Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition

Preprint · March 2021

CITATIONS
0

READS
77

3 authors, including:

Nikhil Churamani
University of Cambridge
29 PUBLICATIONS **184** CITATIONS

SEE PROFILE

Hatice Gunes
University of Cambridge
133 PUBLICATIONS **4,030** CITATIONS

SEE PROFILE

Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition

Nikhil Churamani*, Ozgur Kara†, and Hatice Gunes*

***Abstract*—As Facial Expression Recognition (FER) systems become integrated into our daily lives, these systems need to prioritise making *fair* decisions instead of aiming at higher individual accuracy scores. Ranging from surveillance systems to diagnosing mental and emotional health conditions of individuals, these systems need to balance the *accuracy vs fairness* trade-off to make decisions that do not unjustly discriminate against specific under-represented demographic groups. Identifying *bias* as a critical problem in facial analysis systems, different methods have been proposed that aim to mitigate bias both at data and algorithmic levels. In this work, we propose the novel usage of Continual Learning (CL), in particular, using Domain-Incremental Learning (Domain-IL) settings, as a potent bias mitigation method to enhance the *fairness* of FER systems while guarding against biases arising from skewed data distributions. We compare different non-CL-based and CL-based methods for their classification *accuracy* and *fairness scores* on expression recognition and Action Unit (AU) detection tasks using two popular benchmarks, the RAF-DB and BP4D datasets, respectively. Our experimental results show that CL-based methods, on average, outperform other popular bias mitigation techniques on both *accuracy* and *fairness* metrics.**

***Index Terms*—Fairness, Continual Learning, Bias Mitigation, Facial Expression Recognition, Facial Action Units, Affective Computing.**

I. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) systems are increasingly becoming an important part of human life, monitoring and controlling several aspects of our daily lives, with little to no human oversight. From security and surveillance systems that deploy several ML models such as face detection and recognition systems [1], social media platforms that auto-tag pictures of our friends and family [2], recommender systems that track our digital footprints to show us advertisements of products that we might like to indulge in [3], to banking and finance applications that work on credit approvals based on socio-economical backgrounds of individuals, AI systems are ubiquitous, making 'smart' decisions about several critical aspects of our lives [4], [5]. It is thus important to ensure that these systems make fair and unbiased decisions to avoid potentially catastrophic consequences that adversely affect individuals [6]. In this work, we focus on one such popular application of AI in real life; Facial Expression Recognition (FER) systems.

FER systems (see [7], [8], [9] for a survey) aim to analyse facial expressions either by encoding facial muscle activity as Facial Action Units (AUs) [10] or determining the emotional state being expressed by an individual [11], [12]. Analysing large datasets of human faces, annotated for the expressions represented in the images, these models are heavily data-dependent and thus may be prone to *biases* originating from imbalances in the training data distribution. For a large variety of FER datasets, attributes such as gender, race, age or skin-colour are implicitly encoded in the data which may also be learnt by a (deep) learning model [13]. If these attributes are not balanced across the entire distribution of the dataset, the model may learn to associate such *confounding* attributes with the task of FER. For example, if the training data has a disproportionate number of images of Males expressing 'Happy' than Females, the model may learn to associate gender with the expression, leading to a lot of 'happy female' samples being misclassified.

While the most effective method for preventing biases in FER datasets would be to ensure a balanced and representative data collection, this also turns out to be the most challenging problem. Owing to restrictions with respect to data recording settings, personal preferences, geographic location as well as several social and cultural constraints, it may not always be possible to ensure a balanced data collection. Most recent datasets try to ensure the data collection is fair and unbiased or at the least provide demographic annotations, along with affective labels, that enable researchers to make informed decisions while using these datasets for training ML models [13]. Yet, to ensure fairness despite the inherent imbalances in data distributions, several methods have been proposed that handle these imbalances at the *pre-processing*, *in-processing* or *post-processing* levels [14].

Pre-processing methods focus on *strategically sampling* training data, that is, given the distribution of data with respect to a selected demographic attribute, samples belonging to under-represented groups are either over-sampled compared to dominant groups [15], or scaled penalties are applied when a model incorrectly classifies these samples [16]. Yet, these methods are not perfect and some bias might still creep in. To

*N. Churamani and H. Gunes are with the Department of Computer Science and Technology, University of Cambridge, United Kingdom.
E-mail: {nikhil.churamani, hatice.gunes}@cl.cam.ac.uk

†O. Kara is with the Electrical & Electronics Engineering Department at the Bogazici University, Istanbul, Turkey.
E-mail: ozgur.kara@boun.edu.tr

N. Churamani is funded by the EPSRC under grant EP/R513180/1 (ref. 2107412). H. Gunes' work is supported by the EPSRC under grant ref. EP/R030782/1 and partially by the European Union's Horizon 2020 Research and Innovation programme under grant agreement No. 826232. O. Kara contributed to this work while undertaking a summer research study at the Department of Computer Science and Technology, University of Cambridge. Authors also thank Prof Lijun Yin from Binghamton University (USA) for providing access to the BP4D Dataset and the relevant race attributes; and Shan Li, Prof Weihong Deng and JunPing Du from the Beijing University of Posts and Telecommunications (China) for providing access to RAF-DB.

handle that, changes to the model architecture or the training regime needs to be made. Algorithm-level or *In-Processing* methods achieve this either by explicitly learning domain-specific information such that this can be discounted from the model's learning later [17] or they learn to completely discount domain-specific information by omitting these features from the learnt representations [18]. Post-processing methods, on the other hand, are mostly used to quantify bias in trained algorithms [14] and offer effective tools to evaluate the *fairness* of an ML model.

Interestingly, the underpinning principle behind all the above-mentioned methods is essentially to focus on learning and adapting to the inherent imbalances in data distribution, either by *synthetically* balancing it or adjusting the learning algorithm itself to account for these imbalances. This principle is shared by Continual Learning (CL) methods [19], [20] that also aim to balance learning in the model by being sensitive to shifts in data distributions, ensuring that one particular domain or task does not dominate the model's learning. Their ability to *continually* learn and adapt to novel information, aggregating new knowledge without impacting previously acquired information, may allow them to balance learning across the different learning domains. Domain-Incremental CL settings [21] particularly focus on managing shifts in input data distribution while the task remains the same. This can be considered analogous to solving FER tasks where input data belongs to different domains of gender (male, female) and race (black, white, asian). The challenge for CL models will thus be to maintain performance on FER tasks with respect to one domain while acquiring information about new domains.

Motivated by this notion, we propose the novel use of CL as a learning paradigm that is well-suited for developing *fairer* FER models that can balance learning with respect to different attributes of gender and race. We formulate expression recognition and AU detection across these different domain attributes as a continual learning problem and compare several popular CL approaches with state-of-the-art bias mitigation approaches. To the best of our knowledge, this is the first application of CL learning as a bias mitigation strategy for facial affect analysis tasks. Furthermore, we explore the Domain-Incremental CL settings [21] where the models need to solve the facial analysis tasks across different domains, defined by the demographic attributes of *gender* and *race*. For each attribute, the data is split into different domains, that is, gender annotations are used to define *male* and *female* domains, while race annotations are used to split data into *White/Caucasian, Black/African-American, Asian* and *Latino* domains. We primarily focus on regularisation-based CL methods as these do not require setting up additional memory or computational resources, allowing a fair and direct comparison with other learning methods that focus on mitigating bias arising due to imbalances in data distributions. Our experimental results show that CL-based approaches are, on average, seen to outperform other bias mitigation strategies, both in terms of accuracy as well as fairness scores for both FER and AU detection tasks across the domain-splits.

II. Background and Related Work

A. Understanding Bias

Bias, both in human perception and behaviour as well as ML algorithms can be characterised as an inclination or prejudice towards a person or a group, that may be considered unfair. This may result from an over or under-exposure of an individual towards a certain group of individuals usually characterised by their gender, racial identity, social or economical background or age, amongst other factors. This exposure results in people considering individuals that share similar characteristics as themselves as "in-group" members and others different from them as "out-group" members [22]. It is seen that people tend to be *biased* in favour of in-group members, evaluating them more positively on dimensions of judgement while being negatively *biased* or *prejudiced* against out-group members [23].

Understanding how humans consider in-group and out-group members [22] in their immediate surroundings and base their decisions on aspects such as gender, race or age is important to view ML models in the right perspective when applied to real-world settings. Such an understanding will allow researchers to assess what may be considered 'fair' and how to achieve such fairness in algorithms.

1) ***Bias in Human Perception****:* Following the Perception-Action model of Empathy [24], an individual's behaviour, particularly their facial expressions and body gestures stimulate a similar neural activation in the observer, enabling them to empathise with and understand their actions, intentions and emotions. Gutsell et al. [25], through a series of experiments with multiple participants (30 White university students) interacting with in-group (in this case participants with a *Caucasian* ethnic identity) and out-group (excluded from this circle; in this case, *African–Canadian, East-Asian, South-Asian* ethnic identities) members, concluded that such perception-action couplings are reserved only for in-group members. In-group identification, that is, identifying other individuals to be sharing similar characteristics as oneself, causes a *positive association* with them [23]. Furthermore, people have a harder time recognising the faces of out-group members and interpret their facial expressions [26], [27].

