Mitigating Face Recognition Bias via Group Adaptive Classifier

Sixue Gong Xiaoming Liu Anil K. Jain
Michigan State University, East Lansing MI 48824
{gongsixu, liuxm, jain}@msu.edu

https://github.com/gongsixue/GAC

Abstract

Face recognition is known to exhibit bias - subjects in a certain demographic group can be better recognized than other groups. This work aims to learn a fair face representation, where faces of every group could 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

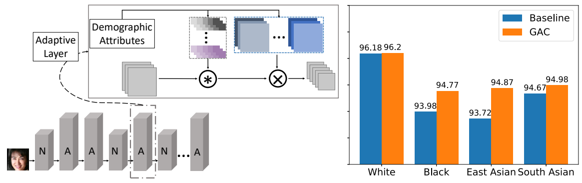


Figure 1: (a) Our proposed group adaptive classifier (GAC) automatically chooses between non-adaptive ("N") and adaptive ("A") layer in a multi-layer network, where the latter uses demographic-group-specific kernel and attention. (b) Compared to the baseline with the 50-layer ArcFace backbone, GAC improves face verification accuracy in most groups of RFW dataset [72], especially under-represented groups, leading to mitigated FR bias, *e.g.*, GAC reduces biasness from 1.11 to 0.60.

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:

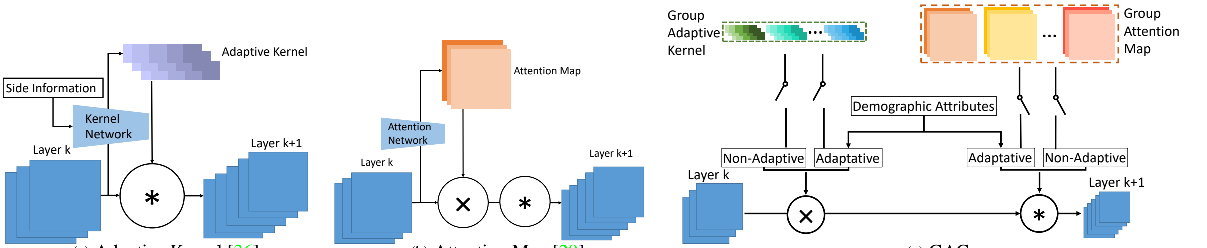
- A new face recognition algorithm that reduces demographic bias and tailors representations for faces in every demographic group by adopting adaptive convolutions and attention techniques;
- A new adaptation mechanism that automatically determines the layers to employ dynamic kernels and attention maps;
- 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 [67], segmentation [51], and image forgery [12, 49]. Normally, attention weights are estimated by feeding images or fea-



(a) Adaptive Kernel [36] (b) Attention Map [29] (c) GAC

Figure 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.

tures into a shared network, composed of convolutional and pooling layers [3, 9, 46, 64] or multi-layer perceptron (MLP) [29, 47, 61, 76]. Apart from feature-based attention, Hou *et al*. [27] propose a correlation-guided cross attention map for few-shot classification where the correlation between the class feature and query feature generates attention weights. The work of [79] introduces a cross-channel communication block to encourage information exchange across channels. To accelerate channel interaction, Wang *et al*. [73] propose a 1D convolution across channels for attention prediction. Different from prior work, our attention maps are constructed by demographic information (see Figs. 2b, 2c), which improves the robustness of face representations in every demographic group.

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 $X \in \mathbb{R}^{c\times h^X \times w^X}$ is convolved with a single kernel matrix $K \in \mathbb{R}^{k\times c\times h^K \times w^K}$, where c is the number of input channels, k the number of filters, h^X and w^X the input size, and h^K and w^K 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 $K^M \in \mathbb{R}^{n\times c\times h^K \times w^K}$, 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 $K_i \in \mathbb{R}^{c\times h^K \times w^K}$ denote the i^{th} channel of the shared filter. The i^{th} channel of adaptive filter for group y_{Demo} is:

$$K_i^{y_{Demo}} = K_i \bigotimes K^M_{y_{Demo}}, \tag{1}$$

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

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.

