Balancing the Scales: Enhancing Fairness in Facial Expression Recognition with Latent Alignment*

Syed Sameen Ahmad Rizvi**[0000−0002−3919−5074], Aryan Seth**[0009−0005−4171−4901], and Pratik Narang[0000−0003−1865−3512]

Birla Institute of Technology and Science, Pilani
{p20190412,f20212221,pratik.narang}@pilani.bits-pilani.ac.in

Abstract. Automatically recognizing emotional intent using facial expression has been a thoroughly investigated topic in the realm of computer vision. Facial Expression Recognition (FER), being a supervised learning task, relies heavily on substantially large data exemplifying various socio-cultural demographic attributes. Over the past decade, several real-world in-the-wild FER datasets that have been proposed were collected through crowd-sourcing or web-scraping. However, most of these practically used datasets employ a manual annotation methodology for labelling emotional intent, which inherently propagates individual demographic biases. Moreover, these datasets also lack an equitable representation of various socio-cultural demographic groups, thereby inducing a class imbalance. Bias analysis and its mitigation have been investigated across multiple domains and problem settings; however, in the FER domain, this is a relatively lesser explored area. This work leverages representation learning based on latent spaces to mitigate bias in facial expression recognition systems, thereby enhancing a deep learning model's fairness and overall accuracy.

Keywords: Bias Mitigation · Facial Expression Recognition · Fairness

1 Introduction

Facial expression recognition (FER) has been an extensively explored problem in the field of deep learning and computer vision. In the past decade, numerous proposed FER datasets have made it easier to approach facial expression recognition as a supervised deep-learning task. Deep learning requires large and diverse datasets for efficaciously modelling data distribution. However, such a supervised learning strategy necessitates substantial training data that reflects a wide range of socio-cultural demographic characteristics.

Over the past decade, various real-world, in-the-wild datasets have been proposed using web-scraped/crowd-sourced images. A crucial drawback of employing such a data-driven method for expression recognition lies in its susceptibility to biases present in the datasets, particularly those that disproportionately affect

* Supported by Kwikpic AI Solutions.

** denotes equal contribution and joint first authorship specific demographic groups.[3, 11]. Facial Expression Recognition requires human annotations per image, which propagates annotative biases and prejudices. Moreover, most real-world in-the-wild datasets lack proportionate representation of different demographic attributes such as race, age, and gender. Another crucial factor contributing to bias in FER datasets is crowd-sourced annotation. Each annotator possesses their own bias with respect to understanding facial expressions in varied demographics. However, given the enormous size of datasets, these biases are often assumed to be components of random noise.[2, 47].

In practice, however, people may harbour systematic and demographic biases, especially when inadequately trained with proper demographic and psychological knowledge; they may incorporate such biases into their annotations [6]. Bias is defined as systematic mistakes that result in unjust outcomes during a decision-making process. In the realm of deep learning, this can originate from multiple factors, such as data collection methodology, algorithm design, and biased human annotation [7]. A deep learning model trained on such datasets would inherently propagate bias, thus making it unfair. Fairness in the context of deep learning refers to the absence of bias or discrimination in deep learning systems; however, achieving it can be difficult since deploying a real-world deep learning solution can propogate biases that can emerge in such systems.

Annotative biases combined with class and demographic imbalances increase bias and reduce equal-odds fairness for attributes such as gender, ethnicity, etc. Therefore, examining the biases within datasets and designing algorithms to mitigate them becomes crucial. Considering age as a protected attribute in datasets, we observe that adolescents are represented positively (such as happy) [6]; on the contrary, senior citizens are represented more negatively (such as sad and disgusted). This causes models to be biased, with adolescents being classified more frequently to positive expressions, viz-a-viz, and senior citizens being predicted to negative expressions.

Bias analysis and its mitigation strategies have gained good traction among researchers working in the facial analysis domain. However, in the FER domain, this is a relatively less explored area [34, 42]. This work is our attempt to tackle and mitigate this bias, therefore increasing fairness in a deep learning model. The major contributions of this work include:

- A novel latent alignment technique with an architecture that generates improved latent representations, mitigates bias, and improves accuracy for FER.
- A novel training technique and loss function that uses Variational Autoencoders and an adversarial discriminator with perceptual loss for bias mitigation and a CNN backbone for expression classification.
- Conducting extensive evaluation on two commonly used datasets (RAF-DB [26] and CelebA[28]) and multiple protected attributes in both separate and combined techniques, mitigating bias towards gender, race, and age, setting new state-of-the-art results and competitive performance.

