Demographic Fairness Transformer for Bias Mitigation in Face Recognition

Ketan Kotwal[1] *and* Sébastien Marcel[1,2]
[1] Idiap Research Institute, Switzerland
[2] University of Lausanne, Switzerland
{ketan.kotwal, sebastien.marcel}@idiap.ch

Abstract

Demographic bias in deep learning-based face recognition systems has led to serious concerns. Often, the biased nature of models is attributed to severely imbalanced datasets used for training. However, several studies have shown that biased models can emerge even when trained on balanced data due to factors in the data acquisition process. Considering the impact of input data on demographic bias, we propose an image to image transformer for demographic fairness (DeFT). This transformer can be applied before the pretrained recognition CNN to selectively enhance the image representation with the goal of reducing the bias through overall recognition pipeline. The multi-head encoders of DeFT provide multiple transformation paths to the input which are then combined based on its demographic information implicitly inferred through soft-attention mechanism applied to intermittent layers of DeFT. We compute probabilistic weights for demographic information, as opposed to conventional hard labels, simplifying the learning process and enhancing the robustness of the DeFT. Our experiments demonstrate that in a cross-dataset testing (pretrained as well as locally trained models), integrating the DeFT leads to fairer models, reducing the variation in accuracies while often slightly improving average recognition accuracy over baselines.

1. Introduction

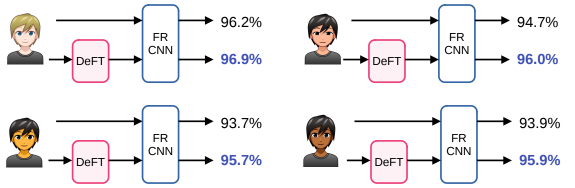


Figure 1. Illustration of demographic bias in the form variation in recognition accuracy across different demographic groups in RFW dataset [36]. In each example, the upper path and corresponding numbers refer to the baseline performance; while lower path and numbers indicate the output of the proposed DeFT.

The issue of demographic bias in Face Recognition (FR) systems has implications that extend beyond technical concerns to include social, societal, and ethical dimensions [2, 32, 27]. Certain demographic groups based on ethnicity/race or gender often experience unequal treatment from modern FR systems utilizing deep convolutional neural networks (CNNs) [25, 31]. For instance, as indicated in Fig. 1 (upper path in each example), one of the state-of-the-art FR CNN, using improved ResNet-50 [9] architecture, results in variation of around 2% in recognition accuracy across four ethnic groups (Asian, African, Caucasian, and Indian) in the RFW dataset [36]. These disparities in recognition inaccuracies can potentially put these groups at risk of negative outcomes in social, criminal, and financial contexts. In access control systems, such biased treatment can lead to undue inconvenience [8, 15].

The imbalance in training data has been regarded as one of the primary cause for demographically biased FR systems [4, 17, 12]. Most publicly available training datasets show significant imbalance in terms of demographic representation. For example, the CASIA-WebFace dataset [38] is comprised of more than 80% Caucasian subjects, while Asians and Indians together make up less than 5% [36]. Similarly, out of the $55k$ samples in the MORPH-II dataset, more than $46k$ samples are male and only around $8k$ samples are female [1]— which indicates the extent of skewness in the distribution of demographic groups in the training data.

Recent studies, however, have shown that even with demographically balanced datasets, the FR CNNs may still exhibit demographic bias [14, 17]. Therefore, achieving balance alone in the dataset does not ensure complete elimination of demographic bias. As demonstrated by several works [5, 21, 28], the process of data acquisition also significantly influences several aspects of input presentations that can contribute to bias in FR systems. Several comprehensive studies conducted by the Maryland Test Facility

(MdTF) have revealed that image acquisition plays a crucial role in shaping demographic disparities [28] and that the characteristics of skin tone in acquired presentations are significant demographic covariates [6]. Thus, not only extent of balance in data, but also their acquisition and characteristics play important role in contributing to biased outcomes of the subsequent FR CNN.

In a typical FR pipeline, the input presentations (images of subjects' faces) are preprocessed (detection, cropping, alignment, etc.) to meet the requirements of the FR CNN. These preprocessed images are then passed through the layers of deep CNN to obtain feature representations (embeddings) for matching and score generation. It is commonly known that early layers in deep CNNs are tailored to the specific data or domain, while later layers tend to be task-specific. Therefore, if demographic bias primarily stems from the data acquisition process, it would be more effective to address this issue during the initial stages or even at the preprocessing phase. If bias exists within the characteristics of input data itself, transforming its representation before subjecting it to a pretrained FR CNN could be beneficial in producing less biased outcomes at lower computational cost. Otherwise, addressing bias at higher network layers or post-processing stages where pixel-level relationships are no longer present would be necessary.

In this work, we propose an image preprocessing module- ***D**emographic **F**airness **T**ransformer* (**DeFT**) to address demographic bias in FR CNN. The DeFT consists of a multi-head attention-based module with a shared decoder, transforming input images into representations that selectively enhance their subtle features while preserving identity-related information. Our approach involves demographic-specific image enhancement before processing them through the FR CNN. While individual subject-specific enhancement would be ideal for higher accuracy, it is computationally challenging. Therefore, our model is designed to enhance input image representations to improve recognition accuracy as a function of their demographic attributes. The multi-head encoder offers diverse transformation functions for each input image, tailored to different demographic groups' characteristics. These outputs are combined using an attention-based fusion module that dynamically generates weight vectors responsive to the image content. By utilizing probabilistic weights instead of hard-label demographic classifiers, our learning process becomes more simplified and resilient against label noise. Unlike previous works [13, 11], our model eliminates the need for a separate extra demographic classifier. The relative improvement in recognition accuracy, on the RFW dataset, for FR CNN using ResNet-50 backbone has been indicated in Fig. 1.