In the case of ML algorithms, we may understand such 'inter-group bias' to result from the imbalances in data distributions where certain groups may be considered to constitute 'in-group' attributes due to their dominance in the data, while under-represented attributes can be considered as members of the 'out-group'. Thus, having witnessed a lot of samples from certain groups, the models are more capable of correctly classifying such samples while performing poorly for the so-called out-group samples.

2) ***Bias in Machine Learning****:* Owing to similar reasons as in the case of human perception, over or under-exposure to experiences (or in this case, data) characterised by specific features, ML models are seen to acquire biases that prejudice model performance for one or more data attributes. Facial analysis models particularly are seen to be affected by biases with respect to demographic attributes of gender, race or age, where samples belonging to one group dominate the

data distribution. In such situations, the under-represented groups get adversely impacted by the model misclassifying samples from these groups. Buolamwini et al. in their seminal work [28], highlighted how popular face recognition algorithms disproportionately misclassified darker females either misgendering them or not being able to detect their faces. Another study by Klare et al. [29] highlighted how face recognition algorithms employed by some law-enforcement agencies significantly underperform for people labelled as black or female compared to other demographics. Such biases in critical systems may lead to unnecessary targeting and exploitation of people from under-represented groups, further disadvantaging their opportunities in society.

B. Mitigating Bias in Facial Analyses

The origins of bias in most ML-based facial analyses algorithms can be traced back to imbalances in data distributions. Collating balanced datasets that enable a fair evaluation of ML models [30] despite being the most effective solution towards mitigating such biases, may not be as straightforward to achieve as a varied and diverse subject-pool might not always be available. As a result, several strategies have been proposed for mitigating the effects of bias on the training and evaluation of ML algorithms. We use a similar nomenclature as [14] to discuss these strategies.

1) ***Pre-Processing Approaches****:* The most simplistic of these is achieved by selectively sampling training data in a manner that balances learning. Samples from under-represented domains are over-sampled while dominant domains are under-sampled to balance learning the training data [18], [31]. This results in the training set to effectively have a balanced data distribution. However, this may not be possible in really small-scale datasets as under-sampling already limited data might not be efficient. An alternative approach is to use data-augmentation techniques to synthetically generate additional data for the under-represented groups [32], [33], [34], to balance training data distribution.

2) ***In-Processing Approaches****:* Another popular approach to mitigate the effects of imbalances in data distributions is to weight model prediction loss differently for the different domain attributes. A weighting factor is applied to the training loss computation based on the occurrence rate for the different classes or domains [15], [16], [35] penalising misclassifications for the under-represented groups more than others. This reduces the effect of these imbalances, mitigating biases in learning.

More recently, several learning strategies have been proposed that, while handing imbalances in data distributions using the above-mentioned techniques, also deal with biases in ML models at the algorithm-level. Howard et al. [5] propose a hierarchical approach that combines outputs from the cloud-based Microsoft Emotion API algorithm with a specialised learner, offering a 17.3% improvement in recognition results on a minority class, in this case, children's facial expressions. Other approaches focus on explicitly separating the decision boundaries with respect to different sensitive domain attributes ensuring that imbalances in data with respect to these attributes are not perpetuated while training the model, to achieve *'fairness through awareness'* [17]. Alternatively, the model can be trained to ignore domain-specific information, making it *unaware* or *blind* towards domain differences and focus only on the task at hand. Adversarial learning has been used to achieve such *'fairness through blindness'* [18] using a min-max training regime that maximises sensitivity towards the task at hand while minimising learning of domain-specific information. Xu et al. [36] implement a disentangled approach [37] that uses a similar strategy to mitigate bias with respect to sensitive domain attributes of gender and race for FER by ensuring that the feature representations learnt by the model do not contain any domain-specific information. The model is split into two parts with a shared feature extraction sub-network. The first part focuses on facial expression analysis, while the other part consists of separate branches for each domain, designed to suppress domain-specific information.

3) ***Post-Processing Approaches****:* Despite several methods proposed for training *fair* ML systems, as described above, it may not always be possible to completely eradicate bias in the model. In such cases, it is still important to examine whether a model is biased and quantify the bias to mitigate it and make fairer decisions. Post-processing approaches (see [14], [38] for a general discussion) focus on quantifying bias in existing algorithms and attempt to counter the effects on classification tasks.

C. Continual Learning

Learning to detect and manage shifts in data distributions, Continual Learning (CL) methods (see [19], [20] for an overview) can effectively learn with incrementally acquired data, offering an improvement over traditional ML models, especially for real-world application.

Typically, CL models are evaluated on 3 different learning scenarios [21]. The first scenario is termed as *Task-Incremental Learning (Task-IL)* where the model incrementally learns to solve several tasks, explicitly being informed about the task identity. Learning is split into different tasks, each corresponding to learning some sub-tasks or classes. The model is evaluated on its ability to preserve its knowledge across several tasks. The second scenario focuses on *Domain-Incremental Learning (Domain-IL)* where the task to be learnt by the model does not change but the input data distribution changes. While the model still needs to solve the same task but the inherent data distribution shifts and the model is evaluated on its ability to manage such a shift. The third and the most complex learning scenario is the *Class-Incremental Learning (Class-IL)* scenario where the model needs to learn a new class without being given any information on the tasks. The model incrementally learns one class at a time, sequentially receiving input data for only that class.

In recent years, several CL approaches have been proposed that employ (deep) ML architectures and equip them with learning capabilities such that they can incrementally integrate novel information while preserving past knowledge [19]. The most common and straightforward approach to achieve this is by regulating model updates in a manner that en-

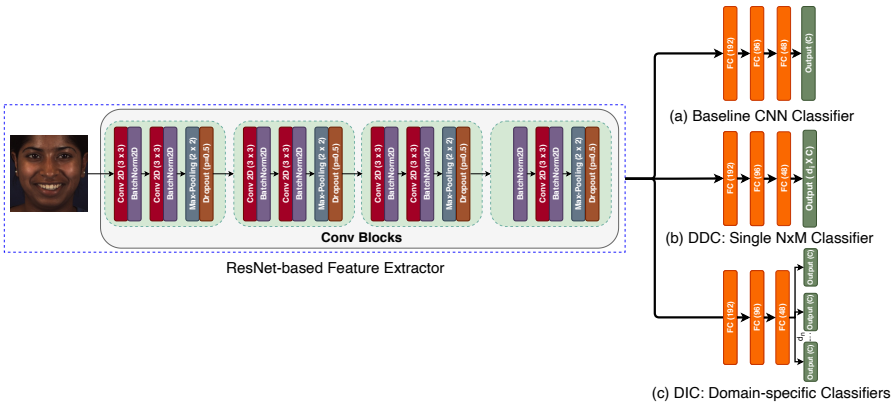


Fig. 1. Model Architectures for (a) the Baseline CNN, (b) Domain Discriminative Classification (DDC) [18], (c) Domain Independent Classification (DIC) [18]. The Baseline CNN is also used to implement all CL methods for a fair comparison.

ables the preservation of knowledge. Such *regularisation-based* methods minimise *destructive interference* by freezing those parts of the model that are most sensitive to previous tasks [39] and updating the rest of the model, selectively. Alternatively, weight-update constraints and penalties are applied that discourage changes in network parameters that deteriorate model performance on previous tasks [40], [41]. A priority or importance term may also be applied to network parameters based on their relevance to a given task and only those parameters are allowed to be updated which have lower importance [42]. Despite the competitive performance of regularisation-based methods, they become computationally expensive as the number of tasks or classes grow, limiting their performance and applicability in Task-IL (in extreme cases) and Class-IL scenarios.

Other CL-based approaches include rehearsal-based methods [43] that aim to simulate offline batch-learning based settings by either physically storing previously encountered data samples; commonly known as *Naive Rehearsal (NR)* [44], or learning a generative or probabilistic model that learns data statistics to simulate *pseudo-samples* for previously seen tasks [45], [46], [47]. Yet, as the number of tasks increase, it becomes extremely difficult to train these models. Furthermore, additional memory and computational resources need to be allocated to either store data samples or generate simulated pseudo-samples making it challenging to implement these approaches.

In this work, we focus on regularisation-based methods evaluated under Domain-IL settings where these models are required to learn to solve expression recognition and AU detection tasks across different domains of gender and race.

III. Methodology

In order to understand bias in FER algorithms, it is important to determine how the implicit data distribution affects model performance. For this, we need to understand which domain attributes dominate the data and how an algorithm performs with respect to these attributes. In this section, we present the problem formulation, the learning scenario as well as the different methods employed in this work, comparing them with popular CL-based methods.

A. Problem Formulation

We aim to measure the variance in model performance on a specific task with respect to *gender* and *race* as the different domain attributes and compare model performances for expression recognition and AU detection. Given a set of input images x_i with task labels y_i and domain label d_i, we wish to determine how the performance of an algorithm $\mathcal{A}(x_i|y_i, d_i)$, varies with respect to different variations of the domain label d_i.

