

# Towards Fair Affective Robotics: Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition

Ozgur Kara

Electrical & Electronics Engineering  
Department  
Bogazici University  
Istanbul, Turkey  
ozgur.kara@boun.edu.tr

Nikhil Churamani

Department of Computer Science and  
Technology  
University of Cambridge  
Cambridge, United Kingdom  
nikhil.churamani@cl.cam.ac.uk

Hatice Gunes

Department of Computer Science and  
Technology  
University of Cambridge  
Cambridge, United Kingdom  
hatice.gunes@cl.cam.ac.uk

## ABSTRACT

As affective robots become integral in human life, these agents must be able to fairly evaluate human affective expressions without discriminating against specific demographic groups. Identifying *bias* in Machine Learning (ML) systems as a critical problem, different approaches have been proposed to mitigate such biases in the models both at data and algorithmic levels. In this work, we propose Continual Learning (CL) as an effective strategy to enhance *fairness* in Facial Expression Recognition (FER) systems, guarding against biases arising from imbalances in data distributions. We compare different state-of-the-art bias mitigation approaches with CL-based strategies for *fairness* on expression recognition and Action Unit (AU) detection tasks using popular benchmarks for each; RAF-DB and BP4D. Our experiments show that CL-based methods, on average, outperform popular bias mitigation techniques, strengthening the need for further investigation into CL for the development of *fairer* FER algorithms.

## 1 INTRODUCTION

From security and surveillance systems [16], monitoring emotional and mental wellbeing [3] and assisting in medical interventions [4] to law enforcement [21], robots are becoming closely embedded in our society, making ‘smart’ decisions about several critical aspects of our lives [36]. Therefore, it is crucial that they make fair and unbiased decisions [21] in order to avoid potentially catastrophic consequences that adversely affect individuals [17].

Fair and unbiased analysis and interpretation of human affective behaviours are among the factors that can contribute to the realisation of effective long-term Human-Robot Interaction (HRI). Successful long-term HRI can be used to provide physical and emotional support to the users, engaging them in a variety of application domains that require personalised human-robot interaction, including healthcare, education and entertainment [9].

Facial Expression Recognition (FER) algorithms (see [25, 30, 38] for a detailed survey) aim to analyse human facial expressions either by encoding facial muscle activity as Facial AU [14] or determining the emotional state being expressed by an individual [12, 13].

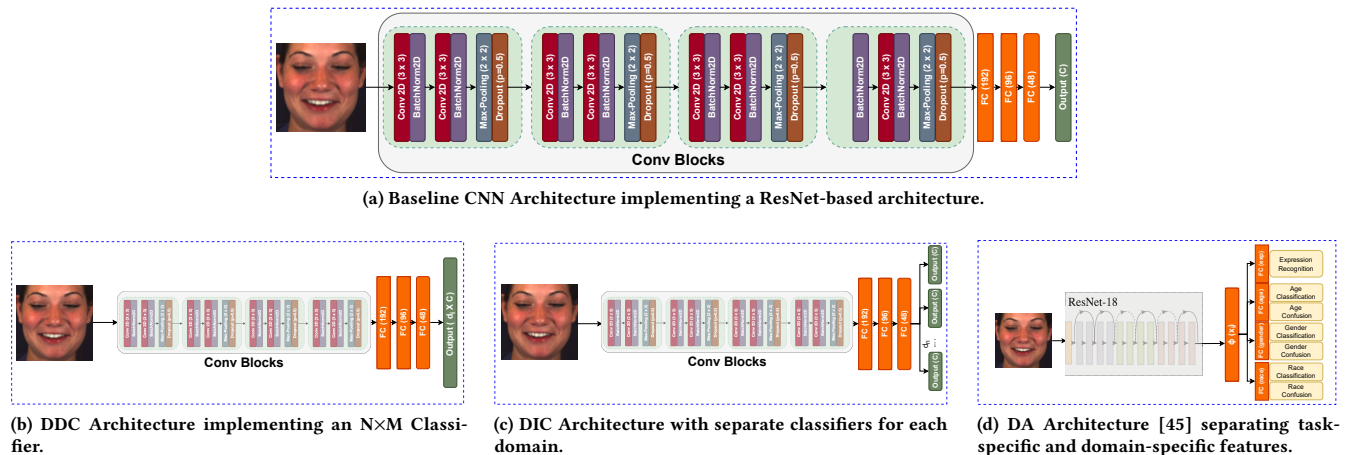
Analysing large datasets of human faces, annotated for the expressions represented in the images, training FER models becomes heavily data-dependent and may be prone to *biases* originating from imbalances in the data distribution with respect to attributes like gender, race, age or skin-color, implicitly encoded in the data [26]. These imbalances may result in the models learning to associate such *confounding* attributes with the task of FER. With affective robots increasingly becoming an integral part of daily human life, ‘bias’ in FER models, as described above, may result in unfavourable and prejudiced consequences for many under-represented groups.