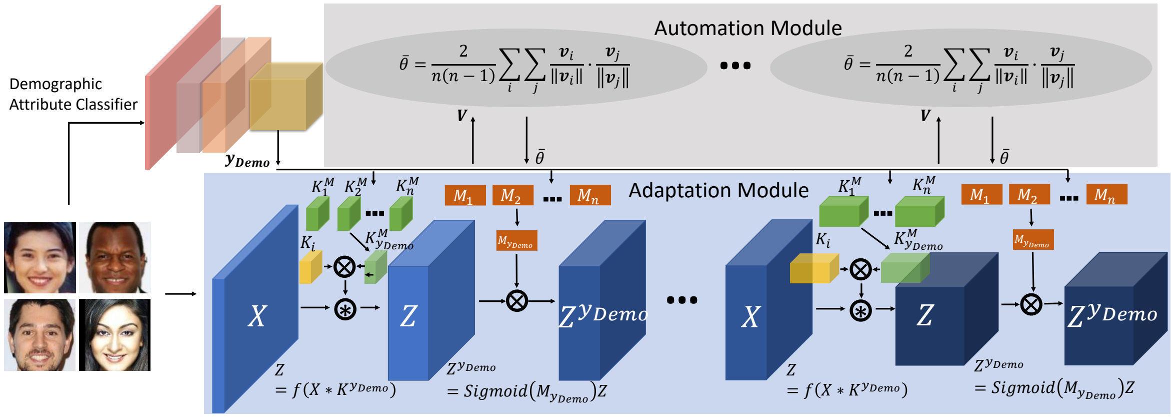


Figure 3: Overview of the proposed GAC for mitigating FR bias. GAC contains two major modules: the adaptive layer and the automation module. The adaptive layer consists of adaptive kernels and attention maps. The automation module is employed to decide whether a layer should be adaptive or not.

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_{Demo} and the current feature map $Z \in \mathbb{R}^{k\times h^Z\times w^Z}$, where h^Z and w^Z are the height and width of Z, the i^{th} channel of the new feature map is calculated by:

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

where $M_{y_{Demo}i}$ is the entry in the $y_{Demo}{}^{th}$ row of M for the demographic group y_{Demo} at i^{th} column. In contrast to the adaptive convolution, elements of each demographic attention map $M_{y_{Demo}}$ diverge in channel-wise manner, while the single attention weight $M_{y_{Demo}i}$ is spatially shared by the entire matrix $Z_i \in \mathbb{R}^{h^Z\times w^Z}$. 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 $n\times c\times h^K\times w^K$, while that of [36] is $s\times k\times c\times h^K\times w^K$ 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 $\frac{s\times k}{n}$ 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 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$, where $\mathbf{v}_i \in \mathbb{R}^l, l = c\times w^K\times h^K$ is the kernel mask of the i^{th} demographic group. Next, we compute Cosine similarity between two kernel vectors, $\theta_{ij} = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}\cdot\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}$, where $1 \leq i, j \leq n$. The average similarity of all pair-wise similarities is obtained by $\overline{\theta} = \frac{2}{n(n-1)}\sum_i\sum_j \theta_{ij}, i \neq j$. If the dissimilarity $-\overline{\theta}$ is lower than a pre-defined threshold τ, 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 τ 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 τ will be adaptive. Thus, when τ 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 $\overline{\theta}$ is stable.

3.4. De-biasing Objective Function

Apart from the objective function for face identity classification, we also adopt a regress 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 $\mathbf{r}_{ijg} = g(I_{ijg}, \mathbf{w})$, where $\mathbf{w}$ 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:

$$\boldsymbol{\mu}_{jg} = \frac{1}{N}\sum_{i=1}^{N} g(I_{ijg}, \mathbf{w}), \tag{3}$$

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

$$Dist_{jg} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{r}_{ijg} - \boldsymbol{\mu}_{jg})^T (\mathbf{r}_{ijg} - \boldsymbol{\mu}_{jg}). \tag{4}$$

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

$$\mathcal{L}_{bias} = \frac{\lambda}{Q \times n}\sum_{g=1}^{n}\sum_{j=1}^{Q}\Big| Dist_{jg} - \frac{1}{n}\sum_{g=1}^{n} Dist_g \Big|, \tag{5}$$

where λ 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 [1]. Each group contains ∼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 112×112 pixels via landmarks detected by RetinaFace [14].

[1] RFW [72] uses Caucasian, African, Asian, and Indian to name demographic groups. We adopt these groups and accordingly rename to White, Black, East Asian, and South Asian for clearer race/ethnicity definition.

