EVOLVE: Enhancing Unsupervised Continual Learning with Multiple Experts

Xiaofan Yu[1], Tajana Rosing[1], Yunhui Guo[2]
[1]University of California San Diego, [2]University of Texas at Dallas
{x1yu, tajana}@ucsd.edu, yunhui.guo@utdallas.edu

Abstract

Recent years have seen significant progress in unsupervised continual learning methods. Despite their success in controlled settings, their practicality in real-world contexts remains uncertain. In this paper, we first empirically investigate existing self-supervised continual learning methods. We show that even with a replay buffer, existing methods cannot preserve the critical knowledge on videos with temporal-correlated input. Our insight is that the primary challenge of unsupervised continual learning stems from the unpredictable input and the absence of supervision as well as prior knowledge. Drawing inspiration from hybrid AI, we introduce EVOLVE*, an innovative framework employing multiple pretrained models in the cloud, as experts, to bolster existing self-supervised learning methods on local clients.* EVOLVE *harnesses expert guidance through a novel expert aggregation loss, calculated and returned from the cloud. It also dynamically assigns weights to experts based on their confidence and tailored prior knowledge, thereby offering adaptive supervision for new streaming data. We extensively validate* EVOLVE *across several real-world data streams with temporal correlation. The results convincingly demonstrate that* EVOLVE *surpasses the best state-of-the-art unsupervised continual learning method by 6.1-53.7% in top-1 linear evaluation accuracy across various data streams, affirming the efficacy of diverse expert guidance. The codebase is at https://github.com/Orienfish/Evolve.*

1. Introduction

Unsupervised continual learning (UCL), continuously extracting information from unlabeled data streams, has emerged as a crucial area of investigation in the field of machine learning [2, 16, 26, 29, 61, 73, 74, 86, 88]. This approach holds significant value for real-world applications like self-driving vehicles [81] and robotics [52], where a mobile agent gathers ongoing data and an algorithm is consistently trained on the agent. These environments are notably dynamic, and obtaining real-time labels for samples proves exceedingly costly. Hence the ultimate goal of UCL is to train a learner capable of acquiring knowledge *solely* from raw sensory data while retaining crucial past concepts. The phenomenon of forgetting important concepts, termed *catastrophic forgetting* [33, 62], can lead to safety risks and potential catastrophic losses.

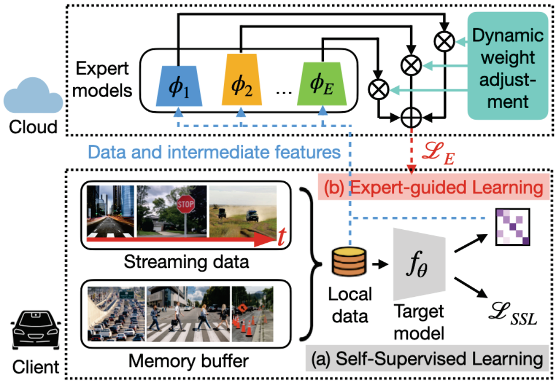

Figure 1. EVOLVE is a novel hybrid learning framework to tackle the UCL problem. EVOLVE enhances UCL with multiple experts sitting in the cloud with dynamic weight adjustment.

Despite significant efforts to achieve the goal, the feasibility of applying these methods in natural environments remains uncertain. The barrier primarily stems from two factors: (1) the *unpredictable* streaming input and (2) the lack of *supervision* and *prior knowledge*. State-of-the-art self-supervised learning methods [7, 20, 23, 35, 92] thrive on large volume of iid data for the optimal results, which does not align with the streaming setting where concepts might appear incrementally. Most of the existing UCL works, even in the absence of class labels, still depend on strong assumptions about the online data stream. For example, works like [2, 74, 88] run multiple passes over iid data, while others works take for granted the availability of class transition boundaries [29, 61]. In contrast, our approach deliberately refrains from such prior knowledge, as the streaming input can be unpredictable.

In order to boost the deployment of UCL methods in a natural environment, we propose to enhance UCL with diverse pre-trained models treated as *experts*. These large pretrained models inherently encompass extensive visual prior knowledge, which can be utilized to address the absence of such knowledge in the context of UCL. An example of an expert can be the Swin Transformer [58] pre-trained on ImageNet [79]. Leveraging the experts, the smaller on-device model with random initial parameters can be guided effectively based on the most recent streaming data. Nevertheless, one major challenge with the scheme is the substantial computational costs to execute the experts locally, which can lead to degraded continual learning performance under computational or delay budgets [32, 69].

In this paper, we propose EVOLVE, a hybrid framework using local and cloud computing for strengthening unsupervised continual learning with multiple experts. EVOLVE draws inspiration from the novel hybrid AI scheme [71] that has emerged recently to leverage the strong capabilities of large language models while considering the resource constraints on local devices. Hybrid AI aims to leverage the advancements in network technologies like 5G for splitting the computation between the device and cloud [38, 95]. As shown in Fig. 1, EVOLVE trains a target model locally, while transmitting a small set of data and intermediate features to the cloud, where multiple experts reside. Within the cloud, the experts conduct inference on the client's data. The output features are utilized to compute an expert aggregation loss that is returned to the client.

EVOLVE has two key designs: (1) the novel expert aggregation loss, distilling invaluable guidance from the expert ensemble, and (2) a dynamic weight adjustment strategy that fine-tunes the impact of each expert according to the latest data. EVOLVE employs multiple experts to provide comprehensive guidance to the continual learner, particularly in unpredictable scenarios, thereby boosting the performance over relying on a single expert model. Furthermore, motivated by *online optimization* [15], we update the weight assigned to each expert model dynamically during training to ensure the continual learner receives the most appropriate guidance.

For optimal guidance in the client's context, EVOLVE necessitates the local client to transmit select data and intermediate features to the cloud. Previous works have shown that data encryption techniques (e.g., Fully Homomorphic Encryption [31]) can effectively safeguard the client's privacy when sharing data for cloud computing services [4, 96]. EVOLVE lends itself seamlessly to data encryption integration, a prospect that we leave as future work.

In summary, the contributions of the paper are:

- We conduct an empirical study and demonstrate that current self-supervised learning methods experience a significant decrease in accuracy when applied in unpredictable and natural environments, thus impeding their practical utility for real-world applications.
- We propose a general expert-guided continual learning framework, called EVOLVE. To the best of the authors' knowledge, EVOLVE represents the *first* effort in leveraging diverse experts to enhance continual learning, managing resource constraints through a meticulously designed hybrid learning scheme. The proposed framework seamlessly integrates with established self-supervised learning methods.
- We extensively validate EVOLVE across various challenging continual learning benchmarks. Our findings highlight that EVOLVE fosters superior representation learning, elevating top-1 linear evaluation accuracy by 6.1-53.7% and kNN accuracy by 3.6-20.0% when compared to existing UCL approaches.