Most recent and popular FER datasets provide annotations for different demographic attributes, along with affective labels, allowing for *fairer* FER models using these annotations explicitly [26]. This can be done at the *pre-processing* level [46] either modifying the data distribution in favour of underrepresented groups while training by strategically sampling [15] the data, or at *in-processing* level [46] by adapting the model architecture or the training process to handle these imbalances. Some methods achieve this by either forcing the model to explicitly learn domain-specific information such that this can be discounted from the model’s learning later [11] or by discounting domain-specific information by omitting these features from the learnt representations [44]. Other popular strategies include using data-augmentation strategies to synthetically generate additional data for the under-represented groups [1, 7, 8, 18] to balance training data distribution or weighting model prediction loss differently for the different domain attributes. A weighting factor may be applied to the loss computation based on the occurrence rate for the different classes or domains [10, 15, 40] penalising misclassifications for under-represented groups more than others. The underpinning principle behind these methods, however, remains the same, focusing on balancing data distributions or learning to adapt to the imbalances by adjusting the learning algorithm.

Continual Learning (CL) approaches [24, 31] focus on this very challenge of managing shifts in data distributions by *continually* learning and adapting to novel information without forgetting previously learned information. As agents interact with their environments and gather more information about specific tasks, they need to be able to remember previously learnt tasks while acquiring new skills. CL may allow them to balance learning across different domains or tasks as well as being robust against the changes in the data distributions. In particular, Domain-Incremental CL [43] deals with managing shifts in the input data distributions, while the task to be learnt remains the same. This can be considered analogous to managing affective interactions with users belonging to different

---

O. Kara contributed to this work while undertaking a summer research study at the Department of Computer Science and Technology, University of Cambridge. 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. The 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, Profs Weihong Deng and JunPing Du from Beijing University of Posts and Telecommunications (China) for providing access to RAF-DB.



**Figure 1: Architectures for (a) the Baseline CNN, (b) the Domain Discriminative Classifier (DDC) [44], (c) the Domain Independent Classifier (DIC) [44], and (d) the Disentangled Approach (DA) [45].**

domain groups of gender (male or female) or race (black, white or asian), trying to analyse their expressions.

We propose the use of CL algorithms, benefiting from their ability for *lifelong* learning, to develop *fairer* FER models for affective robots that can balance learning with respect to different attributes of *gender* and *race*. In particular, we formulate expression recognition and AU detection, across these different domain attributes, as continual learning problems. We compare several popular regularisation-based CL approaches with state-of-the-art bias mitigation methods by splitting the data into different domain labels for each attribute; *male* and *female* for gender, *White/Caucasian*, *Black/African-American*, *Asian* and *Latino* for race, provided by the RAF-DB and BP4D datasets. CL-based approaches are, on average, seen to outperform other bias mitigation strategies both in terms of accuracy as well as fairness for both the RAF-DB dataset (across gender and race splits) and the BP4D dataset (for race splits).

## 2 METHODOLOGY

To investigate the problem of bias in FER systems, 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 briefly describe the different non-CL and CL-based methods compared in this work.

### 2.1 Problem Formation

As our objective is to evaluate bias in FER with respect to different domain attributes, we focus on evaluating model performances on datasets split across *gender* and *race* domain attributes on two different facial analysis tasks: Expression recognition and AU detection. Thus, given a set of samples  $x_i$ , ground truth labels  $y_i$  and domain labels  $d_i$  we are evaluate the performance of the model  $\mathcal{A}(x_i|y_i, d_i)$  across the different domain labels.

