Training Against Disguises: Addressing and Mitigating Bias in Facial Emotion Recognition with Synthetic Data

Aadith Sukumar[1], Aditya Desai[1], Peeyush Singhal[1], Sai Gokhale[2],
Deepak Kumar Jain[3], Rahee Walambe[1,2], Ketan Kotecha[1],[2]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Pune, India
[2] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed) University, Pune, India
[3] Dalian University of Technology, Dalian, China

Abstract—**Facial Emotion Recognition (FER) is a challenging problem due to various challenges such as variability in expressions and ambiguity in data. Several popular benchmarking datasets, specifically employed for FER tasks exhibit bias towards ethnicity, demography and image capture mechanisms. More specifically, the images in such datasets are captured in a controlled environment and are taken in good light, with straight head orientation, no occlusion or other facial artefacts. When employed for FER, these biases may impair a model's generalizability, rendering it ineffective for FER in novel and unseen datasets. Especially, in applications involving security (access control) and identification of mal-intentions from facial expressions, it may prove inefficient. A criminal may disguise their face with make-up, headgear, and religious facial accessories and can fool the FER models trained on these biased datasets.**

To that end, this work focuses on understanding these datasets better by identifying such "good-image" bias. Methods to mitigate such bias which allows the FER models to perform better and improve the robustness are also demonstrated. A simple yet effective FER framework for studying bias mitigation is proposed. Using this framework, the performance on popular dataset is analyzed and a significant difference in model performance is observed. Additionally, a knowledge transfer technique and a synthetic image generation technique are proposed to mitigate the identified bias. Finally, using the SFEW dataset, the findings are validated on the FER task, demonstrating the effectiveness of our techniques in mitigating real-world "good-image" bias. The experiments show that the proposed techniques outperform baseline methods by averaged fourfold improvement.

Keywords: Facial Expression Recognition, Mitigating Bias, Synthetic data, disguise, good-image bias, security

Introduction

In a world where facial emotion (or expression) recognition (FER) technologies are increasingly used for behavior prediction, security and access control, identification of mal-intentions, and education, amongst others, it is imperative that such techniques work accurately and reliably irrespective of image discrepancies. Many a times, the FER techniques are used for the identification of criminals or persons with mal-intentions. It is extremely critical, especially in such security-specific applications to recognize the person and their emotion, irrespective of the disguise they may wear. Disguises can be rendered by applying make-up to look older or change skin color, wearing headgear such as a turban, wearing religious facial accessories such as a Hijab etc. Many standard benchmarking datasets and also in-the-wild datasets such as SFEW [1] contain images which are captured in clear light, with straight head orientation and without any headgear or disguise in lab-controlled environments or studio settings. The FER models trained on such data specifically have the "good-image" bias. The bias can manifest in different ways, such as overemphasizing cultural norms in emotion recognition or misinterpreting expressions and emotions from certain demographics.

The models trained on these biased datasets may not generalize well for real-world deployment where people may wear disguises to evade security systems. The disguise could be in the form of makeup, headgear, spectacles, beard etc. This leads to misclassifications and reduced performance reliability. Embracing diversity in general and in terms of training datasets of FER systems is a key factor in alleviating demographic bias. Alleviating bias in FER systems is vital for ensuring fairness, accountability and responsible use. This research focuses on exploring techniques to address bias in FER. The key contributions of this work are:

1. Mitigation of the good-image bias in benchmarking datasets in FER using synthetically generated data
2. Demonstration of techniques and strategies for the generation of rich synthetic data using Gen-AI tools.
3. Experimental proof that that model trained on both real and generated content is more generalized in nature.

The paper is organized in four sections. Section I.A contains the related work in FER. Section 0 discusses the methods, dataset and bias mitigation techniques. Section **Error! Reference source not found.** includes the system design of the proposed methodology. Results and Experiment details are included in section 0.