2. Related Work

Self-Supervised Learning (SSL). SSL has demonstrated superb performance in representation pre-training, which is reported to obtain even more robust representations than its supervised counterparts [29, 57]. Previous SSL methods can be categorized into generative models [48, 49, 63], progressive clustering [12, 13, 18, 37, 75, 87], contrastive learning [20–22, 41, 65] and information maximization [7, 28, 45, 47, 54, 92]. The recent contributions of Cha *et al.* [17] and Purushwalkam *et al.* [70] indicated the potential of SSL methods with a replay buffer to alleviate catastrophic forgetting on non-iid data streams. However, the question of how to effectively maximize the hidden power of SSL and achieve continual learning under *general real-world* scenarios remains open.

Unsupervised Continual Learning. In contrast to traditional continual learning problems, which assume prior knowledge of data independence, task labels, and class labels in the incoming data (e.g., [1, 3, 5, 11, 19, 24, 36, 46, 50, 51, 60, 67, 72, 76, 77, 80, 83–85, 91, 93, 94]), *Unsupervised Continual Learning* without task and class labels presents the most challenging scenario. The primary challenge is learning useful information solely from the data stream without supervision. Some approaches use variational autoencoders and generative replay to mitigate catastrophic forgetting (e.g., [2, 73, 74, 86, 88]). However, these methods raise concerns about scalability in large datasets and computational costs. STAM [82] offers an expandable memory architecture but requires dataset-specific tuning. LUMP [61] utilizes data augmentation and interpolation of new samples with buffered ones to address forgetting. Knowledge distillation-based methods, such as He *et al.* [40], CCSL [55], and CaSSLe [29], retain critical knowledge from past models but rely on task boundaries to capture model snapshots. Nonetheless, as our experiments demonstrate, state-of-the-art UCL methods struggle with forgetting in temporally correlated data.

To excel in UCL in dynamic environments, we design EVOLVE to leverage and manage the guidance from pre-trained models, which is orthogonal to all above works.

Online Optimization. Numerous studies have explored online optimization, focusing on minimizing regret when making sequential decisions [10, 15, 66]. One prominent approach in online optimization is the ”learning with multiple experts” algorithm [6, 14, 15]. In this method, a learner leverages advice from a set of K experts to choose actions in each round. The experts' weights, denoted as w_e, are updated dynamically using the Multiplicative Weights Update Method [6, 30, 56], with the formula $w_e = w_e(1 - \eta)$, where η is a hyperparameter. Alternative weight update strategies also exist [6, 14, 15]. The learner's worst-case regret is proven to be $O(\sqrt{T \log K})$, where T represents the number of rounds [14, 15]. Despite the shared objective of predicting sequential tasks, online optimization and continual learning have traditionally been studied as distinct domains. This paper illustrates how learning with multiple experts can effectively guide a continual learner in challenging environments.

3. Preliminary

Input Data Setups. Following on the trajectory of several recent research efforts on continual learning [61, 82], we consider a practical and general setting for UCL in the wild. Our problem setting follows state-of-the-art works except that we get rid of the unrealistic prior knowledge, such as (i) iid and multi-pass data as in [2, 74, 88], (ii) class labels as in [26, 53, 89] and (iii) task labels as in [29, 40, 55]. We assume that the data comes in a *class-incremental* manner, i.e., new classes emerge in a sequential manner. Such a setup reflects the changing environment over time. For example, as the autonomous vehicle moves from the city to the suburbs, new classes such as buildings, stop signs, pedestrians and trees are emerging sequentially. We assume there are T training steps. In each step t, we have access to a batch of samples $\{\mathbf{x}_{t,i}\}_{i=1}^{b}$. Each individual sample appears at most once in the training stream. We assume each sample is drawn from a sequence of D classes with each class corresponding to a unique distribution $\mathcal{P}^d$ in $\left\{\mathcal{P}^1, ..., \mathcal{P}^D\right\}$. We consider three typical single-pass streams with *no class label, no task label and unknown boundaries*:

(1) **iid**: we first consider an iid stream as a reference, in which each sample is drawn independently and identically from the entire dataset.

(2) **Seq**: Sequential class-incremental stream where the classes are introduced one-by-one and are balanced (i.e., the number of samples from different classes are identical),

(3) **Seq-imb**: Imbalanced sequential class-incremental stream which introduces a largely varied number of samples from each class incrementally.

Learning Protocol and Evaluation Metrics. The goal of the UCL problem is to train a base model $f_\theta : \mathcal{X} \rightarrow \mathcal{H}$ which is a mapping from the input space $\mathcal{X}$ to a low-dimensional feature space $\mathcal{H}$. θ are the learnable parameters. For evaluation, we construct a separate testing dataset $\mathcal{E} = \{(\mathbf{x}_j, y_j)\}$ by randomly sampling an equal amount of labeled samples from all classes in $\left\{\mathcal{P}^1, ..., \mathcal{P}^D\right\}$. We also create another validation dataset $\mathcal{V} = \{(\mathbf{x}_q, y_q)\}$ for hyperparameter selection. It is important to note that, irrespective of whether a class has appeared in the training sequence or not, it is always included in both $\mathcal{E}$ for ground-truth evaluation. During testing, given a snapshot of model θ_t at time t, we first compute the learned latent representations $\mathbf{h}_j = f_{\theta_t}(\mathbf{x}_j)$ for each testing sample $(\mathbf{x}_j, y_j) \in \mathcal{E}$. Following previous protocols [29, 61, 74, 82], we train the kNN and linear classifiers respectively to evaluate the quality of learned representations. We use kNN and top-1 linear classification accuracies as metrics.

4. An Empirical Study of Existing SSL

Recent studies have indicated that combining SSL with memory replay holds great promise for continual representation learning in the wild [17, 61, 70]. However, it is unclear whether SSL is sufficiently *practical* in unsupervised continual learning scenarios, especially with *dynamic environment* and *temporal-correlated streams*.

In this section, we investigate this question by conducting an empirical study of existing SSL methods with memory replay on image- and video-based continuous data streams. Using the setups described in Section 3, we construct the **iid** and **Seq-imb** data streams from CIFAR-10 (image-based, 10 classes) [64] and Stream-51 (video-based, 51 classes) [78]. Notably, the Seq-imb data stream from Stream-51 follows the original time stamps and imbalanced class appearances, creating a realistic temporal-correlated stream. For the Seq-imb stream from CIFAR-10, the data in each class is sampled randomly. We employ a memory buffer $\mathcal{M}$ to store and replay a finite number of historical samples. Following [61], we update the buffer after each incoming batch $\{\mathbf{x}_{t,i}\}_{i=1}^{b}$ with reservoir sampling.

