



HAL
open science

Fairer analysis and demographically balanced face generation for fairer face verification

Alexandre Fournier-Montgieux, Michael Soumm, Adrian Popescu, Bertrand Luvison, Hervé Le Borgne

► **To cite this version:**

Alexandre Fournier-Montgieux, Michael Soumm, Adrian Popescu, Bertrand Luvison, Hervé Le Borgne. Fairer analysis and demographically balanced face generation for fairer face verification. 2025. cea-04910990

HAL Id: cea-04910990

<https://cea.hal.science/cea-04910990v1>

Preprint submitted on 24 Jan 2025


HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Fairer Analysis and Demographically Balanced Face Generation for Fairer Face Verification

Alexandre Fournier-Montgieux^{*1} 

alexandre.fourniermontgieux@cea.fr

Michaël Soumm^{*1} 

michael.soumm@cea.fr

Adrian Popescu¹ 

adrian.popescu@cea.fr

Bertrand Luvison¹ 

bertrand.luvison@cea.fr

Hervé Le Borgne¹ 

herve.le-borgne@cea.fr

¹Université Paris-Saclay, CEA, LIST,F-91120, Palaiseau, France

Abstract

Face recognition and verification are two computer vision tasks whose performances have advanced with the introduction of deep representations. However, ethical, legal, and technical challenges due to the sensitive nature of face data and biases in real-world training datasets hinder their development. Generative AI addresses privacy by creating fictitious identities, but fairness problems remain. Using the existing DCFace SOTA framework, we introduce a new controlled generation pipeline that improves fairness. Through classical fairness metrics and a proposed in-depth statistical analysis based on logit models and ANOVA, we show that our generation pipeline improves fairness more than other bias mitigation approaches while slightly improving raw performance.

1. Introduction

Face recognition and verification technologies (FRT and FVT) have seen significant advancements in recent years, with applications ranging from security and surveillance to personal device authentication [22, 65, 72]. However, the widespread adoption of face recognition models has also raised concerns about fairness and potential biases in these systems [14, 42]. Studies have shown that FRT and FVT can exhibit disparities in performance across different demographic groups, particularly along the lines of gender, ethnicity, and age [60, 74]. To address these fairness challenges, researchers have explored various approaches, including the development of demographically diversified datasets [29, 74], and debiasing methods [57, 78]. In parallel, synthetic datasets, generated using computer graphics technique [8, 77] and generative AI models [8, 19, 45, 82], offer the potential to mitigate privacy and copyright issues [34] associated with real datasets [15, 30, 43].

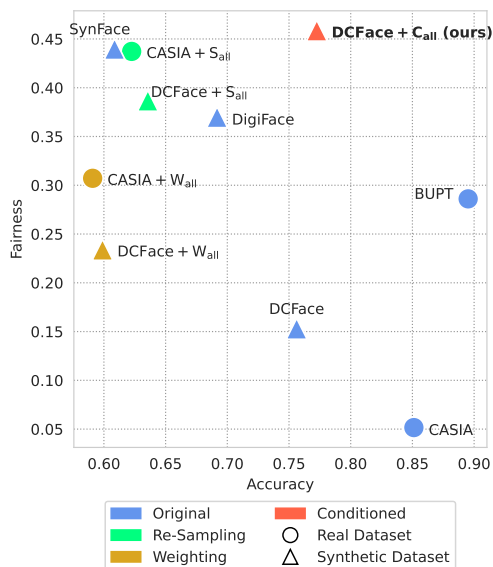


Figure 1. Comparison of the face verification fairness (equalized odds ratio) and micro-average accuracy metrics for models trained with real and synthetic images on the RFW dataset [74]. The proposed pipeline improves the generation fairness and accuracy compared to other synthetic approaches and shows major potential for fairness mitigation compared to real sets. A performance gap in accuracy still subsists with real images.

Nonetheless, the effectiveness of synthetic datasets in improving fairness remains an open question. While existing studies highlighted the potential for generated data to reproduce or even exacerbate the biases present in real datasets [51], most recent works still do not sufficiently analyze the fairness impact on models trained with their synthetic images [8, 45, 54], despite encouraging initiatives [18, 50]. However, these approaches can theoretically provide greater control over data distribution and diversity.

Our first contribution, therefore, introduces a new generation control component based on the existing DCFace pipeline [45]. The resulting approach increases the diversity of sensitive attributes such as `gender`, `ethnicity`, and `age`, and also varies the `pose`, resulting in two new synthetic datasets `DCFace + Cge`, `DCFace + Call`. We compare models trained on these proposed sets with models trained using existing generation datasets, with or without bias mitigation techniques applied.

We employ a range of common metrics to measure fairness. Still, we find them insufficient for an in-depth analysis of the origins of the biases since they do not decorrelate the impacts of the considered attributes. We consequently introduce, as a second contribution, a new analysis approach based on logit regression models that unveils the impact of individual attributes. Furthermore, we use an Analysis of Variance (ANOVA) to examine the relation between attributes and distance in the models' latent space.

As highlighted in [Figure 1](#), our results demonstrate that the proposed controlled generation approach significantly improves fairness metrics while maintaining accuracy. The logit regression and ANOVA analyses draw coherent conclusions and reveal the effectiveness of the proposed controlled generation method in reducing attribute-based biases in both the model predictions and the latent space representations.

The code and data are released to facilitate the adoption of fairness in FRT and FVT: <https://github.com/afm215/FaVGen> (generation) and <https://github.com/MSoumm/FaVFA> (stat. analysis).

2. Related Work

Face verification is a classical yet still open research topic. Following [57], a model is trained to perform face recognition. Then, given a pair of images, the evaluation task is determining whether they belong to the same identity using the trained model as an embedding extractor. A threshold is optimized to separate and predict the positive and negative pairs. Following [52, 53, 74], we advocate for selecting hard negative images to make verification more realistic and consider datasets including difficult negatives to evaluate the models' performance. We also advocate for more efforts to integrate fairness in the verification evaluation process. Fairness evaluation can be improved by designing demographically-diversified verification datasets [29, 53, 74] and integrating demographic metadata in them [60]. Demographic attributes balance deserves particular attention because it is required for analyzing potentially serious discrimination [22, 60].

Real training datasets for face recognition are usually created by scraping a large number of images from publicly available sources [43, 64] and then cleaning them [15, 30, 79] to reduce the number of unrepresentative samples. How-

ever, these datasets face several challenges. First, obtaining subjects' consent at scale is impossible, posing a serious legal challenge when collecting sensitive data such as identified faces. Second, most datasets [15, 30, 79] include copyrighted photos, raising licensing issues. The lawfulness of distributing copyrighted content is a longstanding discussion that applies to other computer vision tasks [55] and was recently revived by the success of foundation models trained with very large datasets [63]. Third, existing large datasets exhibit demographic (gender, ethnicity, age) [53, 60, 74], face characteristics (size, makeup, hairstyle) [4, 5, 69], and visual biases [81], mostly reflecting the sampling bias affecting images datasets [24]. These biases affect underrepresented segments [14, 42, 60] and should be addressed to improve fairness. These problems make the sustainable publication of real datasets very complicated, as proven by the withdrawal of most resources [15, 30, 43] following public pressure [72].

Synthetic datasets have the potential to reduce or remove privacy, copyright, and unfairness issues compared to real datasets [18, 45, 50]. Computer graphics techniques are used in [8, 77] to render diversified face images, and strong augmentations are added to increase accuracy. Most works rely on generative AI, with [82] being an early example that uses dual-agent GANs to generate photorealistic faces. The authors of [54] identify the lack of variability of generated images as a central challenge and propose identity and domain mixup to improve synthetic datasets. Diffusion models were used very recently [45] to create identities and to diversify their samples based on a style bank. Synthetic datasets have the advantage of including fictitious identities, alleviating privacy and copyright issues associated with real-face datasets. However, privacy issues can remain regarding data replication in GANs [25] and diffusion models [67] but can be controlled and mitigated as shown in [9, 20, 21]. When uncontrolled, synthetic datasets are also likely to reproduce and even exacerbate the biases of real datasets in a constrained evaluation setting [51].

Debiasing methods have been proposed to mitigate biases in face verification. One approach is to adapt the verification process to demographic segments. The authors of [57, 70] propose adaptive threshold-based approaches to improve fairness. Another approach is to address ethnicity-related bias by learning disparate margins per demographic segment in the representation space [73, 75, 78] or by suppressing attribute-related information in the model [59]. While technically interesting, these methods are ethically and legally problematic in practice since they assume disparate treatment of human subjects by AI-based systems. We advocate for bias mitigation directly within model training sets, which we show to have a very concrete consequence on model biases.

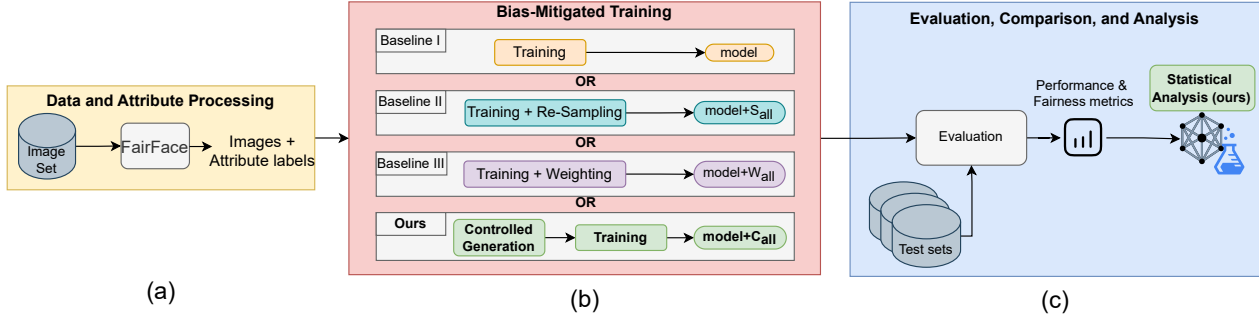


Figure 2. Global pipeline overview for training and evaluating models with the baselines and our proposed generative approach. Critical attributes are collected on image sets (a) that enable using bias mitigation techniques before or during model training (b). Models are then evaluated on FVT evaluation sets (c), and their biases are then analyzed using fairness metrics and our proposed statistical analysis. Contributions of this paper are colored green.