Our contributions can be summarized as follows:

- We propose a lightweight image transformer, DeFT, that can be prepended to an existing FR system to obtain fairer outcomes (less bias) without impacting the recognition performance. The default configuration consists of 0.6M parameters.
- We propose the use of probabilistic weights to implicitly infer demographic information of the input. This mechanism based on soft attention-based fusion module underscores our system's flexibility and robustness to noisy training samples.
- The DeFT achieves state of the art performance on the RFW dataset (well-referenced benchmark) in terms of recognition accuracy and fairness to demographic groups. These results are consistent for different combinations of training datasets.

2. Related Work

The existing literature on bias mitigation in FR can be categorized into three main approaches: data-processing, in-processing, and post-processing. Given that the focus of the present work is more relevant to data-processing-based methods, we aim to provide a detailed explanation of these techniques.

In data-processing approaches, the main goal is to address bias by manipulating the training data before it is fed into the FR system [3]. This can be done through various techniques such as data augmentation, sampling strategies, and feature transfer methods.

In [18], Kortylewski *et al.* investigated the advantages of using synthetic data during the initial training phases of an FR CNN, followed by the application of real-world data for fine-tuning to mitigate biases. It should be noted that their study focuses on addressing biases related to yaw and pose of the face rather than demographic factors. Nonetheless, this approach highlights the potential of synthetic data in preparing the training process and demonstrates that a fractional amount of real-world data may suffice to mitigate bias issues compared to what would be needed for training from scratch.

Wang *et al.* introduced a new technique for augmenting features with large margins [37]. The aim of their technique is to balance demographic-class distributions in FR systems, leading to a more equitable representation of different classes. In addition to enriching the feature set, such augmentation also improves the model's ability to generalize across diverse facial representations.

To enhance the representation of under-represented groups in the feature space, Yin *et al.* proposed a feature transfer technique [39]. Their technique involves transferring features from well-represented individuals to those who are less represented, with the goal of reducing disparities observed in FR dataset distributions. This transfer process facilitates a more equitable representation and recognition accuracy across diverse demographic groups.

Wang *et al.* introduced an in-processing bias mitigation technique based on reinforcement learning to adjust margins for demographics, to obtain balanced performance across different races [35]. In this work, they also introduced the BUPT-GlobalFace and BUPT-Balancedface datasets which are often used to train the bias mitigation CNNs. A group-adaptive training strategy incorporating adaptive convolution kernels and attention mechanisms into FR CNN backbones was proposed in [13]. The work in [11] introduced an adversarial network for debiasing employing one identity classifier and three demographic classifiers (gender, age, race) to achieve unbiased FR. In [23], authors proposed a two-stage method for adversarial bias mitigation through disentangled representations and additive adversarial learning. In [20], the FR CNN has been finetuned by imposing regularization constraints based on score-calibrations for each demographic groups. Li *et al.* formulated the bias mitigation problem as a signal-denoising problem and proposed a progressive cross-transformer architecture that removes race-induced identity-unrelated components from identity-related ones [22]. Some of the recent works have advocated contrastive setup to enhance intra-class similarity and diminish similarity between negative samples [24, 40]. It may be noted that most aforementioned works, based on training or finetuning the FR CNNs, have considered positive and negative samples (same and different identities) from the same demographic groups.

Post-processing techniques for bias mitigation have received less attention compared to the other techniques. These methods focus on score calibration [30] or score normalization [26] to mitigate demographic bias. Terhöst *et al.* proposed a use of shallow network instead of conventional cosine or Euclidean distance based score computation [33]. A variety of score fusion and ensemble strategies were studied by Srinivas *et al.* to address age related bias in FR at post-processing stage [29].

3. DeFT: Transformer for Bias Mitigation

We begin with motivation towards design choices of the proposed image transformer and then describe the actual architecture and functioning of the DeFT. Following that, we explain our loss functions towards training of the DeFT.

If FR CNNs exhibit non-equitable performance for subjects from particular demographic groups, tailored transformations based on demographic attributes could potentially reduce this disparity. Thus, we develop a novel approach involving preprocessing input images to enhance subtle features essential for improving matching accuracy and thereby minimizing demographic bias. While unique transformations may be necessary for individual subjects or presentations to achieve optimal recognition, grouping based on demographics significantly simplifies the enhancement process. The concept of domain alignment (DA) through a prepended modules is a well-established technique in the field of adaptation problems [41, 42]. A similar transformation of presentations acquired from different modalities has been employed for heterogeneous face recognition [10]. These modules primarily aim to align the subspaces of target domain input data with those of the training set (source domain), achieved by minimizing Kullback-Leibler (KL) or Maximum Mean Discrepancy (MMD) loss. However, it is important to note that our objective in addressing bias in FR differs distinctly from DA methods: aligning distributions within the same demographic group may inadvertently reduce inter-subject distances in feature space whereas our goal is enhancing demographic-specific features to improve recognition accuracy without compromising subject-specific discriminatory information.

For mitigation approaches based on demographic-specific processing, accurate knowledge of the demographic label of the data by implicit or explicit means is critical. However, learning the corresponding classifier with noisy training labels presents a complex challenge. Additionally, certain demographic variables such as race or skin color often tend to be non-discrete. To address these complexities, we calculate a probabilistic weight vector for demographic information obtained from the intermediate layers of the DeFT. Taking inspiration from multi-head attention mechanisms that are prominent in transformer architectures, we develop a dynamic, data-dependent module to generate a probabilistic representation of demographic group membership—which reduces the necessity for near-perfect accuracy in demographic classification.

