Bias and Diversity in Synthetic-based Face Recognition

Marco Huber[1,2], Anh Thi Luu[1], Fadi Boutros[1], Arjan Kuijper[1,2], Naser Damer[1,2]
[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany
Email: marco.huber@igd.fraunhofer.de

Abstract

Synthetic data is emerging as a substitute for authentic data to solve ethical and legal challenges in handling authentic face data. The current models can create real-looking face images of people who do not exist. However, it is a known and sensitive problem that face recognition systems are susceptible to bias, i.e. performance differences between different demographic and non-demographics attributes, which can lead to unfair decisions. In this work, we investigate how the diversity of synthetic face recognition datasets compares to authentic datasets, and how the distribution of the training data of the generative models affects the distribution of the synthetic data. To do this, we looked at the distribution of gender, ethnicity, age, and head position. Furthermore, we investigated the concrete bias of three recent synthetic-based face recognition models on the studied attributes in comparison to a baseline model trained on authentic data. Our results show that the generator generate a similar distribution as the used training data in terms of the different attributes. With regard to bias, it can be seen that the synthetic-based models share a similar bias behavior with the authentic-based models. However, with the uncovered lower intra-identity attribute consistency seems to be beneficial in reducing bias.

1. Introduction

Recently, training face recognition (FR) models on synthetic data and using FR models trained on synthetic data has gained attention and importance [15]. Motivated by legal and ethical issues in some jurisdictions (e.g. the European Union [68]) regarding using and sharing authentic face images without consent, synthetic data might be a substitute due to its remarkable quality and similarity to authentic data. Additionally, several large face datasets such as the widely-used MS-Celeb-1M [31], VGGFace2 [16] or DukeMTMC-reID [49] are not available anymore from an official source. However, the existing accuracy gap between authentic and synthetic-based FR requires novel solutions to train suitable well-performing FR models [12, 14].

In authentic FR systems and datasets, the presence of bias, unfair behavior by the systems based on demographic (e.g. gender, age, ethnicity) or non-demographic (e.g. head-pose, accessories) attributes, has been studied extensively in the past [3, 17, 67] and mitigation strategies to increase data diversity and reduce unfair behavior have been explored [3, 30, 66]. Having fairer models is of high importance for biometric applications as the decisions of FR systems are often of high impact on the individuals' lives, e.g. in access control or law enforcement identification.

For synthetic FR data and models trained on synthetic data, unfair behavior is still an unexplored area [15]. The diversity and bias in synthetic data and models trained on synthetic data are of special interest as it provides novel opportunities but also new problems. A major advantage of synthetic data is that new data can be generated depending on certain attributes. For example, a lack of variety with respect to a certain ethnicity can be compensated by synthetic images of these underrepresented ethnicities [41]. While there are works [2, 7] that have shown that the distribution of training data plays an exclusive role for gender bias, there are works [73, 74] that showed that the distribution of training data plays a major role in ethnicity bias. However, a drawback of synthetic data is, that there is no factual or self-reported ground truth regarding specific attributes as the person depicted does not exist in reality and definite statements cannot be made. This is especially important as studies have shown that there are gaps between self-reported and genomic ancestry ethnicity [44], self-reported and observed ethnicity [57] and biases and errors in human age estimation [18, 32, 69].

In this work, we investigate the demographic and non-demographic diversity of existing synthetic face datasets and also the bias in existing synthetic-based FR models. To do this, we investigate the distribution of head pose, gender, ethnicity, and age in three recent synthetic datasets as well as on the authentic datasets used to train the synthetic data generators using attribute predictors. We also investigate the bias of the synthetic-based FR models regarding gender, age, ethnicity, and pose in comparison to an authentic data-based baseline model. Our results show, that the synthetic face generators generate data similarly, in terms of diversity, to their training data and that the models trained on synthetic models suffer higher bias than the model trained on authentic data, motivating new solutions for more diverse and better synthetic data generators or bias mitigation techniques in synthetic FR models.

2. Related Work

2.1. Synthetic Data in Face Biometrics

In recent years, utilizing synthetic data for face biometric tasks has become quite popular [12, 14, 42, 47] as ethical and legal requirements have to be met for biometric data using, sharing, and collection in several jurisdictions (such as the European Union [68]). Although the legal requirements may vary culturally and geographically, the recent ethics guidelines of established international venues, such as the International Conference on Computer Vision (ICCV)[1] or the Conference on Computer Vision and Pattern Recognition (CVPR)[2], also require special considerations when publishing or using databases containing personal data. Moreover, the collection and annotation of large-scale authentic data are expensive and time-consuming and might lead to datasets with low diversity. On the issue of identity relation between synthetic data and the authentic data used to train the generators, Boutros et al. [12] have shown close to no relation.

