Towards Unbiased Continual Learning: Avoiding Forgetting in the Presence of Spurious Correlations

Giacomo Capitani, Lorenzo Bonicelli, Angelo Porrello,
Federico Bolelli, Simone Calderara, and Elisa Ficarra

Università degli Studi di Modena e Reggio Emilia, Italy

{name.surname}@unimore.it

Abstract

Continual Learning (CL) has emerged as a paramount area in Artificial Intelligence (AI) because of its ability to learn multiple tasks sequentially without significant performance degradation. Despite the growing interest in CL frameworks, a critical aspect must be addressed: the inherent biases within training data. In this work, we show that, if overlooked, these biases can significantly impair the efficacy of continual learning models by inducing reliance on suboptimal shortcuts during data stream and memory retention, exacerbating catastrophic forgetting. In response, we present Learning without Shortcuts (LwS), which sets forth two primary objectives: (i) to identify and mitigate the exploitation of spurious correlations within the data stream and (ii) to develop a novel mechanism that constructs a fair memory buffer used in replay-based CL strategies. Our buffer construction strategy exploits the model confidence in a given example to balance the portion of samples per class, hence their contribution when replay activates. Unlike existing methods, LwS is agnostic to protected attributes, and results highlight that the proposed solution is indeed resilient to spurious correlations in CL settings. Code is available at https://github.com/aimagelab/mammoth

1. Introduction

The implications of biases in Artificial Intelligence (AI) are profound, raising significant practical concerns. As such systems are increasingly integrated into society, they can exacerbate societal stereotypes, leading to significant ethical challenges. Notably, recent studies demonstrate how algorithms can exhibit racial bias and lead to disparities in patient care and treatment outcomes [17, 41]. Beyond healthcare, biases in AI can also affect other critical areas, such as criminal justice, financial services, and employment, where algorithms might reinforce existing inequalities [39].

Moreover, modern AI systems are trained on an ever-increasing volume of data, much of which may not be available during the initial training phase, *e.g*. new tasks or classes can be discovered as the system evolves. For this purpose, **Continual Learning** (CL) has become a prominent paradigm, especially when privacy concerns or limited resources constrain access to previous data. In CL, models learn tasks sequentially, facing the challenge of mitigating *catastrophic forgetting* [35,42], where the model forgets previously acquired knowledge while learning new tasks. In this respect, numerous CL methods exploit a *rehearsal* mechanism to protect against forgetting [2,5,8,9,33]. These methods utilize a small memory buffer to store past data and alternate training between the current task and the examples stored within the buffer. The sampling strategy typically employed to add or remove examples is *reservoir sampling* [47, 58], a stochastic method that ensures equal representation of previous tasks within the buffer.

Due to its broad applicability, the intersection between bias-related issues and CL has been recently studied in [24]. We build on this research line, arguing that rehearsal methods have significant limitations when applied to tasks influenced by bias and spurious correlations. Since the memory buffer holds only a small random subset of past examples, it will likely be dominated by instances that exhibit spurious correlations, which may lead to the under-represented groups being unfairly penalized. As the samples from the buffer offer the only source of wisdom regarding past tasks, a buffer poisoned with spurious correlation could further amplify existing biases, creating a compounding effect.

To illustrate the issue, we direct the attention of the reader to Fig. 2. In CelebA [32], attributes like *Wearing Necklace* exhibit strong correlations with latent variables like *Gender*, *i.e*. wearing a necklace is more common among women. As a result, the model is prone to learning shortcuts [20], associating the presence of a necklace with female traits and its absence with male traits. Such a shortcut can lead the model to predict a necklace on a woman even when she is not wearing one, or, conversely, fail to recognize a necklace on a man who is. To avoid shortcuts,

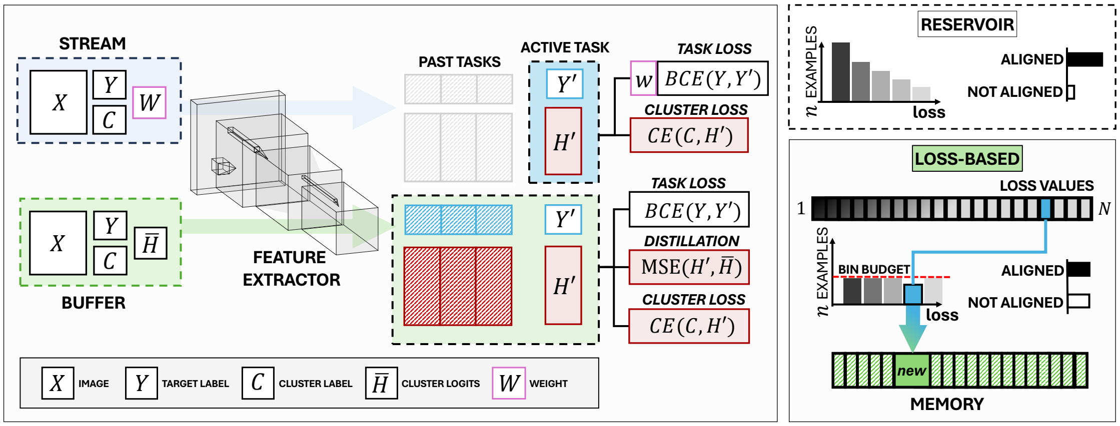

Figure 1. Overview of the proposed framework called Learning without Shortcuts (LwS). *Left*: during training, LwS employs tailored optimization objectives to relieve both forgetting and shortcut learning. Specifically, LwS couples the standard cross-entropy loss on labels Y with an auxiliary self-supervised term (*i.e.*, cluster loss). Importantly, the training loss for each example is dynamically adjusted to amplify the contribution of under-represented groups (e.g., women who do not wear a necklace) during training. *Right*: a visual of the loss-based criterion used by LwS to insert new elements within the memory buffer. The loss values serve as an effective proxy for distinguishing between bias-aligned and unaligned examples, a feature we leverage to achieve balanced representation across groups.

