Faces are Domains: Domain Incremental Learning for Expression Recognition

1st Rahul Singh Maharjan
Manchester Centre for Robotics and AI
University of Manchester
Manchester, United Kingdom
rahulsingh.maharjan@manchester.ac.uk

2nd Marta Romeo
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, United Kingdom
m.romeo@hw.ac.uk

3rd Angelo Cangelosi
Manchester Centre for Robotics and AI
University of Manchester
Manchester, United Kingdom
angelo.cangelosi@manchester.ac.uk

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191542

***Abstract*—Since most existing facial expression recognition methods depend on deep learning models trained in isolation on a facial expression image corpora, once employed in scenarios that are different from those in the corpora, they usually demand ad-hoc retraining to be able to perform better in the expression recognition task for new scenarios. Furthermore, most of these facial expression recognition methods are inconsistent when recognising person-specific expressions or are incapable of adjusting to real-world scenarios where data is exclusively obtainable incrementally. In this paper, we present a face incremental expression recognition model, where we utilise domain incremental learning methods to learn individual facial features of facial expressions. We assume that each individual's facial expression (domain) is presented to the model one domain at a time. We assessed our model's ability to remember previously seen domains (individual's facial expression) and incrementally perform on new face domains. Our model improves performance compared to a non-incremental learning model and an incremental learning model in facial expression recognition for individual data with different expression classes.**

***Index Terms*—Facial Expression Recognition, Deep Learning, Incremental Learning, Affective Computing**

I. Introduction

Facial expressions represent a vital nonverbal signal for human-human interactions [1]. Through facial expressions, human beings can convey their inner states. Due to their importance and potential applications in various domains, facial expression recognition has attracted significant interest. For example, facial expression recognition is used in the automotive industries for driver affective status analysis [2], in medical treatments for pain analysis [3], and in novel human-machine interactions for social robots [4] [5]. Recently, facial expression recognition evolved as a multidisciplinary research area that explores methods extracting hierarchical feature representations of facial expression with hand-crafted features [6], [7] or employs deep learning [8], [9] to learn directly from human data. Facial expression recognition seeks to learn facial expression by encoding facial action units movement [10] or choose the emotional state that is being represented by an individual [11].

Although the existing deep learning models can perform well on facial expression recognition on benchmark datasets [12], [13], they are not capable of translating that performance to real-world conditions where the model incrementally interacts with new humans. Every human displays their emotional state differently via facial expression, accordingly to their personality traits [14] and complex cultural environments [15], which causes slight inter-class similarities; as a result, a shift in domain distribution occurs. Although facial expression has long been thought to be the universal language to represent the internal emotional state, recent works [15], [16] contend that an individual's facial expression of emotion is not universal and is based on an individual's cultural background and personality traits. Due to such heterogeneity in the knowledge space of facial expression, conventional deep-learning models for facial expression recognition encounter challenges outside the standard datasets. Nevertheless, most existing deep learning models try to solve this challenge with train-once-test-all approaches without considering individuals and with high reliance on large-scale balanced, enormously labelled datasets. To tackle these issues, Churamani [17] proposed incremental learning for affective computing to addresses the diversity of individual aspects for personalisation.

From the human perspective, humans internally collect, analyse and distinguish other humans' facial expressions as a part of their decision process [18]. The interplay between recognising an individual's facial characteristics, such as how an individual moves specific muscles in their face, an individual's mouth and eye position, and congregating them into the emotional state is crucial for modelling human expression recognition [19]. An artificial agent able to mimic this capability could enhance its interaction skills and adapt to new humans. In this sense, enabling agents to sense, model and adapt to the facial expressions of specific individuals will improve their performance across various human-machine interaction applications. However, with existing deep learning models, adapting to different facial expression scenarios takes much work due to the re-training process. As soon as most deep learning models for expression recognition need to learn a novel expression representation, they must be re-trained or re-designed. Such a problem occurs as a result of understanding facial expression recognition as a conventional computer vision problem rather than adapting the solutions to distinctive features of individual facial expressions.

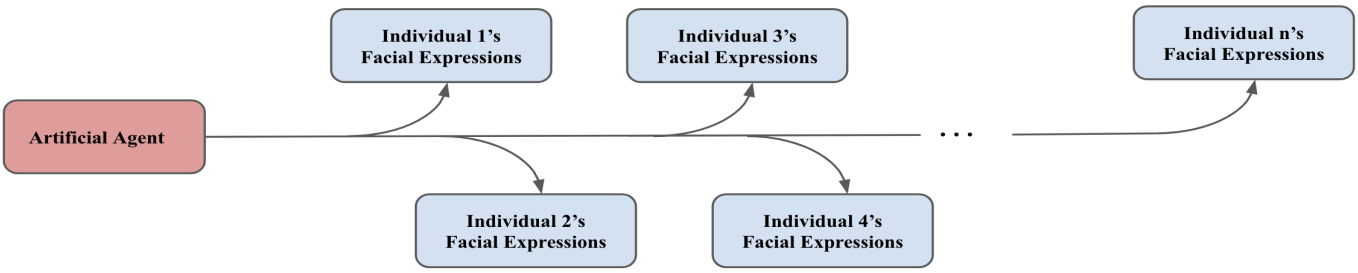

Fig. 1. Artificial agents with incremental facial expression recognition: The agent adapts its facial expression recognition to different users in an incremental paradigm where the full data of the current individual (domain) is only available.

