May the Forgetting Be with You: Alternate Replay for Learning with Noisy Labels

Monica Millunzi[1,2]
monica.millunzi@unimore.it
Lorenzo Bonicelli[1]
lorenzo.bonicelli@unimore.it
Angelo Porrello[1]
angelo.porrello@unimore.it
Jacopo Credi[2]
jacopo.credi@axyon.ai
Petter N. Kolm[3]
petter.kolm@nyu.edu
Simone Calderara[1]
simone.calderara@unimore.it

[1] AImageLab
University of Modena and Reggio Emilia
Modena, Italy

[2] AxyonAI
Viale dell'Autodromo, 210
Modena, Italy

[3] Courant Institute of Mathematical Science
New York University (NYU)
New York, USA

Abstract

Forgetting presents a significant challenge during incremental training, making it particularly demanding for contemporary AI systems to assimilate new knowledge in streaming data environments. To address this issue, most approaches in Continual Learning (CL) rely on the replay of a restricted buffer of past data. However, the presence of noise in real-world scenarios, where human annotation is constrained by time limitations or where data is automatically gathered from the web, frequently renders these strategies vulnerable. In this study, we address the problem of CL under Noisy Labels (CLN) by introducing Alternate Experience Replay (AER), which *takes advantage of forgetting* to maintain a clear distinction between clean, complex, and noisy samples in the memory buffer. The idea is that complex or mislabeled examples, which hardly fit the previously learned data distribution, are most likely to be forgotten. To grasp the benefits of such a separation, we equip AER with Asymmetric Balanced Sampling (ABS): a new sample selection strategy that prioritizes purity on the current task while retaining relevant samples from the past. Through extensive computational comparisons, we demonstrate the effectiveness of our approach in terms of both accuracy and purity of the obtained buffer, resulting in a remarkable average gain of 4.71% points in accuracy with respect to existing loss-based purification strategies. Code is available at https://github.com/aimagelab/mammoth.

1 Introduction

Despite the latest breakthroughs, modern AI still struggles to learn in a continuous fashion and suffers from *catastrophic forgetting* [39], *i.e.* the new knowledge quickly replaces

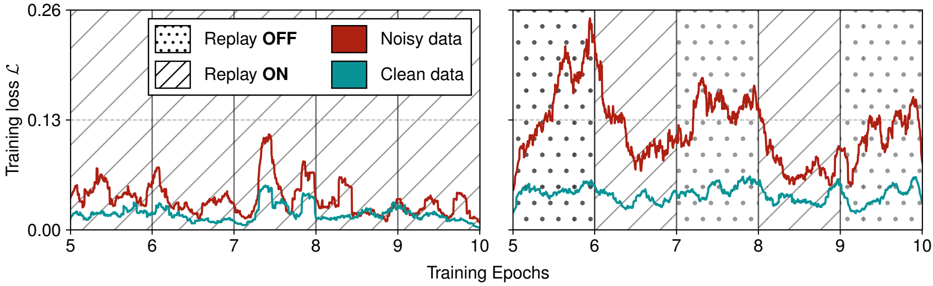

Figure 1: Training loss of clean and noisy during the second task of Seq. CIFAR-10 with 40% noise. The loss is computed on examples from the first task stored in the memory buffer. Standard replay makes the two indistinguishable (*left*) but alternating epochs of replay and forgetting maintain a significant loss separation (*right*).

all past progress. Therefore, Continual Learning (CL) has recently gathered an increasing amount of attention: among the others, one prominent strategy is to interleave examples from the current and old tasks (rehearsal). To do so, a small selection of past data is retained in a memory buffer [17, 55], as in Experience Replay (ER) [47, 49].

Intuitively, the effectiveness of these methods depends strictly on the memory content: the larger the gap between the memory and the true distribution underlying all the previous tasks, the lower the chances of learning a reliable model. In this respect, several factors may intervene and degrade the snapshot portrayed by the buffer. Several works have highlighted the shortcomings of low-capacity buffers and their link to severe overfitting [9, 56]. More recently, the plausible presence of annotation errors has emerged as an engaging factor [6, 30], due to the subsequent poisoning the memory buffer would be subjected to. Indeed, not only would a few observations of past tasks be available for the learner, but they might be erroneously annotated. It is noted that the presence of *noisy annotations* [8, 33, 35, 60] is an inescapable characteristic of CL: to allow the learner to digest incoming training examples on-the-fly, data has to be annotated within a restricted temporal window, leading to poor human annotations and hasty quality controls. In light of this, preliminary works [6, 30] focus on purifying the memory buffer and then consolidate their knowledge at the end of the current task (or while learning the task itself [28]). To do so, they spot clean samples by leveraging the popular *small-loss criterion* [4, 21, 27] and the fact that the most trustworthy examples are those favoured during the first training stages (*memorization effect*), thus exhibit a lower value of the loss function. However, despite its effectiveness in the offline scenario, such a criterion may be weak in incremental settings. Indeed, as learning does not re-start from scratch but builds upon previous knowledge, the adaption is faster and hence the loss-value separation between clean and noisy samples tends to vanish [4, 62].

To overcome this limitation, we explore a radically different approach which could be summarized by a quote ascribed to Julius Caesar "if you can't defeat your enemy, have him as your friend": while existing methods see forgetting only as an issue to solve, we use it to identify noisy examples within the data stream. We build upon the work of [38, 54], which theoretically demonstrates that mislabeled examples are quickly forgotten, whereas complex or rare instances tend to be retained for longer periods or may not be forgotten at all.

