Abstract

Face recognition is known to exhibit bias - subjects in a certain demographic group can be more equally represented. Our proposed group adaptive classifier mitigates bias by using adaptive convolution kernels and attention mechanisms on faces based on their demographic attributes. The adaptive module comprises kernel masks and channel-wise attention maps for each demographic group so as to activate different facial regions for identification, leading to more discriminative features pertinent to their demographics. Our introduced automated adaptation strategy determines whether to apply adaptation to a certain layer by iteratively computing the dissimilarity among demographic-adaptive parameters. A new de-biasing loss function is proposed to mitigate the gap of average intra-class distance between demographic groups. Experiments on face benchmarks (RFW, LFW, IJB-A, and IJB-C) show that our work is able to mitigate face recognition bias across demographic groups while maintaining the competitive accuracy.

1. Introduction

Face recognition (FR) systems are known to exhibit discriminatory behaviors against certain demographic groups [21, 28, 40]. The 2019 NIST Face Recognition Vendor Test [21] shows that all 106 tested FR algorithms exhibit varying biased performances on gender, race, and age groups of a mugshot dataset. Deploying biased FR systems to law enforcement is potentially unethical [11]. Given the implication of automated FR-driven decisions, it is crucial to develop fair and unbiased FR systems to avoid the negative societal impact. Note that, differ from the inductive bias in machine learning [15], we define FR bias as the uneven recognition performance w.r.t. demographic groups.

State-of-the-art (SOTA) FR algorithms [13, 33, 34, 48, 69, 70] rely on convolutional neural networks (CNNs) trained on large-scale face datasets. Biases in data are transmitted to the FR models through network learning. For example, to minimize the overall loss, a network tends to learn a better representation for faces in the majority group whose data dominate the training set, resulting in unequal discriminabilities. However, the imbalanced demographic distribution of face data is not the only trigger of FR bias. Prior works have shown that even using a demographic balanced dataset [71] or training separate classifiers for each group [40], the performance on some groups is still inferior to the others. By studying non-trainable FR algorithms, [40] introduced the notion of inherent bias, i.e., certain groups are inherently more susceptible to errors in face matching.

To tackle the dataset-induced bias, traditional methods re-weight either the data proportions [6] or cost values [1]. Such methods are limited when applied to large-scale imbalanced datasets. Recent imbalance learning methods focus on novel objective functions for class-skewed datasets. For instance, Dong et al. [18] propose a Class Rectification Loss to incrementally optimize on hard samples of the classes with under-represented attributes. Alternatively, researchers strengthen the decision boundary to impede perturbation from other classes by enforcing margins between hard clusters via adaptive clustering [31], or between rare classes via Bayesian uncertainty estimates [38]. To adapt the aforementioned methods to racial bias mitigation, Wang et al. [71]
modify the large margin based loss functions by reinforcement learning. However, [71] requires two auxiliary networks, an offline sampling network and a deep Q-learning network, to generate adaptive margin policy for training the FR network, which hinders the learning efficiency.

To mitigate FR bias, our main idea is to optimize the face representation learning on every demographic group in a single network, despite demographically imbalanced training data. Conceptually, we may categorize face features into two types of patterns: general pattern is shared by all faces; differential pattern is relevant to demographic attributes. When the differential pattern of one specific demographic group dominates training data, the network learns to predict identities mainly based on that pattern as it is more convenient to minimize the loss than using other patterns, leading to bias towards that specific group. One mitigation is to give the network more capacity to broaden its scope for multiple face patterns from different groups. An unbiased FR model shall rely on both unique patterns for recognition of different groups, and general patterns of all faces for improved generalizability. Accordingly, as in Fig. 1, we propose a group adaptive classifier (GAC) to explicitly learn these different feature patterns. GAC includes two modules: the adaptive layer and automation module. The adaptive layer comprises adaptive convolution kernels and channel-wise attention maps where each kernel and map tackle faces in one demographic group. We also introduce a new objective function to GAC, which diminishes the variation of average intra-class distance between demographic groups.