In this work, we tackle facial expression recognition by focusing on the differences between individual facial expressions, as shown in Fig. 2 and learning individual facial expressions, where each individual's facial expressions are available only incrementally, as illustrated in Fig. 1. This problem is interesting and important because of its real-world relevance, where the diversity of facial expression becomes even more problematic when considering interpersonal factors that include age group, personality traits and cultural background. For instance, social robots are starting to be developed to understand human social-emotion signals. Such robots are anticipated to be keen to individual contrasts and offer natural and engaging interaction specific to each individual. When such social robots are deployed in real-world scenarios, they are desired to learn from new individuals' facial expressions in an incremental setting [20].

Throughout this paper, we assume that the facial expressions of each individual represent a unique domain in the form of individual facial expression domain distribution. When an agent is required to learn from new individuals incrementally, the agent faces a domain distribution shift problem [21]. We tackle this problem by developing a new model called **F**ace **I**ncremental e**X**pression **R**ecognition model (**FIXR**) that combines the benefits of a pre-trained deep learning model with incremental learning methods.

II. Related Works

A. Facial Expression Recognition

The ability to recognise human expressions is an interesting - yet challenging issue spreading across several topics, including but not confined to human-machine interaction, health, education, and multimedia recommendation. Hence, facial expression recognition is one of the core goals of affective computing and artificial intelligence. In recent years, researchers have made considerable progress in developing deep learning-based facial expression recognition models that perform better than hand-crafted models, thanks to their capacity to learn robust high and low-level features. For example, some facial expression recognition models classify the static images of the face into one of the six basic emotions [22], [23]. Others attempt to recognise the individual muscle movements that the face produces in order to provide an emotional description [24].

Due to the small set of facial expression datasets, training deep learning models from scratch inclines to over-fitting. To mitigate this, many use extra task-oriented data on well-known pre-trained models [25]–[27]. Fine-tuned EfficientNet [25] for facial expression recognition was used to outperform previous state-of-the-art models by using robust optimisation technique [28] on *in-the-wild* expression datasets [13]] [12]. Differently, Distract Your Attention Network (DAN) [29] presented a facial expression recognition method consisting of three sub-networks to maximise class separability for backbone features while capturing various attentions and penalising overlapping attentions.

These state-of-the-art deep learning models take advantage of the *in-the-wild* expression datasets [13]] [12] which are assumed to increase the variability of expression representation. Further, the datasets used to train these state-of-the-art models are supposed to be Independent and Identically Distributed (IID), with unseen facial expressions lying on the same IID distribution. Fig. 2 presents *t-SNE* visualisation of the "happy" class data from three facial expression datasets: AffectNet dataset [13], individual actor's faces from the RAVDESS [30] dataset and the MEAD [31] dataset. This visualisation underlines the similarity between feature vectors from the datasets mentioned earlier by using VGGNet's [26] penultimate fully connected layer. Here, VGGNet is trained on VGGFace2 [32] dataset. In the figure, we can see that the data from the AffectNet dataset are clustered in the top left, whereas data from MEAD and RAVDESS are distant from every other dataset. This visualisation points out the difference in domain (face) distribution of "happy" classes for three datasets.

The new paradigm for human-robot interaction based on continual learning, and relying on facial expression recognition [20], [34], argues that stationary datasets such as AffectNet [13] and Aff-Wild2 [12] are not well suited for dynamic settings where individual faces change continuously, due to their inability of adaptation. Working toward this effort, the P-AffMem [35] model is based on conditional adversarial autoencoder [36] and learns to represent and edit general expression with *Grow-When-Required* networks for personalising the memory of individual aspects of emotional expressions. The

