The Challenges of Continuous Self-Supervised Learning

Senthil Purushwalkam[1⋆], Pedro Morgado[1,2⋆], and Abhinav Gupta[1]

[1] Carnegie Mellon University
[2] University of Wisconsin-Madison
www.senthilpurushwalkam.com/publication/continuousssl/

Abstract. Self-supervised learning (SSL) aims to eliminate one of the major bottlenecks in representation learning - the need for human annotations. As a result, SSL holds the promise to learn representations from data in-the-wild, i.e., without the need for finite and static datasets. Instead, SSL should exploit the continuous stream of data being generated on the internet or by agents exploring their environments. In this work, we investigate whether traditional self-supervised learning approaches would be effective deployed in-the-wild by conducting experiments on the *continuous self-supervised learning problem*. In this setup, models should learn from a continuous (infinite) non-IID data stream that follows a non-stationary distribution of visual concepts. The goal is to learn representations that are robust, adaptive yet not forgetful of concepts seen in the past. We show that a direct application of current methods to continuous SSL is 1) inefficient both computationally and in the amount of data required, 2) leads to inferior representations due to temporal correlations (non-IID data) in the streaming sources and 3) exhibits signs of catastrophic forgetting when trained on sources with non-stationary data distributions. We study the use of replay buffers to alleviate the issues of inefficiency and temporal correlations, and enhance them by actively maintaining the least redundant samples in the buffer. We show that minimum redundancy (MinRed) buffers allow us to learn effective representations even in the most challenging streaming scenarios (*e.g.*, sequential frames obtained from a single embodied agent), and alleviates the problem of catastrophic forgetting.

1 Introduction

We are witnessing yet another paradigm shift in the field of computer vision: from supervised to self-supervised learning (SSL). This shift promises to unleash the true potential of data, as we are no longer bound by the cost of manual labeling. Unsurprisingly, recent work has begun to scale current methods to extremely large datasets of up to 1 billion images [8, 9, 24–26] with the hope of learning better representations. In this paper, we pose the question: *Are we ready to deploy SSL in-the-wild to harness the full potential of unlimited data?*

⋆ Equal contribution.

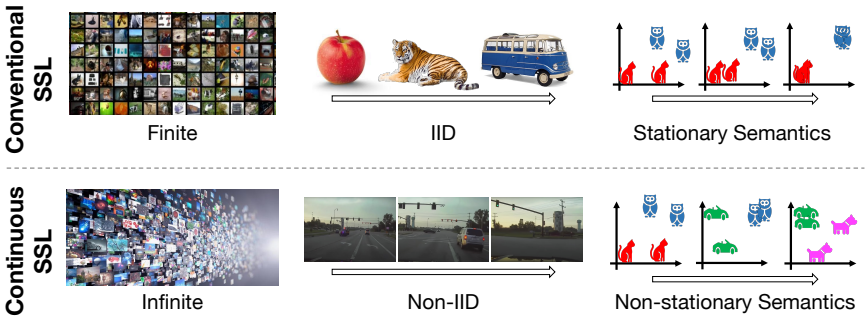


Fig. 1: Conventional vs. Continuous Self-Supervised Learning. The conventional setup of fixed datasets for SSL violates key properties exhibited by data continuously gathered in-the-wild: infinite, non-IID and non-stationary semantics. Hence, the conventional setup serves as a poor benchmark for SSL methods that aim to be deployed in-the-wild. In this work, we introduce the problem of continuous self-supervised learning to facilitate the evaluation of such methods and expose novel challenges. Image Credit: Image of bus was taken from https://i.imgur.com/XtY1cTV.jpg

While SSL promises to exploit the infinite stream of data generated on the internet or by a robotic agent, current practices in SSL still rely on the traditional dataset setup. Images and videos are accumulated to create a training corpus, followed by optimization on hundreds of shuffled passes through the data. The primary reason for working with datasets is the need for reproducible benchmarks, but one question remains: is this traditional static learning setup right for benchmarking self-supervised learning? Does this setup accurately reflect the challenges of a self-supervised system deployed in the wild? We believe the answer is NO. For example, consider a self-supervised system attempting to learn representations of cars over the years from the web. Current setups only evaluate static learning and do not evaluate the ability to adapt representations to new car models (and not forget old ones). Another example is to consider a deployed robotic self-supervised learning agent that actively collects frames from its video feed. This data is heavily structured and correlated due to temporal coherence. However, existing SSL benchmarks do not reflect this challenge since they rely on datasets that can be randomly sampled to produce *IID* samples.

In this paper, we move past dataset-driven SSL and investigate the efficacy of existing methods on the **Continuous Self-Supervised Learning** problem. We explore the challenges faced in two possible deployment scenarios: (a) an internet-based SSL model which relies on continuously acquired images/videos; (b) an agent-based SSL system that learns directly from an agent's sensors. As summarized in Figure 1, both settings rely on a streaming data source that continuously generates new data, presenting three unique challenges that should be reflected when benchmarking SSL approaches.

First, storing infinite data is infeasible and obtaining data in the wild often incurs a cost due to bandwidth or sensor speed limitations. As a result, epoch-based

training is impossible, and a naive deployment of conventional SSL approaches, using each sample only once, would lead to inefficient learners, often waiting for data to be made available, while under-utilizing the data at its disposal. One solution is to rely on replay buffers to decouple data acquisition from the training pipeline. The first question we pose is how effective replay mechanism are at allowing representations to continue to improve while data is being collected?

Second, streaming data sources cannot be "shuffled" to create mini-batches of IID samples. Instead, the ordering of samples is dictated by the source itself. This creates challenges for conventional representation learning approaches, as training data is not necessarily IID. Hence, we pose the question of how to adapt existing SSL methods to learn under non-IID conditions?