This paper is an extended version of our Student Abstract published at AAAI-24[35], which, to the best of our knowledge, is the first attempt to explore repre-

sentation learning using latent spaces in mitigating biases in the facial expression domain. This paper provides more comprehensive experimentation with an additional dataset (CelebA[28]), detailed results on the interplay between different protected attributes, and better insights into our methodology and training approach.
The rest of the paper is organised as follows: Section 2 discusses some recent notable works in bias mitigation. Section 3 describes the methodology adopted, including the training methodology, loss functions, and classification model employed. Section 4 presents our experimental results, the evaluation metric and analysis of datasets. Section 5 provides a component-wise ablation study of our proposed architecture. Section 6 concludes the work and presents directions for future work.

2 Recent Works

Bias in Machine learning has attracted wider attention in recent years, with the rapid growth in the deployment of real-world machine learning applications. Extensive surveys[29, 17, 9, 32] have been done to study bias and its mitigation strategies. In this section, we discuss some of the notable methods for mitigating biases. In literature[9] three types of bias mitigation strategies have been discussed, namely, pre-processing, in-processing, and post-processing methods.

Pre-processing Methods: Calmon et al. [4] proposed an optimized pre-processing strategy that modifies the data features and labels. Zemel et al. [43] proposed a mitigation strategy that learns fair representations by formulating fairness as an optimization problem of finding good representations of the data while obfuscating any information about membership in the protected group. Feldman et al. [14] proposed disparate impact remover, where feature values were modified while preserving rank ordering to improve overall fairness.

In-processing: Kamishima et al. proposed a prejudice remover mechanism [23] that leverages a discrimination-aware regularization approach to the learning objective that can be applied to any prediction algorithm with probabilistic discriminative models. Zhang et al. [45] proposed a strategy that learns fair representations by including a variable for the group of interest and simultaneously learning a predictor and an adversary. Meta Fair Classifier [5] proposes a meta-algorithm for classification that takes fairness constraints as input and returns an optimised classifier.

Post-processing: Reject option Classification [22] presents a discriminative aware classification, which essentially aims at the prediction that carries a higher degree of uncertainty and thereby assigns favourable outcomes to unprivileged groups and unfavourable outcomes to privileged groups. The strategy of calibrated equalized odds [33] is designed to optimise the calibrated classifier score outputs. Its goal is to identify probabilities that can be used to alter output labels while maintaining an objective of equalized odds.

Some other techniques to tackle *dataset bias* include transfer learning[31], adversarial mitigation[46, 39], and domain adaptation [36–38]. Various strategies have been proposed to eliminate or prevent models from acquiring misleading or unwanted correlations. A post-hoc correction technique [15] that imposes an equality of odds constraint on previously learnt predictor. In the domain of deep learning, two popular techniques are the tweaking of loss functions to impose penalties on unfairness[1], and adversarial learning [45, 20, 30]. These techniques aim to learn a fair representation that is devoid of any information related to protected attributes.

Bias mitigation in Facial Expression Recognition: Bias mitigation in facial expression recognition is a relatively less-explored area. With the exponential increase in computing capabilities over the past decade, many datasets and algorithms have been proposed for automatically recognizing facial expressions. However, most of these in the wild real-world datasets are either web-scraped or crowd-sourced. These datasets often have two major limitations [25]. Firstly, most datasets have class imbalances; i.e. people with varied socio-cultural-ethnic identities are inadequately represented among various classes. Secondly, since these large numbers of scraped images are manually labelled by a group of annotators, a personal bias is inherently a part of the dataset.

Some of the existing works that have tackled bias and it's mitigation in facial expression recognition include a facial Action Unit (AUs) calibrated FER approach [8], an attribute aware and a disentangled method [42]. Zeng et al. [44] proposed a Meta-Face2Exp framework that utilized large unlabelled facial recognition datasets.