We implement a ResNet-based Convolutional Neural Network (CNN) [20] architecture composed of 4 conv blocks each of which consists of 2 convolutional layers, a max pooling layer, with dropout

and batch normalisation. The output of the last conv block is attached to a three-layered Multi-layered Perceptron (MLP) with a classifier output. ReLU activation is used after all conv and dense layers. We use the same CNN architecture as the basis for all the approaches, except for the Disentangled Approach (DA) [45] where we take the results from the original paper.

**2.1.1 Baseline Model.** As our baseline approach, above described CNN model (see Fig. 1) is incrementally trained on the domain-based splits of the dataset and model performance is reported after the training for each split. This method is commonly referred to as *finetuning* [2].

**2.1.2 Offline Training.** As a second baseline, we train the above-described *baseline* CNN model with all the training data, *off-line*, at once but report its performance scores individually on domain-specific test-splits.

## 2.2 Non-CL-based Bias Mitigation Strategies

In this section, we describe four popular and state-of-the-art bias mitigation methods, grouped under ‘non-CL-based’ approaches, to be compared with CL-based solutions.

**2.2.1 Domain Discriminative Classifier (DDC):** A popular approach for bias mitigation in ML systems is referred to as “*fairness-through-awareness*” [11] where the domain information is explicitly learnt and encoded in feature representations making the model more ‘aware’ of the different domain labels in order to discriminate between each of them. The model implements a  $N \times M$ -way classifier [44] where  $N$  is the number of domains and  $M$  is the number of classes to be learned per domain (see Fig 1).

**2.2.2 Domain Independent Classifier (DIC):** One of the concerns with DDC-based models is that they may learn decision boundaries for the same class across different domains, that is, even if the model predicts the correct class, it may be penalised unnecessarily due to differences in domain-specific features. Wang et al. [44] propose a different approach by training separate classifiers for each of the domain while sharing the same top-level architectures. This,

with the different model heads (see Fig 1), it can learn to solve the task for each domain group individually and independently.

**2.2.3 Strategic Sampling (SS):** The simplest approach to mitigate bias arising from skewed data distributions, without changing model architecture, is to balance the effect of each domain distribution by *strategically sample* data [15] for each domain-class mapping. This can be achieved by increasing the sampling frequency of the images from under-represented distributions or equivalently, for each domain  $d_i$  assigning a loss weight  $w_i$  inversely proportional to the rate of occurrence of sample for that domain. We follow the weighted-loss approach for strategically sampling data.

**2.2.4 The Disentangled Approach (DA):** Xu et al. [45] implement the Disentangled Approach (DA) [28] that aims to mitigate bias across sensitive domains by making sure that the feature representations learnt by the model do not contain any domain-specific information. The network is split into two parts with a shared ResNet-based feature extraction sub-network. The first part focuses on facial affect analysis, while the other part consists of separate branches for each domain, designed to suppress domain-specific information. For our experiments, the results for the FER tasks are taken from the original paper [45].

## 2.3 Continual Learning Approaches

In this work, we primarily explore regularisation-based CL approaches, under Domain Incremental Learning (Domain-IL) settings, as these can be implemented with the least additional computational and memory overhead. For a comparison with rehearsal-based methods, we also implement a simple Naive Rehearsal (NR) [22] method that physically stores previously encountered data samples for rehearsal. Other CL-based methods that improve model performance using a generative or probabilistic model to simulate *pseudo-samples* for previously seen tasks [8, 35, 41] or by dynamically expanding model architectures by adding dedicated neurons (Growing Neural Networks [8, 32]) or neural layers (Progressive Networks [37]) sensitive to specific domains or tasks are omitted from this evaluation as they require additional memory and computational resources to be allocated, giving them an unfair advantage over non-CL-based bias mitigation methods. For our experiments, the models need to learn AU detection and FER tasks as the input data distributions change with respect to domain-specific attributes of *gender* and *race*. We use the same baseline CNN model and apply the learning protocol as described by the following approaches:

**2.3.1 Elastic Weight Consolidation (EWC):** As proposed by Kirkpatrick et al. [23], the EWC approach introduces a quadratic penalty determined by the relevance of each parameter of the model with respect to old and new tasks, penalising updates in parameters relevant for old tasks, in order to avoid forgetting previously learnt information. The importance of different model parameters are determined using a Fisher Information Matrix, updating the loss function.

