Race Bias Analysis of Bona Fide Errors in Face Anti-spoofing

Latifah Abduh(✉) and Ioannis Ivrissimtzis

Durham University, Durham DH1 3LE, UK
{latifah.a.abduh,ioannis.ivrissimtzis}@durham.ac.uk

Abstract. The study of bias in Machine Learning is receiving a lot of attention in recent years, however, few only papers deal explicitly with the problem of race bias in face anti-spoofing. In this paper, we present a systematic study of race bias in face anti-spoofing with three key features: we focus on the classifier's bona fide errors, where the most significant ethical and legal issues lie; we analyse both the scalar responses of the classifier and its final binary outcomes; the threshold determining the operating point of the classifier is treated as a variable. We apply the proposed bias analysis framework on a VQ-VAE-based face anti-spoofing algorithm. Our main conclusion is that race bias should not necessarily be attributed to different mean values of the response distributions over the various demographics. Instead, it can be better understood as the combined effect of several possible characteristics of these distributions: different means; different variances; bimodal behaviour; the existence of outliers.

Keywords: Face presentation attacks · face anti-spoofing · race bias

1 Introduction

Face recognition is the method of choice behind some of the most widely deployed biometric authentication systems, currently supporting a range of applications, from passport control at airports to mobile phone or laptop login. A key weakness of the technology, is its vulnerability to *presentation attacks*, where imposters attempt to gain wrongful access by presenting in front of the system's camera a photo, or a video, or by wearing a mask resembling a registered person. As a solution to this problem, algorithms for presentation attack detection (PAD) are developed, that is, binary classifiers trained to distinguish between the bona fide samples coming from live subjects, and those coming from imposters.

Here, we deal with the problem of race bias in face anti-spoofing algorithms. The proposed race bias analysis process has three key characteristics. First, the focus is on the bona fide error, that is, on genuine people wrongly classified as imposters. Biases in this type of error have significant ethical, legal and regulatory ramifications, and as it has recently pointed out "creates customer annoyance and inconvenience", [12]. Secondly, we do not analyse just the final binary

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
N. Tsapatsoulis et al. (Eds.): CAIP 2023, LNCS 14185, pp. 23–32, 2023.
https://doi.org/10.1007/978-3-031-44240-7_3

classification outcome, but also the scalar responses of the network prior to thresholding. Thirdly, we treat the value of the threshold, which determines the classifier's operating point on the ROC curve, as a variable. We do not assume it is fixed by the vendor of the biometric verification system in a black-box process.

We demonstrate the proposed bias analysis approach on a face anti-spoofing algorithm based on the Vector Quantized Variational Autoencoder (VQ-VAE) architecture, [20]. The network is trained and validated on the SiW database, and tested for bona fide racial bias on the SiW and RFW databases. Hypotheses are tested using the chi-squared test on the binary outcomes, the Mann-Whitney U test on the scalar responses, and the Hartigan's Dip for testing bimodality in the response distributions.

Our main finding is that racial bias can be attributed to several characteristics of the response distributions at the various demographics: different means; different variances; bimodality; outliers. As a secondary contribution, we also demonstrate that a database which does not specialise in face anti-spoofing, such as RFW, can nevertheless be used to analyse face anti-spoofing algorithms.

The rest of the paper is organised as follows. In Sect. 2, we review the relevant literature. In Sect. 3, we describe the experimental setup. In Sects. 4, and 5 we present the bias analysis on the SiW and RFW databases, respectively. We briefly conclude in Sect. 6.

2 Background

We briefly review the area of face anti-spoofing, and then focus on previous studies of bias in machine learning, and PAD in particular.

2.1 Face Anti-spoofing

The state-of-the-art in face anti-spoofing [5,14,25–28,30,31], is based on various forms of deep learning, such as Central Difference Convolutional Networks (CDCN) [27,28], or transformers [23]. Following some earlier approaches [4,15], the state-of-the-art may also utilise depth information [22,24,25,30], usually estimated by an independently trained neural network, while the use GAN estimated Near Infrared (NIR) information was proposed [14].

Regarding the face anti-spoofing databases we use in this paper, our training dataset is from the *SiW* database, introduced in [15]. It comprises videos of 165 subjects of four types of ethnicities: 35% of Asian and 35% Caucasian and 23% Indian, and 7% African American. The bias analysis is performed on SiW with the subject annotated for ethnicity type by us, and the already annotated RFW database [21], which is widely used in the bias analysis literature. RFW again comprises four types of ethnicities: Caucasian, Asian, Indian, and African.

2.2 Bias in Machine Learning

Because of the ethical, legal, and regulatory issues associated with the problem of bias within human populations, there is a considerable amount of research on the subject, especially in face recognition (FR). A recent comprehensive survey can be found in [17], where the significant sources of bias are categorised and discussed, and the negative effect of bias on downstream tasks is pointed out.

In one of the earliest studies of bias in FR, predating deep learning, [18] reported differences in the performance on humans of Caucasian and East Asian descent between Western and East Asia developed algorithms. In [9], several deep learning-based FR algorithms are analysed and a small amount of bias is detected in all of them.

In [10], the authors compute cluster validation measures on the clusters of the various demographics inside the whole population, aiming at measuring the algorithm's potential for bias, rather than actual bias. Their result is negative, and they argue for the need of more sophisticated clustering approaches. In [19], the aim is the detection of bias by analysing the activation ratios at the various layers of the network. Similarly to our work, their target application is the detection of race bias on a binary classification problem, gender classification in their case. Their result is positive in that they report a correlation between the measured activation ratios and bias in the final outcomes of the classifier. However, it is not clear if their method can be used to measure and assess the statistical significance of the expected bias.