3 Methodology

We propose a two-part model for mitigating bias. Recognizing that CNNs tend to learn from all input features, for the first part of the model we propose a Variational Autoencoder (VAE) to encode the images into a latent space. The images corresponding to each protected attribute in the dataset will each have a corresponding latent space. Our goal is to minimize the distance between these latent spaces so that each latent encodes only the information relevant to expression classification. We propose to utilize a Variational Autoencoder with shared weights for all protected attributes where the inter-latent domain gap is reduced using an adversarial discriminator. We denote the Encoder part as E and the Generator part as G. We introduce a two-part model to address bias mitigation. Given CNNs' propensity to assimilate all input features, our initial model component employs a Variational Autoencoder (VAE) to encode images belonging to protected attributes into the common latent space. The goal is to minimise disparities between these latent spaces, ensuring they contain information relevant to expression classification.
Summarising the methodology:

- The main cause of bias is that models tend to learn protected attributes as features.
- Our model solves this by generating a latent that has forgotten the protected attribute.

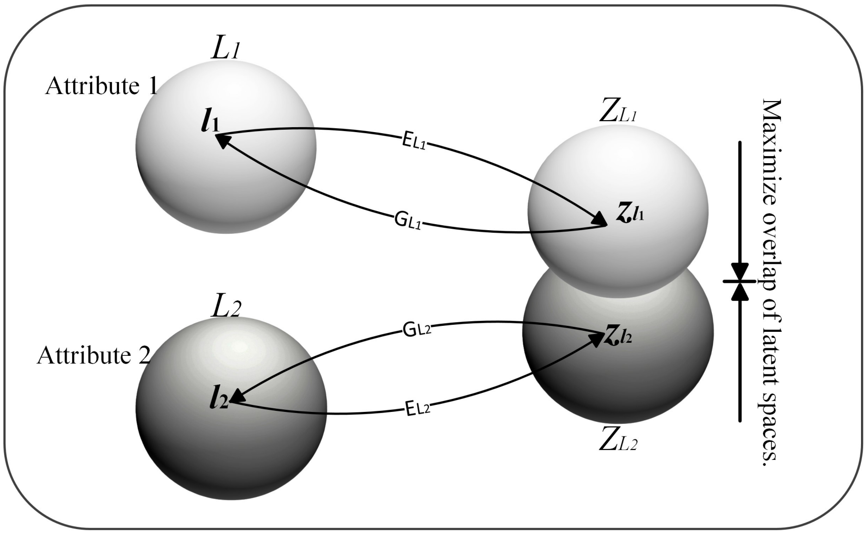


Fig. 1. Architecture for Attribute Disentanglement. L_i represents data having the attribute q_i. Z_{L_i} is the latent representation of L_i. E_{L_i} is a VAE with shared weights $\forall i$. 'E' refers to the Encoder module, which compresses the input image into a latent that does not contain information about the protected attribute. 'G' refers to the Generator, which is a reconstruction module that converts the latent back to the original image.

- This is done by overlapping the latent spaces of data points belonging to different protected attributes; this overlap is done using the discriminator.

Attribute Disentanglement - We propose a shared-weight Variational Autoencoder across all protected attributes, mitigating inter-latent domain disparities through an adversarial discriminator. In this context, we denote the Encoder and Generator components as 'E' and 'G'. This is demonstrated in Fig. 3, where q_i is a protected attribute such as gender.

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}(x) =& \mathrm{KL}\left(z_x \mid x\right) \| \mathcal{N}(0, I)) + \mathcal{L}_{\mathrm{VAE,D}}^{\mathrm{Latent}}(x) \\ &+ \alpha \left\| G_j^{\phi}(\hat{y}) - G_j^{\phi}(y) \right\|_F^2\end{aligned} \tag{1}$$

Equation 1 is the objective function for the VAE. The first component consists of KL-divergence that penalizes deviation of the latent distribution from a Gaussian Distribution. The second component is discriminator loss, which measures whether the discriminator can predict the protected attribute class. The final component is Style-Reconstruction Loss [21].