**2.3.2 EWC-Online:** A disadvantage for the EWC method is that, as the number of tasks increase, the number of quadratic terms for regularisation grows. To handle this, Schwarz et al. [39] 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.

**2.3.3 Synaptic Intelligence (SI):** Similar to EWC, the SI approach also penalises changes to relevant weight parameters (synapses) such that new tasks can be learnt without forgetting the old [47]. To avoid forgetting, importance for solving a learned task is computed for each parameter and changes in most important parameters are discouraged.

**2.3.4 Memory Aware Synapses (MAS):** Memory Aware Synapses (MAS) also tries to alleviate forgetting by calculating the importance of each parameter by looking at the sensitivity of the output function instead of the loss [2]. Parameters that have the most impact on model predictions are given a high importance and changes to these parameters are penalised. However, unlike EWC and SI, parameter importance is calculated in an unsupervised manner with using only unlabelled data.

**2.3.5 Naive Rehearsal (NR):** Inspired by the model used in [22], we implement a (naive) rehearsal-based approach that implements a small replay-buffer to randomly store a fraction of previous seen data. This *old data* along with the new data is used to create mini-batches using equal number of samples from both old and new data and used to train the model ensuring that old knowledge is not overwritten by new data.

## 3 EXPERIMENTATION AND RESULTS

### 3.1 Set-up

For our experiments, we conduct two separate evaluations, comparing the different bias mitigation approaches (see Section 2) for facial expression recognition and AU detection tasks with the RAF-DB [27] and BP4D [48] datasets, respectively. Each approach is compared in terms of the *Fairness Scores* achieved, both with and without data-augmentation. All evaluations are repeated 3 times and results are averaged across the repetitions, except for DA where the results are taking from the original paper.

### 3.2 Fairness Measure

To compare the different approaches for their *fairness* with respect to model performance (in this case, *Accuracy*) for different attributes of gender and race, we use the ‘equal opportunity’ definition of *fairness*, as proposed by Hardt et al. [19].

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*  $\mathcal{F} \in [0, 1]$  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.

$$\mathcal{F} = \min \left( \frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_0, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})}, \dots, \frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_n, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})} \right) \quad (1)$$

**Table 1: Experiment 1: Fairness Scores across Gender and Race 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	0.944	0.925	0.954	0.974
<b>Non-CL-based Bias Mitigation Methods</b>				
DDC [44]	0.968	0.985	0.961	0.976
DIC [44]	0.938	0.989	0.962	0.965
SS [15]	0.955	0.961	0.954	0.975
DA [45]	0.975	0.858	[0.997]	0.919
<b>Continual Learning Methods</b>				
EWC [23]	0.972	0.987	0.983	0.990
EWC-Online [39]	0.970	0.987	0.974	0.990
SI [47]	<b>0.990</b>	<b>0.996</b>	<b>0.999</b>	<b>0.996</b>
MAS [2]	[0.980]	[0.990]	0.990	[0.994]
NR [22]	0.928	0.974	0.923	0.974

### 3.3 Experiment 1: Facial Expression Recognition

To investigate the applicability of CL-based methods as ‘fair’ FER systems, we train and test the approaches described in Section 2 on the RAF-DB dataset and compare their performance (both without and with data augmentation) on learning to correctly categorise the 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). The different approaches are trained with samples belonging to one domain-split at a time, actively trying to preserve the knowledge from previously seen splits while learning to classify samples for the new domain-split. As a result, CL approaches, on average, outperform all other methods on their Fairness Scores with respect to both gender and race domains with SI yielding the best results both with and without data-augmentation (see Table 1). In addition, although data-augmentation has an overall positive effect on model accuracy for all approaches, no significant shift is witnessed in model fairness scores.

Furthermore, for the non-CL-based methods, except for DA, complete knowledge of the domain-groupings is required a priori in order to design the architectures of the models, limiting their real-world applicability. For CL-based methods, however, learning can be adapted to new domains *on-the-fly* as the models are designed to be sensitive to random and sudden changes in data distributions encountering samples from different domain groups over time.