3. Methodology

The overall training and evaluation pipeline (Figure 2) comprises three parts: Part (a) regroups training sets and their attributes. These training sets may or may not be combined with bias mitigation techniques to train models (part (b)). These techniques include our proposed controlled data generation (in green). Finally, as explained in section 2, these models are used in part (c) to perform FVT using the setup of [39, 57]. The results obtained on FAVCI2D [53], RFW [17], and BFW [58] are analyzed in terms of raw performance (accuracy), fairness metrics, and using the statistical approach we introduce in this paper.

Following recent face recognition work [8, 45], we train models using a ResNet50 architecture [35] with a loss designed specifically for this task [44]. We create face recognition models with different training sets. We ensure comparability between these training sets by using the same structure and similar size, compatible with previous studies [8, 54, 79]. They contain 10,000 unique identities and 50 samples per identity.

3.1. Considered Biases

We balance the created datasets for four attributes: ethnicity, gender, age, and pose. The first three are sensitive attributes contributing directly to demographic fairness and are usually employed in the literature [3, 58, 61, 80]. The fourth ensures face appearance variability and augments model performance. ethnicity and gender are attributes associated with each identity. When unavailable in the datasets’ metadata, these attributes are inferred using FairFace [42]. In this case, ethnicity and gender are categorical (Asian, Black, Indian, White) and binary variable (female/male). Since they are supposed to be consistent across the images of the same identity, we mitigate the potential inference errors by averaging the FairFace outputs per identity. Age is also inferred at the image level using

FairFace.

The pose attribute is extracted using the model introduced in [36]. We use face rotation around the pitch, the yaw, and the roll axes (i.e., the rotations around the x , y , and z axes) to characterize pose.

3.2. Proposed Balanced Dataset Generation

Our controlled approach relies on the DCFace [45] generation pipeline. It applies the style of a real picture (style image) to a synthetic face picture (ID image) using a dual-conditioned diffusion model. DCFace combines a single ID with several style images to produce the samples representing each synthetic identity in the training set. The identity-level attributes (ethnicity and gender) are, therefore, controlled by the choice of the ID image. The picture-level attributes (age and pose) are controlled by the choice of the style images.

We thus introduce a joint diversification process on gender, ethnicity, age, and pose attributes. We select a list of ID images generated with DDPM [37] whose joint gender×ethnicity distribution is perfectly balanced. We diversify pose and age by iteratively populating the less-represented age and pose categories of each identity. We also match the demographic segment (gender×ethnicity) of ID and style images to facilitate the loss convergence process. We implemented this matching following initial tests, which showed that convergence is not guaranteed without anything else. We create two versions of the balanced dataset to assess the influence of identity-level and image-level attributes. **DC-Face + C_{ge}** uses only gender×ethnicity, **DC-Face + C_{all}** considers all four attributes.

3.3. Training Set Baselines

We compare **DC-Face + C_{ge}** and **DC-Face + C_{all}** with a representative set of real and synthetic datasets: **CA-**

SIA [79] - real dataset representing celebrities from the IMDB dataset. **BUPT** [75] - real dataset that is balanced for ethnicity. Note that the full version includes more than 1M images. We subsample BUPT to match the structure of other baselines [8, 54, 79]. **SynFace** [54] - synthetic dataset created with a GAN architecture using identity and domain mixup to diversify generated faces. **DigiFace** [8] - synthetic dataset created using rendering technique to obtain diversified representations of faces of each identity. **DCFace** [45] - synthetic dataset generated using the default uncontrolled pipeline of [45].

3.4. Dataset Biases Analysis

We report the attribute diversity a for a dataset \mathcal{D} computed as the normalized entropy applied on the frequency p_{a_i} for the attribute sub-groups $a_i \in [1, m]$.

$$Diversity_a(\mathcal{D}) = -\frac{1}{\ln(N)} \sum_{i=0}^N p_{a_i} \ln(p_{a_i}) \quad (1)$$

Table 1 enables a data-oriented comparison of our datasets and baselines. It highlights the proposed pipeline’s effectiveness and the need for joint attribute balancing to avoid unwanted side effects. For instance, balancing ethnicity and gender alone induces a notable lack of age diversity, and our pose balancing indeed results in more pose diversity. For instance, only balancing on ethnicity and gender reduces age diversity and does not affect pose, while balancing for all attributes results in a better global trade-off.

Attribute	CASIA	BUPT	DigiFace	SynFace	DCFace	DCFace + C _{ge}	DCFace + C _{all}
Gender	1.00	0.93	0.93	0.99	0.99	1.00	1.00
Ethnicity	0.47	0.92	0.65	0.40	0.56	0.93	0.90
Age	0.59	0.71	0.42	0.64	0.64	0.61	0.69
Pose	0.61	0.57	0.67	0.58	0.51	0.51	0.58

Table 1. Inferred diversity for several training datasets. The degree of balance is quantified by the entropy for the considered attributes across the dataset. **Datasets introduced in this paper are in bold.**

3.5. Baseline Debiasing Methods

We compare the proposed dataset bias mitigation pipeline with two classical baseline methods: resampling [6, 10, 46–48, 71, 85] and loss weighting [26, 38, 76]. We apply these common debiasing techniques on imbalanced sets (CASIA and DCFace). The frequency of the considered classes determines an image’s sampling probability and sample weight, which are used in resampling and weighting, respectively. We detail these methods in the supplementary material. We add +S_{ge} and +W_{ge} to initial dataset names for resampling and loss weighting limited to gender and ethnicity. We add +S_{all} and +W_{all} when all attributes are debiased.

4. Toward a Fairer Analysis of FVT evaluation

4.1. Evaluation Sets and Protocol

We use RFW [17], FAVCI2D [53], and BFW [57] in our fairness analysis. We selected the two first face verification datasets because they have sufficient identities per demographic segment for rigorous analysis, and the third one because of its balancing. We provide additional accuracy-oriented results using classical datasets, such as LFW [39], AgeDB [49], and CPLFW [83] in the supplementary material. These datasets are either too small or demographically imbalanced to enable robust fairness in assessment.

Similar to training datasets, we extract FairFace attributes whenever they are not provided. For RFW and BFW, we use the included ethnicity (as well as gender for BFW) attribute since the datasets are already balanced for it. Figure 4 presents a brief description of the pair attributes in the RFW, FAVCI2D, and BFW datasets. While the three datasets have similar balancing on age and pose attributes, they exhibit different characteristics in terms of gender and ethnicity distributions. FAVCI2D has a relatively balanced gender distribution but a skewed ethnicity distribution, with the white ethnicity being the most prevalent. In contrast, RFW has a more balanced representation of ethnicity, with a uniform distribution, like BFW, across African, Indian, Asian, and Caucasian ethnicities, but is unbalanced in terms of gender, unlike BFW. Despite being balanced, BFW has very few identities that might introduce singularity. This can explain the surprising behavior of fairness metrics on this dataset (e.g. CASIA being fair for most metrics). These differences allow for a comprehensive evaluation of face verification models’ fairness and performance across diverse demographic groups, assessing how well the models handle gender and ethnicity variations and identifying potential biases arising from imbalanced training data.

4.2. Fairness and Performance Metrics

To evaluate face recognition performance, we consider the following metrics: **Micro-Average Accuracy** [56] is commonly used for evaluating the overall performance of a face recognition model. It is particularly useful when dealing with unbalanced data, as it gives equal weight to each dataset segment, regardless of the group size. Consequently, the overall accuracy is not biased toward the majority group. **True Match Rate (TMR)**¹, or **TPR**, measures the proportion of actual positive cases that are correctly identified. **False Match Rate (FMR)**, or **FPR**, measures the proportion of negative cases incorrectly identified as positive by the face recognition model. We follow existing face recognition literature [22, 72] and consider FMR as a more critical metric compared to TMR.

¹is equivalent to 1–FNMR

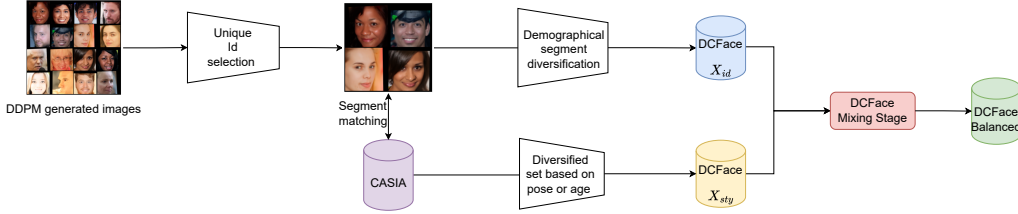


Figure 3. Detailed view of our controlled generation method.

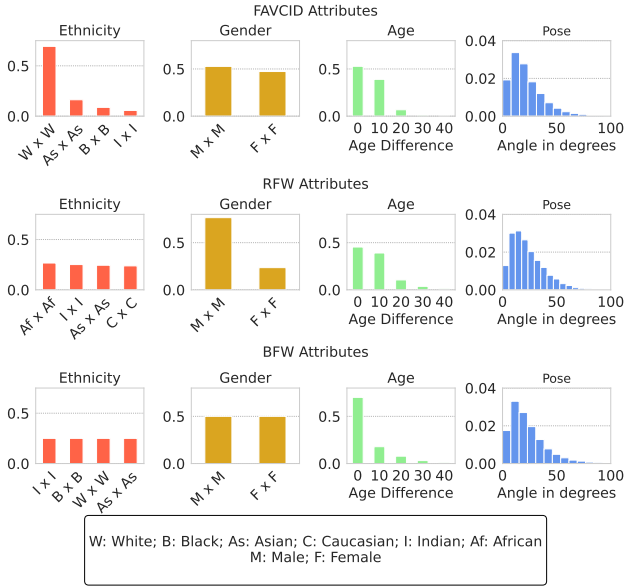


Figure 4. Attribute analysis of the evaluation datasets. Attributes are generated using FairFace [42], except for the gender of BFW, and ethnicity for RFW and BFW included with the datasets.