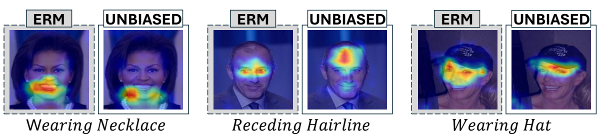

Figure 2. Attention heatmaps for *Wearing Necklace*, *Receding Hairline*, and *Wearing Hat* attributes in CelebA, using Empirical Risk Minimization (ERM) and ClusterFix [12]. ERM models often concentrate on irrelevant areas, exploiting shortcuts. Conversely, CFIX shows focused attention on more pertinent features.

current *debiasing methods* [49] exploit expensive auxiliary annotated metadata (*e.g.*, gender or ethnicity), or training paradigms whose outputs are invariant to biases [30]. However, none of these approaches were designed to handle a continuous stream of evolving and potentially biased tasks.

In light of these intuitions, we propose a novel approach —**Learning without Shortcuts** (**LwS**), see Fig. 1— to mitigate the effect of spurious correlations in CL without relying on latent variables supervision. LwS introduces *i)* an unsupervised objective against shortcuts while training on the current task, and *ii)* a loss-based sampling algorithm to ensure a fair representation across the groups in the buffer population. We conducted experiments on three benchmarks and achieved a notable improvement in average and worst-group accuracy, with our results even sometimes surpassing methods that employ latent variable supervision.

2. Related Works

2.1. Learning without Spurious Correlations

The field of *debiasing* has attracted significant attention, as it is crucial for ensuring fairness and robustness in machine learning models. The primary focus of debiasing methods is to mitigate the impact of *spurious correlations*, which can lead to biased predictions. Traditional methods like Distributionally Robust Optimization (DRO) [45] and Group Distributionally Robust Optimization (GDRO) [49] aim to optimize performance across varying data distributions. However, discovering which attributes cause bias-related issues often presents practical challenges in real-world scenarios. Consequently, there has been a shift towards unsupervised methods, which do not need protected-group labels, offering a more pragmatic approach for diverse application scenarios [29, 38, 40].

Unsupervised Debiasing Techniques. Recent research trends have focused on unsupervised methods for scenarios where access to protected group labels is lacking. Other debiasing approaches employ cluster-based assignments as a proxy of sensitive attribute supervision [53, 55]. Following this intuition, ClusterFix [12] integrates cluster-based DRO and a re-weighting sample importance strategy. Based on this, we introduce a novel loss-based buffer management approach, tackling a crucial shortfall of these models, which were not originally designed for incremental environments.

2.2. Continual Learning

Continual Learning (CL) is a research field focused on enabling models to learn from continuous streams of non-i.i.d. data. To do this without incurring the *catastrophic forgetting* [35] phenomenon, many methods adopt a *rehearsal* strategy, in which a subset of the incoming data is stored and replayed during the training of the model [2, 8, 11, 23, 46].

Rehearsal has proven successful in many CL scenarios, due to its effectiveness [13, 15] and flexibility in complex scenarios, such as those with annotation noise [4, 26, 37], partial lack of supervision [7,25], or absence of task boundaries [8, 14]. Other methods, such as regularization-based approaches [1,27,44,63], or architectural solutions [34,48], have also been proposed, but they are generally less effective [8, 57]. One notable mention is the recently introduced prompt-based methods [36, 54, 59, 60], which have shown effectiveness in mitigating forgetting. However, their applicability remains limited by the need for a huge initial pretraining and Transformer-based architectures.

The CL literature can be broadly categorized into three primary scenarios [57], based on whether the model has access to task identifiers during inference (Task-Incremental and Class-Incremental) or the presence of domain shifts in the input distribution (Domain-Incremental). Among these, the Class-Incremental (Class-IL) scenario is by far the most widely adopted [6, 23, 61], as it is usually regarded as the most challenging and realistic setting for real-world applications [2, 15, 57]. In Class-IL, the model is trained on a sequence of tasks, each containing a separate set of classes, and is evaluated on all the sequences.

Debiasing Continual Learning. Despite the recent advances, CL methods are known to be sensitive to spurious correlations. In particular, preliminary works [28, 51] have shown that the problem of transfer bias is exacerbated in CL, with the former influencing both future and past tasks. Since the current literature regarding the issue of CL under spurious correlations is still in its infancy, methods currently employed to mitigate this issue build upon well-established *rehearsal* CL. Notably, [28] introduces a group-aware Balanced Greedy Sampler (BGS) technique to adjust the last classification layer of the model after the end of each task. However, this method serves as an effective proof of concept but relies on the availability of group labels, often a limitation due to the need for privileged and costly information. Differently, LwP [24] separates the feature extractor from the classification network, with the latter trained from data obtained by a generative model. However, the effectiveness of such a strategy is still limited in CL, as recent works [18,43,56] have highlighted the difficulty of training a generative model from a changing data stream.

3. Problem Definition

Spurious Correlations. AI methods often focus on the interaction between an input space, represented as X (*e.g.* an image), and its associated output space, Y (e.g., ground truth label). In this context, we introduce the notion of a latent variable, referred to as z. This variable captures a unique attribute of an element $x \in X$, ranging from broader aspects like the presence of artifacts to more detailed image features such as the green grass in the background. To define this concept precisely, we can describe an element x with a set of binary attributes $A = \{z_1, z_2, \ldots, z_n\}$.