To illustrate such a phenomenon, we depict the loss trend of clean and noisy samples in a memory buffer produced by a rehearsal baseline (ER-ACE [16]). In particular, Fig. 1

(left) shows the loss value sampled during standard training; differently, in Fig. 1 (right) we alternatively switch replay regularization on and off at each epoch. As can be seen, stopping replay has a distinct impact: while the loss value of clean samples remains low, it hugely increases for mislabeled ones. We remark that this gap holds even when replay regularization turns on, as the model easily adapts to clean samples and hence learns them faster [4, 27, 58].

In light of this, our main contribution is **Alternate Experience Replay** (**AER**), a novel CL optimization scheme that alternates steps of *buffer learning* and *buffer forgetting* to encourage the separation of clean and noisy samples in the buffer. To the best of our knowledge, our work is the first one that exploits forgetting to purify the memory buffer while learning from an online stream. Furthermore, to take advantage of the enhanced separation brought by AER, we propose **Asymmetric Balanced Sampling** (**ABS**): a new sample selection strategy designed to select mostly clean samples while keeping the most informative ones from the past. Through extensive experiments, we *i)* evaluate our method across various noise types (both synthetic and real) and rates, *ii)* compare with existing CLN methods adapted to our setting, *iii)* prove the applicability of AER and ABS on other rehearsal-based methods, ***iv)*** validate the impact of each component on performance and buffer quality.

2 Related works

Continual Learning methods can be broadly categorized into regularization-based – these limit changes to key task-related parameters [31, 61] – and rehearsal-based methods [55].

Rehearsal. In most existing CL scenarios [14, 17, 55], it has been shown that supplementing current training data with past samples is more effective at mitigating forgetting than any regularization-based methods. A simple yet effective method is Experience Replay (ER) [47, 49] which interleaves the current training batch with past examples. Otherwise, GDumb [46] pushes this concept to the extreme by greedily storing incoming samples and subsequently training a model from scratch using only the samples stored in the buffer.

Sampling strategies. Given their low capacity, buffers need to contain a balanced outlook of all seen classes. For this purpose, many employ *reservoir sampling* [57] to update the memory [8, 14, 16]. The outcome is an independent and identically distributed snapshot of the incoming tasks. However, not every sample comes with the same significance or robustness against forgetting. As highlighted by [2, 5, 15], retaining complex samples is crucial for preserving the performance, which they detect through their loss value or model uncertainty, respectively.

2.1 Learning with Noisy Labels

Noisy data can originate from multiple sources, including systematic or measurement errors encountered when retrieving historical data [10, 29, 45], errors introduced by human annotators [23, 53], or the presence of outliers [13]. Furthermore, noisy labels pose a significant challenge in medical imaging [29], where small and noisy validation sets can hinder the effectiveness of model calibration techniques [18, 43]. Various approaches have been proposed to mitigate the impact of data noise, including adversarial training, regularization and robust loss functions [25, 36, 63]. Additionally, ensemble methods [37, 41, 44] have been proven effective in reducing the impact of noisy data on model performance.

Noisy label detection. The prevalent approach for identifying noisy data is grounded on the memorization effect [4, 27], according to which correctly labelled (*clean*) instances tend to produce a smaller loss than mislabeled (*noisy*) ones during the initial stages of training. However, as the training ensues and the model starts to learn wrong patterns from the noisy data, its predictions become less reliable (*confirmation bias*). In this regard, [22] performs gradient ascent on the noisy samples, building on top of existing sample selection strategies and enhancing loss correction algorithms. Other works exploit separate models to perform sample selection, training either on a probably clean subset (CoTeaching [21], MentorNet [27]) or on all seen samples with semi-supervised objectives (DivideMix [34], [3]).

2.2 Continual Learning under Noisy Labels

Recent studies [6, 28, 30] conducted in the online CL setup have shown that existing sampling strategies fail to produce meaningful gains in noisy scenarios. In this respect, the authors of PuriDivER [6] propose a sampling strategy that promotes a trade-off between *purity* and *diversity* for samples in the buffer. Methods like SPR [30] and CNLL [28] use multiple buffers to gradually isolate clean samples: an auxiliary – usually larger – buffer gathers data from the stream, while a refinement procedure based on the small-loss criterion extracts only the clean samples into a purified buffer. Afterwards, SPR trains a network using a self-supervised loss on samples from both buffers, while CNLL adopts a semi-supervised approach inspired by FixMatch [51]. These models are limited to online settings due to either high computational demands or the tendency to overfit, losing effectiveness.