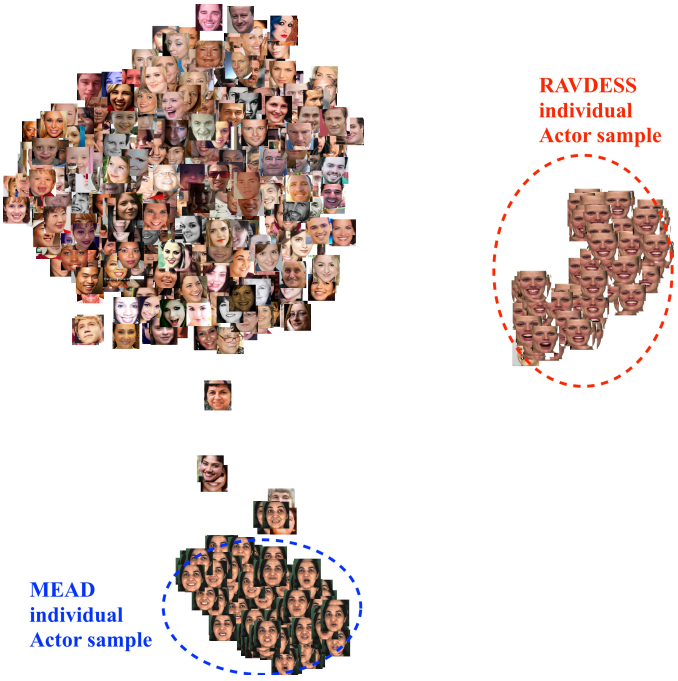

Fig. 2. *t-SNE* [33] result on AffectNet [13], RAVDESS [30] and MEAD [31] datasets and underlines the similarity between feature vectors of "happy" class from the datasets mentioned earlier by using VGGNet.

CLIFER framework [37] focuses on the assortment of human facial expressions by incorporating imagination to imitate expression data for specific humans and combine it into complementary learning-based dual memory. Stoychev *etal* [38] propose a latent generative replay-based incremental learning method to address the problem of using large amounts of memory by replay-based continual learning for facial expression recognition. In addition, Churamani *etal* [39] propose domain incremental learning to address biases in facial expression recognition. It is therefore clear that Continual Learning has demonstrated to offer great opportunities to advance robots' personalisation abilities.

B. Incremental Learning

When training a deep learning model with new data incrementally, it suffers from catastrophic forgetting [40]. Models capable of incrementally incorporating new data knowledge, similar to how humans accumulate knowledge through memories, will be more efficient than models that need re-training every time they need to learn in a new domain. However, without having all the old and new domain datasets while training, catastrophic forgetting in deep learning happens due to the stability-plasticity dilemma [41]. The model needs good plasticity to extract knowledge from a new domain. However, significant changes in the model's parameter weights will cause failure by disrupting earlier learned representation. Maintaining the model's weights prevents formerly learned knowledge from being forgotten; on the contrary, too much stability stops the model from learning from the new domain.

Under domain incremental scenario [42], the model encounters the data from the novel domain at a time. Therefore at any given instance, the model has data accessibility only for the present domain. When the deep learning model is trained in such scenarios, its weights adapt to fine-tune towards the present domain alone, and the prior knowledge is forgotten. A straightforward way to reduce forgetting in the incremental scenario is to periodically rehearse past domain experience, incorporating it with new knowledge. This lets the model balance present and prior domain knowledge as it is trained on combined sets of prior exemplars and present domain data. Rehearsal-based domain incremental learning approaches comprise accumulating sample data from the prior domain in-memory buffers and randomly selecting them, mixing them with new data. Rehearsal-based domain incremental learning methods [43] maintain a memory buffer to accumulate domain samples from prior seen domains allowing the model to rehearse prior knowledge and new domain knowledge. Experience Replay [44] combines old domain samples with current domain data in training batches as a rehearsal.

Further, several methods use the teacher model to leverage knowledge distillation [45] to mitigate forgetting. iCaRL [46] uses a buffer as a training set for the closest mean of exemplars classifier while discouraging the model from forgetting in the later learning process. However, iCaRL is not able to learn in an incremental domain setting. Similarly, regularisation-based methods [47], [48] augment the loss function with a term that discourages the network's weights from altering abruptly.

Building on previous works [38], [39], [49], [50], we present **FIXR**, an Incremental Learning model [47], [51]–[53] for facial expression recognition where each face represents a new domain. The idea behind our model is to continuously learn the facial expressions of new individuals while the domain distribution shifts.

III. METHOD

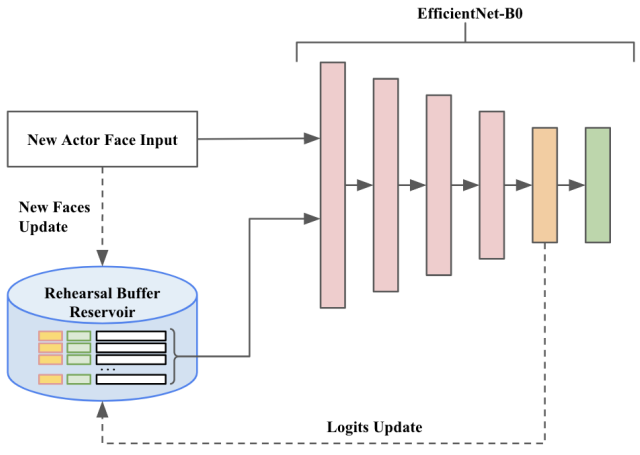

Fig. 3. **Proposed model**: During incremental training based on Dark Experience Replay [54], the model stores exemplars and model logits from the current domain in the rehearsal buffer reservoir and rehearses some of the items from the reservoir in merge with the current training domain.