Recently, a number of works [12, 14, 42, 47] proposed the use of privacy-friendly synthetic data to train FR models as an alternative to privacy-sensitive authentic data. Qiu et al. [47] proposed a synthetic-based FR model, namely SynFace, that utilized synthetic face images generated by attribute-conditional GAN, DiscoFaceGAN [22], to train FR models. Each of the synthetic identities in the proposed approach is generated by fixing the identity condition and randomizing the attribute conditions i.e. pose, illumination, and expression. SynFace [47] also analyzed the performance gap between synthetic and authentic images as training data and identified poor intra-class variations and the domain gap as possible reasons for verification performance differences. To mitigate this, SynFace proposed to use identity and domain mixup, where identity mixup refers to interpolating between two identities and domain mixup refers to interpolating between authentic and synthetic data in the training data. USynthFace [14] also utilized attribute-conditional GAN to train an FR model in an unsupervised manner. UsynthFace proposed a contrastive learning framework that is trained to maximize the distance between two augmented synthetic images of the same synthetic instance. To achieve that, USynthFace proposed a large set of geometric and color transformations as well as a GAN-based augmentation for their contrastive learning framework. SFace [12] and IDNet [42] proposed synthetic-based FR models based on class conditional GANs. Each of the synthetic identities in SFace and IDNet is generated by fixing the class label and randomizing the latent code. SFace proposed to train StyleGAN-ADA under class conditional settings on CASIA-WebFace. SFace also proposed to improve the synthetic-based FR performances by transferring the knowledge from a model trained on authentic data to the model trained on synthetic data without compromising the authentic identities. Unlike SynFace and UsynthFace, the intra-class variations in SFace are not limited to a predefined set of attributes. However, the generated data by SFace suffers from low identity distinctiveness. IDNet very recently extended SFace by incorporating identity information in the GAN training, aiming at enhancing identity discrimination in the generated data. DigiFace-1M [5] proposed synthetic-based FR, where the synthetic images are generated by rendering digital faces using a computer graphics pipeline. Each synthetic identity in DigiFace is created by randomizing the facial geometry, texture, and hairstyle. Although DigiFace-1M achieved relatively competitive FR verification performances, the generation process is computationally expensive, and the generated images do not match the quality and realistic appearance of authentic images. Most recently, IDiff-Face [11] and ExFaceGAN [13] were proposed, leading to more realistic face variations and huge advancement in the performance of the synthetic-based FR.

Besides FR, synthetic face data has also been used for other biometric-related tasks, such as 3D face reconstruction [48], presentation attack detection [27], morphing attack detection [19], FR model quantization [9] or face image manipulation [46]. In contrast to Fu et al. [29], who investigated the diversity in terms of face image quality of synthetic face images and how they relate to the training images, we investigate the diversity and bias regarding specific demographic and non-demographic attributes.

2.2. Bias in Face Recognition

Exact definitions of bias, implications, and its causes vary between sources. A common understanding is that it relates to differences in performance ratings that are influenced by a particular sub-population [50]. Several studies showed that the recognition performance of females is weaker than the performance of male faces when using FR trained on authentic data [1, 2]. Regarding age, studies analyzing the impact of age demonstrated a lower biometric performance for children, than for adults [20, 62]. Research investigating the impact of ethnicity showed that faces of under-presented ethnicities perform worse [34]. In a comprehensive study, Terhörst et al. [67] expanded bias also to non-demographics attributes, such as expression, pose, or illumination.

[1] https://iccv2023.thecvf.com/ethics.guidelines-362600-2-25-26.php

[2] https://cvpr.thecvf.com/Conferences/2024/EthicsGuidelines

To the best of our knowledge, no work so far investigated the bias in synthetic FR models and the diversity of the generated data regarding demographic and non-demographic attributes.

Besides FR, bias can also be observed in other biometric tasks, such as presentation attack detection [26, 28], face image quality assessment [65], biometric systems explanations [37], or face detection [45].

3. Investigation Methodology & Setup

In this section, we describe our investigation methodology and the investigation setup. We start with the approach we use to analyze the diversity of authentic FR datasets and synthetic FR datasets. We describe the utilized datasets, synthetic face generators, and attribute predictors. After that, we describe our approach to investigate the bias of FR models trained on synthetic faces, including the FR models and evaluation methods used.

3.1. Diversity Investigation

To investigate the diversity of the datasets, we use different high-performing attribute predictors to predict the three demographics attributes gender, age, and ethnicity, and the non-demographic attributes head pose. We then report the different distributions of the attributes of the different authentic and synthetic datasets.

The investigation of the diversity of the authentic datasets is also of special interest, as the investigated datasets are used to train the generative models that then create the synthetic datasets. Investigating this allows possible insights into how the distribution changes based on the utilized training data, as some might assume that the generative model in general follows the distribution that has been used to train it.

3.1.1 Attribute Predictors

Since the investigated datasets do not provide human-labeled attributes, labeling large datasets is expensive, and there is also no factual self-reported ground truth in synthetic images regarding specific attributes, we utilize attribute predictors to get automatic attribute labels. In our experiments, we limit ourselves to the most investigated attributes gender, age, ethnicity, and head-pose [24]. The results based on four additional well-established [4, 6, 25, 52] attribute estimator (including another non-demographic attribute face emotion) based on an open source project [59, 60] are provided in the supplementary material due to space. For each of the attributes, we utilize different well-performing attribute predictors.

The **gender predictor** [53] is based on VGG-16 architecture [61] and was pre-trained on ImageNet [55] and fine-tuned on the IMDb-Wiki dataset [53]. It achieved a classification accuracy of 88.50% on the Balanced-Faces-in-the-Wild (BFW) [51] dataset. We decided to limit our prediction classes to two genders $(male, female)$, being aware that there are more than two genders people identify themselves with.

The **age predictor** we utilized a support vector machine (SVM) trained on feature embeddings extracted using ElasticFace-Arc [8] from the Adience [56] dataset. It achieved a mean accuracy of $60.51\% \pm 2.28$ in a five cross-fold evaluation setup on Adience [56], which is comparable to other works on the hard Adience dataset [10, 53, 63, 64]. The Adience dataset provides 8 classes, which are defined as: $(0, 2)$, $(4, 6)$, $(8, 12)$, $(15, 20)$, $(25, 32)$, $(38, 43)$, $(48, 53)$, and $(60, 100)$.

As the **ethnicity predictor**, we also utilize an SVM trained on feature embeddings extracted using ElasticFace-Arc [8]. The images to create the feature embeddings are taken from BUPT-Balancedface [74] dataset, which provides a large set of nearly equally distributed ethnicities. The datasets distinguished between the ethnicities $African/Black$, $Asian$, $Caucasian/White$, and $Indian$. The SVM is trained on randomly selected 10% of the data while keeping the distribution of the ethnicities equal. To evaluate the performance, we test on 1,300 images also randomly selected from BUPT-Balanceface while ensuring that the identities from the training set are not part of the test set. On this test set, we achieved an accuracy of 90.91%.