3 Method

Problem setting. We define the Continual Learning framework as the process of learning from a sequential series of T tasks. During each task $t \in \{0, 1, \dots, T-1\}$, input samples $\mathbf{X_t}$ and their annotations $\mathbf{Y_t}$ are drawn from an i.i.d. distribution $\mathcal{D}_t$. We follow the well-established class-incremental scenario [14, 19, 55] in which $\mathbf{Y}_{t-1} \cap \mathbf{Y}_t = \emptyset$ and at task t the learner f_θ is required to distinguish between all observed classes. Ideally, we wish to minimize:

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}_t \left[\mathop{\mathbb{E}}_{\mathcal{B}\sim\mathcal{D}_t} \left[\mathcal{L}(f_\theta(\mathbf{x}), y) \right] \right], \tag{1}$$

where $\mathcal{L}$ is the cross-entropy loss and $\mathcal{B} = (\mathbf{x}, y)$. As in CL the objective above is inaccessible, we leverage a fixed-size buffer $\mathcal{M}$ to store and replay part of the incoming samples. As a result, the generalized objective for rehearsal CL can be defined as:

$$\theta^* = \operatorname*{argmin}_\theta \mathop{\mathbb{E}}_{\mathcal{B}\sim\mathcal{D}_t} \left[\mathcal{L}(f_\theta(\mathbf{x}), y) \right] + \mathcal{L}_R, \tag{2}$$

where the *replay regularization* term $\mathcal{L}_R$ depends on the choice of the replay-based method. Although our approach can be equally applied to advanced choices of $\mathcal{L}_R$ [9, 12, 14, 26] (see Sec. 5), in this work we build upon the simplest strategy and leverage Experience Replay [47, 49]:

$$\mathcal{L}_R = \mathop{\mathbb{E}}_{(\mathbf{x}_r, y_r)\sim\mathcal{M}} \left[\mathcal{L}(f_\theta(\mathbf{x}_r), y_r) \right]. \tag{3}$$

As the objective in Eq. (3) could result in bias accumulation toward the current task [1], we adopt the asymmetric cross-entropy loss introduced in [16].

Algorithm 1 Overall procedure of AER with ABS

Input: stream data $\mathcal{D}_t$, buffer $\mathcal{M}$, training epochs T

1: **for** epoch in $1,\ldots,T$ **do**
2: $\quad\theta_{\text{CHK}} \leftarrow \theta_{\text{epoch}}$ $\quad\triangleright$ save current parameters of $f_\theta(\cdot)$
3: $\quad$**for** batch $\mathcal{B} \sim \mathcal{D}_t$ **do**
4: $\quad\quad p \leftarrow \text{normalize}(\{s(x); \forall (x,\tilde{y}) \in \mathcal{M}\})$ $\quad\triangleright$ compute asymmetric scores (*selection*)
5: $\quad\quad$**if** epoch in T_{on} **then**
6: $\quad\quad\quad$train on $\mathcal{B} \cup (\mathbf{x},\tilde{y}) \sim \mathcal{M}$ $\quad\triangleright$ buffer learning
7: $\quad\quad$**else**
8: $\quad\quad\quad$train on $\mathcal{B} \sim \mathcal{D}_t$ $\quad\triangleright$ buffer forgetting
9: $\quad\quad\quad\mathcal{R} \leftarrow \text{reservoir}(\{(\mathbf{x},\tilde{y}) \in \mathcal{B} : \mathcal{L}(\mathbf{x},\tilde{y}) < \mathcal{L}_\alpha\})$ $\quad\triangleright$ sample insertion
10: $\quad\quad\quad\mathcal{M}[z \sim p] \leftarrow \mathcal{R}$ $\quad\triangleright$ replace data sampled with p with stream data $\mathcal{R}$
11: $\quad$**if** epoch in T_{off} **then**
12: $\quad\quad\theta_{\text{epoch}} \leftarrow \theta_{\text{CHK}}$ $\quad\triangleright$ restore previous model checkpoint

3.1 Alternate Experience Replay (AER)

As discussed above, we tackle the challenge of **Continual Learning under Noisy Labels** (**CLN**), where each incoming dataset is affected by noise in the labeling process, leading to mislabeled training examples. For a given instance $\mathbf{x}_i \in \mathcal{D}_t$, we indicate with $\tilde{y}_i \sim \tilde{Y}_i$ the labels corrupted with annotation noise and with $\Pr(\tilde{y}_i \neq y_i)$ the respective *noise rate*. In this setting, we must simultaneously address the challenges posed by both noisy labels $\tilde{\mathbf{Y}}_t$ and the problem of forgetting. To achieve this, our main focus is on constructing a memory set $\mathcal{M}$ that is as clean and representative as possible. Since this objective involves distinguishing between noisy and clean examples when populating the memory set, our methodology seeks to maintain a significant gap between the losses of clean and noisy samples. This gap is indeed crucial for filtering examples through the widely used small-loss criterion. However, with no countermeasure, the loss gap starts to deteriorate as the replay of a small selection of data ensues (see Fig. 1, left). This effect is exacerbated in the popular offline (*i.e.* multi-epoch) CL setting [14, 48, 59], where we might be forced to trade-off convergence on the current task to avoid overfitting the mislabeled samples [4, 62].

To counteract the vanishing effect of the small-loss criterion and encourage the separation between the losses of noisy and clean samples, our novel methodology named **Alternate Experience Replay** (**AER**) induces forgetting of buffer datapoints. We refer the reader to Algo. 1 for a summary of the overall procedure. Specifically, we divide the training epochs for the current task into two categories: **buffer learning** and **buffer forgetting** epochs. The training process involves alternating between these two modes of learning.

- **Buffer learning.** In this regime, we train the model with standard replay (line 6) as in Eq. (2). Importantly, we do not modify the samples stored in the memory buffer $\mathcal{M}$ (no insertion or removal operations are performed).
- **Buffer forgetting.** In this case (line 8), we omit regularization on the memory buffer and focus the training exclusively on data from $\mathcal{D}_t$. By halting regularization and causing the subsequent forgetting of buffer datapoints, the loss of noisy examples is likely to increase more rapidly than that of clean ones [38, 54]. This, in turn, makes the small-loss criterion reliable once again (see Fig. 1, right). On top of that, we update $\mathcal{M}$ (line 9) through a loss-based selection strategy during these epochs (see Sec. 3.2).