Even though these attributes may correlate with Y, they do not necessarily correspond to an attribute of interest. For example, the presence of a cow (Y) might be correlated with a background of green grass, where $z = 0$ indicates no green grass and $z = 1$ indicates the presence of green grass. While this correlation exists, relying on it can lead to harmful shortcuts in learning: recognizing the presence of grass may be easier, but it does not indicate the presence of a cow. Hence, relying solely on this correlation could lead to misinterpretation. This discrepancy is often referred to as spurious correlations [19]: associations in the data do not imply a causal relationship with the outcome.

Continual Learning with Spurious Correlations. In an incremental setup, the model is trained sequentially on different datasets $D_1, \ldots, D_T$, where each $D_t = (X_t, Y_t)$ represents a supervised classification task. Each dataset introduces some variation compared to the others, making the tasks distinct from one another. For example, each task could involve classifying a different visual attribute. The objective is to develop a function $f : X_t \to Y_t$ that effectively integrates new knowledge from successive tasks without losing performance on previously learned ones.

Within this context, each dataset D_t may be influenced by different biases. Consequently, the presence of spurious correlations has a detrimental effect on CL, especially on those methods that build upon a memory buffer, like replay-based approaches. Indeed, their effectiveness relies heavily on the quality of samples stored in the buffer, with significant degradation as it becomes contaminated by bias.

4. Method

We herein present *Learning without Shortcuts* (LwS), a continual debiasing approach that relieves the harmful effect of bias on learning from a data stream while preventing catastrophic forgetting. In particular, we exploit an auxiliary self-supervised approach to reduce the incidence of bias. This approach is popular in offline settings [12, 53, 55, 62] and exploits pseudo-labeling to regularize the latent representation of the model. Specifically, the pseudo-labels are obtained by clustering the latent space with k-means. Notably, this strategy poses technical challenges in continual learning due to the emergence of new tasks and associated cluster sets. To overcome these issues, we introduce the following **two modules**.

Data Stream. We start by extracting cluster assignments for the samples of the current task. These will be used throughout the task to ensure alignment with the initial representation. Here, the primary goal is to minimize the distance among samples that belong to the same inferred group (cluster) yet share the same class, thereby reducing the mutual information between spurious correlations and target labels within the data stream [53, 55]. This auxiliary task has also been shown to enhance the smoothness of the latent space [62], a property that facilitates the reuse and transfer of features across tasks [5, 16, 52].

Memory Buffer. To address the shortcomings of traditional replay-based methods, we propose a loss-based strategy to update the memory buffer. Specifically, the magnitude of the loss value is utilized to select which examples to store in the buffer. Secondly, we build upon knowledge distillation [22] to form the replay regularization objective. In contrast to common techniques, our method uses the output of the cluster classifier as the teaching signal for knowledge preservation. By doing so, we can maintain cluster coherence across current and future tasks, thereby mitigating forgetting and enriching transfer capabilities.

4.1. Data Stream Objective

Cluster Assignment. At the start of each task t, our approach assigns a cluster c to every element within the dataset D_t. This step involves partitioning D_t based on target labels y and performing k-means for each partition with features from a pre-trained frozen model $F_{pre} : X \to R^d$. Notably, the model F_{pre} remains the same across all tasks.

Debiased Training. From the samples of the data stream, our model is given a twofold objective. Firstly, it solves the binary classification problem of the task t, where y represents the ground truth label. Secondly, it adheres to a specific objective that constrains the feature space. This objective requires the model to remain consistent with the original cluster assignments c.

To ensure that minority groups are not overlooked, we modify the optimization objective to re-weigh the importance of each example. In practice, we assign a weight, w_c in Eq. (3), proportional to the average *error* and the cardinality of its cluster. The *error* considers the original and pseudo labels y and c.

Formally, let $F : X \to R^d$ be the feature extractor and let $\mathcal{T}_t : R^d \to R^1$ and $C_t : R^d \to R^C$ indicate, respectively, the task head and the cluster classifiers. The latter are two linear projections; while the first outputs the logits of the classes of the t-th task, the second is instead relevant for the auxiliary self-supervised objective. The parameters of the feature extractor F and the task head C_t are updated continuously across tasks. Differently, the parameters of the cluster classifier $\mathcal{T}_t$ are optimized only during task t. Formally, the clustering-structural loss is defined as follows:

$$\mathcal{L}_{cluster} = \mathcal{L}_{\text{CE}}(C_t \circ F(x), c) \tag{1}$$

The main objective of the optimization process is learning to classify y, which is achieved through a task-weighted classification loss. Indeed, the loss is weighted by a factor w_c, which reflects the "importance" of the cluster to which the sample belongs. Finally, the task-weighted classification loss is defined as follows:

$$\mathcal{L}_{\text{target}} = w_c \mathcal{L}_{\text{BCE}}(\mathcal{T}_t \circ F(x), y) \tag{2}$$

where w_c is:

$$\frac{1}{N_c}\mathbb{E}_{(x,y)\sim P_c}[\mathcal{L}_{\text{BCE}}(\mathcal{T}_t \circ F(x), y) + \gamma \mathcal{L}_{\text{CE}}(C_t \circ F(x), c)] \tag{3}$$

Then, the overall stream objective:

$$\mathcal{L}_{stream} = \mathcal{L}_{cluster} + \mathcal{L}_{target} \tag{4}$$

4.2. Memory Buffer Objective

We now present our approach to managing the memory buffer. During the execution of task $t \in \{1, 2, \dots, T\}$, the buffer memory serves as a temporary storage area. Its capacity, denoted as $\mathcal{M}$, sets the maximum number of elements it can hold at any given time. Further, we can allocate a maximum number of elements for each task, known as the budget of the task $\mathcal{B}$. To select which example from the current task to insert, our insertion strategy considers the loss value – defined as $\mathcal{L}_{\text{BCE}}(x_i)$ – for the target label y. Afterwards, we use a set of intervals, called bins, to categorize these loss values into distinct ranges. An example within the memory buffer is hence associated with a specific bin, determined by the interval in which its loss value falls.

