Metric-Agnostic Continual Learning for Sustainable Group Fairness

Heng Lian[1], Chen Zhao[2], Zhong Chen[3], My T. Thai[4], Xingquan Zhu[5], Yi He[1]*

[1] William & Mary, [2] Baylor University, [3] Southern Illinois University, [4] University of Florida, [5] Florida Atlantic University
{hlian01, yihe}@wm.edu, chen_zhao@baylor.edu, zhong.chen@cs.siu.edu, mythai@cise.ufl.edu, xzhu3@fau.edu

Abstract

Group Fairness-aware Continual Learning (GFCL) aims to eradicate discriminatory predictions against certain demographic groups in a sequence of diverse learning tasks. This paper explores an even more challenging GFCL problem – how to sustain a fair classifier across a sequence of tasks with *covariate shifts* and *unlabeled data*. We propose the `MacFRL` solution, with its key idea to optimize the sequence of learning tasks. We hypothesize that high-confident learning can be enabled in the optimized task sequence, where the classifier learns from a set of prioritized tasks to glean knowledge, thereby becoming more capable to handle the tasks with substantial distribution shifts that were originally deferred. Theoretical and empirical studies substantiate that `MacFRL` excels among its GFCL competitors in terms of prediction accuracy and group fairness metrics. Our code and datasets are released publicly at *https://github.com/X1aoLian/MacFRL*.

Introduction

Group Fairness-aware Continual Learning (GFCL) has recently garnered significant attention due to its applications in societal decision-making (Chowdhury and Chaturvedi 2023; Truong et al. 2023b; Zhao et al. 2023). GFCL enables learning model to adapt to shifting data distributions, jointly optimizing classification accuracy and group fairness (Mehrabi et al. 2021). The key idea of GFCL is to eradicate superficial correlation between class labels and protected characteristics, such as gender, age, or ethnicity (Malleson 2018), across various tasks. To do that, Fair Representation Learning (FRL) has emerged as an important technique due to its ease of implementation and reproduction (Oh et al. 2022; Zhao et al. 2023; Chowdhury and Chaturvedi 2023; Truong et al. 2023b). Given a sequence of tasks, FRL aims to represent data from any task into a shared latent space so that 1) the data representations are invariant across all tasks, enabling continual learning, and 2) the protected feature information is not included, ensuring fair predictions.

Unfortunately, the existing FRL methods for GFCL mostly suffer from two major drawbacks. First, they heavily depend on the availability of labeled data, which are costly and time-intensive to obtain across multiple tasks, making them non-sustainable over time. Second, they falter in handling tasks with substantial covariate shifts. Consider, for example, a hiring system with different occupations as tasks. The FRL models require to learn very different representations to disentangle correlations between protected feature (*e.g.,* gender) and label (*e.g.,* hire or not), which can vary significantly across occupations. Enforcing an FRL model learned from one occupation (*e.g.,* marketing) to make prediction on another very different occupation (*e.g.,* engineering) may incur a brittle trade-off between stability and plasticity (Kim et al. 2023). That is, the model either persists in the previously learned representations and hence makes highly unfair or inaccurate predictions in the new occupation, or it adapts to the new occupation completely and forgets how to make fair predictions on all previous tasks.

In this paper, we aim to overcome the two drawbacks at once by exploring a new GFCL problem, namely, how to *sustain* a fair classifier across a sequence of tasks with *covariate shifts* and *unlabeled data*?

Our key insight to resolve the problem is drawn from human learning behaviors. We argue that, like human-beings who rarely handle tasks in arbitrary orders (Elman 1993), GFCL should follow an optimized sequence – tasks with data distributions similar to those previously seen by the classifier should be prioritized, while tasks with substantially different distributions should be deferred. Such optimized task sequence encourages high-confident learning, where the classifier after learning from multiple prioritized tasks can gradually become more knowledgeable to handle the deferred task. We cast this insight into a novel GFCL approach, termed *Metric-agnostic continual Fair Representation Learning* (`MacFRL`), which proceeds in three main steps. First, a fair classifier is initialized with its training dataset retained, and a set of subsequent tasks are fetched in the buffer. Second, `MacFRL` gauges the distance between the retained dataset and each buffered task, selecting the most similar task as the next candidate to be learnt. A shared representation space is induced from the classifier and the selected task using domain adaptation with group fairness constraints. Third, data instances predicted with high confidence from the selected task are merged into the retained dataset, enlarging the classifier's knowledge scope to prepare it for the less similar tasks. The learned task is then replaced by an incoming task in the buffer. The second and

*Corresponding author: Dr. Yi He (yihe@wm.edu)

third steps iterates until the task sequence ends.

The remaining questions are 1) how to gauge task distance with no label in the selected task and 2) how to measure the prediction confidence. For 1), `MacFRL` employs an elastic representation learning network with adaptive learning capacity, allowing it to operate without relying on any specific distance metric. The task distance is measured by the dynamics of learning invariant fair representations between the retained dataset and the selected task. If a highly complex network is required for invariance extraction, their distance is large. For 2), `MacFRL` uses density-based confidence measurement based on the representations learned from the elastic networks in different capacities, where the predicted instances with high confidence are those falling into high-density regions with the same predicted class.

Specific contributions of this paper are as follows:

i) We explore a new GFCL problem with sustainable labeling effort, where only one task is labeled to initialize the fair classifier, and all other tasks with shifted distributions are unlabeled.

ii) We propose a novel `MacFRL` algorithm with optimized task sequence to mitigate the stability and plasticity trade-off. Task distances are measured through learning dynamics of a tailored elastic network, making `MacFRL` label independent and metric agnostic.

iii) We analyze the empirical risk bounds of the elastic network design and the usefulness of task reordering, deferred to Section 2 of supplementary material.

iv) Our empirical studies on eight benchmark datasets substantiate that `MacFRL` outperforms its five state-of-the-art competitors on average by 12.7%, 42.8%, and 28.4% in terms of prediction accuracy, demographic parity, and equalized odds, respectively.

Preliminaries

Given a sequence of tasks $\{\mathcal{T}_i \mid i = 0, 1, \ldots, N\}$, in which only the first task $\mathcal{T}_0 = (\mathbf{X}_0, \mathbf{y}_0, \mathbf{p}_0)$ has labeled data, and the other tasks $\{\mathcal{T}_i = (\mathbf{X}_i, \mathbf{p}_i)\}_1^N$ remain unlabeled. Let $\mathbf{X}_i \in \mathbb{R}^{|\mathcal{T}_i| \times d}$ and $\mathbf{p}_i \in \{0,1\}^{|\mathcal{T}_i|}$ denote the d-dimensional instance vectors and the *protected* feature of the i-th task, respectively, and $\mathbf{y}_0 \in \{0,1\}^{|\mathcal{T}_0|}$ be the true labels of task $\mathcal{T}_0$. Let the joint probabilities $\mathbb{P}(Y_i, P_i) \neq \mathbb{P}(Y_j, P_j)$ for any $i \neq j$, reflecting the shifting distributions across tasks.

