SCALE: Online Self-Supervised Lifelong Learning without Prior Knowledge

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

Abstract

Unsupervised lifelong learning refers to the ability to learn over time while memorizing previous patterns without supervision. Although great progress has been made in this direction, existing work often assumes strong prior knowledge about the incoming data (e.g., knowing the class boundaries), which can be impossible to obtain in complex and unpredictable environments. In this paper, motivated by real-world scenarios, we propose a more practical problem setting called online self-supervised lifelong learning without prior knowledge. The proposed setting is challenging due to the non-iid and single-pass data, the absence of external supervision, and no prior knowledge. To address the challenges, we propose Self-Supervised ContrAstive Lifelong LEarning without Prior Knowledge (SCALE) which can extract and memorize representations on the fly purely from the data continuum. SCALE is designed around three major components: a pseudo-supervised contrastive loss, a self-supervised forgetting loss, and an online memory update for uniform subset selection. All three components are designed to work collaboratively to maximize learning performance. We perform comprehensive experiments of SCALE under iid and four non-iid data streams. The results show that SCALE outperforms the state-of-the-art algorithm in all settings with improvements up to 3.83%, 2.77% and 5.86% in terms of kNN accuracy on CIFAR-10, CIFAR-100, and TinyImageNet datasets. We release the implementation at https://github.com/Orienfish/SCALE.

1. Introduction

Lifelong learning, or continual learning, refers to the ability to continuously learn over time by acquiring new knowledge and consolidating past experiences. One major challenge of lifelong learning is to combat *catastrophic forgetting*, i.e., updating the model using new samples degrades existing knowledge learned in the past [26, 51].

Existing work has assumed various levels of *prior knowledge* about the input data stream. *Supervised Lifelong Learning* presumes the presence of task and class labels along with samples [14, 29, 41, 48]. *General Continual Learning* or *task-free continual learning* eliminates the task labels and boundaries to focus on real-time adaptation to non-stationary continuum with limited memory, but still using class labels [6, 9, 44, 78]. *Unsupervised Lifelong Learning* completely removes all labels; therefore, the algorithm needs to distill the knowledge from raw samples or streaming structure on its own [2, 37, 75].

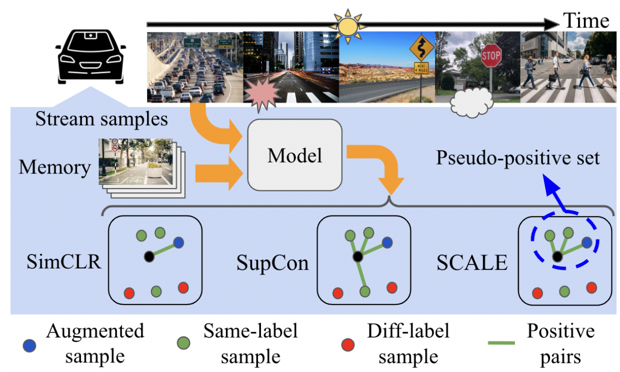

Figure 1. SCALE functions on a self-driving vehicle where the order of the input image sequence can be unforeseeable due to environmental or operational factors. SCALE learns self-supervisedly by contrasting with memory samples. SCALE's pseudo-contrastive loss is inspired from the InfoNCE objective [54]. While SimCLR [15] only uses an augmented sample and SupCon [40] uses samples with the same label to form a positive set, to improve similarity within the set, SCALE self-distills a pseudo-positive set based on pairwise similarity. SCALE does not rely on any supervision or prior knowledge.

While great progress has been made in lifelong learning, it is still challenging to deploy the existing algorithms in the wild to learn over time. One of the reasons is that even in the pure unsupervised setting, existing works assumed knowing the class boundary or the total number of classes in advance [57, 61, 70]. Such prior knowledge greatly eases the difficulty of learning without forgetting. For example, if the class boundary is distinct and known, the learning algorithm can expand the network or create a new memory buffer whenever detecting a class shift. But these prior knowledge is extremely difficult, if not impossible, to obtain in real-world environments which are *complex* and *unpredictable*. Specifically, consider a camera mounted on a vehicle and an application of continuously training an image classification algorithm as the vehicle moves around (Figure 1). The sequence of incoming samples depends on the environment and the trajectory of the vehicle, hence, is very hard to predict when and how smooth the shift is.

In this paper, to align with the unpredictable real-world scenarios, we extend the current unsupervised learning setting to a more challenging and practical case: online unsupervised lifelong learning without prior knowledge. In particular, we make no assumption on the input streams:

(i) Unlike offline self-supervised learning [11, 15], the input data is non-*iid* and *single-pass*, i.e., all data samples appear only once.

(ii) Unlike *General Continual Learning* [9, 44] and task-based lifelong learning [24, 31, 46, 49], the class and task labels are not given (*no external supervision*).

(iii) Unlike VAE-based design [61] and KMeans-based progressive clustering [31, 69], the task or class boundaries and the number of classes are unknown in advance (*no prior knowledge*).

Additionally, the input stream can have distinct/blurred class boundaries or an imbalanced class appearance, all of which are not revealed to the algorithm. Our problem setting reflects the complexity and difficulty of lifelong learning problems in the real world[1].

Recognizing the unique challenges, we propose Self-Supervised ContrAstive Lifelong LEarning without Prior Knowledge (SCALE). SCALE is designed around three major components: a pseudo-supervised contrastive loss for contrastive learning, a self-supervised forgetting loss for lifelong learning, and an online memory update for uniform subset selection. All components are critical to the final learning performance: the contrastive loss enhances the similarity relationship by contrasting with memory samples, the forgetting loss prevents catastrophic forgetting, and the memory buffer retains the most "representative" raw samples within the limited buffer size. Our loss functions utilize pairwise similarity among the feature representations, thus eliminating the dependency on labels or prior knowledge. Moreover, contrastively learned representations have been shown to be more robust against catastrophic forgetting compared to the use of end-to-end cross-entropy loss [12].