3.1. Architecture of DeFT

Figure 2a shows the architecture of the DeFT comprising of an encoder module, fusion module, and a decoder module. Let the DeFT be prepended to a pre-trained FR CNN, $\mathcal{F}$ which accepts an RGB input image, $\mathbf{x} \in \mathbb{R}^{3hw}$, where h and w denote the height and width of the image, respectively. We assume that $\mathbf{x}$ has undergone necessary preprocessing to align with the requirements of $\mathcal{F}$. The DeFT also accepts the same input image $\mathbf{x}$.

Encoder Module: The multi-head encoder provides several transformation paths to the input through a sequence of `conv-ReLU-conv-ReLU-conv` operations. The convolutional operations use a kernel size of 3 and a stride of 1 to maintain the spatial dimensions ($h \times w$) of the intermediate feature representations. For a dataset containing D demographic groups, we structure the DeFT with K heads (encoder blocks) such that $K \geq D$. This design has three advantages: (a) It avoids enforcing 1:1 correspondence between encoder paths and demographic groups, facilitating diverse configuration possibilities and enabling ablation studies. (b) The final transformation for any demo-

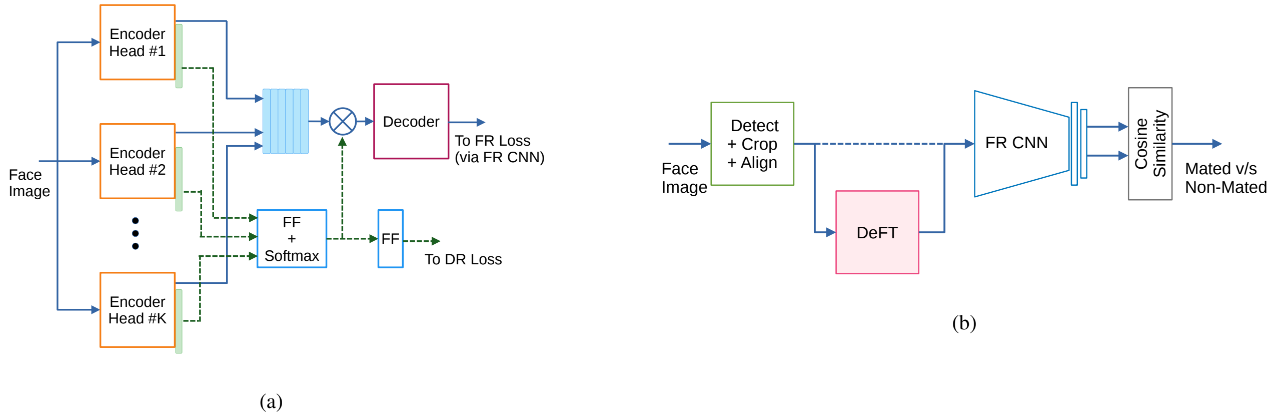


Figure 2. The architecture of the proposed DeFT with encoder, fusion, and decoder modules is shown in (a). The schematic of overall FR pipeline inclusive of the DeFT is provided in (b).

graphic group is achieved as a linear combination of individual block transformations, simplifying the training process. (c) This structure reduces the learning requirements for the demographic classifier by utilizing an over-complete decomposition *dictionary* framework underlying features.

Fusion Module: It employs a channel-wise attention mechanism to combine the outputs of encoder blocks. If each block (output of encoder) consists of J channels or feature maps, for the j-th channel in the k-th encoder block, the feature map $\mathbf{C}_{jk}$ is assigned a weight z_{jk} through channel attention (*ChAtt*), formulated as,

$$z_{jk} = \mathit{ChAtt}(\mathbf{C}_{jk}), \quad j = 1 \text{ to } J;\ k = 1 \text{ to } K. \tag{1}$$

A weight matrix $\mathbf{Z} \in \mathbb{R}^{J \times K}$ is generated and processed by a shallow feed-forward (FF) network followed by a softmax operation across the dimensions representing the number of heads (K). If $\mathbf{Z}'$ represents the output of FF network, this softmax operation can be described by Eq. 2.

$$z''_{jk} = \frac{\exp\left(z'_{jk}\right)}{\sum_{k=1}^{K} \exp\left(z'_{jk}\right)}, \qquad \text{for } j = 1, 2, \dots, J, \tag{2}$$

The normalized weights $\mathbf{Z}''$ are then used to fuse the outputs of K encoders into a single feature map as given by Eq 3.

$$\mathbf{C}'_{j} = \sum_{k=1}^{K} \mathbf{C}_{jk} \cdot z''_{jk}, \qquad \text{for } j = 1, 2, \dots, J. \tag{3}$$

Additionally, we obtain an auxiliary output from the fusion module to facilitate learning of probabilistic weights indicating demographic information using a fully connected (fc) layer on individual channels of normalized fusion weights $\mathbf{Z}''$ to output a D-dimensional output. This layer has K and D as the input and output dimensions respectively. It may be noted that the weight vector is not directly tied to the labels; rather, it serves as a linear combination mechanism for encoder outputs. The auxiliary output ($\in \mathbb{R}^D$) is used to compute the demographic loss as explained in the next Section.

