Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 210 (2022) 212–217

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International Workshop on Artificial Intelligence Methods for Smart Cities (AISC 2022)
October 26-28, 2022, Leuven, Belgium

The More Secure, The Less Equally Usable: Gender and Ethnicity (Un)fairness of Deep Face Recognition along Security Thresholds

Andrea Atzori, Gianni Fenu, Mirko Marras*

Department of Mathematics and Computer Science, University of Cagliari, V. Ospedale 72, 09124 Cagliari, Italy

Abstract

Face biometrics are playing a key role in making modern smart city applications more secure and usable. Commonly, the recognition threshold of a face recognition system is adjusted based on the degree of security for the considered use case. The likelihood of a match can be for instance decreased by setting a high threshold in case of a payment transaction verification. Prior work in face recognition has unfortunately showed that error rates are usually higher for certain demographic groups. These disparities have hence brought into question the fairness of systems empowered with face biometrics. In this paper, we investigate the extent to which disparities among demographic groups change under different security levels. Our analysis includes ten face recognition models, three security thresholds, and six demographic groups based on gender and ethnicity. Experiments show that the higher the security of the system is, the higher the disparities in usability among demographic groups are. Compelling unfairness issues hence exist and urge countermeasures in real-world high-stakes environments requiring severe security levels.

Keywords: Authentication; Bias; Biometrics; Fairness; Equity; Causality; Face Recognition; Security; Usability.

1. Introduction

Sensors, cameras, and artificial intelligence are all going to play a huge role in making city environments smarter, safer and more efficient. One emerging technology becoming important for smart cities is face recognition. Both public and private sector organizations can enhance operational productivity by integrating this capability. For instance, vast deployments have been seen in airports to enhance safety and speed up passenger boarding. Face recognition can also be useful for recognizing unusual behavior and identifying known offenders throughout the city environment.

Face recognition systems have achieved impressively high accuracy by leveraging latent representations extracted from deep neural networks. As in other domains, obtaining the highest possible accuracy has been seen as the ultimate goal for years [1]. Despite this advance in performance, face recognition has been proved to be susceptible to algo-

* Corresponding author. Tel.: +39-070-675-87-56; E-mail address: mirko.marras@acm.org

1877-0509
10.1016/j.procs.2022.10.139

rithmic bias. When a bias impacts on legally protected groups (e.g., based on their gender and ethnicity), the resulting inequalities might lead to severe societal consequences like discrimination and unfairness [2, 3].

There is growing evidence of studies on definitions, criteria, metrics, and countermeasures against unfairness in face recognition. For instance, prior work has shown that women suffer from worse performance than men and that children's faces are less likely to be recognized than those of adults [4, 5]. To counter these issues, concerted efforts have been devoted to the creation of demographically balanced datasets [6, 7, 8]. Subsequent research put attention to the origin of the bias, for instance by analyzing the influence of image distortions [9] or covariates beyond demographics [10, 11]. However, the influence of the security level, decreased (increased) by implementing a lower (higher) recognition threshold, on demographic disparities has by no means been researched exhaustively.

In this paper, we therefore investigate the extent to which disparities among demographic groups change under different security levels (*RQ1*) and whether any co-relationships between face characteristics and error rates across security levels exist (*RQ2*). To answer these questions, in a first step, we conducted a systematic study on disparities in false rejection rates emphasized by ten state-of-the-art face encoders against gender and ethnicity groups under three fixed security levels (the false acceptance rate is fixed). In a second step, we analyzed the co-relationships between over fifteen face characteristics, including image (e.g., noise or occlusions) and appearance perspectives (e.g., presence of beard or make-up), and the false rejection rates obtained under the same three fixed security levels.

Our results revealed that the higher the security level is, the higher the disparate usability among demographic groups is (*RQ1*) and that key co-relationships between face characteristics and false rejection rates are present (*RQ2*). Use cases leveraging high security thresholds hence require particular attention with respect to disparate impacts.

2. Data and Method

To answer the two research questions, our method includes two main steps. In a first step, we instantiated a range of face encoders and assessed their overall performance. In a second step, we conducted an analysis on demographic disparities and the impact of face characteristics on such performance estimates.