As the **head-pose predictor** we use Hopenet [54], which is one of the publicly available top-performing head pose estimators. It uses a multi-loss convolutional neural network and predicts intrinsic Euler angles (yaw, pitch, and roll) of the head pose. For simplification, we only consider yaw and evaluate on the Annotated Facial Landmark in the Wild (AFLW) [43] dataset, which results in a mean absolute error of the predictor of 8.26. This implies that the predicted yaw angle differs from the real angle by about 8.26. Since we mainly care about the general head pose, we divide the obtained yaw into five classes: $0°$ (frontal), $22.5°$, $45°$, $67.5°$, and $90°$ (profile) based on the yaw angle.

3.1.2 Diversity Evaluation Metric

To measure and evaluate the diversity of the authentic and synthetic training data, we report the overall data distribution in terms of gender, age, ethnicity, and head pose as predicted by our attribute predictors. Since in some cases, different samples of the same identity might have conflicting attribute predictions, we also want to gain insights into the intra-identity distribution of the attributes. Differences in these attributes might be natural (e.g. aging process) or due to less distinct features. We propose the novel *Intra-Identity Attribute Consistency Ratio (IIACR)*. The IIACR is calculated as:

$$IIACR = \frac{1}{N} \sum_{n=1}^{N} \frac{max_{\alpha \in S}(m_{n,a})}{m_n}, \qquad (1)$$

where N is the number of identities, S is the set of different classes for an attribute, α is a class of S, m_n is the total amount of images of an identity n, and $m_{n,a}$ is the amount of images of an identity n that belongs to class α. The IIACR, therefore, provides insights into the consistency of an attribute within the images of an identity. For some attributes such as gender or ethnicity, this value is expected to be 1, as individuals rarely change their gender or ethnicity. In synthetically generated face images, where the creators rather aim at preserving or creating the synthetic identity than maintaining the similar attribute for the different samples of a synthetic identity, this might vary more often. For other attributes, such as head pose and age, a lower value might be favorable, as this indicates more variety in the face data. In our experiments, we randomly select 1% of the identities and calculate the IIACR based on all available images of the randomly selected individuals. On USynthFace-400k, we only applied it to the GAN-augmented samples, as the geometric and color transformation had been done online.

3.2. Bias Investigation

To investigate the bias in FR models trained on synthetic or authentic data, we analyze performance differences depending on the investigated demographic attributes (gender, cross-age, ethnicity) and non-demographic attributes (head pose). To obtain these performance differences, we evaluate different datasets that make a distinction regarding the specific attributes, e.g. we compare between the verification accuracy of a female subset to that of a male subset. On all benchmarks, we follow the provided evaluation protocol to produce reproducible results.

Finally, we combine the results from the diversity and consistency analysis with the results from the bias analysis to investigate the influence of the synthetic or authentic training data on the different verification performances.

3.2.1 Bias Evaluation Datasets & Metric

For the evaluation in terms of bias, we utilize six different datasets. The different datasets provide different labels to create attribute-based subsets and are often used in bias studies [50, 73]. For the evaluation in terms of gender bias, we report the performance on Balanced-Faces-in-the-Wild (BFW) [51]. To investigate the ethnicity bias, we evaluate the different models on BFW [51] as well as Racial-Faces-in-the-Wild (RFW) [73]. To investigate the performance difference on images with different ages, we compare LFW [35] (smaller age gap) with Cross-Age LFW [76] (larger age gap) as they are based on the same data but instead of random comparisons (LFW), the Cross-Age LFW dataset consist of genuine and imposter pairs with higher age gaps. For the head-pose performance difference, we compare the performance on CFP-FF (frontal-frontal) [58] to CFP-FP (frontal-profile) [58] which are also based on the same data but provide a different pose evaluation scenario, were the pairs of CFP-FF both show frontal images, while in the CFP-FP datasets, one face image is a profile image.

To investigate the bias in synthetic FR models, we report the verification accuracy on the different subsets following the defined protocol of each dataset, as well as the mean accuracy (mAcc) and the standard deviation (STD), similar to other bias analyses works [36, 71, 72]. Furthermore, we also report the Skewed Error Ratio (SER) [71]. Error skewness is computed by the ratio of the highest error rate to the lowest error rate among different attributes and is therefore calculated as:

$$SER = \frac{max_a Err(a)}{min_b Err(b)} \qquad (2)$$

where a, b are classes of the investigated attribute.

3.3. Face Recognition Models & Datasets

To investigate the bias in FR models trained on synthetic data, we utilized three different recently proposed models trained on their associated synthetic training data. The utilized models are SFace$_{synth}$ [12], SynFace [47], and USynthFace [14]. We chose these models because they provide state-of-the-art synthetic FR and are of different natures in their approach. All models used are publicly available and have been released by the authors of the respective works.

As a **Baseline**, we investigate the bias in FR models trained on authentic data, ResNet50 [33] trained on CASIA-WebFace [75] with CosFace loss [70]. This model is considered in our evaluation as it was used by the utilized synthetic-based FR models [12, 47] as a baseline for comparing their verification performances with the model trained on authentic data.

SFace$_{synth}$ [12] is a model trained on the SFace-60 dataset. The model architecture is ResNet-50 [33] trained with CosFace loss [70]. During the training phase, SFace proposed to transfer the knowledge from a model trained on authentic data, CASIA-WebFace [75]. This aims at guiding the synthetic FR model to learn to produce feature representations that are similar to the ones learned by the model trained on authentic data. It should be noted that the training process of SFace does not include any authentic data. The SFace training dataset consists of 634,320 images of 10,572 identities (60 images per identity). The images have been generated by a StyleGAN2-ADA [38] conditionally trained on CASIA-WebFace [75].

