Mitigating Face Recognition Bias via Group Adaptive Classifier (Supplementary Material)

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

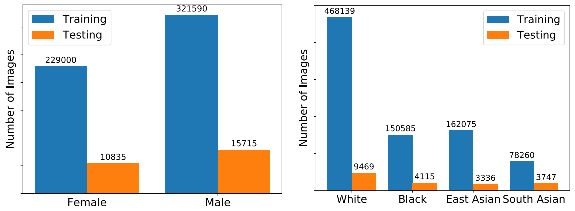
In this supplementary material we include: (1) Section 1: the statistics of datasets used in the experiments; (2) Section 2: performance of the pre-trained gender and race/ethnicity classifiers to provide GAC with demographic information; (3) Section 3: the study on demographic proportions in training set and the intrinsic bias; (4) Section 4: additional experimental results on RFW and IJB-C; (5) Section 5: comparisons of network complexity and FLOPs; (6) Section 6: ablation study on the automation module in GAC.

1. Datasets

Tab. 1 summarizes the datasets we adopt for conducting experiments, which reports the total number of face images and subjects (identities), and the types of demographic annotations. In the cross-validation experiment in Tab. 2, we report the statistics of each data fold for the cross-validation experiment on BUPT-Balancedface and RFW datasets.

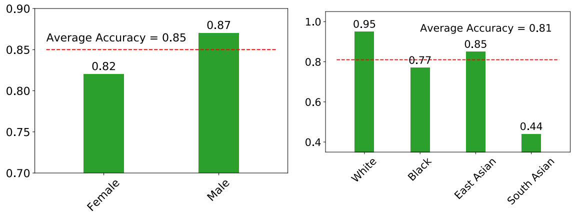
2. Demographic Attribute Estimation

We train a gender classifier and a race/ethnicity classifier to provide GAC with demographic information during both training and testing procedures. We use the same datasets for training and evaluating the two demographic attribute classifiers as the work of [3]. The combination of IMDB, UTKface, AgeDB, AFAD, and AAF is used for gender estimation, and the collection of AFAD, RFW, IMFDB-CVIT, and PCSO is used for race/ethnicity estimation. Fig. 1 shows the total number of images in each demographic group of the training and testing set. Fig. 2 shows the performance of demographic attribute estimation on the testing set. For gender estimation, we see that the performance in the male group is better than that in the female group. For race/ethnicity estimation, the white group outperforms the other race/ethnicity groups.



(a) Gender Distribution (b) Race Distribution

Figure 1: Statistics of the datasets for training and testing demographic attribute estimation networks. (a) The number of images in each gender group of the datasets for gender estimation; (b) The number of images in each race/ethnicity group of the datasets for race/ethnicity estimation.



(a) Gender Estimation (b) Race/Ethnicity Estimation

Figure 2: Performance of the demographic attribute estimation networks. (a) The classification accuracy in each gender group; (b) The classification accuracy in each race/ethnicity group. The red dashed line shows the average performance.

3. Analysis on Intrinsic Bias and Data Bias

For all the algorithms listed in Tab. 1 of the main paper, the performance is higher in White group than those in the other three groups, even though all the models are trained on a demographic balanced dataset, BUPT-Balancedface [12]. In this section, we further investigate the intrinsic bias of face recognition between demographic groups and the impact of the data bias in the training set. *Are non-White faces inherently difficult to be recognized for existing algorithms? Or, are face images in BUPT-Balancedface (the training set) and RFW [13] (testing set) biased towards the White group?*

To this end, we train our GAC network using training sets with different race/ethnicity distributions and evaluate them on RFW. In total, we conduct four experiments, in which we gradually reduce the total number of subjects in the White