Face Dataset Preparation. For our analyses, we considered `DiveFace` [12], a well-known dataset for studying fairness in face recognition. This dataset, with 140,000 images from 24,000 identities, is annotated with protected attribute labels and demographically balanced. The protected attributes cover both ethnicity (Asian, Black, Caucasian) and gender (Men and Women). Consequently, there are six demographic groups: Asian Men (AM), Asian Women (AW), Black Men (BM), Black Women (BW), Caucasian Men (CM), and Caucasian Women (CW). The dataset has been divided by the original authors into a training set and a test set, containing 70% and 30% of the identities respectively. Demographic groups are equally represented in both training ($18,584 \pm 2968$ images per group) and test ($4,706 \pm 875$ images per group) sets. Faces were detected through DeepFace [13], before clipping, aligning, and resizing them.

Face Characteristics Extraction. Being generally interested in analyzing the relationships between face characteristics and error rates, we then augmented the descriptive information accompanying each image in the dataset with a range of face characteristics that might influence the system's performance. Specifically, given an image, we extracted a vector $c \in \mathbb{R}^f$, including $f = 20 \in \mathbb{N}$ face characteristics, as reported in Table 1. Except for the protected attributes gender and ethnicity, these face characteristics were extracted through the *Microsoft Cognitive Services*[1]. The selected face characteristics describe images from a wide range of perspectives, emerged by reviewing recently studied influential covariates in the literature [9, 14]. Finally, given an individual u, we computed a fixed-length representation $c_u \in \mathbb{R}^f$, obtained by considering the average (in case of continuous values) or the mode (in case of discrete values) for each characteristic reported across the vectors c extracted from images of the individual u.

Face Encoder Creation. For extracting a latent representation of each image in the considered dataset, we built and trained a range of face encoders based on CNN backbones (`ResNet152`, `AttentionNet`, `ResNeSt`, `RepVGG`, `HRNet`), proved to perform well in recent face recognition benchmarks [1]. Specifically, `ResNet152` is a variant of the well-known ResNet architecture [15] that uses shortcut connections to obtain the residual counterpart. `AttentionNet` [16] is a neural network with residual attention, but enriched with attention modules. Each consists of (i) a mask branch

[1] https://azure.microsoft.com/it-it/services/cognitive-services/face/

Attribute Type	Attribute Name(s)[1]	Short Description
Demographic	*Gender* (Man, Woman); *Ethnicity* (Asian, Black Caucasian); *Age* [1,100]	Characteristics protected by law.
Facial Hair	*Mustache* [0, 1]; *Beard* [0, 1]; *Sideburns* [0, 1]	Presence of facial hair.
Makeup	*Eye makeup* [0, 1]; *Lip makeup* [0, 1]	Presence of cosmetics in the face.
Accessory	*Head wear* [0, 1]; *Glasses* [0, 1]	Presence of any facial accessory.
Face Orientation	*Head roll* [-180, 180]; *Head yaw* [-180, 180]; *Head pitch* [-180, 180]	Spatial orientation of the face.
Face Occlusion	*Occluded forehead* {0, 1}; *Occluded eyes* {0, 1}; *Occluded mouth* {0, 1}; *Face exposure* [0, 1]	Occlusion of face parts.
Image Distortion	*Blur* [0, 1]; *Noise* [0, 1]	Noise and blur in the image.
Emotional	*Smile* [0, 1]	Presence of smile in the represented face.

[1] Value ranges for continuous variables are reported as [X,Y], whereas value ranges for discrete variables are reported as {X,Y}.

Table 1. Face characteristics whose influence on error rates is investigated in our study.

acting as a gradient update filter during training and as a feature selector during inference, and (ii) a trunk branch for feature processing during both phases. `ResNeSt` [17] is characterized by split-attention blocks, each with a feature map group and split attention operations. `RepVGG` [18] uses the relative simplicity of its structure as its strength. The inference is performed through a series of 3×3 and ReLU convolutions, while the training layers follow a multi-branch topology. Finally, `HRNet` [19] maintains a high resolution throughout the series of convolutions, thanks to parallel connections of the convolutional streams and continuous exchange of information between different resolutions.

