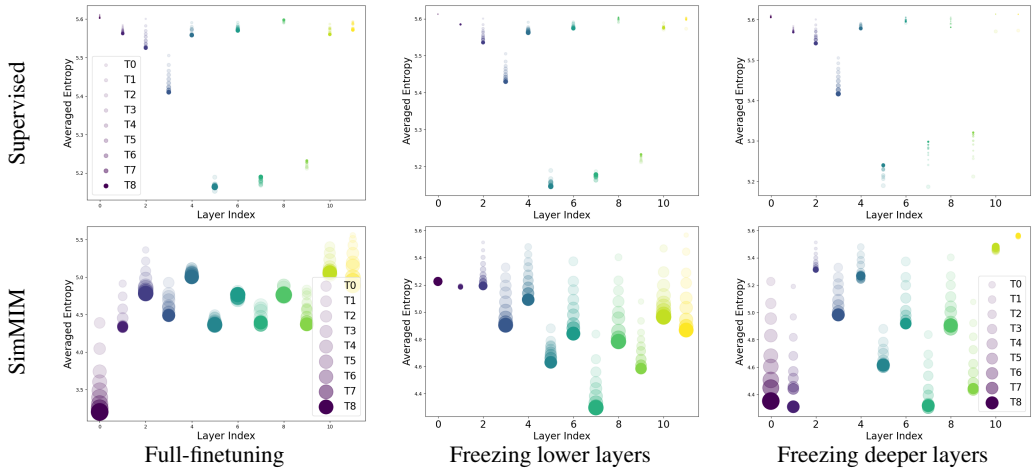


Figure 13: **Visualization of aggregated attention entropy** on an OOD task (T9) at each layer at the end of each continual pre-training task phase (T0→T8). We freeze the two lowest or deepest layers after pre-training the first task (T0). The radius of the marker indicates the standard deviation over attention heads in the corresponding layer.

significant change in supervised continual pre-training. This is because supervised learning is prone to rigidly focus on different features according to the layer depths (i.e., global to local), and therefore cannot flexibly cope with capturing locality inductive bias.