Third, real-world data is non-stationary. For example, a higher number of football-related images are seen during the world cup. Also, robots exploring indoor environments observe temporally clustered semantic distributions - a sequence of bedroom objects, followed by kitchen objects, and so on. An intelligent lifelong learning system should be able to continuously learn new concepts without forgetting old ones from non-stationary data distributions. However, we show empirically that conventional contrastive learning approaches can overfit to the current distribution, displaying signs of forgetting. We thus pose the question of how to design SSL methods that learn under non-stationary conditions?

Overall, the main contributions of this work are the following. We identify three critical challenges that arise in the continuous self-supervised learning setup, namely, training efficiency, robustness to non-IID data streams and learning under non-stationary semantic distributions. For each challenge, we construct a curated data stream that simulates this challenge and quantitatively demonstrates the shortcomings of existing SSL methods. We also propose initial solutions to these problems, with the goal of encouraging further research along these directions. We explore the idea of Buffered SSL, which involves augmenting existing approaches with a *replay buffer* to improve training efficiency. Second, we propose a novel method to handle non-IID data streams by minimizing redundancy among stored samples. Finally, we show that *minimum redundancy buffers* prevent forgetting and improve continual learning under non-stationary data distributions.

2 Related Work

Self-supervised visual representation learning is a mature research area, capable of producing models that outperform fully supervised methods when transferred to a variety of downstream tasks [9, 13, 27, 29]. Despite not relying on labeled data, these methods are still trained on fixed-size curated datasets originally developed for the supervised setting. This paper explores the various challenges of deploying self-supervised learning systems truly in-the-wild.

Self-supervised learning has a long history in computer vision [7, 35, 42, 50, 70, 71] aiming to learn representations of visual data by solving tasks that can be defined without human annotations. A breadth of methodologies has been proposed from generative models such as denoising auto-encoders [78], sparse

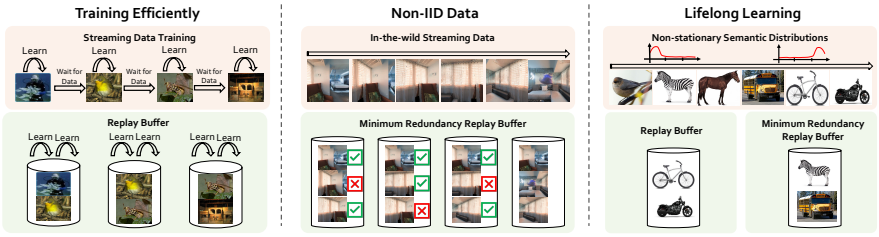


Fig. 2: Overview: We investigate continuous self-supervised learning, exposing three challenges faced by SSL methods deployed in-the-wild. First, representations should be learned in a single pass, as streaming sources do not repeat data samples. We show that augmenting an existing SSL method [13] with replay buffers can significantly alleviate data and computational inefficiencies of training in a single pass. Second, data gathered continuously in-the-wild is often temporally correlated and non IID. We show that actively maintaining minimally redundant samples in the replay buffers yields less correlated training data. Finally, semantic distributions of data gathered in-the-wild are non-stationary. This can cause models to "forget" concepts seen in past distributions. We show that minimizing redundancy also mitigates "forgetting" by focusing on unique samples from various semantic groups.

coding [36, 54, 55], inpainting [59] and colorization [18, 34, 86], to methods that learn representations predictive of spatial context [19, 22, 53], temporal context [21, 47, 58, 61, 79, 80], or concurrent modalities like audio [4, 52, 57, 70], text [17, 23, 62] or speech [43, 44]. One successful approach is to learn transformation invariant representations [12, 20, 28, 29, 46, 56, 65, 81]. After relentless improvements in image augmentations [12, 46], backbone models [10, 24], stable (slow-moving) learning targets [9, 15, 29], and transformation invariant loss functions [10, 13, 27, 56, 84], augmentation invariance has produced impressive models that improve state-of-the-art on a diverse set of downstream tasks like recognition [9, 10], detection [29] and video object segmentation [10].

Given its success, a few attempts have been made to scale SSL to large uncurated datasets, such as YFCC-100M [8, 26] and Instagram-1B [9, 24]. Goyal *et al.* [26] showed that tasks such as colorization [86], context prediction [53] and rotation [22] have diminishing returns on large datasets, due to the low complexity of the task, and argued for the development of more complex tasks. Transformation invariance objectives, coupled with heavy data augmentations, have increased the task's complexity substantially. As a result, recent attempts of scaling up augmentation invariance [9, 24, 25] have seen some performance gains. However, we argue that these methods are still not ready to be deployed truly in-the-wild. Beyond the difficulties of training on uncurated data, already studied in prior work [9, 24], training on fixed datasets ignores important challenges of streaming data, such as the non-iid nature of streaming sources, data acquisition costs, and model saturation due to its fixed capacity.

Continual and lifelong learning: The ability to continuously learn new concepts or tasks over time is often referred to as lifelong learning [75] or never-

ending learning [14, 48]. Lifelong learning has traditionally been studied in supervised and reinforcement learning settings. In both cases, the model is expected to learn distinct tasks presented sequentially, without forgetting previous ones [32, 38, 64, 76, 85]. However, these works usually assume access to full supervision (class labels or external rewards) not available in the wild.

Techniques developed for supervised continual learning are nevertheless useful for the Continuous SSL problem. Rehearsal techniques [2, 6, 60, 67, 68, 73] store and replay a small set of training samples from previous tasks to avoid forgetting previously learned skills or concepts. While there is no notion of well-defined tasks in Continuous SSL, we show that replay buffers help improve training efficiency. We also propose replay buffers that minimize the redundancy of stored memories to decorrelate highly correlated streaming sources. Beyond rehearsal techniques, expandable models [69, 83] have also been used to reduce catastrophic forgetting. This is often accomplished either by progressively growing the model each time a new task is added [37, 69, 83], or adapting a common backbone model to each task separately using small task-specific adaptation blocks [41, 51, 67]. The lack of well-defined tasks in streaming SSL makes lifelong learning more challenging, as it needs to learn from data distributions that may shift over time.