Each backbone was plugged into a head network for the final classification, during training. Considering the same face recognition benchmark [1], we selected `MagFace` and `NPCFace` as head networks. `MagFace` is a refined implementation of the well-known ArcFace [20]. This head adds an additive angular margin penalty to enhance intra-class compactness and inter-class discrepancy. On the other hand, `NPCFace` [21] emphasizes the training on both negative and positive hard cases via the collaborative-margin mechanism in softmax logits.

We finally trained one face encoder for each combination of backbone and head network (5 backbones × 2 head networks = 10 face encoders) on images from the `DiveFace` training set. Each face encoder was trained for a maximum of 80 epochs (early stop, patience 5), with a batch size of 64. We used Categorical Cross-entropy as the loss function, SGD as the optimizer, with momentum 0.9, weight decay $1e-8$, and initial learning rate 0.1.

Face Encoder Evaluation. Once trained, for each face encoder, we unplugged the head network such that each face encoder would return as an output the latent representation (size: 512) of the face image given as an input. With the face images of individuals included in the test set (disjoint set of individuals with respect to the training set), we then simulated a face verification task by creating a range of trial verification pairs for each individual: 6 positive pairs[2] with both images coming from the same person and 50 negative pairs with the second image in the pair coming from another person. Subsequently, for each face encoder and trial verification pair, we extracted the 512-sized latent representations of the two face images and then computed the Cosine similarity (range $[-1, 1]$) between them.

Once we collected all the Cosine similarity scores resulting from a given face encoder, we determined the three thresholds that would lead to three well-known security levels, implemented to achieve a fixed false acceptance rate (FAR) of 1%, 0.1%, and 0.01% respectively. The false acceptance rate measures the likelihood that the system will incorrectly accept an access attempt by an unauthorized user. A false acceptance is often the most serious error as it gives access to unauthorized users. These security levels were selected due to their popular adoption in prior work [1].

Finally, using the threshold resulted for a given security level, we computed the corresponding false rejection rate (FRR@FAR 1%, 0.1%, and 0.01%, respectively) for each individual. The false rejection rate at a given false acceptance rate measures the likelihood that the system will incorrectly reject an access attempt by an authorized individual, by using a threshold that would force the system to maintain a target false acceptance rate. The higher the false rejection rate is, the lower the usability of the corresponding face recognition system at that security level is.

3. Results

Our experiments analyzed the extent to which disparities among groups change based on the security level (*RQ1*) and whether any co-relationships between face characteristics and false rejection rates exist across such levels (*RQ2*).

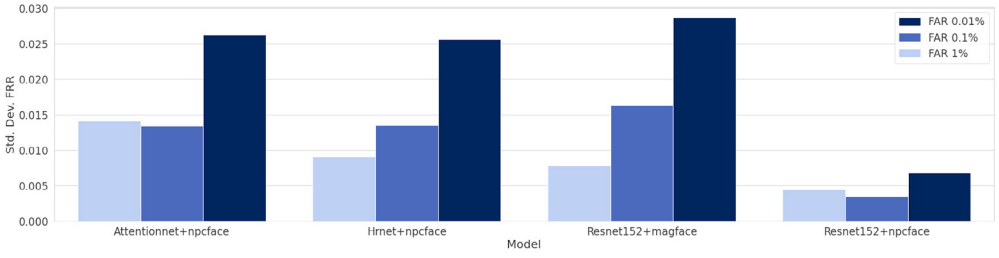
[2] Since the minimum number of images per person was 4, we could generate $(4 \times 3)/2$ positive pairs from the images belonging to each person.