Our **contributions** can be summarized as follows:

(1) We propose a more practical setting for unsupervised lifelong learning which assumes that the input data streams are non-*iid* and single pass, and no external supervision or prior knowledge is given.

(2) We design SCALE to extract and memorize knowledge on-the-fly without supervision and prior knowledge. SCALE uses contrastive lifelong learning based on self-distilled pairwise similarity, along with an online memory update to retain the "representative" raw samples on imbalanced streams.

(3) We perform comprehensive experiments on five different types of single-pass data stream sampled from CIFAR-10, CIFAR-100 and TinyImageNet datasets. SCALE outperforms state-of-the-art algorithms in all settings.

[1] In this paper we focus on image classification while the same setup and methodology can be easily extended to other applications as well.

2. Related Work

Self-Supervised Learning (SSL) has been developed to learn low-dimensional representations on offline datasets without class labels, for various downstream tasks. Variational autoencoder (VAE)-based designs aimed for data reconstruction assuming various prior models in the latent space [37, 39, 52]. Progressive clustering-based methods alternated between network update and clustering for self-labeling until convergence [10, 11, 13, 30, 62, 76]. Information theory-based techniques maximized the mutual information between representations of augmented samples to retain invariance and avoid degenerate solutions [8, 22, 34, 36, 45, 80]. Contrastive learning draws closer the augmented representation pairs while pushing away the others [15–17, 32, 54]. Recent architecture techniques such as BYOL, SimSiam and OBoW [18, 25, 27] used asymmetric networks to prevent learning trivial representations. However, all the above-mentioned works are designed for offline *iid* data and do not address catastrophic forgetting.

Supervised Lifelong Learning has been widely explored in three lines: dynamic architecture [1, 43, 55, 59, 65, 73], regularization [3, 4, 41, 64, 79, 81, 82], and experience replay using a memory buffer [9, 14, 19, 29, 35, 48, 63, 71, 74]. Recently, a large amount of effort has been invested in online supervised lifelong learning. Most works used memory replay, such as Co2L [12], CoPE [20], GMED [38], DualNet [56], ASER [67], SCR [50], OCM [28], ODDL [78], OCD-Net [44]. Nevertheless, the problem is significantly simplified with the presence of class labels.

Unsupervised Lifelong learning (ULL) is mostly studied under offline *iid* data with multiple passes on the entire dataset during training [2, 37, 75, 77]. In contrast, online ULL is more challenging due to the non-*iid* and single-pass data continuum. Lifelong generative models leveraged mixture generative replay to mitigate catastrophic forgetting during online updates [60, 61]. However, these VAE-based methods were computationally expensive. Many recent works have applied self-supervised knowledge distillation on task-based online ULL. **He *et al.*** [31] utilized pseudo-labels from KMeans clustering to guide knowledge preservation from the previous task. **CCSL** [46] em-

Table 1. Comparison of previous work and SCALE (this paper) on assumed prior knowledge.

Papers	Single-pass	Non-*iid*	No task labels	No class labels
VASE [2], CURL [61], L-VAEGAN [77]	×	✓	✓	✓
He *et al.* [31], CCSL [46], CaSSLe [24], LUMP [49]	✓	✓	×	✓
Tiezzi *et al.* [70], KIERA [57]	✓	✓	✓	×
STAM [68], SCALE (this paper)	✓	✓	✓	✓

ployed self-supervised contrastive learning for intra- and inter-task distillation. **CaSSLe** [24] proposed a general framework for SSL backbones, which extracted the best possible representations that are invariant to task shifts. **LUMP** [49] mitigated forgetting by interpolating the current task's samples with the finite memory buffer. But all of these works relied on task boundaries to generate good results. **Tiezzi *et al.*** [70] developed a human-like attention mechanism for continuous video streams with little supervision. **KIERA** [57] and **STAM** [68] employed expandable memory architecture for single-pass data using online clustering, novelty detection and memory update. KIERA required labeled samples in the initial batch of each task for cluster association. The problem definition of STAM is most similar to ours. Yet, STAM's memory architecture cannot be trained with common optimizers, and thus is limited in fine-tuning for downstream tasks.

We summarize the existing contributions for online ULL in Table 1 based on the assumed prior knowledge. The proposed SCALE excels existing works in that SCALE learns low-dimension representations online without any external supervision or prior knowledge about task, class or data; thus, it better adapts unpredictable real-world environments.

3. Online Unsupervised Lifelong Learning without Prior Knowledge

In this section, we present the online unsupervised lifelong learning problem without prior knowledge. Our setup is motivated by real-world applications and extended from previous studies by removing certain assumptions.

Input streams. We assume the data comes in a class- (or distribution-) incremental manner. Such a setup mimics continuous and periodic sampling while the surrounding environment changes over time. Suppose that the input samples are drawn from a sequence of T classes with each class corresponding to a unique distribution in $\{\mathcal{P}^1, ..., \mathcal{P}^T\}$. The complete input sequence can then be represented as $\mathcal{D} = \{\mathcal{D}^1, ..., \mathcal{D}^T\}$ where $\mathcal{D}^t$ denotes a series of n_t batches of samples, i.e., $\mathcal{D}^t = \{X_1^t, ..., X_{n_t}^t\}$. With t denoting the class ID and u representing the batch ID in the current class, each batch of data X_u^t is a set of samples $\{\mathbf{x}_1^t, ..., \mathbf{x}_{|X_u^t|}^t\}$, where $\mathbf{x}_i^t \sim \mathcal{P}^t(\mathbf{X})$. In the rest of the paper, we use capital letters to denote batches and lowercase letters for individual samples. Each training batch $X_u^t \in \mathcal{D}^t$ appears once in the entire stream (single-pass) while the task and class labels are not revealed. The total number of classes T, the transition boundaries and the batch numbers n_t are not known by the learning algorithm either. Our goal is to learn a model that distinguishes classes or distributions $\{\mathcal{P}^1, ..., \mathcal{P}^T\}$ *at any moment* throughout the stream, without supervision by external labels or prior knowledge.