Lifelong Generative Models: Discriminative self-supervised representation learning has never been investigated in the continuous learning setup (streaming, non-IID and non-stationary data). However, recent works [1, 63, 66, 82] have attempted to address a sub-problem of ours, *i.e.*, learning self-supervised representations using generative models in a continual learning setting where the domain of data exhibits significant shifts during training. These works present approaches to locate domain shifts in order to avoid the problem of catastrophic forgetting. These techniques are made possible by the fact that training data is constructed by collecting samples from images in significantly different datasets - for example, [82] uses Celeb-A[40] faces followed by 3D-Chair[5] images). In contrast, we consider a more realistic setting of ImageNet images with a smoothly changing distribution of classes. Furthermore, as highlighted above, these works do not address other critical challenges of deploying SSL in-the-wild, as they are limited to epoch-based optimization, do not consider non-curated and/or high correlated streaming sources, data efficiency, or the issue of early convergence.

3 Problem Setup and Challenges

Our goal is to investigate the efficacy of self-supervised representation learning on a naturally occurring source of streaming data, which we refer to as the *continuous self-supervised learning problem*. First, we describe the distinction between conventional training and the continuous self-supervised learning setup. We then discuss the various unique challenges that appear in the continuous case.

3.1 Streaming vs Conventional Self-Supervised Learning

Existing self-supervised learning methods rely on fixed-size datasets. These datasets $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are finite (*i.e.*, $N << \infty$), immutable (*i.e.*, $\mathcal{D}$ does

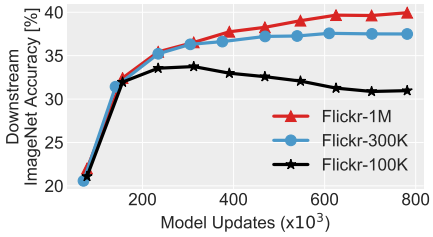


Fig. 3: ImageNet downstream accuracy of a SimSiam model trained on datasets of different sizes with a ResNet-18 backbone.

not change) and readily available (*i.e.*, all its samples $\mathbf{x}_i$ can be easily accessed at all times). Due to these properties, samples can be indexed, shuffled, and accessed at any point in training. Conventional SSL takes advantage of these possibilities by iterating over the datasets multiple times (epochs).

In contrast, Continuous SSL relies on a *streaming source* $\mathcal{S}$, defined as a time-series of unlabeled sensory data $\mathcal{S} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$, potentially of infinite length $T \to \infty$. At any given moment in time t, fetching data from a streaming source $\mathcal{S}$ yields the current sample $\mathbf{x}_t$. Future samples $\{\mathbf{x}_\tau \forall \tau > t\}$ are not accessible at time t, and past samples $\{\mathbf{x}_\tau \forall \tau < t\}$ are only accessible if stored when fetched.

In the Continuous SSL setup, one important parameter is the ratio between the data loading time t_{data} and the time taken to perform one optimization step t_{opt}. In most deployment setups $t_{\text{data}} > t_{\text{opt}}$, due to slower data transfer speed or low sensor frame rates. Therefore, even with parallelization, optimization algorithms can wait idle for $t_{\text{idle}} = t_{\text{data}} - t_{\text{opt}}$. Therefore, SSL methods developed for the continuous setup should be able to efficiently and continually build better representations, while training on samples obtained from a streaming source.

3.2 Why Continuous SSL? Does scaling the number of unique images help representation learning?

To understand the effect of increasing the scale of training data (potentially to infinite), we indexed all Creative Commons images uploaded to the photo-sharing website Flickr.com between 2008 and 2021. We then used this index to create datasets of varying sizes, and train visual representations through self-supervision over multiple epochs in the Conventional SSL setup.

We adopt SimSiam [13] as a prototypical example of contrastive learning methods, which have been shown to be effective for Conventional SSL. SimSiam learns representations by optimizing the augmentation invariance loss

$$\mathcal{L}(x_1, x_2) = -\texttt{sg}(\mathbf{z}_1)^T g(\mathbf{z}_2) - \texttt{sg}(\mathbf{z}_2)^T g(\mathbf{z}_1) \tag{1}$$

where x_1 and x_2 are two random transformations of an image x, $\mathbf{z}_i = f(x_i)$ is the model output representations, $\texttt{sg}(\cdot)$ the stop gradient and $g(\cdot)$ a prediction head. Refer to [13] for full details. Figure 3 shows the linear classification accuracy on ImageNet for models trained on different datasets as a function of the number of model updates. Unsurprisingly, training with more diverse data leads to better

representations. This highlights the benefits of scaling *unique* images, which Continuous SSL will take to the extreme.

3.3 Challenges of Continuous SSL

Streaming sources available in the wild do not allow revisiting past samples. Since storing the full data stream is infeasible due to the potentially infinite length, Continuous SSL methods should learn representations in a **single pass** over the data (instead of learning over multiple epochs). This setup poses novel challenges that Conventional SSL methods do not face.

Computational and Data Efficiency: Sampling data from streaming sources in the real world can be significantly slower (when compared to sampling from static datasets) due to sensor frame rates or bandwidth limitations. Thus, current SSL approaches could become inefficient learners, as optimization algorithms may have to wait idly while waiting for new data to be made available, while under-utilizing the data at their disposal.

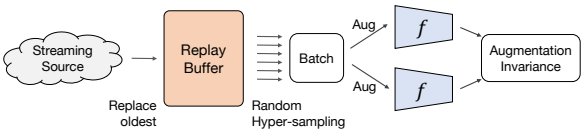
Correlated Samples: Many streaming sources in the wild exhibit temporal coherence. For example, consecutive frames from online videos or from a robot exploring its environment display minimal changes. Such correlations break the *IID* assumption on which conventional optimization algorithms rely.

