Fairer Analysis and Demographically Balanced Face Generation for Fairer Face Verification
- Supplementary Materials -

Alexandre Fournier-Montgieux*[1]
alexandre.fourniermontgieux@cea.fr

Michaël Soumm*[1]
michael.soumm@cea.fr

Adrian Popescu[1]
adrian.popescu@cea.fr

Bertrand Luvison[1]
bertrand.luvison@cea.fr

Hervé Le Borgne[1]
herve.le-borgne@cea.fr

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

1. Parameters for training and generation

For training the face classifier, we use the Adaface training pipeline [14]. 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 [14]. The margin parameter m is set to 0.4, and the control concentration constant h to 0.333 as recommended by [14]. On each training set, the training lasts 60 epochs.

For generating the DCFACE set and its variants, we use the generation pipeline of [15]. 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 [15], 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 files [1] 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 [15]. 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 [15]. 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 [31] 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 [4].

[1]The provided center_ir_101_adaface_webface4m_faces_webface _112x112.pth file doesn't have a required "similarity_df" field. Also, the dcface_3x3.ckpt file doesn't seem to store the following property: recognition_model.center.weight.data

2. Performance in Accuracy on other sets

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

In addition to FAVCI2D , BFW, and RFW, we report in Table 1 the raw accuracy results on 5 common evaluation sets used in prior work on the FR task [2, 14, 15, 20] : (1) Labeled Faces in the Wild (LFW) [11], the reference dataset for the task (2) CALFW [29], a version of LFWwith a larger age variability, (3) CPLFW [28], a version of LFWwith pose variability, (4) AgeDB [19], a dataset designed for maximizing age variability, and (5) CFP-FP [22] 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 1 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).

3. Bias Mitigation techniques details

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

3.1. 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 [13, 23].

Oversampling [1, 3, 16, 30] 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 [5, 6, 9] can be used in order to tackle this over-fitting 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 reduction of the majority classes so that their representativity matches the underrepresented classes. [17, 18, 24]. 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.

3.2. 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 [14], 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 3.1, i.e. using the invert of the frequency/count [8, 10, 26]. 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:

$$\mathcal{L}(x_1, ..., x_K) = \frac{\sum_k w_k \mathcal{L}(x_k)}{\sum_k w_k} \tag{1}$$

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

4. 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.
- 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 [7] 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 1-left).
- Residual vs. predicted plot: Helps detect heteroscedasticity and nonlinearity (Figure 1-right).

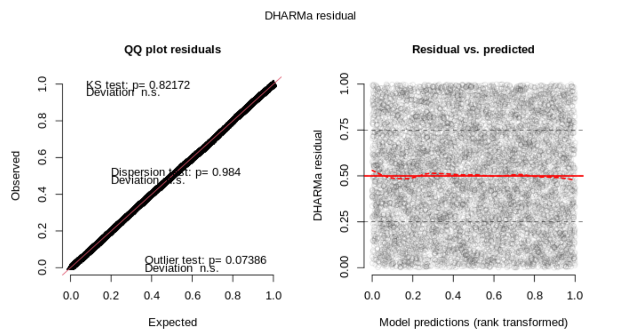


Figure 1. QQ-plot of residuals and Residual vs. predicted plot: logit model is adapted and log-odds are linear in the variables.

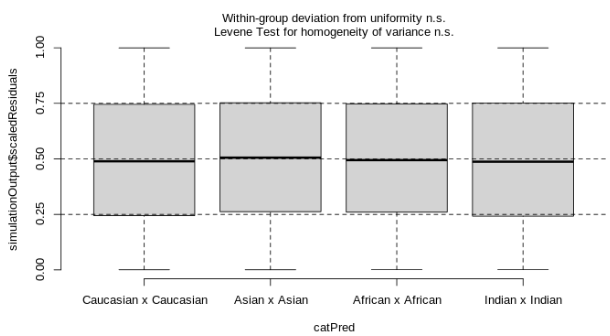


Figure 2. Residual vs. predictor plots: exogeneity is verified.