To evaluate face recognition fairness, we consider the following metrics: **Degree of Bias (DoB)** [28] is the standard deviation of accuracy across different subgroups, which is higher when the performance varies a lot w.r.t each subgroup. However, datasets with low accuracy tend to have a smaller overall variance inherently. Moreover, DoB does not allow for fine-grained error analysis, which is central to understanding performance variations in our case. **Demographic Parity Difference (DPD)** and **Demographic Parity Ratio (DPR)** [1, 2] require that the probability for individuals to receive a positive outcome should be the same across all demographic groups. DPD is the absolute difference between the highest and lowest probability across all subgroups, whereas DPR is the ratio between the lowest and highest. The closer the DPD is to zero and the closer the DPR is to one, the fairer the results are. **Equalized Odds Difference (EOD)** and **Equalized Odds Ratio (EOR)** [1, 32] require that the face recognition model’s TMR and FMR are independent of the demographic groups, thus ensuring consistent accuracy across groups. EOD is

calculated as the maximum absolute difference between the TMRs or FMRs across groups. EOR is the minimum between the ratios of the TMRs and FMRs across groups. The closer the EOD is to zero and the closer the EOR, the fairer the results are.

4.3. Proposed Statistical Analysis Approach

Our statistical analysis pipeline comprises logit regression [7] and Analysis of Variance (ANOVA) [27]. These methods provide complementary insights into the impact of the studied attributes on fairness.

Logit regression [7] models the relation between attributes and the binary outcome of a model. It is a generalized linear model that estimates the probability of a binary outcome based on one or more independent variables using:

$$\ln \frac{\mathbb{P}[y = 1|X]}{\mathbb{P}[y = 0|X]} = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k \quad (2)$$

where y is a target binary variable to explain, X_1, \dots, X_k are the k explanatory variables; and β_0, \dots, β_k are the fitted coefficients. In the fairness analysis context, for an image pair given as input to a face verification algorithm, the binary outcome represents the quantity $\mathbb{1}(y_{pred} = y_{true})$. When properly fitted, the logit regression coefficients represent the impact on the binary outcome for a unit change in the corresponding attribute, holding other attributes constant.

ANOVA [27] is used to determine whether significant differences exist among the means of multiple groups. In fairness analysis, ANOVA can be applied to a continuous variable, such as the distance between face representations in the latent space, to measure the importance of each attribute in explaining the observed variations. By treating the latent space distance as the dependent variable and the attributes as independent variables or factors, ANOVA can partition the total variance in the distances into components attributable to each attribute. Additionally, a quantity named η^2 can be computed for each variable and used to represent the variance the variable explains.

ANOVA identifies the overall importance of attributes in explaining variations in the latent space, while logit regression quantifies the specific impact of attributes on binary identification outcomes. Section 5 presents the detailed application of these methods to the datasets and fairness metrics, along with the interpretation of the results.

	RFW [74]						FAVCI2D [53]						BFW [57]					
	DoB↓	DPD↓	EOD↓	DPR↑	EOR↑	Acc↑	DoB↓	DPD↓	EOD↓	DPR↑	EOR↑	Acc↑	DoB↓	DPD↓	EOD↓	DPR↑	EOR↑	Acc↑
BUPT	30.3	23.6	11.9	68.4	28.6	89.5	38.4	13.0	14.4	75.2	19.3	81.8	25.7	5.5	12.5	88.8	26.1	92.6
CASIA	<u>35.3</u>	19.0	22.0	<u>71.1</u>	5.2	<u>85.1</u>	<u>39.0</u>	21.2	28.5	66.3	16.5	<u>81.1</u>	<u>29.1</u>	<u>9.2</u>	<u>14.9</u>	<u>82.2</u>	1.3	<u>90.3</u>
CASIA + S _{eg}	39.4	11.5	<u>18.8</u>	80.4	29.4	79.9	43.1	<u>15.5</u>	<u>18.7</u>	<u>70.6</u>	31.6	75.1	31.8	10.0	18.9	80.4	11.3	88.0
CASIA + S _{all}	48.2	17.8	24.0	68.8	43.7	62.3	48.6	22.0	24.1	60.3	43.8	61.8	43.5	19.3	33.6	67.8	23.1	74.0
CASIA + W _{eg}	43.5	<u>17.3</u>	22.8	70.2	23.8	74.0	45.3	22.3	21.0	59.2	<u>32.7</u>	71.1	35.4	12.5	18.5	77.0	18.5	84.9
CASIA + W _{all}	49.1	26.7	36.2	54.7	<u>30.7</u>	59.1	49.1	28.1	31.2	47.1	31.0	59.4	46.4	29.6	38.5	52.4	<u>23.5</u>	68.2
SynFace	48.6	13.8	24.9	73.6	<u>44.0</u>	60.9	48.5	22.7	26.4	57.3	37.4	62.0	45.4	20.4	23.2	63.3	<u>36.4</u>	70.7
DigiFace	45.9	15.5	25.6	73.6	37.0	69.2	47.3	21.0	22.2	62.1	40.4	66.0	45.7	16.0	21.1	70.1	44.8	70.0
DCFace	42.7	17.2	32.7	71.4	15.3	75.6	45.1	20.0	18.9	62.8	32.1	71.6	35.4	14.2	21.5	74.4	11.7	85.0
DCFace + S _{eg}	44.0	13.7	36.7	76.5	18.2	72.3	45.9	15.5	21.2	68.4	31.5	69.5	37.2	18.6	29.7	68.3	10.1	82.9
DCFace + S _{all}	48.0	16.7	23.8	69.5	38.7	63.6	48.1	22.1	23.0	58.2	43.8	63.4	42.9	16.8	25.8	68.7	21.8	75.3
DCFace + W _{eg}	44.2	16.7	33.4	70.7	18.9	72.7	46.0	19.2	20.9	62.1	29.9	69.4	36.9	14.6	20.1	72.3	12.5	83.5
DCFace + W _{all}	49.0	19.4	31.6	59.9	23.4	59.9	48.5	23.7	26.0	54.9	36.5	61.8	44.4	24.0	24.3	56.5	27.3	72.8
DCFace + C_{eg}	<u>42.2</u>	<u>12.7</u>	<u>13.7</u>	<u>77.1</u>	41.2	<u>76.4</u>	<u>44.7</u>	<u>14.3</u>	<u>15.6</u>	<u>71.1</u>	66.0	<u>72.4</u>	<u>34.7</u>	11.3	<u>13.8</u>	77.8	23.0	<u>85.7</u>
DCFace + C_{all}	41.6	11.2	<u>14.6</u>	80.3	45.9	77.3	44.5	14.2	14.9	<u>70.9</u>	<u>58.6</u>	72.7	34.2	<u>11.5</u>	13.5	<u>77.5</u>	24.1	86.1

Table 2. Fairness metrics and Micro-average accuracy scores of tested datasets and bias mitigation techniques. Real and synthetic datasets are separated. Groups are defined as a combination of `gender` and `ethnicity`. DPD: Demographic Parity Difference; EOD: Equalized Odds Difference; DPR: Demographic Parity Ratio; Equalized Odds Ratio; Acc: Micro-average Accuracy. The best results for each dataset type are in **bold**, and the second-to-best results are underlined.

5. Results and Analysis

We report and discuss fairness metrics on both FAVCI2D and RFW sets. Statistical and ANOVA analysis is performed on RFW and is reported for FAVCI2D in the supplementary material.

5.1. Performance & Fairness Comparison

We report the fairness metrics and micro-average accuracy for all training approaches on RFW, FAVCI2D, and BFW, both for real and synthetic datasets (Table 2).

Among the real datasets, the model trained on BUPT achieves higher accuracy than models trained on CASIA on RFW, FAVCI2D and BFW. BUPT also gets the best fairness metrics on FAVCI2D and BFW but surprisingly, on RFW, it shows a mitigated behavior, being first in terms of EOD only. Overall on RFW, CASIA+S_{ge} shows the best behavior in terms of fairness (DPR, DPD, EOD), at the cost of 5.2 points of accuracy compared to the original CASIA set. This surprising behavior is not noticed with our in-depth analysis (especially in Figure 5), which draws other conclusions for BUPT model sensitivity, advocating for the usefulness of our analysis approach. Findings are similar for BFW, with CASIA fairness being close to BUPT’s. This surprising behavior might come from the few identities that compose the dataset. BUPT accuracy is also better than CASIA’s, with some variability between the three tested verification datasets. This is probably the result of a different degree of shift between train and verification data.

Among synthetic datasets, the proposed DCFace + C_{ge} and DCFace + C_{all} show the most promising results across the evaluation sets. These balanced variants improve fairness compared to DCFace, the original generation pipeline they build upon. The fairness gains are large for DPD, EOD, DPR, and EOR and less important for DoB. The differences between DCFace + C_{all} and DCFace

+ C_{ge} are small for most fairness metrics, but DCFace + C_{all} provides a mild accuracy gain. The results demonstrate that the proposed balancing pipeline, particularly DCFace + C_{all}, substantially improves fairness metrics across different verification datasets. Importantly, a small accuracy gain compared with the original DCFace dataset is also observed, along with fairness improvement. The models trained with balanced datasets probably benefit from a smaller shift between training and verification datasets, reflected in the micro-average accuracy measured during evaluation. Similar results are obtained in terms of raw accuracy and are reported in the supplementary on five additional verification sets used in prior works [8, 45].

5.2. Logit Model for Bias Quantification

To quantify the biases in face recognition outcomes more precisely, we employ a logit model that estimates the impact of person attributes on face verification model predictions. Hence, we examine the relationship between the studied attribute and the face recognition system’s performance in terms of FMR and TMR. The two logit regressions are:

$$\begin{aligned} \text{(TMR)} \mathbb{1}(\hat{y} = 1 | y = 1) &= \sigma(\beta_0 + \beta_1 \cdot \text{ethnicity} \\ &\quad + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{age} \\ &\quad + \beta_4 \cdot \text{pose}) \end{aligned}$$

$$\begin{aligned} \text{(FMR)} \mathbb{1}(\hat{y} = 1 | y = 0) &= \sigma(\beta_0 + \beta_1 \cdot \text{ethnicity} \\ &\quad + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{age} \\ &\quad + \beta_4 \cdot \text{pose}) \end{aligned}$$

where \hat{y} is the prediction of the model; y is the ground-truth label of the pair; σ is the sigmoid function; `ethnicity` and `gender` are categorical variables implemented with the dummy variable coding [33]; `age` and `pose` are handled as continuous variables.