Prior work on dynamic CNNs introduce adaptive convolutions to either every layer [36,73,79], or manually specified layers [27,51,67]. In contrast, we propose an automation module to choose which layers to apply adaptations. As we observed, not all convolutional layers require adaptive kernels for bias mitigation (see Fig. 4a). At any layer of GAC, only kernels expressing high dissimilarity are considered as demographic-adaptive kernels. For those with low dissimilarity, their average kernel is shared by all inputs in that layer. Thus, the proposed network progressively learns to select the optimal structure for the demographic-adaptive learning. Both non-adaptive layers with shared kernels and adaptive layers are jointly learned in a unified network.

The contributions of this work include:

- A new face recognition algorithm that reduces demographic bias and tailors representations for faces in every demographic group by adopting adaptive convolutions and attention techniques;
- A new adaptation mechanism that automatically determines the layers to employ dynamic kernels and attention maps;
- The proposed method achieves SOTA performance on a demographic-balanced dataset and three benchmarks.

2. Related Work

Fairness Learning and De-biasing Algorithms. A variety of fairness techniques are proposed to prevent machine learning models from utilizing statistical bias in training data, including adversarial training [2,26,52,74], subgroup constraint optimization [37,75,86], data pre-processing (e.g., weighted sampling [22], and data transformation [4]), and algorithm post-processing [39,58]. Another promising approach learns a fair representation to preserve all discerning information about the data attributes or task-related attributes but eliminate the prejudicial effects from sensitive factors [11,24,55,65,82]. Locatello et al. [50] show the feature disentanglement is consistently correlated with increasing fairness of general purpose representations by analyzing 12,600 SOTA models. Accordingly, a disentangled representation is learned to de-bias both FR and demographic attribute estimation [20]. Other studies address the bias issue in FR by leveraging unlabeled faces to improve the performance in minority groups [59,72]. Wang et al. [71] propose skewness-aware reinforcement learning to mitigate racial bias. Unlike prior work, our GAC is designed to customize the classifier for each demographic group, which, if successful, would lead to mitigated bias.

Adaptive Neural Networks. Three types of CNN-based adaptive learning techniques are related to our work: adaptive architectures, adaptive kernels, and attention mechanism. Adaptive architectures design new performance-based neural functions or structures, e.g., neuron selection hidden layers [30] and automatic CNN expansion for FR [84]. As CNN advances many fields, prior works propose dynamic kernels to realize content-adaptive convolutions [19,45,78]. Li et al. [43] propose a shape-driven kernel for facial trait recognition where each landmark-centered patch has a unique kernel. The works of [17,44] use a kernel selection scheme to automatically adjust the receptive field size based on inputs. To better suit input data, [16] splits training data into clusters and learns an exclusive kernel per cluster. Alternatively, one may feed input images into a kernel function to dynamically generate kernels [35,42,66,81]. Despite its effectiveness, such individual adaptation may not be suitable given the diversity of faces in demographic groups. Our work is most related to the side information adaptive convolution [36], where in each layer a sub-network inputs auxiliary information to generate filter weights. We mainly differ in that GAC automatically learns where to use adaptive kernels in a multi-layer CNN (see Figs. 2a and 2c), thus more efficient and capable in applying to a deeper CNN.

As the human perception naturally selects the most pertinent piece of information, attention mechanisms are designed for many tasks, e.g., detection [83], recognition [9], image captioning [8], tracking [7], pose estimation [57], segmentation [51], and image forgery [12,49]. Normally, attention weights are estimated by feeding images or fea-
Fig. 2: A comparison of approaches in adaptive CNNs. In contrast to (a) and (b), the adaptive units in GAC are constructed by demographic information and are automatically applied to corresponding layers.

3. Methodology

3.1. Overview

Our goal is to train a FR network that is impartial to individuals in different demographic groups. Unlike image-related variations, e.g., large-poses or low-resolution faces are harder to be recognized, demographic attributes are subject-related properties with no apparent impact in recognizability of identity, at least from a layman’s perspective. Thus, an unbiased FR system should be able to obtain equally salient features for faces across demographic groups. However, due to imbalanced demographic distributions and inherent face differences between groups, it was shown that certain groups achieve higher performance even with hand-crafted features [40]. Thus, it is impractical to extract features from different demographic groups that exhibit equal discriminability. Despite such disparity, a FR algorithm can still be designed to mitigate the difference in performance.