- Residual vs. predictor plots: Useful for identifying problems with specific predictors (similar to exogeneity) (Figure 2).
- Overdispersion Test: helps to identify if there's more variation in the data than expected under the binomial distribution (Figure 3).
- Zero-inflation Test: check for an excess of zeros or ones (Figure 4).

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.

5. 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. 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 2). With that in mind, we

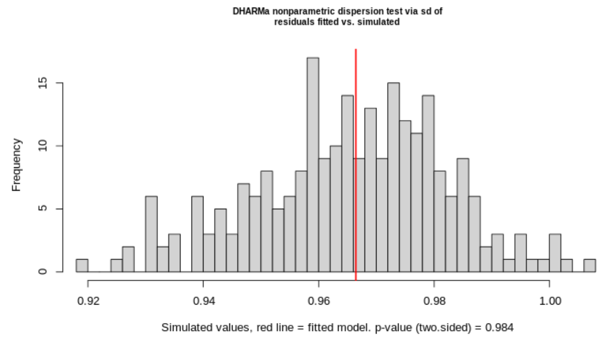


Figure 3. Overdispersion Test: Correct Specification and no autocorrelation.

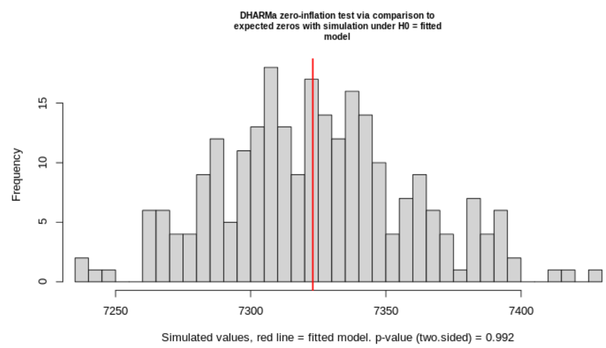


Figure 4. Zero-inflation Test: the model correctly predicts the probability of the outcome.

still get consistent results for the effects of demographic attributes on the models (Figure 5). 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 6) and FMR are coherent with the idea that models tend to predict positive outcomes for certain protected ethnical 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 7 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.

6. Statistical Analysis on BFW

To tackle the issue of the lack of metadata, in addition to BFW, other alternatives exist such as BFW [21] and DemogPairs [12]. 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 10 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 9 and 8 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 2. 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.

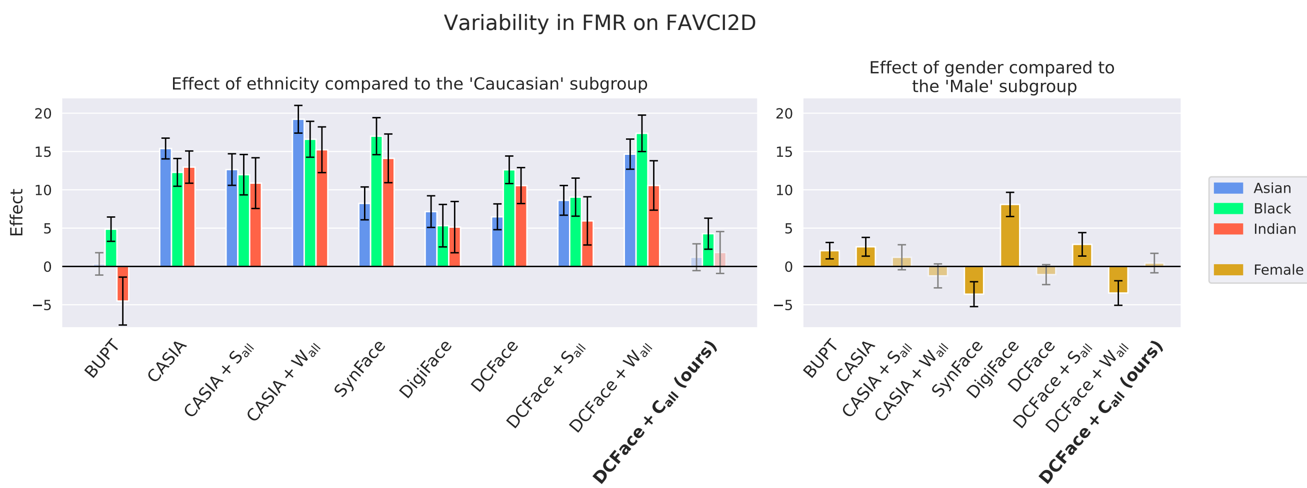


Figure 5. Marginal effect on FMR (lower is better) for each method compared to the unprotected group. Analysis done on FAVCI2D