Classification Model We feed the latent representation generated by E into a custom classification module using MBConv[18] blocks. This is demonstrated in Fig. 2.

$$\min_{E_{\mathcal{X}_i},G_{\mathcal{X}_i}} \max_{D_{\mathcal{X}_i}} = \mathcal{L}_{\text{VAE}}(x) + \mathcal{L}_{\text{VAE,D}}^{\text{latent}}(x_{q_i}) \quad \forall q \tag{2}$$

Training Method The Encoder and the Discriminator are trained jointly with a min-max objective function (Equation 2) with a categorical cross-entropy loss for the Discriminator. The classification model is trained after the VAE with a symmetric cross-entropy loss for robustness.

Training Configuration The training was conducted on 2 NVIDIA Tesla V100s with 32 GB of GPU memory. A Stochastic Gradient Descent Optimizer with a learning rate set to 0.0001 and momentum set to 0.9 was used. Hyper-parameter α from L_{VAE} from Equation 1 in the paper was set to 10 after grid search.

RAF-DB [26] provides images resized to 128x128 pixels. We applied basic augmentations to our dataset, including horizontal flips with a probability of 50% and random rotations by a maximum angle of 15°.

Loss Functions The proposed model has a novel loss function (Equation1), which consists of three parts. The first part is the KL Divergence between the latent and a sample from a Gaussian distribution with mean 0 and variance 1 according to [24]. This is used to provide denser representations in the latent space, improving accuracy and mitigating bias (as shown later in Section 5).
The second component is the loss from the discriminator's ability to predict the protected attribute accurately. The Encoder's goal is to be able to fool the discriminator into not knowing the protected attribute. This is the main component that aligns the latent spaces and ensures the Encoder does not learn the protected attribute features.
The final component is the Style-Reconstruction Loss from [21], which is added to ensure that the semantic emotion-level features are not lost on the Generator's reconstruction of the image. This is used instead of a pixel-wise loss because expression is a subjective concept, and a pixel-wise loss does not necessarily represent semantic consistency.

$$G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \tag{3}$$

Equation 3 is the Gram matrix of the j_{th} feature map for a network ϕ where $\phi_j(x)$ represents the activations of the $j_t h$ layer of the network. The final loss is the squared Frobenius norm of the input and output feature matrices.

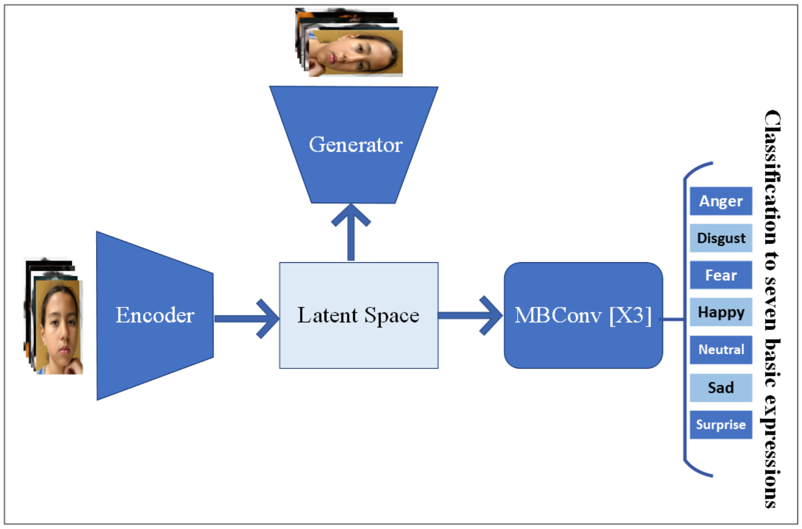


Fig. 2. Classification backbone uses the latent representation generated by the encoder to classify into the 7 emotions.

Classification Model: We have used 3 sequential MBConv [19] modules which use the latent representation generated by the Latent Alignment VAE and classify it into the seven basic expressions. The MBConv block has been extensively explored in many areas of deep learning and is a versatile and efficient building block. We have also experimented with using Residual Blocks [16] and found that they have a minor reduction in accuracy (described further in Section 5).

