Bias and Diversity in Synthetic-based Face Recognition - Supplementary Material

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

1. Introduction

This is the supplementary material to the paper: Bias and Diversity in Synthetic-based Face Recognition. To validate that our conclusions still hold when using other attribute estimator, in the first section, we describe the additionally used estimators for the results provided in the supplementary material. After that, we shortly describe and provide the diversity investigation based on these estimators that supplement the experiments and results provided in the paper.

2. Additional Attribute Estimators

In addition to the attribute estimators presented in the main paper, we also utilized four established attribute predictors. The additional attribute predictors cover the attribute gender, age, ethnicity, and face emotion. They all have been used in several publications [4,6,23,50]. The gender predictor is a pre-trained gender classifier provided by HpyerExtended LightFace [58] and is also based on the VGG-16 architecture [59] It achieved a classification accuracy of 90.82% on the BFW [49] dataset.

The age predictor is a pre-trained age classifier provided by HyperExtended LightFace [58] and is also based on the VGG-16 architecture [59]. It predicts an age between 0 and 100. We adjusted the age classes of Adience [54] to close the gaps $((0, 3)$, $(4, 7)$, $(8, 13)$, $(14, 22)$, $(23, 34)$, $(35, 45)$, $(46, 56)$, and $(57, 100))$ and achieved an accuracy of 27.52% on the Adience dataset [54], which is better than random, but worse than the utilized predictor in the paper (60.51%).

The ethnicity predictor E_2 is a pre-trained model from HyperExtended LightFace [58] and predicts six ethnicities: $Asian$, $Black$, $Indian$, $LatinoHispanic$, $MiddleEastern$, and $Indian$. We achieved an accuracy of 85.46% on 1,300 randomly selected test images from BUPT-Balanceface when merging $LatinoHispanic$ to $White$ and $MiddleEastern$ to $Indian$, which led to the highest merged accuracy. The accuracy of our SVM-based ethnicity predictor used in the paper is better with an accuracy of 90.91% on the same evaluation data.

As the emotion predictor, we use a pre-trained model provided by HyperExtended LightFace [58]. It consists of 12-layer architecture and the details are provided in [LF]. It outputs $Angry$, $Disgust$, $Fear$, $Happy$, Sad, $Surprise$, and $Neutral$. The emotion model achieved an accuracy of 57.42% on the FER-2013 dataset according to the model provider [58].

3. Diversity Investigation

The distribution of the gender attribute is visualized in Figure 1 and provided in Table 1. The predicted gender distribution of the authentic FFHQ and CASIA-Webface dataset are very similar to the predicted distribution in Figure 1. In contrast to the predicted distribution in Figure 1, the USynthFace-400k dataset and the Syn_10K_50 are more biased. The observation, that the synthetic data generator may tend to create more samples from the majority class remains, as the datasets created with a generator trained on the gender-unbalanced CASIA-WebFace (SFace-60 and USynthFace-400k) show higher gender imbalance.

The ethnicity distribution of the alternative ethnicity is provided in Figure 2 and Table 1. The mis-match to the established [49, 71] four ethnicities (white, black, indian, asian) hamper the comparison with Figure 2. Nevertheless, the high imbalance regarding white individuals can be still observed. Similar is true for the decreased diversity, as less indians and black individuals are predicted in the synthetic datasets than in the original authentic datasets.

The age distribution of the alternative age estimator is provided in Figure 3 and Table 2. The distribution shows a high imbalance regarding middle-aged (23-45) individuals, similar to the distribution observed in Figure 3. The diversity in the synthetic dataset is further reduced compared to the authentic datasets. Since the age estimator used in the paper has a higher classification accuracy, the results in the paper are more reliable than the results presented based on

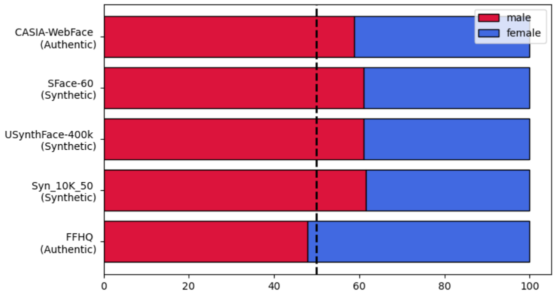

Figure 1. **Gender Distribution - Alternative Gender Estimator:** The imbalance in terms of gender of SFace-60 and USynthFace-400k increased in comparison to the authentic base dataset CASIA-WebFace. This supports the results based on the other gender estimator. The distribution of FFHQ is very similar when comparing both predicted gender distributions. A larger difference can be observed regarding the imbalance of the Syn_10k_50 dataset.

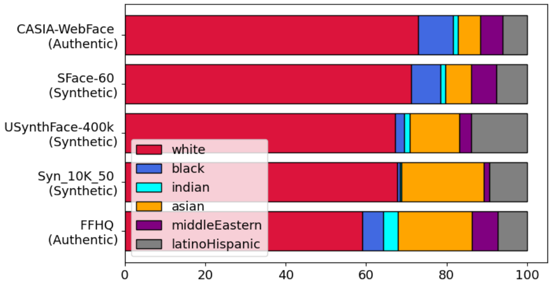

Figure 2. **Ethnicity Distribution - Alternative Ethnicity Estimator:** The alternative ethnicity estimator indicates a high imbalance towards white individuals, similar to the results of the other utilized estimator. The difference in ethnicity classes hamper the comparison between both results.

this estimator.