Setups. During training, we randomly sample a finite subset $\{\mathbf{x}_{\mathcal{M},i}\}_{i=1}^{b_\mathcal{M}}$ to concatenate the new incoming samples. We assume the size of the memory buffer is $m = 256$, the finite number of samples for replay is $b_\mathcal{M} = 128$, the streaming batch size is $b = 128$. At each time stamp t, we feed the stacked input $\mathbf{x}_t = [\{\mathbf{x}_{t,i}\}_{i=1}^{b}, \{\mathbf{x}_{\mathcal{M},i}\}_{i=1}^{b_\mathcal{M}}]$ into two ways of random augmentations $t^A \sim \mathcal{T}^A, t^B \sim \mathcal{T}^B$ and obtain two augmented views: $\mathbf{x}_t^A = t^A(\mathbf{x}_t), \mathbf{x}_t^B = t^B(\mathbf{x}_t)$. We input the augmented view into the base encoder f_θ and obtain $\mathbf{h}_t^A = f_\theta(\mathbf{x}_t^A)$. In all SSL methods, a small projection head f_h is used to map the representations to the embeddings $\mathbf{z}_t^A = f_h(\mathbf{h}_t^A)$. Similarly, $\mathbf{z}_t^B$ is produced from another augmented view. More details about the setups and baseline implementations are reported in the Appendix.

We next detail the state-of-the-art SSL methods from various categories, as shown in Tab. 1, which all employ

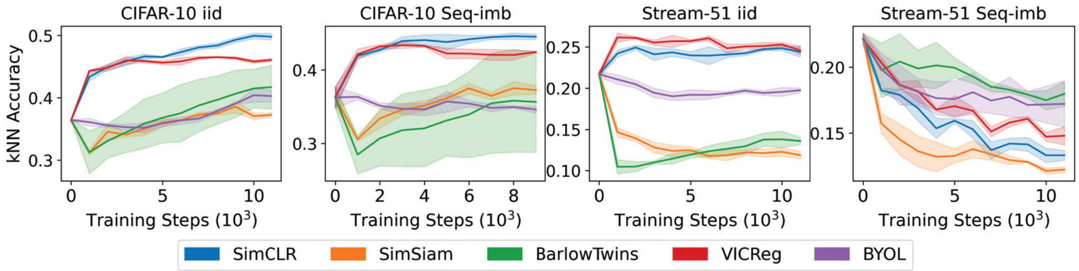

Figure 2. The kNN accuracy during training on CIFAR-10 and Stream-51 under the **iid** and **Seq-imb** streaming data, using various SSL baselines with a replay buffer. The accuracy at step 0 is evaluated on the randomly initialized model. Shaded areas show the standard deviation of measurements after three random trials.

a variant of the Siamese networks [9].

•**SimCLR** [20] uses the InfoNCE loss [65] to maximize the similarity between augmented embeddings $\mathbf{z}_t^A, \mathbf{z}_t^B$ (*positives*) while contrasting against the other embeddings in the same batch (*negatives*). In Tab. 1, we consider the cosine similarity and τ is a hyperparameter called temperature.

•**BYOL** [35] maintains a momentum encoder from the past and **SimSiam** [23] applied stop gradient, both operating on one branch of the Siamese structure, to avoid degenerated solutions. An additional predictor head f_p is used to generate another view $\mathbf{q}_t^A = f_p(\mathbf{z}_t^A)$, and similar for $\mathbf{q}_t^B$. BYOL and SimSiam employ MSE-based losses between the two views. SimSiam uses the negative cosine similarity $\mathcal{D}(\mathbf{q}_t^A, \mathbf{z}_t^B) = -\frac{\mathbf{q}_t^A}{\left\|\mathbf{q}_t^A\right\|_2} \cdot \frac{\mathbf{z}_t^B}{\left\|\mathbf{z}_t^B\right\|_2}$ which is equivalent to the MSE of ℓ_2-normed vectors. Tab. 1 only shows a simplified form of the loss while the complete symmetric loss can be constructed with an additional term obtained by swapping two views.

•**BarlowTwins** [92] and **VICReg** [7] learn self-supervised with losses inspired by information theory. BarlowTwins computes the cross-correlation matrix of the two-view embeddings: $\mathcal{C}_{uv} = \frac{\sum_j \mathbf{z}_{j,u}^A \mathbf{z}_{j,v}^B}{\sqrt{\sum_j \left(\mathbf{z}_{j,u}^A\right)^2} \sqrt{\left(\sum_j \mathbf{z}_{j,v}^B\right)^2}}$, and builds the loss function as shown in Tab. 1, where the first term enhances the invariance between augmented pairs and the second term decorrelates non-identical samples. VICReg employs a combination of three losses: invariance $s(\cdot, \cdot)$ (MSE between two views), variance $v(\cdot)$, and co-variance $c(\cdot)$ (off-diagonal coefficients of $\mathcal{C}$), weighted by appropriate hyperparameters. For simplicity, we only show one-view of $v(\cdot)$ and $c(\cdot)$ in Tab. 1.

Results and Discussions. Fig. 2 shows the kNN accuracy during training on three random trials of the CIFAR-10 and Stream-51 streams. Our empirical results on CIFAR-10 share similar patterns with the latest studies that report the continual learning ability of SSL with memory replay on imaged-based datasets [17, 61, 70]. Among all SSL methods, SimCLR and VICReg (with replay buffer) show better results on CIFAR-10 with progressively increased accuracies in terms of both iid and Seq-imb streams. In contrast, SimSiam and BarlowTwins learn less effectively while the performance of BarlowTwins is unstable.

Table 1. Overview of state-of-the-art SSL methods and losses.

Methods	**Loss**	**Loss Function** $\mathcal{L}_{SSL}$
SimCLR [20]	InfoNCE	$-\log \frac{\exp(\mathbf{z}_i^A \cdot \mathbf{z}_j^B / \tau)}{\sum_{k \neq i} \exp(\mathbf{z}_i^A \cdot \mathbf{z}_k / \tau)}$
BYOL [35]	MSE	$\left\|\mathbf{q}_t^A - \mathbf{z}_t^B\right\|_2^2$
SimSiam [23]	MSE	$\mathcal{D}(\mathbf{q}_t^A, \mathbf{z}_t^B)$
Barlow Twins [92]	Cross-Correlation	$\sum_u (1 - \mathcal{C}_{uu})^2 +$ $\psi \sum_u \sum_{v \neq u} \mathcal{C}_{uv}^2$
VICReg [7]	MSE + Variance + Cross-Correlation	$\psi s(\mathbf{z}_t^A, \mathbf{z}_t^B) +$ $\mu v(\mathbf{z}_t^A) + \nu c(\mathbf{z}_t^A)$

However, the performances are significantly different when deploying those SSL methods on Stream-51 data with a temporal order. While VICReg and SimCLR show positive accuracy gains with iid streams, *all* SSL baselines struggle to learn and retain knowledge in sequential and imbalanced streams, consistently displaying declining kNN accuracies during training. Remarkably, even the top two SSL methods on Seq-imb Stream-51, BarlowTwins and BYOL, yield notably inferior kNN accuracies compared to a randomly initialized network (18.0% and 17.2% versus 21.8%). Our findings demonstrate the limited ability of current SSL methods, even with replay buffers, to continually learn from practical data streams within real-world contexts. We hypothesize that this limitation is primarily due to the incapability of memorizing critical knowledge purely from the unsupervised streams, which motivates the design of EVOLVE.