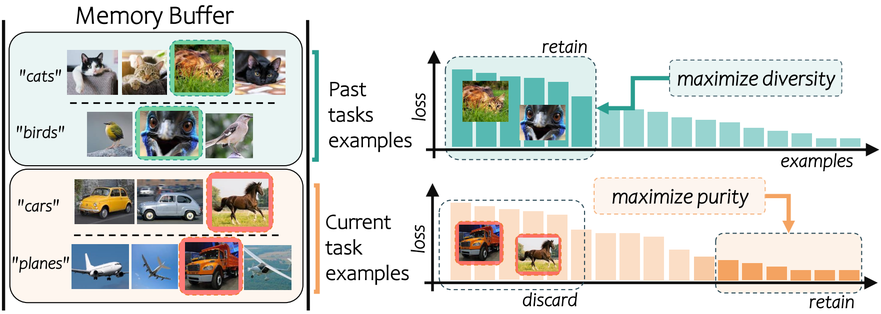

Figure 2: Asymmetric Balanced Sampling (ABS). Past examples are chosen to retain the most complex ones, while the criterion is reversed for the current task to maximize purity.

This way, at the end of each buffer forgetting epoch, we get a cleaner version of the memory buffer. However, cycling between buffer learning and forgetting could result in the buffer being under-optimized, as it is effectively exploited only during the former epochs. We avoid this through *model checkpointing*: specifically, at the start of each forgetting epoch, we save the parameters of the model f_θ (line 12) and restore them at the end of the same epoch (line 2). While this option results in the model being optimized for only half of the epochs, we prove in Sec. 4 that the trade-off significantly enhances the final accuracy of the model.

3.2 Asymmetric Balanced Sampling (ABS)

In this section, we outline the sampling strategy used to *insert* and *delete* examples into and from the memory buffer during each buffer forgetting epoch.

Sample insertion. Given a batch of data $\mathcal{B}$ from the current task, the first step is to determine which examples should be included in the buffer. To encourage the inclusion of clean examples, we exploit the memorization effect and employ a simple criterion that involves applying a threshold to the loss value. Formally, let α denote the percentage of samples within the current batch that we intend to exclude from the insertion procedure, we compute:

$$\mathcal{R} = \{(\mathbf{x}, \tilde{y}) \in \mathcal{B} : \mathcal{L}(\mathbf{x}, \tilde{y}) < \mathcal{L}_\alpha\} \tag{4}$$

where $\mathcal{L}_\alpha$ is the loss value at the α-th percentile of the loss distribution over $\mathcal{B}$. For our experiments, we set α to 75, thus discarding the 75% of samples with the highest loss and treating the remaining 25% as candidates to be inserted in the buffer (lines 9-10 of Algo. 1).

Sample selection. We approach the selection process by sampling from a probability distribution $p(\mathbf{x})$ defined over all exemplars $\forall \mathbf{x} \in \mathcal{M}$ in the buffer. To model such a distribution, as carried out by most methods [5, 6, 15], we leverage the score $s(x) = \mathcal{L}(x, y) \geq 0$ given by the loss function. It is noted that a valid distribution can be then obtained by normalizing these scores, such that $p(x) = \frac{s(x)}{Z}$ where $Z = \sum_{x \in \mathcal{M}} s(x)$. We refer to this strategy as **Loss-Aware Symmetric Sampling** (**LASS**). In LASS, examples with higher loss are more likely to be replaced, leading to a memory buffer that maintains greater **purity**.

However, we argue that a replacement criterion based on loss value like LASS could be detrimental in terms of **diversity**, as it discourages the retention of complex yet clean samples into the memory buffer. Indeed, these examples tend to exhibit higher loss value, as they lie in the proximity of the decision boundary [2, 8, 15]. Conversely, we argue that the small-loss criterion should be slightly revised when employed for memory replacement. Indeed, an ideal replacement strategy should preserve challenging yet representative examples from past tasks, in order to ensure an effective regularization signal.

In light of this, we propose a novel replacement strategy called **Asymmetric Balanced Sampling** (**ABS**, see Fig. 2) that looks for a compromise between two contrasting objectives. Namely, it aims to ensure the inclusion of both *ii)* high-loss (i.e., complex) samples from the past and *i)* small-loss (i.e., clean) samples from the present. To do so, it builds upon an **asymmetric score** (line 4) that differentiates whether a given example in the memory buffer belongs to a past task or the current one. Specifically, the score $s(x)$ is equal to the loss $\mathcal{L}(x,y)$ if the example comes from the current task $\mathcal{M}_{cur}$ (as in LASS). Conversely, to encourage diversity, the criterion is **reversed** for examples from past tasks $\mathcal{M}_{past}$, with the score being equal to $-\mathcal{L}(x,y)$. By taking this approach, we accept that a few mislabeled examples from past tasks might remain in the memory buffer; however, given the joint effect of the insertion policy and the LASS-like side of the replacement criterion – both of which tend to favor purity among examples from the current task – we can be cautiously optimistic that the examples from past tasks are correctly annotated (see Sec. 5 for an empirical analysis).