4 Expermineation, Results, and Analysis

4.1 Evaluation Metric

We formulate our metric for fairness as [42] and use the "equal odds" philosophy.

$$\mathcal{F} = \min(\frac{\sum_{c=1}^{C} p\left(\hat{y} = c \mid y = c, q = q_i, \mathbf{x}\right)}{\sum_{c=1}^{C} p(\hat{y} = c \mid y = c, q = d, \mathbf{x})}. \tag{4}$$

$$\forall i \in (1, 2....N)$$

In equation 4, "d" is the protected attribute that has the highest accuracy. We add the accuracy for each class per attribute and use the minimum value as our metric for fairness. For completeness, we also use the mean per-class per-attribute accuracy as in [40].

4.2 Experimentation and Analysis

Experimentation was conducted on the RAF-DB [26] and CelebA [28] datasets similar to [42]. The RAF-DB dataset has 7 human-annotated classes. The model is trained on the provided train-test split consisting of 12271 train images, and inference is run on 3068 test images. Table 1 and Table 5 show that our model achieves state-of-the-art results on RAF-DB for both metrics and demonstrates significant bias mitigation.

Our methodology and setup is based on the hypothesis that protected attributes can be forgotten without information loss of other facial attributes. Ideally, a network would be able to perfectly distinguish attributes if these attributes were completely separable from the rest of the informative features of the image. However, since they are not, we hypothesize that if subsets of a dataset partitioned on the basis of the protected attribute are aligned or brought closer in a latent space, these attributes are considered to be forgotten.

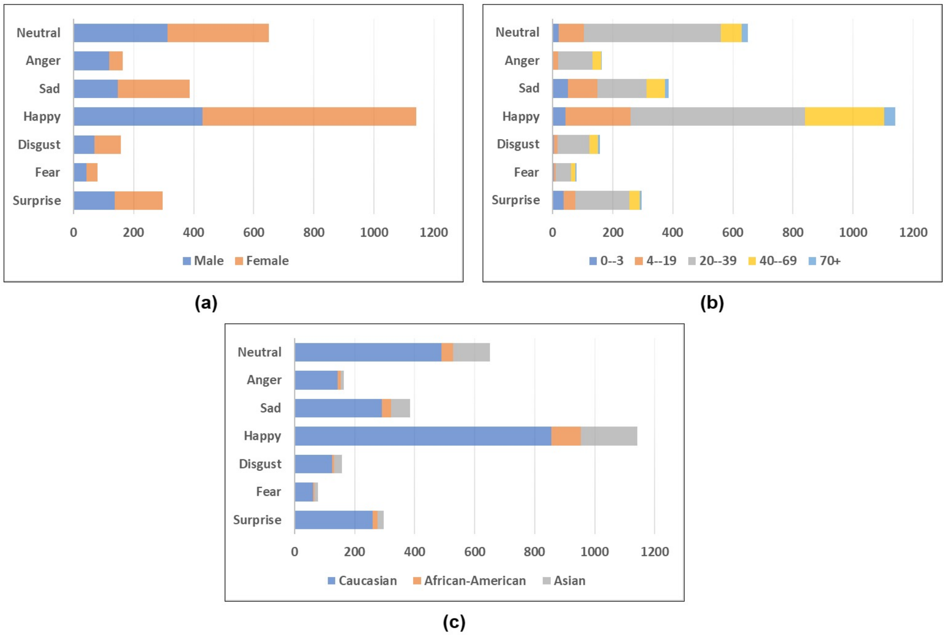
To achieve this, we use a discriminator module to classify the latents into their respective protected attributes. When this discriminator cannot determine membership of a latent into a particular protected attribute subset, then fairness can be achieved since the classification would be done solely on the basis of a latent which does not contain information about the protected attribute.

Table 1. Comparison of expression-wise accuracies on RAF-DB.

Expression	**Accuracy(%)**	
	Xu et al.	**Ours**
Anger	81.0	83.2
Disgust	54.1	57.7
Fear	53.8	60.2
Happy	93.3	92.0
Neutral	82.1	81.0
Sad	77.7	76.0
Surprise	81.8	82.9
Mean	**74.8**	**76.1**

Table 2. Mean class-wise accuracy broken down by Gender and Race attributes on RAF-DB.