At each round i, the model predicts a task $\mathcal{T}_i$ and returns the predicted labels $\widehat{\mathbf{y}}_i$. After N rounds, the true labels of all tasks are revealed $\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_N$. The goal of sustainable group fairness learning is to learn a fair classifier $f : \mathbf{X} \mapsto \mathbf{y}$ with empirical risk minimization (ERM) constrained by group fairness measurement (GFM), defined as:

$$\min_f \mathbb{E}_{(\mathbf{X}_i, \mathbf{y}_i) \in \{\mathcal{T}_i\}_{i=1}^N} [\mathbf{y}_i \neq f(\mathbf{X}_i)], \text{subject to } \sum_{i=1}^{N} GFM(\mathcal{T}_i) \leq \epsilon, \tag{1}$$

where ϵ is the fairness threshold and the GFM constraint can be implemented with demographic parity (DP) (Feldman et al. 2015), equalized odds (EO) (Hardt, Price, and Srebro 2016; Alghamdi et al. 2022), or other metrics based on the domain requirements. In this paper, we employ DP and EO differences:

$$\Delta_{DP}(\mathcal{T}_i) = |\mathbb{P}(\widehat{Y}_i = 1 \mid P_i = 0) - \mathbb{P}(\widehat{Y}_i = 1 \mid P_i = 1)|,$$
$$\Delta_{EO}(\mathcal{T}_i) = \max_{Y=\{0,1\}} \{|\mathbb{P}(\widehat{Y}_i = 1 \mid P_i = 0, Y) - \mathbb{P}(\widehat{Y}_i = 1 \mid P_i = 1, Y)|\},$$

where minimizing Δ_{DP} ensures all groups enjoy equal probability of being predicted as positive. Note, Δ_{DP} focuses on the predicted results only, regardless of the prediction accuracy (*e.g.*, can incur many $\widehat{Y}_i \neq Y_i$ cases). To avoid such cases, Δ_{EO} requires the all groups have equal probabilities to be classified (or misclassified) as positive, which eliminates the negative affect.

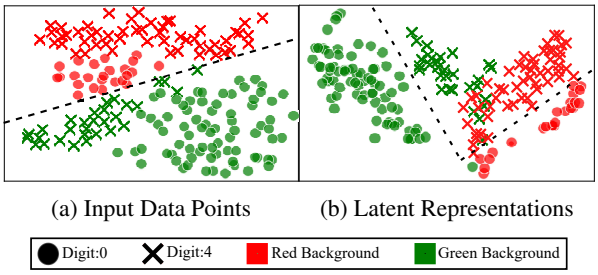
Note, our method can be easily extended to handle multi-class scenarios, including multiple protected features and multiple label categories. First, if the data contains various protected features, *e.g.*, gender and age, the group fairness difference for each feature is calculated, and the maximum value among them is reported. Second, if a protected feature or label consists of multiple classes, for instance (*e.g.*, the ethnicity feature includes various ethnic groups), the difference for each pair of ethnic groups is computed. The overall fairness measure is determined by taking the largest difference among these pairs (Denis et al. 2021).

Proposed Approach

Unsupervised Domain Adaptation with Protected Feature Obfuscation

In scenarios where new tasks arrive without labels and under different distributions, our method ensures continuous learning without performance degradation. We illustrate this with an example involving two tasks in an unsupervised domain adaptation (UDA) regime (Ganin and Lempitsky 2015; Madras et al. 2018a; Truong et al. 2023a), showcasing how our approach adapts to these challenges. Specifically, let $\mathrm{R}^{(i)}$ denote a retained dataset at the i-th round, and the labeled $\mathcal{T}_0 = \mathrm{R}^{(0)}$ for initialization. Given a new task $\mathcal{T}_i$, UDA seeks to learn a latent representation that aligns closely between $\mathrm{R}^{(0)}$ and $\mathcal{T}_i$, even in the presence of distributional discrepancies. Let $\mathbf{m} = \{0, 1\}$ denote the task membership, where $m = 0$ corresponds to instances originating from $\mathrm{R}^{(i)}$ and $m = 1$ to those from $\mathcal{T}_i$. The target is for the model to produce representations such that a classifier $D : \mathbf{X} \mapsto \mathbf{m}$ is unable to distinguish the task membership of any instance. A representation is considered *task-invariant* if a mapping ϕ effectively obfuscates its task membership.

We leverage this idea to debias the protected feature information in latent representation as well. Consider any instance $\mathbf{x} \in \mathbb{R}^{d+1}$ that includes the protected variable $p \in P$. Upon its learned task-invariant representation $\phi(\mathbf{x})$, we train a classifier $g : \mathbf{X} \mapsto \mathbf{p}$ that treats the protected feature as target variable. A maximin game ensues between ϕ and g (Zemel et al. 2013; Madras et al. 2018a; Rezaei et al. 2021), where g endeavors to maximize its prediction accuracy for p while ϕ strives to obfuscate g by minimizing its performance. The crux of this adversarial setup lies in the ability of g to predict $g(\mathbf{x}) = p$ by discerning demographic information, such as $p = 1$ or $p = 0$, from the original $\mathbf{x}$.



(a) Input Data Points (b) Latent Representations

Figure 1: T-SNE visualization of (a) original data and (b) latent representations of the `Bias-MNIST` dataset, adapted from Section 4. Shape and color represent different digital types and background colors (*i.e.,* protected feature), respectively. The dash line indicates classification boundary. The classifier in (a) is biased, because the digits are classified based on their different background color. Adversarial learning results in more fair classifier in (b), where the protected feature information is obfuscated, with classifications made based on digital types (*i.e.,* class labels).

However, if g struggles to predict the representation $\phi(\mathbf{x})$, it indicates that the protected feature information has been debiased from $\phi(\mathbf{x})$. We frame this intuition in an objective function, defined as follows.

$$\begin{aligned}\max_{g,D}\min_{\phi,f}\mathcal{L}_{\text{FairUDA}} &= \mathbb{E}_{(\mathbf{x},y)\in \mathrm{R}^{(i)}}\left[\ell(y, f(\phi(\mathbf{x})))\right] \\ &- \lambda_1 \mathbb{E}_{(\mathbf{x},p,m)\in \mathrm{R}^{(i)}\cup\mathcal{T}_i}\left[\ell(m, D(\phi(\mathbf{x}))) + \lambda_2 \ell(p, g(\phi(\mathbf{x})))\right]\end{aligned} \tag{2}$$

where f, D, and g are classifiers trained by treating the groundtruth label y, task membership m, and protected feature p as target variables, respectively. The loss function $\ell(\cdot,\cdot)$ gauges the discrepancies between the true and predicted variables. Two positive parameters λ_1 and λ_2 balance the scales of different terms. Note the minus sign that indicates the maximization of the prediction losses on m and p. After optimizing Eq. (2), we envision the learned representation $\phi(\mathbf{x})$ to i) satisfy f by enabling accurate prediction on the labeled retained dataset $\mathrm{R}^{(i)}$ and ii) obfuscate D and p so as to make $\phi(\mathbf{x})$ a task-invariant and debiased representation of the original input $\mathbf{x}$ across $\mathrm{R}^{(i)}$ and $\mathcal{T}_i$. To validate, we adapt experimental results from Section 4, as shown in Figure 1. Digit numbers 0 and 4 after optimizing Eq. (2) are represented in a latent space, where the superficial correlation between protected features (*i.e.,* background color) and labels (*i.e.,* digit types) is eliminated.