To enable a fair comparison between the different bias mitigation methods, for our experiments, we implement the same ResNet architecture-based [48] CNN model for all the methods consisting of 4 convolutional blocks, each with 2 conv layers, a max-pooling layer and implementing drop-out with batch-normalisation. The output of the last conv block is connected to three dense layers and a classification layer making model predictions. ReLU activation is used for each conv as well as dense layer. The same architecture is used to implement all the approaches compared in this work (see Fig. 1) with the exception of the Disentangled Approach for which the results from the original paper [36] are used directly for comparison purposes.

*1) **The Baseline**:* For baseline evaluations, we split the dataset into different subsets based on the domain attributes. For example, for gender, the datasets are split into male and female splits and model performance is reported when trained incrementally on these data splits. This is sometimes also referred to as *finetuning* [49]. The model (see Fig. 1a), without any explicit mechanism to preserve knowledge, is expected to suffer from forgetting old tasks while preference is given to the new tasks.

*2) **Off-line Training**:* Providing another baseline evaluation, the above-described Convolutional Neural Network (CNN) model (see Fig. 1a) is trained on all the training data, *off-line*, at once but its performance scores are reported individually on domain-specific test-splits. Off-line training provides a fair comparison with traditional ML-based learning models and is a popularly used benchmark for evaluating the performance of CL-based methods.

B. Non-CL-based Bias Mitigation Strategies

Here, we investigate some of the popular bias mitigation strategies that are found in the literature and implement 4 different methods for comparison. We group them under 'non-CL-based' strategies to differentiate them from our baselines as well as the CL approaches.

*1) **Domain Discriminative Classification (DDC):*** A popular method for mitigating bias is to focus on achieving 'fairness through awareness' [17] where information about sensitive attributes (or domains) is explicitly learnt in feature encodings. This information later allows models to account for bias in learning by being more 'aware'. One way to achieve this to create an $N \times M$-way discriminative classifier where N denotes the number of domains and M is the number of classes to be learnt [18]. For example, for FER classifying 7 different expression classes for samples encoding 3 different race labels, a classifier is used with each output unit corresponding to a unique expression-race label pair (in this case, $7 \times 3 = 21$ label pairs). This allows the model to be more 'aware' of the different domains in order to learn discriminative features for each of them. For our experiments, we use the same model architecture (see Fig 1b) only replacing the output layer.

*2) **Domain Independent Classification (DIC):*** A major concern with the DDC method is that the network may implicitly learn decision boundaries within the same class across different domains. This may be redundant as, despite the different domain attributions, the class-boundaries may remain the same and the network may be unnecessarily penalised due to incorrect domain predictions even if it predicts the task correctly. Wang et al. [18] offer a solution to this by training separate classifiers for each domain, sharing the feature extraction layers. For our experiments, we make use of the same model architecture (see Fig. 1c), connecting separate dense-layered classifiers for each of the domains. The DIC model consists of different *heads*, each consisting of the same number of output units but corresponding to different domain attribute labels.

*3) **Strategic Sampling (SS):*** A simple approach for handling bias arising from imbalanced data distributions is to strategically sample data [15] for each domain-class mapping such that the resultant data distribution 'appears' to be balanced. Samples from under-represented distributions can be sampled more often during training or equivalently, prediction loss can be appropriately weighted to account for the under-represented classes. For our experiments, samples (s) for each of the N domains (d_i) are assigned a weight w_i *inversely proportional* to the rate of occurrence of samples for that domain, scaling the loss function to appropriately account for imbalances in the training set distribution. The scaled cross-entropy loss function is given as:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N} w_i \sum_{s=1}^{S} y_s^{(i)} \log \hat{y}_s^{(i)} \tag{1}$$

*4) **Disentangled Feature Learning (DA):*** Xu et al. [36] implement the Disentangled Feature Learning (DA) approach of [37] for facial expression recognition. This approach ensures that the feature representations learnt by the model do not contain any domain-specific information. The two sub-parts of the model focus on analysing facial expressions while learning to suppress domain-specific information. For our experiments, we sue the results from the original paper [36] for comparison.

C. Continual Learning Approaches

Domain-Incremental CL deals with scenarios where the structure of the tasks remains the same albeit with the input distribution is changing [21]. For our experiments, we model the tasks of expression recognition and AU detection in a domain-incremental manner where the models learn to solve these tasks as the input data distributions change with respect to domain attributes of gender and race. For example, in the case of gender, the models first learn to classify expression classes or predict activated AUs for 'male' samples and then, sequentially, learn to solve these tasks for 'female' samples (or vice-versa), without forgetting the previous task. In our experiments, each approach is implemented using the Baseline CNN architecture as shown in Fig 1a. All implementations were based on the CL code-benchmarks provided by [21], [44].

*1) **Elastic Weight Consolidation (EWC):*** The EWC approach, as proposed by Kirkpatrick et al. [41] imposes a quadratic penalty on parameter updates between old and new tasks in order to avoid forgetting previously learnt information. For each parameter θ, its relevance is calculated with respect to a task's training data $\mathcal{D}$, modelled as the posterior distribution $p(\theta|\mathcal{D})$. Thus, for two data distributions $\mathcal{D}_A$ and $\mathcal{D}_B$, corresponding to two independent tasks A and B, according to Bayes' rule, the posterior probability is given as:

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B), \tag{2}$$

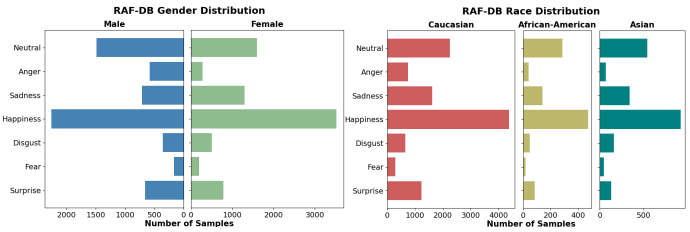
such that $\log p(\theta|\mathcal{D}_A)$ embeds all information about previously learnt tasks. As this term becomes intractable, Laplace approximation is used to approximate it as a Gaussian Distribution with its mean given by parameters θ_A^* (referring to parameters of task A) and the importance of the parameters determined by the diagonal of the Fischer Information Matrix. The loss function for the EWC method thus becomes:

$$L(\theta) = L_B(\theta) + \frac{1}{2}\lambda \sum_i F_i(\theta_i - \theta_{A,i}^*)^2, \tag{3}$$

where L_B is the loss for task B, λ is the regularisation coefficient that determines the relevance of old tasks with respect to the new one, i denotes the index of the parameter θ and F_i is the i^{th} diagonal element of the Fischer Matrix.

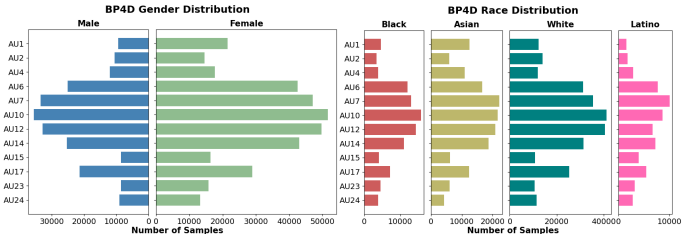
*2) **EWC-Online**:* A disadvantage for the EWC method is that as the number of tasks increase, the number of quadratic terms in the regularisation term grows linearly. To handle this, Schwarz et al. [50] proposed a modification to EWC where instead of many quadratic terms, a single quadratic penalty is applied, determined by a running sum of the Fischer Information Matrices of the previous tasks. Thus, the updated regularisation term of the proposed EWC-online approach is given as:

$$L_{reg}^T = \sum_i \tilde{F}_i^{(T-1)}(\theta_i - \theta_i^{(T-1)})^2, \tag{4}$$



(a) Gender (b) Race

Fig. 2. RAF-DB data distribution for Gender and Race Attributes.



(a) Gender (b) Race

Fig. 3. BP4D data distribution for Gender and Race Attributes.

TABLE I
REGULARISATION COEFFICIENT VALUES FOR FER EXPERIMENTS WITH THE RAF-DB DATASET ($\times 10^3$).

Method	W/O Data-augmentation		W/ Data-augmentation	
	Gender	*Race*	*Gender*	*Race*
EWC (λ)	5	10	10	5
EWC-Online (γ)	10	1	10	5
SI (c)	1	10	1	5
MAS (λ)	0.5	5	1	0.2
NR (λ)	0.4	1.1	0.1	1.1

TABLE II
REGULARISATION COEFFICIENT VALUES FOR AU DETECTION EXPERIMENTS WITH THE BP4D DATASET ($\times 10^3$).

Method	W/O Data-augmentation		W/ Data-augmentation	
	Gender	*Race*	*Gender*	*Race*
EWC (λ)	5	5	5	1
EWC-Online (γ)	0.05	0.5	0.05	0.01
SI (c)	5	5	0.05	0.01
MAS (λ)	1	0.5	0.1	1
NR (λ)	0.4	0.1	0.1	0.1

where $\theta_i^{(T-1)}$ is the i^{th} parameter after learning task $T-1$ and $\tilde{F}_i^{(T-1)}$ is the running sum of the diagonal elements of the Fischer Matrices of all previous tasks calculated as:

$$\tilde{F}_i^{(T)} = \gamma \tilde{F}_i^{(T-1)} + F_i^T, \tag{5}$$

where γ controls the contribution of previously learnt tasks.