The logit model coefficients β_k represent the change

Variability in FMR on RFW

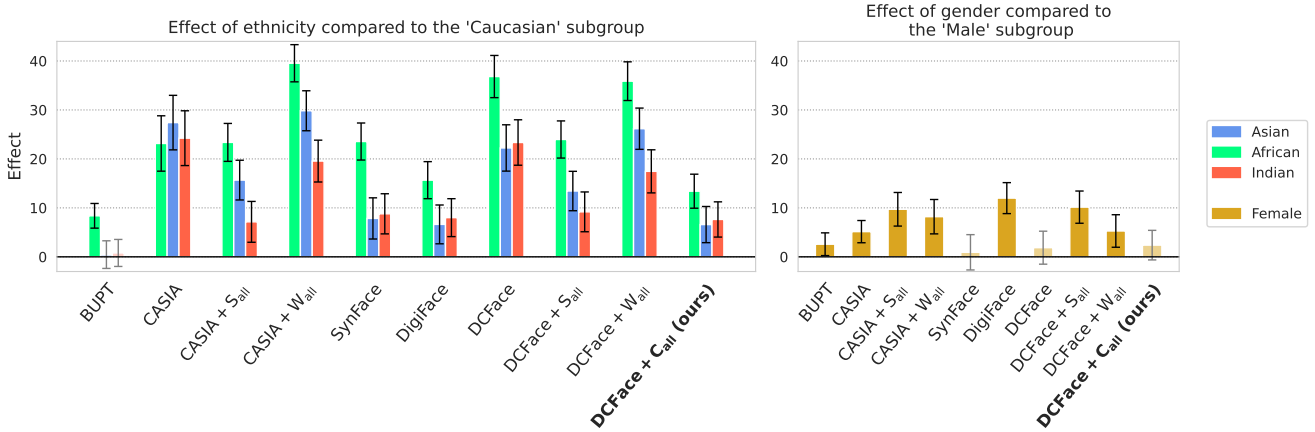


Figure 5. Marginal effect on FMR (lower is better) for each method compared to the unprotected group. Example: "When using CASIA , on average and other things being equal, two people from the African subgroup are 22% more likely to be wrongly misidentified than two people from the Caucasian subgroup". Non-significant effects are shown in transparency. Our controlled generation reduces biases of DCFace more effectively than other bias mitigation techniques.

in the log odds of the binary outcome (e.g., false positive or true positive) for a unit change in the corresponding attribute, holding other attributes constant. The unit change is computed with respect to the unprotected group (Caucasian for ethnicity and Male for gender), which is the reference level in the dummy coding. Since the β_k are not easily interpretable by themselves, we then compute the mean marginal effects of each attribute, i.e., how much the TMR or FMR change when we shift from the unprotected value to a protected one (for instance Male to Female). Since we control for all other variables simultaneously, this effect can be interpreted as an effect with all other attributes kept constant. Therefore, the marginal effect estimates the effective demographic biases while accounting for confounding factors.

Figure 5 presents the logit model results for the ethnicity and gender attributes on RFW, showing the computed marginal effects on FMR. The marginal effects are calculated relative to each attribute’s unprotected reference group. The higher the bar, the higher the bias against the protected subgroup. For example, when using DCFace, our analysis shows that the FMR for the African subgroup is 35 points higher than for the White subgroup, independently of the other considered attributes. The addition of re-weighting does not affect this bias, while re-sampling reduces it to 22 points. Our method further reduces it to 12 points. Concerning gender bias, despite decreasing the bias for ethnicity, re-sampling increases the bias for gender. The proposed controlled generation reduces biases for ethnicity while keeping the bias in gender non-significant. The results of the logit model on TMR and on FAVCI2D are provided in the supp. material.

The logit model results provide valuable insights into the fairness implications of different face recognition methods and datasets. By comparing the marginal effects across attributes and methods, we identify the extent and nature of biases of each approach. The significantly smaller marginal effects observed in Figure 5 shows our controlled dataset generation reduces biases compared to the original DCFace dataset and baseline mitigation techniques. The interpretation of the logit model results highlights the disparities in face recognition performance across different attribute subgroups, showing the importance of considering fairness in the development and evaluation of face recognition systems and the need for effective bias mitigation strategies.

5.3. ANOVA on Latent Space

The variation of the performance and fairness metrics across demographic segments can be seen as a consequence of the variability in the distribution of feature vectors in the model’s latent space. Therefore, we utilize ANOVA to investigate the influence of personal attributes on the distances in this latent space. In our case, the groups are defined by the person’s attributes, such as gender, age, and ethnicity, and the explained variable is the distance between face representations in the latent space. We use the sum of squares computed during ANOVA to extract the η^2 associated with each attribute. Each η^2 value represents the impact of the variable on the distance variance in the latent space. The η^2 of each attribute sum to the R^2 of the ANOVA, i.e. the total variance explained by the model.

Figure 6 shows the result of ANOVA on the distances in the latent space of the RFW dataset, both on the positive and negative pairs. As expected, the explained variance

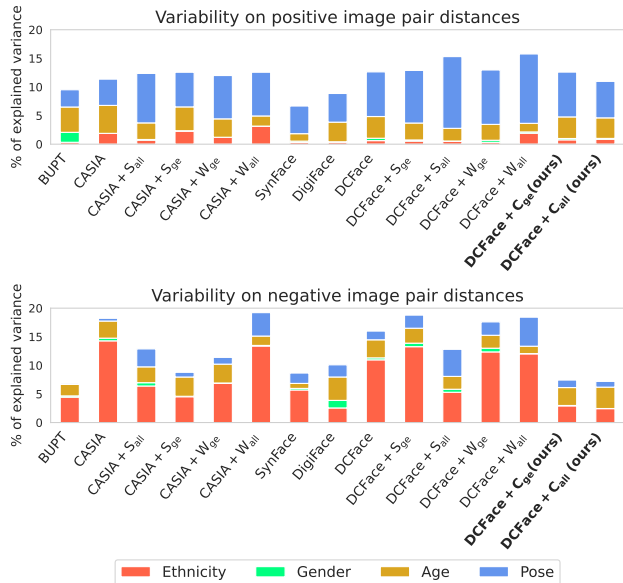


Figure 6. ANOVA results on RFW: total height corresponds to R^2 , the explained variance by the variables. Each bar is decomposed into multiple η^2 , i.e. the individual contributions to the variance.

on the positive pairs is generally smaller than the explained variance of the negative pairs, since two images of different people are likely to have more variability than two images of the same person. Moreover, the total $R^2 = 0.18$ of the ANOVA shows that 18% of variance in the distances in the latent space can be attributed solely to the considered people’s attributes. `pose` has the strongest influence on the positive pairs, a finding explained by the strong pose variability for the same person. However, neither `ethnicity` nor `gender` play a big role, meaning that across demographic segments, the spread of the latent vectors of a single person is very similar. This is expected since the training loss tries to bring closer the latent vectors of the same individual, who has only one `ethnicity` and `gender` value.

On the negative pairs, `ethnicity` is the attribute having the highest impact on the latent vectors. This means that the distances for negative pairs are much higher for some demographic segments than others. This result quantifies how much the demographic imbalance translates into the geometry of the latent space. Confirming previous works [42, 61] with another approach, our analysis shows a significant impact of the demographic attributes on the spread of the latent vectors. Once more, the impact of the proposed datasets, DCFace + C_{all} and DCFace + C_{age}, on the η^2 shows the effectiveness of our controlled generation. By contrast, traditional training strategies such as re-sampling and loss-weighting are not as good at mitigating the biases in the latent space.

6. Limitations

Attribute inference tools are needed to obtain demographic attributes but introduce prediction errors. FairFace is widely used in the field [45, 62, 75] and could be improved, particularly for a finer-grained `ethnicity` detection. The statistical bias analysis is sensitive to the variability and size of evaluation sets. Consequently, it should only be applied to datasets having sufficient samples for each demographic segment to obtain significant results. This discards using classical datasets, such as LFW [39], for fairness analysis. We present a dataset-balancing pipeline combining several attributes and implement it for DCFace [45], a recent and competitive face generator. The pipeline can be adapted to other generators, such as IDiff-Face [12]. `ethnicity` and `gender` are the most sensitive attributes, and their balancing focuses on seed ID images needed to drive identity generation in most existing methods.

7. Conclusion

We addressed FVT fairness by evaluating the performance and bias of models trained on various real and synthetic datasets. We proposed a novel controlled generation approach to create balanced synthetic datasets, DCFace + C_{ge} and DCFace + C_{all}, which prioritize attribute diversity. Our experiments demonstrated that models trained on balanced datasets significantly improved face verification fairness metrics while maintaining competitive accuracy. The proposed analysis based on logit regression and ANOVA revealed that the controlled generation method effectively reduces attribute-based biases in both model predictions and latent space representations. It also highlights a persistent disparity in fairness across all considered approaches, which penalizes the African subgroup in particular.

Our findings have important implications for developing fairer and more inclusive FVT systems. By demonstrating the effectiveness of attribute balancing in synthetic data generation and providing a comprehensive evaluation framework, we advocate for more efforts in addressing bias issues in computer vision applications. Future research could explore integrating our approach with other bias mitigation techniques and investigate the generalizability of our findings to other computer vision tasks and datasets.

Acknowledgement This work was partly supported by the SHARP ANR project ANR-23-PEIA-0008 in the context of the France 2030 program and the STARLIGHT project funded by the European Union’s Horizon 2020 research and innovation program under grant agreement No 101021797. It was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council.