Elastic Fair Representation Learning Network

To avoid negative transfer, the model must select the new task most similar to the current retained dataset distribution, even without labels. To achieve this, we propose the elastic fair representation learning (EFRL) network, which is tailored to intermediately gauges task-wise distances. A key trait of our EFRL design is its adaptive learning capacity (Ganin and Lempitsky 2015; Long et al. 2015), which differs from traditional neural networks that mostly employ static and predetermined architecture. Unlike traditional networks that use a fixed number of hidden layers for representation learning, the network depth of EFRL is a *learnable* parameter. Intuitively, EFRL expands or contracts its depth in response to the complexity of the task at hand. Specifically, if the incoming $\mathcal{T}_i$ is more distant from $\mathrm{R}^{(i)}$ in terms of protected feature distribution, a more complex mapping ϕ is necessitated to satisfy the equilibrium of task-invariance and fairness between them as outlined in Eq. (2). Hence, EFRL responds by deepening its representation layers to approximate the required complex ϕ mapping.

To implement the intuition behind EFRL, we build an over-complete neural network consisting of L layers (with L sufficiently large). Each of its l-th layer is assigned with a weight parameter $\alpha^{(l)}$. The output of each intermediate layer is fed into the classifiers f, D, and g. The predictions from f, D, and g at the l-th layer are denoted as $\hat{y}^{(l)} = f^{(l)}(\phi^{(l)}(\mathbf{x}))$, $\hat{m}^{(l)} = D^{(l)}(\phi^{(l)}(\mathbf{x}))$, and $\hat{p}^{(l)} = g^{(l)}(\phi^{(l)}(\mathbf{x}))$, respectively. The mapping $\phi^{(l)}$ receives the representation learned from the previous layer in a recursive formulation:

$$\phi^{(l)}(\mathbf{x}) = \sigma\big(\theta_l^\top \phi^{(l-1)}(\mathbf{x})\big), \quad \phi^{(0)}(\mathbf{x}) = \mathbf{x}, \quad \forall l \in [L], \tag{3}$$

where θ_l parametrizes the l-th representation layer and σ denotes a non-linear activation such as sigmoid, ReLU, etc.

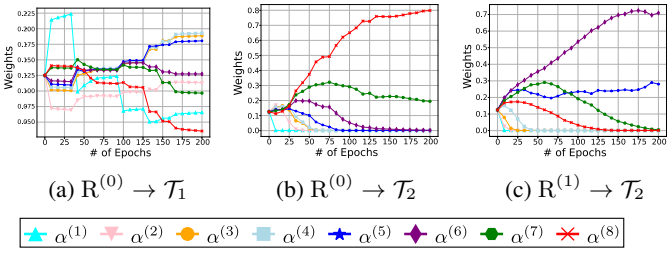
The learned depth of EFRL is reflected by the dynamics of the layer weights $\{\alpha^{(1)}, \ldots, \alpha^{(L)}\}$. In this paper, we leverage Hedge BackPropagation (HBP) (Freund and Schapire 1997; Sahoo et al. 2018) to update the weight $\alpha^{(l)}$ for each layer, with the updating function defined as follows.

$$\alpha^{(l)} \leftarrow \max\left\{\frac{s}{L}, \frac{\alpha^{(l)}\beta^{\mathcal{L}^{(l)}_{\text{EFRL}}}}{\sum_{l=1}^{L}\alpha^{(l)}\beta^{\mathcal{L}^{(l)}_{\text{EFRL}}}}\right\}, \forall \alpha^{(l)} \in (0,1), \tag{4}$$

$$\mathcal{L}^{(l)}_{\text{EFRL}} = \sum_{T}\Big[\ell\big(y, \hat{y}^{(l)}\big) - \lambda_1\big[\ell\big(m, \hat{m}^{(l)}\big) + \lambda_2 \ell\big(p, \hat{p}^{(l)}\big)\big]\Big]. \tag{5}$$

where T denotes a fixed number of training epochs across all layers. The parameters $\beta \in (0,1)$ and $s \in (0,1)$ are the *discount rate* and *smoothing threshold* of HBP, respectively, which control the aggressiveness for updating layer weights. It is trivial to observe from Eq. (4) that $\sum_L \alpha^{(l)} = 1$. We further draw comparison between Eq. (2) and Eq. (5) to observe that the EFRL network strives to expedite convergence in the maximin optimization by instilling a competitive dynamic among representations derived from all intermediate layers. This is achieved by weighting representations based on their loss performance $\mathcal{L}^{(l)}_{\text{EFRL}}$ within an epoch window of size T. Intermediate representations that ensure accurate label prediction, task invariance, and debiasing of protected features are prioritized, with higher weights assigned to the layers producing these representations.

Intuition: the learning dynamics of EFRL network. We conceptualize the layer weight dynamics during three EFRL learning phases as follows. First, at initial stage, shallower layers (denoted by smaller l) tend to dominate due to faster convergence rates. This is attributed to the *diminishing feature reuse* phenomenon (Huang et al. 2016; Larsson, Maire, and Shakhnarovich 2017), where deeper layers can dilute the



(a) $\mathrm{R}^{(0)} \rightarrow \mathcal{T}_1$ (b) $\mathrm{R}^{(0)} \rightarrow \mathcal{T}_2$ (c) $\mathrm{R}^{(1)} \rightarrow \mathcal{T}_2$

Figure 2: Layer weight dynamics of EFRL network during training. Three tasks $\mathcal{T}_0$ (*i.e.*, $\mathrm{R}^{(0)}$), $\mathcal{T}_1$, and $\mathcal{T}_2$ are reduced from the Bias-MNIST dataset, where $\mathrm{R}^{(0)}$ is more similar to $\mathcal{T}_1$ and distant from $\mathcal{T}_2$. Given different UDA settings, the layers that converge with large weights are *(a)* shallow or *(b)* deep. After learning , $\mathrm{R}^{(1)} = \mathrm{R}^{(0)} \cup \mathcal{T}_1$ gleans knowledge from it, and EFRL now *(c)* uses shallower layers for UDA.

semantic meanings of raw inputs through random parameter initialization. Second, as learning progresses, deeper layers begin to take over by gradually increasing their associated weights $\alpha^{(l)}$. This is because deeper layers with expansive learning capacity are adept at yielding representations that obscure protected features and extract task-invariant information, *i.e.*, enlarging $\ell(p, \hat{p}^{(l)})$ and $\ell(m, \hat{m}^{(l)})$, respectively. This dual capacity empowers these layers to optimize Eq. (5) at their respective depths. Third, in post-convergence phase, the weights $\alpha^{(l)}$ of excessively deep layers (*i.e.*, very high l) remain minimal. Despite their depth, these layers accrue substantial loss $\sum_T \mathcal{L}^{(l)}_{\text{EFRL}}$ over the epoch window. This loss culmination results in an experiential (albeit discounted by β) decrease in the value of $\alpha^{(l)}$.