To this end, we propose a CNN-based group adaptive classifier that utilizes dynamic kernels and attention maps to boost FR performance in all demographic groups considered here. Specifically, GAC has two main modules, an adaptive layer and an automation module. In an adaptive layer, face images or feature maps are convolved with a unique kernel for each demographic group, and multiplied with adaptive attention maps to obtain demographic-differential features for faces in a certain group. The automation module determines in which layers of the network adaptive kernels and attention maps should be applied. As shown in Fig. 3, given an aligned face, and its identity label $y_{ID}$, a pre-trained demographic classifier first estimates its demographic attribute $y_{Demo}$. With $y_{Demo}$, the image is then fed into a recognition network with multiple demographic adaptive layers to estimate its identity. In the following, we present these two modules.

3.2. Adaptive Layer

Adaptive Convolution. For a standard convolution in CNN, an image or feature map from the previous layer $X \in \mathbb{R}^{c \times h \times w}$ is convolved with a single kernel matrix $K \in \mathbb{R}^{k \times k \times h \times w}$, where $c$ is the number of input channels, $k$ the number of filters, $h$ and $w$ the input size, and $h$ and $w$ the filter size. Such an operation shares the kernel with every input going through the layer, and is thus agnostic to demographic content, resulting in limited capacity to represent minority groups. To mitigate the bias in convolution, we introduce a trainable matrix of kernel masks $K^M \in \mathbb{R}^{n \times k \times k \times w \times w}$, where $n$ is the number of demographic groups. In the forward pass, the demographic label $y_{Demo}$ and kernel matrix $K^M$ are fed into the adaptive convolutional layer to generate demographic adaptive filters. Let $K_{i} \in \mathbb{R}^{c \times k \times w \times w}$ denote the $i$th channel of the shared filter. The $i$th channel of adaptive filter for group $y_{Demo}$ is:

$$K^{y_{Demo}}_{i} = K_{i} \otimes K^{M}_{y_{Demo}}. \quad (1)$$

where $K^{M}_{y_{Demo}} \in \mathbb{R}^{c \times k \times w \times w}$ is the $y_{Demo}^{th}$ kernel mask for group $y_{Demo}$ and $\otimes$ denotes element-wise multiplication. Then the $i$th channel of the output feature map is given by $Z_{i} = f(X \ast K^{y_{Demo}}_{i})$, where $\ast$ denotes convolution, and $f(\cdot)$ is activation. Unlike conventional convolution, samples in every demographic group have a unique kernel $K^{y_{Demo}}_{i}$.

Adaptive Attention. Each channel filter in a CNN plays an important role in every dimension of the final representation, which can be viewed as a semantic pattern detector [8]. In the adaptive convolution, however, the values of a kernel mask are broadcast along the channel dimension, indicating that the weight selection is spatially varied but channel-wise joint.
Hence, we introduce a channel-wise attention mechanism to enhance the face features that are demographic-adaptive. First, a trainable matrix of channel attention maps \( M \in \mathbb{R}^{n \times k} \) is initialized in every adaptive attention layer. Given \( y_{\text{Demo}} \) and the current feature map \( Z \in \mathbb{R}^{k \times h^2 \times w^2} \), where \( h^2 \) and \( w^2 \) are the height and width of \( Z \), the \( i^{th} \) channel of the new feature map is calculated by:

\[
Z_i^{\text{Demo}} = \text{Sigmoid}(M_{y_{\text{Demo}}}) \cdot Z_i, \tag{2}
\]

where \( M_{y_{\text{Demo}}} \) is the entry in the \( y_{\text{Demo}} \)\(^{th}\) row of \( M \) for the demographic group \( y_{\text{Demo}} \) at \( i^{th}\) column. In contrast to the adaptive convolution, elements of each demographic attention map \( M_{y_{\text{Demo}}} \) diverge in channel-wise manner, while the single attention weight \( M_{y_{\text{Demo}}} \) is spatially shared by the entire matrix \( Z_i \in \mathbb{R}^{h^2 \times w^2} \). The two adaptive matrices, \( K^M \) and \( M \), are jointly tuned with all the other parameters supervised by the classification loss.