Decoder Module: The architecture of the decoder is same as that of an individual encoder head, except that it produces an output with 3 channels (same as input). This output is subsequently processed by $\mathcal{F}$ to yield an *embedding* that is expected to enhance recognition accuracy and mitigate demographic biases.

It is important to clarify that while image enhancement typically aims for visually pleasing results or noise reduction, our goal is to enhance the representation in a manner that enables the subsequent pre-trained and frozen FR CNN to extract higher quality features across layers. This ultimately enhances recognition performance for the given input.

3.2. Loss Functions

We consider ArcFace, an angular margin-based variant of of cross entropy loss [7], that operates on embedding (output of $\mathcal{F}$) and subject's identity label. This loss, L_{FR}, ensures that the recognition accuracy is improved: Initially, an ArcFace classifier head is appended to the FR CNN and trained to align it to the pre-trained FR CNN parameters. Post alignment, the cross-entropy loss is computed using Eq. 4 and backpropagated to the DeFT via frozen layers of the FR CNN.

$$L_{\text{FR}} = -\sum_{i=1}^{N} y_i \log\left(\frac{\exp\left(\mathbf{f}_i\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{f}_j\right)}\right), \tag{4}$$

where $\mathbf{f}$ and y represent the normalized embedding and subject identity (label) respectively. N is the batch size.

We use a supervisory loss function to learn the soft-weights (related to demographic information) from the auxiliary output of the fusion module. We consider cross-entropy loss with label smoothing, serving as a regularization strategy that enhances tolerance towards errors in demographic labels of training data. Label smoothing prevents the DeFT from being overly confident about its predictions- which is particularly useful when working with noisy or uncertain demographic labels. Eq. 5 provides formula to compute the aforementioned demographic recognition loss, L_{DR}.

$$L_{\text{DR}} = -\sum_{i=1}^{N} \tilde{d}_i \log\left(\frac{\exp\left(\mathbf{zz}''_i\right)}{\sum_{j=1}^{N}\exp\left(\mathbf{zz}''_j\right)}\right), \tag{5}$$

where $\mathbf{zz}''$ denotes the auxiliary output of the fusion module, and $\tilde{d}$ is the smoothed demographic label.

To minimize the difference between the scores of constituent demographic groups, we first compute cosine similarity between the mated pairs of each demographic in the training batch. Following which we obtain the average intra-group score $\bar{s_d}$ as follows:

$$\bar{s_d} = \frac{1}{N_d}\sum_{\#\text{pairs}} 1 - \cos(\mathbf{f}_i,\, \mathbf{f}_j), \tag{6}$$

where $\mathbf{f}_i, \mathbf{f}_j$ represent the feature vectors of the same subject (*i.e.*, mated pair) belonging to demographic group d, and N_d is the number of such pairs in the given batch. Inspired from loss functions in [13], we design our bias loss function (L_{bias}) to minimize the absolute variation in average intra-group scores of constituent demographic groups. Our loss term, however, differs from [13] in terms of distance function as well as formulation of intra-group scores. Eq. 7 describes the bias loss function.

$$L_{\text{bias}} = \sum_{d=1}^{D}\left|\bar{s_d} - \frac{1}{D}\sum_{g=1}^{D}\bar{s_g}\right| \tag{7}$$

The overall loss L_{tot} to train the DeFT is a weighted combination of the FR, DR, and bias losses, where λ_{w1} and λ_{w2} are the relative weights of the DR and bias loss, respectively. The overall loss is given as:

$$L_{\text{tot}} = L_{\text{FR}} + \lambda_{w1}L_{\text{DR}} + \lambda_{w2}L_{\text{bias}} \tag{8}$$

By optimizing L_{tot}, it ensures that the DeFT learns to generate embeddings that are not only discriminative for recognition purposes but also sensitive to the nuances of demographic diversity, thereby helping in reducing in bias within the FR CNN.

4. Experimental Results

We first present details related to the experimental setup and then discuss the results of the proposed DeFT-based enhancement towards mitigating demographic bias in FR. We conduct three different sets of experiments with specific goals: In the first set of experiments, we utilize the same dataset for training both FR CNN and subsequently the DeFT, followed by a comparison of their effectiveness in terms of recognition accuracy and bias mitigation. The objective is to investigate whether training the DeFT with same training data as that of the FR CNN allows improvement in its fairness. If access to the original training data is no longer possible, we explore whether it is feasible to use a different dataset to train only DeFT (while freezing FR CNN) and achieve enhanced and equitable performance from this combined model (DeFT + backbone). Our second set of experiments was conducted to evaluate this practical scenario where we essentially test our proposed method on pre-trained models. In the final set of experiments, we conduct ablation study to understand how number of encoder heads in DeFT and relative weight factors (from Eq. 8) impact the performance of overall FR pipeline.

4.1. Experimental Setup

FR CNN Backbones: The ResNet architectures, which incorporate adapted residual blocks, have demonstrated state-of-the-art performance in FR [9]. For evaluation of the proposed DeFT, we utilized the FR CNNs based on the ResNet architecture with variations of 34, 50, and 100 layers.

FR Pipeline:[1] The FR pipeline remains consistent regardless of the dataset and FR backbone used. Prior to training and testing, each input image underwent a standard preprocessing procedure. This involved initial face detection and landmark identification using MTCNN, followed by aligning and resizing the face region to 112×112 (RGB) to meet the specified requirements of each ResNet-based FR CNN architecture. This preprocessed input requirement is consistent across all experiments described in this work. The aligned fixed-size input images were then passed to either the FR CNN or DeFT; the FR CNN produced a 512-d feature vector or *embeddings* which were then matched using cosine similarity.