Attribute Labels	**Mean Class wise Accuracies**						
	Xu et al.	**Offline[10]**	**Focal Loss[27]**	**DDC[12]**	**DIC[41]**	**SS[13]**	**Ours**
Male	74.2	72.0	71.0	71.0	72.0	72.0	**76.3**
Female	74.4	75.0	75.0	74.0	75.0	76.0	**76.0**
Caucasian	75.6	74.0	73.0	72.0	74.0	74.0	**76.15**
African-American	76.6	76.0	75.0	73.0	76.0	75.0	**77.1**
Asian	70.4	76.0	75.0	74.0	77.0	76.0	**75.5**



(a)

(b)

(c)

Fig. 3. Data Distribution of the test test of RAF-DB. (a) represents the gender-wise distribution, (b) represents the age group distribution, and (c) represents the ethnic distribution of the test set of RAF-DB.

RAF-DB Bias Analysis. Most FER datasets do not have the respective age, gender, and ethnic labels; therefore, to substantiate our results, we conducted experiments on RAF-DB [26], one of the most popular benchmark FER datasets. RAF-DB contains 15,339 images of diverse facial expressions downloaded from the internet and annotated manually by crowd-sourcing and reliable estimation; this dataset consists of seven basic expressions and eleven compound expressions.

RAF-DB provides labels that include expression, gender type, ethnicity, and age group. Fig. 3 showcases the attribute-wise breakdown of each label class in the test data. Since the distribution of test and training data is kept similar, we can draw few inferences from this distribution.

- Considering "race" as an attribute, we observe that almost 77% of the images belong to a single class i.e. Caucasian, rest, 23% are then distributed among two attributes, namely African-American and Asian.
- Similarly, for the age attribute, almost 57% of the images belong to one of the five age brackets, namely {20-39}. The rest of the 43% of images are distributed among the remaining four classes. Moreover, senior citizens from

Table 3. Mean class-wise accuracy broken down by Age and Gender-Race attributes on RAF-DB.

Attribute Labels	**Mean Class wise Accuracies**	
	Xu et al.	**Ours**
0-3	80.2	**82.4**
4-19	69.9	**72.3**
20-39	76.4	**77.0**
40-69	74.4	**75.7**
70+	62.2	**70.1**
M-Caucasian	74.5	**76.0**
M-African-American	80.2	**81.1**
M-Asian	70.2	**73.4**
F-Caucasian	75.5	**76.2**
F-African-American	87.6	**81.1**
F-Asian	69.0	**71.7**

the 70+ age bracket and infants from {0-3} age bracket are highly under-represented, consisting of about 3% and 5% of the total images, respectively.

- Observing the expression attribute, we can infer that 39.7% of the total images belong to one of the seven expression classes, i.e. happy; the rest of the six classes are then distributed among the remaining six expressions. Moreover, expressions like fear, disgust and surprise are highly under-represented, consisting of about 2.7%, 5% and 10% of the total images, respectively.

Table 4. Comparison of mitigation of bias (higher is better) on RAF-DB broken down by attribute labels.

Protected attributes	**Mitigation of Bias**						
	Xu et al.[42]	**Offline[10]**	**Focal Loss[27]**	**DDC[12]**	**DIC[41]**	**SS[13]**	**Ours**
Gender	**99.97**	95.4	96.1	96.2	95.4	95.4	99.51
Race	91.9	97.4	97.2	**97.6**	96.5	97.5	94.2
Age	82.1	-	-	-	-	-	**84.8**

This further substantiates our claim and establishes the need to mitigate bias in most FER datasets. The expression accuracy shown in Table 1 does not sufficiently portray the performance variation of classifiers across different demographics; therefore, in Table 2,3, we comprehensively compare accuracies broken down by each demographic group. Furthermore, to substantiate the inter-play of "gender" and "race" attributes we also provide results of joint "Gender-Race" groups in Table 3. From Table 2,3 it can be inferred, that our proposed method outperforms for mean class-wise accuracies broken down by attributes "age", "gender", "race" and "gender-race".To provide a numerical assessment of mitigation of bias for sensitive attributes such as age, gender, and race, in Table 4, we provide comparisons with [42, 10, 27, 12, 41, 13] using our evaluation metric for fairness (using Equation 4). From Table 4 we can infer that with regards to bias mitigation, our approach performs almost at par with Xu et al. [42] for "gender" attribute, whereas for "age" class it outperforms [42].