The authentic **CASIA-WebFace** [75] dataset used to train the FR baseline and the SFace generative model consists of 494,414 images of 10,575 different identities. The images in CASIA-WebFace have been collected semi-automatically from the web. The authors of the datasets did not make any statement on the diversity of their dataset, but it is regularly used to train FR systems [8,21,40], especially when compared to synthetic-based FR [12,47].

SynFace [47] was trained by the authors on the Syn_10k_50 dataset using identity-mixup. The backbone architecture is ResNet-50 [33] trained with ArcFace loss [21]. **Syn_10k_50** [47] consists of 500,000 synthetic images of 10,000 different identities. Syn_10k_50 utilized DiscoFaceGAN [22] to generate the synthetic face images. DiscoFaceGAN is an attribute conditional GAN model trained on the FFHQ dataset [39]. Synthetic identities in Syn_10k_50 are generated by fixing the identity condition and randomly sampling latent variables from the standard normal distribution for expression, pose, and illumination. The FFHQ dataset contains 70k images collected from Flickr and encompasses variation in ethnicity, age, image background, and accessories [39].

USynthFace [14] is an unsupervised FR model trained with unlabeled synthetic data. USynthFace FR model architecture is ResNet-50 [33] trained with contrastive learning. Similar to the SynFace model, USynthFace utilized a DiscoFaceGAN trained on FFHQ to generate the synthetic dataset, USynthFace-400k. The USynthFace-400k dataset consists of 400,000 images of 400,000 synthetic identities.

In the diversity analysis we analyze two authentic datasets CASIA-Webface [75], FFHQ [39] that had been used during the generator training, and the three different synthetic datasets, SFace-60 [12], Syn_10k_50 [47] and USynthFace-400k [14] that has been created by the generators. In the bias analysis, we investigate the bias of the authentic baseline model and the three synthetic models, SFace$_{synth}$, SynFace, and USynthFace.

4. Results

In this section, we present the results of our investigation. First, we provide the results of our data diversity investigation starting with the distributions of gender, ethnicity, age, and head pose on the five different datasets. Later on, we present and discuss the results of the intra-identity attribute consistency analysis, to evaluate how consistent or diverse the different attributes are in an authentic dataset and synthetic dataset. The diversity distributions for the predictions of four additional attribute estimators are provided in the supplementary material.

4.1. Data Diversity

4.1.1 Attribute Distribution

The distribution of the gender attribute is visualized in Figure 1 and the percentages are also shown in Table 1. The distribution of male and female individuals in the dataset is close to even, while the authentic FFHQ dataset shows more female individuals, the other authentic dataset, CASIA-WebFace shows more male individuals. SFace-60, the synthetic dataset which has been created utilizing a generative model trained on CASIA-Webface revealed a higher imbalance regarding gender, which might indicate that the generative model tends to create samples from its majority class regarding gender distribution. This is especially interesting, as the synthetic identities of the SFace-60 dataset have been created based on the original CASIA-WebFace identities. The generative model, therefore, at least in some cases, flipped the gender of some images or identities in their synthetic counterparts from female to male.

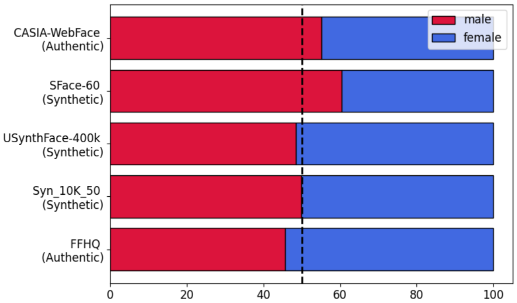

Figure 1. **Gender Distribution:** The dotted line indicates an equal distribution. While the FFHQ dataset, USynthFace-400k dataset, and Syn_10K_50 dataset are nearly balanced, the imbalance of the gender distribution in the SFace-60 dataset increased in contrast to its generator training data, CASIA-WebFace. This might indicate a bias regarding generating individuals from the majority class.

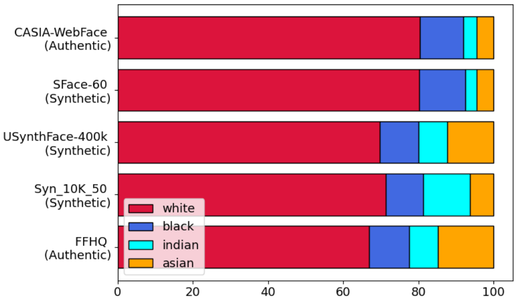

Figure 2. **Ethnicity Distribution:** All datasets show a high ethnicity imbalance. The synthetic datasets seem to inherit the general distribution regarding ethnicities from the authentic training set to train the generators. (FFHQ was training data for the Syn_10K_50 generator and the USynthFace-400k generator, CASIA-WebFace was used to train and generate the SFace-60 dataset.

The ethnicity distribution is shown in Figure 2 and the numerical values in Table 1. The figure shows that there is a highly imbalanced distribution in the authentic, but also in the synthetic face datasets. White/Caucasian individuals are highly over-represented while other ethnicities are under-represented. One can also note that the synthetic datasets follow, to some degree, the overall distribution of the respective generator training data of the generator. The generative model trained on the slightly less imbalanced FFHQ dataset led to less imbalanced synthetic datasets USynFace-400k and Syn_10K_50 in contrast to the more imbalanced

Dataset	Gender		Ethnicity			
	Male	Female	African/Black	Asian	White/Caucasian	Indian
CASIA-WebFace (auth.)	55.23	44.77	11.49	4.45	80.45	3.61
SFace-60 (syn.)	66.52	39.48	12.20	4.44	80.20	3.16
USynthFace (syn.)	48.56	51.44	10.44	12.26	69.64	7.66
Syn_10K_50 (syn.)	49.80	50.20	9.93	6.21	71.40	12.46
FFHQ (auth.)	45.62	54.38	10.70	14.85	66.87	14.85

Table 1. **Distribution of Gender and Ethnicity in %:** Similar to Figure 1, the values show that the synthetic SFace-60 is more imbalanced than its authentic origin dataset CASIA-WebFace. Regarding the ethnicity distribution, the synthetic datasets inherit the general balance from their authentic origin dataset (see also Figure 2).