5. EVOLVE

Inspired by online optimization, we introduce the EVOLVE framework which enhances unsupervised continual learning methods through multiple experts. Our central idea involves leveraging the extensive knowledge embodied by pre-trained large expert models, readily available online. By distilling expert insights with self-supervised learning techniques, the unsupervised continual learner quickly adapts to new environments while circumventing catastrophic forgetting.

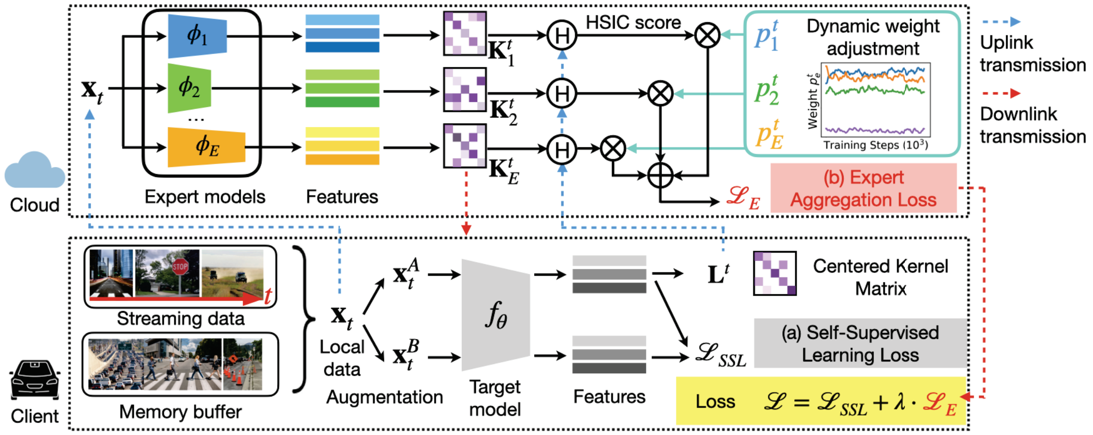

Figure 3. The workflow of EVOLVE that has two components: self-supervised learning and expert-guided learning, with two loss terms. EVOLVE dynamically updates the weight associated to each expert.

5.1. Overview

The general workflow of EVOLVE is depicted in Fig. 3. Given the constrained resources of the local client, executing extensive expert models becomes a challenge. To address this, we adopt a hybrid AI [71] approach, confining SSL training to the local client while storing and executing all expert models on the cloud. The local client share a small subset of data and intermediate information with the cloud (blue dotted lines in Fig. 3), while the computation of expert aggregation loss takes place in the cloud, with the resulting insights transmitted back to the client for learning (red dotted line in Fig. 3). Essentially, the training process of EVOLVE uses a sum of two loss terms:

$$\mathcal{L} = \mathcal{L}_{SSL} + \lambda \cdot \mathcal{L}_E, \tag{1}$$

where $\mathcal{L}_{SSL}$ is the SSL objective as summarized in Tab. 1, $\mathcal{L}_E$ is our novel expert aggregation loss, and λ is a hyperparameter to balance the two.

In the following section, we provide a detailed explanation of the primary design of EVOLVE: (1) the computation of $\mathcal{L}_E$ using diverse experts, and (2) the dynamic adjustment of expert weights.

5.2. Expert Aggregation Loss

Diverse Experts. Transferring knowledge from a universal visual model has become the dominant paradigm in computer vision [20, 41, 43]. We introduce a novel approach to guide the continual learner: harnessing a diverse library of pre-trained models as experts, contrasting the reliance on a single model in traditional transfer learning. Our motivation is rooted in the fact that the transferability of pre-trained models is heavily dependent on the target task [8,90]. When confronted with new visual inputs, we anticipate that at least one of the experts will be able to accurately capture the underlying semantics of the image.

Experts Aggregation. Given a library of E expert models, we aim to define a new loss term for aggregating the knowledge across multiple experts unsupervisedly. Due to the absence of labels, knowledge distillation-based methods [44] cannot be used. Instead, we propose to use Hilbert-Schmidt independence criterion (HSIC) for assessing the models' ability in differentiating two augmentations of the same image. HSIC was proposed in [34] as a measure of dependence between two random variables X and Y. Assume that $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ are drawn from the joint distribution (X, Y). $\mathbf{K}$ and $\mathbf{L}$ are the centered kernel matrices computed on $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ respectively. The empirical HSIC can be computed as $\mathrm{HSIC}(X, Y) = \mathrm{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2}\,\mathrm{Tr}(\mathbf{KHLH})$, where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix.

In our setup, using a subset of shared samples $\mathbf{x}_t$, each expert model ϕ_e calculates representations $\phi_e(\mathbf{x}_t)$ via a forward pass. These representations contribute to the computation of the kernel matrix $\mathbf{K}_e^t$. Similarly, the local continual learner generates the kernel matrix $\mathbf{L}^t$ for sharing with the cloud. Denoted as $\mathrm{HSIC}(\mathbf{K}_e^t, \mathbf{L}^t)$, the HSIC score between the expert model ϕ_e and the continual learner signifies their similarity. Intuitively, the continual learner is trained to emulate the higher-order representation similarities observed in the expert models. To account multiple experts, we introduce an expert aggregation loss as shown in Fig. 3:

$$\mathcal{L}_E = -\sum_{e=1}^{E} p_e^t \cdot \mathrm{HSIC}(\mathbf{K}_e^t, \mathbf{L}^t), \tag{2}$$

where p_e^t is the weight of the expert e which is learned in the next section. $\mathcal{L}_E$ and $\mathbf{K}_e^t$ are then transmitted back to the client to compute the final loss and gradients. We analyze the communication overhead of EVOLVE in Section 6.5.