Within the domain incremental learning setting [42] for facial expression recognition, the objective of a domain incremental learning model with parameter θ is to learn to perform expression recognition on an unending stream of individual domain distribution of $J = \{J_1, J_2, ..., J_3\}$ of facial images x and expression classes y where $J = \{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, ..., \langle x_n, y_n\rangle$. The model with parameter θ is trained on an initial individual base domain and is then incrementally adapted on the stream J in mini-batches.

Inspired by existing facial expression recognition methods and domain incremental learning methods, our proposed model FIXR combines the benefits of pre-trained EfficientNet [25] with experience replay, knowledge distillation of past experiences and regularisation of previously observed human facial expression. The FIXR algorithm 1 is based on Dark Experience Replay [54]. For our implementation, we adapt the EfficientNet [25] for feature extraction in combination with dark experience replay [54].

Algorithm 1: FIXR

Input: Individual facial expression J dataset, parameter θ, scalar α, learning rate λ
Output: Facial Expression e
Reservoir: $R \leftarrow \{\}$
for *each Individual* I **do**
  $\langle x_i, y_i\rangle \leftarrow$ Current individual's input face and expression classes
  $x_i^t \leftarrow transform(x_i)$
  if $M \neq \varnothing$ **then**
    $\langle x^{'}, z^{'}\rangle \leftarrow$ samples from replay buffer reservoir R
    $x^{'t} \leftarrow transform(x^{'})$
  $z \leftarrow logit_\theta(x_i^t)$ `// calculate the logit of current facial expressions`
  $\mathcal{L} \leftarrow cross\_entropy(y_i, f_\theta(x_i))$
  $reg \leftarrow \alpha \parallel z^{'} - logit_\theta(x^{'t}) \parallel_2^2$
  $\theta \leftarrow \theta + \lambda * \nabla_\theta[\mathcal{L} + reg]$
  $R \leftarrow (R, \langle x, z\rangle)$

FIXR caches exemplars and model logits from the current training domain in the rehearsal buffer reservoir. During each training stage, it incorporates some of those items with the current domain batch; as a result, the model rehearses previous domains as it is trained on current data. By employing logit sampling throughout the model optimisation trajectory, the model converges to flatter minima by resembling several different teacher parameterisations [54]. Likewise, FIXR uses previous domain exemplars and logits to align prior domain and present domain results to regularise the model parameters θ. It is to be noted that during training, all the expression classes (*i.e.* 6 classes) of the current individual are available to FIXR.

IV. Experiments

A. Datasets

To perform domain incremental learning for facial expressions, individual data with various expression labels is required, which is then made available incrementally to the model during training. We performed experiments on two datasets: RAVDESS [30] and MEAD [31] and pre-processed the images using approach [37]. In both experiments, we focused on frontal faces and a small rehearsal buffer reservoir size; we resized the facial images into 112x112 with only frontal faces. We focus on the six basic facial expressions: angry, sad, happy, surprise, disgust, and fear. We did not use neutral, calm, and contempt as the number of data for each individual varied a lot or was not provided in both datasets.

RAVDESS [30] is a gender balanced dataset composed of 7356 recordings where 24 actors vocalise lexically-matched statements. We randomly selected 20 actors for training and evaluation of our model. As the dataset was video-level annotated, we extracted each frame from each recording to represent each class. We further extracted aligned faces from each frame using MTCNN [55]. Since the RAVDESS dataset does not provide train-test splits, we split it into a train-test set with an 80:20 ratio.

MEAD [31] is a dataset composed of videos featuring 60 actors talking using eight different emotions at three different intensity levels. Out of 60 actors, we randomly selected 40 actors for the training process. Similar to RAVDESS pre-processing, we split the video into frames, extracted aligned faces using MTCNN, and split the data into six basic expression classes. **MEAD** was split into an 80% training set and a 20% test set.

B. Implementation Details and Evaluation Protocol

For training on both RAVDESS and MEAD datasets, we use all convolutional layers of Efficient-B0 [25], pre-trained on the ImageNet dataset [56]. Following the convolutional layers, we added one fully connected layer of 512 nodes with the final output of six classes. We employ random reservoir sampling [57] for storing the previous individual's face examples, labels and logits in the memory buffer reservoir. To provide a fair evaluation, we compare the performance of FIXR with Function Distance Regularisation (**FDR**) [58], an incremental learning method that fights forgetting similar to FIXR by using past exemplars and network outputs to align past and current outputs. Further, we perform **finetuning** of Efficient-B0 [25] with 512 nodes and the final output of six classes and compare its performance with both FIXR and FDR. We followed the same training regime for all three models (FIXR, FDR and Finetuning). We use a batch size of 64 with a learning rate of 0.01. We trained all three models for 20 epochs for each actor on the RAVDESS dataset, while we trained the models for 30 epochs for each actor on the MEAD dataset. For FIXR and FDR, we used a buffer size of 64 for the replay buffer reservoir, and we assigned α to be 1.0.