Datasets	# of Images	# of Subjects	Demographic Annotations
IMDB [10]	460,723	20,284	Gender, Age
UTKFace [14]	24,106	-	Gender, Age, Race/ethnicity
AgeDB [8]	16,488	567	Gender, Age
AFAD [9]	165,515	-	Gender, Age, Ethnicity (East Asian)
AAF [1]	13,322	13,322	Gender, Age
RFW [13]	665,807	-	Race/Ethnicity
BUPT-Balancedface [12]	1,251,430	28,000	Race/Ethnicity
IMFDB-CVIT [11]	34,512	100	Gender, Age Groups, Ethnicity (South Asian)
MS-Celeb-1M [4]	5,822,653	85,742	No Demographic Labels
PCSO [2]	1,447,607	5,749	Gender, Age, Race/Ethnicity
LFW [5]	13,233	5,749	No Demographic Labels
IJB-A [6]	25,813	500	Gender, Age, Skin Tone
IJB-C [7]	31,334	3,531	Gender, Age, Skin Tone

Table 1: Statistics of training and testing datasets for the experiments in the paper.

Fold	White (#)		Black (#)		East Asian (#)		South Asian (#)	
	Subjects	Images	Subjects	Images	Subjects	Images	Subjects	Images
1	1,991	68,159	1,999	67,880	1,898	67,104	1,996	57,628
2	1,991	67,499	1,999	65,736	1,898	66,258	1,996	57,159
3	1,991	66,091	1,999	65,670	1,898	67,696	1,996	56,247
4	1,991	66,333	1,999	67,757	1,898	65,341	1,996	57,665
5	1,994	68,597	1,999	67,747	1,898	68,763	2,000	56,703

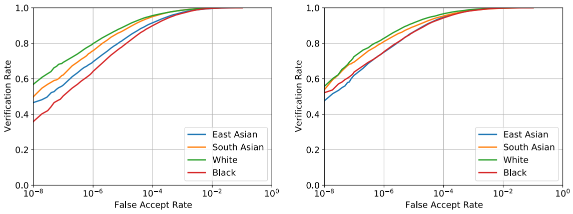
Table 2: Statistics of dataset folds in the cross-validation experiment.

Training Ratio	White	Black	East Asian	South Asian	Avg (↑)	STD (↓)
$7:7:7:7$	96.20	94.77	94.87	94.98	95.21	0.58
$5:7:7:7$	96.53	94.67	94.55	95.40	95.29	0.79
$3.5:7:7:7$	96.48	94.52	94.45	95.32	95.19	0.82
$1:7:7:7$	95.45	94.28	94.47	95.13	94.83	0.48
$0:7:7:7$	92.63	92.27	92.32	93.37	92.65	0.44

Table 3: Verification accuracy (%) on the RFW protocol [13] with varying race/ethnicity distribution in the training set.

group from the BUPT-Balancedface dataset. To construct a new training set, subjects from the non-White groups in BUPT-Balancedface remain the same, while a subset of subjects is randomly picked from the White group. As a result, the ratios between non-White groups are consistently the same, and the ratios of White, Black, East Asian, South Asian are $\{5:7:7:7\}$, $\{3.5:7:7:7\}$, $\{1:7:7:7\}$, $\{0:7:7:7\}$ in the four experiments, respectively. In the last setting, we completely remove White from the training set.

Tab. 3 reports the face verification accuracy of models trained with different race/ethnicity distributions on RFW. For comparison, we also put our results on the balanced dataset here (with ratio $\{7:7:7:7\}$), where all images in BUPT-Balancedface are used for training. From the results, we see several observations: (1) It shows that the White group still outperforms the non-White groups for all the first three experiments. Even without any White subjects in the training set, the accuracy on the White testing set is still higher than those on the testing images in Black and East Asian groups. This suggests that White faces are either intrinsically easier to be verified or face images in the White group of RFW are less challenging. (2) With the decline in the total number of White subjects, the average performance declines as well. In fact, for all these groups, the performance suffers from the decrease in the number of White faces. This indicates that face images in the White groups are helpful to boost the face recognition performance for both White and non-White faces. In other words, faces from the White group benefit the representation learning of global patterns for face recognition in general. (3) Opposite to our intuition, the biasness is lower with less number of White faces, while the data bias is actually increased by adding the unbalancedness to the training set.



(a) Baseline (b) GAC

Figure 3: ROC of (a) baseline and (b) GAC evaluated on all pairs of RFW.