979-8-3503-9494-8/24/$31.00 © 2024 IEEE

A. Related Work

1) Facial Emotion Recognition (FER)

Humans convey more using facial expressions and emotions rather than speech. Surveys ([2], [3], [4]) mention the ever-increasing usage of FER systems in everyday life – healthcare, security, driving and the importance of affective computing. They also mention the usage of laboratory-controlled datasets and in-the-wild datasets, with a constraint of data annotation consistencies and demographic bias (e.g., age, gender, race) engrained. Biases in the FER systems [5] and their growing popularity now demand validation of datasets, applications, approaches, architectures and bias mitigation methods in FER technology. The growing awareness of bias and the use of deep learning in modern FER architectures [2] has seen a contribution to bias mitigation in deep FER systems ([6], [7]).

2) Mitigating bias in FER

Biases in FER systems stem from models [8], processes, people and datasets [9]. The deep models learn class (emotion labels) discriminative features and can learn strongly correlated biases; these biases get accentuated when there is data imbalance in datasets [8]. At a high level, the approaches that are used for mitigating bias in deep FER systems include dataset augmentation, attribute-aware algorithms, feature-disentangled approaches [7] and adversarial-based approaches. If the protected attributes are labelled then, it is possible to employ statistical methods for correcting the bias [10] or employ contrastive-based approaches [11]. However, the availability of labelled protected information is not possible most of the time and in those cases typically adversarial-based approaches ([7], [12], [13] [14]) include domain adaptation ([15], [16]) are employed.

3) Use of Synthetic Images in FER

GANs [17] based generated synthetic images in FER are used due to few or no labels in many datasets. Endeavour of GAN-based ([17], [18]) strategies is not only to learn domain-invariant features and to mitigate domain shift but also to transfer local features. Abbasnejad et al [19] created synthetic images by using 3D-CNN Network. A multi-label variant of SMOTE - MLSMOTE [20] aimed to produce samples linked to the minority class. This work includes the use of generative AI for synthetic image generation to augment the existing dataset.

4) Bias Mitigation Techniques

The bias mitigation techniques can be segregated into 3 main categories depending on stage of treatment/intervention of the model training process.

a) Pre-training

Wang et al. [10] mention the use balanced data (mentioned as strategic sampling) achieved as under sampling of the majority class or over sampling of minority classes, as one of the simplistic techniques to alleviate bias. Data augmentation is also a beneficial strategy to support over sampling, especially in case of low-resource settings. In this work, we use synthetic images using generative AI rather than existing strategies ([19], [20]) to increase the number of data available for network to learn, further these images include obfuscations and accessories to help learn the model learn better.

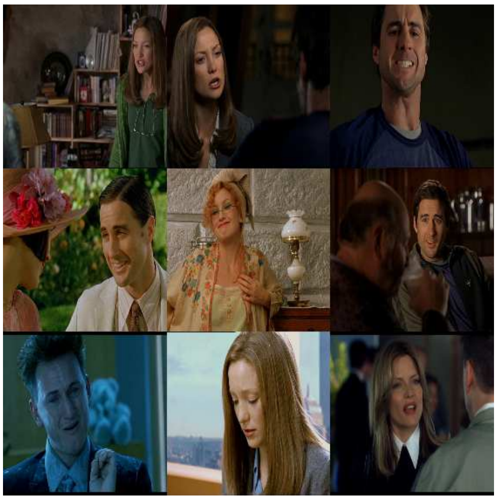

Fig. 1. A Sample of the Training Data from real images of class angry, happy and surprised respectively.

b) In-training

Data augmentation strategies are also used during in-training phase e.g., this work uses random cut-out, rotation, flip, blur, grayscale, contrast, equalize, sharpness, and jitter. Weighted loss is another strategy to alleviate dataset imbalance related effects. Other strategies include:

- Adversarial training: In this strategy, task classifier ability to predict the correct class is maximized while the adversary's ability to predict the protected variable is minimized. This strategy is also known as *Fairness through blindness* [21], because the model does not look at the information related with protected variable lead to alleviated bias.
- Domain discriminative training: In this strategy, explicit information about protected variables is learnt. This strategy is also therefore also known as *Fairness through awareness* [22] as the model is more 'aware' of each protected variable which may help to account from bias.
- Domain independent training: In this strategy, separate classifiers are trained for each protected variable, however feature extraction layers are shared [10].
- Domain adaptation: This extends adversarial training strategy, with a difference that we have multiple datasets and the bias alleviation is due to inclusion of "less" bias dataset [23]. Xu et al. [7] created a disentangled approach, very similar to domain adaptation but with one domain only, where they made sure that the representations do not contain specific information about protected variables.

c) Post-Training

This strategy relies on quantifying bias and then attempting to counter the effects of classification, taking into account the quantified bias. Michael et al [24] have used this strategy for improved fairness in diverse applications.