For the train-test regime, we incrementally provide the individual actor data to the models. For example, for the first

actor J_1, we train models with its train set and evaluate it on the same actor's test set. In the next stage, both models are provided with the second actor's datasets. The models are trained on the second actor's train set but are evaluated on the first and second actor's test sets. Following this regime, when trained on the twentieth actor on the RAVDESS dataset and the fortieth actor on the MEAD dataset, all models are tested on all previous actor's test sets.

We follow the following protocols for the evaluation. First, we evaluate, in terms of linear accuracy, if the models can perform facial expression recognition on previously seen actors. By doing this evaluation, we examine the model's ability to prevent itself from catastrophic forgetting. Second, we evaluate the performance on the current actor; as a result, we can verify the plasticity of all models. Finally, we evaluate all the model's performance based on overall accuracy. During the testing phase, all the models are domain-agnostic, meaning that at evaluation time the models are not aware from which face domain the test set is evaluated.

V. Results

We compare the performance of FIXR against FDR and Finetuning after being incrementally trained from actor 1 to actor 20 on the RAVDESS and MEAD datasets. On the RAVDESS, We observed that the overall accuracy on FIXR after training 20 actors was $(75.13 \pm 0.5)\%$, whereas, with the FDR and Finetuning, we observe the overall accuracy to be $(64.91 \pm 0.5)\%$ and $(52.15 \pm 0.5)\%$ respectively. Fig. 4 clearly shows that the performance of FIXR, when trained incrementally, outperforms all other models.

Confusion matrices of Finetuning, FDR and FIXR trained incrementally till actor 20 on RAVDESS dataset are shown in Fig. 5. Based on Fig. 4 and Fig. 5, we noticed that FIXR is able to remember (*plasticity*) all the previous actors better than Finetuning. Similarly, FIXR is able to integrate (*elasticity*) the new actor's facial expression better than other comparison models. Overall, FDR and FIXR performed better than Finetuning because these two models benefit from storing the small set of exemplars and labels, which contains lots of information about the prior domain and enables rehearsing those exemplars while training on current domains. From Fig. 5, we can see that Finetuning can perform similar to FDR and FIXR on current actor 20 however suffers severely as we go back and evaluate very old domains such as actors 11 and 1. This is because Finetuning is finetuned for the current domain and does not care about any previous domains.

Although both FDR and FIXR distil knowledge from past results based on past exemplars, when compared, FDR exploits the network appointed at the end of each domain as the single distillation signal, which causes high sensitivity to specific domains. In contrast, FIXR stores sampled logits throughout the optimization trajectory.

Regarding the comparison on the MEAD dataset, there is no difference from the previous result; FIXR performed better on the MEAD dataset than Finetuning and FDR. Fig. 6 lists out the performance of all the actor's test set concerning the accuracy, with the apparent result of FIXR performing better than the compared models.

However, if we look closer at confusion matrices in Fig. 7, the performance of FIXR degraded severely when randomly tested on actors 19, 7 and 1 of the MEAD dataset. This is primarily because of the reservoir strategy, which weakens FIXR when a sudden domain distribution change occurs in input facial data. Additionally, FIXR sample prior logits, stored in the rehearsal buffer reservoir, are severely biased by the training on the prior domain for rehearsal.

VI. Conclusion and Future Work

Our **F**ace **I**ncremental e**X**pression **R**ecognition model (**FIXR**) is inspired by the fact that each person's facial expressions are distinct from others and that incremental integration of diverse individual's facial expression is fundamental. Domain incremental learning was presented as a method for the facial expression recognition task. We presented that combining rehearsal-based, regularisation-based and knowledge distillation-based incremental learning methods can help develop individual facial expression recognition where all the individual's data are not available at the beginning of training. As each individual's facial expression data is provided incrementally, FIXR learns and adapts to new individual facial expressions with the ability to recall previously seen domains (individuals). Our experiments involve the training, evaluation and comparison of FIXR, FDR and finetuning, and we observe improvement in performance in all the evaluation protocols.

Finally, we would like to discuss our work's shortcomings, which are worth further investigation. First, our model stores exemplars, labels and model logits. However, in many real-world cases storing real images is not desirable due to privacy issues or memory constraints. To minimise this problem, a possible approach with domain incremental learning for facial expression recognition will be to adapt based on generative models [37], [38] to act as the pseudo-rehearsal mechanism. Secondly, one of the most common issues regarding data-driven methods is limited labelled data. Unfortunately, facial expression recognition is not secured from that issue. Therefore, it would be better to leverage the semi-supervised continual learning method [59] that can be integrated into the domain incremental facial expression recognition model such that the model can incrementally learn from labelled and unlabelled facial expression data.

Acknowledgment

The PERSEO project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778.

References

[1] K. R. Scherer and P. H. Tannenbaum, "Emotional experiences in everyday life: A survey approach," *Motivation and Emotion*, vol. 10, pp. 295–314, 12 1986.

[2] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–30, 2020.