Figure 4 and Table 2 provide the distribution of the non-demographic face emotion attribute. The results show, that all the datasets suffer from a high imbalance regarding neutral or happy face expressions. On the synthetic datasets, this effect intensifies.

With this additional investigation based on other established estimators, it can be observed that the results support the findings of the experiments in the paper and lead to similar conclusions, despite using different attribute estimators. With this additional investigation based on other established estimators, it can be observed that the results support the findings of the experiments in the paper and lead to similar conclusions.

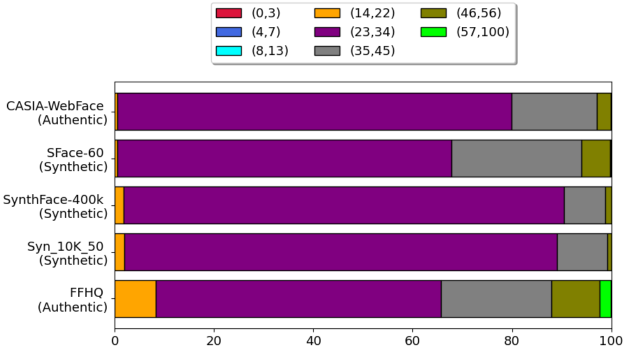

Figure 3. **Age Distribution:** A high representation of the age range of (23,34) and the adjacent age range (35,45) can be observed. The age distribution seems also to be inherited from the authentic datasets, while reducing diversity.

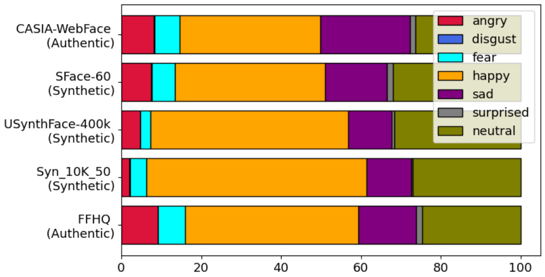

Figure 4. **Emotion Distribution:** The emotion distribution indicates a high imbalance regarding the face expression of the individual in authentic and synthetic datasets that might lead to bias. The large majority of the faces has been predicted as happy or neutral, with an higher imbalance to this emotions in the synthetic datasets.

Dataset	Gender		Ethnicity					
	Male	Female	Asian	Indian	Black	White	Middle Eastern	Latino Hispanic
CASIA-WebFace (auth.)	58.90	41.10	5.57	1.29	8.53	73.03	5.62	5.96
SFace-60 (syn.)	61.13	38.87	6.46	1.16	7.41	71.15	6.15	7.68
USynthFace (syn.)	61.09	38.91	12.31	1.48	2.34	67.14	2.87	13.87
Syn_10K_50 (syn.)	61.60	38.40	20.46	0.43	0.65	67.75	1.33	9.38
FFHQ (auth.)	47.78	52.22	18.30	3.67	5.23	59.09	6.40	7.31

Table 1. **Distribution of Gender and Ethnicity based on the alternative estimators in %: :** The values show that the synthetic SFace-60 and the USynthFace-400k 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.

Dataset	Age								Emotion						
	(0,3)	(4,7)	(8,13)	(14,22)	(23,34)	(35,45)	(46,56)	(57,100)	Angry	Disgust	Fear	Happy	Sad	Surprised	Neutral
CASIA-WF. (auth.)	0.00	0.00	0.00	0.52	79.44	17.10	2.90	0.03	8.20	0.16	6.29	35.22	22.49	1.40	26.24
SFace-60 (syn.)	0.00	0.00	0.00	0.62	67.22	26.16	5.74	0.26	7.55	0.07	5.89	37.50	15.53	1.45	31.99
USynthFace (syn.)	0.00	0.00	0.00	1.76	88.69	8.25	1.29	0.00	4.82	0.03	2.44	49.51	11.00	0.68	31.53
Syn_10K_50 (syn.)	0.00	0.00	0.00	2.00	87.13	10.04	0.81	0.00	2.11	0.04	4.16	55.09	11.30	0.30	26.30
FFHQ (auth.)	0.00	0.00	0.05	8.33	57.36	22.20	9.77	2.30	9.16	0.06	6.77	43.43	14.47	1.51	24.59

Table 2. **Distribution of Age and Emotion based on the alternative estimators in %:** The percentages in the table show that their is a high imbalance towards the age ranges (23,34) and the adjacent age ranges in all datasets. The infant and elderly classes are under-represented in all datasets, similar to the observation on Figure 3 and Table 2. The performance of the alternative age predictor is far worse than the predictor utilized in the paper, therefore, the results provided in the paper are more reliable. Regarding the emotion distribution, also a high imbalance towards haooy and neutral can be observed in both, authentic and synthetic datasets.