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 1080Ti GPU is 1.4ms and 1.1ms, and the number of model parameters is 44.0M and 43.6M, 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

Method	White	Black	East Asian	South Asian	Avg (↑)	STD (↓)
RL-RBN [71]	96.27	95.00	94.82	94.68	95.19	0.63
ACNN [36]	96.12	94.00	93.67	94.55	94.58	0.94
PFE [63]	**96.38**	**95.17**	94.27	94.60	95.11	0.93
ArcFace [13]	96.18*	94.67*	93.72*	93.98*	94.64	0.96
CosFace [70]	95.12*	93.93*	92.98*	92.93*	93.74	0.89
DebFace [20]	95.95	93.67	94.33	94.78	94.68	0.83
GAC	96.20	94.77	**94.87**	**94.98**	**95.21**	**0.58**

Table 1: Performance comparison with SOTA on the RFW protocol [72]. The results marked by (*) are directly copied from [71].

Method	White	Black	East Asian	South Asian	Avg (↑)	STD (↓)
Baseline	96.18	93.98	93.72	94.67	94.64	1.11
GAC-Channel	95.95	93.67	94.33	94.78	94.68	0.83
GAC-Kernel	96.23	94.40	94.27	94.80	94.93	0.78
GAC-Spatial	95.97	93.20	93.67	93.93	94.19	1.06
GAC-CS	96.22	93.95	94.32	**95.12**	94.65	0.87
GAC-CSK	96.18	93.58	94.28	94.83	94.72	0.95
GAC-($\tau = 0$)	96.18	93.97	93.88	94.77	94.70	0.92
GAC-($\tau = -0.1$)	**96.25**	94.25	94.83	94.72	95.01	0.75
GAC-($\tau = -0.2$)	96.20	**94.77**	**94.87**	94.98	**95.21**	**0.58**

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

Method	White	Black	East Asian	South Asian	Avg (↑)	STD (↓)
Number of Layers						
ArcFace-34	96.13	93.15	92.85	93.03	93.78	1.36
GAC-ArcFace-34	96.02	94.12	94.10	94.22	94.62	0.81
ArcFace-50	96.18	93.98	93.72	94.67	94.64	1.11
GAC-ArcFace-50	96.20	94.77	94.87	94.98	95.21	0.58
ArcFace-100	96.23	93.83	94.27	94.80	94.78	0.91
GAC-ArcFace-100	96.43	94.53	94.90	95.03	95.22	0.72
Race/Ethnicity Labels						
Ground-truth	96.20	94.77	94.87	94.98	95.21	0.58
Estimated	96.27	94.40	94.32	94.77	94.94	0.79
Random	95.95	93.10	94.18	94.82	94.50	1.03

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

λ	White	Black	East Asian	South Asian	Avg (↑)	STD (↓)
0	96.23	94.65	94.93	95.12	95.23	0.60
0.1	96.20	94.77	94.87	94.98	95.21	0.58
0.5	94.89	94.00	93.67	94.55	94.28	0.47

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

PFE [63] by training it on BUPT-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 ($\tau = -0.2$) 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 biasness. 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-($\tau = *$): set τ to $*$.

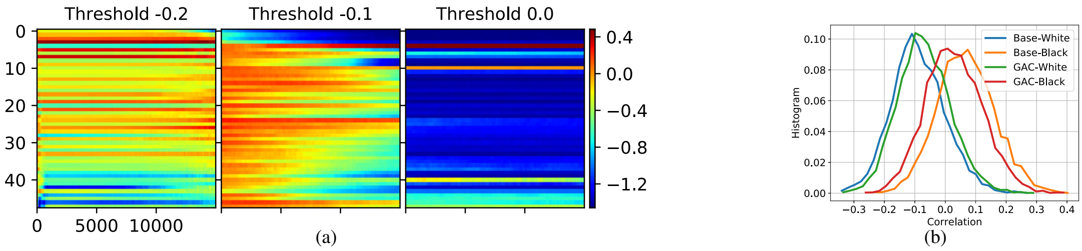
From Tab. 2, we make several observations: (1) the baseline model is the most biased across race groups. (2) spatial attention mitigates the race bias at the cost of verification accuracy, and is less effective on learning fair features than other adaptive techniques. This is probably because spatial 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 ($\lambda = 0.1$, $\tau = -0.2$) 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