Lifelong learning: Access to infinite streams of data provides us the opportunity to continuously improve visual representations. However, the non-stationary nature of data streams in the wild cause conventional SSL methods to quickly forget features that are no longer relevant for the current distribution. Continuous SSL methods should therefore be able to integrate new concepts in their representations without forgetting previously learned ones.

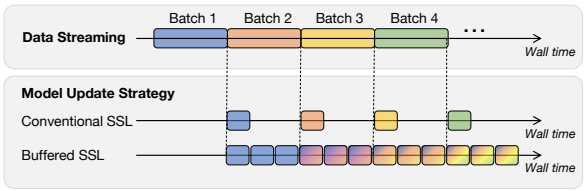
While all these challenges co-exist in the wild, our goal is to analyze each one comprehensively and in isolation. Therefore, we disentangled each challenge by designing a set of data streams that highlight each problem separately, and assess its effect on existing SSL methods. This helps us building a thorough characterization of each challenge and inform us on how to tackle them. We believe a disentangled analysis will help the community build intuitions about the impact of each challenge on continuous SSL as a whole. Section §4 introduces the challenge of one pass training and computational efficiency. Section §5 introduces the non-iid data setup, and Section §6 analyses the lifelong learning setting.

4 Efficient Training

Computational and data efficiency are two challenges that currently prevent SSL from being deployed on continuous data streams in-the-wild. For most practical applications, $t_{\text{data}} : t_{\text{optim}}$ might be high, so SSL methods should use idle time to improve the models. Second, fetching new samples can still be costly. For example, exploration robots often run on batteries, and web crawlers are limited by network bandwidths. Trivially deploying current SSL methods to the streaming setup would discard each batch of data after being used once. However,



(a) Overview of Buffered SSL



(b) Optimization under limited streaming bandwidth.

Fig. 4: Buffered Self-Supervised Learning. Buffered SSL introduces a replay buffer, which allows the model to continuously train even under limited bandwidth settings.

current deep learning optimization practices show that iterating over the same samples over multiple epochs helps learn better representations. For example, supervised learning on ImageNet [30, 33] iterates over the dataset 100 times, and SSL approaches [12] have been shown to keep improving even after seeing each sample 800 times. Therefore, we would like to answer the question of how to improve data efficiency while still following the streaming setting.

4.1 Buffered Self-Supervised Learning

We present a simple solution to the challenges above. The key idea is to maintain a fixed-size *replay buffer* that stores a small number of recent samples. This idea is inspired by experience replay [39] commonly used in reinforcement learning [3, 49, 72] and supervised continual learning [31, 68]. As shown in Figure 4a, the replay buffer decouples the streaming source from the training pipeline. The streaming data can be added to the replay buffer when available, replacing the oldest samples (*i.e.* first-in-first-out (FIFO) update rule). Simultaneously, mini-batches of training data can be generated at any time by randomly sampling from the buffer. As shown in Figure 4b, replay buffers allow us to continue training during the otherwise idle wait time t_{idle}. Replay buffers also allow us to reuse samples by sampling them multiple times, hence reducing the total data cost. We refer to this approach as *Buffered Self-Supervised Learning*.

4.2 Single-pass training experiments

We study the effectiveness of replay buffers when training with a single pass of the data. We trained ResNet-18 SimSiam models with and without replay buffers, with various amounts of idle time $t_{\text{idle}} = t_{\text{data}} - t_{\text{optim}}$. All models were trained using the first 20 million images in our Flickr index as the streaming source.

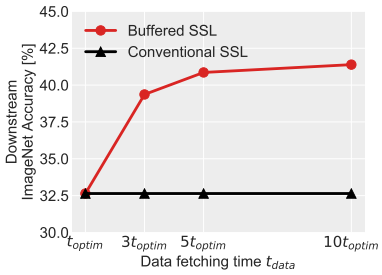


Fig. 5: Streaming SSL with limited bandwidth. Comparison of buffered and non-buffered approaches for various limited bandwidth settings. $t_{\text{data}} : t_{\text{optim}}$ denotes the ratio of data acquisition time to the optimization time. Buffered SSL can take advantage of the idle time to effectively improve the learned representations instead of waiting idly for new data.

Table 1: Data Efficiency: Augmenting SSL methods with replay buffers can improve efficiency allowing us to train on data streams with one pass. We show that Buffered SSL methods outperform the Conventional SSL methods and achieve performances close to training for multiple epochs.

	Epochs	Hyper Sampling	Memory Size	ImageNet Top1 Acc	iNaturalist Top1 Acc
Training DB: Flickr 20M					
Conventional SSL	1	-	-	32.2	12.2
Buffered SSL	1	10	16K	41.4	16.7
Buffered SSL	1	10	64K	41.8	17.4
Buffered SSL	1	10	256K	41.5	17.5
Epoch-based SSL*	10	-	-	41.7	17.3
Training DB: Flickr 5M					
Conventional SSL	1	-	-	14.5	2.8
Buffered SSL	1	40	16K	39.9	16.1
Buffered SSL	1	40	64K	41.0	17.1
Buffered SSL	1	40	256K	41.5	17.3
Epoch-based SSL*	40	-	-	41.8	17.0
Training DB: Flickr 1M					
Conventional SSL	1	-	-	8.0	1.5
Buffered SSL	1	200	16K	30.5	9.5
Buffered SSL	1	200	64K	36.4	14.3
Buffered SSL	1	200	256K	38.8	15.5
Epoch-based SSL*	200	-	-	41.7	17.3

** Epoch-based SSL violates the streaming setting (reference only).*

Figure 5 shows the ImageNet linear classification performance for increasing t_{data}. By maintaining a small replay buffer (containing only the most recent 64k images), Buffered SSL was able to make good use of the idle time and improve representations significantly (41.4% accuracy on ImageNet) over the bottlenecked Conventional SSL approach (32.5% ImageNet accuracy). Replay buffers also improve data efficiency in the Continuous SSL setup, as each sample can be reused multiple times. Data usage is proportional to the hyper-sampling rate K, defined as the ratio between the number of mini-batches generated for training and acquired from the streaming source.