3.1. Disparate Impact on Usability across Security Levels (RQ1)

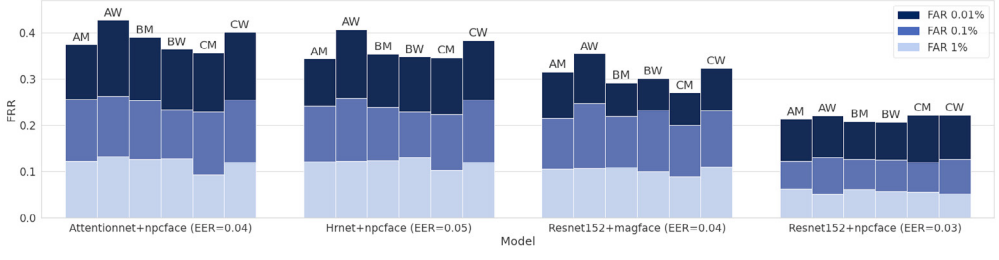
In a first analysis, we investigated whether the disparities among demographic groups change under different security levels and, if so, whether there is any relationship between the degree of disparity and the security level.

To this end, for each trained face encoder, we first computed the standard deviation of false rejection rates among demographic groups under each of the three security levels (light blue bar: FRR@FAR 1%; mid blue bar: FRR@FAR 0.1%; dark blue bar: FRR@FAR 0.01%). The higher the standard deviation in false rejection rate is, the higher the disparate impact on usability is among demographic groups (and therefore the higher the unfairness). Figure 1a collects the standard deviation of false rejection rates for the four best performing face encoders previously trained (due to space constraints). The other face encoders showed on average coherent result patterns. Specifically, it can be observed that the highest standard deviation of false rejection rates was measured under the most secure level (dark blue bars) for all the four face encoders. This standard deviation ranged between 0.025 and 0.030, except for ResNet152 + npcface (0.006). Considering the other two less secure thresholds, the increasing trend in standard deviation was less evident but still present. In some cases (AttentionNet + npcface and ResNet152 + npcface), conversely, there was a minimal negligible decrease in standard deviation for FAR 0.1%, with respect to FAR 1%. We conjecture that the backbones behind those two face encoders might be more robust to a security level change, when such security levels are not relatively high, for instance due to a smaller threshold change between the two. We can conclude that there exists a general trend showing that the higher the security level is, the higher the disparate impact is.

To have a more detailed picture, we analyzed the results at group level. Specifically, Figure 1b reports the false rejection rates experienced by individuals of the six demographic groups, namely Asian Men (AM), Asian Women (AW), Black Men (BM), Black Women (BW), Caucasian Men (CM), and Caucasian Women (CW), under the three security thresholds. As expected, the false rejection rate increased as the security level increased for all the four face encoders, though this increment was smaller for ResNet152 + npcface. Under our analyses at the most secure threshold, Asian Women (AW) suffered from the highest false rejection rate whereas the lowest error rate was, in three out of four face encoders, reported for Caucasian Men (CM). This pattern tended to be confirmed but less evident under the other two security levels. Interestingly, within the same gender group, Asian (Black) individuals often had the highest error rate within Women (Men). On the other hand, within the same ethnicity group, Women (Men) were disadvantaged within Asians and Caucasians (Black). The disadvantaged ethnicity group depended on the gender and viceversa. This confirmed that disparate impacts are a complex phenomenon involving multiple covariates.



(a) Std. Dev. of False Rejection Rates (FRRs) among groups under a given False Acceptance Rate (FAR) threshold.



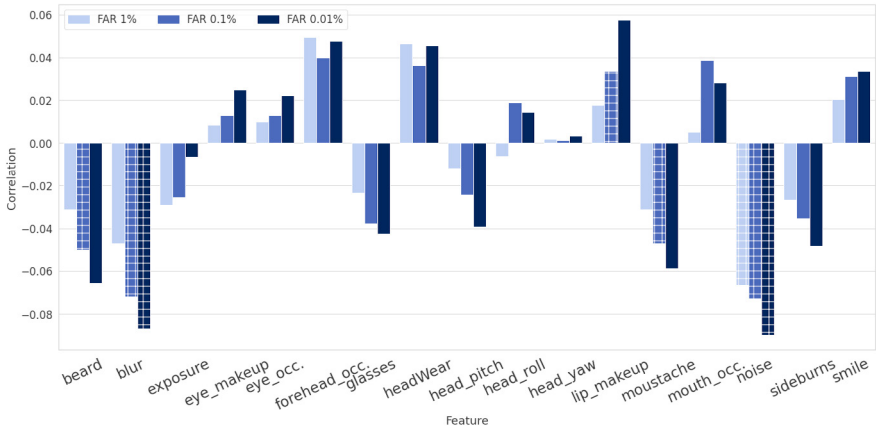
(b) False Rejection Rate (FRR) under a given False Acceptance Rate (FAR) threshold for each group.