Method	Gender	White	Black	East Asian	South Asian	Avg (↑)	STD (↓)
Baseline	Male	97.49 ± 0.08	96.94 ± 0.26	97.29 ± 0.09	97.03 ± 0.13	96.96 ± 0.03	0.69 ± 0.04
	Female	97.19 ± 0.10	97.93 ± 0.11	95.71 ± 0.11	96.01 ± 0.08		
AL+Manual	Male	98.57 ± 0.10	98.05 ± 0.17	98.50 ± 0.12	$\mathbf{98.36 \pm 0.02}$	98.09 ± 0.05	0.66 ± 0.07
	Female	98.12 ± 0.18	$\mathbf{98.97 \pm 0.13}$	96.83 ± 0.19	97.33 ± 0.13		
GAC	Male	$\mathbf{98.75 \pm 0.04}$	$\mathbf{98.18 \pm 0.20}$	$\mathbf{98.55 \pm 0.07}$	98.31 ± 0.12	$\mathbf{98.19 \pm 0.06}$	$\mathbf{0.56 \pm 0.05}$
	Female	$\mathbf{98.26 \pm 0.16}$	98.80 ± 0.15	$\mathbf{97.09 \pm 0.12}$	$\mathbf{97.56 \pm 0.10}$		

Table 5: Verification Accuracy (%) of 5-fold cross-validation on 8 groups of RFW [72].



(a) (b)

Figure 4: (a) For each of the three τ in automatic adaptation, we show the average similarities of pair-wise demographic kernel masks, *i.e.*, $\overline{\theta}$, at 1-48 layers (y-axis), and 1-15K training steps (x-axis). The number of adaptive layers in three cases, *i.e.*, $\sum_1^{48}(\overline{\theta} > \tau)$ at $15K^{th}$ step, are 12, 8, and 2, respectively. (b) With two race groups (White, Black in PCSO [40]) and two models (baseline, GAC), for each of the four combinations, we compute pair-wise correlation of face representations using any two of 1K subjects in the same race, and plot the histogram of correlations. GAC reduces the difference/bias of two distributions.

face recognition compared to baseline. This indicates the efficacy of GAC even in the absence of ground-truth labels.

Ablation on λ. We use λ to control the weight of de-biasing loss. Tab. 4 reports the results of GAC trained with different values of λ. When $\lambda = 0$, de-biasing loss is removed in training. The results indicate a larger λ leads to lower biasness at the cost of overall accuracy.

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%[2], respectively. Then, these fixed classifiers are affiliated with GAC to provide demographic information for learning adaptive kernels and attention maps. We merge BUPT-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

Method	LFW (%)	Method	IJB-A (%) 0.1% FAR	IJB-C @ FAR (%) 0.001%	0.01%	0.1%
DeepFace+ [68]	97.35	Yin *et al.* [80]	73.9 ± 4.2	-	-	69.3
CosFace [70]	99.73	Cao *et al.* [5]	90.4 ± 1.4	74.7	84.0	91.0
ArcFace [13]	**99.83**	Multicolumn [77]	*92.0 ± 1.3*	<u>77.1</u>	<u>86.2</u>	<u>92.7</u>
PFE [63]	*99.82*	PFE [63]	$\mathbf{95.3 \pm 0.9}$	**89.6**	**93.3**	**95.5**
Baseline	99.75	Baseline	90.2 ± 1.1	80.2	88.0	92.9
GAC	<u>99.78</u>	GAC	<u>91.3 ± 1.2</u>	*83.5*	*89.2*	*93.7*

Table 6: Verification performance on LFW, IJB-A, and IJB-C. [Key: **Best**, *Second*, <u>Third</u> Best]

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 τ. A lower τ 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

[2]This seemingly low accuracy is mainly due to the large dataset we assembled for training and testing gender/race classifiers. Our demographic classifier has been shown to perform comparably as SOTA on common benchmarks. While demographic estimation errors impact the training, testing, and evaluation of bias mitigation algorithms, the evaluation is of the most concern as demographic label errors may greatly impact the biasness calculation. Thus, future development may include either manually cleaning the labels, or designing a biasness metric robust to label errors.