*3) **Synaptic Intelligence (SI)**:* Similar to EWC, this approach also penalises changes to relevant weight parameters (synapses) in a manner that new tasks can be learnt without forgetting the old [42]. To alleviate catastrophic forgetting, the importance for solving a learned task is computed for each individual synapses and changes in the most important synapses are discouraged. A modified cost function L_n^* is used with a surrogate loss term which approximates the summed loss functions of all the previous tasks L_o^*:

$$L_n^* = L_n + c\sum_i \Omega_k^n (\theta_k^* - \theta_k)^2, \tag{6}$$

where θ_k represents the parameters for the new task, θ_k^* represents the parameters at the end of the previous task, Ω_k^n is the parameter regulation strength and c is the weighting factor balancing new vs. old learning.

*4) **Memory Aware Synapses (MAS)**:* Similar to EWC and SI, MAS also calculates the importance of each parameter although by looking at the sensitivity of the output function instead of the loss [49]. For each new sample, MAS updates the importance of each parameter by evaluating how sensitive the model prediction is to the changes in that parameter. Parameters that have the most impact on model predictions are given high importance and changes to these parameters are penalised. Different from EWC and SI, parameter importance is computed only using unlabelled data by measuring changes in model performance. For each new task (T_n), in addition to the task-loss ($L_n(\theta)$), changes to parameters important for previous tasks are penalised:

$$L(\theta) = L_n(\theta) + \lambda \sum_{i,j} \Omega_{i,j} (\theta_{i,j} - \theta_{i,j}^*)^2, \tag{7}$$

where λ is the hyperparameter balancing new vs. old task losses, and θ^* denotes the old network parameters.

*5) **Naive Rehearsal (NR)**:* For the Naive Rehearsal (NR) approach, we implement a straightforward rehearsal-based method that combines new data with previously seen data while training the model. [44]. A small replay buffer is implemented to randomly store a fraction of previously seen data samples that can be replayed to the model. Each mini-batch of data is constructed using an equal number of samples from the new as well as previously seen data. This interleaving of data pertaining to previously learnt tasks with new data ensures that old knowledge is not overwritten by new data.

IV. EXPERIMENT SET-UP

A. Datasets

For evaluating the different bias mitigation strategies and comparing them with CL-based methods, we use two popular benchmark datasets; the RAF-DB dataset for FER *in-the-wild* and the BP4D dataset recorded for AU detection in controlled settings . These datasets are selected due to (i) the diversity in their data acquisition settings, (ii) providing labels not only for expression/AU recognition but also gender and race attributes, and (iii) containing notable imbalances in the data distributions with respect to class and domain attribute labels. These factors make the RAF-DB and the BP4D datasets a good choice for our evaluation.

*1) **RAF-DB Dataset**:* The RAF-DB dataset [51] consists of $\approx 15K$ facial images labelled for six expression classes namely, *Surprise, Fear, Disgust, Happy, Sad* and *Anger* along with *Neutral* to denote absence of any expression. Additionally, it provides demographic attribute labels such as

gender (Male, Female, Unsure) and race (Caucasian, African-American, Asian) labels. For our experiments, we split the dataset using multiple grouping strategies based on the gender and race Labels. For gender-based grouping, we exclude images labelled as 'Unsure' and only use the 'Male' and 'Female' samples. As shown in Fig. 2, not only is the dataset imbalanced with respect to the different expression categories, there exist stark imbalances with respect to different demographic attributes as well. The majority of the samples in the training set represent the "Happy" expression class and belong to "Female" and "Caucasian" categories.

*2) **BP4D Dataset**:* The BP4D dataset [52] consists of video sequences from 41 subjects performing 8 different affective tasks to elicit emotional reactions. Each video is annotated frame-wise for the occurrence and intensity of the activated AUs. In our experiments, we only use occurrence labels for 12 most frequent AU resulting in $\approx 150K$ labelled frames, in total. Other than the frame-wise AU labels, demographic attribute labels for gender (Male, Female) and race (Black, White, Latino, Asian) have been provided to us, specifically for this research. Fig. 3 shows the data distribution of the BP4D dataset for the 12 AU labels with respect to the gender and race attributes. As can be seen, the majority of the samples in the dataset represent "White" and "Female" attribute labels.

B. Pre-processing and Data-Augmentation

Both RAF-DB and BP4D datasets provide face-centred RGB images which are resized to $(100 \times 100 \times 3)$ and normalised to be used as input for all the models. Training deep neural networks requires a lot of training data for each of the classes to be learnt. Due to the inherent imbalances in the dataset with respect to the different expression classes (RAF-DB) or the AU labels (BP4D), we increase the overall training data by performing data-augmentation by randomly $(p = 0.5)$ flipping images horizontally to create additional samples. For each experiment, we present the results *with* and *without* data-augmentation separately, for clarity.

C. Experiment Settings

*1) **Evaluation Metrics**:* To compare the different methods on their ability to balance classification performance within individual domain-splits while remaining consistent across the domains, we evaluate them both in terms of their accuracy scores as well as *fairness*. Furthermore, for the CL methods, we also report Catastrophic Forgetting (CF) scores, measuring the ability of the models to maintain performance on previously seen tasks while learning new tasks.

Accuracy (Acc): Accuracy is defined as the fraction of correctly classified samples. Given that TP = True Positives, FP = False Positives, TN = True Negatives and FN = False Negatives, Accuracy (Acc) can be computed as:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

In our experiments, we report accuracy scores separately for different gender and race attributes to highlight differences in model performance for these domains, underlining bias in the models' performance.

Fairness Measure ($\mathcal{F}$): To evaluate different approaches for their *fairness* with respect to model performance for gender and race attributes, we use the 'equal opportunity' definition of *fairness*, as proposed by Hardt et al. [54].

Let $\mathbf{x}$, $\mathbf{y}$, $\hat{\mathbf{y}}$ be the variables denoting input, ground truth label and the predicted label, respectively, $s \in S_i$ be the sensitive (domain) attribute (for example, $(S_i = \{\text{male, female}\})$, f be a function computing the accuracy score for a given sensitive attribute s and d be the dominant attribute which has the highest accuracy score, then the *Fairness Measure* $\mathcal{F}$ of a model is defined as the largest accuracy gap among all sensitive attributes computed as the minimum of the ratios of the accuracy scores of each sensitive attribute with respect to the dominant attribute.

$$\mathcal{F} = \min\left(\frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_0, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})}, ..., \frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_n, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})}\right) \tag{9}$$

In other words, $\mathcal{F}$ is defined as the ratio of the lowest accuracy for a sensitive attribute with respect to the highest accuracy value for that sensitive attribute.

Catastrophic Forgetting (CF): *Catastrophic forgetting* [55] occurs when learning a new task negatively impacts previously learnt information. For our experiments, we also report the *CF* metric score [56] for the CL methods, measuring the average change in the accuracy scores of the CL model for each previous task right after learning a new task. This is computed as follows:

$$CF = \frac{\sum_{j=1}^{i-1} a_{j,j} - a_{i,j}}{i-1}, \; A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}$$

where $a_{i,j}$ denotes the accuracy of i^{th} task right after learning j^{th} task, A is the matrix storing accuracy scores with dimensions $(n \times n)$ and n is the number of classes.

*2) **Implementation Details**:* All models are trained using the *adam* optimiser with a learning rate of $1.0e^{-4}$ and a batch-size of 24. For the experiments with the RAF-DB dataset, all models are trained for 25 epochs while for the BP4D dataset, due to a higher number of data samples, training converged after only 10 epochs for all the approaches. All experiments are *repeated* 3 times and the results are *averaged* across the repetitions to account for the random seeds. All models are implemented using the PyTorch Python Library based on the Continual Learning benchmarks provided by [21], [44].

Table I and II report the regularisation coefficient values for the CL methods for experiments with the RAF-DB and BP4D datasets, respectively. These values are set based on separate hyper-parameter searches for each model and selecting the best-performing values.