Based on previous problem formulations [57,61,68], five particular types of input streams are considered: (i) *iid* data that is sampled *iid* from all classes. (ii) Sequential class-incremental stream where the observed classes are balanced in length and are introduced one-by-one with clear boundaries that are not known by the algorithm. (iii) Sequential class-incremental stream with blurred boundaries. The boundary is blurred by mixing the samples from two consecutive classes, mimicking class shift is smooth and difficult to detect. (iv) Imbalanced sequential class-incremental stream uses different batch sizes in each class, mimicking distribution shifts at unpredictable times. (v) Sequential class-incremental stream with concurrent classes where more than one class is incrementally introduced at the time. In this paper two classes are revealed concurrently and $\mathcal{P}^i$ refers to their combined distribution. To aid understanding, we use a self-driving vehicle with a mounted camera as an example to visualize all five input streams as shown in Figure 2.

Training and evaluation protocol. The training and evaluation setup is similar to [61, 68] and is detailed in Figure 2. The model is a representation mapping function to a low-dimensional feature space, i.e., $f_\theta : \mathcal{X} \to \mathcal{Z}$ where θ represents learnable parameters and $\mathcal{Z}$ refers to the low-dimensional feature space. The training proceeds self-supervisedly based on the feature representation batch $Z_u^t = f_{\theta_u^t}(X_u^t)$. As for evaluation, we periodically test the frozen model θ_u^t on a separate dataset $\mathcal{E} = \{(\mathbf{x}_j, y_j)\}$ as the training progresses. We randomly sample an equal amount of labeled samples from each class possibly seen in $\{\mathcal{P}^1, ..., \mathcal{P}^T\}$ and add them to $\mathcal{E}$. Thus even when the class has not shown up in the sequence, it is always included in $\mathcal{E}$. For each testing sample $(\mathbf{x}_j, y_j) \in \mathcal{E}$, we first compute the learned latent representations $\mathbf{z}_j = f_{\theta_u^t}(\mathbf{x}_j)$. We then apply a classifier $g : \mathcal{Z} \to \mathcal{Y}$ on $\mathbf{z}_j$ to generate the predicted labels $\hat{y}_j$. The classifier g can be unsupervised or supervised to evaluate different aspects of the representation learning ability. Following previous protocols [61, 68], we use spectral clustering, an unsupervised clustering method,

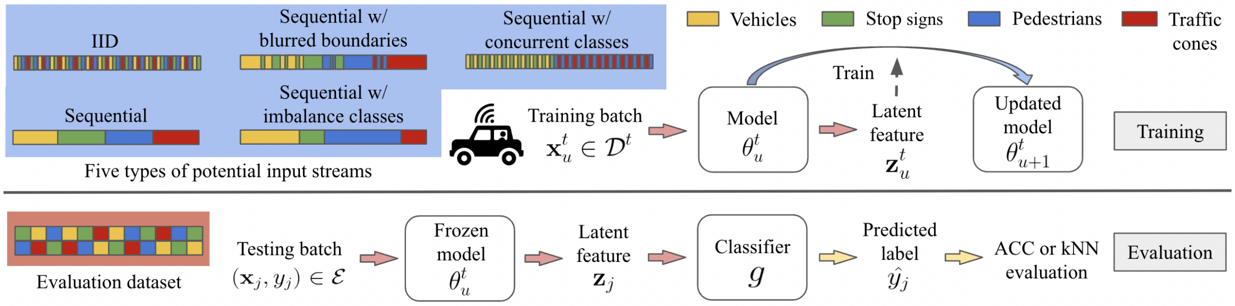

Figure 2. The training and evaluation setup of online ULL taking the self-driving vehicle as an example. The input samples are from classes of vehicles, stop signs, pedestrians and traffic cones. We consider five typical input streams of *iid*, sequential classes, sequential classes with blurred boundaries, imbalanced sequential classes, and sequential classes with concurrent class appearance. Each training step updates the model self-supervisedly based on the feature representations, while periodic evaluation is triggered on a separate evaluation dataset using supervised or unsupervised classifiers on the learned feature representations.

and employ *unsupervised clustering accuracy* (ACC) as the accuracy metric. ACC is defined as the best accuracy among all possible assignments between clusters and target labels:

$$ACC = \max_{\psi} \frac{\sum_{j=1}^{|\mathcal{E}|} \mathbf{1}\{y_j = \psi(\hat{y}_j)\}}{|\mathcal{E}|}. \tag{1}$$

Here, the predicted label $\hat{y}_j$ is the cluster assignment to sample $\mathbf{x}_j$, ψ ranges over all possible one-to-one mappings between $\hat{y}_j$ and y_j. For supervised classification, we employ k-Nearest Neighbor (kNN) classifier.

Challenges. The major difference between our online unsupervised lifelong learning and previous problems is the prior assumptions about the input stream. Online ULL is more challenging than previous ULL problems as shown in Table 1 from three aspects:

(C1) **The non-*iid* and single-pass input data streams** require online knowledge distillation, which is largely different from offline self-supervised learning with *iid* data and multi-pass training [11, 15, 80].

(C2) **The lack of task or class labels** differs our online ULL from *General Continual Learning* (with class labels) [9, 44] and task-based lifelong learning (with task labels) [24, 31, 46, 49]. The model must distill the knowledge from the stream on its own without external supervision.

(C3) **The absence of prior knowledge.** Existing ULL methods rely on class boundaries [24, 31, 46, 49] or maintain and update class prototypes after detecting a shift [57, 61]. However, these approaches do not apply when there is no prior knowledge, especially with smooth transitions, imbalanced streams or simultaneous classes as in our online ULL.