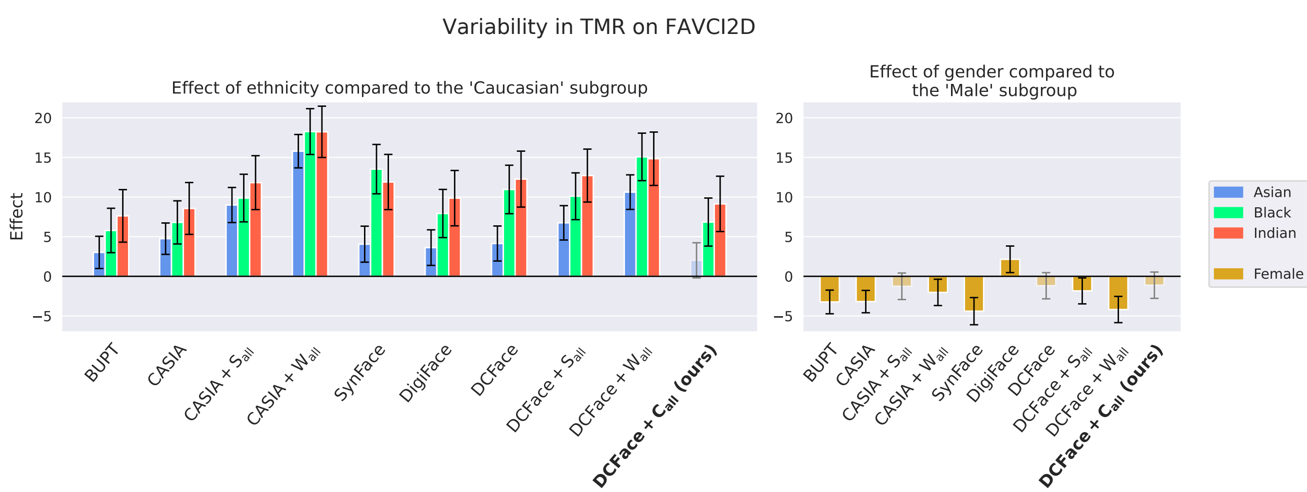


Figure 6. Marginal effect on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on FAVCI2D