Unlike dynamic CNNs [36] where additional networks are engaged to produce input-variant kernel or attention map, our adaptiveness is yielded by a simple thresholding function directly pointing to the demographic group with no auxiliary networks. Although the kernel network in [36] can generate continuous kernels without enlarging the parameter space, further encoding is required if the side inputs for kernel network are discrete variables. Our approach, in contrast, divides kernels into clusters so that the branch parameter learning can stick to a specific group without interference from individual uncertainties, making it suitable for discrete domain adaptation. Further, the adaptive kernel masks in GAC are more efficient in terms of the number of additional parameters. Compared to a non-adaptive layer, the number of additional parameters of GAC is \( n \times c \times h^2 \times w^2 \), while that of [36] is \( s \times k \times c \times h^2 \times w^2 \) if the kernel network is a one-layer MLP, where \( s \) is the dimension of input side information. Thus, for one adaptive layer, [36] has \( \frac{k \times h^2}{n} \) times more parameters than ours, which can be substantial given the typical large value of \( k \), the number of filters.

3.3. Automation Module

Though faces in different demographic groups are adaptively processed by various kernels and attention maps, it is inefficient to use such adaptations in every layer of a deep CNN. To relieve the burden of unnecessary parameters and avoid empirical trimming, we adopt a similarity fusion process to automatically determine the adaptive layers. Since the same fusion scheme can be applied to both types of adaptation, we take the adaptive convolution as an example to illustrate this automatic scheme.

First, a matrix composed of \( n \) kernel masks is initialized in every convolutional layer. As training continues, each kernel mask is updated independently to reduce classification loss for each demographic group. Second, we reshape the kernel masks into 1D vectors \( V = [v_1, v_2, \ldots, v_n] \), where \( v_i \in \mathbb{R}^l, l = c \times w^2 \times h^2 \) is the kernel mask of the \( i^{th}\) demographic group. Next, we compute Cosine similarity between two kernel vectors, \( \theta_{ij} = \frac{v_i}{\|v_i\|} \cdot \frac{v_j}{\|v_j\|} \), where \( 1 \leq i, j \leq n \). The average similarity of all pair-wise similarities is obtained by \( \theta = \frac{1}{n(n-1)} \sum_{i < j} \theta_{ij} \). If the dissimilarity \( -\theta \) is lower than a pre-defined threshold \( \tau \), the kernel parameters in this layer reveal the demographic-agnostic property. Hence, we merge the \( n \) kernels into a single kernel by averaging along the group dimension. By definition, a lower \( \tau \) implies more adaptive layers. Given an array of \( \{-\theta_i\}^t \) (\( t \) is the total number of convolutional layers), we first sort the elements from smallest to highest, and this way, layers whose \( -\theta_i \) values are larger than \( \tau \) will be adaptive. Thus, when \( \tau \) decreases, more layers will be adaptive. In the subsequent training, this single kernel can still be updated separately for each demographic group, as the...
kernel may become demographic-adaptive in later epochs. We monitor the similarity trend of the adaptive kernels in each layer until it is stable.

3.4. De-biasing Objective Function

Apart from the objective function for face identity classification, we also adopt a regression loss function to narrow the gap of the intra-class distance between demographic groups. Let \( g(\cdot) \) denote the inference function of GAC, and \( I_{ijg} \) is the \( i^{th} \) image of subject \( j \) in group \( g \). Thus, the feature representation of image \( I_{ijg} \) is given by \( r_{ijg} = g(I_{ijg}, w) \), where \( w \) denotes the GAC parameters. Assuming the feature distribution of each subject is a Gaussian distribution with identity covariance matrix (hyper-sphere), we utilize the average Euclidean distance to every subject center as the intra-class distance of each subject. In particular, we first compute the center point of each identity-sphere:

\[
\mu_{jg} = \frac{1}{N} \sum_{i=1}^{N} g(I_{ijg}, w), \tag{3}
\]

where \( N \) is the total number of face images of subject \( j \). The average intra-class distance of subject \( j \) is as follows:

\[
\text{Dist}_{jg} = \frac{1}{N} \sum_{i=1}^{N} (r_{ijg} - \mu_{jg})^T (r_{ijg} - \mu_{jg}). \tag{4}
\]

We then compute the intra-class distance for all subjects in group \( g \) as \( \text{Dist}_g = \frac{1}{Q} \sum_{j=1}^{Q} \text{Dist}_{jg} \), where \( Q \) is the number of total subjects in group \( g \). This allows us to lower the difference of intra-class distance by:

\[
L_{\text{bias}} = \frac{\lambda}{Q \times N} \frac{1}{n} \sum_{g=1}^{Q} \left| \sum_{j=1}^{Q} \text{Dist}_{jg} - \frac{1}{n} \sum_{g=1}^{Q} \text{Dist}_g \right|. \tag{5}
\]

where \( \lambda \) is the coefficient for the de-biasing objective.