### 3.4 Experiment 2: Action Unit Detection

Action Unit (AU) detection poses a multi-label classification problem where the models need to predict multiple AUs activated in a given sample. As in the case of Experiment 1, we report and compare the different bias CL-based and non-CL-based methods on the BP4D dataset with respect to their average Fairness Scores across 12 AUs, individually for gender and race domain-splits. We see that even though the CL-based methods are able to achieve highest individual accuracy scores for most of the gender and race groups, this comes at the cost of balancing learning across the different attributes. For the gender- splits, the Disentangled Approach (DA) [45] achieves the highest fairness scores, despite performing moderately in terms

**Table 2: Experiment 2: Fairness Scores across Gender and Race 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
<b>Non-CL-based Bias Mitigation Approaches</b>				
DDC [44]	[0.990]	0.920	0.991	0.924
DIC [44]	0.979	0.925	0.985	0.922
SS [15]	0.977	0.920	0.983	0.919
DA [45]	<b>0.994</b>	[0.954]	<b>0.995</b>	[0.962]
<b>Continual Learning Approaches</b>				
EWC [23]	0.981	0.949	0.992	0.943
EWC-Online [39]	0.976	0.937	[0.994]	0.957
SI [47]	0.986	0.946	0.965	0.954
MAS [2]	0.966	0.920	0.967	0.909
NR [22]	0.983	<b>0.966</b>	0.954	<b>0.974</b>

of accuracy on individual splits (see Table 2). In the case of race splits, we see that even though the NR approach achieves the highest fairness cores, this is owed to the memory-intensive rehearsal mechanism that physically stores and replays samples from previously seen domains to retain model performance. Even though the regularisation-based approaches target accuracy and trade-off fairness in the process, they still perform better than most non-CL-based methods. On the contrary, for the non-CL methods we see that more importance is given to fairness than individual accuracy with DA achieving consistently high fairness scores.

## 4 CONCLUSION AND FUTURE WORK

In this work, we focused on the problem of bias in facial analysis tasks and proposed a novel application of continual learning as a bias mitigation strategy for FER systems. We highlight how using CL can help develop *fairer* expression recognition and AU detection algorithms with our experiments with popular benchmark datasets; RAF-DB for expression recognition and BP4D for AU detection showcasing 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 domains, not only achieving high accuracy scores but also maintaining *fairness* across the different splits. Yet, in our experiments we primarily focus on regularisation-based CL methods due their efficacy and economic implementation for real-world application. It will be interesting to contrast these methods to other resource-hungry yet, improved algorithms [29, 34] that are able to better handle long-term retention of knowledge. Combining regularisation-based model adaptation with latent replay strategies [33] may prove helpful in implementing *fairer* facial analysis systems for long-term HRI.

Furthermore, future work for us also entails conducting long-term HRI studies, comparing CL vs. non-CL-based methods, by embedding these models onto a humanoid robot. Implementing long-term social interactions with under-represented population groups such as children [42], ethnic and racial minorities [6] and the elderly [5] can help evaluate how CL-based FER systems respond to users from different demographics.