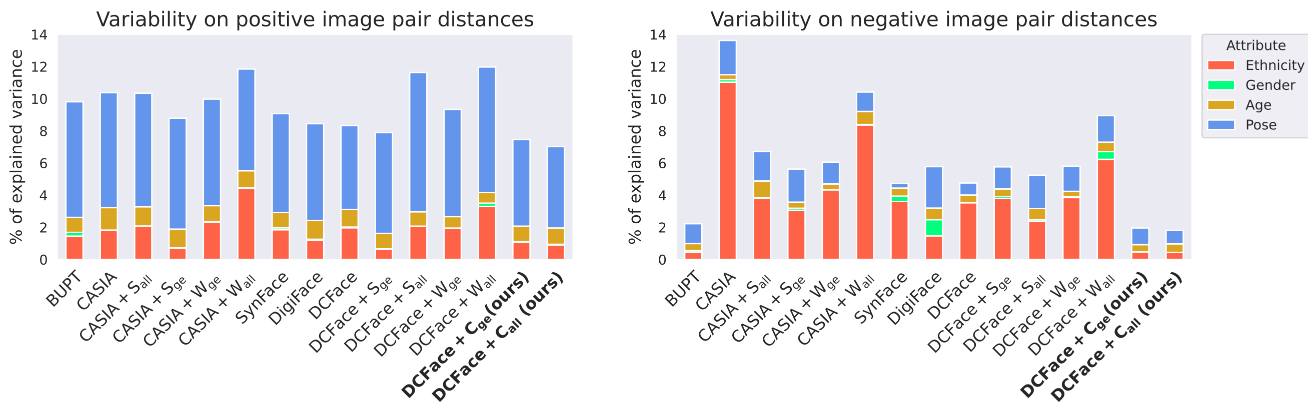


Figure 7. ANOVA results on FAVCI2D : total height corresponds to R^2, the explained variance by the variables. Each bar is decomposed into multiple η^2, i.e. the individual contributions to the variance

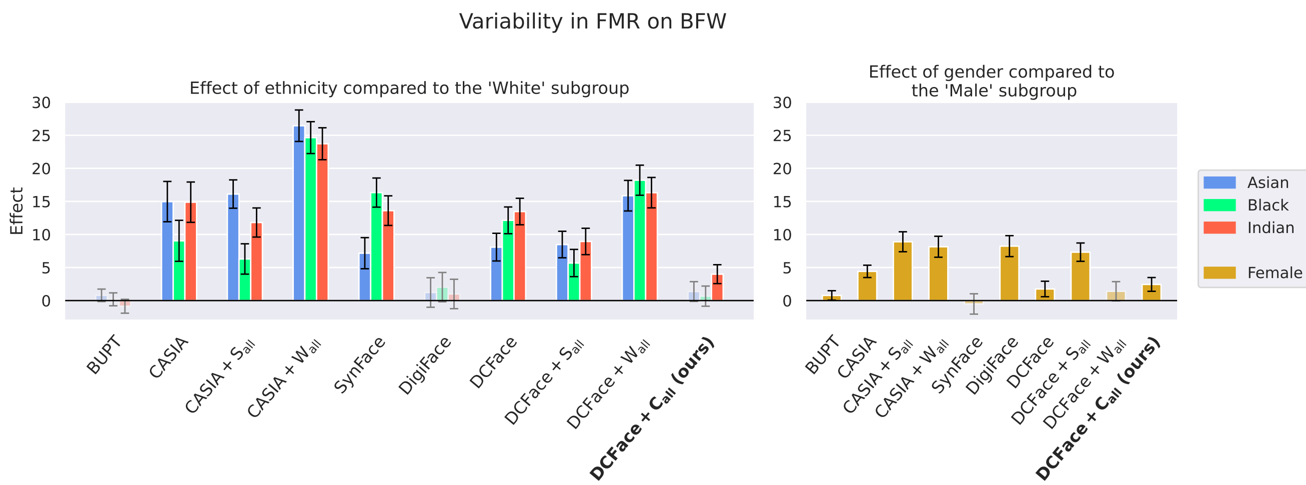


Figure 8. Marginal effect on FMR (lower is better) for each method compared to the unprotected group. Analysis done on BFW