To train the DeFT (or the FR CNN for selected experiments), we used an SGD-based optimizer with an initial learning rate of 1e-2 and momentum of 0.9. A multistep rate scheduler was implemented to decrease the learning rate by a factor of 0.2 after every 30–40 epochs. We also implemented an early stopping criteria based on training loss with a patience of 5. For each backbone architecture (ResNet-34, ResNet-50, and Resnet-100), batch size was set at 64 during SGD-based optimization and ten positive samples and ten negative samples per subject were predetermined for contrastive setup required by loss functions.

For training the FR CNN from scratch, we used up to

[1] The source code of DeFT experiments described in this work is available at https://gitlab.idiap.ch/bob/bob.paper.deft_ijcb2024.

80 epochs, while the DeFT was trained for up to another 80–100 epochs. To provide initial warm-up, the DeFT is trained with higher learning rate (10–20×) by freezing the Arcface-head for initial 20 epochs. Fig. 3 summarizes our training process across epochs.

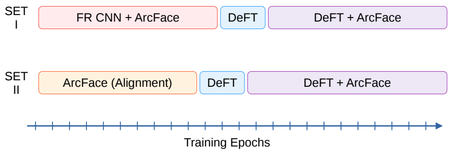


Figure 3. Timeline of training procedure for Set-I and Set-II of our experiments outlined in this Section.

For the second set of experiments (using pretrained FR CNNs), we obtained the weights for our models from the InsightFace repository[2]. These models were trained on MS1MV2 dataset with ArcFace loss.

Datasets: We used the BUPT-BalancedFace dataset [34] for training. This dataset categorizes individual subjects into four ethnicity labels: African, Asian, Caucasian, and Indian. For evaluation, we employed the RFW dataset [36], which has become a de-facto benchmark in the academic community for evaluation for demographic bias in FR. This dataset provides a well-balanced protocol for four demographic groups almost 6000 comparisons per group—resulting in a total of 24k comparisons[3].

Performance Evaluation: To assess the effectiveness of the proposed approach towards mitigating demographic bias, we follow the methods adopted in [11, 13, 22]. These methods utilize average recognition accuracy and standard deviation in accuracy across demographic groups as performance metrics. A higher average accuracy with reduced standard deviation indicates a more equitable FR system with respect to demographic fairness. It also ensures that bringing fairness has not led to severe degradation in the recognition accuracy. Additionally, we compute the Skewed Error Ratio (**SER**), which represents the ratio of the highest error rate to the lowest error rate among all demographic groups [35]. We calculate these measures for baseline models (*i.e.*, FR CNNs without any specific preprocessing for demographics) and compare them with results obtained after applying our enhanced preprocessing technique using DeFT. We also compare the performance of our method against various SOTA methods namely, GAC [13], MTL [22], PCT [22], and ScoreReg [20]. We omit the description of other methods due to brevity of space.

4.2. Results: Set I

In the first set of experiments, we utilized the BUPT-Balancedface dataset to train the FR CNN backbones from scratch and obtained the corresponding baselines. Subsequently, we employed the same BUPT-Balancedface dataset to train the DeFT prepended to the frozen FR CNN and evaluated its effectiveness in mitigating demographic bias using the RFW dataset. This experimental formulation aligns with previous studies in this field [22, 11, 13], facilitating the use of results reported by the respective publications.

Method	African	Asian	Caucasian	Indian	Avg (↑)	STD (↓)	SER (↓)
Baseline	93.15	92.85	96.13	93.03	93.78	1.36	1.84
GAC [13]	94.12	94.10	96.02	94.22	94.62	0.81	**1.48**
MTL [22]	94.82	**94.47**	**96.60**	95.23	95.28	0.93	1.62
ScoreReg [20]	94.64	94.43	96.41	95.05	95.13	0.77	1.55
DeFT (ours)	**95.08**	94.38	96.31	**95.56**	**95.33**	**0.71**	1.52

Method	African	Asian	Caucasian	Indian	Avg (↑)	STD (↓)	SER (↓)
Baseline	93.98	93.72	96.18	94.67	94.64	1.11	1.64
GAC [13]	94.77	94.87	96.20	94.98	95.21	**0.59**	**1.37**
MTL [22]	**96.05**	95.25	**97.20**	**96.05**	96.13	0.70	1.69
ScoreReg [20]	95.42	95.31	96.92	95.57	95.80	0.65	1.52
DeFT (ours)	95.90	**95.73**	97.14	95.75	**96.13**	**0.59**	1.49

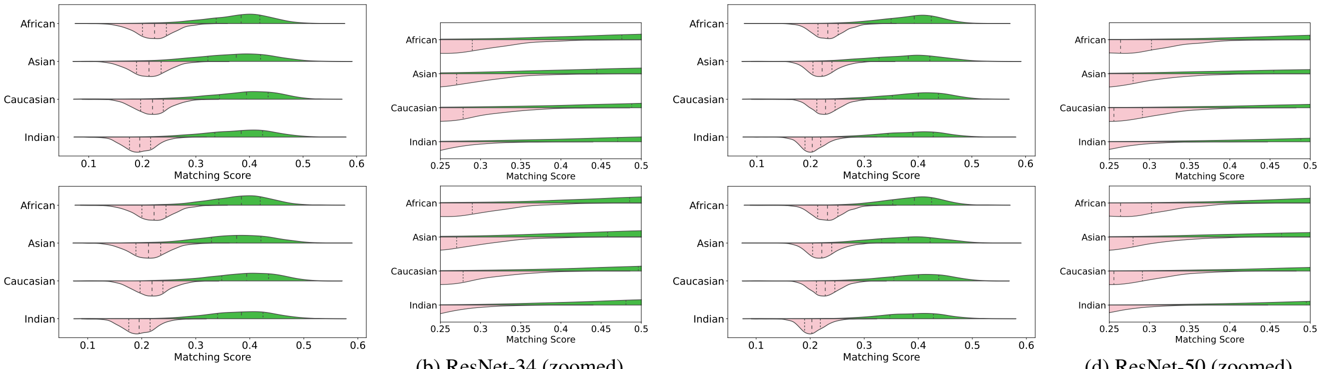
Table 1. Performance evaluation of the proposed method (train: BUPT-BalancedFace, test: RFW dataset). top: ResNet-34 and bottom: ResNet-50 FR backbones. All accuracy values are indicated as percentages.