Datasets, Methods And System Design

Datasets- SFEW

Inspired by [24] the dataset employed for this project was Static Facial Expressions in the Wild (SFEW 2.0 [1]). It was developed by choosing static frames from the AFEW dataset by calculating keyframes based on the concept of facial landmark clustering. The SFEW 2.0 has been segregated into 3 separate sets based on the usage. The train set has 958 samples, validation samples amount to 436 and the test set holds 372 samples each. The images are further differentiated into 7 expressions (6+Neutral). The expressions are namely; anger, disgust, fear, neutrality, happiness, sadness, and surprise. Only Train and Validation sets are used as labels for the Test set are not available.

Sythethic Image Generation

Synthetic images were generated keeping the data from SFEW 2.0 as reference images. The Synthetic facial images were created using a stable diffusion/realistic-vision-v51 model. For focus on the facial details, ControlNet was utilized to capture nuances of emotions of various images. The control strength parameter was set to 0.75 to ensure accurate control over the facial expressions. Furthermore, the resolution of the images was set to 512 x 512 pixels to maintain clarity and detail so as to match the original data. The generation process and observation were meticulous, ensuring a thorough progression in the image synthesis. Additionally, a guidance scale of 9 was implemented to provide adequate guidance throughout the generation process. Finally, the Deformable Part based Model [DPM] Solver++ sampler was utilized to optimize the sampling method, ensuring high-quality and relevant results. We randomly selected images from each of the 7 emotions and generated 20 synthetic images with various objects. The objects ensured minimal hindrance of emotions; the objects employed varied from hats, sunglasses, hijab and so on.

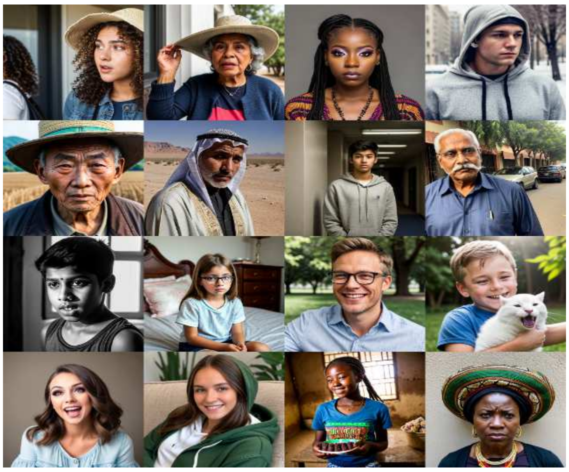

Fig 2. Generated Images of Various Emotions

B. *Image Pre-processing*

In order to improve the effectivity of emotion recognition, face recognition and cropping to 224 x 224 were done. The preprocessing (see Image preprocess in Fig. *3*) was done using pretrained pytorch face detection model (pytorch-facenet) - Multi Task Cascaded Convolutional Neural Network or MTCNN, based on FaceNet [25] . The process is same for SFEW as well as for synthetic generated images. Further, the pre-process is re-used at time of validation and inference too.

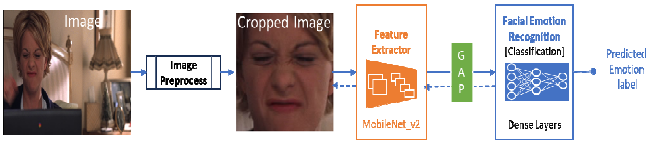

Fig 3. Image pre-processing and training pipeline for facial emotion recognition. Feature extractor is a pre-trained Mobilenet_v2 backbone which feeds into global average pooling (or GAP) layer. Finally, a set of dense layers effectively do the classification task of emotion recognition.