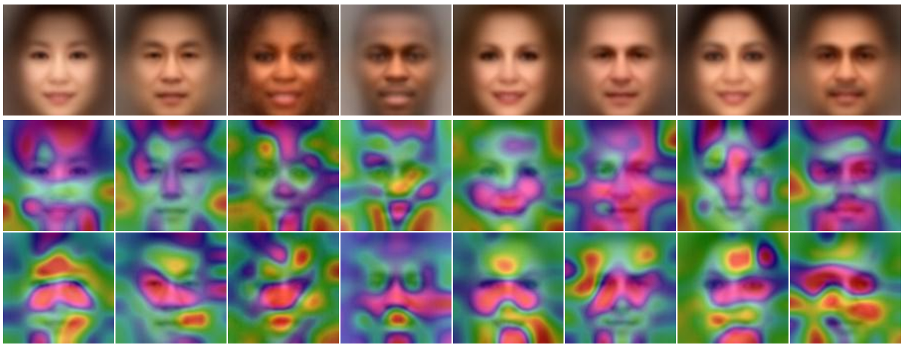


Figure 5: The first row shows the average faces of different groups in RFW. The next two rows show gradient-weighted class activation heatmaps [62] at the 43^{th} convolutional layer of the GAC and baseline. The higher diversity of heatmaps in GAC shows the variability of parameters in GAC across groups.

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 import 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 pair-wise 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 pair-wise 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.

Race	Mean		StaD		Relative Entropy	
	Baseline	GAC	Baseline	GAC	Baseline	GAC
White	1.15	1.17	0.30	0.31	0.0	0.0
Black	1.07	1.10	0.27	0.28	0.61	0.43
East Asian	1.08	1.10	0.31	0.32	0.65	0.58
South Asian	1.15	1.18	0.31	0.32	0.19	0.13

Table 7: Distribution of ratios between minimum inter-class distance and maximum intra-class distance of face features in 4 race groups of RFW. GAC exhibits higher ratios, and more similar distributions to the reference.

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 (#60NANB19D154), National Institute of Standards and Technology.

References

[1] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. Applying support vector machines to imbalanced datasets. In *ECML*. Springer, 2004. 1

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCV*, 2018. 2

[3] Alexei A Bastidas and Hanlin Tang. Channel attention networks. In *CVPR Workshops*, 2019. 3

[4] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *NeurIPS*, 2017. 2

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGface2: A dataset for recognising faces across pose and age. In *FRGC*. IEEE, 2018. 7

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002. 1

[7] Boyu Chen, Peixia Li, Chong Sun, Dong Wang, Gang Yang, and Huchuan Lu. Multi attention module for visual tracking. *Pattern Recognition*, 2019. 2

[8] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017. 2, 3

[9] Zetao Chen, Lingqiao Liu, Inkyu Sa, Zongyuan Ge, and Margarita Chli. Learning context flexible attention model for long-term visual place recognition. *IEEE Robotics and Automation Letters*, 2018. 2, 3

[10] Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3):333–345, 2019. 5

[11] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *ICML*, 2019. 1, 2

[12] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. In *CVPR*, 2020. 2

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1, 5, 6, 7

[14] Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *arxiv*, 2019. 5

[15] Thomas G Dietterich and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University, 1995. 1

[16] Chen Ding, Ying Li, Yong Xia, Wei Wei, Lei Zhang, and Yanning Zhang. Convolutional neural networks based hyperspectral image classification method with adaptive kernels. *Remote Sensing*, 2017. 2

[17] Chen Ding, Ying Li, Yong Xia, Lei Zhang, and Yanning Zhang. Automatic kernel size determination for deep neural networks based hyperspectral image classification. *Remote Sensing*, 2018. 2

[18] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1

[19] Jian Du, Shanghang Zhang, Guanhang Wu, José MF Moura, and Soummya Kar. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*, 2017. 2

[20] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. *ECCV*, 2020. 2, 5, 6

[21] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (FRVT) part 3: Demographic effects. In *Technical Report, National Institute of Standards and Technology*, 2019. 1, 5

[22] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *NeurIPS*, 2019. 2

[23] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 7

[24] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, 2016. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[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

[27] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *NeurIPS*, 2019. 2, 3

[28] J Howard, Y 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 *IEEE BTAS*, 2019. 1

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[30] Ting-Kuei Hu, Yen-Yu Lin, and Pi-Cheng Hsiu. Learning adaptive hidden layers for mobile gesture recognition. In *AAAI*, 2018. 2

[31] Chen Huang, Yining Li, Change Loy Chen, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[32] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008. 7