4. Experiments

Datasets Our bias study uses RFW dataset [72] for testing and BUPT-Balancedface dataset [71] for training. RFW consists of faces in four race/ethnic groups: White, Black, East Asian, and South Asian. Each group contains \( \sim 10K \) images of 3K individuals for face verification. BUPT-Balancedface contains 1.3M images of 28K celebrities and is approximately race-balanced with 7K identities per race. Other than race, we also study gender bias. We combine IMDB [60], UTKFace [85], AgeDB [54], AAF [10], AFAD [56] to train a gender classifier, which estimates gender of faces in RFW and BUPT-Balancedface. All face images are cropped and resized to \( 112 \times 112 \) pixels via landmarks detected by RetinaFace [14].

Implementation Details We train a baseline network and GAC on BUPT-Balancedface, using the 50-layer ArcFace architecture [13]. The classification loss is an additive Cosine margin in Cosface [70], with the scale and margin of \( s = 64 \) and \( m = 0.5 \). Training is optimized by SGD with a batch size 256. The learning rate starts from 0.1 and drops to 0.0001 following the schedule at 8, 13, 15 epochs for the baseline, and 5, 17, 19 epochs for GAC. We set \( \lambda = 0.1 \) for the intra-distance de-biasing. \( \tau = -0.2 \) is chosen for automatic adaptation in GAC. Our FR models are trained to extract a 512-dim representation. Our demographic classifier uses a 18-layer ResNet [25]. Comparing GAC and the baseline, the average feature extraction speed per image on NVIDIA 1080 Ti GPU is 1.4 ms and 1.1 ms, and the number of model parameters is 44.0 M and 43.6 M, respectively.

Performance Metrics The common group fairness criteria like demographic parity distance [50] are improper to evaluate fairness of learnt representations, since they are designed to measure independence properties of random variables. However, in FR the sensitive demographic characteristics are tied to identities, making these two variables correlated. The NIST report uses false negative and false positive for each demographic group to measure the fairness [21]. Instead of plotting false negative vs. false positives, we adopt a compact quantitative metric, i.e., the standard deviation (STD) of the performance in different demographic groups, previously introduced in [20, 71] and called “biasness”. As bias is considered as systematic error of the estimated values compared to the actual values, here, we assume the average performance to be the actual value. For each demographic group, its biasness is the error between the average and the performance on demographic group. The overall biasness is the expectation of all group errors, which is the STD of performance across groups. We also report average accuracy (Avg) to show the overall FR performance.

4.1. Results on RFW Protocol

We follow RFW face verification protocol with 6K pairs per race/ethnicity. The models are trained on BUPT-Balancedface with ground truth race and identity labels.

Compare with SOTA We compare the GAC with four SOTA algorithms on RFW protocol, namely, ACNN [36], RL-RBN [71], PFE [63], and DebFace [20]. Since the approach in ACNN [36] is related to GAC, we re-implement it and apply to the bias mitigation problem. First, we train a race classifier with the cross-entropy loss on BUPT-Balancedface. Then the softmax output of our race classifier is fed to a filter manifold network (FMN) to generate adaptive filter weights. Here, FMN is a two-layer MLP with a ReLU in between. Similar to GAC, race probabilities are considered as auxiliary information for face representation learning. We also compare with the SOTA approach.
Table 1: Performance comparison with SOTA on the RFW protocol [72]. The results marked by (*) are directly copied from [71].