Buffer Management. As we initiate a task t, the budget allocation for the task is determined as $\mathcal{B} = \frac{\mathcal{M}}{t}$, taking considering the total capacity $\mathcal{M}$ and the task index t. The allocation for each bin is then defined as $\frac{\mathcal{B}}{n}$, which ensures a proportional distribution of memory resources across the predefined number of n bins. For each instance of data x_i in the training set D_t, a loss value $l_x = \mathcal{L}_{\text{BCE}}(\mathcal{T}_t \circ F(x), y)$ is calculated and stored, after a warm-up of a few epochs. This follows [38], which shows that the gap between bias-aligned and bias-not-aligned emerges during the initial training epochs. Thus, we choose to compute the loss value after 5 epochs to take advantage of a more significant gap. These values range from a minimum $\mathcal{L}_{\min}$ to a maximum $\mathcal{L}_{\max}$, which establishes the scope of the bins. The allocation of loss ranges to specific bins is determined based on their relative position within this range $\mathcal{L}_{\max} - \mathcal{L}_{\min}$, divided into n equal intervals.

Buffer Population. To determine whether to include an instance x in the buffer, we first check the current number of elements in its corresponding bin. If this number is below the allocated budget $\frac{\mathcal{B}}{N}$, the instance is included. This method ensures a fair representation of instances within the buffer, including both low and high loss values. This way, we ensure that the memory buffer always contains examples that are both aligned and not aligned with spurious correlations. Indeed, there is a significant empirical correlation between the value of the loss and potential biases (see Fig. 3). We leverage this correlation to maintain a balanced buffer.

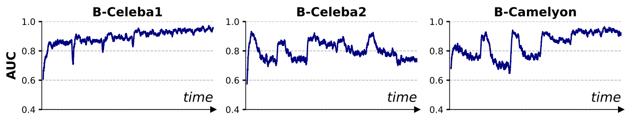

Figure 3. The AUC trend using binary cross-entropy loss to distinguish between the 'bias-aligned' and 'non-bias-aligned' groups. Notably, a higher AUC indicates that the loss is more effective at separating examples aligned with spurious correlations from those that are not. This result supports our strategy of achieving a balanced representation of bias-aligned and non-aligned groups, promoting fairer and more equitable sampling.

Knowledge Distillation from Buffer Memory. Our implementation of knowledge distillation involves classifying samples stored in a buffer and comparing the cluster logits values saved to those computed for the current model. Let $y' = \mathcal{T}_t \circ F(x)$ represent the model output for task t (current or past), and $h' = C_t \circ F(x)$ represent the cluster classifier logits as well. We define two distinct terms:

$$\mathcal{L}_{\text{buf}} = \mathcal{L}_{\text{BCE}}(y', y) + \mathcal{L}_{\text{BCE}}(h', c), \tag{5}$$

The loss function $\mathcal{L}_{\text{buf}}$ combines the target and cluster classification loss. Additionally, we define the knowledge distillation objective for the buffer as:

$$\text{KD}_{\text{buf}} = \mathbb{E}_{(x,\bar{h})\sim\mathcal{M}} \left[\left\| \bar{h} - h' \right\|_2^2 \right], \tag{6}$$

KD_{buf} stands for the expected euclidean distance between stored logits $\bar{h}$ and the computed current logits h' over the distribution of samples $(x, \bar{h})$ drawn from the buffer memory $\mathcal{M}$. Finally, the overall objective function combines the stream, buffer, and knowledge distillation objectives:

$$\mathcal{L} = \mathcal{L}_{\text{stream}} + \mathcal{L}_{\text{buf}} + \text{KD}_{\text{buf}} \tag{7}$$

5. Experiments

Assessing debiasing methods in an environment affected by spurious correlations is challenging. Many works use synthetic datasets or custom splits to regulate a latent attribute z in a controlled scenario [38, 53, 55, 64]. In our continual setting, we face a similar challenge as in [24, 28]. Here, we deal with a sequence of tasks occurring successively, each influenced by a certain degree of bias. We extend the setting [28] by increasing the number of tasks.

To comply with standard metrics used in literature about debiasing [12, 30, 50, 53, 55, 64], we used the worst-case accuracy (not employed in [24]). Namely, we compute the average and worst accuracies across groups, where a group is defined as $g = (y, z)$. The group-specific accuracy is denoted as $acc_g(f_T, D_{test_t})$, representing the accuracy of the final model f_T on group g in the t-th task. The metrics for average and worst-group accuracies are defined as follows:

$$Acc_{avg}(f_T, D_{test}) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|G|}\sum_{g\in G} acc_g(f_T, D_{test_t}) \tag{8}$$

$$Acc_{worst}(f_T, D_{test}) = \frac{1}{T}\sum_{t=1}^{T}\min_{g\in G} acc_g(f_T, D_{test_t}) \tag{9}$$

where G represents the set of all groups across tasks. Notably, each task t comes with its test unbiased dataset D_{test_t}, employed for evaluation.