References

- [1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. 5
- [2] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 09–15 Jun 2019. 5
- [3] Vítor Albiero and Kevin Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are, 08 2020. 3
- [4] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. How does gender balance in training data affect face recognition accuracy? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 2
- [5] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2021. 2
- [6] Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Y. A. Hawalah, and Amir Hussain. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, 4:7940–7957, 2016. 4, 13
- [7] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009. 5
- [8] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023. 1, 2, 3, 4, 6, 13, 19
- [9] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8001–8010, 2023. 2
- [10] Kwabena Ebo Bennin, Jacky Wai Keung, Passakorn Phannachitta, Akito Monden, and Solomon Mensah. Mahakil: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. *IEEE Transactions on Software Engineering*, 44:534–550, 2018. 4, 13
- [11] Mantas Birškus. Glasses Detector, 3 2024. 13
- [12] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fuzzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023. 8
- [13] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Thanaruk Theeramunkong, Boonserm Kijssirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 475–482, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. 13
- [14] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 1, 2
- [15] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 1, 2
- [16] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002. 13
- [17] Jean-Rémy Conti, Nathan Noiry, Stephan Clemençon, Vincent Despiegel, and Stéphane Gentric. Mitigating gender bias in face recognition using the von mises-fisher mixture model. In *International Conference on Machine Learning*, pages 4344–4369. PMLR, 2022. 3, 4
- [18] Ivan Deandres-Tame, Rubén Tolosana, Pietro Melzi, R. Vera-Rodríguez, Minchul Kim, C. Rathgeb, Xiaoming Liu, A. Morales, Julian Fierrez, J. Ortega-Garcia, Zhizhou Zhong, Y. Huang, Yuxi Mi, Shouhong Ding, Shuigeng Zhou, Shuai He, Lingzhi Fu, Heng Cong, Rongyu Zhang, Zhihong Xiao, Evgeny Smirnov, Anton Pimenov, A.P. Grigorev, Denis Timoshenko, K. Asfaw, C. Low, Hao Liu, Chuyi Wang, Qing Zuo, Zhixiang He, Hatef Otroschi Shahreza, Anjith George, A. Unnervik, Parsa Rahimi, Sébastien Marcel, Pedro C. Neto, Marco Huber, J. Kolf, N. Damer, Fadi Boutros, Jaime S. Cardoso, Ana F. Sequeira, A. Atzori, G. Fenu, M. Marras, Vitomir vStruc, Jiang Yu, Zhangjie Li, Jichun Li, Weisong Zhao, Zhen Lei, Xiangyu Zhu, Xiao-Yu Zhang, Bernardo Biesseck, Pedro Vidal, Luiz Coelho, Roger Granada, and David Menotti. Second edition frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data. 2024. 1, 2
- [19] Moreno D’Inca, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Improving fairness using vision-language driven image augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4695–4704, 2024. 1
- [20] Perla Doubinsky, Nicolas Audebert, Michel Crucianu, and Hervé Le Borgne. Multi-attribute balanced sampling for disentangled GAN controls. *Pattern Recognition Letters*, 162:56–62, 2022. 2
- [21] Perla Doubinsky, Nicolas Audebert, Michel Crucianu, and Hervé Le Borgne. Wasserstein loss for semantic editing in the latent space of gans. In *International Conference on Content-Based Multimedia Indexing*, 2023. 2
- [22] Ho Daniel E., Emily Black, Maneesh Agrawala, and Fei-Fei Li. Domain shift and emerging questions in facial recognition technology. *HAI Policy Brief*, 2020. 1, 2, 4

- [23] Hartig F. Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models., 2018. [14](#)
- [24] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022. [2](#)
- [25] Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix M Martinez. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6701–6710, 2021. [2](#)
- [26] K. Ruwani M. Fernando and Chris P. Tsokos. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, 2022. [4](#), [14](#)
- [27] James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. *An introduction to statistical learning: with applications in R*. Springer, 2013. [5](#)
- [28] Sixue Gong, Xiaoming Liu, and A Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*, pages 330–347, 2020. [5](#)
- [29] Patrick J Grother, Mei L Ngan, Kayee K Hanaoka, et al. Face recognition vendor test part 3: demographic effects. Technical report, National Institute of Standards and Technology, 2019. [1](#), [2](#)
- [30] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. [1](#), [2](#)
- [31] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In De-Shuang Huang, Xiaoping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. [13](#)
- [32] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. [5](#)
- [33] Melissa A Hardy. *Regression with dummy variables*. Number 93. Sage, 1993. [6](#)
- [34] Jules. Harvey, Adam. LaPlace. Exposing.ai, 2021. [1](#)
- [35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [3](#)
- [36] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022. [3](#)
- [37] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. [3](#)
- [38] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5375–5384, 2016. [4](#), [14](#)
- [39] Gary B. Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014. [4](#), [8](#), [13](#)
- [40] I. Hupont and Carles Fernández Tena. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. *IEEE International Conference on Automatic Face & Gesture Recognition*, 2019. [15](#)
- [41] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 336–351. PMLR, 11–13 Apr 2022. [13](#)
- [42] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. [1](#), [2](#), [3](#), [5](#), [8](#)
- [43] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016. [1](#), [2](#)
- [44] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022. [3](#), [13](#), [14](#)
- [45] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dc-face: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023. [1](#), [2](#), [3](#), [4](#), [6](#), [8](#), [13](#), [19](#)
- [46] Felix Last, Georgios Douzas, and Fernando Bação. Over-sampling for imbalanced learning based on k-means and smote. *ArXiv*, abs/1711.00837, 2017. [4](#), [13](#)
- [47] Daniel Lehmann and Marc Ebner. Subclass-based under-sampling for class-imbalanced image classification. In *VISIGRAPP*, 2022. [4](#), [14](#)
- [48] Xu-ying Liu, Jianxin Wu, and Zhi-hua Zhou. Exploratory under-sampling for class-imbalance learning. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 965–969, 2006. [4](#), [14](#)
- [49] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. [4](#), [13](#)
- [50] Pedro C. Neto, Eduarda Caldeira, Jaime S. Cardoso, and Ana F. Sequeira. Compressed models decompress race biases: What quantized models forget for fair face recognition. *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2023. [1](#), [2](#)

- [51] Malsha V Perera and Vishal M Patel. Analyzing bias in diffusion-based face generation models. *arXiv preprint arXiv:2305.06402*, 2023. [1](#), [2](#)
- [52] P Jonathon Phillips, J Ross Beveridge, Bruce A Draper, Geof Givens, Alice J O’Toole, David Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer. The good, the bad, and the ugly face challenge problem. *Image and Vision Computing*, 30(3):177–185, 2012. [2](#)
- [53] Adrian Popescu, Liviu-Daniel Ștefan, Jérôme Deshayes-Chossart, and Bogdan Ionescu. Face verification with challenging imposters and diversified demographics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3357–3366, 2022. [2](#), [3](#), [4](#), [6](#)
- [54] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021. [1](#), [2](#), [3](#), [4](#), [13](#), [19](#)
- [55] Jenny Quang. Does training ai violate copyright law? *Berkeley Tech. LJ*, 36:1407, 2021. [2](#)
- [56] Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation, 2021. [4](#)
- [57] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020. [1](#), [2](#), [3](#), [4](#), [6](#)
- [58] Joseph P. Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Transactions on Image Processing*, 32:4365–4377, 2023. [3](#), [15](#)
- [59] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Flac: Fairness-aware representation learning by suppressing attribute-class associations. *arXiv preprint arXiv:2304.14252*, 2023. [2](#)
- [60] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Towards fair face verification: An in-depth analysis of demographic biases. *arXiv preprint arXiv:2307.10011*, 2023. [1](#), [2](#)
- [61] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Towards fair face verification: An in-depth analysis of demographic biases, 2023. [3](#), [8](#)
- [62] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Towards fair face verification: An in-depth analysis of demographic biases, 2023. [8](#)
- [63] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. [2](#)
- [64] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. [2](#)
- [65] Neil Selwyn, Mark Andrejevic, Gavin J. D. Smith, Xin Gu, and Christopher O’Neill. Facial recognition technology: key issues and emerging concerns, 1 2023. [1](#)
- [66] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. [13](#)
- [67] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. [2](#)
- [68] Chakkrit Kla Tantithamthavorn, A. Hassan, and Ken ichi Matsumoto. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. *IEEE Transactions on Software Engineering*, 46:1200–1219, 2018. [13](#)
- [69] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *arXiv preprint arXiv:2103.01592*, 2021. [2](#)
- [70] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *2020 8th international workshop on biometrics and forensics (IWBF)*, pages 1–6. IEEE, 2020. [2](#)
- [71] Chih-Fong Tsai, Wei-Chao Lin, Ya-Han Hu, and Guan-Ting Yao. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, 477:47–54, 2019. [4](#), [14](#)
- [72] Richard Van Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834):354–358, 2020. [1](#), [2](#), [4](#)
- [73] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020. [2](#)
- [74] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019. [1](#), [2](#), [6](#)
- [75] Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8433–8448, 2021. [2](#), [4](#), [8](#), [20](#)
- [76] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 7032–7042, Red Hook, NY, USA, 2017. Curran Associates Inc. [4](#), [14](#)
- [77] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. [1](#), [2](#)
- [78] Zhanjia Yang, Xiangping Zhu, Changyuan Jiang, Wenshuang Liu, and Linlin Shen. Ramface: Race adaptive mar-

- gin based face recognition for racial bias mitigation. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 1, 2
- [79] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2, 3, 4, 20
- [80] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P. Breckon. Measuring hidden bias within face recognition via racial phenotypes. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3202–3211, 2022. 3
- [81] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018. 2
- [82] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. *Advances in neural information processing systems*, 30, 2017. 1, 2
- [83] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018. 4, 13
- [84] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 13
- [85] Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. Oversampling method for imbalanced classification. *Comput. Informatics*, 34:1017–1037, 2015. 4, 13
- [86] Zheng Zhu et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 13

Supplementary Material

A. Parameters for training and generation

For training the face classifier, we use the Adaface training pipeline [44]. We apply the same augmentations, crop, and low-resolution augmentations, for all training sets, with an exception on DigiFace, where we also use the augmentation of the authors to reach optimal performances. We perform the training on 4 GPUs with a batch size of 256 (i.e. 64 per GPU), the optimizer is the standard SGD with a learning rate of 0.1 and a momentum of 0.9. We use as a scheduler a multi-step learning rate decay whose milestones are the epochs 12,20,24 and the decay coefficient is 0.1. The training loss is that of Adaface [44]. The margin parameter m is set to 0.4, and the control concentration constant h to 0.333 as recommended by [44]. On each training set, the training lasts 60 epochs.