4. The Design of SCALE

To address the above challenges of online ULL, we propose SCALE, an unsupervised lifelong learning method that can learn over time without prior knowledge. An overview of SCALE is shown in Figure 3. SCALE is designed around three major components (shown in yellow boxes): a pseudo-supervised contrastive loss, a self-supervised forgetting loss, and an online memory update module that emphasizes uniform subset selection. By combining stored memory samples with the streaming samples during learning, SCALE addresses challenge (C1). Secondly, SCALE uses the newly proposed pseudo-supervised contrastive learning paradigm that distills the relationship among samples via pairwise similarity. Pseudo-supervised distillation works without task or class labels thus handles challenge (C2). Learning from pairwise similarity does not depend on class boundaries or the number of classes, therefore SCALE responds to challenge (C3).

We emphasize that all components are carefully designed to work collaboratively and maximize learning performance: the contrastive loss is responsible for extracting the similarity relationship by contrasting with memory samples, the forgetting loss retains the similarity knowledge thus prevents catastrophic forgetting, finally the online memory update maintains a memory buffer with representative raw samples in the past. We record the raw input samples rather than feature representations in the memory buffer because feature representations might change during training. The quality or the "representativeness" of memory samples can significantly affect learning performance, as demonstrated by our results in the evaluation section.

Figure 3 shows the pipeline of SCALE in detail. Memory buffer is assumed to have maximum size of M, and the stored memory samples are represented by $\{\mathbf{e}_i\}_{i=1}^{M}$. Each streaming batch X_u^t with batch size of $n = |X_u^t|$ is stacked with a randomly sampled subset of m memory samples to form a combined batch $\{\mathbf{x}_i\}_{i=1}^{m+n}$ as input to SCALE. We apply double-view augmentation to the stacked data and obtain $\{\tilde{\mathbf{x}}_i\}_{i=1}^{2(m+n)}$ where $\tilde{\mathbf{x}}_{2k-1}, \tilde{\mathbf{x}}_{2k}$ denote two randomly augmented samples from $\mathbf{x}_k$. The augmented samples are fed into the representation learning model f_θ to obtain normalized low-dimensional features $\tilde{\mathbf{z}}_i = f_\theta(\tilde{\mathbf{x}}_i), \forall i \in \{1, ..., 2(m+n)\}$. SCALE distills pairwise similarity from

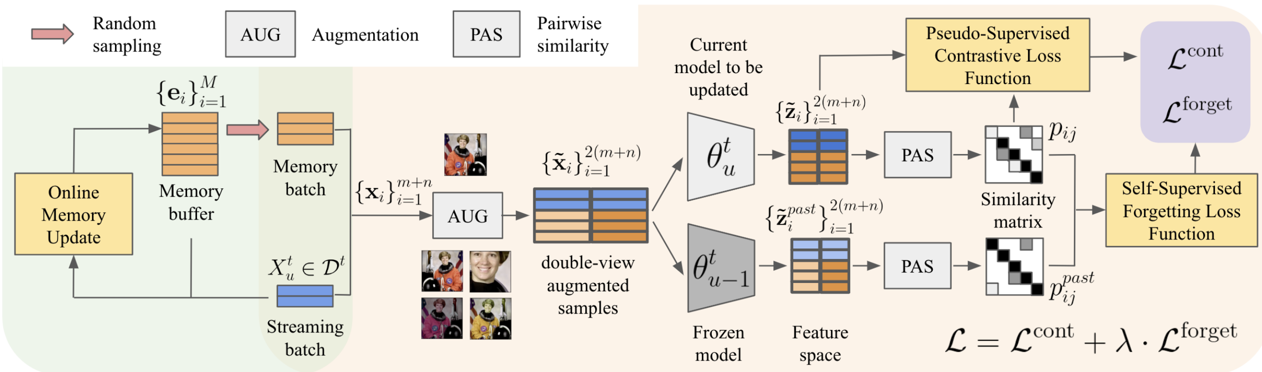

Figure 3. The pipeline of SCALE is designed around three major components depicted by yellow boxes. The right-hand portion in orange includes the operations related to self-supervised contrastive and lifelong learning. The left-hand portion in green contains the procedures related to online memory update. SCALE requires careful design for all three components to distill and memorize knowledge on the fly.

$\{\tilde{\mathbf{z}}_i\}_{i=1}^{2(m+n)}$, which are then used to compute the pseudo-supervised contrastive and forgetting losses to update the current model θ_u^t. On the other hand, online memory update takes previous memory buffer $\{\mathbf{e}_i\}_{i=1}^M$ and the streaming batch X_u^t as input, selects a subset of M samples to store in the updated memory buffer. We discuss the details below.

4.1. Pseudo-Supervised Contrastive Loss and Self-Supervised Forgetting Loss

The loss function of SCALE has two terms: a novel pseudo-supervised contrastive loss $\mathcal{L}^{\text{cont}}$ for learning representations and a self-supervised forgetting loss $\mathcal{L}^{\text{forget}}$ for preserving knowledge:

$$\mathcal{L} = \mathcal{L}^{\text{cont}} + \lambda \cdot \mathcal{L}^{\text{forget}}. \tag{2}$$

A hyperparameter λ is used to balance the two losses. Both loss functions rely on pairwise similarity hence do not need prior knowledge and adapt to a variety of streams.