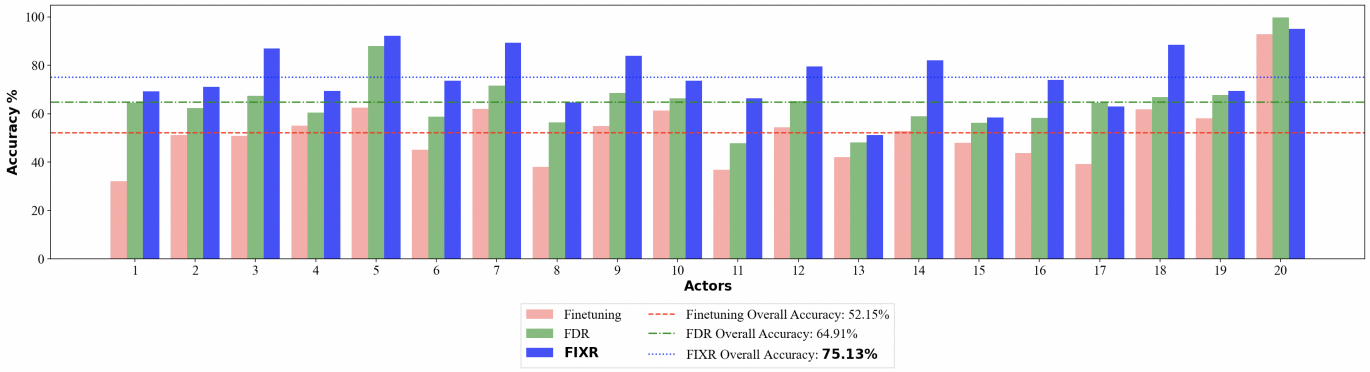

Fig. 4. Classification accuracy results on all the RAVDESS actor's test where Finetuning, FDR and FIXR are incrementally trained until Actor 20. **FIXR** can recognise the facial expressions of all the actors with 50% or above accuracy while **Finetuning** performs worse among all the other models.

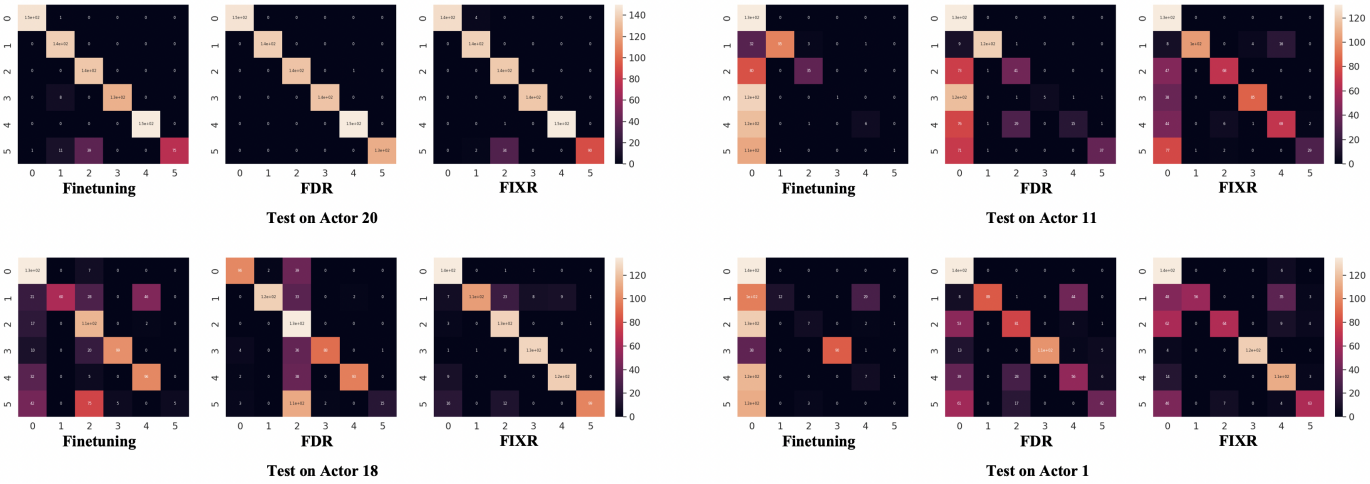

Fig. 5. Confusion matrices result of randomly selected actor test sets of finetuning, FDR and FIXR incrementally trained on Actor 20. **Note:** 0 - Happy, 1 - Angry, 2 - Sad, 3 - Fear, 4 - Surprise and 5 - Disgust