Verification of EFRL intuitions. We carry out both theoretical and empirical studies to rationale our EFRL design. *Theoretically*, we derive Theorem 1 in Section 2.1 in Appendix, which suggests the existence of an optimal, intermediate layer $l^\star \in [1, L]$, and that our EFRL network can approximate a network trained with fixed-depth $l^\star$, while knowing the exact value of $l^\star$ across all tasks is impossible beforehand. *Experimentally*, we follow the study by (He et al. 2021) to visualize the dynamics of layer weights during training, as shown in Figure 2. We make three observations. First, while the incoming task $\mathcal{T}_1$ is close to $\mathrm{R}^{(0)}$, shallow layers would suffice to approximate a simple mapping ϕ, with deeper layers stay non-activated. Second, enforcing UDA from $\mathrm{R}^{(0)}$ to a dissimilar task $\mathcal{T}_2$ is likely to incur negative transfer, as the EFRL network ends up with assigning large weights on the deepest layers. Third, with $\mathcal{T}_2$ postponed after $\mathcal{T}_1$ learning, the shallower layers in EFRL can approximate a comparatively simpler mapping ϕ, where the deepest layer has a lower weight value. These observations support using learned layer weights in EFRL to quantify the relative distance between retained data and incoming tasks. Note, an over-complete network adjusts model depth adaptively by HBP, focusing on learning dynamics rather than enhancing the learning capacity. While NAS (Liu et al. 2018) offers a similar function, its higher computational cost limits the efficiency in CL.

Sustaining Group Fairness with Task Reordering

The layer weights, computed using labels, protected attributes, and task membership, can represent the model's depth through the largest weight. However, using only the learned EFRL network depth to quantify similarity is insufficient, considering the case that two tasks require the same number of layers for UDA with the current retained dataset $\mathrm{R}^{(i)}$. We propose to use weighted entropy Q (Guiaşu 1971) for this scalar, defined as $\mathrm{Q} = -\sum_{l=1}^{L} l \cdot \alpha^{(l)} \log \alpha^{(l)}$. Conceptually, a larger Q reflects two possible converged states – either that 1) deep layers are with large weights thus dominate the predictions or that 2) the weights of all layers follow a uniform distribution. Otherwise if Q is small, all large weights are converged to shallow layers, making deep layers trivial in learning representations. As such, a buffer with size k is used to contain k new coming tasks with k values of Q's, each of which is resulted from training EFRL between its corresponding task $\mathcal{T}_i$ and $\mathrm{R}^{(i)}$. We can now prioritize the task with the lowest Q value for UDA, transferring label information from $\mathrm{R}^{(i)}$ to $\mathcal{T}_i$ with respect to group fairness constraints. Tasks with substantial distribution shifts are deferred. Upon completion, the instances from the learned task are assimilated into $\mathrm{R}^{(i)}$ for knowledge augmentation, enriching the subsequent task selection and learning phases. The size of the buffer balances the efficacy of task reordering and the sequential nature of CL.

There are two reasons for using a selective approach to augment the retained dataset $\mathrm{R}^{(i)}$. First, the default/original task sequence may present in more harsh setting for CL. In practice, most tasks waiting in the buffer could be considerably distant from $\mathrm{R}^{(i)}$, to the extent that no instance fosters a task-invariant and debiased representation between them. Enforcing UDA between two tasks with different distributions can lead to erroneous and unfair predictions, which could cascade through subsequent tasks and compromise the entire CL pipeline. Second, as our model learns from an increasing number of tasks, the size of the retained dataset increases sharply, potentially leading to unmanageable memory overhead. To enable selective knowledge augmentation for better task sequencing, we integrate the instances predicted with high confidence, with the confidence level gauged by margin (Elsayed et al. 2018; Yan, Guo, and Zhang 2019):

$$\max_{\mathbf{x}\sim\mathcal{T}_i} \sum_{l=1}^{L} \alpha^{(l)} \frac{\sum_{\hat{y}^{(l)}=\hat{y}^j} \exp\left[-\|\phi^{(l)}(\mathbf{x}) - \phi^{(l)}(\mathbf{x}^j)\|_2\right]}{\sum_{\mathbf{x}^j\in\mathcal{N}(\mathbf{x})} \exp\left[-\|\phi^{(l)}(\mathbf{x}) - \phi^{(l)}(\mathbf{x}^j)\|_2\right]}, \quad (6)$$

where $\mathcal{T}_i$ is the selected task. At the l-th layer, $\mathbf{x}^j \in \mathcal{N}(\mathbf{x})$ is a data point in $\mathbf{x}$'s nearest neighbors, predicted as $\hat{y}^j = f^{(l)}(\phi^{(l)}(\mathbf{x}^j))$. From a geometric perspective, the prediction confidence of an instance correlates with its margin from the decision boundary. Eq. (6) reflects this measurement. In Eq. (6), an instance $\mathbf{x}$ garners high value (confidence) if 1) it is close to its nearest neighbors, indicating dense cluster (low denominator) and 2) its neighbors mostly fall within the same class, signifying a large margin (high numerator). Integrating instances with a large margin thus equates to choos-

ing those predicted with high confidence. The continuum of EFRL training, task prioritizing, and selective knowledge augmentation executes until the task sequence is exhausted.

Experiments

Data Sets. Eight real-world datasets from various domains set up the benchmark, with their statistics summarized in the table below. We follow (Le Quy et al. 2022) to define the protected features. Details of the studied datasets are deferred to Section 3 of supplementary material.

No.	Dataset	# Samples	# Features	# Tasks	$\mathbf{y}\|0:1$	$\mathbf{p}\|0:1$
1	Adult	30010	15	12	75:25	32:68
2	KDD Census-Income	199523	41	9	94:6	52:48
3	Bank marketing	31647	17	12	88:12	40:60
4	Dutch census	42125	12	10	52:48	50:50
5	Diabetes	71236	50	9	54:46	46:54
6	Law School	14298	23	6	5:95	16:84
7	Bias-MNIST	60000	$28 \times 28 \times 3$	5	10: . . . :10	68:32
8	CelebA	100000	$178 \times 218 \times 3$	5	49:51	42:58

Table 1: Statistics of the 8 datasets.