Pseudo-Supervised Contrastive Loss. Our contrastive loss is inspired from the InfoNCE objective [54] which enhances the similarity between positive pairs over negative pairs in the feature space. SimCLR [15] and SupCon [40] are the typical offline contrastive learning techniques using InfoNCE loss. Different from SimCLR (treats only the augmented pair as positive, unsupervised) and SupCon (forms the positive set based on labels, supervised), SCALE establishes a pseudo-positive set based on pairwise similarity. Given a feature representation $\tilde{\mathbf{z}}_i$, its pseudo-positive pair $\tilde{\mathbf{z}}_j$ is selected from the self-distilled pseudo-positive set Γ_i. Negative pairs are all non-identical representations in the augmented batch $\{\tilde{\mathbf{z}}_i\}_{i=1}^{2(m+n)}$. Formally, the pseudo-supervised contrastive loss is defined as:

$$\mathcal{L}^{\text{cont}} = \sum_{i=1}^{2n} \frac{-1}{|\Gamma_i|} \sum_{j \in \Gamma_i} \log \frac{\exp(\tilde{\mathbf{z}}_i \cdot \tilde{\mathbf{z}}_j / \tau)}{\sum_{k=1, k \neq i}^{2(m+n)} \exp(\tilde{\mathbf{z}}_i \cdot \tilde{\mathbf{z}}_k / \tau)}, \tag{3}$$

where $\tau > 0$ is a temperature hyperparameter. Note, that all memory samples only act as negative contrasting pairs to avoid overfitting. Without task or class labels, SCALE distills the pairwise similarity p_{ij} and forms the pseudo-positive set as:

$$\Gamma_i = \{j \in \{1, ..., 2n\} \,|\, j \neq i, p_{ij} > \mu\}, \tag{4}$$

where p_{ij} (defined later) indicates the pairwise similarity among feature representations and $\mu > 0$ is a hyperparameter as similarity threshold. Our contrastive loss is unique and different from traditional contrastive loss functions [15,32,40] due to the self-distilled pseudo-positive set Γ_i, which maximizes the effectiveness of unsupervised representation learning in an online setting.

Self-Supervised Forgetting Loss. To combat catastrophic forgetting, we construct a self-supervised forgetting loss based on the KL divergence of the similarity distribution:

$$\mathcal{L}^{\text{forget}} = \sum_{i=1}^{2(m+n)} \sum_{j=1, j \neq i}^{2(m+n)} -p_{ij} \cdot \log \frac{p_{ij}}{p_{ij}^{past}}, \tag{5}$$

where p_{ij}, p_{ij}^{past} are the pairwise similarity among feature representations $\{\tilde{\mathbf{z}}_i\}_{i=1}^{2(m+n)}$ and $\{\tilde{\mathbf{z}}_i^{past}\}_{i=1}^{2(m+n)}$, which are mapped by the model θ_u^t and frozen model θ_{u-1}^t. To form a valid distribution, we enforce the pairwise similarity of a given instance to sum to one: $\sum_{j=1, j \neq i}^{2(m+n)} p_{ij} = 1, \forall i \in \{1, ..., 2(m+n)\}$. The same rule applies to p_{ij}^{past}. In SCALE, the learned knowledge is stored by pairwise similarity. Hence penalizing the KL divergence of pairwise similarity distribution from a past model can prevent catastrophic updates. As we are not aware of class or task boundaries, we use the frozen model from the previous batch. Note, that a similar distillation loss is used in [12, 31, 46] but for supervised or task-based lifelong learning.

Pairwise Similarity. Pairwise similarity is the key of SCALE hence picking the suitable metric is of critical importance. An appropriate pairwise similarity metric should

(i) consider the global distribution of all streaming and memory samples, and (ii) sum to one for a given instance as required by the forgetting loss. We adopt the symmetric SNE similarity metric from t-distributed stochastic neighbor embedding (t-SNE), which was originally proposed to visualize high-dimensional data by approximating the similarity probability distribution [72]:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}, \; p_{j|i} = \frac{\exp(\tilde{\mathbf{z}}_j \cdot \tilde{\mathbf{z}}_i / \kappa)}{\sum_{k=1, k\neq i}^{2(m+n)} \exp(\tilde{\mathbf{z}}_k \cdot \tilde{\mathbf{z}}_i / \kappa)}, \quad (6)$$

where $\kappa > 0$ is a temperature hyperparameter. Since the form of Equation (6) is similar to Equation (3), in practice, the computation can be reused to improve efficiency. The symmetric SNE similarity captures the global similarity distribution among all features without using supervision or prior knowledge.

4.2. Online Memory Update

The goal of online memory update is to retain the most "representative" raw samples from historical streams to obtain the best outcome in contrastive learning. One major challenge is that the input streams are non-*iid* and possibly imbalanced. Existing work has proposed various memory update strategies to extract the most informational samples, e.g., analyzing interference or gradients information [5, 7, 19, 38, 67]. However, most previous works rely on class labels thus are not applicable in online ULL. Without labels and prior knowledge, we cannot make any assumption (for example, clusters) on the manifold of the feature representations that are fed to memory update. Purushwalkam *et al.* [58] were the first to bring up a similar problem setting and proposed minimum redundancy (MinRed) memory update, prioritizing dissimilar samples without considering the global distribution. Unlike MinRed, we propose to perform distribution-aware uniform subset sampling for memory update.

The input to memory update is the imbalanced combined batch $\{\mathbf{x}_i\}_{i=1}^{M+n}$ of the previous memory buffer $\{\mathbf{e}_i\}_{i=1}^{M}$ and streaming batch X_u^t. We first map the raw samples to the feature space, i.e., $\mathbf{z}_i = f_\theta(\mathbf{x}_i), \forall i \in \{1, ..., M+n\}$. Then we select a subset of M samples from $\{\mathbf{z}_i\}_{i=1}^{M+n}$ and store the corresponding raw samples in the limited-size memory buffer, while discard the rest. Aiming at extracting the representative samples from non-*iid* streams without supervision, SCALE employs the Part and Select Algorithm (PSA) [66] for uniform subset selection. PSA first performs M partition steps which divide all samples into M subsets, then picks one sample from each subset. Each step partitions the existing set with the greatest dissimilarity among its members, thus PSA selects a subset of samples with uniform distribution in the spanned feature space. To the best of the authors' knowledge, this is the first time using uniform subset selection in lifelong learning problems.