For generating the DCFACE set and its variants, we use the generation pipeline of [45]. We impose the X_{id} image and the X_{sty} to be of the same demographic group as we found that mismatching is likely to induce non-convergence of the resnet50 model when training on the resulting dataset (in particular when mismatching in gender). Randomly sampling the style image within the CASIA dataset thus results in a non-decreasing loss of the ResNet network. Within the code of [45], there is a sampling strategy we haven't tested: combining DDPM images with the closer CASIA faces. This approach was and still is, unfortunately, non-usable due to incomplete critical files² Moreover, this strategy is not mentioned in the original paper and, since it combines similar CASIA and DDPM faces in a resnet100 latent space, it seems to be in contradiction with what is stated within the ID Image Sampling subsection of [45]. We thus chose to ignore this strategy, our study being primarily an analysis of fairness and improvement research in this regard.

For all methods, similarly to what the original paper did, we introduce variability within the considered DDPM X_{id} pictures by using a similar F_{eval} model as in [45]. However, one should be aware that the Cosine Similarity Threshold might vary depending on the training of the F_{eval} network. We used the network trained on [86] provided by the Adaface Github repository and found 0.6 as an effective threshold to filter similar images. We also get rid of faces wearing glasses with the following solution [11].

B. Performance in Accuracy on other sets

In addition to FAVCI2D, BFW, and RFW, we report in Table 3 the raw accuracy results on 5 common evaluation

²The provided center_ir_101_adaface_webface4m_faces_webface_112x112.pth file doesn't have a required "similarity_df" field. Also, the dface_3x3.ckpt file doesn't seem to store the following property: recognition_model.center.weight.data

Verif. dataset	Real dataset				Synthetic datasets		
	CASIA	BUPT	SynFace	DigiFace	DCFace	DCFace + C_{ge}	DCFace + C_{all}
LFW	99.46	99.55	87.28	94.88	98.13	98.24	98.25
CFP-FP	94.87	90.03	67.01	83.4	80.92	80.03	81.28
CPLFW	90.35	85.98	64.91	76.61	79.94	79.32	80.17
AgeDB	94.95	94.3	61.78	78.26	87.96	86.77	86.53
CALFW	93.78	94.38	73.53	79.78	90.39	90.6	90.03
RFW	86.38	90.35	64.3	72.73	76.95	78.51	79.5
FAVCI2D	82.77	81.81	61.19	67.17	72.84	73.31	73.73
BFW	89.3	92.48	70.08	77.27	84.47	85.45	88.53
AVG	91.48	91.11	68.76	78.76	83.95	84.03	84.75

Table 3. Raw Accuracy obtained for the different used sets on 8 datasets including five commonly used datasets in addition to BFW, RFW and FAVCI2D

sets used in prior work on the FR task [8, 44, 45, 54]: (1) Labeled Faces in the Wild (LFW) [39], the reference dataset for the task (2) CALFW [84], a version of LFW with a larger age variability, (3) CPLFW [83], a version of LFW with pose variability, (4) AgeDB [49], a dataset designed for maximizing age variability, and (5) CFP-FP [66] that is designed for pose variability.

Raw accuracy differs from the micro accuracy reported on the paper. Micro accuracy gives the same importance to each demographic segment, whereas raw accuracy performs a simple mean across all images, without any distinction.

Table 3 confirms the performance gain of DCFace + C_{all} over the original generation pipeline: The generation pipeline slightly improves accuracy for four of these datasets (+0.12, +0.36, +0.23, and +0.89 for LFW, CFP-FP, CPLFW, and FAVCI2D) and slightly degrades performance for the other two (-1.43 and -0.36 points for Age-DB and CALFW). On the balanced sets, (i.e. RFW and BFW) the pipeline induces important gains in accuracy (+2.55 for RFW and +4.06 for BFW).

C. Bias Mitigation techniques details

We provide implementation details about the baselines, re-sampling, and loss weighting used to compare with our approach.

C.1. Re-sampling

Data re-sampling balances class distribution within training data by employing strategies other than the default uniform sampling. These strategies can consist of over-sampling the data from the under-represented classes and/or under-sampling majority classes [41, 68].

Oversampling [6, 10, 46, 85] increases the number of samples by replicating existing data. However, duplicating data by sampling the several times can lead to over-fitting. On tabular data, interpolating techniques such as SMOTE and its variants [13, 16, 31] can be used in order to tackle this overfitting issue. Still, such approaches are not trivial and more costly for non-tabular data such as images.

Undersampling, on the other hand, consists in the re-

duction of the majority classes so that their representativity matches the underrepresented classes. [47, 48, 71]. The main drawback of such an approach is that it results in unused data, which is not an optimal setup.

Here we use Re-Sampling as a baseline for bias mitigation by combining over-sampling and under-sampling. Specifically, for each attribute a with values a_j , we count n_j , the number of images with value a_j . We then assign a weight $w_j = 1/n_j$ to each image sharing value a_j . For each image x_i , we compute its weight w_i as the product of the weights of all attributes associated with the image. The sampling probability for each image is calculated as $p_i = w_i / \sum_k w_k$. At each beginning of a training epoch, we sample N images according to the probability distribution $\{p_i\}$, where N is the size of the original dataset.

Note that this approach, coupled with the set of random image augmentations used during training, should mitigate to a certain extent the mentioned limitations of both over-sampling and under-sampling.

C.2. Loss Weighting

Loss weighting tries to adapt the loss scale depending of the characteristics of the sample. This weighting can be linked to the difficulty of the sample as done implicitly by the Adaface Loss [44], which can be induced by the class imbalance or in our use case, by the corresponding attributes representativity. A common way to weight the loss is to use the same weights computed in subsection C.1, i.e. using the invert of the frequency/count [26, 38, 76]. We thus use the same weights w_i for weighting the loss. The weights are normalized batch-wise to ensure the same order of gradient amplitude. The loss of the batch is defined as:

$$\mathcal{L}(x_1, \dots, x_K) = \frac{\sum_k w_k \mathcal{L}(x_k)}{\sum_k w_k} \quad (3)$$

where $\mathcal{L}(x_k)$ is the sample-wise loss for image x_i .

D. Diagnostics on the regressions

To be valid, a linear regression needs to satisfy a few properties, mainly:

- **Correct specification:** The model is correctly specified, meaning all relevant variables are included, and no irrelevant variables are included.
- **Normal distribution of errors:** While not strictly necessary for estimation, the assumption that errors are normally distributed allows for valid hypothesis testing and the construction of confidence intervals.
- **Zero conditional mean (exogeneity):** The expected value of the error term is zero for any given value of the independent variables. This implies that the independent variables are uncorrelated with the error term.

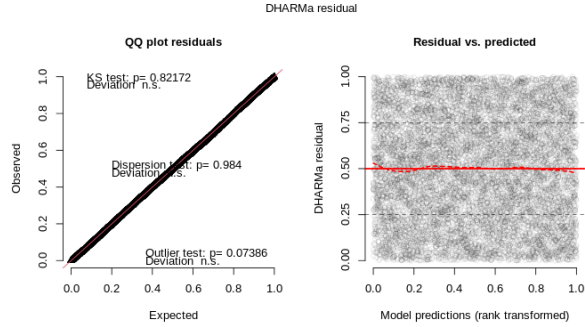


Figure 7. QQ-plot of residuals and Residual vs. predicted plot: logit model is adapted and log-odds are linear in the variables.

- **Homoscedasticity:** The variance of the error term is constant across all levels of the independent variables.

For a generalized linear model, such as the logit model, these assumptions are not possible to verify strictly due to the non-linearity of the model. Therefore, we use the DHARMA package [23] in R to run diagnostics on our models and verify the validity of our regressions. DHARMA uses simulation-based residuals. It creates new data from the fitted model and then calculates the empirical cumulative density function for each observation. This approach allows for standardized residual calculation even for non-normal distributions like in logit models.

The package provides several diagnostic plots:

- **QQ-plot of residuals:** Checks for overall deviations from the expected distribution (Figure 7-left).
- **Residual vs. predicted plot:** Helps detect heteroscedasticity and nonlinearity (Figure 7-right).
- **Residual vs. predictor plots:** Useful for identifying problems with specific predictors (similar to exogeneity) (Figure 8).
- **Overdispersion Test:** helps to identify if there’s more variation in the data than expected under the binomial distribution (Figure 9).
- **Zero-inflation Test:** check for an excess of zeros or ones (Figure 10).

Here, we will show the diagnostics only for the model DCFace + C_{all} on RFW, but diagnostics graphs are constant across all tested models on both test datasets.

E. Statistical Analysis on FAVCI2D

We present here the results of our statistical analysis on FAVCI2D. Be aware that while the metadata of this dataset contains gender information, it doesn’t provide ethnicity.

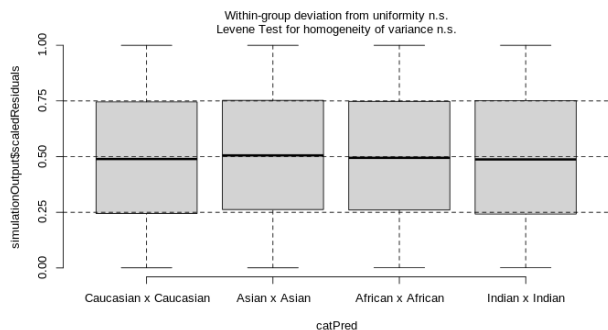


Figure 8. Residual vs. predictor plots: exogeneity is verified.