<table>
<thead>
<tr>
<th>Method</th>
<th>White</th>
<th>Black</th>
<th>East Asian</th>
<th>South Asian</th>
<th>Avg (%)</th>
<th>STD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground-truth</td>
<td>96.20</td>
<td>94.77</td>
<td>94.87</td>
<td>94.98</td>
<td>95.21</td>
<td>0.58</td>
</tr>
<tr>
<td>Estimated</td>
<td>96.27</td>
<td>94.40</td>
<td>94.32</td>
<td>94.77</td>
<td>94.94</td>
<td>0.79</td>
</tr>
<tr>
<td>Random</td>
<td>95.95</td>
<td>93.10</td>
<td>94.18</td>
<td>94.82</td>
<td>94.56</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 2: Ablation of adaptive strategies on the RFW protocol [72].

<table>
<thead>
<tr>
<th>Method</th>
<th>White</th>
<th>Black</th>
<th>East Asian</th>
<th>South Asian</th>
<th>Avg (%)</th>
<th>STD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground-truth</td>
<td>96.20</td>
<td>94.77</td>
<td>94.87</td>
<td>94.98</td>
<td>95.21</td>
<td>0.58</td>
</tr>
<tr>
<td>Estimated</td>
<td>96.27</td>
<td>94.40</td>
<td>94.32</td>
<td>94.77</td>
<td>94.94</td>
<td>0.79</td>
</tr>
<tr>
<td>Random</td>
<td>95.95</td>
<td>93.10</td>
<td>94.18</td>
<td>94.82</td>
<td>94.56</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 3: Ablations on CNN depths and demographics on RFW protocol [72].

<table>
<thead>
<tr>
<th>Method</th>
<th>White</th>
<th>Black</th>
<th>East Asian</th>
<th>South Asian</th>
<th>Avg (%)</th>
<th>STD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground-truth</td>
<td>96.20</td>
<td>94.77</td>
<td>94.87</td>
<td>94.98</td>
<td>95.21</td>
<td>0.58</td>
</tr>
<tr>
<td>Estimated</td>
<td>96.27</td>
<td>94.40</td>
<td>94.32</td>
<td>94.77</td>
<td>94.94</td>
<td>0.79</td>
</tr>
<tr>
<td>Random</td>
<td>95.95</td>
<td>93.10</td>
<td>94.18</td>
<td>94.82</td>
<td>94.56</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 4: Ablations on λ on RFW protocol (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>White</th>
<th>Black</th>
<th>East Asian</th>
<th>South Asian</th>
<th>Avg (%)</th>
<th>STD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground-truth</td>
<td>96.20</td>
<td>94.77</td>
<td>94.87</td>
<td>94.98</td>
<td>95.21</td>
<td>0.58</td>
</tr>
<tr>
<td>Estimated</td>
<td>96.27</td>
<td>94.40</td>
<td>94.32</td>
<td>94.77</td>
<td>94.94</td>
<td>0.79</td>
</tr>
<tr>
<td>Random</td>
<td>95.95</td>
<td>93.10</td>
<td>94.18</td>
<td>94.82</td>
<td>94.56</td>
<td>1.03</td>
</tr>
</tbody>
</table>

PFE [63] by training it on BUP-T-BalancedFace. As shown in Tab. 1, GAC is superior to SOTA w.r.t. average performance and feature fairness. Compared to kernel masks in GAC, the FMN in ACNN [36] contains more trainable parameters. Applying it to each convolutional layer is prone to overfitting. In fact, the layers that are adaptive in GAC (τ = -0.2) are set to be the FMN based convolution in ACNN. As the race data is a four-element input in our case, using extra kernel networks adds complexity to the FR network, which degrades the verification performance. Even though PFE performs the best on standard benchmarks (Tab. 6), it still exhibits high bias. Our GAC outperforms PFE on RFW in both biasness and average performance. Compared to DebFace [20], in which demographic attributes are disentangled from the identity representations, GAC achieves higher verification performance by optimizing the classification for each demographic group, with a lower biasness as well.

Ablation on Adaptive Strategies. To investigate the efficacy of our network design, we conduct three ablation studies: adaptive mechanisms, number of convolutional layers, and demographic information. For adaptive mechanisms, since deep feature maps contain both spatial and channel-wise information, we study the relationship among adaptive kernels, spatial and channel-wise attentions, and their impact to bias mitigation. We also study the impact of τ in our automation module. Apart from the baseline and GAC, we ablate eight variants: (1) GAC-Channel: channel-wise attention for race-differential feature; (2) GAC-Kernel: adaptive convolution with race-specific kernels; (3) GAC-Spatial: only spatial attention is added to baseline; (4) GAC-CS: both channel-wise and spatial attention; (5) GAC-CSK: combine adaptive convolution with spatial and channel-wise attention; (6,7,8) GAC-(τ = *): set τ to *.