V. Experiments and Results

A. Experiment 1: Mitigating Bias in FER

In our experiments, we compare state-of-the-art bias mitigation approaches (see Section III-B) with popular CL-based methods (see Section III-C) on their ability to classify

TABLE III
EXPERIMENT 1: GENDER-WISE ACCURACY AND FAIRNESS SCORES ON RAF-DB DATASET. ACCURACY SCORES ARE REPORTED AFTER TRAINING THE MODELS ON BOTH MALE AND FEMALE SUBSETS. **BOLD** VALUES DENOTE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	Accuracy W/O Data-Augmentation		Fairness	Accuracy W/ Data-Augmentation		Fairness
	Male	*Female*		*Male*	*Female*	
Baseline	0.596±0.025	0.714±0.014	0.834	0.596±0.017	0.730±0.008	0.816
Offline	0.704±0.011	0.746±0.007	0.944	0.724±0.006	0.759±0.007	0.954
Non-CL-based Bias Mitigation Methods						
DDC [17]	0.699±0.013	0.722± 0.008	0.968	0.717±0.013	0.746±0.007	0.961
DIC [18]	0.698±0.014	0.744± 0.006	0.938	0.729±0.008	0.758±0.002	0.962
SS [15]	0.716±0.010	[*0.750±0.008*]	0.955	0.729±0.013	[*0.764±0.011*]	0.954
DA [36]	0.625	0.610	0.975	[*0.742*]	0.744	[*0.997*]
Continual Learning Methods						
EWC [41]	**0.723± 0.006**	0.744±0.006	0.972	0.735±0.007	0.748±0.012	0.983
EWC-Online [50]	0.721± 0.008	0.743±0.006	0.970	0.736±0.003	0.756±0.010	0.974
SI [42]	0.718± 0.007	0.725±0.004	**0.990**	0.739±0.008	0.739±0.005	**0.999**
MAS [49]	0.721± 0.008	0.735±0.012	[*0.980*]	**0.745±0.006**	0.753±0.009	0.990
NR [53]	[*0.722± 0.001*]	**0.778±0.006**	0.928	0.738±0.004	**0.799±0.005**	0.923

TABLE IV
EXPERIMENT 1: CATASTROPHIC FORGETTING (CF) AND OVERALL ACCURACY (PREVIOUS TASKS) AFTER EACH TASK FOR GENDER-ORDERED LEARNING ON RAF-DB DATASET. **BOLD** VALUES DENOTE THE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	W/O Data-Augmentation				W/ Data-Augmentation			
	Male		Female		Male		Female	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [41]	**0.730**	X	[*0.734*]	**-0.028**	**0.746**	X	0.742	**-0.032**
EWC Online [50]	0.728	X	0.733	-0.015	[*0.745*]	X	0.746	[*-0.024*]
SI [42]	0.721	X	0.728	-0.003	0.741	X	0.738	-0.004
MAS [49]	0.721	X	0.731	[*-0.024*]	0.743	X	[*0.749*]	-0.013
NR [53]	[*0.729*]	X	**0.753**	-0.010	0.736	X	**0.772**	-0.010

facial expressions without being affected by imbalances in data distributions. Furthermore, to evaluate the applicability of CL strategies as *'fair'* FER systems, we train and test these approaches on the RAF-DB dataset and compare their performance (both without and with data-augmentation) on learning to categorise 7 expression classes, namely, *surprise, sadness, happiness, fear, anger, disgust* and *neutral*, with respect to 2 different domain groups; gender (Male, Female) and race (Caucasian, African-American, Asian).

*1) **Bias Across Gender Attributes**:* For the RAF-DB dataset, approximately 53.4% of the samples are labelled as 'Female' while the 'Male' group constitutes about 40.3% of the total samples. The rest of the samples are labelled as 'Unsure' and omitted from our evaluations (see Fig. 2a). As a result, the effective split of the dataset with respect to gender is somewhat balanced, 56.3% Female against 43.7% Male samples. For the non-CL-based methods, the models are trained on the entire dataset and tested individually on the Male and Female subsets. For the CL evaluations, however, the learning is split into two tasks corresponding to expression recognition for the Male (Task 1) followed by expression recognition for the Female (Task 2) sub-sets. Task-ordering is discussed further in Section VI-A.

Table III presents the experimental results comparing the different methods on their Accuracy as well as Fairness Measure scores. It can be seen that the CL methods, overall, outperform all other methods both in accuracy as well as fairness scores while the baseline method performs the worst. Furthermore, although the accuracy scores of all the approaches increase when data-augmentation is used, not all of them are able to maintain fairness. CL methods (with the exception of NR) on the other hand, improve upon their fairness scores as well, with SI [42] achieving the highest fairness scores both without and with data-augmentation.

In order to fully appreciate how CL enables the models to retain their performance across the two tasks, it is important to understand how learning each new tasks impacts the model's performance on previously learnt tasks. Table IV further reports the overall accuracy and CF scores for the CL methods after each task, evaluating the performance of these methods in terms of their ability to classify expressions for both male and female sub-sets. We observe that both with and without augmentation, NR [44] achieves the highest overall accuracy score after both tasks are learnt, while EWC experiences the least forgetting. Furthermore, negative CF scores for all CL methods indicate that after learning task 2, that is, to predict expressions on Female samples, the overall accuracy of the model increased for both Male and Female samples, without any forgetting occurring in the model.

*2) **Bias Across Race Attributes**:* The data distribution of the RAF-DB dataset is highly imbalanced with respect to Race with a majority (77.4%) of the samples labelled as Caucasian, while the African-American and Asian subsets correspond to only 7.1% and 15.5% of the samples, respectively (see Fig. 2b). Similar to the evaluations across gender, for the Non-CL-based methods, the model is trained on the entire dataset

TABLE V
EXPERIMENT 1: RACE-WISE ACCURACY AND FAIRNESS SCORES ON RAF-DB DATASET. ACCURACY SCORES ARE REPORTED AFTER TRAINING THE MODELS ON ALL THE SUBSETS. **BOLD** VALUES DENOTE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	Accuracy W/O Data-Augmentation			Fairness	Accuracy W/ Data-Augmentation			Fairness
	Caucasian	*African American*	*Asian*		*Caucasian*	*African American*	*Asian*	
Baseline	0.750±0.019	0.764±0.029	**0.795±0.010**	0.943	0.758±0.004	0.778±0.023	**0.809±0.008**	0.937
Offline	0.727±0.010	0.750±0.011	0.735±0.025	0.969	0.743±0.006	0.762±0.017	0.763±0.007	0.974
Non-CL-based Bias Mitigation approaches								
DDC [17]	0.714±0.009	0.710±0.009	0.721±0.009	0.985	0.729±0.006	0.736±0.001	0.747±0.007	0.976
DIC [18]	0.724±0.004	0.730±0.015	0.732±0.016	0.989	0.745±0.007	0.768±0.012	0.772±0.013	0.965
SS [15]	0.734±0.005	0.728±0.015	0.757±0.014	0.961	0.748±0.002	0.752±0.019	0.767±0.023	0.975
DA [36]	0.634	0.584	0.544	0.858	0.756	0.766	0.704	0.919
Continual Learning approaces								
EWC [41]	0.764±0.011	0.758±0.002	0.768±0.016	0.987	**0.796±0.006**	[*0.788±0.009*]	0.794±0.007	0.990
EWC-Online [50]	[*0.773±0.018*]	0.763±0.003	0.773±0.002	0.987	0.777±0.017	0.780±0.002	0.785±0.014	0.990
SI [42]	0.769±0.010	[*0.766±0.006*]	0.769±0.008	**0.996**	0.785±0.013	0.783±0.003	0.782±0.009	**0.996**
MAS [49]	0.762±0.007	0.756±0.001	0.764±0.010	[*0.990*]	0.781±0.017	0.776±0.012	0.781±0.007	[*0.994*]
NR [53]	**0.779±0.015**	**0.772±0.017**	[*0.793±0.002*]	0.974	[*0.787±0.012*]	**0.796±0.005**	[*0.808±0.014*]	0.974

TABLE VI
EXPERIMENT 1: CATASTROPHIC FORGETTING (CF) AND OVERALL ACCURACY (PREVIOUS TASKS) AFTER EACH TASK FOR RACE-ORDERED LEARNING ON RAF-DB DATASET. **BOLD** VALUES DENOTE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	W/O Data-Augmentation						W/ Data-Augmentation					
	Task 1		Task 2		Task 3		Task 1		Task 2		Task 3	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [41]	0.777	X	**0.754**	[*0.025*]	0.764	0.019	0.797	X	**0.768**	**0.030**	**0.795**	[*0.000*]
EWC-Online [50]	[*0.778*]	X	0.738	0.046	[*0.779*]	0.012	0.798	X	[*0.763*]	[*0.038*]	0.779	0.026
SI [42]	**0.782**	X	0.736	0.047	0.769	**-0.003**	**0.802**	X	0.759	0.044	0.784	0.007
MAS [49]	0.766	X	[*0.744*	**0.023**	0.762	[*0.003*]	[*0.801*]	X	0.762	0.039	0.780	0.004
NR [53]	[*0.778*]	X	0.734	0.049	**0.781**	0.009	0.784	X	0.744	0.043	[*0.790*]	**-0.006**

and evaluated individually for the different race attributes. For the CL evaluations, the learning is again split into three tasks corresponding to learning to predict expressions for Caucasian (Task 1), African-American (Task 2) and Asian (Task 3) faces.

Table V presents the results of the experiments comparing Accuracy and Fairness Measure scores. The imbalances in the data distribution affect all the approaches such that the model accuracy varies across the race groupings. Yet, the CL methods seem to handle this best, achieving comparable accuracy with high fairness scores. Even though the CL approaches are not always the best performing ones, particularly for the *Asian* subset, all of them achieve high fairness scores, with SI performing the best. This underlines their ability to balance learning across the different tasks. They are able to give preference to being consistent and *fair*, trading-off higher accuracy scores for any individual race label. The NR approach, on the other hand, achieves the highest accuracy scores on Task 1 and Task 2, owing to the explicit replay mechanism, but sacrifices fairness across all groups in the process. Additionally, as RAF-DB is a relatively small dataset, data-augmentation has a positive effect on accuracy scores of all the models but the fairness scores do not change significantly.