[33] Yuge Huang, Pengcheng Shen, Ying 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

[34] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: Adaptive curriculum learning loss for deep face recognition. In *CVPR*, 2020. 1

[35] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NeurIPS*, 2016. 2

[36] Di Kang, Debarun Dhar, and Antoni Chan. Incorporating side information by adaptive convolution. In *NeurIPS*, 2017. 2, 3, 4, 5, 6

[37] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019. 2

[38] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *CVPR*, 2019. 1

[39] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *AAAI/ACM*, 2019. 2

[40] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Trans. Information Forensics and Security*, 7(6):1789–1801, 2012. 1, 3, 7, 8

[41] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 2015. 7

[42] Benjamin Klein, Lior Wolf, and Yehuda Afek. A dynamic convolutional layer for short range weather prediction. In *CVPR*, 2015. 2

[43] Shaoxin Li, Junliang Xing, Zhiheng Niu, Shiguang Shan, and Shuicheng Yan. Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition. In *CVPR*, 2015. 2

[44] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, 2019. 2

[45] Xudong Li, Mao Ye, Yiguang Liu, and Ce Zhu. Adaptive deep convolutional neural networks for scene-specific object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 2

[46] Hefei Ling, Jiyang Wu, Junrui Huang, Jiazhong Chen, and Ping Li. Attention-based convolutional neural network for deep face recognition. *Multimedia Tools and Applications*, 2020. 3

[47] Drew Linsley, D Schiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. In *ICLR*, 2019. 3

[48] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1

[49] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. In *arXiv:2103.10596*, 2021. 2

[50] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *NeurIPS*, 2019. 2, 5

[51] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 2

[52] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018. 2

[53] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. IARPA janus benchmark-c: Face dataset and protocol. In *2018 ICB*, 2018. 7

[54] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*, 2017. 5

[55] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *NeurIPS*, 2018. 2

[56] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, 2016. 5

[57] Utku Ozbulak. Pytorch cnn visualizations. `https://github.com/utkuozbulak/pytorch-cnn-visualizations`, 2019. 8

[58] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *NeurIPS*, 2017. 2

[59] Haoyu Qin. Asymmetric rejection loss for fairer face recognition. *arXiv preprint arXiv:2002.03276*, 2020. 2

[60] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2018. 5

[61] Muhammad Sadiq, Daming Shi, Meiqin Guo, and Xiaochun Cheng. Facial landmark detection via attention-adaptive deep network. *IEEE Access*, 2019. 3

[62] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 8

[63] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, 2019. 5, 6, 7

[64] Vishwanath A Sindagi and Vishal M Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 2019. 3

[65] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *ICAIS*, 2019. 2

[66] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, 2019. 2

[67] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *CVPR*, 2019. 2

[68] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 7

[69] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017. 1

[70] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 1, 5, 6, 7

[71] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, 2020. 1, 2, 5, 6

[72] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, 2019. 1, 2, 5, 6, 7

[73] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. *arXiv preprint arXiv:1910.03151*, 2019. 2, 3

[74] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *CVPR*, 2019. 2

[75] Zeyu Wang, Klint Qinami, Yannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. *arXiv preprint arXiv:1911.11834*, 2019. 2

[76] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 3

[77] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. *arXiv preprint arXiv:1807.09192*, 2018. 7

[78] Chao Xiong, Xiaowei Zhao, Danhang Tang, Karlekar Jayashree, Shuicheng Yan, and Tae-Kyun Kim. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, 2015. 2

[79] Jianwei Yang, Zhile Ren, Chuang Gan, Hongyuan Zhu, and Devi Parikh. Cross-channel communication networks. In *NeurIPS*, 2019. 2, 3

[80] Xi Yin and Xiaoming Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Trans. Image Processing*, 27(2):964–975, 2017. 7

[81] Julio Zamora Esquivel, Adan Cruz Vargas, Paulo Lopez Meyer, and Omesh Tickoo. Adaptive convolutional kernels. In *ICCV Workshops*, 2019. 2

[82] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013. 2

[83] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018. 2

[84] Yuanyuan Zhang, Dong Zhao, Jiande Sun, Guofeng Zou, and Wentao Li. Adaptive convolutional neural network and its application in face recognition. *Neural Processing Letters*, 2016. 2

[85] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017. 5

[86] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 2