5. Evaluation

5.1. Experimental Setup

Datasets: We construct the online single-pass data streams from CIFAR-10 (10 classes) [53], CIFAR-100 (20 coarse classes) [42] and a subset of TinyImageNet (10 classes) [21]. For each dataset, we construct five types of streams: ***iid***, sequential classes (**seq**), sequential classes with blurred boundaries (**seq-bl**), sequential classes with imbalance lengths (**seq-im**), and sequential classes with concurrent classes (**seq-cc**).

Networks: For all datasets, we apply ResNet-18 [33] with a feature space dimension of 512.

Baselines. Since SCALE uses an InfoNCE-based loss, we compare with **SimCLR** [15] and **SupCon** [40] and the following lifelong learning baselines using SimCLR as backbone:

- From the group of supervised lifelong learning, we select **PNN** [65], **SI** [81] and **DER** [9] with necessary modifications for online ULL.
- For task-based ULL, we use the source code of **CaSSLe** [24] after removing the task labels.
- Finally, we also compare with **STAM** [68], using their original data loader and parameters, and **LUMP** [49].

We did not compare with VAE-based methods such as [37, 61] since they are reported to scale poorly on medium to large image datasets [23].

Metrics. We use spectral clustering with T as the number of clusters and compute the ACC. kNN classifier is used to evaluate the supervised accuracy with $k = 50$.

Implementation details of SCALE and baselines are presented in the Appendix. All memory methods use a buffer of size $M = 1280$. The size of the sampled memory batch is $m = 128$, which is the same as the size of the streaming batch n. For the similarity threshold, we use an adaptive threshold of $mean + \mu(max - mean)$ where $mean$ and max are the mean and max pairwise similarity in p_{ij}. With an adaptive threshold, we alleviate the effects of variations in absolute similarity. SCALE employs the Stochastic Gradient Descent optimizer with a learning rate of 0.03.

5.2. Accuracy Results

Final Accuracy. The final ACC and kNN accuracy on all datasets and all data streams are reported in Figure 4. Both mean and standard deviation of the accuracy are reported after 3 random trials. ACC values are generally lower than their kNN counterparts. It should be noted that SCALE outperforms all state-of-the-art ULL algorithms on almost all streaming patterns, both in terms of ACC and kNN accuracy. In all settings in CIFAR-10, SCALE improves 1.69-4.62% on ACC and 1.32-3.83% on kNN comparing with the best performed baseline. For CIFAR-100, SCALE achieves improvements of up to 2.15% regarding ACC and 2.77%

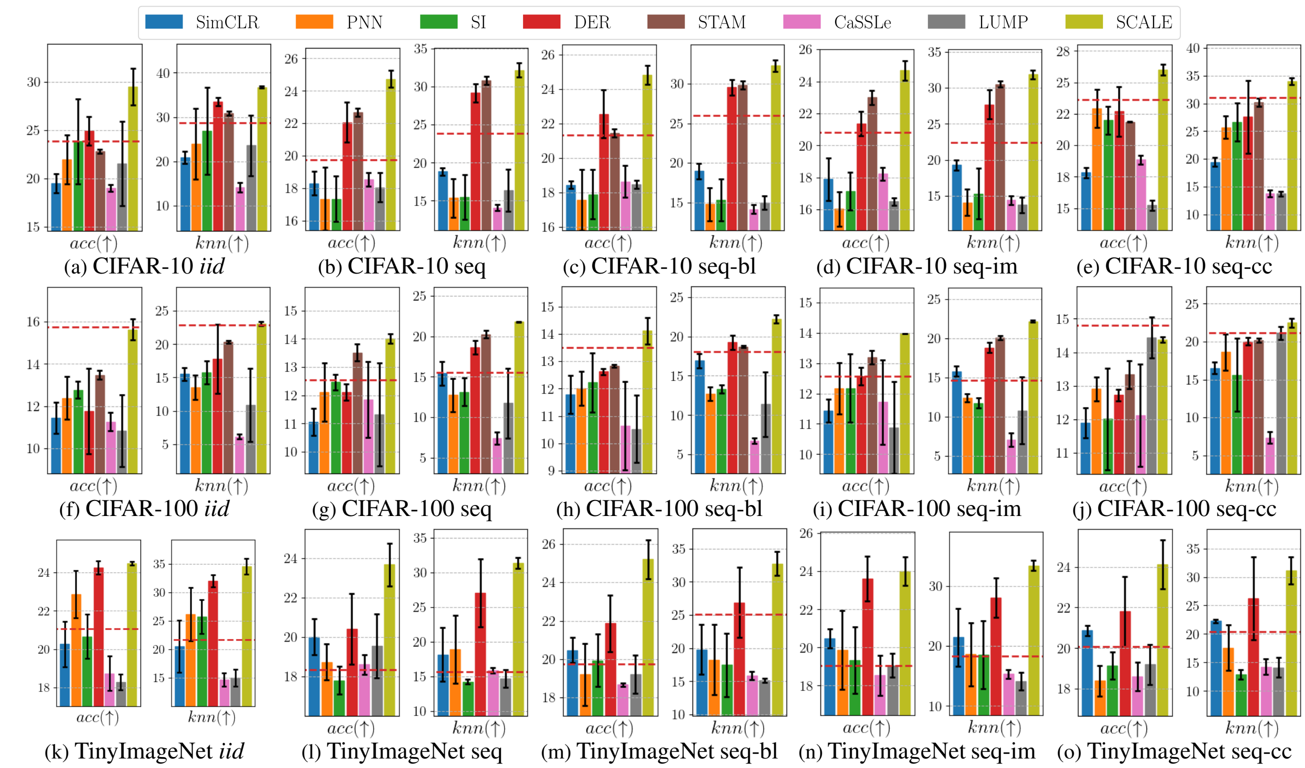

Figure 4. **SCALE improves kNN accuracy over the best state-of-the-art baseline by up to 3.83%, 2.77% and 5.86% kNN on CIFAR-10, CIFAR-100 and TinyImageNet datasets.** The figures show final accuracy results on five different streams sampled from CIFAR-10 (first row), CIFAR-100 (second row) and TinyImageNet (third row) datasets. For each data stream setting, the left figure displays ACC while the right figure shows the kNN accuracy at the end of the stream. The red dashed line depicts the ACC or kNN accuracy of SupCon.

regarding kNN comparing with the best baseline. For TinyImageNet, SCALE enhances 0.2-3.33% on ACC and 2.53-5.86% on kNN accuracy over the best baseline. Out of all data streams, *iid* and seq-cc streams are easier to learn while the single-class sequential streams are more challenging and result in lower accuracy. Our results demonstrate the strong adaptability of SCALE which does not require any prior knowledge about the data stream.