Dataset	Age								Head Pose				
	(0,2)	(4,6)	(8,12)	(15,20)	(25,32)	(38,43)	(48,53)	(60,100)	0°	22.5°	45°	67.5°	90°
CASIA-WF. (auth.)	0.01	0.31	1.79	5.21	67.74	19.02	2.97	2.96	51.01	36.12	8.14	3.65	1.06
SFace-60 (syn.)	0.00	0.18	1.10	5.16	63.30	23.92	2.99	3.36	48.37	37.07	10.21	3.88	0.47
USynthFace (syn.)	1.87	6.70	3.79	14.55	53.34	18.32	1.17	0.57	59.85	35.78	4.18	0.21	0.00
Syn_10K_50 (syn.)	0.74	6.27	4.43	15.73	54.62	17.34	0.69	0.19	63.97	33.71	2.26	0.06	0.00
FFHQ (auth.)	2.88	7.40	5.45	8.70	49.80	17.10	4.65	4.03	59.89	36.03	3.79	0.28	0.02

Table 2. **Distribution of Head Pose and Age in %:** Similar to Figure 3, the percentages in the Table show that there is a high imbalance towards the age ranges (25,32) and the adjacent age ranges in all datasets. The infant and elderly classes are under-represented in all datasets, but the imbalance increased when using the generators trained on FFHQ to create the USynthFace and the Syn_10K_50 datasets. Regarding head pose, the CASIA-WebFace and also SFace-60 provide higher diversity, but the diversity is reduced in the synthetic training dataset as fewer profile or next-to-profile images are present in the dataset (see also Figure 4).

CASIA-WebFace and SFace-60 dataset.

The age distribution is presented in Figure 3 and the distribution percentages in Table 2. Again, a high imbalance can be seen across the datasets. The majority of the images are classified as showing individuals in the age range of (25-32). Since the age predictor is not perfect, but also generally provides a high one-off accuracy [10,64], the high amount of samples in the adjacent age ranges of (15,20) and (38,42) might be explained. The small number of children and adolescents in the data sets is noticeable, although it is also noticeable that, similar to ethnicity, the general distribution of the attributes remains the same in the respective generator training dataset.

The distribution of the non-demographic head pose attribute is shown in Figure 4 and Table 2. To simplify the evaluation, we neglect the direction of the yaw angle, i.e. we do not differentiate between the right or left profile of the face. From Figure 4 it can be observed, that there is a high imbalance regarding frontal or nearly frontal images (yaw angle of around 22.5°) on all datasets. Interestingly, SFace-60 seems to contain more images with a semi-profile view (45°) than the original CASIA-WebFace increasing the diversity of the training data. The datasets based on FFHQ show a very similar head pose distribution to the authentic dataset used to train their generator.

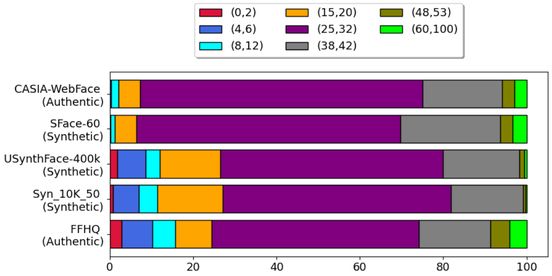

Figure 3. **Age Distribution:** A relatively higher representation of the age range of (25,32) and the adjacent age ranges can be observed. The age distribution seems also to be inherited from the authentic datasets, as SFace-60 mimics the distribution of CASIA-WebFace. USynthFace-400k and Syn_10K_50 are also roughly similar to FFHQ with some differences in the elderly and infant class, which are more under-represented in the synthetic datasets.

4.1.2 Intra-Identity Attribute Consistency

We also investigate the consistency of our investigated attributes to gain insights into how variable the specific attributes are and how good the synthetic models can maintain these, as they normally aim at preserving synthetic identity information and not the specific attributes. With the IIACR we measure the mean intra-identity attribute consistency over the analyzed subsets of the datasets. For attributes such as gender and ethnicity, we expect it to be beneficial for the FR training, if the IIACR is high, as this means that less noise or domain difference is introduced into the model. For attributes such as age and pose, we assume a lower consistency to be beneficial as this means a higher natural intra-class variability is present in the training dataset.

Table 3 shows the IIACR values for the authentic CASIA-WebFace dataset as a reference dataset for an authentic dataset and the three synthetic datasets. The high values of 0.95 for gender and 0.92 for ethnicity indicate

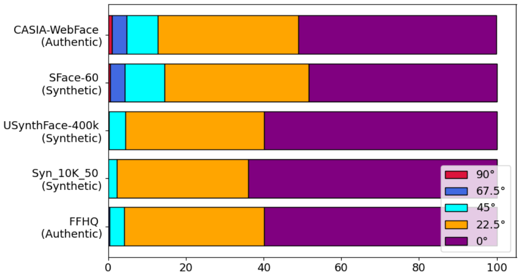

Figure 4. **Headpose Distribution:** The large majority of the face images in the authentic and synthetic datasets are frontal or near frontal. Profile and near profile images are under-represented and the percentage decreases at least when using the CASIA-WebFace-based generator to create SFace-60 (see also Table 2).