Table 5. Comparison of accuracy broken down by smiling attribute on CelebA dataset.

Expression	Accuracy	
	Xu et al. [42]	Ours
Smiling	92.2	92.9
Not-Smiling	94.1	94.8
Mean	93.15	93.85

Table 6. Mean class-wise accuracy broken down by attributes on CelebA.

Attribute Labels	Mean Class-wise Accuracy	
	Xu et al.[42]	Ours
Female	93.6	94.5
Male	91.9	93.4
Old	91.6	92.5
Young	93.6	94.3
Female-Old	92.7	93.3
Female-Young	93.8	94.9
Male-Old	90.7	92.1
Male-Young	92.8	93.7

CelebA Bias Analysis

We conduct experimentation for images in CelebA for the binary attribute of "smiling". This is done to facilitate the expression recognition of happy. We use the CelebA dataset since it is much larger as compared to RAF-DB with 39920 images in a subset of CelebA as compared to 12271 in all of RAF-DB. The protected attributes we use for fairness are Gender and Age.

The Smiling/No Smiling attribute is evenly distributed with exactly 50% of the images having the smiling attribute. The image distribution for Gender and Age are not evenly distributed, with a 22.8% gap between the number of Male and Female images, and a 51.4% gap between the number of Young and Old images. The comparison of accuracies with "Smiling" vs "No Smiling" is shown in Table 5. Since this is a simple binary classification task, accuracies are almost comparable. Table 6 provides comparable class-wise (i.e. "Smiling" vs "No Smiling") accuracies broken down by attribute labels ("gender", "age", and "Gender-Age"). Table 7 provides comparisons with [42] using our evaluation metric for fairness (using Equation 4) on sensitive attributes.

Table 7. Comparison of mitigation of bias (higher is better) on CelebA broken down by attribute labels.

Protected Attribute	Mitigation of Bias	
	Xu et al.[42]	Ours
Gender	98.3	99.1
Age	98.1	98.9
Gender-Age	96.9	98.0

5 Ablation Study

Table 8. Component-wise Ablation Study of our model.

Component	**Mean Accuracy**	**Bias (Gender)**	**Bias (Race)**
VAE+MBConv+Discriminator (Ours)	76.1	99.93	94.2
Auto Encoder+Discriminator+MBConv	74.2	97.6	91.2
VAE+Discriminator+ResBlock	74.5	99.91	93.8
VAE+MBConv	76	91.4	79.2
VAE+ResBlock	73	91	79.3

We demonstrate the importance and effectiveness of each technical contribution through this ablation study on RAF-DB [26]. We first look at the impact of using a Variational Autoencoder as compared to a standard Autoencoder or other dimensional reduction techniques. We can see a significant drop in accuracy and a corresponding drop in bias mitigation when an Autoencoder is used in place of a VAE. We believe this is due to the ability of VAEs to generate denser representations due to the KL-Divergence loss from the Gaussian distribution present in VAEs.

We further look at the impact of the Discriminator module on latent space alignment and examine whether it increases fairness. From Table 8, we see that there is a significant decrease in bias mitigation when the VAE is trained without the min-max objective jointly with the discriminator. This demonstrates that the Discriminator is highly impactful for latent space alignment and that the sensitive attributes are encoded in the latent without it.
We further analyze the impact of the CNN classifier backbone on accuracies. We find that the MBConv block[18] performs superior as compared to ResBlock [16]. In recent works, MBConv blocks have been known for their superior expressive power in CNNs. MBConv outperforms ResBlocks given all other parameters remain the same. However, this difference is minimal given that the largest contributor to our model is the VAE+Discriminator architecture for latent alignment.

6 Conclusion

With the exponential increase of real-world artificial intelligence systems deployed in our daily lives, accounting for fairness has become a crucial factor in the design and research of such systems. AI systems can be deployed in various critical settings to make important life-changing decisions; hence, ensuring that these decisions do not exhibit bias or discriminatory behaviour against specific groups or demographics is of utmost importance. As a result, bias mitigation investigation and its mitigating strategies have gained good traction among researchers. Recently, many works have proposed bias mitigation strategies through traditional machine learning and deep learning in various subdomains; however, this is a relatively less-explored area in facial expression recognition. In this work, we propose a new method for mitigating bias in FER systems by using a Variational Autoencoder with an Adversarial Discriminator followed by an MBConv-based classification module. We surpass the results presented [42] and provide an adaptable framework that can be extended to other image classification tasks. To the best of our knowledge, this is the first work that uses latent alignment for debiasing in FER systems. We hope that our work will pave the way for a more extensive exploration of latent space manipulation for bias reduction in a wider range of image classification scenarios.