Baseline Performances. SimCLR produces low accuracy as it is originally designed for offline unsupervised representation learning with multiple epochs. Interestingly, the supervised contrastive learning baseline, SupCon (shown by red dashed line in Figure 4), does not always result in superior accuracy and can be attributed to overfitting on the limited memory buffer. Such result aligns with the recent findings that self-supervisedly learned representations are more robust than supervised counterparts under non-*iid* streams [24, 47]. Among the techniques adapted from supervised lifelong learning, DER achieves relatively good results on all datasets but is still not comparable with SCALE. The recently proposed ULL module, CaSSLe , significantly relies on task boundary knowledge to preserve the classification semantics from previous tasks, thus showing poor results in our online ULL setup. LUMP utilizes a mixup data augmentation technique and may not work well for certain image datasets. STAM outperforms the rest of the baselines. However, STAM utilizes a unique memory architecture and cannot be fine-tuned for downstream tasks.

Accuracy Curve. To examine the dynamics of online learning, we summarize the kNN accuracy curves during training on blurred sequential CIFAR-10 and CIFAR-100 streams in Figure 5 (more results in the Appendix). We can observe that SCALE enjoys gradually increasing kNN accuracy as we introduce new classes, which demonstrates SCALE's ability to consistently learn new knowledge while consolidating past information, all without supervision or prior knowledge. Most baselines are subject to collapse or forgetting, and are not able to distill or remember the knowledge in online ULL. The expandable-memory baseline STAM is incapable of learning without effective novelty detection.

5.3. Ablation Studies

Loss Functions. We experiment with various combinations of contrastive loss and forgetting loss on the *sequential* streams, as shown in Table 2. Even with a replay buffer, SimCLR and SupCon do not lead to satisfying results on online ULL. Co2L [12] is a supervised lifelong learning baseline using constrastive and forgetting losses. For fair comparison, we remove its dependence on class labels. With the pseudo-supervised contrastive loss, SCALE gains 1.55% in

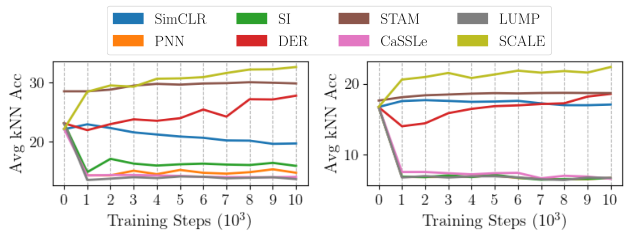

Figure 5. The average kNN accuracy during training on the *blurred sequential* streams sampled from CIFAR-10 (left) and CIFAR-100 (right) datasets. Each training trial contains 10k training steps while each class spans 1k steps.

Table 2. Average final kNN accuracy on the *sequential* streams, under different combinations of loss functions.

Contrast Loss	Forget Loss	CIFAR-10	TinyImageNet
SimCLR [15]	×	18.84	18.13
SupCon [40]	×	23.83	15.67
Co2L [12]	✓	30.63	30.80
SCALE	×	30.45	30.40
SCALE	✓	**32.18**	**31.33**

Table 3. Average final kNN accuracy on the *imbalanced sequential* streams using different memory update policies in SCALE.

Memory update	CIFAR-10	CIFAR-100	TinyImageNet
w/ label	32.41	21.21	27.73
random	29.80	20.10	23.67
KMeans	31.59	22.15	29.07
MinRed [58]	23.66	19.75	25.13
PSA (this paper)	**32.21**	**23.16**	**31.33**

terms of kNN accuracy on CIFAR-10 compared to Co2L with a traditional contrastive loss. With the forgetting loss, SCALE gets 1.78% kNN accuracy gain on CIFAR-10.

Memory Update Policies. We experiment with SCALE on the *imbalanced sequential* stream with different memory update policies and summarize the results in Table 3. With the distribution-aware uniform PSA memory update, SCALE surpasses the rest unsupervised strategies. KMeans-based memory selection does not lead to the best result on sequential streams as the representations are not separable. MinRed [58] prioritizes dissimilar samples regardless of global distribution, thus leads to biased selection and degraded performance on imbalanced data. All components of SCALE are necessary for the best overall learning performance.

5.4. Hyperparameters

We experiment with the important parameters in SCALE. The weight balancing coefficient λ plays an important role in the balance between pseudo-contrastive loss and self-supervised forgetting loss in SCALE. The accuracy

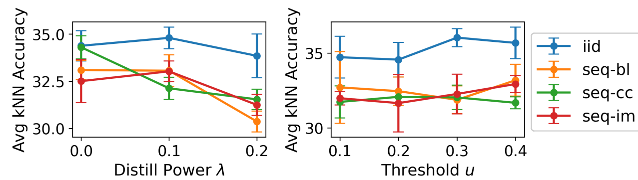

Figure 6. Hyperparameters experiments on CIFAR-10 streams under various λ (left) and u (right).

on various CIFAR-10 streams after 3 random trials, under various λ, are plotted in Figure 6 (left). In most settings, $\lambda = 0.1$ produces the best results. A smaller λ places less weight on the forgetting loss thus leads to forgetting; conversely, a larger λ may over-emphasize the memorizing effect and prevent learning meaningful representations.