Finally, to achieve a balanced representation of both current and previous tasks in the memory buffer, we decide whether to replace a sample from the current or previous task based on their relative sizes. Considering $|\mathcal{M}_{cur}|$ as the number of samples from the current task in $\mathcal{M}$, we define a Bernoulli distribution with probability $\frac{|\mathcal{M}_{cur}|}{|\mathcal{M}|}$, where a success corresponds to sampling from $\mathcal{M}_{cur}$. This approach ensures that the likelihood of replacing a sample is proportional to the current task's sample size within the entire memory buffer.

3.3 Buffer consolidation

While combining AER and ABS achieves a balance between sample purity and complexity preservation, a further reduction in noise levels can be achieved by optimally selecting samples from $\mathcal{M}$ at the end of each task. We employ a MixMatch-based [7] approach to enhance model robustness, utilizing uncertain samples as unlabeled ones (*i.e.* buffer *consolidation* in the following). Further details about this phase are provided in the supplementary materials.

4 Experiments

Datasets and noise settings. We conduct experiments on five distinct datasets and various levels of noise. Specifically, we use the **Seq. CIFAR-100** dataset [32], which contains 32×32 images from 100 categories, split into 10 tasks, and the **Seq. NTU RGB+D** [50] dataset for 3D skeleton-based human action recognition, featuring 60 classes divided into 6 tasks. On these datasets, we inject two types of synthetic noise commonly employed in literature [21, 27, 34]: *symmetric* and *asymmetric* noise. In the first scenario, we replace the ground-truth label with probability $r \in [0,1]$ determined by the designated noise rate. The asymmetric or class-dependent noise setting, instead, is an approximation of real-world corruption patterns, altering labels within the same superclass as in [42, 63]. To further address real-world label noise, we evaluate our method on **Seq. Food-101N** [33] (5 tasks), composed of images gathered from the web, thus containing *instance-level* annotation noise. Additionally, ResNet18 [24] is used for Seq. CIFAR-100 with 50 epochs per task, ResNet34 [24] for Food-101N with 20 epochs, and EfficientGCN-B0 [52] for Seq. NTU-60 with 30 epochs.

Benchmark	**Seq. CIFAR-100**					**Seq. NTU-60**	
	symm			*asymm*		*symm*	
Noise rate	20	40	60	20	40	20	40
Joint	$54.77_{\pm 0.61}$	$38.46_{\pm 0.92}$	$23.36_{\pm 1.09}$	$56.70_{\pm 0.57}$	$42.61_{\pm 0.92}$	$68.26_{\pm 0.69}$	$63.02_{\pm 0.88}$
Finetune	$08.65_{\pm 0.13}$	$07.55_{\pm 0.14}$	$06.15_{\pm 0.17}$	$07.78_{\pm 0.14}$	$05.73_{\pm 0.09}$	$14.30_{\pm 0.51}$	$11.73_{\pm 1.07}$
ER [57]	$25.14_{\pm 0.28}$	$14.64_{\pm 0.23}$	$8.92_{\pm 0.23}$	$29.42_{\pm 0.39}$	$18.91_{\pm 0.86}$	$29.95_{\pm 2.16}$	$16.02_{\pm 0.27}$
+ *CoTeaching* [21]	$25.79_{\pm 0.61}$	$14.46_{\pm 0.49}$	$8.92_{\pm 0.30}$	$32.18_{\pm 2.55}$	$20.76_{\pm 2.44}$	$43.87_{\pm 0.78}$	$30.71_{\pm 1.86}$
+ *DivideMix* [34]	$33.31_{\pm 0.27}$	$22.91_{\pm 0.43}$	$13.58_{\pm 1.0.2}$	$36.98_{\pm 0.78}$	$26.10_{\pm 1.10}$	$40.92_{\pm 0.97}$	$32.07_{\pm 1.73}$
GDumb [46]	$16.96_{\pm 0.61}$	$11.31_{\pm 0.45}$	$7.62_{\pm 0.28}$	$17.25_{\pm 0.28}$	$11.75_{\pm 0.06}$	$11.34_{\pm 0.21}$	$6.86_{\pm 0.86}$
+ *CoTeaching* [21]	$17.02_{\pm 0.50}$	$13.17_{\pm 0.31}$	$8.17_{\pm 0.99}$	$17.07_{\pm 0.54}$	$12.05_{\pm 0.62}$	$12.37_{\pm 2.04}$	$8.82_{\pm 0.51}$
+ *DivideMix* [34]	$19.26_{\pm 0.97}$	$15.67_{\pm 0.97}$	$10.51_{\pm 0.32}$	$18.80_{\pm 1.55}$	$13.29_{\pm 0.29}$	$15.96_{\pm 1.16}$	$7.49_{\pm 1.11}$
PuriDivER [6]	$27.53_{\pm 0.53}$	$24.36_{\pm 0.40}$	$17.81_{\pm 0.43}$	$25.46_{\pm 1.44}$	$18.84_{\pm 0.64}$	$39.33_{\pm 1.59}$	$38.86_{\pm 0.79}$
PuriDivER.ME[†]	$41.25_{\pm 0.63}$	$37.61_{\pm 0.85}$	<u>27.18</u>$_{\pm 0.76}$	<u>41.65</u>$_{\pm 0.49}$	<u>30.22</u>$_{\pm 0.74}$	$43.10_{\pm 1.11}$	$38.07_{\pm 1.06}$
OURs	<u>44.34</u>$_{\pm 0.48}$	<u>38.64</u>$_{\pm 0.57}$	$26.34_{\pm 0.85}$	$41.24_{\pm 0.40}$	$29.26_{\pm 0.91}$	<u>47.71</u>$_{\pm 0.89}$	<u>43.11</u>$_{\pm 2.12}$
w. consolidation	$\mathbf{46.11}_{\pm 1.46}$	$\mathbf{40.27}_{\pm 0.40}$	$\mathbf{34.81}_{\pm 1.63}$	$\mathbf{43.67}_{\pm 0.73}$	$\mathbf{32.64}_{\pm 0.48}$	$\mathbf{48.73}_{\pm 1.20}$	$\mathbf{45.19}_{\pm 0.05}$