Fig. 1. Analysis of the impact of the security threshold on the false rejection rates and their standard deviation over groups.

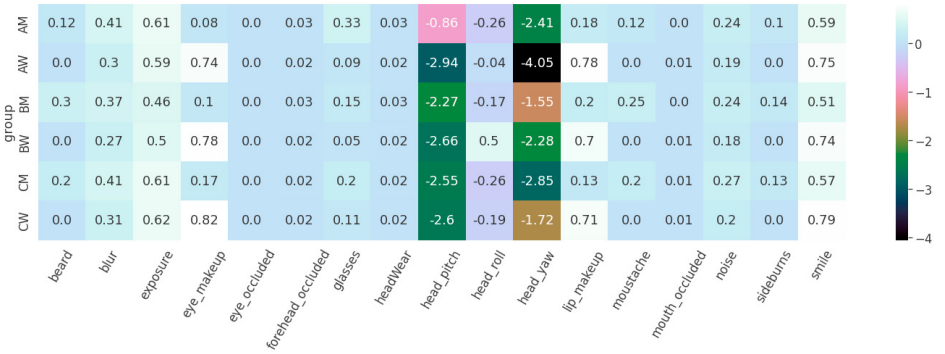
3.2. Co-Relationships between Face Characteristics and Disparate Usability across Security Levels (RQ2)

In a second analysis, we investigated the existence of co-relationships between face characteristics and false rejection rates under the same three security levels and the extent to which such characteristics are present in certain groups in the test set. With this in mind, for each face encoder, we computed the Pearson correlation between the value of a given face characteristic score (across images of the same individual) and the corresponding false rejection rate measured for that individual. Figure 2a shows the correlation scores averaged across face encoders, under a given security level. The bar texture is 'o' ('+') in case of p-value < 0.05 (< 0.03). Our results show that false rejection rates had a weak positive correlation with smile, make-up, and face occlusion. Weak negative correlations were observed for facial hair, glasses, and image distortion. These correlations tended to be stronger as much as the security increased.

To link back our correlation analysis to the disparate impacts observed in the previous section, Figure 2b collects the average score (for conciseness) for each face characteristic, measured on images coming from a certain demographic group. For instance, within each ethnicity group, our previous analysis showed that Women consistently experienced the highest false rejection rate. Indeed, face images representing Women tended to report the highest average score for smile, make-up, and face occlusion (except for Caucasians), that were characterized by a positive correlation with the false rejection rate. Women images also had lower scores for those face characteristics having a weak negative correlation with the false rejection rate, such as facial hair, glasses, noise, and blur. Similarly, within each gender group, Asians were often the most disadvantaged group. The observations on their face characteristics tended to be in line with those made for Women. Again, the presence of such disparate impacts depends on a complex combination of face characteristics. Therefore, in future work, we will give emphasis on the analysis of multiple face characteristics jointly, going beyond their individual inspection.



(a) Pearson correlation between a face characteristic score for images of an individual and their corresponding false rejection rate under a given security level, averaged across face encoders.



(b) The average (for conciseness) face characteristic score for images from a certain group in the test set.

Fig. 2. Analysis of the relationship between face characteristics and false rejection rates.

4. Conclusions and Future Work

In this paper, we analyzed the influence of the security level on the disparate impacts emphasized by a range of face recognition systems. We also investigated the co-relationships between relevant face characteristics, the demographic group membership, and the estimated false rejection rates. Our results revealed a general trend that the higher the security level is, the higher the disparate usability among groups is. Moreover, there are key face characteristics more present in certain groups, whose correlation with false rejection rates increases across security levels.