Competitors. Five rival models are employed for comparative study. *ULLC* (He et al. 2021) is a CL method which only focuses on maximizing accuracy. Group fairness constraints are applied on learned representations after training. *FaDL* (Zhang, Lemoine, and Mitchell 2018) employs adversarial training to debias intermediate representation with fully labeled data. *FaIRL* (Chowdhury and Chaturvedi 2023) prevents forgetting in CL with data replay. Partial instances from previous tasks are randomly sampled for later tasks. It postulates full access to labels. *UnFaIRL* ablates FaIRL by removing the labels of subsequent tasks. To wit the performance skyline, we let the method *Skyline* jointly learn all tasks in an offline, multi-task learning setting with full labels, building an upper bound in both accuracy and fairness.

Metrics. We use three metrics tailored for continual learning. For prediction accuracy, we use the average accuracy (Lopez-Paz and Ranzato 2017) across all tasks up to the current task $\mathcal{T}_i$, defined as Accuracy $= 1/N \sum_i^N \text{Acc}(\mathcal{T}_i)$, where $\text{Acc}(\mathcal{T}_i)$ returns the accuracy on $\mathcal{T}_i$. We extend the group fairness metrics Δ_{DP} and Δ_{EO} in CL contexts as follows: DP $= \sum_i^N |\omega_i \Delta_{DP}(\mathcal{T}_i)|$, EO $= \sum_i^N |\omega_i \Delta_{EO}(\mathcal{T}_i)|$, and $\omega_i = |\mathcal{T}_i| / \sum_i^N |\mathcal{T}_i|$, where $\Delta_{DP}(\mathcal{T}_i)$ and $\Delta_{EO}(\mathcal{T}_i)$ return the demographic parity and equalized odds differences on the predicted $\mathcal{T}_i$, respectively, as defined in Section 2. $|\mathcal{T}_i|$ denotes the number of instances in $\mathcal{T}_i$. ω_i represents the proportion of each task's sample size in the entire dataset, which alleviates the negative impact of different task sizes. For multiple-class datasets such as `Bias-MNIST`, we follow (Hardt, Price, and Srebro 2016) to take the most unfair class that returns maximal EO value for calculation.

RQ 1: How does our `MacFRL` approach compare to the state-of-the-art group fairness methods?

We make three observations from Table 2 and Figure 3. We compare our `MacFRL` with three fairness-oriented competitors, FaDL, FaIRL and UnFaIRL. To ensure level comparison, confidently labeled instances are replayed across all three methods. Against FaDL, `MacFRL` outperforms on all 8 datasets in terms of accuracy, exceeding FaDL by over 5% on `KDD Census-income`, `Dutch`, `Diabetes`, and `CelebA`. `MacFRL` leads on 4 and 7 datasets for DP and EO, respectively. Compared to FaIRL, `MacFRL` surpasses in 20 out of 21 results, notably achieving an average DP reduction of 12.4%. While UnFaIRL matches our accuracy, it falls short in DP and EO in 14 of 16 cases. Second, Skyline and ULLC, which are evaluated in less restrictive settings, outperform our `MacFRL` only at minor margins. Skyline only outperforms `MacFRL` on both EO and accuracy in 2 out of 8 datasets, with 1% and 5% increases, respectively. `MacFRL` enjoys the highest accuracy with 93.9% and lowest DP and EO with 2% and 9.6% on `Law School`, respectively. ULLC only excels in accuracy on two datasets, however, `MacFRL` outperforms ULLC in DP and EO by decreasing them by 12.3% and 12.7% on average, respectively. Third, in `Bias-MNIST`, `MacFRL` only ties with Skyline by achieving 92.9% prediction accuracy and 14.1% EO, which outperforms all other methods on average by 3.0% and 17.9%, respectively. In `CelebA`, `MacFRL` lowers the values of DP and EO of ULLC, from 40.9% and 26.3% to 22.3% and 19.1%, respectively. FaDL outperforms `MacFRL` in DP but sacrifices accuracy and EO, which are 9% and 16.2% lower, respectively. These results demonstrate the superior generalization performance of our `MacFRL` on both traditional tabular data and high-dimensional images in a CL context with one-time labeling effort only.

RQ 2: How does similarity-based task reordering sustain group fairness in continual learning?

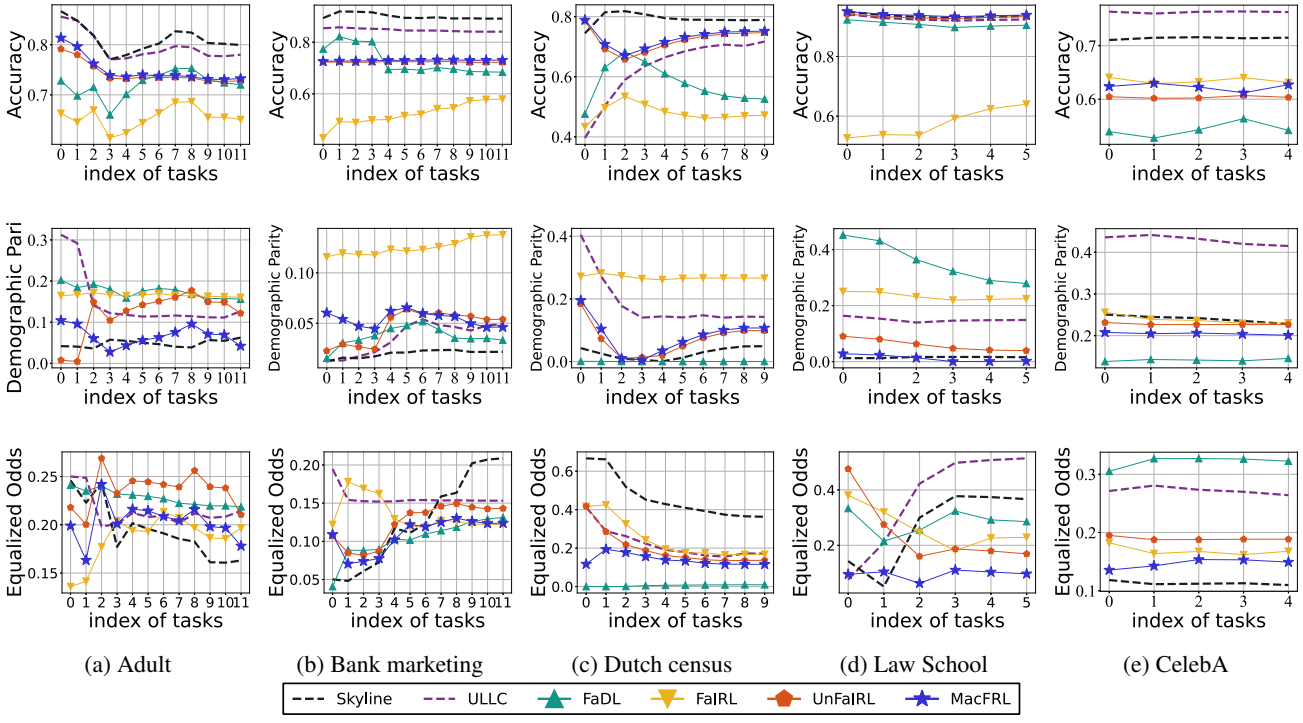
We answer this question by comparing our `MacFRL` with FaDL, FaIRL, and UnFaIRL. First, although none of the three competitors use re-ordering, FaIRL performs worst, with higher DP and standard deviations across all settings, indicating its inability to ensure fairness and accuracy in new tasks with distribution shifts. Second, FaDL performs better than FaIRL but stays inferior to our `MacFRL`. `MacFRL` outperforms FaDL in 20 out of 24 settings. Particularly in `Dutch census`, the accuracy of FaDL is only 52.6% while that of `MacFRL` is 75.2%. Although FaDL performs better accuracy on the first four tasks of `Bank marketing`, its result declines after undergoing the learning of $\mathcal{T}_4$, and becomes worse than ours in Figure 3b. Without re-ordering, their learning process cannot avoid negative transfer caused by $\mathcal{T}_4$. To compare, although our method makes more mistakes during $\mathcal{T}_1$ and $\mathcal{T}_2$, these mistakes do not impact the model performance on other tasks. Third, as shown in Table 2, `MacFRL` outperforms UnFaIRL across most settings and achieves higher accuracy on all datasets. This advantage is especially evident in `Diabetes` and `Law School`, where `MacFRL` achieves DP of 0.4% and 0.2% compared to UnFaIRL's 4.7% and 3.4%, respectively. These results suggest that `MacFRL` can offer superior prediction accuracy and group fairness because of the reorder method.