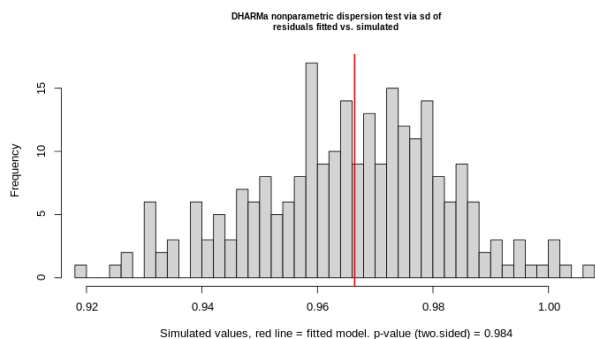


Figure 9. Overdispersion Test: Correct Specification and no auto-correlation.

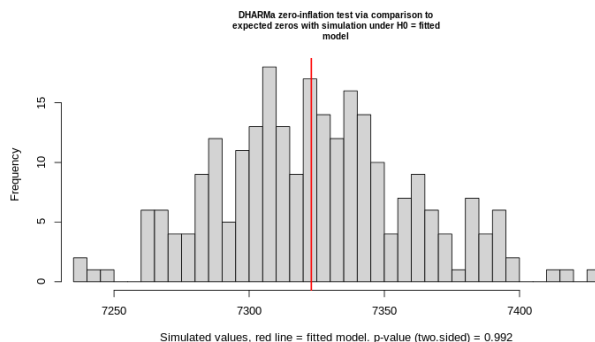


Figure 10. Zero-inflation Test: the model correctly predicts the probability of the outcome.

We infer it using FairFace. We consider the prediction of FairFace robust enough to compute macro metrics such as the Diversity metric of the main paper however for a finer study such as ours, it might introduce some uncertainty due to model prediction error (Table 4). With that in mind, we still get consistent results for the effects of demographic attributes on the models (Figure 11). Our approach shows even more insensitiveness on FAVCI2D than BUPT, by contrast to the results obtained on RFW. The increase of the

BUPT-trained model’s sensitivity with regard to the inferred labels on FAVCI2D might come from the dataset balancing done on the same labeling system as RFW. Results obtained regarding the TMR (Figure 12) and FMR are coherent with the idea that models tend to predict positive outcomes for certain protected ethnic sub-groups. They thus have a high recall for these groups (high TMR and high FMR). With the gender provided by the metadata, we can thus confirm the impact of the balancing on fairness relative to this attribute. While most of the models are sensitive to gender, the model trained on DCFace + C_{all} DCFace has close to no sensitivity for this attribute, both being close to perfectly balanced concerning gender.

Figure 13 shows the result of ANOVA on the distances in the latent space of the FAVCI2D dataset, both on the positive and negative pairs. The results are coherent with the ANOVA computed on RFW. It furthermore highlights the sensitivity of some models’ latent space to gender, while our balancing approach allows for more insensitivity about demographic attributes.

F. Statistical Analysis on BFW

To tackle the issue of the lack of metadata, in addition to BFW, other alternatives exist such as BFW [58] and DemogPairs [40]. While these datasets provide some ground-truth metadata, they are composed of significantly fewer identities compared to datasets like FAVCI2D or RFW. This is a limitation of our analysis: Having too few identities might bring instability within Anova or marginal effect studies due to redundancy. We report the results obtained with BFW on as similar number of pairs as RFW and FAVCI2D (24k), meaning every single identity appears in around 30 evaluated pairs. The impact of the number of identities within benchmarking should be studied in future works as this might affect the obtained analysis of performance and fairness.

Figure 16 shows the ANOVA analysis performed on BFW. As before, on the negative image pairs, our conditional generation methods greatly reduces the variance explained by the sensitive attributes.

Figures 15 and 14 present the marginal effects of the attributes, respectively, on TMR and FMR. As we see, the fairness gain mostly comes from a fairer FMR between ethnicities: the FMR of the Asian and Black subgroups are 8 and 12 points higher than for the White subgroup in the original DCFace, and become non-significant with DCFace + C_{all} . For the TMR, however, just as for RFW, becomes slightly more unfair between ethnicities. Still, as shown in Table 2 of the paper, on all fairness metrics except EOR, our method outperforms the other synthetic data approaches on BFW.

ethnicity	Black	White	East-Asian	Indian	Latino-Hispanic	Middle-Eastern	South-Asian
Prediction accuracy	0.863	0.777	0.784	0.724	0.581	0.631	0.641

Table 4. FairFace model accuracy when inferring on the Fairface validation set. Available Metadata only provides the race7 variable ground truth while we are considering the race variable (whose values are White, Black, Asian, and Indian). The robustness of the model for this latter should be thus greater.

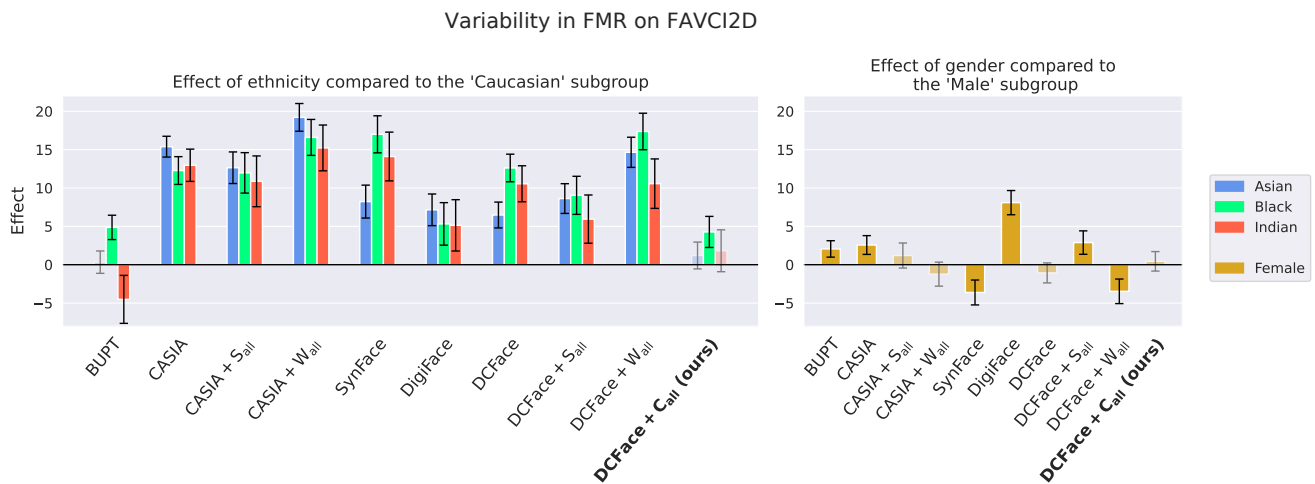


Figure 11. Marginal effect on FMR (lower is better) for each method compared to the unprotected group. Analysis done on FAVCI2D

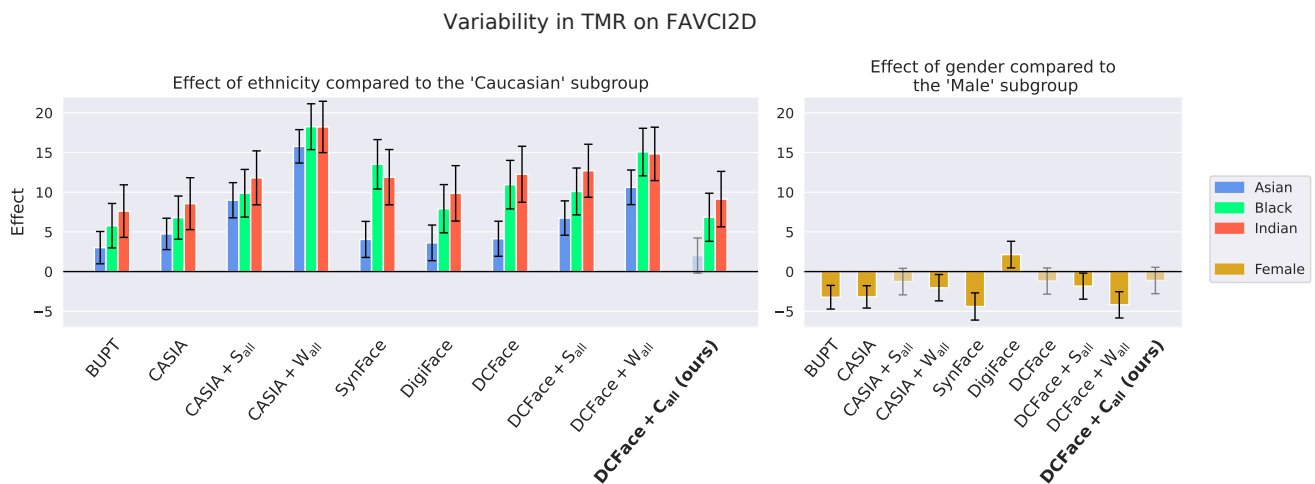


Figure 12. Marginal effect on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on FAVCI2D

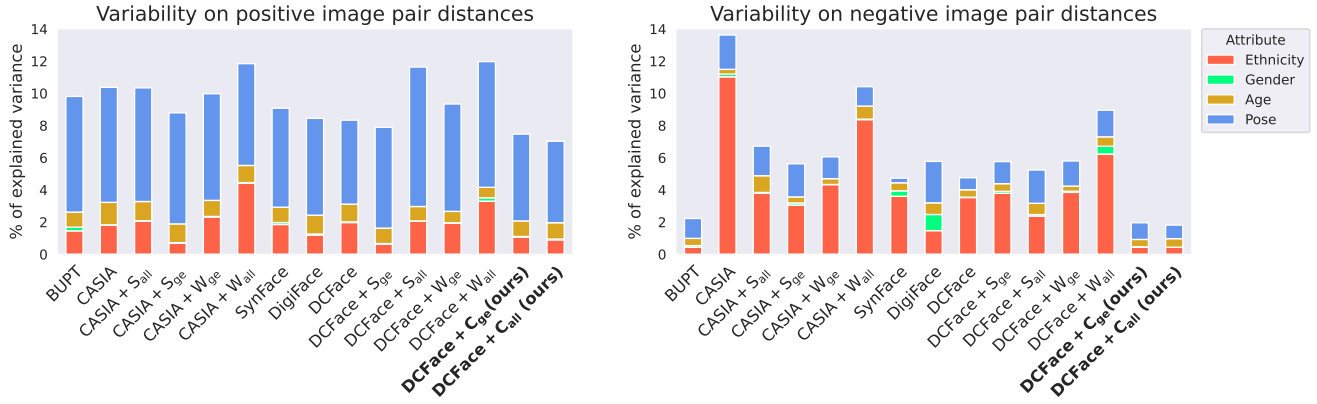


Figure 13. ANOVA results on FAVCI2D : total height corresponds to R^2 , the explained variance by the variables. Each bar is decomposed into multiple η^2 , i.e. the individual contributions to the variance