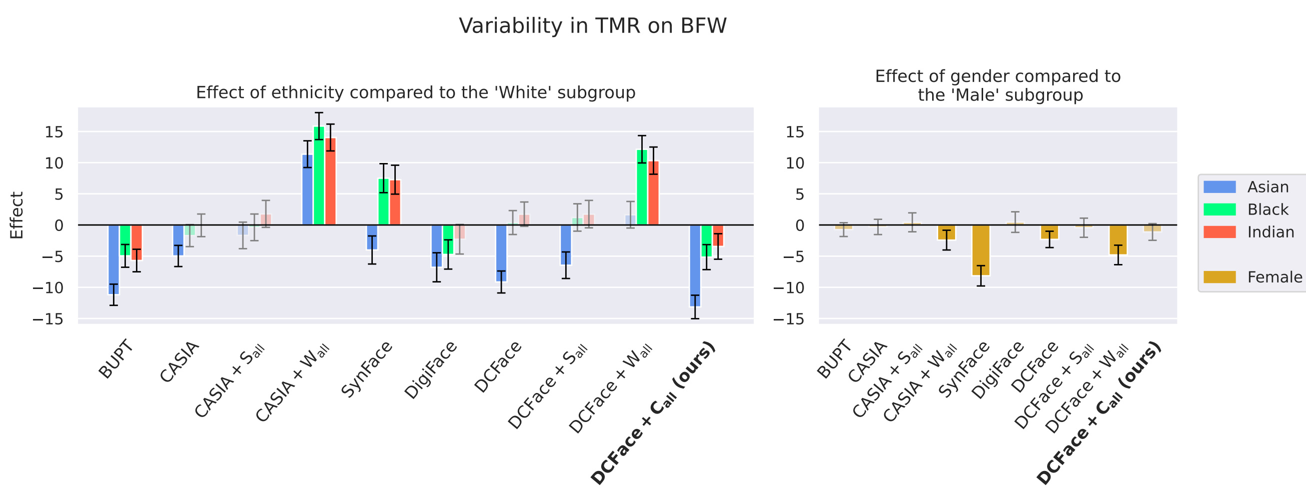


Figure 9. Marginal effect on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on BFW

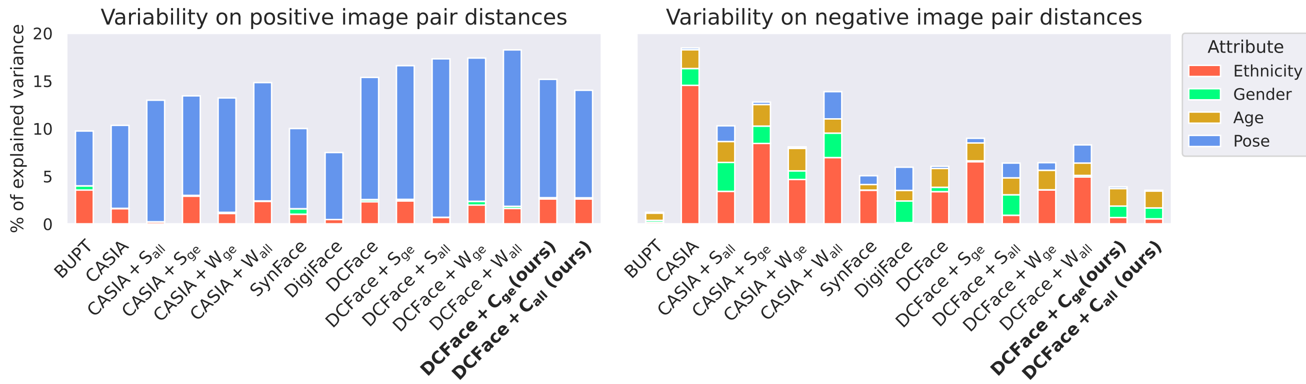


Figure 10. ANOVA results on BFW: total height corresponds to R^2, the explained variance by the variables. Each bar is decomposed into multiple η^2, i.e. the individual contributions to the variance