To understand the limits of hyper-sampling, we trained a ResNet-18 SimSiam model with a replay buffer for a fixed amount of updates (780 000 iterations). Table 1 shows a comparison of Buffered SSL at varying hyper-sampling rates K,

to Conventional SSL trained on the same amount of data, and Epoch-based SSL methods trained for K epochs. Epoch-based SSL and Buffered SSL are optimized with the same number of updates, but the former violates the streaming setup. Despite being required to train on a single pass of the data, Buffered SSL with a hyper-sampling rate of $K = 10$ achieved similar performance to epoch-based training, even for buffers as small as 64K images (0.3% of the 20M unique images seen). Table 1 also shows that, as hyper-sampling rates increase, the size of the replay buffer becomes critical. For example, for $K = 200$, Buffered SSL still improves significantly over Conventional SSL on the same amount of data, regardless of buffer size. However, better representations are learned as the buffer size increases. Since, in high hyper-sampling regimes, the buffer is updated slowly with new images from the streaming source, increasing the buffer size prevents the model from quickly overfitting to the samples in the buffer.

5 Correlated Data Sources

Visual data obtained in-the-wild is often correlated and non-*IID*. For example, video feed from a self-driving car collects very similar consecutive frames. This is in stark contrast to the data used in Conventional SSL methods. For example, the ImageNet dataset allows sampling images from a collection of 1000 uniformly distributed object classes. Even methods trained on larger datasets like Instagram-1B [24, 26] are less likely to encounter heavily correlated samples in the mini-batches. However, the constant flow of data in the Continuous SSL setup generally violates these assumptions even in the static image setup (images uploaded near events are likely to be highly correlated).

Let $(x_i : i \in \mathcal{D})$ be a sequence of samples. When x_i is generated by randomly sampling from a large dataset, samples are close to IID. Hence, the probability p_c that two samples x_i and x_j are highly correlated is low, $p_c \approx 0$. Correlated samples may indicate images that are visually very similar or visually dissimilar but depict similar semantic content. However, in the Continuous SSL setup, the IID assumption is generally violated, leading to $p_c >> 0$. Under the assumption that consecutive samples in a continuous stream of data have the same correlation probability p_c, the likelihood of a random pair in a batch $(x_i, \ldots, x_{i+b})$ of size b being correlated (*correlation likelihood*) is large, and given by

$$\mathcal{L}_{\text{Seq}} = P_c(b, p_c) = \frac{2}{b(b-1)} \sum_{i=1}^{b-1} \sum_{j=i+1}^{b} p_c^{j-1} = \frac{2p_c}{b(b-1)} \Big(\frac{p_c^b - 1}{(1-p_c)^2} + b \frac{p_c}{1-p_c} \Big). \quad (2)$$

Introducing a replay buffer of size $B >> b$, as proposed in §4.1, lowers the correlation likelihood to $\mathcal{L}_{\text{FIFO}} = P_c(B, p_c) \approx \frac{b}{B} \mathcal{L}_{\text{Seq}} < P_c(b, p_c)$[3], and enables more effective representation learning.

5.1 Minimum Redundancy Replay Buffer

While replay buffers are able to reduce the correlation likelihood, prohibitively large replay buffers ($B >> b$) are required to significantly lower $\mathcal{L}_{\text{FIFO}}$ in heavily

[3] Approximation holds for large values of B and b, and $p_c \neq 1$.

correlated setups ($p_c \approx 1$). In order to overcome this, we propose a modified replay buffer to only retain de-correlated samples, thereby actively reducing p_c. We call this the Minimum Redundancy Replay Buffer (MinRed).

To accomplish this, we rely on the learned embedding space to identify redundant samples. Consider a replay buffer $\mathcal{B}$ with a maximum capacity of B, already containing B samples with representation $\bar{\mathbf{z}}_i$. To add a new sample x to $\mathcal{B}$, we rely on the cosine distance between all pairs of samples to discard the most redundant:

$$\mathcal{B} \leftarrow \mathcal{B} \backslash i^* \cup \{x\} \quad \text{where} \quad i^* = \arg\min_{i \in \mathcal{B}} \min_{j \in \mathcal{B}} d_{\cos}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j). \tag{3}$$

In other words, we discard the sample with minimum distance to its nearest neighbor. To represent instances, we track the features $\bar{\mathbf{z}}_i$ of all samples in the buffer using a moving average $\bar{\mathbf{z}}_i = \alpha \bar{\mathbf{z}}_i + (1-\alpha)\mathbf{z}_i$, where $\mathbf{z}_i = f(\mathbf{x}_i)$ is the current feature of the i^{th} sample, and α the moving average coefficient. Since redundant samples are dropped from the buffer, the probability p_c of two consecutive samples in the buffer being correlated decreases. If this probability decreases from p_c to ηp_c where $\eta << 1$, the correlation likelihood is lowered to $\mathcal{L}_{\text{MinRed}} = P_c(B, \eta p_c) < P_c(B, p_c)$, which facilitates representation learning.