Table 1 shows the results of our experiments using the BUPT-Balancedface dataset to train the FR CNNs and DeFT. Following application of our enhanced preprocessing technique using DeFT, the overall recognition accuracy for the 34- as well as 50-layered FR CNN backbones increased by 1.55% and 1.48% respectively, over baselines. This increment may possibly be attributed to additional parameters being added to the overall model. However, along with this improvement in the FR accuracy, we also observed significant reduction in the standard deviation among constituent demographic groups for each of the FR backbone—which is the primary objective of DeFT's incorporation. For instance, when input presentations were transformed by DeFT, the standard deviation (STD) in recognition accuracy for ResNet-34 backbone decreased from 1.36 to 0.70, while for ResNet-50, this value reduced from 1.11 to 0.47- which amounts to more than 50% reduction from baseline. Similar trends were observed regarding SER where the ratio of extreme errors reduced by approximately 0.30 for each FR CNN backbone. While our method did not necessarily outperform every compared work in the accuracy; it can be ranked first when a combined impact of accuracy and its STD are taken into account.

Fig. 4 shows the comparison of distributions of matching scores for the baseline FR CNNs and the models with enhanced preprocessing using DeFT. While the performance

[2] https://github.com/deepinsight/insightface

[3] We removed identities that were incorrectly labeled as suggested in bob.bio.face



(a) ResNet-34

(b) ResNet-34 (zoomed)

(c) ResNet-50

(d) ResNet-50 (zoomed)

Figure 4. The violin plots, along with their zoomed versions, representing distribution of mated (in green) and non-mated (in red) scores for each demographic group of the RFW dataset for different FR CNNs. Top row shows baselines and bottom row shows the results of our work. The BUPT-Balancedface dataset was used to train both DeFT and FR CNN.

evaluation uses accuracy metrics, it refers to binarized decisions (match or no-match), whereas evaluation based on actual scores, independent of classification threshold, shed light on different perspective– also known as differential performance [16, 19]. The areas of overlap of mated and non-mated distributions correspond to the sample pairs that are likely to be incorrectly classified depending on the score threshold to be applied. It can be seen from the comparison of baseline and resultant score distributions that the use of DeFT has slightly reduced this overlap, especially at the tails of distributions (tapered edges in the zoomed versions). These improvements are visually subtle as the overall improvement often results in 1–2% increase in the accuracy. Fairness of the FR system can be understood by observing the alignment of inflection points (where both distributions tend to crossover) across demographics.

4.3. Results: Set II

In our second series of experiments, we utilized three different datasets: the MS1MV2 for training the FR CNN, BUPT-Balancedface for training the DeFT, and RFW dataset for evaluating performance. However, in these experiments, instead of training the FR CNN by ourselves, we employed pre-trained models (weights) from the Insightface repository. The results of these experiments are presented in Table 2. The use of DeFT resulted in measurable improvement in recognition accuracy across all FR CNN backbones (34, 50, and 100 layers), with an increase from 0.4–1.6% compared to baselines trained using MS1MV2. Additionally, our enhancement reduced standard deviation among recognition accuracies within individual demographic groups by 0.37, 0.28, and 0.13 for backbones with 34, 50, and 100 layers respectively. We observed that the relative improvement was highest for the smallest FR backbone and vice-versa. This could possibly be due to near-saturated performance of deep FR CNNs. The SER which indicates highest ratio of error rates between different demographic groups, also showed a nominal decrease of 0.06 in two out of three experiments. The violin plots in Figure 5 provides a visual representation of the improvements in recognition accuracy indicated by lesser extent of overlap in the scores via slightly tapered tails of distributions; and fairness indicated by a better alignment of scores distributions across demographic groups.