Implementation Details. All reported results are the average of three runs. We use ResNet-18 [21] pre-trained on ImageNet-1K; **for fairness**, we apply this backbone to all tested methods. Each task was trained for 25 epochs using Stochastic Gradient Descent (SGD), with a learning rate of 1×10^{-3}. We performed k-means with $k = 8$ for all experiments. More details are provided in the supplementary.

5.1. Experimental Setup and Benchmarks

To model the presence of spurious correlations, we use CelebA [31] and Camelyon17 [3] from the WILDS benchmark [50]. We split the datasets into tasks such that a latent attribute z correlates with a target attribute, quantified by a given factor p_{corr}. We set p_{corr} to 0.95, indicating that 95% of images with a specific attribute y (*e.g.*, a necklace) are of a particular latent attribute z (*e.g.*, gender). In the supplementary materials, we provide an extensive graphical analysis illustrating the correlation factor between the variables y and z within our experimental settings. During the training process, we do not have access to latent variables z, using them only for evaluation.

Biased CelebA. The CelebA dataset [31] was divided into eight separate tasks for our study. These tasks focus on the binary classification of various target attributes y. We made two variants: **B-Celeba1** includes $\{$*Heavy Makeup, Blond Hair, Receding Hairline, Young, Wearing Necklace, Bags Under Eyes, Smiling, Eyeglasses*$\}$ while **B-Celeba2** includes $\{$*Chubby, Pale Skin, Bald, Gray Hair, Wearing Necktie, Wearing Hat, Arched Eyebrows, Mouth Slightly Open*$\}$. Each task contains 4 480 images in the training set, evenly distributed in terms of y. The latent attribute z is the gender label as in [12, 38, 53]. Each task has a test data D_{test_t} with 100 samples per group (there are 4 groups for each task) to assess model debiasing performance.

Biased Camelyon. This dataset is derived from the Camelyon17 dataset [3]. It consists of 4 tasks, each involving binary classification of tumors. The hidden variable z represents the hospital from which the images were sourced. The presence of a tumor is indeed correlated with the hospital where the images were taken, thereby creating a spurious correlation between the two variables. The training phase includes 4 hospitals, while the test phase includes a fifth hospital not present in the training data. Each task con-

Table 1. Comparison between unbiasing and CL methods, in terms of worst-group accuracy [↑] and average accuracy [↑]. The symbol † recalls that BGS uses auxiliary data during training, *i.e*. the label groups annotations.

	B-CelebA1		**B-CelebA2**		**B-Camelyon**	
Method	Acc$_{worst}$[%]	Acc$_{avg}$[%]	Acc$_{worst}$[%]	Acc$_{avg}$[%]	Acc$_{worst}$[%]	Acc$_{avg}$[%]
Random	50.00	50.00	50.00	50.00	50.00	50.00
SGD	14.87 ± 1.56	60.12 ± 0.68	8.12 ± 0.57	56.06 ± 0.09	48.53 ± 6.47	85.8 ± 1.51
Debiasing						
BPA	15.08 ± 1.56	61.69 ± 0.47	9.16 ± 0.47	56.33 ± 0.53	62.13 ± 2.73	88.06 ± 1.11
CFIX	18.00 ± 2.04	64.00 ± 1.25	17.65 ± 1.97	61.26 ± 0.96	59.56 ± 0.83	87.88 ± 0.57
Replay (1024)						
BGS†	55.68 ± 2.92	74.64 ± 0.34	56.56 ± 2.74	76.45 ± 0.51	77.55 ± 0.07	91.89 ± 0.41
ER-ACE	16.37 ± 1.76	60.75 ± 0.77	13.12 ± 1.10	59.03 ± 0.59	56.80 ± 1.70	88.48 ± 0.08
DER++	21.79 ± 1.06	61.34 ± 0.54	18.03 ± 1.37	60.87 ± 0.36	53.40 ± 1 41	82.56 ± 0.57
BPA + replay	16.04 ± 0.90	60.92 ± 0.51	11.37 ± 0.43	58.19 ± 0.66	65.33 ± 1 02	88.88 ± 0.52
CFIX + replay	17.80 ± 0.04	61.57 ± 0.39	19.79 ± 0.75	62.62 ± 0.58	55.93 ± 0.34	86.48 ± 0.58
LwP	19.40 ± 0.91	62.33 ± 1.31	13.44 ± 3.94	57.47 ± 1.47	54.40 ± 9.54	86.39 ± 4.24
LwS (ours)	**58.84** ± **2.42**	**71.43** ± **1.67**	**50.91** ± **3.52**	**70.73** ± **0.65**	**80.40** ± **1.74**	**92.47** ± **0.21**

tains 4,096 images, and the test sets are balanced for tumor presence and hospital origin, with 500 images per hospital.

5.2. Baseline and Competing Methods

Rehearsal Methods. While **SGD** does not incorporate measures against forgetting, **ER-ACE** [10] enhances traditional Experience Replay (ER) by applying distinct loss functions for the stream (considering the logits of incoming data) and the buffer. **DER++** [8] adopts self-distillation by encouraging consistency in the model's output, minimizing the L2 norm between the logits of current and past iterations. However, they do not consider the potential contamination of the buffer by spurious correlations, which could affect future knowledge retention and subsequent tasks.

Continual Debiasing Methods. To mitigate spurious correlations in both the stream and buffer, several methods have been proposed. **LwP** [24] aims to prevent spurious correlations by using self-supervised learning with feature-level augmentation. **BGS**† [28] constructs the buffer to store group-class balanced examples across all encountered tasks. In this context, BGS acts as an oracle by leveraging latent variable z supervision to structure the buffer.