In Table VI, the accuracy and CF scores can be seen for all the CL methods reporting model performance on all previous tasks computed at the end of each new task. We see that all models tend to forget as they learn new tasks, yet the SI method is able to mitigate forgetting the best after having learnt all the tasks. When data-augmentation is used, the individual accuracy scores are enhanced but the CF scores do not improve. Owing to the explicit replay mechanism, the NR method performs the best in terms of mitigating forgetting when using data-augmentation.

B. Experiment 2: Mitigating Bias in AU Detection

As more than one AU may be *activated* at the same time (for example, AU $1, 2$ and 26 together may depict *surprise*), predicting facial AUs poses a multi-label classification problem. Imbalances in data distributions with respect to the different *gender* and *race* attributes become even more prominent with certain AU classes having much more data samples than the others (see Fig. 3). We compare different bias mitigation strategies (see Section III-B) with CL-based methods (see Section III-C) to understand how they cope with imbalances in data distributions while retaining model performance.

Similar to Experiment 1, we compare the performance of all models (without and with data-augmentation) on detecting activations for the 12 AUs for *gender* (Male, Female) and *race* (White, Black, Asian, Latino) groupings.

*1) **Bias Across Gender Attributes**:* Similar to Experiment 1, for the non-CL-based methods, we train the individual models on the entire dataset but evaluate them individually for Male and Female subsets. The BP4D data distribution is skewed in favour of Female samples constituting 60.96% of the data while only 39.04% samples belong to the Male subset (see Fig. 3a). For CL methods, the learning is split into two *tasks*: Task 1: Male and Task 2: Female, incrementally

TABLE VII

EXPERIMENT 2: GENDER-WISE ACCURACY AND FAIRNESS SCORES ON BP4D DATASET. ACCURACY SCORES ARE REPORTED AFTER TRAINING THE MODELS ON BOTH MALE AND FEMALE SUBSETS. **BOLD** VALUES DENOTE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	Accuracy W/O Data-Augmentation		Fairness	Accuracy W/ Data-Augmentation		Fairness
	Male	*Female*		*Male*	*Female*	
Baseline	0.691±0.007	0.718±0.004	0.962	0.692±0.009	0.735±0.004	0.941
Offline	0.701±0.007	0.712±0.006	0.984	0.731±0.005	0.735±0.008	[*0.994*]
Non-CL-based Bias Mitigation approaches						
DDC [17]	0.740±0.008	0.733±0.005	[*0.990*]	0.752±0013	0.745±0.007	0.991
DIC [18]	0.734±0.010	0.750±0.008	0.979	0.741±0.012	0.752±0.009	0.985
SS [15]	0.715±0.003	0.732±0.002	0.977	0.735±0.005	0.748±0.005	0.983
DA [36]	0.731	0.735	**0.994**	0.776	**0.772**	**0.995**
Continual Learning approaches						
EWC [41]	[*0.769±0.006*]	0.754± 0.008	0.981	0.772±0.006	0.766±0.006	0.992
EWC-Online [50]	0.762±0.009	0.744± 0.006	0.976	0.772±0.012	[*0.767±0.004*]	[*0.994*]
SI [42]	0.755±0.021	0.745± 0.016	0.986	[*0.778±0.017*]	0.751±0.022	0.965
MAS [49]	**0.788±0.007**	**0.761± 0.009**	0.966	**0.784±0.003**	0.758±0.012	0.967
NR [53]	[*0.769±0.010*]	[*0.756± 0.004*]	0.983	0.774±0.009	0.739±0.016	0.954

TABLE VIII

EXPERIMENT 2: CF AND OVERALL ACCURACY (PREVIOUS TASKS) AFTER EACH TASK FOR GENDER-ORDERED LEARNING ON BP4D DATASET. **BOLD** VALUES DENOTE THE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	W/O Data-Augmentation				W/ Data-Augmentation			
	Task 1		Task 2		Task 1		Task 2	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [41]	[*0.745*]	X	[*0.768*]	[*-0.020*]	0.729	X	[*0.761*]	**-0.024**
EWC Online [50]	0.744	X	**0.769**	**-0.022**	0.728	X	0.753	[*-0.016*]
SI [42]	**0.753**	X	0.764	0.002	[*0.736*]	X	0.749	-0.008
MAS [49]	0.734	X	0.765	-0.011	**0.745**	X	**0.766**	-0.002
NR [53]	0.740	X	0.759	0.000	[*0.736*]	X	0.758	-0.006

learning with the two data splits. Task-ordering and its affect is discussed in Section VI-B.

Table VII compares the different methods on their Accuracy and Fairness Measure scores for both Male and Female subsets. CL methods are shown to outperform other methods in terms of accuracy, yet, DA [36] attains the highest Fairness score. Although CL methods perform better than others on individual tasks, they are not able to balance this learning across tasks as compared to DA. The MAS outperforms other CL methods on accuracy, yet SI and EWC-Online methods score higher on fairness, without and with data-augmentation, respectively. Data-augmentation, overall, has a positive impact on model accuracy scores but much like Experiment 1, it does not impact model fairness, significantly. Individual class accuracy between Male and Female splits does not vary significantly for the 12 AU labels with AU 2 and AU 12 achieving the lowest and highest accuracy scores, respectively, across the models for both the splits. This can be due to these classes consisting of the lowest and highest number of data samples in BP4D distribution across both gender and race splits (see Fig 3).

Comparing different CL methods on their ability to maintain performance across the tasks, Table VIII reports the overall accuracy and CF scores for the models. Owing to the complex multi-label nature of the tasks as well as the high gender disparity in the data distribution, we see a high variation in performance scores of the different CL methods. While EWC-Online performs the best without employing data-augmentation achieving a negative CF score, the MAS model performs the best with data-augmentation.

*2) **Bias Across Race Attributes**:* The majority of the samples in the BP4D dataset are labelled as White (approximately 46.76%) with other samples corresponding to Asian (26.08%), Black (16.56%) and Latino (10.6%) groups (see Fig. 3b). For our evaluations, we split the dataset into 4 subsets based on these labels, representing the 4 tasks, that is, Task 1: Black, Task 2: Asian, Task 3: White and Task 4: Latino, for the CL models. Task-orderings are discussed further in Section VI-B.

It can be seen in Table IX that CL methods, overall, achieve high accuracy and Fairness scores, with the NR approach outperforming other methods both without and with data-augmentation. While NR achieves the highest accuracy scores for Black, Asian and White subsets even without data-augmentation, none of the approaches is able to beat the offline baseline for the Latino subset. This may be owing to the extremely low sample-rate for the Latino subset that the CL methods, instead of focusing on improving performance on this sub-set, focus on maintaining performance across all the sub-sets. Data-augmentation has a positive impact on all the models in terms of improvements both in accuracy and fairness scores.

Different CL models handle the high variance in data distribution with respect to racial identity labels with varying levels of success. Table X shows how, at different points during the learning, different models perform better than others, while NR achieves the highest accuracy and CF scores after all tasks are learnt, both with and without data-augmentation. The negative CF scores for all the approaches at the end of Task 4 signifies that all the CL models were able to mitigate forgetting and the overall model performance improved as they incrementally learnt new tasks.

TABLE IX
EXPERIMENT 2: RACE-WISE ACCURACY AND FAIRNESS SCORES ON BP4D DATASET. ACCURACY SCORES ARE REPORTED AFTER TRAINING THE MODELS ON ALL THE SUBSETS. **BOLD** VALUES DENOTE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	Accuracy W/O Data-Augmentation				Fairness	Accuracy W/ Data-Augmentation				Fairness
	Black	***Asian***	***White***	***Latino***		***Black***	***Asian***	***White***	***Latino***	
Baseline	0.659±0.048	0.720±0.016	0.771±0.018	0.764±0.008	0.855	0.654±0.042	0.685±0.022	0.763±0.009	0.737±0.019	0.858
Offline	0.694±0.009	0.724±0.019	[*0.781±0.012*]	**0.790±0.011**	0.878	0.706±0.008	0.750±0.017	0.783±0.011	0.777±0.019	0.901
Non-CL-based Bias Mitigation Approaches										
DDC [17]	0.722±0.018	0.729±0.007	0.775±0.015	[*0.785±0.005*]	0.920	0.727±0.004	0.731±0.013	0.781±0.010	[*0.787±0.005*]	0.924
DIC [18]	0.712±0.011	0.731±0.014	0.767±0.017	0.770±0.015	0.925	0.719±0.009	0.742±0.009	0.778±0.005	0.780±0.012	0.922
SS [15]	0.722±0.007	0.729±0.012	0.775±0.012	[*0.785±0.013*]	0.920	0.723±0.018	0.731±0.007	0.781±0.002	[*0.787±0.006*]	0.919
DA [36]	0.706	0.736	0.740	0.736	[*0.954*]	**0.767**	0.774	0.771	**0.797**	[*0.962*]
Continual Learning Approaches										
EWC [41]	[*0.742±0.012*]	0.760±0.016	0.769±0.007	0.731±0.007	0.949	0.721±0.035	0.754±0.019	0.764±0.003	0.738±0.009	0.943
EWC-Online [50]	[*0.742±0.004*]	0.769±0.024	0.762±0.008	0.720±0.000	0.937	0.748±0.004	0.776±0.008	0.781±0.009	0.747±0.006	0.957
SI [42]	0.731±0.024	[*0.773±0.000*]	0.762±0.000	0.770±0.007	0.946	0.752±0.018	**0.788±0.004**	[*0.782±0.008*]	0.765±0.001	0.954
MAS [49]	0.710±0.023	[*0.773±0.006*]	0.753±0.009	0.772±0.003	0.920	0.715±0.021	[*0.786±0.008*]	0.763±0.009	0.775±0.012	0.909
NR [53]	**0.767±0.023**	**0.779±0.009**	**0.788±0.001**	0.762±0.012	**0.966**	[*0.763±0.006*]	0.780±0.005	**0.783±0.012**	0.764±0.002	**0.974**