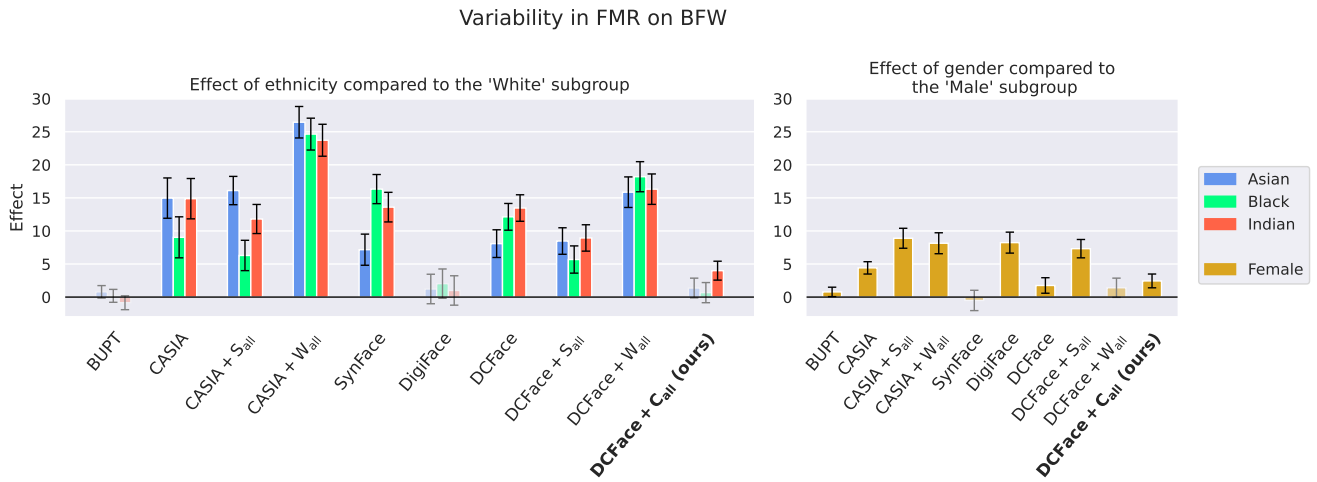


Figure 14. Marginal effect on FMR (lower is better) for each method compared to the unprotected group. Analysis done on BFW

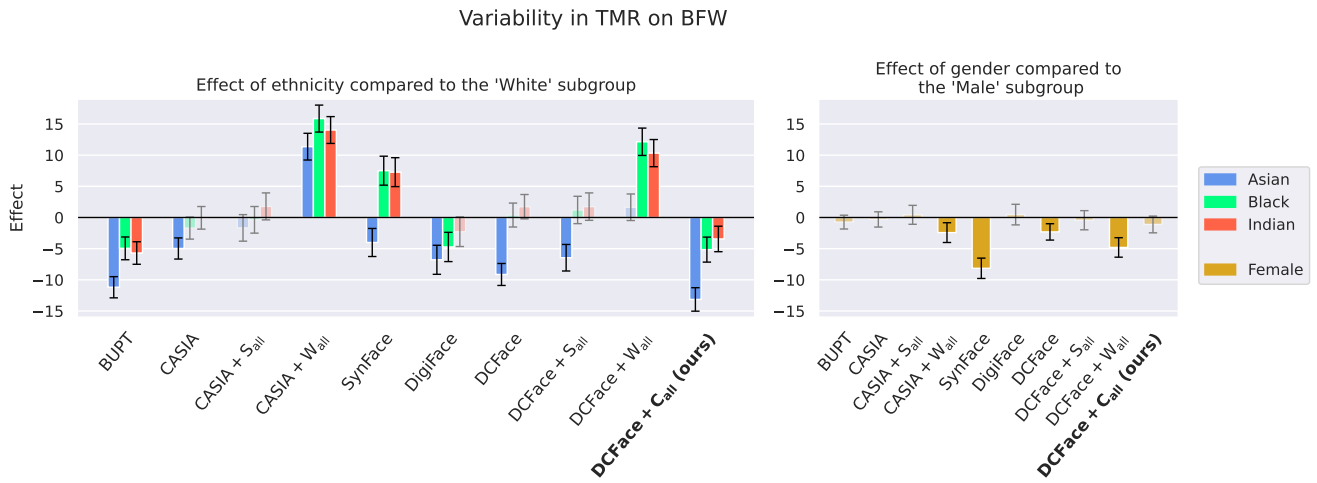


Figure 15. Marginal effect on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on BFW

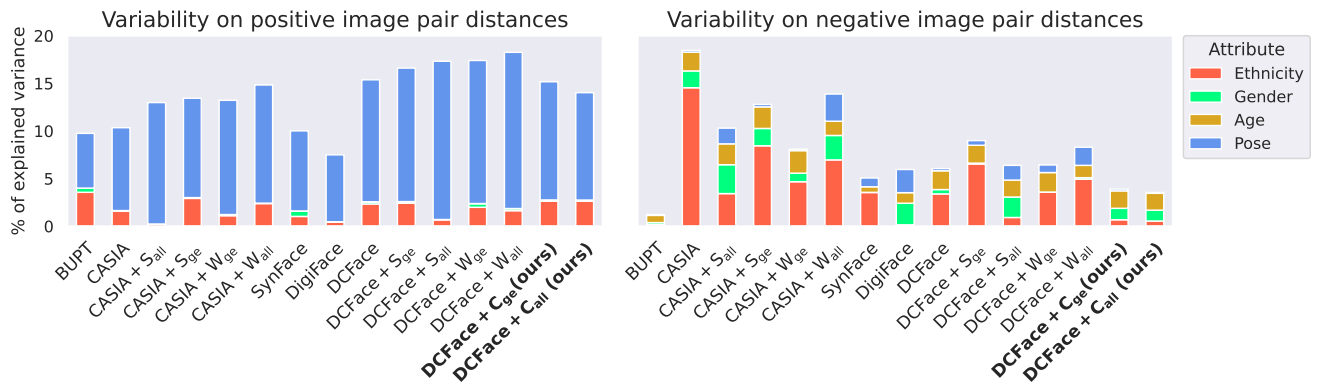


Figure 16. ANOVA results on BFW: total height corresponds to R^2 , the explained variance by the variables. Each bar is decomposed into multiple η^2 , i.e. the individual contributions to the variance

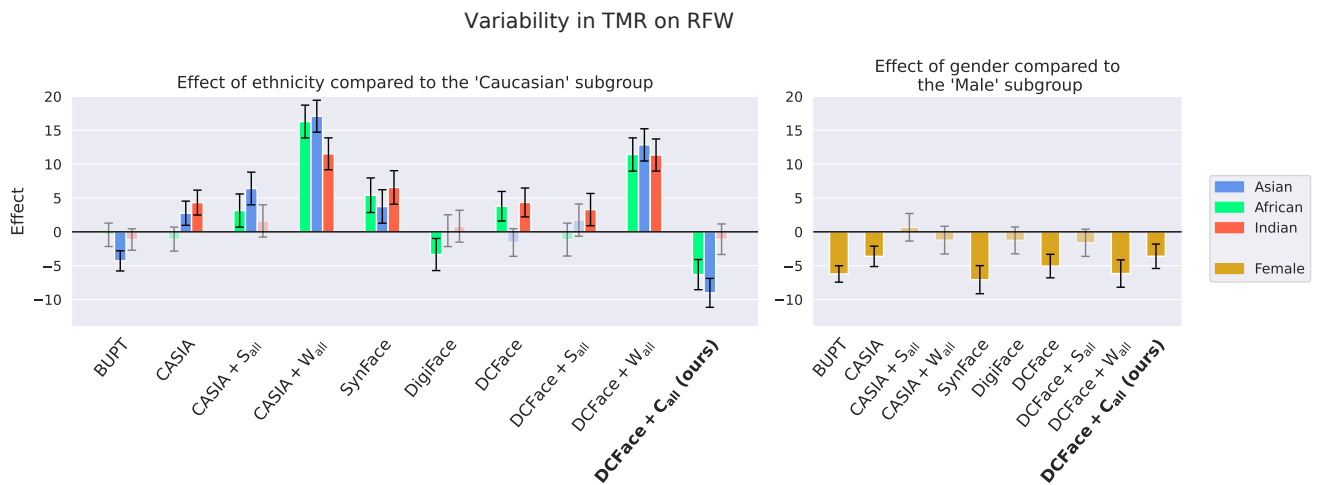


Figure 17. Marginal effects on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on RFW

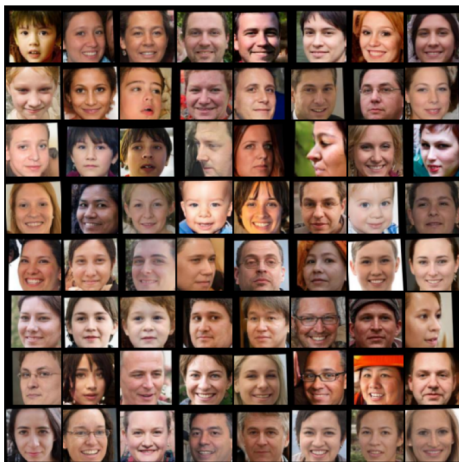
G. Datasets Images examples



(a) Examples of images within our proposed DCFace + C_{all} approach. We notice a greater diversity of images.



(b) Examples of images generated with the original DCFace [45] pipeline



(c) Examples of images generated with the SynFace pipeline [54]



(d) Examples of images within the DigiFace dataset [8]



(e) Examples of images within the CASIA dataset [79]



(f) Examples of images within the BUPT dataset [75]