Table 1: Final Average Accuracy (FAA) [↑] on multiple datasets and noise rates. [†] Additional baselines adapted to the multi-epoch scenario.

Benchmarking. In line with notable CL works [5, 11, 20, 26, 40, 48, 59], we adhere to a class-incremental and **multi-epoch** setting, in which samples can be experienced multiple times within the respective task. The results are presented in terms of Final Average Accuracy (FAA), computed at the end of the last task. All results are averaged across 5 runs.

We compare against PuriDivER [6], the current state-of-the-art selection strategy for CLN, as well as common rehearsal CL baselines. For the latter, we follow [6] and apply both CoTeaching [21] and DivideMix [34] to consolidate the buffer of ER [47, 49] and GDumb [46]. Since current CLN methods are designed for the online setting (*i.e.* a single training epoch is allowed), a direct comparison would be problematic: based on Sec. 3.1, we hence refine PuriDivER by suspending memory updates after the first epoch, naming such method as **PuriDivER.ME**. We also compare with SPR [30] and CNLL [28], adapted for offline CLN and with the same overall memory budget for fairness. Given the huge computational demands of SPR and CNLL, evaluating them on complex datasets like CIFAR-100 and NTU proved impractical: hence, we employ the smaller Seq. CIFAR-10 dataset (5 tasks).

Finally, the upper bound is attained by training jointly on all tasks (*Joint*), while the lower bound is attained by training without any countermeasure to forgetting or noise (*Finetune*).

4.1 Comparison with State-of-the-Art

The results of our main evaluation are presented in Tab. 1. To streamline the discussion, we first compare our approach with traditional continual learning baselines, followed by an analysis of methods designed for continual learning under noisy labels (*e.g.* PuriDivER).

Comparison with rehearsal baselines. As outlined by Tab. 1, the approaches relying solely on buffer consolidation – such as ER and GDumb – are poorly effective, especially as noise levels rise. Regarding GDumb, its training phase is limited to the content of the memory buffer, preventing it from utilizing the data variety available throughout the task. This limitation is also evident from the comparison with standard ER, which consistently

Seq. CIFAR-10 – 40% *symm*			
Buffer size (total)		2500	*unlimited*
CNLL	*1 epoch*	38.14	57.26
CNLL	*50 epochs*	35.46	43.43
OURs	*50 epochs*	**67.10**	**76.83**
Buffer size (total)		1000	
SPR‡	*25 epochs*	26.34	
OURs	*25 epochs*	**63.65**	

Table 2: Comparison with SPR and CNLL. ‡training iterations spread across epochs.

ER	w. ACE	α	AER	ABS	FAA
✓	✓				11.65
✓	✓	✓			19.97
✓	✓	✓	✓		24.19
✓	✓	✓		✓	21.68
✓	✓	✓	✓	✓	**26.34**
✓		✓	✓	✓	22.02

Table 3: Ablation study for each component of our proposal – 60 % symmetric noise. (Seq. CIFAR-100 – 60% *symm*)

outperforms GDumb when noise levels are low. These outcomes highlight the benefits of performing multiple training iterations. However, this advantage turns into a double-edged sword as the stream becomes noisier, leading to a significant drop in performance.

Comparison with CNL methods. Firstly, we highlight the substantial improvement achieved by our adapted PuriDivER.ME, which outperforms PuriDivER by an average of 8.36%. Both versions perform buffer consolidation [6] at the end of each CL task; however, PuriDivER relies on a model trained over multiple epochs, which leads to the degradation of the small-loss criterion, an issue outlined in Sec. 3.1. Moreover, both PuriDivER.ME and ER + DivideMix are consistently surpassed by our proposal. In particular, we measure an average 1.50% gain over the best competitor's performance without any buffer consolidation, suggesting that our proposal improves the purity and diversity of samples in the buffer. However, as the sample selection is not perfect, applying an additional buffer consolidation technique tends to be more effective in more complex noise scenarios, with an average improvement of 4.71%.

We conduct additional comparisons with CNLL and SPR (Tab. 2) and PuriDivER.ME (Tab. 4). For the latter, we adopt the more realistic Food-101N dataset (*i.e.*, images collected from the web and automatically labeled). Even in these scenarios, our approach remains superior, both with and without buffer consolidation. We remark that these considerable gains come with a remarkable speed-up in terms of both time and resources used (see supplementary materials), making it more suitable for a multi-epoch incremental scenario.