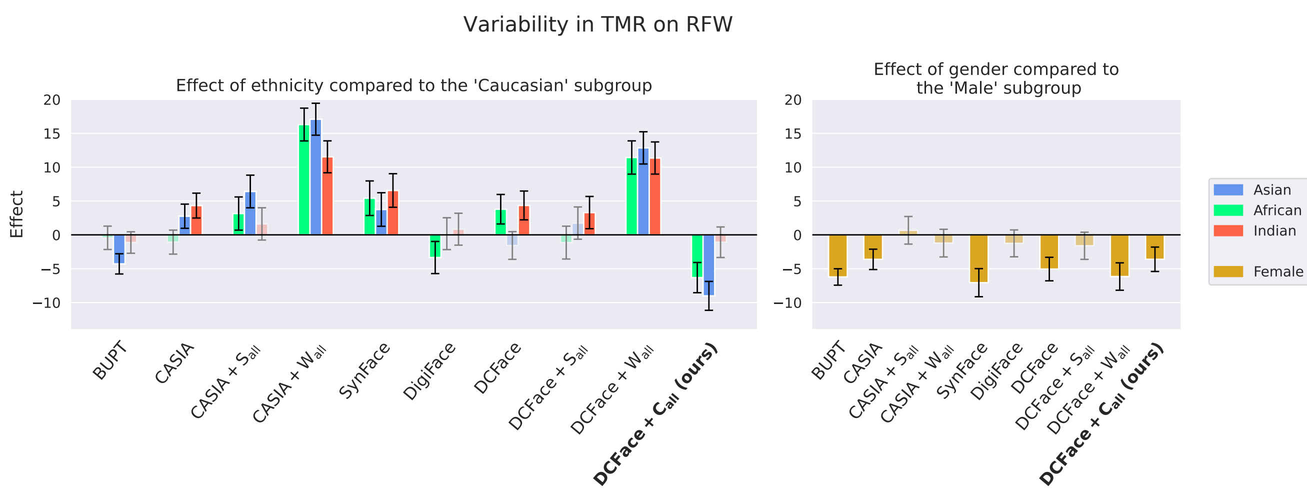
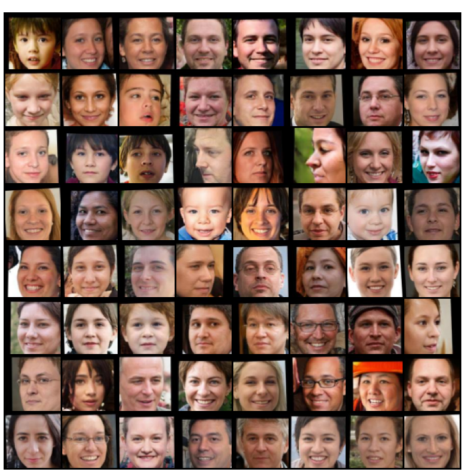


Figure 11. Marginal effects on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on RFW

7. Datasets Images examples

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

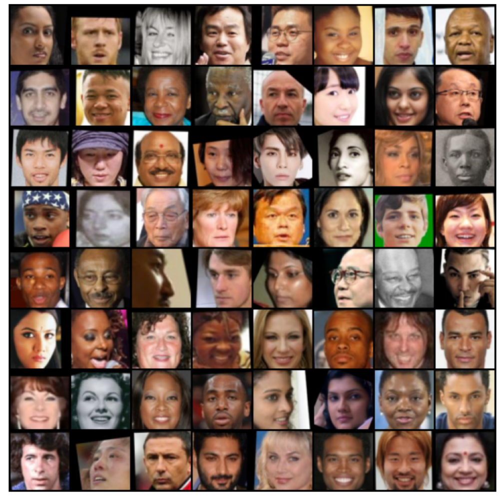
(b) Examples of images generated with the original DCFace [15] pipeline



(c) Examples of images generated with the SynFace pipeline [20]

(d) Examples of images within the DigiFace dataset [2]

(e) Examples of images within the CASIA dataset [27]



(f) Examples of images within the BUPT dataset [25]

References

[1] 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. 2

[2] 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, 8

[3] 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. 2

[4] Mantas Birškus. Glasses Detector, 3 2024. 1

[5] 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 Kijsirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 475–482, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. 2

[6] 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. 2

[7] Hartig F. Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models., 2018. 2

[8] 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. 2

[9] 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, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. 2

[10] 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. 2

[11] 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. 1

[12] 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. 4

[13] 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. 2

[14] 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. 1, 2

[15] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: 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, 8

[16] Felix Last, Georgios Douzas, and Fernando Bação. Oversampling for imbalanced learning based on k-means and smote. *ArXiv*, abs/1711.00837, 2017. 2

[17] Daniel Lehmann and Marc Ebner. Subclass-based undersampling for class-imbalanced image classification. In *VISIGRAPP*, 2022. 2

[18] 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. 2

[19] 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. 1

[20] 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, 8

[21] 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. 4

[22] 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. 1

[23] 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. 2

[24] 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. 2

[25] 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. 9

[26] 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. 2

[27] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 9

[28] 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. 1

[29] 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. 1

[30] Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. Oversampling method for imbalanced classification. *Comput. Informatics*, 34:1017–1037, 2015. 2

[31] 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. 1