TABLE X
EXPERIMENT 2: CF AND OVERALL ACCURACY (PREVIOUS TASKS) AFTER EACH TASK FOR RACE-ORDERED LEARNING ON BP4D DATASET. **BOLD** VALUES DENOTE THE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	W/O Data-Augmentation								W/ Data-Augmentation							
	Task 1		Task 2		Task 3		Task 4		Task 1		Task 2		Task 3		Task 4	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [41]	0.763	X	0.707	0.059	0.692	[*0.125*]	0.761	[*-0.042*]	0.763	X	0.686	0.087	0.673	**0.129**	0.748	-0.026
EWC-Online [50]	**0.773**	X	0.693	0.079	0.682	**0.122**	0.757	-0.032	0.766	X	0.697	0.081	0.682	[*0.140*]	0.754	-0.026
SI [42]	[*0.765*]	X	**0.734**	**0.044**	**0.714**	0.154	[*0.764*]	-0.027	[*0.774*]	X	0.702	0.091	[*0.686*]	0.147	[*0.768*]	-0.032
MAS [49]	0.759	X	[*0.724*]	[*0.048*]	[*0.707*]	0.144	0.762	-0.035	0.754	X	**0.717**	**0.056**	**0.697**	0.162	0.762	[*-0.033*]
NR [53]	0.759	X	0.710	0.055	0.690	0.143	**0.777**	**-0.053**	**0.775**	X	[*0.705*]	[*0.079*]	0.685	0.156	**0.780**	**-0.035**

VI. DISCUSSION

Our experiments on FER (see Section V-A) and AU detection (see Section V-B) tasks are motivating as they highlight how adopting CL strategies may enable *fairer* facial affect analysis algorithms. Consistently achieving high accuracy as well as fairness measure scores, CL offers an improvement over existing learning strategies for bias mitigation in ML algorithms. Robustly managing imbalances in data distributions, both without and with data-augmentation, CL methods are better equipped to deal with biases owing to their learning strategy of focusing on one domain group at a time. Here, we discuss each task individually and highlight how CL provides a solution towards *fairer* facial affect analyses.

A. Facial Expression Recognition

When applied to FER, CL methods aim to sequentially learn to predict expression categories for the different gender and race groups. The models are trained with one group at a time and as the model experiences samples from other groups, it is actively trying to maintain performance at previously seen groups without forgetting. As a result, for both gender and race groups, CL models are able to achieve high fairness scores by balancing performance across the splits, with the SI model performing the best (see Table XI). Selective updates of network parameters in order to mitigate forgetting allows CL models to maintain high accuracy scores across the different gender and race attributes. This makes them distinct from other approaches, directly focusing on maintaining performance across different domain distributions instead of deciding whether to capture domain-specific features or not. In comparison, non-CL-based methods rely on becoming 'aware' of domain attributes to predict expressions according to the subjects sharing gender or race attributes or learning feature representations that actively 'block' domain discriminative features [36]. Furthermore, for most of the non-CL methods, with the exception of DA, we need to know the domain groupings a priori which may not always be possible in real-world scenarios. For CL methods, however, as models learn sequentially, there is no need to provide any domain information a priori and learning can be extended to new domains.

One concern when applying CL methods to FER tasks is the class-ordering effect where model performance is seen to be sensitive to the order in which it learns different expression classes [47]. In our experiments, as we implement the Domain-IL scenario where all classes are learnt at the same time, albeit one domain-group at a time, class-ordering does not play any role in the learning. Instead, we explore whether different task-orderings, that is, learning with different sequences of gender or race group splits has any effect on the models' ability to maintain performance. For both gender and race domains, we experiment with different orders of learning the tasks but no significant effect of domain ordering is witnessed on the models' performance. Class-wise accuracies are largely consistent between the different learning settings for the CL models with model accuracy being the worst for

TABLE XI
EXPERIMENT 1: FAIRNESS MEASURE SCORES ACROSS GENDER AND RACE DISTRIBUTIONS FOR THE RAF-DB DATASET. **BOLD** VALUES DENOTE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	W/O Data-Augmentation		W/ Data-Augmentation	
	Gender	*Race*	*Gender*	*Race*
Baseline	0.834	0.943	0.816	0.937
Offline Training	0.944	0.925	0.954	0.974
Non-CL-based Bias Mitigation Methods				
DDC [18]	0.968	0.985	0.961	0.976
DIC [18]	0.938	0.989	0.962	0.965
SS [15]	0.955	0.961	0.954	0.975
DA [36]	0.975	0.858	[*0.997*]	0.919
Continual Learning Methods				
EWC [41]	0.972	0.987	0.983	0.990
EWC-Online [50]	0.970	0.987	0.974	0.990
SI [42]	**0.990**	**0.996**	**0.999**	**0.996**
MAS [49]	[*0.980*]	[*0.990*]	0.990	[*0.994*]
NR [53]	0.928	0.974	0.923	0.974

TABLE XII
EXPERIMENT 2: FAIRNESS MEASURE SCORES ACROSS GENDER AND RACE DISTRIBUTIONS FOR THE BP4D DATASET. **BOLD** VALUES DENOTE BEST WHILE [*bracketed*] DENOTE SECOND-BEST VALUES FOR EACH COLUMN.

Method	W/O Data-Augmentation		W/ Data-Augmentation	
	Gender	*Race*	*Gender*	*Race*
Baseline	0.962	0.855	0.941	0.858
Offline	0.984	0.878	[*0.994*]	0.901
Non-CL-based Bias Mitigation Approaches				
DDC [18]	[*0.990*]	0.920	0.991	0.924
DIC [18]	0.979	0.925	0.985	0.922
SS [15]	0.977	0.920	0.983	0.919
DA [36]	**0.994**	[*0.954*]	**0.995**	[*0.962*]
Continual Learning Approaches				
EWC [41]	0.981	0.949	0.992	0.943
EWC-Online [50]	0.976	0.937	[*0.994*]	0.957
SI [42]	0.986	0.946	0.965	0.954
MAS [49]	0.966	0.920	0.967	0.909
NR [53]	0.983	**0.966**	0.954	**0.974**

disgust due to the overall low number of samples, in line with what was reported in [36].

B. Action Unit Detection

Action Unit (AU) detection poses a *harder* multi-label classification problem where the models need to predict all the AUs activated in a given sample. The inherent class-imbalances in the BP4D dataset are further accentuated by the imbalances with respect to gender and race attributes, making it extremely difficult for CL as well as non-CL models to maintain performance across the different groups. The under-represented classes are reduced to even fewer samples per class when split across gender or race, making it even more difficult for these models to cope with data imbalances. Even though CL-based methods are able to achieve the highest individual accuracy scores (averaged across the 12 AUs) for most of the gender and race groups, this comes at the cost of balancing learning across the different attributes. For the gender splits, Disentangled Feature Learning (DA) achieves the highest fairness scores, despite performing moderately in terms of accuracy on individual splits (see Table XII). Blocking individual domain-specific information, allows DA to balance learning across the different splits, resulting in high fairness scores. For CL models, however, the multi-label classification settings cause the models to focus more on overall individual performance rather than on maintaining performance across the gender splits. In the case of race-splits, we see that the NR approach achieves the highest fairness scores, while DA performs second-best. This is due to the memory-intensive rehearsal mechanism that physically stores and replays samples from previously seen domains to retain model performance. Even though regularisation-based approaches target accuracy and trade-off fairness in the process, they still perform better than most non-CL-based methods.

Due to the multi-label settings, all classes are learnt together with no ordering of the classes required. Furthermore, domain-ordering, that is, in which order the gender and race domains should be learnt, does not have any significant effect on model performance for the CL methods. Owing to the highly imbalanced class-distributions, the performance of all models are poor for under-represented classes such as AU $1, 2$ and 4, across all gender and race splits. On the other hand, the highest model performances are achieved for dominant classes such as AUs 10 and 12. These results are inline with other AU prediction approaches [35], [57], [58] that report similar differences in performance across these AUs.