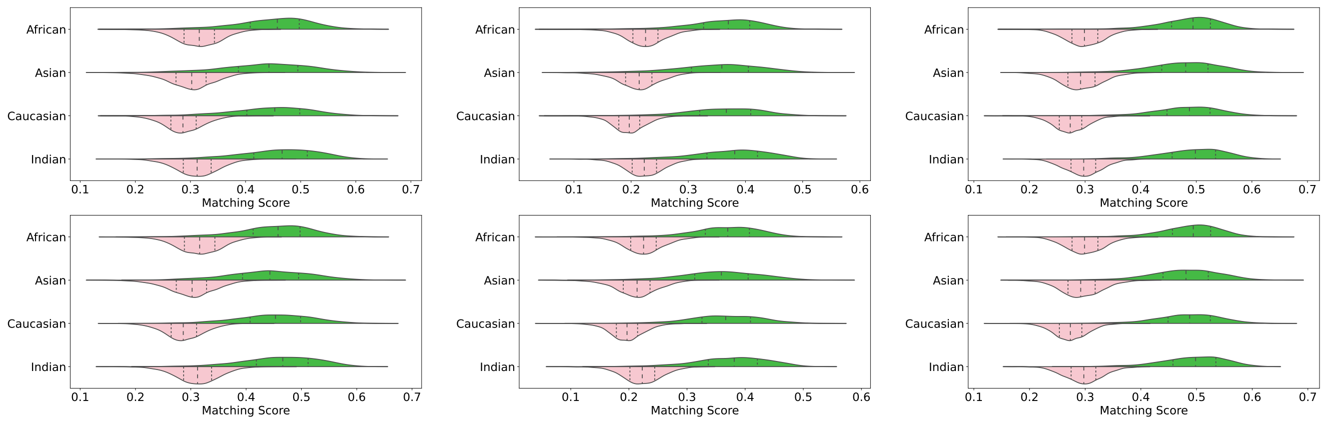
Backbone	Method	African	Asian	Caucasian	Indian	Avg (↑)	STD (↓)	SER (↓)
34	Baseline	88.47	87.19	91.90	90.39	89.49	1.79	1.58
34	Proposed	89.94	89.43	93.06	91.42	90.96	1.42	1.52
50	Baseline	91.27	89.06	94.40	91.85	91.64	1.90	1.95
50	Proposed	92.57	91.61	95.96	93.42	93.39	1.62	2.07
100	Baseline	96.83	95.53	98.10	96.93	96.84	0.91	2.35
100	Proposed	97.26	96.09	98.30	97.26	97.22	0.78	2.30

Table 2. Performance evaluation of the proposed method (train FR CNN: MS1MV2 (pretrained), train DeFT: BUPT-BalancedFace, test: RFW dataset). top: ResNet-34 and bottom: ResNet-50 FR backbones. All accuracy values are indicated as percentages.

4.4. Results: Set III

Finally, we conducted ablation studies to investigate the efficacy of design choices of the DeFT. The fusion module in the DeFT decouples the requirement of matching the number of encoder blocks with the demographic groups in the data. We leverage this flexibility to study impact of number of blocks (or width) of the encoder towards improving fairness of the FR CNN. Additionally we also study the impact of relative weights (by varying λ_w) while training the DeFT. For the ablation, we report results on the FR CNN with ResNet-34 backbone as it has relatively high margins.

Table 3 provides summary of our experiments for 2 variations in encoder design and 4 values of relative weights to the loss term. Our experiments indicate that when higher weightage is assigned to bias loss, the output of FR CNN results in better fairness, however at the cost of slightly reduced recognition accuracy. However, we did not observe



(a) ResNet-34

(b) ResNet-50

(c) ResNet-100

Figure 5. The violin plots representing distribution of mated (in green) and non-mated (in red) scores for each demographic group of the RFW dataset for different FR CNNs. Top row shows baselines and bottom row shows the results of our work. The BUPT-Balancedface dataset was used to the DeFT and the FR CNN has been pretrained on the MS1MV2 dataset.

any clear pattern between the recognition accuracy (of individual groups and overall) and ablation parameters.

5. Conclusions

In this work, we developed an image-to-image transformer aimed at selectively enhancing the representation of face images such that the demographic fairness of the subsequent FR CNN improves without compromising its recognition accuracy. This lightweight transformer, called DeFT, consists of a multi head encoder that offers multiple transformation paths for input images. The encoder outputs are fused using a dynamically weighted combination determined by an attention mechanism. This approach necessitates knowledge of demographic information—possibly in implicit manner and with minimal computational overheads. Addressing challenges such as noisy training data, learning precise demographic attributes and their non-discrete nature, we propose the use of probabilistic weights for demographic attributes instead of conventional hard labels. This approach not only mitigates the aforementioned challenges but also simplifies the image transformation/enhancement process and decouples the DeFT architecture from the number of demographic groups present in the data.

Encoder	λ_w	African	Asian	Caucasian	Indian	Avg (↑)	STD (↓)	SER (↓)
8	0.01	95.13	94.20	97.31	95.56	95.55	1.13	2.16
8	0.1	95.15	94.15	97.25	95.55	95.52	1.08	2.12
8	1.0	95.08	94.38	96.31	95.56	96.13	0.70	1.52
8	10.0	94.80	94.10	96.26	95.23	95.09	0.78	1.58
12	0.01	94.91	94.10	97.16	95.55	95.43	1.12	2.08
12	0.1	94.86	94.06	97.25	95.53	95.43	1.17	2.15
12	1.0	95.05	94.20	97.23	95.43	95.47	1.10	2.09
12	10.0	94.68	93.83	96.21	95.16	94.97	0.86	1.62

Table 3. Ablation of encoder width (number of heads) and relative weights in loss for the DeFT for ResNet-34 FR CNN backbone. Relative weights $\lambda_w = \lambda_{w1} = \lambda_{w2}$ from Eq. 8. (train FR CNN: MS1MV2 (pretrained), train DeFT: BUPT-BalancedFace, test: RFW dataset). All accuracy values are indicated as percentages.

Through experiments conducted on pretrained and locally trained FR CNNs, we have demonstrated the effectiveness of the DeFT in reducing the non-equitable treatment to various ethnic groups. Our results indicate a reduction of up to 50% in the standard deviation among recognition accuracies of individual demographic groups (ethnicities) without compromising the average recognition accuracy. These results are consistent across different combinations of training datasets and FR CNN backbones, suggesting that the proposed transformer functions as an independent module without requiring the same training data as that of FR backbones.