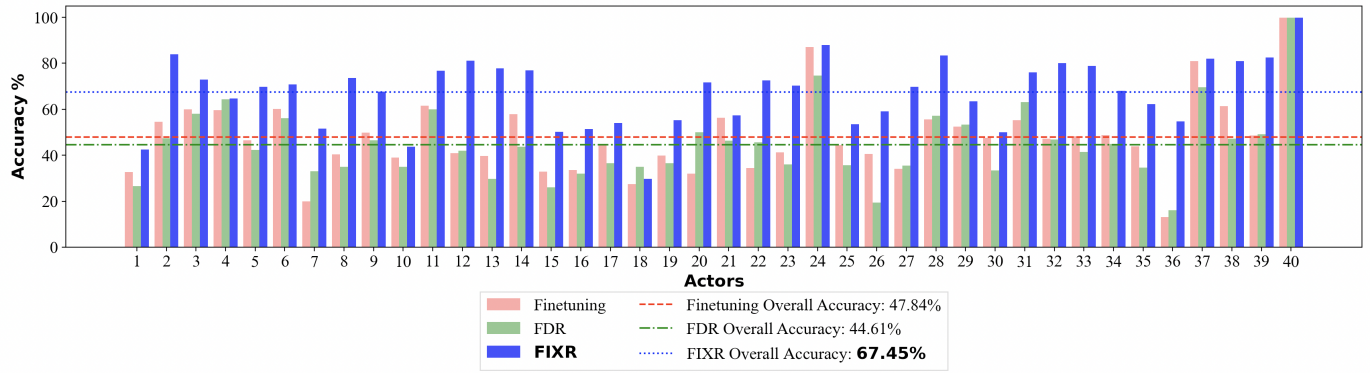

Fig. 6. Classification accuracy results on all the MEAD actor's test where Finetuning, FDR and FIXR are incrementally trained till Actor 40.

[3] R. A. Virrey, C. D. S. Liyanage, M. I. b. P. H. Petra, and P. E. Abas, "Visual data of facial expressions for automatic pain detection," *Journal of Visual Communication and Image Representation*, vol. 61, pp. 209–217, 2019.

[4] S. S. Ge, H. A. Samani, Y. H. J. Ong, and C. C. Hang, "Active affective facial analysis for human-robot interaction," *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, pp. 83–88, 2008.

[5] P. Barros, C. Weber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 582–587, IEEE, 2015.

[6] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*, pp. 200–205, 1998.

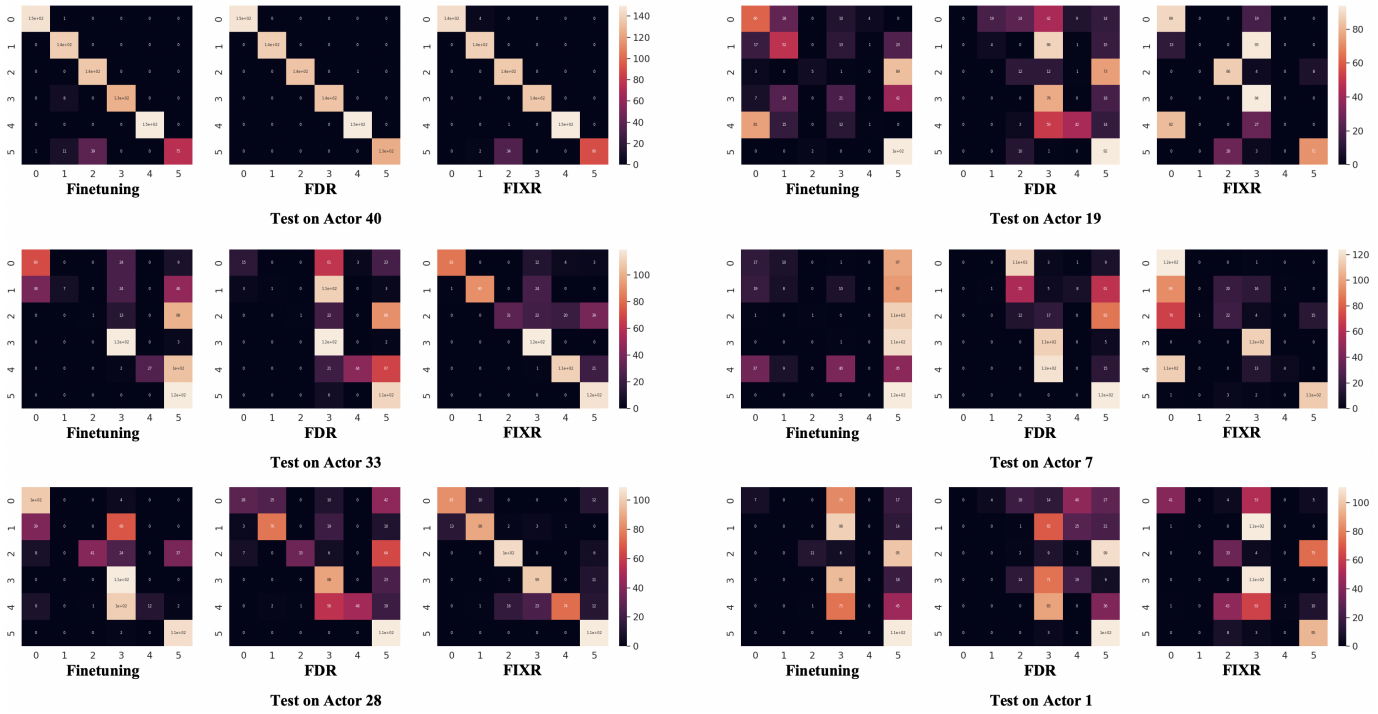

Fig. 7. Confusion matrices result of randomly selected actor test sets of Finetuning, FDR and FIXR incrementally trained on Actor 40. **Note:** 0 - Happy, 1 - Angry, 2 - Sad, 3 - Fear, 4 - Surprise and 5 - Disgust

[7] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[8] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[9] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[10] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[11] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[12] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.