Offline Debiasing Methods. We also assessed standard debiasing algorithms such as **BPA** [53], which employs a per-sample re-weighting strategy. **CFIX** [12] optimizes a dual objective to re-weight sample importance, using cluster classification as an additional regularization to smooth the latent space. Since these methods do not natively support the arrival of new tasks, we also introduce **BPA + replay** and **CFIX + replay**, which refer to our adaptations that incorporate buffer reservoir sampling.

5.3. Experimental Results

Tab. 1 summarize the key findings of our work. LwS boosts average and worst-group accuracy metrics, outperforming rehearsal methods across various scenarios. A notable feature is the gain in worst-group accuracy, highlighting its effectiveness against spurious correlations. Also, the results prove how our mechanism to update the memory buffer allows the retention of unbiased past knowledge.

Baselines. Regarding debiasing methods, CFIX [12] and BPA [53] have effectively improved worst-case accuracy with respect to fine-tuning on the new task (SGD). However, their gains are relatively small compared to LwS, indicating the need for a buffer strategy to avoid forgetting. In this context, offline debiasing algorithms serve as more reliable baselines than naive fine-tuning (SGD).

Rehearsal Methods. Their results are reported in Tab. 1; we refer the reader to Fig. 4 for a in-depth comparison with DER++ [8], one of the most simple yet effective approaches. As can be observed, replay methods surpass their baselines, highlighting the advantage of memory replay. However, the table reveals a crucial issue. If the buffer contains mostly biased elements, it can amplify the bias within new tasks when samples are retained from the buffer. This underscores the limitation of traditional rehearsal methods, which can easily fall into the trap of shortcut learning.

Continual Debiasing Methods. From our results, LwS outperforms a continual debiasing model like LwP [24] and pairs the performance of BGS [28], which presents our upper bound. Indeed, it constructs the buffer **using the latent attribute z supervision** to balance the number of elements for each group in the memory, which is preferable but less

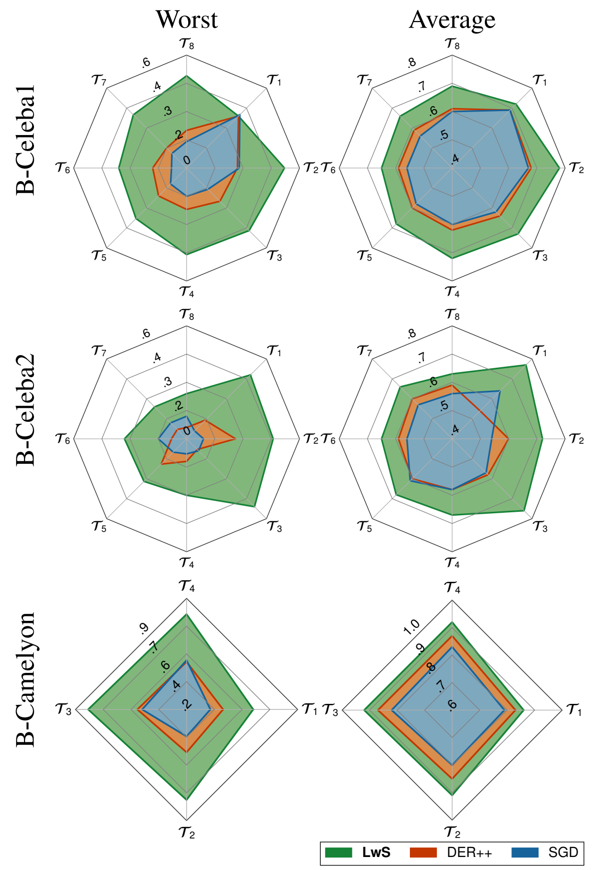

Figure 4. Comparative analysis across tasks, showcasing worst-group accuracy and average accuracy for each dataset.

realistic. Indeed, to identify the group labels, one must *i)* discover the variable z that determines the spurious correlation; *ii)* annotate the training set accordingly. This process is expensive and requires a thorough analysis of the dataset. Furthermore, it becomes even more challenging in continual learning where tasks arrive continuously. While annotating attributes like gender may be easy, it becomes unpractical when the attribute z is hard to inspect (*e.g.* metadata protected by privacy laws or hidden artifacts in images). In such cases, a framework like ours, which avoids relying on group labels, is advantageous.

6. Ablation Studies

Reservoir Sampling Fails with Spurious Correlations. Tab. 2 illustrates the impact of memory buffer size ($\mathcal{M}$) and buffer handling strategies on LwS. The results reveal that the loss-based approach consistently outperforms the *reservoir* method in terms of worst-group and average accuracy across all datasets and buffer sizes (256, 512, 1024). This outcome supports our hypothesis that random strategies like *reservoir* may unintentionally amplify spurious correlations in scenarios with minimal buffer capacity due to the limited representation of non-aligned elements.

Table 2. LwS performance in terms of worst [↑] and average accuracy [↑] across different buffer sizes and management strategies.

	$\mathcal{M}$	Strategy	$\text{Acc}_{worst}[\%]$	$\text{Acc}_{avg}[\%]$
B-CelebA1	256	reservoir	14.14	58.21
		loss-based	**36.29**	**66.73**
	512	reservoir	18.50	61.08
		loss-based	**52.12**	**71.17**
	1024	reservoir	17.87	62.16
		loss-based	**56.98**	**72.57**
B-CelebA2	256	reservoir	18.71	61.10
		loss-based	**51.62**	**72.06**
	512	reservoir	19.37	62.43
		loss-based	**48.50**	**69.46**
	1024	reservoir	20.50	63.06
		loss-based	**53.37**	**71.40**
B-Camelyon	256	reservoir	41.40	81.92
		loss-based	**79.40**	**91.84**
	512	reservoir	36.80	81.92
		loss-based	**79.60**	**92.42**
	1024	reservoir	55.80	86.50
		loss-based	**80.40**	**92.84**

Table 3. LwS performance comparison varying number of bins and usage of knowledge distillation (KD).