C. *Facial Emotion Recognition - Network*

The base model is an ImageNet pretrained mobilenetV2 as a feature extractor backbone. The last layer of the original mobilenetV2 is removed and a global average pooling (GAP) layer which gives 1280 feature points.

$$GAP(x_{ij}) = \frac{1}{H \times W} \sum_{m=1}^{H} \sum_{n=1}^{W} x_{ijmn}$$

where x_{ij} is the input feature map, H is the height, and W is the width.

These feature points are then fed into a task (emotion label) classifier which reduces the number of features in two steps from 1280 to 100 and from 100 to 7 (i.e., number of classes). ReLU is used as the nonlinearity in the task (emotion label) classifier. In Training, data augmentation was used as a tool for regularization and to reduce overfitting. Various data augmentation strategies were applied with randomization. The strategies include resizing and crop, Horizontal flip, Rotation (15 degrees), Gray Scaling, Color Jitter, Gaussian Blur, Sharpness, Auto Contrast, Image equalization and 8 cutouts. During validation, testing, and inference, no data augmentation was used. Cross Entropy (CE) classification loss is used for multi-class emotion classification.

$$CE(p, q) = -\sum_{i=1}^{C=7} p_i \log(q_i)$$

where p is true probability distribution (one-hot encoded label), q is the predicted probability distribution (softmax output) and in our case number of emotion classes C is 7.

Experiments and Results

The experiments were conducted to understand how generated synthetic data can be used for commercial and sensitive FER AI systems. We designed the experiment in 4 steps, with step-1 as our baseline and no generated synthetic data is used. Step-

2, step-3a and step-b use generated synthetic data in training or testing or both training and testing. The flow of the steps is described in Fig 4 pictorially and in following sections.

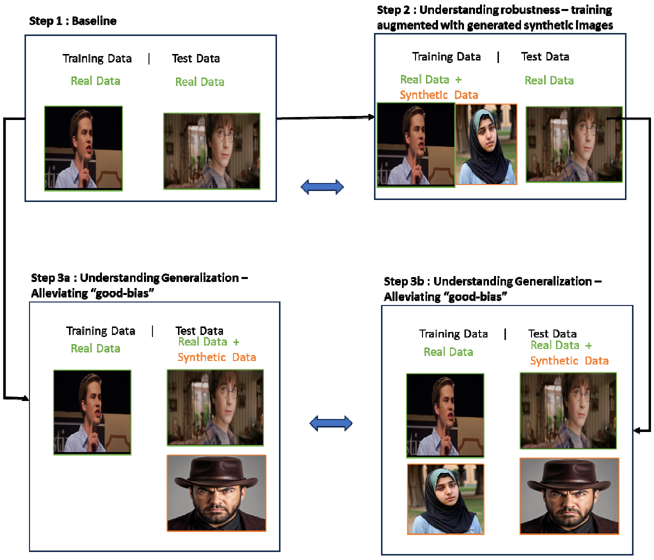
Step 1 : Baseline
Training Data | Test Data
Real Data
Real Data
Step 2 : Understanding robustness – training augmented with generated synthetic images
Training Data | Test Data
Real Data + Synthetic Data
Real Data
Step 3a : Understanding Generalization – Alleviating "good-bias"
Training Data | Test Data
Real Data
Real Data + Synthetic Data
Step 3b : Understanding Generalization – Alleviating "good-bias"
Training Data | Test Data
Real Data
Real Data + Synthetic Data

Fig 4. Experiment steps. Step 1 is baseline model with only real-data usage in both training and testing. Step-2 use synthetic data in training only, step-3 use synthetic data in testing only, and step-4 used synthetic data in both training and testing.

D. Step 1 : Baseline

After image preprocessing, the pre-trained architecture (mentioned in Fig. 3 and Section III.C) is fine-tuned (with all layers unfrozen) with SFEW 2.0 [1] training (see Fig. 1) and validation data. This model provides the baseline value for the metrics – average accuracy across classes.

E. Step 2 : Understanding robustness -Training augmented with Synthetic Images

This step is very similar to the baseline step, with a difference that the training data not only include SFEW 2.0 [1] training data, but also synthetic images generated for various emotions (images from Fig. 1 and Fig. 2). 117 synthetically generated images were added across emotions (Angry-16, Happy- 16, Neutral-15, Sad -15, Surprise-17, Fear -20, Disgust-18). The intention behind this step, is to understand if synthetically generated images improve the baseline accuracy metric.