5.2 Experiments with non-IID data streams

We assess the performance of SSL methods on two data streams with heavy temporal coherence. The first is created by concatenating video frames from the Kinetics dataset [11]. From each video, we sample N_{seq} frames at random and add them sequentially to the data stream. The second data stream is composed by consecutive frames from the KrishnaCAM dataset[4] [74] which records egocentric videos spanning nine months in the life of a computer vision graduate student. On each stream, we train the baseline SimSiam (Conventional SSL), SimSiam augmented with replay buffers (Buffered SSL) and SimSiam augmented with MinRed buffers (Buffered SSL (MinRed)). We evaluate these representations by training a linear classifier on ImageNet [16] and iNaturalist [77]. Results are shown in Table 2. We observe that the correlated nature of the data heavily disrupts training of conventional models. While regular replay buffers alleviate the issue to some extent, the learned representations still suffer when trained on heavy correlated data streams (as in Kinetics $N_{\text{seq}} = 64$ and KrishaCAM). Finally, MinRed buffers are very effective in these setups, learning representations that perform similarly to the "oracle" IID setting (*i.e.*, by randomly sampling from the collection of all frames from all videos, violating the streaming assumption).

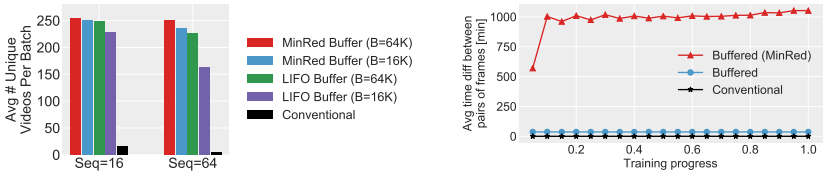
Correlation of training samples: One of the benefits of Buffered SSL is the ability to generate training samples with low correlation likelihood and thus closer to *IID*. We analyzed the contents of the replay buffer during training to track the correlation likelihood (see Figure 6). We confirmed that the contents of MinRed replay buffers are significantly less correlated than FIFO buffers. In

[4] Concatenated videos are looped over 10 times to create a larger stream.

Table 2: Visually Correlated SSL: Linear classification performance of buffered and unbuffered SimSiam representations trained on data sources with high temporal coherence. MinRed buffers learns better representations by decorrelating the data.

	Epochs	Hyper Sampling	Memory Size	ImageNet Top1 Acc	iNaturalist Top1 Acc
Streaming source: Kinetics (N_{seq}=16)					
Conventional SSL	5	-	-	17.7	3.0
Buffered SSL	1	5	64K	25.9	8.4
Buffered SSL (MinRed)	1	5	64K	26.2	7.9
Decorrelated source*	5	-	-	25.9	7.9
Streaming source: Kinetics (N_{seq}=64)					
Conventional SSL	5	-	-	8.0	0.8
Buffered SSL	1	5	64K	9.8	0.8
Buffered SSL (MinRed)	1	5	64K	30.6	9.6
Decorrelated source*	5	-	-	31.4	10.6
Streaming source: Krishna CAM					
Conventional SSL	5	-	-	0.4	0.03
Buffered SSL	1	5	16K	0.5	0.05
Buffered SSL (MinRed)	1	5	16K	15.2	3.43
Buffered SSL	1	5	64K	1.7	0.07
Buffered SSL (MinRed)	1	5	64K	17.9	5.91
Decorrelated source*	5	-	-	19.2	6.94

** Decorrelated sources violate the streaming setting (reference only).*



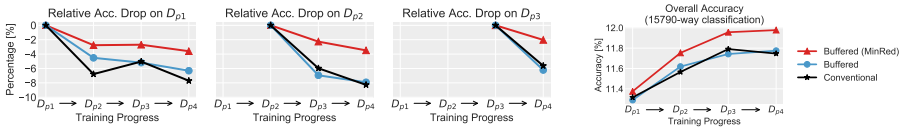
(a) Kinetics **(b)** Krishna CAM

Fig. 6: Estimate of within batch correlation while training w/ and w/o replay buffers.

KrishnaCAM, MinRed buffers tend to maintain memories of past unique frames for longer periods of time. In Kinetics, MinRed buffers also yield mini-batches with frames from a larger number of unique videos.

6 Lifelong Self-Supervised Learning

As we explore the world, we encounter different distributions of object classes, some previously seen and some unseen. For example, we see furniture and appliances every day. But we also encounter novel concepts like zebras when we visit a zoo. This suggests that the distribution of semantic classes is often correlated in time with occasional changes in distribution. However, Conventional SSL methods learn from a limited vocabulary of concepts that is repeatedly seen thousands of times (often uniformly). This provides a simplification of the learning setup that does not reflect the non-stationary nature of concepts in-the-wild.



(a) *Forgetting* **(b)** *Generalization*

Fig. 7: Continual unsupervised representation learning on full ImageNet (14M images). The dataset is partitioned in 4 separate tasks which are seen in sequence $D_{p1} \rightarrow D_{p2} \rightarrow D_{p3} \rightarrow D_{p4}$. Forgetting (7a) is measured by computing the relative accuracy drop on each task after training on data of the task itself. Generalization (7b) is measured as the overall accuracy across all 15790 full ImageNet classes. All results are averaged over 3 different sequences p_i.

6.1 A non-stationary data stream to benchmark SSL

To evaluate deployable SSL, we must use benchmarks that simulate the non-stationary semantic distributions encountered in-the-wild. Inspired by supervised continual learning [32, 38], we created a stream with smooth shifting semantics.

First, we create four datasets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4$ by splitting the classes of the ImageNet-21K dataset [16]. We create the splits based on the Wordnet [45] hierarchy such that each $\mathcal{D}_i$ contains images from semantically similar classes. For each class, we hold out 25 images per class for evaluation. The training stream is created by sampling images from the four splits $\{\mathcal{D}_{p_1}, \mathcal{D}_{p_2}, \mathcal{D}_{p_3}, \mathcal{D}_{p_4}\}$[5] such that images from $\mathcal{D}_{p_i}$ are seen only after most images of $\mathcal{D}_{p_{i-1}}$ are sampled (see Appendix for a detailed description of the sampling procedure), simulating a smooth change in semantic distribution.