References

1. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
2. Beigman, E., Klebanov, B.B.: Learning with annotation noise. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 280–287 (2009)
3. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)
4. Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. Advances in neural information processing systems **30** (2017)
5. Celis, L.E., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: A meta-algorithm with provable guarantees. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 319–328 (2019)
6. Chen, Y., Joo, J.: Understanding and mitigating annotation bias in facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14980–14991 (2021)
7. Chen, Y., Joo, J.: Understanding and mitigating annotation bias in facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14980–14991 (2021)
8. Chen, Y., Joo, J.: Understanding and mitigating annotation bias in facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14980–14991 (2021)

9. Chen, Z., Zhang, J.M., Sarro, F., Harman, M.: A comprehensive empirical study of bias mitigation methods for machine learning classifiers. ACM transactions on software engineering and methodology **32**(4), 1–30 (2023)
10. Churamani, N., Kara, O., Gunes, H.: Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. IEEE Transactions on Affective Computing **14**(4), 3191–3206 (2022)
11. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: A survey on an emerging challenge. IEEE Transactions on Technology and Society **1**(2), 89–103 (2020)
12. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
13. Elkan, C.: The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17, pp. 973–978. Lawrence Erlbaum Associates Ltd (2001)
14. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 259–268 (2015)
15. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hort, M., Chen, Z., Zhang, J.M., Harman, M., Sarro, F.: Bias mitigation for machine learning classifiers: A comprehensive survey. ACM Journal on Responsible Computing (2023)
18. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
20. Jia, S., Lansdall-Welfare, T., Cristianini, N.: Right for the right reason: Training agnostic networks. In: Advances in Intelligent Data Analysis XVII: 17th International Symposium, IDA 2018,'s-Hertogenbosch, The Netherlands, October 24–26, 2018, Proceedings 17. pp. 164–174. Springer (2018)
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
22. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th international conference on data mining. pp. 924–929. IEEE (2012)
23. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23. pp. 35–50. Springer (2012)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

25. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE transactions on affective computing **13**(3), 1195–1215 (2020)
26. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593. IEEE (2017)
27. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
28. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
29. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
30. Morales, A., Fierrez, J., Vera-Rodriguez, R., Tolosana, R.: Sensitivenets: Learning agnostic representations with application to face images. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(6), 2158–2164 (2020)
31. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)
32. Parraga, O., More, M.D., Oliveira, C.M., Gavenski, N.S., Kupssinskü, L.S., Medronha, A., Moura, L.V., Simões, G.S., Barros, R.C.: Fairness in deep learning: A survey on vision and language research. ACM Computing Surveys (2023)
33. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in neural information processing systems **30** (2017)
34. Rhue, L.: Racial influence on automated perceptions of emotions. Available at SSRN 3281765 (2018)
35. Rizvi, S.S.A., Seth, A., Narang, P.: Fair-fer: A latent alignment approach for mitigating bias in facial expression recognition (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 23633–23634 (2024)
36. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE international conference on computer vision. pp. 4068–4076 (2015)
37. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)
38. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 692–702 (2019)
39. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: 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. pp. 5310–5319 (2019)
40. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8919–8928 (2020)
41. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8919–8928 (2020)

42. Xu, T., White, J., Kalkan, S., Gunes, H.: Investigating bias and fairness in facial expression recognition. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 506–523. Springer (2020)
43. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning. pp. 325–333. PMLR (2013)
44. Zeng, D., Lin, Z., Yan, X., Liu, Y., Wang, F., Tang, B.: Face2exp: Combating data biases for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20291–20300 (2022)
45. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
46. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
47. Zhuang, H., Young, J.: Leveraging in-batch annotation bias for crowdsourced active learning. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 243–252 (2015)