5 Model analysis

Ablative study. We herein aim to investigate the impact of each component. Starting from the base rehearsal method used in our research, *i.e.* ER-ACE [16], we gradually introduce our two main contributions, AER and ABS, one at a time. As seen from the results in Tab. 3, each additional feature produces an increase in performance on Seq. CIFAR-100. For an in-depth analysis of the effects of the asymmetric cross-entropy loss function (ACE), we compare against the standard cross-entropy (*i.e.* ER in Tab. 3). The results indicate that the contribution of ACE is significant, aligning with both [16] and our initial expectations.

Purity of the buffer. Considering Seq. CIFAR-10 (40% noise), Fig. 4 depicts the *purity* and the *diversity* of the buffer produced by ABS, PuriDivER.ME, and LASS. For each class, purity is defined as the ratio of examples labeled correctly within the memory buffer. Instead, we model diversity as the intra-class variation within each class, thereby computing

Benchmark	Food-101N
Joint	39.91±1.05
PuriDivER.ME	28.62±0.85
OURs	29.86±1.18
w. buffer fit.	**34.79**±0.64

Table 4: Performances [↑] of our method and main competitor on a real-world noisy dataset.

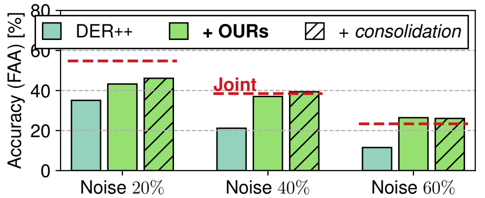

Figure 3: FAA ([↑]) of DER++ with our method and buffer fitting.

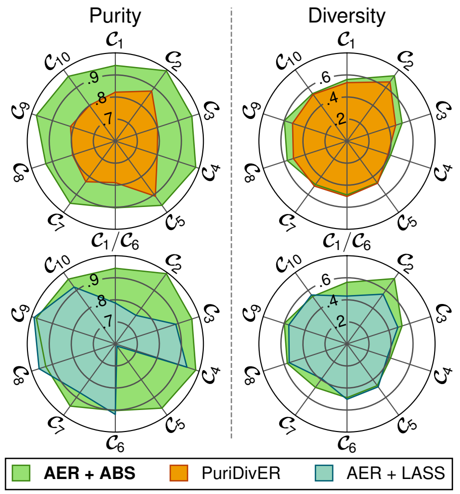

Figure 4: Final composition of the buffer with different choices of sample selection.

the average std. deviation of the features produced by the *Joint* ideal model. Finally, we scale all metrics according to their occurrence rate to account for potential imbalances in the number of examples from different classes. As shown in Fig. 4, ABS clearly outperforms both LASS and PuriDivER.ME in terms of purity and diversity. Unexpectedly, LASS yields a particularly unbalanced buffer, with only the most recent classes showing a good balance. In contrast, PuriDivER.ME achieves better balance but falls short in terms of purity.

Applicability to other methods. To evaluate whether AER/ABS can enhance other rehearsal methods, we apply them on **DER++** [14] and conduct tests on Seq. CIFAR-100. We also report the results with and without the consolidation phase (Sec. 4.1). The gains shown in Fig. 3 support the validity of our AER/ABS on enhancing other CL baselines.

Additional results. In the supplementary materials, we provide: *i)* details about the experimental settings, the adapted baselines, noise injection process and hyperparameters, *ii)* the evaluation of Final Forgetting (FF), *iii)* an analysis of the computational costs, *iv)* an evaluation of the speed at which the model learns the noisy data, *v)* a sensitivity analysis conducted on the hyperparameter α, which controls the purity within the sample insertion strategy.

6 Conclusions

We present an innovative framework for Continual Learning in the presence of Noisy Labels, a common issue in real-world AI applications. We focus on the multi-epoch class-incremental scenario, arguing the shortcomings of current methods leveraging the small-loss criterion. We hence appeal to a long-standing enemy of continual learning – *forgetting* – and propose Alternate Experience Replay to maintain a clear separation between mislabeled and clean samples. Additionally, we introduce Asymmetric Balanced Sampling to enhance sample diversity and purity within the buffer. We demonstrate the merits of our approach through extensive experiments, showcasing its potential in noisy incremental scenarios.

Acknowledgements

This paper has been supported from Italian Ministerial grant PRIN 2020 "LEGO.AI: LEarning the Geometry of knOwledge in AI systems", n. 2020TA3K9N. We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Finally, the authors would like to express their sincere gratitude to Alberto Zurli for his earlier work on this topic and his valuable, constructive contributions to the discussions.

References

[1] Hongjoon Ahn, Jihwan Kwak, S. Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. SS-IL: Separated Softmax for Incremental Learning. In *IEEE International Conference on Computer Vision*, 2021.

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient Based Sample Selection for Online Continual Learning. In *Advances in Neural Information Processing Systems*, 2019.

[3] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, 2019.

[4] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017.

[5] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2021.

[6] Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on a contaminated data stream with blurry task boundaries. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022.

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 2019.

[8] Federico Bolelli, Guido Borghi, and Costantino Grana. XDOCS: an Application to Index Historical Documents. In *Digital Libraries and Multimedia Archives*, volume 806, pages 151–162. Springer, Jan 2018. ISBN 978-3-319-73164-3. doi: https://doi.org/10.1007/978-3-319-73165-0_15.

[9] Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello, Concetto Spampinato, and Simone Calderara. On the Effectiveness of Lipschitz-Driven Rehearsal in Continual Learning. In *Advances in Neural Information Processing Systems*, 2022.