5.3. Dynamic Weight Adjustment

The effectiveness of expert models naturally fluctuates over time in response to varying input data. For instance, as an autonomous vehicle transitions from urban to suburban areas, the most effective expert model shifts from those trained on street views to those specialized in natural landscapes. Consequently, the weight p_e^t in Eq.(2) requires dynamic updating for each expert. However, achieving this within the context of UCL poses challenges due to the absence of labels. To realize this goal, our design necessitates two components: (i) a metric to assess the quality of each expert, and (ii) a dynamic online weight adjustment algorithm based on this metric.

Confidence Metric. Drawing inspiration from Contrastive Predictive Coding [65], we introduce a confidence metric based on an expert model's capability to predict one image view from another. For a given expert e with model ϕ_e, this involves assessing the confidence of expert e via augmented representations $\phi_e(\mathbf{x}_t^A)$ and $\phi_e(\mathbf{x}_t^B)$, referred to as $\mathbf{h}_e^A$ and $\mathbf{h}_e^B$ respectively:

$$q_e^t = \sum_i \frac{\exp(\mathbf{h}_{e,i}^A \cdot \mathbf{h}_{e,i}^B/\tau)}{\sum_{k\neq i}\exp(\mathbf{h}_{e,i}^A \cdot \mathbf{h}_{e,k}^A/\tau) + \sum_k \exp(\mathbf{h}_{e,i}^A \cdot \mathbf{h}_{e,k}^B/\tau)}, \tag{3}$$

where τ is a temperature hyperparameter. Alternatively, the confidence metric can be viewed as the categorical cross-entropy for correctly predicting the augmented view among all representations, utilizing softmax. This metric offers two distinct benefits: Firstly, q_e^t quantitatively gauges expert e's proficiency without relying on labels, always within the bounds of $[0, 1]$. Secondly, the computation of HSIC can be reused to compute cosine similarities among representations, resulting in significant time savings.

Online Adjustment. The online weight adjustment algorithm should account the historical and latest confidence. We maintain a weight w_e^t to each expert e at step t, while initializing $w_e^0 = 1$ for all e. Instead of using the multiplicative update rule [6, 30, 56] for online optimization, we propose updating weights through moving average:

$$w_e^{t+1} = \alpha w_e^t + (1-\alpha) q_e^t. \tag{4}$$

where α is a hyperprameter. We adopt this approach because the multiplicative update rule might allocate an excessively small weight to an initially underperforming expert. Consequently, even if the expert improves later, it struggles to regain a substantial weight. A similar issue is noted in [39]. Conversely, our proposed update method prioritizes the current model's confidence.t α can be fine-tuned to strike a balance between past and present confidence. Throughout training, we directly normalize w_e^t across all experts, i.e., employing $p_e^t = w_e^t / \sum_l w_l^t$ in Eq. (2).

The weight adjustment scheme we propose connects and contrasts with conventional online optimization, with detailed discussion provided in the Appendix.

6. Experiments

6.1. Experimental Setup

Our setups are listed as follows with more implementation details in the Appendix.

Datasets. We conduct comprehensive experiments on four visual datasets: **CIFAR-10** (10 classes) [64], **TinyImageNet** (100 classes) [27], **CORe50** (50 classes) [59] and **Stream-51** (51 classes) [78]. To align with real-world scenarios, we mainly consider the **Seq, Seq-imb** streams in each dataset. CIFAR-10 and TinyImageNet are image-based datasets, where we randomly sample a stream from the image pools. CORe50 and Stream-51 are video-based continual learning datasets collected from streaming scenarios, with CORe50 focusing on hand-held object detection and Stream-51 hosting a variety of classes from animals to vehicles. In Seq and Seq-imb, we preserve the temporal order within the video datasets.

Implementation details of EVOLVE. We use ResNet-18 with a feature space dimension of 512 as the model. Similar to [61], we employ the SGD optimizer with a learning rate of 0.03. The batch sizes are the same as Section 4. Based on validation results, we set $\alpha = 0.95$, $\tau = 0.1$ and set λ to ensure comparable $\mathcal{L}_{SSL}$ and $\mathcal{L}_E$, i.e., $\lambda = \frac{|\mathcal{L}_{SSL}|}{|\mathcal{L}_E|}$. For evaluation, the kNN classifier is trained with a separate subset of testing samples with $k = 50$. The linear classifier is trained offline for 50 epochs. All measurements are averaged after three random trials.

For the experts, we employ 4 large models pretrained on ImageNet [79] with many offline epochs. Hence the expert models capture rich and diverse knowledge in the computer vision domain. Specifically, we use (1) Supervisedly pretrained ResNet-50 [42], (2) Supervisedly pretrained base Swin Transformer [58], (3) Unsupervisedly pretrained ResNet-50 with MoCo [41], and (4) Unsupervisedly pretrained ResNet-50 with MoCo v2 [22]. All models can be easily downloaded from the torchvision package [68] or the original code release. All expert models are frozen.

Baselines. Apart from the SSL methods in Section 4, we also adapt state-of-the-art continual learning baselines which can be added on top of the SSL framework: the regularization-based **SI** [93], the architecture-based **PNN** [80], the replay-based **DER** [11]. All above baselines are originally proposed for supervised continual learning, but are adapted to the unsupervised settings following [61]. **CaSSLe** [29] and **LUMP** [61] are the UCL methods based on SSL backbones. To ensure a fair comparison, CaSSLe does not use task labels and stores the models from the previous batch.

6.2. Comparison with Existing UCL Baselines

Final Accuracy. We first compare the final accuracy results of all methods in combination with various SSL back-

Table 2. Comparison of EVOLVE and unsupervised continual learning baselines on the **Seq-imb TinyImageNet** streams. Bold and underlined values show the best and second best results based on each SSL.

Method	kNN Accuracy(↑)					Linear Evaluation Accuracy(↑)				
	SimCLR	**BYOL**	**SimSiam**	**BarlowTwins**	**VICReg**	**SimCLR**	**BYOL**	**SimSiam**	**BarlowTwins**	**VICReg**
SSL	15.2±0.1	10.8±0.7	10.7±0.7	7.5±0.6	13.9±0.3	11.6±0.9	7.4±0.9	5.2±1.2	3.0±0.7	15.9±0.3
SI	13.7±0.2	10.0±0.5	9.6±0.1	7.6±0.3	12.8±0.5	10.7±0.4	7.7±0.1	4.3±0.5	2.8±0.8	11.9±0.7
PNN	13.7±0.1	9.9±0.3	9.8±0.3	7.0±0.5	12.5±0.3	10.8±0.5	6.4±0.4	4.5±0.4	2.8±0.1	12.1±0.3
DER	13.1±0.2	9.5±0.6	9.1±0.4	7.6±0.5	12.6±0.2	13.0±0.3	7.0±1.3	6.3±0.3	2.9±0.5	12.4±0.2
CaSSLe	8.5±0.8	10.2±0.4	10.1±0.4	7.4±0.1	14.4±0.2	3.2±0.9	6.7±2.0	4.6±0.6	2.1±0.1	15.8±0.2
LUMP	12.2±0.3	7.9±0.4	8.5±0.1	6.8±0.5	14.5±0.4	9.0±1.5	2.5±0.2	3.2±0.6	2.1±0.4	14.7±0.3
EVOLVE	**18.8±0.2**	**18.5±0.1**	**18.4±0.6**	**12.3±2.3**	**19.4±0.5**	**19.5±1.6**	**16.3±2.8**	**19.2±0.9**	**10.5±3.9**	**24.0±0.5**

Table 3. Comparison of EVOLVE and unsupervised continual learning baselines on the **Seq-imb CoRe50** streams. Bold and underlined values show the best and second best results based on each SSL.