Dataset	Gender	Age	Ethnicity	Pose
CASIA-WF	0.95±0.07	0.78±0.16	0.92±0.11	0.57±0.14
SFace-60	0.93±0.10	0.67±0.15	0.91±0.07	0.49±0.10
Syn_10K_50	0.71±0.11	0.57±0.13	0.73±0.16	0.64±0.07
USynthFace	0.92±0.13	0.76±0.18	0.82±0.18	0.60±0.07

Table 3. **Intra-Identity Attribute Consistency Ratio:** SFace-60 and USynthFace-400k show a similar IIACR as CASIA-WebFace regarding gender and ethnicity, indicating that most images of one synthetic individual share the same attribute features. Syn_10K_50 shows a different behavior with less consistency, which might be due to the identity-mixup without checking for matching gender or ethnicity. The age and pose IIACR is lower for SFace-60 than for CASIA-WebFace which might indicate a higher intra-class variation of the synthetic data in contrast to the authentic data.

Model	Male	Female	mAcc+STD	SER
Baseline	93.45	91.52	92.49±0.97	1.29
SFace$_{syth}$	92.38	89.67	91.03±1.35	1.36
SynFace	80.52	77.96	79.24±1.28	1.13
USynthFace	81.27	78.27	79.77±1.50	1.16

Table 4. **Gender Bias on BFW:** All the models, including the authentic model performed better on male faces than female faces. While the STD indicates less bias in the authentic model, the SER indicates higher bias in the authentic data.

the high consistency of the gender and ethnicity attributes on authentic data. SFace-60, which utilizes the identities of CASIA-WebFace to create its synthetic identities also achieves a high consistency in these attributes, which indicates a lower domain gap. The IIACR for Syn_10K_50 dataset shows a lower IIACR regarding gender and ethnicity which might be due to the identity-mixup during the face synthesis which might mix up individuals of different genders or ethnicities, leading to contradicting face features in different images.

To summarize our investigation of the diversity of the training dataset showed, that the overall distribution of the authentic training dataset is to some degree inherited by the synthetic training data, depending on which data the generator has been trained. In some cases, the imbalance is even increased after the face synthesis. The investigation on the intra-identity consistency showed that most synthetic datasets show a high consistency regarding fixed attributes such as gender and ethnicity, while having a similar variability regarding changing attributes such as age and pose.

4.2. Bias in Synthetic-based FR Models

In this subsection, we investigate the bias in the decisions of synthetic-based FR models in comparison to a baseline model trained on authentic data. We investigate this on gender, ethnicity, age, and head pose. To evaluate this, we report the subset-specific accuracy, the mean accuracy (mAcc) including standard deviation (STD), and the SER, following [36, 71, 72]. Following the ”Rule of 30” and its extensions [23], in most cases, a confidence level of 90% with a percent relative error of 10% is achieved regarding the reported accuracies and errors. Only the Baseline and the SFace model are below the required number of errors on the LFW and FF dataset to achieve this confidence and percent relative error. Given the error rates in the Tables 4, 5, 6 and 7 and the level of confidence for the percent relative bias, the bias in most cases is statistically significant.

Table 4 presents the verification accuracy on the gender subsets of the BFW dataset. In all cases, the performance is worse on female than male faces, which is consistent with previous findings [1, 2]. While the performance of the SFace$_{synth}$ is competing with the Baseline model, the other models perform worse.

To analyze ethnicity bias, we provide the verification accuracy, the mean accuracy, standard deviation, and SER on the ethnicity split on RFW and BFW in Table 5. In the reported values we can observe an ethnicity bias as the Caucasian/White ethnicity class is the best-performing attribute class regarding verification accuracy across all investigated models. Similar to the observation on the gender bias, we observe that the STD of the synthetic model on the RFW dataset is higher in contrast to the STD of the authentic models, while the SER is lower, indicating a lower bias. Interestingly, on the BFW dataset, the SER and the STD are lower for SynFace and USynthFace, meaning lower bias. A possible explanation might be, that due to the less constrained identity attributes and a higher inconsistency regarding ethnicity within an identity in the synthetic data (see 3), the synthetic models are less biased.

The results of the age bias evaluation are presented in Table 6. In contrast to the other datasets, the setup on age is slightly different, as the distinction is not made between different age groups, but between comparisons of more similar age groups and cross-age comparisons. The results show, that similar to authentic models, the performance on higher age gaps between compared samples reduced the performance also for synthetic models. The higher variation in the intra-identity age attribute might also be beneficial, as

	RFW					
Model	African	Asian	Caucasian	Indian	mAcc + STD	SER
Baseline	85.52	84.27	92.52	88.03	87.59±3.15	2.10
SFace$_{syth}$	80.17	80.93	90.15	84.32	83.89±3.94	2.01
SynFace	59.75	67.20	70.05	66.00	65.75±3.76	1.34
USynthFace	60.25	67.67	72.40	69.13	67.36±4.45	1.44
	BFW					
Model	Black	Asian	White	Indian	mAcc + STD	SER
Baseline	92.20	87.43	95.56	90.71	91.48±2.92	2.83
SFace$_{syth}$	90.05	85.66	94.04	88.99	89.69±2.99	2.41
SynFace	74.92	72.79	79.47	75.90	75.77±2.41	1.32
USynthFace	75.66	74.01	80.35	77.06	76.77±2.33	1.32

Table 5. **FR performance and ethnicity Bias on BFW and RFW:** The authentic-based baseline model as well as the synthetic-based models show a high positive bias regarding Caucasians/Whites. While the STD and SER on RFW indicate higher bias of the synthetic models toward the ethnicities, the results on BFW might indicate that a lower intra-class ethnicity consistency might be beneficial to reduce bias.

Model	LFW	Cross-Age LFW	mAcc + STD	SER
Baseline	99.38	93.22	96.30±3.80	10.94
SFace$_{syth}$	99.02	92.08	95.55±3.47	8.08
SynFace	91.77	75.18	83.48±8.29	3.02
USynthFace	91.83	76.88	84.35±7.48	2.83

Table 6. **Age Bias on LFW and Cross-Age LFW:** All the models perform worse when confronted with comparison pairs of higher age gaps. The absolute performance drop was even higher on the synthetic-based models than the authentic-based baseline. SER is consistently lower (better) for synthetic-based models.