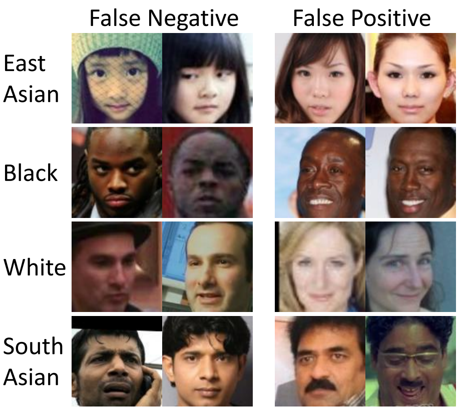


Figure 4: 8 false positive and false negative pairs on RFW given by the baseline but successfully verified by GAC.

Model	Male	Female	Avg (↑)	STD (↓)
Baseline	89.72	79.57	84.64	5.08
GAC	88.25	83.74	86.00	2.26

Table 4: Verification (%) on gender groups of IJB-C (TAR @ 0.1% FAR).

Model	Input Resolution	# Parameters (M)	MACs (G)	Inference (ms)
Baseline	112 × 112	43.58	5.96	1.1
GAC	112 × 112	44.00	9.82	1.4

Table 5: Network complexity and inference time.

4. Additional Experimental Results

To further present the superiority of GAC over the baseline model in terms of bias, we plot Receiver Operating Characteristic (ROC) curves to show the values of True Acceptance Rate (TAR) at various values of False Acceptance Rate (FAR). Fig. 3 shows the ROC performance of GAC and the baseline model on RFW. We see that the curves of demographic groups generated by GAC suggest smaller gaps in TAR at every FAR, which demonstrates the de-biasing capability of GAC. Fig. 4 shows pairs of false positives (two faces falsely verified as the same identity) and false negatives in RFW dataset.

Since IJB-C also provides gender labels, we evaluate our GAC-gender model (see Sec. 4.2 of the main paper) on IJB-C as well. Specifically, we compute the verification TAR at 0.1% FAR on the pairs of female faces and male faces, respectively. Tab. 4 reports the TAR @ 0.1% FAR on gender groups of IJB-C. The biasness of GAC is still lower than the baseline for different gender groups of IJB-C.

5. Network Complexity and FLOPs

Tab. 5 summarizes the network complexity of GAC and the baseline in terms of the number of parameters, multiplier–accumulator, and inference times. While we agree the number of parameters will increase with the number of demographic categories, it will not necessarily increase the inference time, which is more important for real-time applications.

Method	White	Black	East Asian	South Asian	Avg (↑)	STD (↓)
Ada-All	93.22	90.95	91.32	92.12	91.90	0.87
Ada-8	96.25	94.40	94.35	95.12	95.03	0.77
GAC	96.20	94.77	94.87	94.98	95.21	0.58

Table 6: Ablations on the automation module on RFW protocol (%).

6. Ablation on Automation Module

In this section, we ablate GAC with two variants to show the efficiency of its automation module: i) *Ada-All*, *i.e.*, all the convolutional layers are adaptive and ii) *Ada-8*, *i.e.*, the same 8 layers as GAC are set to be adaptive starting from the beginning of the training process, with no automation module (our best GAC model has 8 adaptive layers). As in Tab. 6, with automation module, GAC achieves higher average accuracy and lower biasness than the other two models.

References

[1] 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. 2

[2] Debayan Deb, Lacey Best-Rowden, and Anil K Jain. Face recognition performance under aging. In *CVPRW*, 2017. 2

[3] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. *ECCV*, 2020. 1

[4] 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. 2

[5] 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. 2

[6] 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. 2

[7] 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. 2

[8] 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. 2

[9] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, 2016. 2

[10] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2018. 2

[11] Shankar Setty, Moula Husain, Parisa Beham, Jyothi Gudavalli, Menaka Kandasamy, Radhesyam Vaddi, Vidyagouri Hemadri, J C Karure, Raja Raju, Rajan, Vijay Kumar, and C V Jawahar. Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations. In *NCVPRIPG*, 2013. 2

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

[13] 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

[14] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017. 2