6.2 Experiments with non-stationary distributions

We train representations using conventional SimSiam, SimSiam with replay buffers (§4.1) and SimSiam with minimum redundancy buffers (§5.1) on a single pass of this stream of data. These models are initialized from SimSiam trained on Flickr200M. During evaluation, we train a linear classifier on the learned representations to recognize all classes in the ImageNet-21k dataset, and measure the accuracy on the held-out set of each $\mathcal{D}_{p_i}$ separately. All results are averaged over 3 permutations of $p_1, \ldots, p_4$.

Figure 7a shows the drop in classification accuracy on each dataset D_{p_i} after the representation is trained on new data $D_{p_{i+1}}, D_{p_{i+2}}$, *etc.*, relative to the initial accuracy at the end of D_{p_i}. This serves as a measure of forgetting - a larger drop indicates that the representation lost its ability to discriminate older classes. As can be seen, all methods suffer from forgetting. However, SimSiam with MinRed buffers displays less forgetting compared to conventional and buffered SimSiam. Intuitively, this can be attributed to the MinRed criteria which naturally retains instances from previous semantic distributions. Figure 7b also depicts the accuracy

[5] $[p_1, p_2, p_3, p_4]$ denotes a permutation of the sequence $[1, 2, 3, 4]$.

on all classes as training progresses, showing that SimSiam with MinRed buffers also yields slightly better overall generalization, consistently throughout training. In supplementary material, we also evaluated the learned representations on unseen classes, by testing only on future data streams $D_{p_{i+t}}$. Since MinRed buffers maintain training buffers with wider coverage of semantics, the learned representations were also shown to generalize better to unseen concepts.

7 Discussion and Future Work

In this work, we exposed three challenges that require investigation to build robust deployable self-supervised learners. We improve the efficiency of Continuous SSL by leveraging replay buffers to revisit old samples. In future work, developing approaches for quickly rejecting samples by preemptively evaluating their value might yield improved data efficiency. We also propose a novel minimum redundancy buffer to discard correlated samples allowing us to mimic the generation of IID training data, even in highly correlated settings. An alternative future direction could focus on learning representations that take advantage of the correlated nature of the data stream to learn from fine-grained discrepancies.

In data streams with non-stationary semantic distributions, we show that MinRed buffers alleviate the issue of catastrophic forgetting, as they are capable of maintaining unique samples from past distributions. However, we observed signs of saturating generalization as new concepts are introduced. Some possible reasons could be: 1) the cosine decay learning rate schedule and 2) the fixed capacity of our models that prohibits learning a large sequence of novel concepts. In preliminary experiments (see supplementary material), we saw that training with a constant learning rate (on 100M images from Flickr) does not lead to significant improvements in performance. We also observed that trivially expanding the architecture at regular intervals does not lead to noticeable improvements. However, we believe that further exploration in this direction is required to continually learning novel concepts in a self-supervised manner.

8 Conclusion

One of the grand goals of self-supervised learning is to build systems capable of continually learning from unlimited sources of unlabelled data. However, due to the need for benchmarking, existing SSL methods have primarily focused on curated datasets of limited size. Unfortunately, while the existing approaches work well in the dataset setup, we are still not close to deployable continual self-supervised methods. In this work, we advocate for a more realistic SSL setup that will facilitate deployment, while retaining the benefits of benchmarking. To this end, we identified three broad challenges of deployable SSL - training efficiency, correlated data, and lifelong learning, - and proposed potential solutions to address them. We believe however that further research is needed to develop deployable systems that deliver on the promise of self-supervised learning, and hope future efforts in SSL research focus on these challenges.

Bibliography

[1] Achille, A., Eccles, T., Matthey, L., Burgess, C., Watters, N., Lerchner, A., Higgins, I.: Life-long disentangled representation learning with cross-domain latent homologies. Advances in Neural Information Processing Systems **31** (2018) 5

[2] Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. Advances in Neural Information Processing Systems **32**, 11816–11825 (2019) 5

[3] Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., Zaremba, W.: Hindsight experience replay. arXiv preprint arXiv:1707.01495 (2017) 8

[4] Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 609–617 (2017) 4

[5] Aubry, M., Maturana, D., Efros, A.A., Russell, B.C., Sivic, J.: Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3762–3769 (2014) 5

[6] Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. In: Advances in Neural Information Processing Systems (2020) 5

[7] Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 132–149 (2018) 3

[8] Caron, M., Bojanowski, P., Mairal, J., Joulin, A.: Unsupervised pre-training of image features on non-curated data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2959–2968 (2019) 1, 4

[9] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS) (2020) 1, 3, 4

[10] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 4

[11] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 11

[12] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning. pp. 1597–1607. PMLR (2020) 4, 8

[13] Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021) 3, 4, 6
[14] Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge from web data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1409–1416 (2013) 5
[15] Chen*, X., Xie*, S., He, K.: An empirical study of training self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 4
[16] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255. Ieee (2009) 11, 13
[17] Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11162–11173 (2021) 4
[18] Deshpande, A., Rock, J., Forsyth, D.: Learning large-scale automatic image colorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 567–575 (2015) 4
[19] Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1422–1430 (2015) 4
[20] Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans. on Pattern Analysis and Machine Intelligence **38**(9), 1734–1747 (2015) 4
[21] Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3636–3645 (2017) 4
[22] Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018) 4
[23] Gomez, L., Patel, Y., Rusiñol, M., Karatzas, D., Jawahar, C.: Self-supervised learning of visual features through embedding images into text topic spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4230–4239 (2017) 4
[24] Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al.: Self-supervised pretraining of visual features in the wild. arXiv preprint arXiv:2103.01988 (2021) 1, 4, 10
[25] Goyal, P., Duval, Q., Seessel, I., Caron, M., Singh, M., Misra, I., Sagun, L., Joulin, A., Bojanowski, P.: Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv preprint arXiv:2202.08360 (2022) 4
[26] Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6391–6400 (2019) 1, 4, 10