F. Step 3 : Understanding Generalization – Alleviating "good-bias"

Typically, networks work best when training and testing data are from the same distribution. However, for commercial deployments and real-world settings, the model should provide performance on un-seen and real-world images (including occluded, images having head-gears, and other diversities). With that intention, models created in step-2 and step-3 for generalization or "good-bias" alleviation were tested on synthetic images. Efforts were taken to have this synthetic testing set to be different from synthetic training used in step-2. 89 test images across emotions were used: Angry-14, Happy- 14, Neutral-12, Sad -14, Surprised-12, Fear -8, Disgust-15. Both models created as part of step-1 (called step-3a) and step-2 (called step 3b) are used to understand generalization and mitigation of "good-bias" and are called step-3a and step-3b in TABLE I.

G. Discussion and Analysis

TABLE I shows the performance metrics of all the steps. From average accuracy result difference between step-1 and step-2, we can conclusively say that there is an improvement in robustness of the model. This may be attributed to both quantity and quality of data. On one hand, more images (part of synthetic generated data) lead to larger epochs resulting in more training translated into better testing metric. On other hand, the model now witnessed images from different distribution which help make the model learn better. This is an example of pre-training bias-mitigation strategy.

Improvement in the average accuracy between steps-3a and steps-3b, points to better generalization (including out-of-domain generalization). Even though the synthetic testing data used in training and testing was vastly different, the model created as part of step-2 is able to generalize better to unseen / different distribution images than model trained as part of step-1.

TABLE I : PERFORMANCE METRIC OF EXPERIMENT STEPS.

Experiment Steps	*Data used*	*Average Accuracy (%)*
Step-1: Baseline	• Training: SFEW 2.0 • Testing: SFEW 2.0	45.58
Step-2: Understanding Robustness - Training augmented with Synthetic generated data	• Training: SFEW 2.0 + Synthetic generated data • Testing: SFEW 2.0	50.59
Step-3a: Understanding Generalization – alleviating good-bias	• Training: SFEW 2.0 • Testing: Synthetic generated data	21.34
Step-3b: Understanding Generalization – alleviating good-bias	• Training: SFEW 2.0 + Synthetic generated data • Testing: Synthetic generated data	38.20

TABLE II : CLASS (EMOTIONS) ACCURACY COMPARISON BETWEEN STEP-1 AND STEP-2 : UNDERSTANDING ROBUSTNESS.

Emotions	*Test Data: SFEW 2.0*	
	Step 1 (%)	*Step 2 (%)*
Angry	77.33	52.00
Disgust	31.82	31.82
Fear	11.63	9.30
Happy	70.83	87.50
Neutral	17.31	42.31
Sad	47.62	58.33
Surprise	29.58	39.44

TABLE III : CLASS (EMOTIONS) ACCURACY COMPARISON BETWEEN STEP-3A AND STEP-3B: UNDERSTANDING GENERALIZATION – ALLEVIATING "GOOD-BIAS"

Emotions	***Test Data: Synthetic generated data***	
	Step 3a (%)	***Step 3b (%)***
Angry	14.29	28.57
Disgust	6.67	60.00
Fear	12.50	12.50
Happy	85.71	85.71
Neutral	0.00	25.00
Sad	0.00	0.00
Surprise	25.00	41.67

From TABLE II, it can be seen the class accuracy has improved for 5 out of 7 classes. Again, this can be attributed towards the quantity and quality of data that step 2 model has seen over step 1. Similarly, from TABLE *III*, we can see that there is an improvement of class accuracy of emotionfs across the board. An example in this case is emotion – Neutral: The model not trained on synthetic data is not able to recognize the emotion at all. The model trained on synthetic data is able to infer much better. However, for emotion – Sad, none of the models is able to give a good performance. This possibly can be attributed to sad emotion overlap with other emotions.