Method	kNN Accuracy(↑)					Linear Evaluation Accuracy(↑)				
	SimCLR	**BYOL**	**SimSiam**	**BarlowTwins**	**VICReg**	**SimCLR**	**BYOL**	**SimSiam**	**BarlowTwins**	**VICReg**
SSL	24.7±0.8	26.8±1.2	20.6±1.3	30.9±1.2	25.9±1.0	13.6±1.0	15.4±6.1	8.2±3.3	12.8±10.3	20.8±0.9
SI	19.4±1.9	25.0±0.6	19.5±0.1	23.0±3.1	16.3±1.4	8.6±3.3	11.2±2.4	5.4±0.2	16.9±2.7	11.4±1.2
PNN	17.8±1.2	26.5±0.5	19.8±1.4	20.1±2.6	15.9±1.0	9.2±3.2	16.8±3.6	8.7±3.3	12.1±4.1	11.1±1.3
DER	23.0±0.8	26.6±0.3	25.2±1.0	20.0±2.2	16.9±3.3	16.6±1.1	18.3±0.6	11.7±3.4	12.5±2.7	11.4±2.3
CaSSLe	14.7±0.9	26.5±1.6	21.9±0.8	16.7±0.3	26.7±0.4	4.1±1.0	13.8±5.4	8.5±4.0	4.6±0.1	21.7±0.1
LUMP	27.1±0.3	23.2±0.6	16.0±0.6	17.6±1.1	30.2±0.4	16.3±2.3	9.8±1.2	4.9±1.0	7.5±1.1	27.6±0.6
EVOLVE	**44.2±1.4**	**44.2±1.8**	**45.2±0.5**	**41.6±3.9**	**40.9±1.9**	**51.3±3.1**	**54.3±1.4**	**50.8±4.0**	**45.4±7.4**	**42.9±2.1**

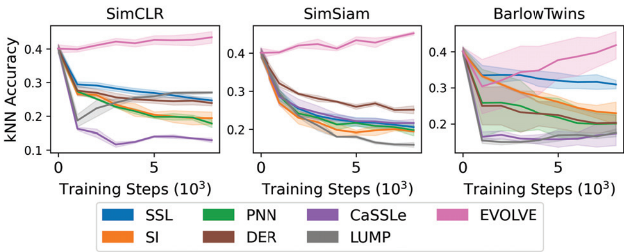

Figure 4. The kNN accuracies during training on **Seq-imb CORe50** data streams using EVOLVE and other baselines.

bones. Tab. 2, 3 and 4 present the results on Seq-imb streams from TinyImageNet, CORe50 and Stream51 respectively. EVOLVE significantly improves the learning performance on all settings including both image- and video-based datasets, when accompanied with various SSL bases. Specifically, **EVOLVE outperforms the top baseline using the same SSL by 3.6-20.0% in kNN accuracy and 6.1-53.7% in top-1 linear evaluation accuracy across diverse data streams.** This demonstrates that EVOLVE, by distilling the guidance from experts, can significantly enhance the unsupervised continual learning capability of existing SSL methods, even on the challenging video-based streams with temporal correlations.

Accuracy during Training. To better visualize the continual learning dynamics, we plot the the evolution of kNN accuracies on CORe50 Seq-imb streams in Fig. 4 using SimCLR, SimSiam and BarlowTwins. In the UCL setting, the accuracy curves of existing continual learning baselines can deteriorate faster than the SSL method with a replay buffer, indicating catastrophic forgetting. EVOLVE outperforms the other methods, being able to preserve critical knowledge and continually improve throughout the process.

6.3. Comparison with Other Weight Update Policies for Using the Experts

In Section 6.2, we contrast EVOLVE with baselines not employing pretrained experts. In this section, we assess our dynamic weight update design when utilizing the same set of experts. We compare with three other commonly used policies: (1) **EW**: equally assigning weights for all experts throughout training, namely $w_e^t = 1, \forall e$. (2) **Single**, assigning $p_e^t = 1$ to the best expert e and 0's to the rest at each timestamp t. (3) **MW**: using the Multiplicative Weight Update Method [6, 30, 56] and updating with $w_e^{t+1} = w_e^t(1 + \eta q_e^t)$. We show the final kNN accuracies when using various policies on Seq streams in Tab. 5. EVOLVE with the moving-average weight update algorithm surpasses the other policies on TinyImageNet and Stream-51. On CORe50, the Swin Transformer prevails among other experts, leading to the best outcome when employing a single expert, with EVOLVE trailing by only 0.5%. In summary, our moving-average weight update emerges as the most robust design within the dynamic UCL context, effectively leveraging multiple experts.

For deeper insights into the intricate weight adjustment dynamics, we visualize the normalized weight p_e^t during MW and EVOLVE training in Fig. 5. The left plot demonstrates that MW establishes weights based on historical confidence accumulation, which can converge to extremes and struggle to adapt to new samples. Conversely, EVOLVE monitors the latest confidence while retaining a balanced influence from the past, yielding a dynamic weight pattern as depicted in Fig. 5 (right).

Table 4. Comparison of EVOLVE and unsupervised continual learning baselines on the **Seq-imb Stream-51** streams. Bold and underlined values show the best and second-best results based on each SSL.