Model	F-F	F-P	mAcc + STD	SER
Baseline	99.33	95.84	97.59±1.74	6.21
SFace$_{syth}$	98.80	91.90	95.35±3.45	6.75
SynFace	90.33	74.53	82.43±7.90	2.63
USynthFace	90.34	78.20	84.27±6.07	2.26

Table 7. **Headpose Bias on CFP-FF (frontal-frontal) and CFP-FP (frontal-profile)**: The results show that cross-pose verification accuracy is worse on every model, authentic and synthetic-based. The absolute performance decrease is higher on the synthetic-based models than on the authentic baseline.

a lower SER can be reported for the synthetic models than the authentic baseline.

Finally, the results on variation in head pose are presented in Table 7. We compare two scenarios: Frontal-Frontal and Frontal-Profile comparison pairs. All the models perform worse on the frontal-profile scenario indicating a higher challenge and biased behavior. Analyzing the results of the synthetic models shows that the performance difference between frontal-frontal and frontal-profile increased in contrast to the authentic model.

To summarize, in total we observed similar biases in the synthetic FR models as in the authentic FR model, which motivates the use and development of bias mitigation techniques also on synthetic-based FR models. The intra-identity attribute consistency which might be lower in synthetic data due to fewer constraints might be beneficial to reduce bias.

5. Conclusion

In this work, we investigated the diversity and bias of synthetic data and synthetic-based FR models. As synthetic data and synthetic models are becoming established as a real alternative to authentic data, it is highly necessity to investigate this in more detail to study discriminatory behavior and to reduce it in future work. In our investigation, we analyzed the distribution and intra-identity attribute consistency on three demographic (gender, ethnicity, age) and one non-demographic (head pose) attribute. The results show that the generator models tend to recreate the distribution from the training datasets and might slightly amplify the imbalance in the synthetic datasets. To investigate the bias, and performance differences depending on different subsets, we performed several experiments on gender, ethnicity, age, and pose splits. The results show, that similar biases can be observed in synthetic FR models as in authentic FR models, motivating existing and new works regarding bias mitigation also be applied to the novel synthetic-based models, especially given the possibility of inducing variability in the synthesized images. Furthermore, our investigation showed that the synthetic face recognition models yet do not achieve the same performance as models trained on authentic data.

Acknowledgement This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science, and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Vítor Albiero and Kevin W. Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. 2, 7

[2] 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 2020, Houston, TX, USA, September 28 - October 1, 2020*, pages 1–10. IEEE, 2020. 1, 2, 7

[3] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11129 of *Lecture Notes in Computer Science*, pages 556–572. Springer, 2018. 1

[4] D. Anghelone, S. Lannes, and A. Dantcheva. Anyres: Generating high-resolution visible-face images from low-resolution thermal-face images. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 246–251, Los Alamitos, CA, USA, jul 2023. IEEE Computer Society. 3

[5] Gwangbin Bae, Martin de La Gorce, Tadas Baltrusaitis, Charlie Hewitt, Dong Chen, Julien P. C. Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 3515–3524. IEEE, 2023. 2

[6] Arturo Miguel Russell Bernal and Jane Cleland-Huang. Hierarchically organized computer vision in support of multi-faceted search for missing persons. In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*, pages 1–7. IEEE, 2023. 3

[7] Aman Bhatta, Vítor Albiero, Kevin W. Bowyer, and Michael C. King. The gender gap in face recognition accuracy is a hairy problem. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pages 1–10. IEEE, 2023. 1

[8] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1577–1586. IEEE, 2022. 3, 5

[9] Fadi Boutros, Naser Damer, and Arjan Kuijper. Quantface: Towards lightweight face recognition by synthetic data low-bit quantization. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 855–862, 2022. 2

[10] Fadi Boutros, Naser Damer, Philipp Terhörst, Florian Kirchbuchner, and Arjan Kuijper. Exploring the channels of multiple color spaces for age and gender estimation from face images. In *22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019*, pages 1–8. IEEE, 2019. 3, 6

[11] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19650–19661, October 2023. 2

[12] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–11. IEEE, 2022. 1, 2, 4, 5

[13] Fadi Boutros, Marcel Klemt, Meiling Fang, Arjan Kuijper, and Naser Damer. Exfacegan: Exploring identity directions in gan's learned latent space for synthetic identity generation. *CoRR*, abs/2307.05151, 2023. 2

[14] Fadi Boutros, Marcel Klemt, Meiling Fang, Arjan Kuijper, and Naser Damer. Unsupervised face recognition using unlabeled synthetic data. In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*, pages 1–8. IEEE, 2023. 1, 2, 4, 5

[15] Fadi Boutros, Vitomir Struc, Julian Fiérrez, and Naser Damer. Synthetic data for face recognition: Current state and future prospects. *Image Vis. Comput.*, 135:104688, 2023. 1

[16] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 67–74. IEEE Computer Society, 2018. 1

[17] Jacqueline G. Cavazos, P. Jonathon Phillips, Carlos Domingo Castillo, and Alice J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Trans. Biom. Behav. Identity Sci.*, 3(1):101–111, 2021. 1

[18] Colin W. G. Clifford, Tamara L. Watson, and David White. Two sources of bias explain errors in facial age estimation. *Royal Society Open Science*, 5(10):180841, 2018. 1

[19] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1606–1617, June 2022. 2

[20] Debayan Deb, Neeta Nain, and Anil K. Jain. Longitudinal study of child face recognition. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 225–232. IEEE, 2018. 2

[21] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. 5

[22] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5153–5162. Computer Vision Foundation / IEEE, 2020. 2, 5

[23] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. The nist speaker recognition evaluation – overview, methodology, systems, results, perspective. *Speech Communication*, 31(2):225–254, 2000. 7