	B-CelebA1		**B-CelebA2**		**B-Camelyon**	
# bins	Acc_w	Acc_{avg}	Acc_w	Acc_{avg}	Acc_w	Acc_{avg}
2	58.23	72.71	55.12	75.40	76.40	90.88
4	**61.55**	**73.24**	53.37	71.40	81.20	92.72
8	53.12	70.79	51.25	70.68	**81.40**	92.40
16	55.61	71.83	52.00	71.72	80.40	92.84
32	50.37	70.82	51.00	71.34	79.60	**93.04**
no KD	58.80	**73.40**	50.50	70.18	80.40	**92.84**
w. KD	**61.55**	73.24	**53.37**	**71.40**	**81.20**	92.72

Varying the Correlation Factor p_{corr}. We analyze how the model learns as the correlation factor changes and evaluate the effectiveness of different strategies. On the left side of Fig. 5, the relationship between the loss value and alignment with spurious signals (AUC) is shown as p_{corr} varies. After the warm-up phase, we compute the loss for all training elements of task t, which is then used in the buffer update described in Sec. 4. We observe a gradual decrease in AUC after buffer insertion, which is the desired outcome. As depicted on the right, joint training and

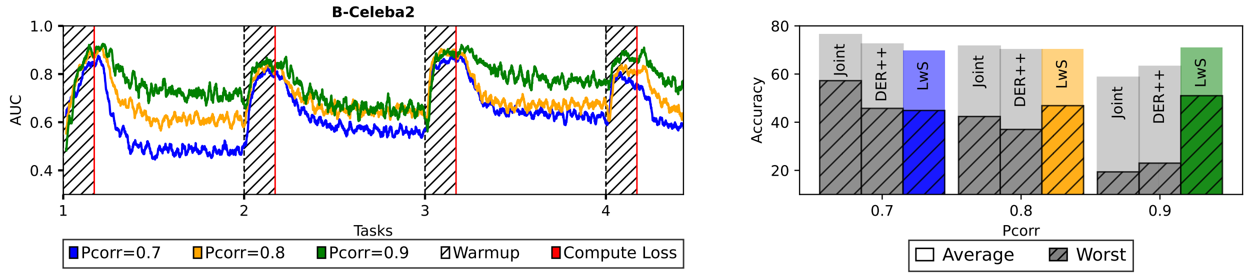

Figure 5. The figure displays AUC curves (left), which show the correlation between the loss value and alignment with spurious signals varying levels of p_{corr}. The shaded regions on the curves show the warm-up phase, followed by target loss computation for all training samples of each task. Loss values are utilized by the buffer insertion strategy, explained in Sec. 4. The right side of the figure presents a comparative accuracy analysis under different p_{corr} values for joint training, DER++, and LwS methods.

Table 4. LwS with adaptive weights w_c and fixed $w_c = 1$.

Dataset	w_c	Acc$_{worst}$	Acc$_{avg}$
B-CelebA1	adaptive	**58.84 ± 2.42**	**71.43 ± 1.67**
	fixed	52.83 ± 1.59	70.90 ± 0.99
B-CelebA2	adaptive	**50.91 ± 3.52**	70.73 ± 0.65
	fixed	47.29 ± 1.43	**70.74 ± 0.68**
B-Camelyon	adaptive	**80.40 ± 1.74**	**92.47 ± 0.21**
	fixed	78.20 ± 1.60	91.85 ± 0.37

Table 5. LwS results on B-CelebA1 using difference values of γ.

Metric	$\gamma = .0$	$\gamma = .2$	$\gamma = .5$	$\gamma = .8$	$\gamma = 1$
Acc$_{worst}$	47.61	51.92	55.62	54.85	**58.25**
Acc$_{avg}$	69.46	70.16	70.91	70.83	**72.12**

DER++ are more susceptible to spurious correlations. As p_{corr} increases, both methods suffer a drop in average and worst-case accuracy while our approach performs robustly across different p_{corr} values.

On the Number of Bins. We investigate the effect of varying the number of bins for the buffer population. As the number of bins increases, we observe a slight decline in worst-case accuracy, as shown in Tab. 3. This trend can be attributed to the fixed buffer size; a greater number of bins entails a reduced allocation budget per bin, potentially leading to an under-representation of elements that diverge from the bias within each bin. Despite this, our strategy maintains competitive performance, even with a higher bin count, as shown for B-Camelyon.

Knowledge Distillation using Cluster Logits. We analyzed the impact of the KD_{buf} term introduced in Eq. (6). Our findings demonstrate that knowledge distillation offers significant advantages in smoothing the feature landscape and facilitating knowledge transfer across future tasks. In particular, Tab. 3 shows that utilizing cluster prediction logits improves the worst-case accuracy performance without negatively affecting the average accuracy.

On the Effect of w. Fixing $w_c = 1$ in Eq. (3) worsened model performance as shown in Tab. 4, demonstrating the effectiveness of our adaptive weighting strategy. As expected, the decrease with $w_c = 1$ was not severe thanks to the buffer population, which serves as a regularization term.

Sensitivity of γ. Tab. 5 shows how increasing the value of γ in Eq. (3) leads to better results. The scalar γ multiplies $L_{cluster}$ term, which indicates heterogeneity within a cluster c, where individuals share the same target label y (e.g., blond hair) but differ in attribute z (e.g., gender). Therefore, we assign a higher weight w_c to a cluster with a high expected error for L_{target} or $L_{cluster}$.