Method	kNN Accuracy(↑)					Linear Evaluation Accuracy(↑)				
	SimCLR	**BYOL**	**SimSiam**	**BarlowTwins**	**VICReg**	**SimCLR**	**BYOL**	**SimSiam**	**BarlowTwins**	**VICReg**
SSL	14.0±0.4	17.1±0.1	13.1±1.0	18.1±0.4	14.8±0.3	33.1±1.5	27.7±2.6	11.9±3.5	51.4±1.1	40.8±0.2
SI	12.9±0.5	17.0±1.0	12.3±0.6	12.2±0.8	11.1±0.4	21.3±1.9	27.0±8.3	13.2±5.9	26.2±0.2	23.5±2.0
PNN	12.0±0.2	17.1±1.1	12.5±0.5	12.7±1.2	11.6±0.8	13.5±0.6	29.9±0.1	9.8±0.9	26.9±4.6	25.4±0.1
DER	13.6±0.6	16.0±0.4	14.4±1.3	13.0±1.2	10.7±0.4	31.5±1.5	37.5±2.4	28.0±5.0	28.9±0.1	24.0±1.7
CaSSLe	14.7±0.9	26.5±1.6	21.9±0.8	16.7±0.3	12.6±2.0	7.5±3.2	27.3±5.0	20.6±6.1	10.0±1.2	38.5±2.4
LUMP	20.5±0.7	14.5±0.5	12.7±0.1	13.9±0.5	20.7±1.3	48.2±0.1	27.2±0.8	8.4±0.1	16.9±3.4	55.1±1.5
EVOLVE	**30.1±1.6**	**31.6±1.3**	**31.5±1.7**	**30.1±1.7**	**24.8±0.4**	**82.2±0.9**	**84.4±1.0**	**81.7±1.0**	**75.7±2.4**	**61.2±1.7**

Table 5. The kNN accuracies when using various online weight update policies in EVOLVE, tested on BYOL and **Seq** streams.

Dataset	EW	Single	MW	EVOLVE
TinyImageNet	19.7±0.2	18.3±0.6	18.0±0.6	**20.2±0.5**
CORe50	41.5±1.8	**44.7±0.6**	44.0±0.3	44.2±1.3
Stream-51	21.5±0.3	22.5±0.3	22.2±0.6	**30.2±0.8**

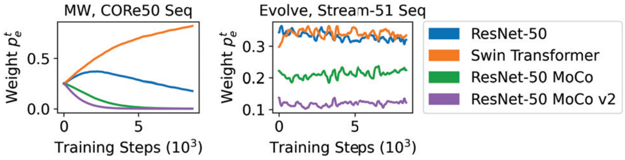

Figure 5. The expert weights p_e^t during training under MW and EVOLVE on **Seq** streams with BYOL.

6.4. Hyperparameters

Number of Experts. We experiment using 1,2,3,4 experts from our model candidates on the Seq-imb Stream-51 streams with BYOL. As shown in Fig. 6 (left), augmenting the number of experts yields consistent accuracy enhancement, highlighting the efficacy of incorporating diverse experts for adaptive responses to dynamic environments. In the Appendix, we show that by using a longer training stream (such as multiple epochs), the performance of EVOLVE can be comparable to that of the best expert, or even surpass it, despite having a much smaller model size.

Hyperparameters λ and α. λ is the hyperparameter to balance $\mathcal{L}_{SSL}$ and $\mathcal{L}_E$. We experiment $\lambda = \psi\frac{|\mathcal{L}_{SSL}|}{|\mathcal{L}_E|}$ with $\psi \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$ on Stream-51 Seq-imb streams using BYOL backbone. Fig. 6 (middle) reveals that $\psi = 1.0$ yields optimal results, indicating a need for balanced weighting between $\mathcal{L}_{SSL}$ and $\mathcal{L}_E$. Another pivotal hyperparameter is α, employed in dynamic weight updates. Fig. 6 (right) showcases results using $\alpha \in \{0.9, 0.93, 0.95, 0.97, 0.99\}$ on Stream-51 Seq-imb with BYOL. A smaller α elevates the importance of recent expert confidence, while a larger α accentuates past confidence. Notably, $\alpha = 0.95$ delivers peak performance, aligning with our validation outcomes.

6.5. Communication Overhead

EVOLVE's hybrid scheme allows for effective utilization of large expert models while necessitating the transmission of a subset of $\mathbf{x}_t$ and its corresponding kernel matrix $\mathbf{L}^t$. We analyze the communication cost of this approach, considering a scenario where $\mathbf{x}_t$ consists of 128 RGB images of size 32×32 from CIFAR-10, and $\mathbf{L}^t$ matrix of size 128×128, using 4 experts. The downlink transmission from the cloud to the local client includes $\mathcal{L}_E$ and the experts' kernel matrices $\mathbf{K}_e^t$. Consequently, a single round of transmission involves around 720kB of uncompressed data. This data exchange takes less than 1 ms using 5G technology at an average speed of 10Gbps [25], making it almost negligible. Communication efficiency can be further enhanced through data compression, a prospect for our future work.

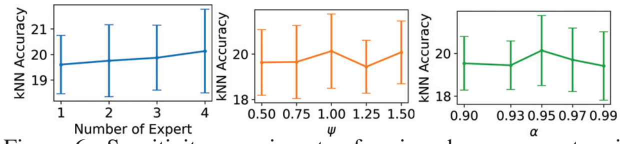

Figure 6. Sensitivity experiments of various hyperparameters in EVOLVE, tested on BYOL and **Seq-imb Stream-51** streams.

7. Summary

In this paper, we demonstrate that when applied to practical temporal-correlated data streams, current UCL methods experience a significant performance drop. We propose a general expert-guided learning framework, called EVOLVE, for enhancing the UCL capability of existing self-supervised learning methods using a hybrid learning scheme. EVOLVE introduces multiple frozen experts, each with dynamically updatable weights, to guide the continual learner in adapting to challenging data streams. Extensive results demonstrate that EVOLVE significantly enhances the performance of existing UCL methods in challenging natural environments.

Acknowledgements

This work was supported in part by National Science Foundation under Grants #2003279, #1826967, #2100237, #2112167, #1911095, #2112665, and in part by SRC under task #3021.001, and in part by PRISM and CoCoSys, centers in JUMP 2.0, an SRC program sponsored by DARPA. This work was also supported in part by a startup funding by the University of Texas at Dallas.