Table 2: Comparative results on 8 datasets with 3 metrics in mean ± standard deviation format. Bold values represent the best results except for Skyline with less restrictive settings.

No.	Dataset	Skyline	ULLC	FaDL	FaIRL	UnFaIRL	MacFRL
Evaluation Metric = Accuracy (↑) where higher values are better.							
1	Adult	.799 ± .000	**.780 ± .006**	.720 ± .001	.651 ± .082	.728 ± .007	.733 ± .013
2	KDD Census-Income	.944 ± .000	**.792 ± .013**	.678 ± .000	.605 ± .133	.714 ± .006	.722 ± .006
3	Bank marketing	.891 ± .000	.710 ± .022	.684 ± .000	.580 ± .072	.722 ± .006	**.730 ± .006**
4	Dutch census	.789 ± .000	.717 ± .011	.526 ± .000*	.472 ± .116	.747 ± .001	**.752 ± .001**
5	Diabetes	.618 ± .000	.565 ± .005	.459 ± .002	.501 ± .030	.584 ± .001	**.590 ± .001**
6	Law School	.936 ± .000	.924 ± .006	.905 ± .000	.640 ± .088	.933 ± .001	**.939 ± .002**
7	Bias-MNIST	.973 ± .087	.913 ± .062	.888 ± .066	N/A**	.895 ± .078	**.929 ± .088**
8	CelebA	.715 ± .005	.762 ± .002	.542 ± .058	**.631 ± .024**	.603 ± .021	.630 ± .032
Evaluation Metric = Demographic Parity (↓) where lower values are better.							
1	Adult	.062 ± .000	.127 ± .038	.155 ± .001	.160 ± .095	.122 ± .023	**.042 ± .007**
2	KDD Census-Income	.003 ± .000	.170 ± .016	.067 ± .000	.122 ± .093	.041 ± .011	**.020 ± .003**
3	Bank marketing	.016 ± .000	.049 ± .020	.032 ± .000	.138 ± .127	**.022 ± .004**	.046 ± .005
4	Dutch census	.049 ± .000	.143 ± .015	**.011 ± .001***	.266 ± .246	.099 ± .002	.107 ± .005
5	Diabetes	.043 ± .000	.042 ± .034	.037 ± .012	.137 ± .106	.047 ± .004	**.004 ± .001**
6	Law School	.016 ± .000	.149 ± .018	.071 ± .000	.225 ± .153	.034 ± .001	**.002 ± .000**
7	Bias-MNIST	.136 ± .022	.302 ± .071	.267 ± .054	N/A**	.239 ± .061	**.151 ± .010**
8	CelebA	.228 ± .017	.409 ± .001	**.145 ± .012**	.231 ± .016	.226 ± .008	.215 ± .005
Evaluation Metric = Equalized Odds (↓) where lower values are better.							
1	Adult	.163 ± .000	.214 ± .018	.218 ± .001	.196 ± .082	.210 ± .014	**.178 ± .005**
2	KDD Census-Income	.220 ± .000	.256 ± .021	.087 ± .000	.168 ± .101	.163 ± .011	**.077 ± .003**
3	Bank marketing	.209 ± .000	.153 ± .031	.131 ± .000	.124 ± .094	.143 ± .005	**.122 ± .008**
4	Dutch census	.362 ± .000	.171 ± .016	**.009 ± .000***	.167 ± .155	.136 ± .001	.115 ± .007
5	Diabetes	.038 ± .000	.078 ± .029	.027 ± .001	.073 ± .032	.060 ± .001	**.022 ± .002**
6	Law School	.366 ± .000	.514 ± .025	.285 ± .000	.228 ± .133	.168 ± .002	**.096 ± .001**
7	Bias-MNIST	.139 ± .043	.333 ± .110	.319 ± .092	N/A**	.243 ± .098	**.141 ± .100**
8	CelebA	.110 ± .008	.263 ± .002	.322 ± .016	.167 ± .019	.190 ± .011	**.160 ± .012**

** N/A indicates that FaIRL is not applicable on the Bias-MNIST dataset as it is tailored for binary classification, while Bias-MNIST has ten class labels.

*: Note, FaDL suffers substantial tradeoff between accuracy and fairness; in settings where FaDL obtains the best DP/EO performance, it incurs substantial accuracy decrease.



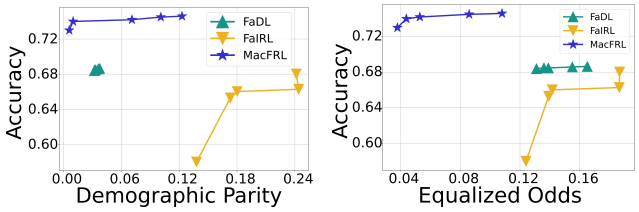
(a) Adult (b) Bank marketing (c) Dutch census (d) Law School (e) CelebA

Figure 3: The trends of Accuracy (top row), Demographic Parity (middle row) and Equalized Odds (bottom row) of our MacFRL approach and its 5 competitors on 5 datasets w.r.t. the input sequence of tasks.

RQ 3: How sensitive are the tuned parameters to the tradeoff between accuracy and fairness?