[27] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B., Guo, Z., Azar, M., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems (2020) 3, 4

[28] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1735–1742. IEEE (2006) 4

[29] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020) 3, 4

[30] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 8

[31] Hsu, Y.C., Liu, Y.C., Ramasamy, A., Kira, Z.: Re-evaluating continual learning scenarios: A categorization and case for strong baselines. arXiv preprint arXiv:1810.12488 (2018) 8

[32] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017) 5, 13

[33] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012) 8

[34] Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: Proceedings of the European Conference on Computer Vision. pp. 577–593. Springer (2016) 4

[35] Le, Q.V.: Building high-level features using large scale unsupervised learning. In: Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8595–8598. IEEE (2013) 3

[36] Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems. pp. 801–808 (2007) 4

[37] Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: Proceedings of the International Conference on Machine Learning. pp. 3925–3934. PMLR (2019) 5

[38] Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2017) 5, 13

[39] Lin, L.J.: Reinforcement learning for robots using neural networks. Carnegie Mellon University (1992) 8

[40] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3730–3738 (2015) 5

[41] Mallya, A., Davis, D., Lazebnik, S.: Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: Proceedings of the European Conference on Computer Vision. pp. 67–82 (2018) 5
[42] Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Proceedings of the International Conference on Artificial Neural Networks. pp. 52–59. Springer (2011) 3
[43] Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020) 4
[44] Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640 (2019) 4
[45] Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998) 13
[46] Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717 (2020) 4
[47] Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: Proceedings of the European Conference on Computer Vision. pp. 527–544. Springer (2016) 4
[48] Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., et al.: Never-ending learning. Communications of the ACM **61**(5), 103–115 (2018) 5
[49] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015) 8
[50] Mobahi, H., Collobert, R., Weston, J.: Deep learning from temporal coherence in video. In: Proceedings of the International Conference on Machine Learning. pp. 737–744 (2009) 3
[51] Morgado, P., Vasconcelos, N.: Nettailor: Tuning the architecture, not just the weights. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3044–3054 (2019) 5
[52] Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021) 4
[53] Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Proceedings of the European Conference on Computer Vision. pp. 69–84. Springer (2016) 4
[54] Olshausen, B.A.: Sparse coding of time-varying natural images. In: Proc. of the Int. Conf. on Independent Component Analysis and Blind Source Separation. vol. 2. Citeseer (2000) 4
[55] Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381**(6583), 607–609 (1996) 4

[56] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 4
[57] Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018) 4
[58] Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B.: Learning features by watching objects move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2701–2710 (2017) 4
[59] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2536–2544 (2016) 4
[60] Prabhu, A., Torr, P.H., Dokania, P.K.: Gdumb: A simple approach that questions our progress in continual learning. In: Proceedings of the European Conference on Computer Vision (2020) 5
[61] Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021) 4
[62] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 4
[63] Ramapuram, J., Gregorova, M., Kalousis, A.: Lifelong generative modeling. Neurocomputing **404**, 381–400 (2020) 5
[64] Rannen, A., Aljundi, R., Blaschko, M.B., Tuytelaars, T.: Encoder based lifelong learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1320–1328 (2017) 5
[65] Ranzato, M., Huang, F.J., Boureau, Y.L., LeCun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007) 4
[66] Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y.W., Hadsell, R.: Continual unsupervised representation learning. Advances in Neural Information Processing Systems **32**, 7647–7657 (2019) 5
[67] Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017) 5
[68] Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T.P., Wayne, G.: Experience replay for continual learning. In: Advances in Neural Information Processing Systems. pp. 350–360 (2019) 5, 8
[69] Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016) 5

[70] de Sa, V.R.: Learning classification with unlabeled data. In: Advances in Neural Information Processing Systems. pp. 112–119. Citeseer (1994) 3, 4
[71] Salakhutdinov, R., Hinton, G.: Deep boltzmann machines. In: Artificial Intelligence and Statistics. pp. 448–455. PMLR (2009) 3
[72] Schaul, T., Quan, J., Antonoglou, I., Silver, D.: Prioritized experience replay. In: Proceedings of the International Conference on Learning Representations (2016) 8
[73] Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Advances in Neural Information Processing Systems. pp. 2994–3003 (2017) 5
[74] Singh, K.K., Fatahalian, K., Efros, A.A.: Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016) 11
[75] Thrun, S.: A lifelong learning perspective for mobile robot control. In: Intelligent robots and systems. pp. 201–214. Elsevier (1995) 4
[76] Titsias, M.K., Schwarz, J., Matthews, A.G.d.G., Pascanu, R., Teh, Y.W.: Functional regularisation for continual learning with gaussian processes. arXiv preprint arXiv:1901.11356 (2019) 5
[77] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8769–8778 (2018) 11
[78] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the International Conference on Machine Learning. pp. 1096–1103 (2008) 3
[79] Wang, J., Jiao, J., Liu, Y.H.: Self-supervised video representation learning by pace prediction. In: Proceedings of the European Conference on Computer Vision. pp. 504–521. Springer (2020) 4
[80] Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2802 (2015) 4
[81] Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018) 4
[82] Ye, F., Bors, A.G.: Learning latent representations across multiple data domains using lifelong vaegan. In: Proceedings of the European Conference on Computer Vision. pp. 777–795. Springer (2020) 5
[83] Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. arXiv preprint arXiv:1708.01547 (2017) 5
[84] Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Proceedings of the International Conference on Machine Learning (2021) 4
[85] Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Proceedings of the International Conference on Machine Learning. pp. 3987–3995. PMLR (2017) 5

[86] Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Proceedings of the European Conference on Computer Vision. pp. 649–666. Springer (2016) 4