[13] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[14] M. G. Calvo and L. Nummenmaa, "Perceptual and affective mechanisms in facial expression recognition: An integrative review," *Cognition and Emotion*, vol. 30, no. 6, pp. 1081–1106, 2016.

[15] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett, "Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture.," *Emotion*, vol. 14, no. 2, p. 251, 2014.

[16] J. Park, V. Barash, C. Fink, and M. Cha, "Emoticon style: Interpreting differences in emoticons across cultures," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, pp. 466–475, 2013.

[17] N. Churamani, "Continual learning for affective computing," *arXiv preprint arXiv:2006.06113*, 2020.

[18] V. C. Pammi and N. Srinivasan, *Decision making: neural and behavioural approaches*. Newnes, 2013.

[19] R. Sprengelmeyer, M. Rausch, U. T. Eysel, and H. Przuntek, "Neural structures associated with recognition of facial expressions of basic emotions," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 265, no. 1409, pp. 1927–1931, 1998.

[20] N. Churamani, M. Axelsson, A. Caldir, and H. Gunes, "Continual learning for affective robotics: A proof of concept for wellbeing," in *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, (Los Alamitos, CA, USA), pp. 1–8, IEEE Computer Society, oct 2022.

[21] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.

[22] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 2983–2991, 2015.

[23] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1805–1812, 2014.

[24] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 19–27, 2015.

[25] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations, ICLR, 2015.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[28] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[29] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *arXiv preprint arXiv:2109.07270*, 2021.

[30] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[31] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional

talking-face generation," in *European Conference on Computer Vision*, pp. 700–717, Springer, 2020.

[32] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74, IEEE, 2018.

[33] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[34] N. Churamani, S. Kalkan, and H. Gunes, "Continual learning for affective robotics: Why, what and how?," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 425–431, IEEE, 2020.

[35] P. Barros, G. I. Parisi, and S. Wermter, "A personalized affective memory model for improving emotion recognition," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 758–767, 2019.

[36] A. Lindt, P. Barros, H. Siqueira, and S. Wermter, "Facial expression editing with continuous emotion labels," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8, IEEE, 2019.

[37] 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, IEEE, 2020.

[38] S. Stoychev, N. Churamani, and H. Gunes, "Latent generative replay for resource-efficient continual learning of facial expressions," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2023.

[39] N. Churamani, O. Kara, and H. Gunes, "Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition," *IEEE Transactions on Affective Computing*, 2022.

[40] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[41] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers in psychology*, vol. 4, p. 504, 2013.

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

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

[44] D. Rolnick, A. Ahuja, J. Schwarz, T. P. Lillicrap, and G. Wayne, "Experience replay for continual learning," *Advances in Neural Information Processing Systems*, vol. 32, no. NeurIPS, 2019.

[45] G. Hinton, O. Vinyals, J. Dean, *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[46] S. A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 5533–5542, 2017.

[47] 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.

[48] M. K. Titsias, J. Schwarz, A. G. d. G. Matthews, R. Pascanu, and Y. W. Teh, "Functional regularisation for continual learning with gaussian processes," *arXiv preprint arXiv:1901.11356*, 2019.

[49] J. Han, Z. Zhang, M. Pantic, and B. Schuller, "Internet of emotional people: Towards continual affective computing cross cultures via audiovisual signals," *Future Generation Computer Systems*, vol. 114, pp. 294–306, 2021.

[50] O. Kara, N. Churamani, and H. Gunes, "Towards fair affective robotics: continual learning for mitigating bias in facial expression and action unit recognition," *arXiv preprint arXiv:2103.09233*, 2021.

[51] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[52] M. Mundt, Y. W. Hong, I. Pliushch, and V. Ramesh, "A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning," *arXiv preprint arXiv:2009.01797*, 2020.

[53] M. B. Ring, "Child: A first step towards continual learning," pp. 261–292, 1998.

[54] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: a strong, simple baseline," vol. 33, pp. 15920–15930, 2020.

[55] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 10 2016.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[57] J. S. Vitter, "Random sampling with a reservoir," *ACM Transactions on Mathematical Software (TOMS)*, vol. 11, no. 1, pp. 37–57, 1985.

[58] A. S. Benjamin, D. Rolnick, and K. P. Kording, "Measuring and regularizing networks in function space," *7th International Conference on Learning Representations, ICLR 2019*, 5 2018.

[59] M. Boschini, P. Buzzega, L. Bonicelli, A. Porrello, and S. Calderara, "Continual semi-supervised learning through contrastive interpolation consistency," *Pattern Recognition Letters*, vol. 162, pp. 9–14, 2022.