From Tab. 2, we make several observations: (1) the baseline model is the most biased across race groups. (2) spatial attention mitigates the race bias at the cost of verification accuracy, and is less effective on learning fair features than other adaptive techniques. This is probably because spatial contents, especially local layout information, only reside at earlier CNN layers, where the spatial dimensions are gradually decreased by the later convolutions and poolings. Thus, semantic details like demographic attributes are hardly encoded spatially. (3) Compared to GAC, combining adaptive kernels with both spatial and channel-wise attention increases the number of parameters, lowering the performance. (4) As τ determines the number of adaptive layers in GAC, it has a great impact on the performance. A small τ may increase redundant adaptive layers, while the adaptation layers may lack in capacity if too large.

Ablation on Depths and Demographic Labels. Both the adaptive layers and de-biasing loss in GAC can be applied to CNN in any depth. In this ablation, we train both the baseline and GAC (λ = 0.1, τ = -0.2) in ArcFace architecture with three different numbers of layers: 34, 50, and 100. As the training of GAC relies on demographic information, the error and bias in demographic labels might impact the bias reduction of GAC. Thus, we also ablate with different demographic information, (1) ground-truth: the race/ethnicity labels provided by RFW; (2) estimated: the labels predicted by a pre-trained race estimation model; (3) random: the demographic label randomly assigned to each face.

As shown in Tab. 3, compared to the baselines, GAC successfully reduces the STD at different number of layers. We see that the model with least number of layers presents the most bias, and the bias reduction by GAC is the most as well. The noise and bias in demographic labels do, however, impair the performance of GAC. With estimated demographics, the biasness is higher than that of the model with ground-truth supervision. Meanwhile, the model trained with random demographics has the highest biasness. Even so, using estimated attributes during testing still improves fairness in
4.2 Results on Gender and Race Groups

We now extend demographic attributes to both gender and race. First, we train two classifiers that predict gender and race/ethnicity of a face image. The classification accuracy of gender and race/ethnicity is 85% and 81%, respectively. Then, we train two classifiers that predict gender and race/ethnicity of a face image. The classification accuracy of gender and race/ethnicity is 85% and 81%, respectively. Then, these fixed classifiers are affiliated with GAC to provide demographic information for learning adaptive kernels and attention maps. We merge BUP-T-BalancedFace and RFW, and split the subjects into 5 sets for each of 8 demographic groups. In 5-fold cross-validation, each time a model is trained on 4 sets and tested on the remaining set.

Here we demonstrate the efficacy of the automation module for GAC. We compare to the scheme of manually design (AL+Manual) that adds adaptive kernels and attention maps to a subset of layers. Specifically, the first block in every residual unit is chosen to be the adaptive convolution layer, and channel-wise attentions are applied to the feature map output by the last block in each residual unit. As we use 4 residual units and each block has 2 convolutional layers, the manual scheme involves 8 adaptive convolutional layers and 4 groups of channel-wise attention maps. As in Tab. 5, automatic adaptation is more effective in enhancing the discriminability and fairness of face representations. Figure 4a shows the dissimilarity of kernel masks in the convolutional layers changes during training epochs under three thresholds \( \tau \). A lower \( \tau \) results in more adaptive layers. We see the layers that are determined to be adaptive do vary across both layers (vertically) and training time (horizontally), which shows the importance of our automatic mechanism.

4.3 Results on Standard Benchmark Datasets

While our GAC mitigates bias, we also hope it can perform well on standard benchmarks. Therefore, we evaluate GAC on standard benchmarks without considering demographic impacts, including LFW [32], IJB-A [41], and IJB-C [53]. These datasets exhibit imbalanced distribution in demographics. For a fair comparison with SOTA, instead of using ground truth demographics, we train GAC on Ms-Celeb-1M [23] with the demographic attributes estimated by the classifier pre-trained in Sec. 4.2. As in Tab. 6, GAC
outperforms the baseline and performs comparable to SOTA.