[10] Lorenzo Bonicelli, Angelo Porrello, Stefano Vincenzi, Carla Ippoliti, Federica Iapaolo, Annamaria Conte, and Simone Calderara. Spotting virus from satellites: Modeling the circulation of west nile virus through graph neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[11] Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[12] Matteo Boschini, Lorenzo Bonicelli, Angelo Porrello, Giovanni Bellitto, Matteo Pennisi, Simone Palazzo, Concetto Spampinato, and Simone Calderara. Transfer without forgetting. In *Proceedings of the European Conference on Computer Vision*, 2022.

[13] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

[14] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark Experience for General Continual Learning: a Strong, Simple Baseline. In *Advances in Neural Information Processing Systems*, 2020.

[15] Pietro Buzzega, Matteo Boschini, Angelo Porrello, and Simone Calderara. Rethinking Experience Replay: a Bag of Tricks for Continual Learning. In *International Conference on Pattern Recognition*, 2020.

[16] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New Insights on Reducing Abrupt Representation Change in Online Continual Learning. In *International Conference on Learning Representations Workshop*, 2022.

[17] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. In *International Conference on Machine Learning Workshop*, 2019.

[18] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6889–6899, 2021.

[19] Sebastian Farquhar and Yarin Gal. Towards Robust Evaluations of Continual Learning. In *International Conference on Machine Learning Workshop*, 2018.

[20] Emanuele Frascaroli, Aniello Panariello, Pietro Buzzega, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Clip with generative latent replay: a strong baseline for incremental learning. *arXiv preprint arXiv:2407.15793*, 2024.

[21] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

[22] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, 2020.

[23] Jasmine Hattab, Angelo Porrello, Anastasia Romano, Alfonso Rosamilia, Sergio Ghidini, Nicola Bernabò, Andrea Capobianco Dondona, Attilio Corradi, and Giuseppe Marruchella. Scoring enzootic pneumonia-like lesions in slaughtered pigs: Traditional vs. artificial-intelligence-based methods. *Pathogens*, 2023.

[24] 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*, 2016.

[25] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations Workshop*, 2018.

[26] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019.

[27] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2018.

[28] Nazmul Karim, Umar Khalid, Ashkan Esmaeili, and Nazanin Rahnavard. Cnll: A semi-supervised approach for continual noisy label learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022.

[29] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.

[30] Chris Dongjoo Kim, Jinseo Jeong, Sangwoo Moon, and Gunhee Kim. Continual learning on noisy data streams via self-purified replay. In *IEEE International Conference on Computer Vision*, 2021.

[31] 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.

[32] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[33] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018.

[34] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations Workshop*, 2020.

[35] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, Jesse Berent, Abhinav Gupta, Rahul Sukthankar, and Luc Van Gool. Webvision challenge: Visual learning and understanding with web data. *arXiv preprint arXiv:1705.05640*, 2017.

[36] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in Neural Information Processing Systems*, 2020.

[37] Yangdi Lu and Wenbo He. Selc: Self-ensemble label correction improves learning with noisy labels. *IJCAI*, 2022.

[38] Pratyush Maini, Saurabh Garg, Zachary Lipton, and J Zico Kolter. Characterizing datapoints via second-split forgetting. *Advances in Neural Information Processing Systems*, 2022.

[39] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 1989.

[40] Martin Menabue, Emanuele Frascaroli, Matteo Boschini, Enver Sangineto, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Semantic residual prompts for continual learning. In *Proceedings of the European Conference on Computer Vision*, 2024.

[41] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations Workshop*, 2020.

[42] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[43] Federico Pollastri, Juan Maroñas, Federico Bolelli, Giulia Ligabue, Roberto Paredes, Riccardo Magistroni, and Costantino Grana. Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification. In *International Conference on Pattern Recognition*, pages 1298–1305. IEEE, Jan 2021. ISBN 978-1-7281-8809-6. doi: https://dx.doi.org/10.1109/ICPR48806.2021.9412685.

[44] Federico Pollastri, Mario Parreño, Juan Maroñas, Federico Bolelli, Roberto Paredes, Daniel Ramos, and Costantino Grana. A Deep Analysis on High Resolution Dermoscopic Image Classification. *IET Computer Vision*, 15(7):514–526, October 2021. ISSN 1751-9632. doi: https://doi.org/10.1049/cvi2.12048.

[45] Angelo Porrello, Stefano Vincenzi, Pietro Buzzega, Simone Calderara, Annamaria Conte, Carla Ippoliti, Luca Candeloro, Alessio Di Lorenzo, and Andrea Capobianco Dondona. Spotting insects from satellites: modeling the presence of culicoides imicola through deep cnns. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2019.

[46] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. GDumb: A simple approach that questions our progress in continual learning. In *Proceedings of the European Conference on Computer Vision*, 2020.

[47] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, 1990.

[48] 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*, 2017.

[49] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 1995.

[50] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[51] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 2020.

[52] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[53] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11244–11253, 2019.

[54] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations Workshop*, 2019.

[55] Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 2022.

[56] Eli Verwimp, Matthias De Lange, and Tinne Tuytelaars. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In *IEEE International Conference on Computer Vision*, 2021.

[57] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985.

[58] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2020.

[59] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019.

[60] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015.

[61] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, 2017.

[62] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations Workshop*, 2017.

[63] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 2018.