References

[1] Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3931–3940, 2020. 2

[2] Alessandro Achille, Tom Eccles, Loic Matthey, Christopher P Burgess, Nick Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. *arXiv preprint arXiv:1808.06508*, 2018. 1, 2, 3

[3] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. *Advances in neural information processing systems*, 32, 2019. 2

[4] Abdulatif Alabdulatif, Ibrahim Khalil, and Xun Yi. Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption. *Journal of Parallel and Distributed Computing*, 137:192–204, 2020. 2

[5] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 2

[6] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012. 3, 6, 7

[7] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 1, 2, 4

[8] Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. *Advances in Neural Information Processing Systems*, 34, 2021. 5

[9] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 4

[10] Sébastien Bubeck. Introduction to online optimization. *Lecture notes*, 2:1–86, 2011. 3

[11] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020. 2, 6

[12] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 2

[13] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2

[14] Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997. 3

[15] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. 2, 3

[16] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9516–9525, 2021. 1

[17] Sungmin Cha, Dongsub Shim, Hyunwoo Kim, Moontae Lee, Honglak Lee, and Taesup Moon. Is continual learning truly learning representations continually? *arXiv preprint arXiv:2206.08101*, 2022. 2, 3, 4

[18] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017. 2

[19] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. 2

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 4, 5

[21] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[22] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 6

[23] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1, 4

[24] Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. In *International Conference on Machine Learning*, pages 1952–1961. PMLR, 2020. 2

[25] Cisco. What are 5g speeds?, 2023. 8

[26] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8250–8259, 2021. 1, 3

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[28] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR, 2021. 2

[29] Enrico Fini, Victor G Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. *arXiv preprint arXiv:2112.04215*, 2021. 1, 2, 3, 6

[30] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997. 3, 6, 7

[31] Craig Gentry. *A fully homomorphic encryption scheme*. Stanford university, 2009. 2

[32] Yasir Ghunaim, Adel Bibi, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader Hammoud, Ameya Prabhu, Philip HS Torr, and Bernard Ghanem. Real-time evaluation in online continual learning: A new hope. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11888–11897, 2023. 2

[33] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1

[34] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005. 5

[35] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 1, 4

[36] Yunhui Guo, Mingrui Liu, Tianbao Yang, and Tajana Rosing. Improved schemes for episodic memory-based lifelong learning. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[37] Philip Haeusser, Johannes Plapp, Vladimir Golkov, Elie Aljalbout, and Daniel Cremers. Associative deep clustering: Training a classification network with no labels. In *German Conference on Pattern Recognition*, pages 18–32. Springer, 2018. 2

[38] Najmul Hassan, Kok-Lim Alvin Yau, and Celimuge Wu. Edge computing in 5g: A review. *IEEE Access*, 7:127276–127289, 2019. 2

[39] Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 128–141, 2016. 6

[40] Jiangpeng He and Fengqing Zhu. Unsupervised continual learning via pseudo labels. *arXiv preprint arXiv:2104.07164*, 2021. 2, 3

[41] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 5, 6

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[43] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 5

[44] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 5

[45] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International conference on machine learning*, pages 1558–1567. PMLR, 2017. 2

[46] Wenpeng Hu, Qi Qin, Mengyu Wang, Jinwen Ma, and Bing Liu. Continual learning by using information of each class holistically. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7797–7805, 2021. 2

[47] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019. 2

[48] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1965–1972, 2017. 2

[49] Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514, 2020. 2

[50] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[51] Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. In *International Conference on Learning Representations*, 2020. 2

[52] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68, 2020. 1

[53] Jin Li, Zhong Ji, Gang Wang, Qiang Wang, and Feng Gao. Learning from students: Online contrastive distillation network for general continual learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3215–3221, 2022. 3

[54] Yazhe Li, Roman Pogodin, Danica J Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556, 2021. 2

[55] Zhiwei Lin, Yongtao Wang, and Hongxiang Lin. Continual contrastive self-supervised learning for image classification. *arXiv preprint arXiv:2107.01776*, 2021. 2, 3

[56] Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994. 3, 6, 7

[57] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021. 2

[58] 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 IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 6

[59] V Lomanco and Davide Maltoni. Core50: a new dataset and benchmark for continual object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 17–26, 2017. 6

[60] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6470—6479, 2017. 2

[61] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 4, 6

[62] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1

[63] Eric Nalisnick and Padhraic Smyth. Stick-breaking variational autoencoders. *arXiv preprint arXiv:1605.06197*, 2016. 2

[64] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 3, 6

[65] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4, 6

[66] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019. 3

[67] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11321–11329, 2019. 2

[68] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[69] Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet K Dokania, Philip HS Torr, Ser-Nam Lim, Bernard Ghanem, and Adel Bibi. Computationally budgeted continual learning: What does matter? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3698–3707, 2023. 2

[70] Senthil Purushwalkam, Pedro Morgado, and Abhinav Gupta. The challenges of continuous self-supervised learning. pages 702–721, 2022. 2, 3, 4

[71] Qualcomm. The future of ai is hybrid, May 2023. 2, 5

[72] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. itaml: An incremental task-agnostic meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13588–13597, 2020. 2

[73] Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Lifelong generative modeling. *Neurocomputing*, 404:381–400, 2020. 1, 2

[74] Dushyant Rao, Francesco Visin, Andrei A Rusu, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Continual unsupervised representation learning. *arXiv preprint arXiv:1910.14481*, 2019. 1, 2, 3

[75] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Lsd-c: Linearly separable deep clusters. *arXiv preprint arXiv:2006.10039*, 2020. 2

[76] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2

[77] Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[78] Ryne Roady, Tyler L. Hayes, Hitesh Vaidya, and Christopher Kanan. Stream-51: Streaming classification and novelty detection from videos. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 3, 6

[79] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2, 6

[80] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2, 6

[81] Khadija Shaheen, Muhammad Abdullah Hanif, Osman Hasan, and Muhammad Shafique. Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks. *Journal of Intelligent & Robotic Systems*, 105(1):1–32, 2022. 1

[82] James Smith, Cameron Taylor, Seth Baer, and Constantine Dovrolis. Unsupervised progressive learning and the stam architecture. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2979–2987. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 2, 3

[83] Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2022. 2

[84] Johannes Von Oswald, Christian Henning, João Sacramento, and Benjamin F Grewe. Continual learning with hypernetworks. 2020. 2

[85] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing Hong, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay with data compression for continual learning. *arXiv preprint arXiv:2202.06592*, 2022. 2

[86] Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Memory replay gans: Learning to generate new categories without forgetting. In *NeurIPS*, 2018. 1, 2

[87] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016. 2

[88] Fei Ye and Adrian G Bors. Learning latent representations across multiple data domains using lifelong vaegan. In *European Conference on Computer Vision*, pages 777–795. Springer, 2020. 1, 2, 3

[89] Fei Ye and Adrian G Bors. Task-free continual learning via online discrepancy distance learning. *arXiv preprint arXiv:2210.06579*, 2022. 3

[90] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pages 12133–12143. PMLR, 2021. 5

[91] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6982–6991, 2020. 2

[92] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 1, 2, 4

[93] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017. 2, 6

[94] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020. 2

[95] Yan Zhang and Yan Zhang. Mobile edge computing for beyond 5g/6g. *Mobile Edge Computing*, pages 37–45, 2022. 2

[96] Feng Zhao, Chao Li, and Chun Feng Liu. A cloud computing security solution based on fully homomorphic encryption. In *16th international conference on advanced communication technology*, pages 485–488. IEEE, 2014. 2