C. Limitations of CL-based Bias Mitigation

Our benchmark experiments with the RAF-DB and BP4D datasets highlight the potentials of CL-based models for creating *fairer* facial expression recognition systems. CL-based models outperform other bias mitigation strategies for evaluations across gender and race domains, managing shifts in data distributions well. However, more work is needed to optimise CL-based models for multi-label settings where they under-perform (see Table VII and IX). Recent work by Kim et al. [59] proposes a new replay-based strategy, the Partitioning Reservoir Sampling (PRS), that aims to tackle continual learning for multi-label classification, balancing both intra- and inter-task imbalances. Yet, they benchmark their approach on classification settings with little-to-no overlap between the tasks. This is not the case for AU detection where the different domains, as well as the classes within each domain, share feature representations, making it even harder for the models.

Furthermore, as regularisation-based CL models assign *importance* to different parameters based on their contribution towards previously learnt tasks, shared feature representations makes it harder for models to incrementally learn different tasks/domains as model parameters may contribute to more than one task or domain. Rehearsal-based methods such as NR, on the other hand, require the models to physically store seen samples from previous tasks, interleaving them with new data to maintain performance. As the number of tasks, or in the case of Domain-IL, data-splits across domains such as

gender or race increase, storing samples from all the domains becomes extremely expensive both in terms of its memory footprint as well as the computational power needed to train the algorithms.

Additionally, as the tasks increase, models may experience saturation [60] requiring stronger regularisation in the models to be able to preserve past knowledge [61]. The performance of the models also takes a hit where the model needs to re-prioritise whether to give more importance to the new task or remembering previous tasks. We see this in race-wise splits for both the datasets (see Table V and IX) where regularisation-based models attain higher accuracy scores for the last split, while the NR method aims to maintain a higher fairness score instead.

VII. Conclusion and Future Work

In this work, we propose the novel use of Domain Incremental CL as a potent bias mitigation method for facial analysis tasks. In particular, we highlight how using Domain-IL settings, regularisation-based CL methods can help develop *fairer* expression recognition and AU detection algorithms. Our experiments with popular benchmark datasets, RAF-DB for expression recognition and BP4D for AU detection, showcase the superlative performance of CL methods at handling imbalances in data distributions with respect to demographic attributes of gender and race. In comparison with state-of-the-art bias mitigation approaches, these methods are able to balance learning across different domain splits, not only achieving high accuracy scores but also maintaining fairness across the different splits.

Yet, this proof-of-concept evaluation was limited to regularisation-based methods only and hence further experimentation is needed to fully understand the benefits of using CL as an effective bias mitigation strategy for facial expression and action unit recognition tasks. With harder problems, as in the case of multi-label AU detection, we see that even though regularisation-based methods achieve high accuracy, they do so by sacrificing fairness across different domain attributes. While a simplistic and naive rehearsal mechanism is able to improve model performance, our future work will aim to investigate other, more complex, pseudo-rehearsal methods [46], [47], [59], [61] or neuro-inspired [60], [62], [63] on their bias mitigation abilities.

References

[1] S. Feldstein, "The global expansion of ai surveillance," Carnegie Endowment for International Peace, Tech. Rep., 2019.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[3] S. Dey, B. R. Duff, N. Chhaya, W. Fu, V. Swaminathan, and K. Karahalios, "Recommendation for video advertisements based on personality traits and companion content," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ser. IUI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 144–154.

[4] D. Roselli, J. Matthews, and N. Talagala, "Managing bias in AI," in *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, May 2019.

[5] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2017, pp. 1–7.

[6] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[7] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, June 2015.

[8] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: a survey," *IEEE Transactions on Affective Computing*, June 2017.

[9] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[10] P. Ekman and W. V. Friesen, *Facial action coding systems*. Consulting Psychologists Press, 1978.

[11] ——, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[12] P. Ekman, "Darwin's contributions to our understanding of emotional expressions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3449–3451, Dec. 2009.

[13] S. Li and W. Deng, "A Deeper Look at Facial Expression Dataset Bias," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[14] S. Yucer, S. Akcay, N. A. Moubayed, and T. Breckon, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation." in *Workshop on Fair, Data Efficient and Trusted Computer Vision, Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2020.

[15] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 973–978.

[16] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment," in *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 725–740.

[17] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 214–226.

[18] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8919–8928.

[19] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[20] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information Fusion*, vol. 58, pp. 52–68, 2020.

[21] G. M. van de Ven and A. S. Tolias, "Three scenarios for continual learning," *arXiv preprint arXiv:1904.07734*, 2019.

[22] H. Tajfel and J. Turner, "An Integrative Theory of Intergroup Conflict," *The Social Psychology of Inter-group Relations. Monterey, CA: Brooks/Cole*, pp. 33–47, 1979.

[23] M. Hewstone, M. Rubin, and H. Willis, "Intergroup bias," *Annual Review of Psychology*, vol. 53, no. 1, pp. 575–604, 2002.

[24] S. D. Preston, "A perception-action model for empathy," *Empathy in mental illness*, vol. 1, pp. 428–447, 2007.

[25] J. N. Gutsell and M. Inzlicht, "Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups," *Journal of experimental social psychology*, vol. 46, no. 5, pp. 841–845, 2010.

[26] S. L. Sporer, "Recognizing faces of other ethnic groups: An integration of theories." *Psychology, Public Policy, and Law*, vol. 7, no. 1, pp. 36–97, Mar. 2001.

[27] H. A. Elfenbein and N. Ambady, "Is there an in-group advantage in emotion recognition?" *Psychological Bulletin*, vol. 128, no. 2, pp. 243–249, 2002.

[28] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91.

[29] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[30] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[31] V. Iosifidis and E. Ntoutsi, "Dealing with bias via data augmentation in supervised learning scenarios," in *International Workshop on Bias in Information, Algorithms, and Systems (BIAS). Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems (BIAS)*. CEUR Workshop Proceedings, 2018, pp. 24–29.

[32] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives," *IEEE CI Magazine*, vol. 14 (2), pp. 68–81, 2019.

[33] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3d convolutional networks," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1609–1618.

[34] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, vol. 89, pp. 385 – 397, 2015.

[35] N. Churamani, S. Kalkan, and H. Gunes, "Spatio-Temporal Analysis of Facial Actions using Lifecycle-Aware Capsule Networks," *arXiv preprint arXiv:2011.08819*, 2020.

[36] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham: Springer International Publishing, 2020, pp. 506–523.

[37] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, "Exploring disentangled feature representation beyond face identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, "Face recognition algorithm bias: Performance differences on images of children and adults," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2269–2277.

[39] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "Pathnet: Evolution channels gradient descent in super neural networks," *arXiv: Neural and Evolutionary Computing*, 2017.

[40] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[41] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[42] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," *Proceedings of machine learning research*, vol. 70, p. 3987, 2017.

[43] A. Robins, "Catastrophic forgetting in neural networks: the role of rehearsal mechanisms," in *Proc. The 1st New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, Nov 1993, pp. 65–68.

[44] Y.-C. Hsu, Y.-C. Liu, A. Ramasamy, and Z. Kira, "Re-evaluating continual learning scenarios: A categorization and case for strong baselines," in *NeurIPS Continual learning Workshop*, 2018.

[45] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[46] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Advances in Neural Information Processing Systems*, 2017, pp. 2990–2999.

[47] N. Churamani and H. Gunes, "CLIFER: Continual Learning with Imagination for Facial Expression Recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020, pp. 322–328.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[49] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.

[50] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *International Conference on Machine Learning*, 2018, pp. 4528–4537.

[51] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[52] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014, best of Automatic Face and Gesture Recognition 2013.

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

[54] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[55] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[56] N. Díaz-Rodríguez, V. Lomonaco, D. Filliat, and D. Maltoni, "Don't forget, there is more than forgetting: new metrics for continual learning," in *Workshop on Continual Learning, NeurIPS 2018 (Neural Information Processing Systems*, 2018.

[57] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 8594–8601.

[58] Z. Shao, L. Zou, J. Cai, Y. Wu, and L. Ma, "Spatio-temporal relation and attention learning for facial action unit detection," *arXiv preprint arXiv:2001.01168*, 2020.

[59] C. D. Kim, J. Jeong, and G. Kim, "Imbalanced Continual Learning with Partitioning Reservoir Sampling," in *ECCV*, 2020.

[60] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization," *Frontiers in Neurorobotics*, vol. 12, p. 78, 2018.

[61] M. K. Titsias, J. Schwarz, A. G. de G. Matthews, R. Pascanu, and Y. W. Teh, "Functional regularisation for continual learning with gaussian processes," in *International Conference on Learning Representations*, 2020.

[62] R. Kemker and C. Kanan, "Fearnet: Brain-inspired model for incremental learning," *CoRR*, vol. abs/1711.10563, 2018.

[63] N. Kamra, U. Gupta, and Y. Liu, "Deep generative dual memory network for continual learning," *CoRR*, vol. abs/1710.10368, 2017.