Conclusion

Bias-mitigation is an imperative requirement for commercial and sensitive FER systems, especially employed for tasks such as access control, detection of abnormal and mal-behavior etc. Typical benchmarking datasets have a bias towards good images which are captured in good light and certain controlled conditions. However, if an individual wears a disguise, then FER systems trained on such good-image biased datasets are incapable of performing robustly. To that end, in this work, we showed that synthetic data, when used in train and test is able to improve the three cornerstones of good AI systems – robustness, "good-bias" alleviation and out-of-domain generalization. Real data is difficult and tardy to procure and process. The use of generated synthetic data can alleviate this problem while improving the performance and bias metrics too.

This work also uncovers the limitations and difficulty in inferring certain emotions -further work must be undertaken to discover the reasons and possibly quantify them.

References

1. A. Dhall, R. Goecke, S. Lucey and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, Spain, 2011.
2. W. D. Shan Li, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing,* vol. 13, pp. 1195-1215, 2022.
3. M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hijji, K. Muhammad and J. J. Rodrigues, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal,* vol. 68, pp. 817-840, 2023.
4. O. S. Ekundayo and S. Viriri, "Facial Expression Recognition: A Review of Trends and Techniques," *IEEE Access,* vol. 9, pp. 136944-136973, 2021.
5. E. Kim, D. Bryant, D. Srikanth and A. Howard, "Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*, New York, New York, 2021.
6. T. Wang, J. Zhao, M. Yatskar, K.-W. Chang and V. Ordonez, "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
7. T. Xu, J. White, S. Kalkan and H. Gunes, "Investigating Bias and Fairness in Facial Expression Recognition.," in *In Computer Vision – ECCV 2020 Workshops*, 2020.
8. V. Suresh and D. C. Ong, "Using Positive Matching Contrastive Loss with Facial Action Units to mitigate bias in Facial Expression Recognition," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Nara, Japan, 2022.
9. Y. Chen and J. Joo, "Understanding and Mitigating Annotation Bias in Facial Expression Recognition," in *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
10. 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 *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
11. Y. Hong and E. Yang, "Unbiased classification through bias-contrastive and bias-balanced learning," in *Advances in Neural Information Processing*, 2021.
12. B. H. Zhang, B. Lemoine and M. Mitchell, "Mitigating unwanted bias with adversarial learning," in *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.
13. D. Madras, E. Creager, T. Pitassi and R. Zemel, "Learning Adversarially Fair and Transferable Representations," in *in International Conference on Machine Learning*, 2018.
14. M. Alvi, A. Zisserman and C. Nellaker, "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings," in *in Proceedings of the European Conference on Computer Vision (ECCV) workshops*, 2018.
15. B. Bozorgtabar, M. S. Rad, H. K. Ekenel and J.-P. Thiran, "Using Photorealistic Face Synthesis and Domain Adaptation to Improve Facial Expression Analysis," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019.
16. P. Singhal, R. Walambe, S. Ramanna and K. Kotecha, "Domain Adaptation: Challenges, Methods, Datasets, and Applications," *IEEE Access,* vol. 11, pp. 6973-7020, 2023.
17. X. Wang, X. Wang and Y. Ni, "Unsupervised Domain Adaptation for Facial Expression Recognition Using Generative Adversarial Networks," *Computational Intelligence and Neuroscience,* pp. 1687-5265, 2018.
18. B. Bozorgtabar, D. Mahapatra and J.-P. T. , "ExprADA: Adversarial domain adaptation for facial expression analysis," *Pattern Recognition,* vol. 100, 2020.
19. 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.

20. F. Charte, A. J. Rivera, M. J. d. Jesús and F. Herrera, "MLSMOTE:
21. Approaching imbalanced multilabel learning through synthetic instance generation," *Kowledge Based Systems,* vol. 89, pp. 385-397, 2015.

22. M. Alvi, A. Zisserman and C. Nellaker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proceedings of European Conference on Computer Vision (ECCV) workshops*, 2018.

23. C. Dwork, M. Hardt, T. Pitassil, O. Reingold and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012.

24. Y. S. Kong, V. Suresh, J. Soh and D. C. Ong, "A Systematic Evaluation of Domain Adaptation in Facial Expression Recognition," *arXiv,* 2021.

25. M. P. Kim, A. Ghorbani and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Conference on AI, Ethics, and Society*, 2019.

26. F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.