Given that this is initial work on using probabilistic weights for demographic information to achieve fairer outcomes, our experiments on designing the weighting strategy have been limited. The primary objective was to demonstrate the effectiveness of employing soft attention without explicit demographic labels (hard attention). Building on this success, future research could explore alternative weighting or decomposition strategies to further improve fairness in FR systems.

Acknowledgement

Authors would like to thank the Hasler foundation for their support through the SAFER project. This work has also been supported by the Swiss Center for Biometrics Research and Testing.

References

[1] G. Bingham, B. Yip, M. Ferguson, and C. Nansalo. MORPH-II: Inconsistencies and Cleaning. *University of North Carolina Wilmington NSF REU*, 2017. 1

[2] D. Castelvecchi. Is facial recognition too biased to be let loose? *Nature*, 587(7834):347–350, 2020. 1

[3] S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020. 2

[4] V. Cherepanova, S. Reich, S. Dooley, H. Souri, J. Dickerson, M. Goldblum, and T. Goldstein. A deep dive into dataset imbalance and bias in face identification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 229–247, 2023. 1

[5] C. Cook, J. Howard, Y. B. Sirotin, J. Tipton, and A. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019. 1

[6] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Demographic effects across 158 facial recognition systems, 2023. 2

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 4

[8] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020. 1

[9] I. C. Duta, L. Liu, F. Zhu, and L. Shao. Improved residual networks for image and video recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 9415–9422. IEEE, 2021. 1, 5

[10] A. George, A. Mohammadi, and S. Marcel. Prepended domain transformer: Heterogeneous face recognition without bells and whistles. *IEEE Transactions on Information Forensics and Security*, 18:133–146, 2022. 3

[11] S. Gong, X. Liu, and A. K. Jain. Debface: De-biasing face recognition. *arXiv preprint arXiv:1911.08080*, 2019. 2, 3, 6

[12] S. Gong, X. Liu, and A. K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 330–347. Springer, 2020. 1

[13] S. Gong, X. Liu, and A. K. Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3414–3424, 2021. 2, 3, 5, 6

[14] M. Gwilliam, S. Hegde, L. Tinubu, and A. Hanson. Rethinking common assumptions to mitigate racial bias in face recognition datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4123–4132, 2021. 1

[15] N. Hallowell, L. Amoore, S. Caney, and P. Waggett. Ethical issues arising from the police use of live facial recognition technology. *Interim Report of the Biometrics and Forensics Ethics Group Facial Recognition Working Group, Rep*, 2019. 1

[16] J. J. Howard, Y. B. Sirotin, and A. Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *Proceedings of the International Conference on Biometrics Theory, Applications and Systems*, pages 1–8, 2019. 7

[17] M. Kolla and A. Savadamuthu. The impact of racial distribution in training data on face recognition bias: A closer look. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 313–322, 2023. 1

[18] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[19] K. Kotwal and S. Marcel. Fairness Index Measures to Evaluate Bias in Biometric Recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 479–493. Springer, 2022. 7

[20] K. Kotwal and S. Marcel. Mitigating demographic bias in face recognition via regularized score calibration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1150–1159, 2024. 3, 6

[21] K. Krishnapriya, V. Albiero, K. Vangara, M. King, and K. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020. 1

[22] Y. Li, Y. Sun, Z. Cui, S. Shan, and J. Yang. Learning fair face representation with progressive cross transformer. *arXiv preprint arXiv:2108.04983*, 2021. 3, 6

[23] J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu. Additive adversarial learning for unbiased authentication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11428–11437, 2019. 3

[24] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022. 3

[25] C. Rathgeb, P. Drozdowski, D. Frings, N. Damer, and C. Busch. Demographic fairness in biometric systems: What do the experts say? *IEEE Technology and Society Magazine*, 41(4):71–82, 2022. 1

[26] T. Salvador, S. Cairns, V. Voleti, N. Marshall, and A. Oberman. Bias mitigation of face recognition models through calibration. *arXiv preprint arXiv:2106.03761*, 2021. 3

[27] R. Singh, P. Majumdar, S. Mittal, and M. Vatsa. Anatomizing bias in facial analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12351–12358, 2022. 1

[28] Y. Sirotin and A. Vemury. Demographic variation in the performance of biometric systems: Insights gained from large-scale scenario testing. *Virtual Events Series–Demographic fairness in biometric systems. EAB*, 2021. 1, 2

[29] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King. Face recognition algorithm bias: Performance differences on images of children and adults. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 0–0, 2019. 3

[30] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, 2020. 3

[31] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021. 1

[32] P. Terhörst, K. Riehl, N. Damer, P. Rot, B. Bortolato, F. Kirchbuchner, V. Struc, and A. Kuijper. Pe-miu: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access*, 8:93635–93647, 2020. 1

[33] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *Proceedings of the International Workshop on Biometrics and Forensics*, pages 1–6. IEEE, 2020. 3

[34] M. Wang and W. Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*, 2019. 6

[35] M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020. 3, 6

[36] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 692–702, 2019. 1, 6

[37] P. Wang, F. Su, Z. Zhao, Y. Guo, Y. Zhao, and B. Zhuang. Deep class-skewed learning for face recognition. *Neurocomputing*, 363:35–45, 2019. 2

[38] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1

[39] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019. 2

[40] F. Zhang, K. Kuang, L. Chen, Y. Liu, C. Wu, and J. Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *Proceedings of the International Conference on Learning Representations*, 2022. 3

[41] L. Zhang and X. Gao. Transfer adaptation learning: A decade survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3

[42] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3