## REFERENCES

- [1] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey. 2017. Using Synthetic Data to Improve Facial Expression Analysis with 3D Convolutional Networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 1609–1618.
- [2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 139–154.
- [3] Indu P. Bodala, Nikhil Churamani, and Haticce Gunes. 2020. Creating a Robot Coach for Mindfulness and Wellbeing: A Longitudinal Study. arXiv:2006.05289 [cs.HC] arXiv 2006.05289.
- [4] Jason Borenstein, Ayanna Howard, and Alan R. Wagner. 2017. *Pediatric robotics and ethics: The robot is ready to see you now, but should it be trusted?* Oxford University Press, United Kingdom, 127–141.
- [5] Joost Broekens, Marcel Heerink, Henk Rosendal, et al. 2009. Assistive social robots in elderly care: a review. *Gerontechnology* 8, 2 (2009), 94–103.
- [6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (*Proceedings of Machine Learning Research, Vol. 81*). Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. <http://proceedings.mlr.press/v81/buolamwini18a.html>
- [7] Francisco Charte, Antonio J. Rivera, Maria J. del Jesus, and Francisco Herrera. 2015. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems* 89 (2015), 385 – 397.
- [8] N. Churamani and H. Gunes. 2020. CLIFER: Continual Learning with Imagination for Facial Expression Recognition. In *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 322–328.
- [9] Nikhil Churamani, Sinan Kalkan, and Haticce Gunes. 2020. Continual Learning for Affective Robotics: Why, What and How?. In *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 425–431.
- [10] Nikhil Churamani, Sinan Kalkan, and Haticce Gunes. 2020. Spatio-Temporal Analysis of Facial Actions using Lifecycle-Aware Capsule Networks. *arXiv preprint arXiv:2011.08819* (2020).
- [11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (Cambridge, Massachusetts) (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226.
- [12] Paul Ekman. 2009. Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (Dec. 2009), 3449–3451.
- [13] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124.
- [14] Paul Ekman and Wallace V. Friesen. 1978. *Facial action coding systems*. Consulting Psychologists Press.
- [15] Charles Elkan. 2001. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (Seattle, WA, USA) (*IJCAI'01*). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 973–978.
- [16] Steven Feldstein. 2019. *The Global Expansion of AI Surveillance*. Technical Report. Carnegie Endowment for International Peace. [iii]–[iv] pages.
- [17] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine* 38, 3 (2017), 50–57.
- [18] J. Han, Z. Zhang, N. Cummins, and B. Schuller. 2019. Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives. *IEEE CI Magazine* 14 (2) (2019), 68–81.
- [19] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
- [20] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
- [21] Ayanna Howard and Jason Borenstein. 2017. The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics* 24, 5 (Sept. 2017), 1521–1536.
- [22] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. 2018. Re-evaluating Continual Learning Scenarios: A Categorization and Case for Strong Baselines. In *NeurIPS Continual Learning Workshop*.
- [23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.
- [24] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. 2020. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information Fusion* 58 (2020), 52–68.
- [25] S. Li and W. Deng. 2020. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* (2020).
- [26] S. Li and W. Deng. 2020. A Deeper Look at Facial Expression Dataset Bias. *IEEE Transactions on Affective Computing* (2020).
- [27] Shan Li, Weihong Deng, and JunPing Du. 2017. 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*. 2852–2861.
- [28] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. 2018. Exploring Disentangled Feature Representation Beyond Face Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [29] David Lopez-Paz and Marc Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in neural information processing systems*. 6467–6476.
- [30] Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic analysis of facial actions: a survey. *IEEE Transactions on Affective Computing* (June 2017).
- [31] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* 113 (2019), 54–71.
- [32] German I. Parisi, Jun Tani, Cornelius Weber, and Stefan Wermter. 2018. Lifelong Learning of Spatiotemporal Representations With Dual-Memory Recurrent Self-Organization. *Frontiers in Neurobotics* 12 (2018), 78.
- [33] Lorenzo Pellegrini, Gabriele Graffieti, Vincenzo Lomonaco, and Davide Maltoni. 2019. Latent Replay for Real-Time Continual Learning. *CoRR abs/1912.01100* (2019). arXiv:1912.01100
- [34] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2001–2010.
- [35] Anthony Robins. 1995. Catastrophic Forgetting, Rehearsal and Pseudorehearsal. *Connection Science* 7, 2 (1995), 123–146.
- [36] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing Bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM.
- [37] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671* (2016).
- [38] E. Sariyanidi, H. Gunes, and A. Cavallaro. 2015. Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 6 (June 2015), 1113–1133.
- [39] Jonathan Schwarz, Wojciech Czarnecki, Jelenka Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & Compress: A scalable framework for continual learning. In *International Conference on Machine Learning*. 4528–4537.
- [40] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. 2018. Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment. In *Computer Vision – ECCV 2018*. Springer International Publishing, Cham, 725–740.
- [41] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*. 2990–2999.
- [42] Lai Poh Emily Toh, Albert Causo, Pei-Wen Tzuo, I-Ming Chen, and Song Huat Yeo. 2016. A Review on the Use of Robots in Education and Young Children. *Journal of Educational Technology Society* 19, 2 (2016), 148–163.
- [43] Gido M van de Ven and Andreas S Tolias. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734* (2019).
- [44] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8919–8928.
- [45] Tian Xu, Jennifer White, Sinan Kalkan, and Haticce Gunes. 2020. Investigating Bias and Fairness in Facial Expression Recognition. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, 506–523.
- [46] S. Yucer, S. Akcay, N. Al Moubayed, and T.P. Breckon. 2020. 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.
- [47] Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. *Proceedings of machine learning research* 70 (2017), 3987.
- [48] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. 2014. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706. Best of Automatic Face and Gesture Recognition 2013.