4.4. Visualization and Analysis on Bias of FR

Visualization To understand the adaptive kernels in GAC, we visualize the feature maps at an adaptive layer for faces of various demographics, via a Pytorch visualization tool [57]. We visualize important face regions pertaining to the FR decision by using a gradient-weighted class activation mapping (Grad-CAM) [62]. Grad-CAM uses the gradients back from the final layer corresponding to an input identity, and guides the target feature map to highlight important regions for identity predicting. Figure 5 shows that, compared to the baseline, the salient regions of GAC demonstrate more diversity on faces from different groups. This illustrates the variability of network parameters in GAC across different groups.

Bias via Local Geometry In addition to STD, we explain the bias phenomenon via the local geometry of a given face representation in each demographic group. We assume that the statistics of neighbors of a given point (representation) reflects certain properties of its manifold (local geometry). Thus, we illustrate the pairwise correlation of face representations. To minimize variations caused by other variables, we use constrained frontal faces of a mug shot dataset, PCSO [40]. We randomly select 1K White and 1K Black subjects from PCSO, and compute their pair-wise correlation within each race. In Fig. 4b, Base-White representations show lower inter-class correlation than Base-Black, i.e., faces in the White group are over-represented by the baseline than the Black group. In contrast, GAC-White and GAC-Black shows more similarity in their correlation histograms.

As PCSO has few Asian subjects, we use RFW for another examination of the local geometry in 4 groups. That is, after normalizing the representations, we compute the pairwise Euclidean distance and measure the ratio between the minimum distance of inter-subjects pairs and the maximum distance of intra-subject pairs. We compute the mean and standard deviation (StaD) of ratio distributions in 4 groups, by two models. Also, we gauge the relative entropy to measure the deviation of distributions from each other. For simplicity, we choose White group as the reference distribution. As shown in Tab. 7, while GAC has minor improvement over baseline in the mean, it gives smaller relative entropy in the other 3 groups, indicating that the ratio distributions of other races in GAC are more similar, i.e., less biased, to the reference distribution. These results demonstrate the capability of GAC to increase fairness of face representations.

5. Conclusion

This paper tackles the issue of demographic bias in face recognition by learning a fair face representation. A group adaptive classifier (GAC) is proposed to improve robustness of representations for every demographic group. Both adaptive convolution kernels and channel-wise attention maps are introduced to GAC. We further add an automation module to determine whether to use adaptations in a given layer. Our findings suggest that faces can be better represented by using layers adaptive to different demographic groups, leading to more balanced performance gain for all groups.

Acknowledgement This work is supported by U.S. Department of Commerce (#00NANB19D154), National Institute of Standards and Technology.

<table>
<thead>
<tr>
<th>Race</th>
<th>Mean Baseline</th>
<th>GAC Baseline</th>
<th>StaD Baseline</th>
<th>GAC Baseline</th>
<th>Relative Entropy Baseline</th>
<th>GAC Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>1.15</td>
<td>1.17</td>
<td>0.30</td>
<td>0.31</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Black</td>
<td>1.07</td>
<td>1.10</td>
<td>0.27</td>
<td>0.28</td>
<td>0.61</td>
<td>0.43</td>
</tr>
<tr>
<td>East Asian</td>
<td>1.08</td>
<td>1.10</td>
<td>0.31</td>
<td>0.32</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td>South Asian</td>
<td>1.15</td>
<td>1.18</td>
<td>0.31</td>
<td>0.32</td>
<td>0.19</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Figure 5: The first row shows the average faces of different groups in RFW. The next two rows show gradient-weighted class activation heatmaps [62] at the 43th convolutional layer of the GAC and baseline. The higher diversity of heatmaps in GAC shows the variability of parameters in GAC across groups.
References

[18] Qi Dong, Shaogang Gong, and Xiaotian Zhu. Imbalanced deep learning by minority class incremental rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. 1
[26] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In ECCV, 2018. 2
[31] Chen Huang, Yining Li, Change Loy Chen, and Xiaou T Ang. Deep imbalanced learning for face recognition and attribute prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. 1
[33] Yuge Huang, Pengcheng Shen, Yng Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. Improving face recognition from hard samples via distribution distillation loss. In ECCV, 2020. 1