The threshold u is critical in defining the pseudo-positive set. The accuracy after 3 random trials on CIFAR-10 streams are shown in Figure 6 (right). The sensitivity to threshold on *iid* and sequential streams is different. For *iid* streams, each incoming batch contains diverse samples from all classes. A higher threshold improves performance by restricting the pseudo-positive set to near-by samples that are more likely to belong to one class. For sequential streams, as the samples from the same batch are from the same class, a positive but lower threshold helps filter sufficiently similar samples into the pseudo-positive set, to boost learning outcome.

6. Conclusion

Existing works in unsupervised lifelong learning assume various prior knowledge thus are not applicable for learning in the wild. In this paper, we propose the online unsupervised lifelong learning problem without prior knowledge that (i) accepts non-*iid*, non-stationary and single-pass streams, (ii) does not rely on external supervision, and (iii) does not assume prior knowledge. We propose SCALE, a self-supervised contrastive lifelong learning technique based on pairwise similarity. SCALE uses a pseudo-supervised contrastive loss for representation learning, a self-supervised forgetting loss to avoid catastrophic forgetting, and an online memory update for uniform subset selection. Experiments demonstrate that SCALE improves kNN accuracy over the best state-of-the-art baseline by up to 3.83%, 2.77% and 5.86% on all non-*iid* CIFAR-10, CIFAR-100 and TinyImageNet streams.

Acknowledgements

This work was supported in part by National Science Foundation under Grants #2003279, #1830399, #1826967, #2100237, #2112167, #2112665, and in part by SRC under task #3021.001. This work was also supported 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] 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

[5] Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. *arXiv preprint arXiv:1908.04742*, 2019. 6

[6] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019. 1

[7] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 6

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

[9] 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. 1, 2, 4, 6

[10] 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

[11] 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, 4

[12] 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. 2, 5, 7, 8

[13] 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

[14] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. 1, 2

[15] 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, 6, 8

[16] 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

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

[18] 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. 2

[19] Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. In *International Conference on Machine Learning*, pages 1952–1961. PMLR, 2020. 2, 6

[20] 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. 2

[21] 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

[22] 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

[23] William Falcon, Ananya Harsh Jha, Teddy Koker, and Kyunghyun Cho. Aavae: Augmentation-augmented variational autoencoders. *arXiv preprint arXiv:2107.12329*, 2021. 6

[24] 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. 2, 3, 4, 6, 7

[25] Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Online bag-of-visual-words generation for unsupervised representation learning. *arXiv preprint arXiv:2012.11552*, 2020. 2

[26] 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

[27] 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. 2

[28] Yiduo Guo, Bing Liu, and Dongyan Zhao. Online continual learning through mutual information maximization. In *International Conference on Machine Learning*, pages 8109–8126. PMLR, 2022. 2

[29] 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. 1, 2

[30] 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

[31] Jiangpeng He and Fengqing Zhu. Unsupervised continual learning via pseudo labels. *arXiv preprint arXiv:2104.07164*, 2021. 2, 3, 4, 5

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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. 1, 2, 6

[38] Xisen Jin, Arka Sadhu, Junyi Du, and Xiang Ren. Gradient-based editing of memory examples for online task-free continual learning. *Advances in Neural Information Processing Systems*, 34:29193–29205, 2021. 2, 6

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

[40] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 1, 5, 6, 8

[41] 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. 1, 2

[42] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[43] 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

[44] 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. 1, 2, 4

[45] 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

[46] Zhiwei Lin, Yongtao Wang, and Hongxiang Lin. Continual contrastive self-supervised learning for image classification. *arXiv preprint arXiv:2107.01776*, 2021. 2, 3, 4, 5

[47] 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. 7

[48] 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. 1, 2

[49] 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. 2, 3, 4, 6

[50] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3589–3599, 2021. 2

[51] 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

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

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

[54] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 5

[55] 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
[56] Quang Pham, Chenghao Liu, and Steven Hoi. Dualnet: Continual learning, fast and slow. *Advances in Neural Information Processing Systems*, 34:16131–16144, 2021. 2
[57] Mahardhika Pratama, Andri Ashfahani, and Edwin Lughofer. Unsupervised continual learning via self-adaptive deep clustering approach. In *International Workshop on Continual Semi-Supervised Learning*, pages 48–61. Springer, 2022. 1, 3, 4
[58] Senthil Purushwalkam, Pedro Morgado, and Abhinav Gupta. The challenges of continuous self-supervised learning. pages 702–721, 2022. 6, 8
[59] 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
[60] Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Lifelong generative modeling. *Neurocomputing*, 404:381–400, 2020. 2
[61] 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, 4, 6
[62] 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
[63] 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
[64] Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31, 2018. 2
[65] 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
[66] Shaul Salomon, Gideon Avigad, Alex Goldvard, and Oliver Schütze. Psa–a new scalable space partition based selection algorithm for moeas. In *EVOLVE-A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation II*, pages 137–151. Springer, 2013. 6
[67] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021. 2, 6
[68] 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. 3, 6
[69] Abu Md Niamul Taufique, Chowdhury Sadman Jahan, and Andreas Savakis. Unsupervised continual learning for gradually varying domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3740–3750, 2022. 2
[70] Matteo Tiezzi, Simone Marullo, Lapo Faggi, Enrico Meloni, Alessandro Betti, and Stefano Melacci. Stochastic coherence over attention trajectory for continuous learning in video streams. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3480–3486. International Joint Conferences on Artificial Intelligence Organization, 2022. 1, 3
[71] 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
[72] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6
[73] Johannes Von Oswald, Christian Henning, João Sacramento, and Benjamin F Grewe. Continual learning with hypernetworks. 2020. 2
[74] 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
[75] 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
[76] 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
[77] 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. 2, 3
[78] Fei Ye and Adrian G Bors. Task-free continual learning via online discrepancy distance learning. *arXiv preprint arXiv:2210.06579*, 2022. 1, 2
[79] 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
[80] 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. 2, 4
[81] 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

[82] 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