We first evaluate the accuracy-fairness tradeoff on Bank marketing by sweeping λ_2 in [0.1, 0.08, 0.06, 0.04, 0.02]. Figure 4 shows the tradeoff curves (left to right) for all three methods as λ_2 decreases. This hyper-parameter also controls

the balance for FaIRL and FaDL, the same range is used. FaDL demonstrates minimal sensitivity to changes in λ_2, maintaining stable accuracy but at the cost of lower fairness, as indicated by small improvements in DP and EO. On the other hand, FaIRL exhibits that its accuracy decreasing as fairness improves, which is more sensitive to λ_2 change. Our method `MacFRL` also shows that the larger (small) λ_2, the better (worse) model fairness and the worse (better) model accuracy. Moreover, MacFRL maintains the highest accuracy among the three methods while still improving fairness. Second, experimental results of `Bank marketing` shown in Table 3 demonstrate the impact of λ_1. We can observe that increasing λ_1 from 0.01 to 0.05 improves fairness that DP drops from 8.4% to 4.6% and EO from 7.9% to 6.4%. However, the model tends to focus more on similar representations with increasing λ_1 to 0.5, degrading DP to 17.3% and EO to 18.1% and accuracy to 71.3%. When λ_1 achieves 1, the model collapses, with accuracy at 65%, DP at 59.8%, and EO at 60.5%.



(a) Accuracy vs DP (b) Accuracy vs EO

Figure 4: Results of accuracy-fairness tradeoffs on `Bank marketing` sweeping over a range of λ_2.

λ_1		0.01	0.03	0.05	0.1	0.5	1
Bank marketing	Acc	.745	.740	.730	.721	.713	.650
	DP	.084	.055	.046	.122	.173	.598
	EO	.079	.070	.064	.142	.181	.605

Table 3: Results of accuracy-fairness tradeoffs on `Bank marketing` sweeping over a range of λ_1

Related Work

Fair Representation Learning (FRL) Fairness issues in data-driven models arise at various stages, including data preparation, model training, and user interaction (Mehrabi et al. 2021). FRL methods focus on ensuring fair predictive modeling by extracting debiased intermediate representations from biased raw data, as introduced by (Zemel et al. 2013). While FRL can address both group and individual fairness, our study emphasizes group fairness, aiming for equitable treatment of protected demographic groups (Malleson 2018; Barocas and Selbst 2016). Later studies (Louizos et al. 2015; Moyer et al. 2018; Jaiswal et al. 2018; Xu et al. 2018; Madras et al. 2018b; Amini et al. 2019) leveraged deep generative models, using discriminators to distinguish protected groups. Adversarial training for group fairness has been explored since (Edwards and Storkey 2016; Beutel et al. 2017; Madras et al. 2018a; Elazar and Goldberg 2018; Zhang, Lemoine, and Mitchell 2018), aiming to obfuscate protected feature information. However, most existing FRL methods focus on a single task (Barrett et al. 2019) and struggle with adapting to tasks with distribution shifts. Recent studies (Jing, Xu, and Ding 2021; Singh et al. 2021; Paul et al. 2022; Deka and Sutherland 2023) propose aligning intermediate representations across tasks based on similarities. Modeling shifted distributions as a weighted combination of training data is proposed in (Mandal et al. 2020) to minimize worst-case fairness loss. However, these methods rely on controlled task-wise distances, which may not hold as tasks diverge significantly in continual task sequences, leading to fair adaptation failures.

Continual Learning (CL) CL aims to build systems that learn incrementally (Kirkpatrick et al. 2017; Li and Hoiem 2017; Rolnick et al. 2019; Hao et al. 2013; Mitchell et al. 2018; Abujabal et al. 2018), addressing catastrophic forgetting, where new knowledge disrupts previously learned information. CL methods generally fall into three categories: First, regularization-based methods, which regularize model parameters to avoid drastic updates, searching for Pareto-effective solutions that balance performance across tasks, thus mitigating forgetting (Kirkpatrick et al. 2017; Aljundi, Chakravarty, and Tuytelaars 2017; Shmelkov, Schmid, and Alahari 2017; Li and Hoiem 2017; Aljundi et al. 2018). Second, rehearsal methods, which store instances from previous tasks in external memory (*i.e.*,the retained dataset) for joint training with current task instances (Gepperth and Karaoguz 2016; Schaul et al. 2016; Rebuffi et al. 2017; Lopez-Paz and Ranzato 2017; Rolnick et al. 2019; Hayes, Cahill, and Kanan 2019). Third, model expansion, which expands the model by increasing the network size (Li et al. 2019; Rao et al. 2019; Zhao et al. 2022), or designing sub-networks for each task (Ke, Liu, and Huang 2020; Mallya and Lazebnik 2018; Serra et al. 2018; Wang et al. 2020). Recent works address gradient interference between tasks via scaled gradient projection (Saha and Roy 2023) or leverage pre-trained models (PTMs) instead of random initialization (McDonnell et al. 2024). However, most CL methods prioritize classification accuracy over group fairness. FaIRL (Chowdhury and Chaturvedi 2023) addresses both issues using task rehearsal but assumes all tasks are fully labeled. Our `MacFRL` removes these assumptions, requiring labels from one initial task, making the CL more cost-effective and sustainable.

Conclusion

This paper presents `MacFRL`, a novel algorithm to sustain group fairness in continual learning, with all incoming tasks unlabeled. The key idea of `MacFRL` lies in its strategic task reordering inspired by human learning, prioritizing similar tasks to glean knowledge and become gradually more capable to handle the originally deferred, more dissimilar tasks. We analyzed the theoretical risk bounds of `MacFRL` to rationalize the design of task sequence optimization. Extensive experiments on eight benchmark datasets substantiate the viability, effectiveness, and sustainability of `MacFRL` in both accuracy and group fairness metrics.

Acknowledgement

This work has been supported in part by the National Science Foundation (NSF) under Grant Nos. IIS-2441449, IIS-2236578, IOS-2446522, IIS-2236579, IIS-2302786, IOS-2430224, SCH-2123809, and IIS-1939725, and also supported in part by the Commonwealth Cyber Initiative (CCI).

References

Abujabal, A.; Saha Roy, R.; Yahya, M.; and Weikum, G. 2018. Never-ending learning for open-domain question answering over knowledge bases. In *WWW*, 1053–1062.

Alghamdi, W.; Hsu, H.; Jeong, H.; Wang, H.; Michalak, P. W.; Asoodeh, S.; and Calmon, F. 2022. Beyond Adult and COMPAS: Fair Multi-Class Prediction via Information Projection. In *NeurIPS*.

Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; and Tuytelaars, T. 2018. Memory Aware Synapses: Learning What (Not) to Forget. In *ECCV*. Berlin, Heidelberg: Springer-Verlag.

Aljundi, R.; Chakravarty, P.; and Tuytelaars, T. 2017. Expert gate: Lifelong learning with a network of experts. In *CVPR*, 3366–3375.

Amini, A.; Soleimany, A. P.; Schwarting, W.; Bhatia, S. N.; and Rus, D. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *AIES*, 289–295.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California law review*, 671–732.