Our findings, together with the limitations of our study, will be the drivers for our next steps. First, we plan to extend the set of face characteristics under consideration and explain their influence through other explainability techniques. Additional datasets will be included in our analysis and we will address both verification and identification scenarios. Though we measured correlation coefficients, they are not really sufficient to explain first- or second-order situations (correlation does not imply causation). We will therefore deepen the analysis of these dependencies with multi-factor causal models. Finally, knowledge about face characteristics and their co-relationships with false rejection rates will be used as a proxy for designing unfairness mitigation methods that do not require protected attribute labels.

References

[1] J. Wang, Y. Liu, Y. Hu, H. Shi, T. Mei, Facex-zoo: A pytorch toolbox for face recognition, in: Proc. of the 29th ACM International Conference on Multimedia (ACM MM 2021), 2021, pp. 3779–3782.

[2] M. Gwilliam, S. Hegde, L. Tinubu, A. Hanson, Rethinking common assumptions to mitigate racial bias in face recognition datasets, in: Proc. of the IEEE/CVF In. Conf. on Computer Vision (CVPR 2021), 2021, pp. 4123–4132.

[3] V. Albiero, K. KS, K. Vangara, K. Zhang, M. C. King, K. W. Bowyer, Analysis of gender inequality in face recognition accuracy, in: Proc. of the IEEE/CVF Winter Conf. on App. of Computer Vision Workshops, 2020, pp. 81–89.

[4] V. Albiero, K. W. Bowyer, Is face recognition sexist? no, gendered hairstyles and biology are, arXiv preprint arXiv:2008.06989 (2020).

[5] K. R. Jr., S. Bhardwaj, M. Sodomsky, A review of face recognition against longitudinal child faces, in: Proc. of the 14th International Conference of the Biometrics Special Interest Group (BIOSIG 2015), Vol. P-245 of LNI, 2015, pp. 15–26.

[6] J. Yu, X. Hao, H. Xie, Y. Yu, Fair face recognition using data balancing, enhancement and fusion, in: Proc. of the European Conference on Computer Vision (ECCV 2020), Springer, 2020, pp. 492–505.

[7] M. Wang, Y. Zhang, W. Deng, Meta balanced network for fair face recognition, IEEE Tran. on Patt. Analysis and Machine Intel. (2021) 1–1.

[8] I. Hupont, C. Fernández, Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition, in: Proc. of the 14th IEEE In. Conf. on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–7.

[9] P. Majumdar, S. Mittal, R. Singh, M. Vatsa, Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models, in: Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), 2021, pp. 3786–3795.

[10] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, A. Kuijper, A comprehensive study on face recognition biases beyond demographics, IEEE Transactions on Technology and Society 3 (1) (2021) 16–30.

[11] A. Atzori, G. Fenu, M. Marras, Explaining bias in deep face recognition via image characteristics, in: Proc. of the 2022 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2022.

[12] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana, Sensitivenets: Learning agnostic representations with application to face images, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (6) (2020) 2158–2164.

[13] S. I. Serengil, A. Ozpinar, Lightface: A hybrid deep face recognition framework, in: Proc. of the Innovations in Intelligent Systems and Applications Conference (ASYU 2020), IEEE, 2020, pp. 1–5.

[14] T. Sixta, J. Jacques Junior, P. Buch-Cardona, E. Vazquez, S. Escalera, Fairface challenge at eccv 2020: Analyzing bias in face recognition, in: Prof. of the European Conf. on Computer Vision (ECCV 2020), Springer, 2020, pp. 463–481.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016, pp. 770–778.

[16] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2017), 2017, pp. 3156–3164.

[17] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Muller, R. Manmatha, M. Li, A. Smola, Resnest: Split-attention networks, arXiv preprint arXiv:2004.08955 (2020).

[18] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: Making vgg-style convnets great again, in: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), 2021, pp. 13733–13742.

[19] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (10) (2020) 3349–3364.

[20] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), 2019, pp. 4690–4699.

[21] D. Zeng, H. Shi, H. Du, J. Wang, Z. Lei, T. Mei, Npcface: Negative-positive collaborative training for large-scale face recognition, arXiv preprint arXiv:2007.10172 (2020).