[24] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *CoRR*, abs/2003.02488, 2020. 3

[25] Noyan Evirgen and Xiang 'Anthony' Chen. Ganzilla: User-driven direction discovery in generative adversarial networks. In Maneesh Agrawala, Jacob O. Wobbrock, Eytan Adar, and Vidya Setlur, editors, *The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022 - 2 November 2022*, pages 75:1–75:10. ACM, 2022. 3

[26] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Demographic bias in presentation attack detection of iris recognition systems. In *28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam, Netherlands, January 18-21, 2021*, pages 835–839. IEEE, 2020. 3

[27] Meiling Fang, Marco Huber, and Naser Damer. Synthaspoof: Developing face presentation attack detection based on privacy-friendly synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1061–1070, June 2023. 2

[28] Meiling Fang, Wufei Yang, Arjan Kuijper, Vitomir Struc, and Naser Damer. Fairness in face presentation attack detection. *Pattern Recognition*, 147:110002, 2024. 3

[29] Biying Fu, Marcel Klemt, Fadi Boutros, and Naser Damer. On the quality and diversity of synthetic face data and its relation to the generator training data. In *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2023. 2

[30] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 330–347. Springer, 2020. 1

[31] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016. 1

[32] Hu Han, Charles Otto, and Anil K. Jain. Age estimation from face images: Human vs. machine performance. In Julian Fiérrez, Ajay Kumar, Mayank Vatsa, Raymond N. J. Veldhuis, and Javier Ortega-Garcia, editors, *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, pages 1–8. IEEE, 2013. 1

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 4, 5

[34] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(11):2781–2794, 2020. 2

[35] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 4

[36] Linzhi Huang, Mei Wang, Jiahao Liang, Weihong Deng, Hongzhi Shi, Dongchao Wen, Yingjie Zhang, and Jian Zhao. Gradient attention balance network: Mitigating face recognition racial bias via gradient attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 38–47. IEEE, 2023. 4, 7

[37] Marco Huber, Meiling Fang, Fadi Boutros, and Naser Damer. Are explainability tools gender biased? a case study on face presentation attack detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 945–949, 2023. 3

[38] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 4

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. 5

[40] Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18729–18738. IEEE, 2022. 5

[41] Minchul Kim, Feng Liu, Anil K. Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12715–12725. IEEE, 2023. 1

[42] Jan Niklas Kolf, Tim Rieber, Jurek Elliesen, Fadi Boutros, Arjan Kuijper, and Naser Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 806–816, June 2023. 2

[43] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, pages 2144–2151. IEEE Computer Society, 2011. 3

[44] Tesfaye B. Mersha and Tilahun Abebe. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Human Genomics*, 9(1):1, Jan 2015. 1

[45] Surbhi Mittal, Kartik Thakral, Puspita Majumdar, Mayank Vatsa, and Richa Singh. Are face detection models biased? In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*, pages 1–7. IEEE, 2023. 3

[46] Richard Plesh, Peter Peer, and Vitomir Struc. Glassesgan: Eyewear personalization using synthetic appearance discovery and targeted subspace modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16847–16857, June 2023. 2

[47] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10860–10870. IEEE, 2021. 2, 4, 5

[48] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 460–469. IEEE Computer Society, 2016. 2

[49] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, volume 9914 of *Lecture Notes in Computer Science*, pages 17–35, 2016. 1

[50] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, 2020. 2, 4

[51] Joseph P. Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Trans. Image Process.*, 32:4365–4377, 2023. 3, 4

[52] Francisco Romero, Johann Hauswald, Aditi Partap, Daniel Kang, Matei Zaharia, and Christos Kozyrakis. Optimizing video analytics with declarative model relationships. *Proc. VLDB Endow.*, 16(3):447–460, 2022. 3

[53] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.*, 126(2-4):144–157, 2018. 3

[54] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2074–2083. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 3

[56] Wojciech Samek, Alexander Binder, Sebastian Lapuschkin, and Klaus-Robert Müller. Understanding and comparing deep neural networks for age and gender classification. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 1629–1638. IEEE Computer Society, 2017. 3

[57] C L Saunders, G A Abel, A El Turabi, F Ahmed, and G Lyratzopoulos. Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the english cancer patient experience survey. *BMJ Open*, 3(6):e002882, June 2013. 1

[58] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Domingo 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 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–9. IEEE Computer Society, 2016. 4

[59] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020. 3

[60] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4, 2021. 3

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[62] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S. Bolme, and Michael King. Face recognition algorithm bias: Performance differences on images of children and adults. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2269–2277. Computer Vision Foundation / IEEE, 2019. 2

[63] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Multi-algorithmic fusion for reliable age and gender estimation

from face images. In *22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019*, pages 1–8. IEEE, 2019. 3

[64] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Ines Zelch, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Reliable age and gender estimation from face images: Stating the confidence of model predictions. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, FL, USA, September 23-26, 2019*, pages 1–8. IEEE, 2019. 3, 6

[65] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020*, pages 1–11. IEEE, 2020. 3

[66] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognit. Lett.*, 140:332–338, 2020. 1

[67] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2022. 1, 2

[68] The European Parliament and the Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (General Data Protection Regulation), 2016. 1, 2

[69] Manuel C Voelkle, Natalie C Ebner, Ulman Lindenberger, and Michaela Riediger. Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age. *Psychol. Aging*, 27(2):265–277, June 2012. 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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. Computer Vision Foundation / IEEE Computer Society, 2018. 4

[71] Mei Wang and Weihong Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *CoRR*, abs/1911.10692, 2019. 4, 7

[72] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9319–9328. Computer Vision Foundation / IEEE, 2020. 4, 7

[73] 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 692–702. IEEE, 2019. 1, 4

[74] Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):8433–8448, 2022. 1, 3

[75] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 4, 5

[76] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. 4