Barrett, M.; Kementchedjhieva, Y.; Elazar, Y.; Elliott, D.; and Søgaard, A. 2019. Adversarial Removal of Demographic Attributes Revisited. In *EMNLP-IJCNLP*, 6330–6335. Association for Computational Linguistics.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Chowdhury, S. B. R.; and Chaturvedi, S. 2023. Sustaining Fairness via Incremental Learning. In *AAAI*, volume 37, 6797–6805.

Deka, N.; and Sutherland, D. J. 2023. MMD-B-Fair: Learning Fair Representations with Statistical Testing. In *AISTATS*, 9564–9576. PMLR.

Denis, C.; Elie, R.; Hebiri, M.; and Hu, F. 2021. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*.

Edwards, H.; and Storkey, A. 2016. Censoring representations with an adversary. In *ICLR*.

Elazar, Y.; and Goldberg, Y. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *EMNLP*, 11–21.

Elman, J. L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1): 71–99.

Elsayed, G.; Krishnan, D.; Mobahi, H.; Regan, K.; and Bengio, S. 2018. Large margin deep networks for classification. *NeurIPS*, 31.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268.

Freund, Y.; and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189. PMLR.

Gepperth, A.; and Karaoguz, C. 2016. A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, 8(5): 924–934.

Guiaşu, S. 1971. Weighted entropy. *Reports on Mathematical Physics*, 2(3): 165–179.

Hao, Y.; Chen, Y.; Zakaria, J.; Hu, B.; Rakthanmanon, T.; and Keogh, E. 2013. Towards never-ending learning from time series streams. In *KDD*, 874–882.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hayes, T. L.; Cahill, N. D.; and Kanan, C. 2019. Memory efficient experience replay for streaming learning. In *ICRA*, 9769–9776. IEEE.

He, Y.; Chen, S.; Wu, B.; Yuan, X.; and Wu, X. 2021. Unsupervised lifelong learning with curricula. In *WWW*, 3534–3545.

Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; and Weinberger, K. Q. 2016. Deep networks with stochastic depth. In *ECCV*, 646–661. Springer.

Jaiswal, A.; Wu, R. Y.; Abd-Almageed, W.; and Natarajan, P. 2018. Unsupervised adversarial invariance. *NeurIPS*, 31.

Jing, T.; Xu, B.; and Ding, Z. 2021. Towards fair knowledge transfer for imbalanced domain adaptation. *IEEE Transactions on Image Processing*, 30: 8200–8211.

Ke, Z.; Liu, B.; and Huang, X. 2020. Continual learning of a mixed sequence of similar and dissimilar tasks. *NeurIPS*, 33: 18493–18504.

Kim, S.; Noci, L.; Orvieto, A.; and Hofmann, T. 2023. Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *CVPR*, 11930–11939.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2017. FractalNet: Ultra-Deep Neural Networks without Residuals. In *ICLR*.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3): e1452.

Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to grow: A continual structure learning framework for

overcoming catastrophic forgetting. In *ICML*, 3925–3934. PMLR.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.-J.; Fei-Fei, L.; Yuille, A.; Huang, J.; and Murphy, K. 2018. Progressive neural architecture search. In *ECCV*, 19–34.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*, 97–105. PMLR.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *NeurIPS*, 6467–6476.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018a. Learning adversarially fair and transferable representations. In *ICML*, 3384–3393. PMLR.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018b. Learning Adversarially Fair and Transferable Representations. In *ICML*.

Malleson, K. 2018. Equality law and the protected characteristics. *The Modern Law Review*, 81(4): 598–621.

Mallya, A.; and Lazebnik, S. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, 7765–7773.

Mandal, D.; Deng, S.; Jana, S.; Wing, J.; and Hsu, D. J. 2020. Ensuring fairness beyond the training data. *NeurIPS*, 33: 18445–18456.

McDonnell, M. D.; Gong, D.; Parvaneh, A.; Abbasnejad, E.; and van den Hengel, A. 2024. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Yang, B.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; et al. 2018. Never-ending learning. *Communications of the ACM*, 61(5): 103–115.

Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; and Ver Steeg, G. 2018. Invariant representations without adversarial training. *NeurIPS*, 31.

Oh, C.; Won, H.; So, J.; Kim, T.; Kim, Y.; Choi, H.; and Song, K. 2022. Learning fair representation via distributional contrastive disentanglement. In *KDD*, 1295–1305.

Paul, W.; Hadzic, A.; Joshi, N.; Alajaji, F.; and Burlina, P. 2022. Tara: training and representation alteration for ai fairness and domain generalization. *Neural Computation*, 34(3): 716–753.

Rao, D.; Visin, F.; Rusu, A.; Pascanu, R.; Teh, Y. W.; and Hadsell, R. 2019. Continual unsupervised representation learning. *NeurIPS*, 32.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *CVPR*, 2001–2010.

Rezaei, A.; Liu, A.; Memarrast, O.; and Ziebart, B. D. 2021. Robust fairness under covariate shift. In *AAAI*, volume 35, 9419–9427.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. In *NeurIPS*, 348–358.

Saha, G.; and Roy, K. 2023. Continual learning with scaled gradient projection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9677–9685.

Sahoo, D.; Pham, Q.; Lu, J.; and Hoi, S. C. 2018. Online deep learning: learning deep neural networks on the fly. In *IJCAI*, 2660–2666.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2016. Prioritized experience replay. In *ICLR*.

Serra, J.; Suris, D.; Miron, M.; and Karatzoglou, A. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, 4548–4557. PMLR.

Shmelkov, K.; Schmid, C.; and Alahari, K. 2017. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 3400–3409.

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2021. Fairness violations and mitigation under covariate shift. In *FACCT*, 3–13.

Truong, T.-D.; Le, N.; Raj, B.; Cothren, J.; and Luu, K. 2023a. Fredom: Fairness domain adaptation approach to semantic scene understanding. In *CVPR*, 19988–19997.

Truong, T.-D.; Nguyen, H.-Q.; Raj, B.; and Luu, K. 2023b. Fairness continual learning approach to semantic scene understanding in open-world environments. In *NeuIPS*, volume 36, 65456–65467.

Wang, Z.; Jian, T.; Chowdhury, K.; Wang, Y.; Dy, J.; and Ioannidis, S. 2020. Learn-prune-share for lifelong learning. In *ICDM*, 641–650. IEEE.

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. Fairgan: Fairness-aware generative adversarial networks. In *Big Data*, 570–575. IEEE.

Yan, Z.; Guo, Y.; and Zhang, C. 2019. Adversarial margin maximization networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(4): 1129–1139.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *ICML*, 325–333.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*, 335–340.

Zhao, C.; Mi, F.; Wu, X.; Jiang, K.; Khan, L.; Grant, C.; and Chen, F. 2023. Towards fair disentangled online learning for changing environments. In *KDD*.

Zhao, T.; Wang, Z.; Masoomi, A.; and Dy, J. 2022. Deep bayesian unsupervised lifelong learning. *Neural Networks*, 149: 95–106.