7. Conclusion

The challenge of shortcut learning in neural networks is a complex and relatively unexplored area. This issue is further exacerbated in Continual Learning, particularly in methods based on rehearsal. Our approach, *Learning without Shortcuts* (**LwS**), tackles this by integrating a debiasing strategy within the data-stream and a sampling mechanism designed to mitigate spurious correlations. Our study lays a solid groundwork for promoting worst-case generalization and algorithmic fairness in online settings.

8. Acknowledgments

This work was partially supported by the Italian Ministerial grants PRIN 2022: “B-Fair: Bias-Free Artificial Intelligence Methods for Automated Visual Recognition” (CUP E53D23008010006) and “AIDA: explAinable multImodal Deep learning for personAlized oncology” (Project Code 20228MZFAA). We acknowledge the CINECA award under the ISCRA initiative for providing high-performance computing resources. This research was also supported by the University of Modena and Reggio Emilia and Fondazione di Modena through the FAR 2023 and FARD-2024 funds (Fondo di Ateneo per la Ricerca).

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018. 3

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019. 1, 2, 3

[3] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018. 5

[4] 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 *CVPR*, 2022. 3

[5] Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello, Concetto Spampinato, and Simone Calderara. On the effectiveness of lipschitz-driven rehearsal in continual learning. *Advances in Neural Information Processing Systems*, 35:31886–31901, 2022. 1, 4

[6] 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. 3

[7] Matteo Boschini, Pietro Buzzega, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Continual semi-supervised learning through contrastive interpolation consistency. *Pattern Recognition Letters*, 2022. 3

[8] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020. 1, 2, 3, 6

[9] Pietro Buzzega, Matteo Boschini, Angelo Porrello, and Simone Calderara. Rethinking experience replay: a bag of tricks for continual learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2180–2187. IEEE, 2021. 1

[10] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. *arXiv preprint arXiv:2104.05025*, 2021. 6

[11] 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 *ICLR*, 2022. 2

[12] Giacomo Capitani, Federico Bolelli, Angelo Porrello, Simone Calderara, and Elisa Ficarra. Clusterfix: A cluster-based debiasing approach without protected-group supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4870–4879, 2024. 2, 3, 5, 6

[13] 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. 2

[14] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[15] Sebastian Farquhar and Yarin Gal. Towards Robust Evaluations of Continual Learning. In *International Conference on Machine Learning Workshop*, 2018. 2, 3

[16] Enrico Fini, Victor G Turrisi Da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *CVPR*, pages 9621–9630, 2022. 4

[17] Chloë FitzGerald and Samia Hurst. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):1–18, 2017. 1

[18] Qiankun Gao, Chen Zhao, Bernard Ghanem, and Jian Zhang. R-dfcil: Relation-guided representation learning for data-free class incremental learning. In *ECCV*, 2022. 3

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 3

[20] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[23] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 2, 3

[24] Myeongho Jeon, Hyoje Lee, Yedarm Seong, and Myungjoo Kang. Learning without prejudices: Continual unbiased learning via benign and malignant forgetting. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3, 5, 6

[25] Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, and Karteek Alahari. A soft nearest-neighbor framework for continual semi-supervised learning. In *ICCV*, 2023. 3

[26] Chris Dongjoo Kim, Jinseo Jeong, Sangwoo Moon, and Gunhee Kim. Continual learning on noisy data streams via self-purified replay. In *ICCV*, 2021. 3

[27] 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 neu-

ral networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[28] Donggyu Lee, Sangwon Jung, and Taesup Moon. Continual learning in the presence of spurious correlations: Analyses and a simple baseline. In *ICLR*, 2024. 3, 5, 6

[29] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *ECCV*, pages 270–288. Springer, 2022. 2

[30] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 2, 5

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 1

[33] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 1

[34] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, 2018. 3

[35] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1, 2

[36] Martin Menabue, Emanuele Frascaroli, Matteo Boschini, Enver Sangineto, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Semantic residual prompts for continual learning. *ECCV*, 2024. 3

[37] Monica Millunzi, Lorenzo Bonicelli, Angelo Porrello, Jacopo Credi, Petter N Kolm, and Simone Calderara. May the forgetting be with you: Alternate replay for learning with noisy labels. In *BMVC*, 2024. 3

[38] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: Training Debiased Classifier from Biased Classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 2, 4, 5

[39] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018. 1

[40] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020. 2

[41] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 1

[42] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019. 1

[43] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 3

[44] Angelo Porrello, Lorenzo Bonicelli, Pietro Buzzega, Monica Millunzi, Simone Calderara, and Rita Cucchiara. A second-order perspective on compositionality and incremental learning. *arXiv preprint arXiv:2405.16350*, 2024. 3

[45] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. 2

[46] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *CVPR*, 2017. 2

[47] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018. 1

[48] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 3

[49] Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *ICLR*, 2020. 2

[50] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021. 5

[51] Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. When does bias transfer in transfer learning? *arXiv preprint arXiv:2207.02842*, 2022. 3

[52] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR, 2018. 4

[53] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised Learning of Debiased Representations with Pseudo-Attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16742–16751, 2022. 2, 3, 4, 5, 6

[54] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *CVPR*, 2023. 3

[55] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020. 2, 3, 4, 5

[56] Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 2020. 3

[57] Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022. 3

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

[59] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, 2022. 3

[60] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, 2022. 3

[61] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019. 3

[62] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *CVPR*, June 2020. 3, 4

[63] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017. 3

[64] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 5