In Cavazos et al. [6], similarly to our approach, most of the analysis assumes a one-sided error cost, in their case the false acceptance rate, and the operating thresholds are treated as user-defined variables. However, the analytical tools they used, mostly visual inspection of ROC curves, do not allow for a deep study of the distributions of the similarity scores, while, here, we give a more in-depth analysis of the response distributions, which is the equivalent of the similarity scores. In Pereira and Marcel [8], a fairness metric is proposed, which can be optimised over the decision thresholds, but again, there is no in-depth statistical analysis of the scores.

The literature on bias in presentation attacks is more sparse. Race bias was the key theme in the competition of face anti-spoofing algorithms on the CASIA-SURF CeFA database [13]. Bias was assessed by the performance of the algorithm under a cross-ethnicity validation scenario. Standard performance metrics, such as APCER, BPCER and ACER we reported. In [2], the standard CNN models Resnet 50 and VGG16, were compared for gender bias against the debiasing-VAE proposed in [3], and several performance metrics were reported. A recent white paper by the ID R&D company presents the results of a large-scale bias assessment experiment conducted by Bixelab, a NIST accredited independent laboratory [12]. Similarly to our approach, they focus on bona fide errors, and their aim is for the BPCER error metric to be below a prespecified threshold across all demographics. Regarding other biometric identification modalities, [7] studied gender bias in iris PAD algorithms.

3 Experimental Setup

We chose the VQ-VAE architecture because of some recently reported impressive results on various computer vision problems. For a more detailed description of the classifier, see our Arxiv preprint [1].

3.1 The VQ-VAE Classifier

The encoder consists of two convolutional layers of kernel size 4, stride step 2, padding 1; followed by a ReLU; one convolutional layer of kernel size 3, stride step 1, padding 1; followed by two residual blocks implemented as ReLU (3×3 conv, ReLU, 1×1 conv for each block). It outputs a 16×16 grid of vectors quantized on a codebook of size 512. The decoder is symmetrical to the encoder, using transposed convolutions. The model was ADAM optimised with learning rate 1e-3, for 100 epochs, with batch size 16. The weight factor β was set to 0.25.

For face detection we used the Multi-Task Cascade Convolutional Neural Network (MTCNN) [29]. The detected faces were horizontally aligned, and cropped at 64×64. As our classifier is based on anomaly detection, the training set consisted of bona fide only data, 124,000 samples. We assessed performance on a test set of 1,600 samples, 400 samples from each race, with equal split between bona fide and attack. At an operating threshold of 0.054, corresponding to the EER value at an independent validation set, we obtained an HTER of 0.169, which indicates satisfactory performance.

3.2 Overview of the Bias Analysis Process

The bias analysis process is summarised in Fig. 1. The binary outcomes of the classifiers are analysed with the chi-squared test, and the scalar responses with the Mann-Whitney U test [16].

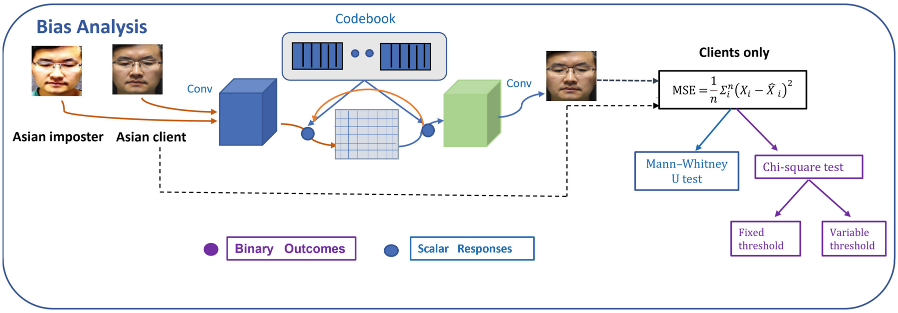

Fig. 1. The bias analysis process. The binary outcome analysis is shown in purple and the scalar response analysis in blue. (Color figure online)

4 Bias Analysis on SiW

We perform bias analysis on the bona fide samples of SiW test set in Sect. 3.

4.1 Statistical Analysis of the Binary Outcomes

First, we analyse the binary outcomes corresponding to the operating threshold 0.054, which was used in the validation of the classifier in Sect. 3. For each pair of races, we form the 2 × 2 contingency tables, and apply the chi-squared test, computing p-values for the hypothesis that samples from the race with the most misclassifications have higher misclassification probability. The results are summarised in Table 1. In several cases, the p-values are low, meaning that for any reasonable threshold of statistical significance, the bias hypothesis is accepted. In other cases, p-values above 0.05 mean that bias has not been detected.

Table 1. p-values of the chi-squared tests for the 0.054 threshold used in Sect. 3.

Af-As	Af-Ca	Af-In	As-Ca	As-In	Ca-In
0.1158	0.0104	0.0147	0.0000	0.0000	1.0000

Next, we treat the operating threshold as a variable. Figure 2 shows the p-values as a function of the threshold for the six pairs of races. We notice that, over the range of all thresholds, there could be several disconnected intervals corresponding to high bias (low p-values), which means that threshold optimisation for low bias should not assume a unique solution, as it is often implicitly assumed in the literature.

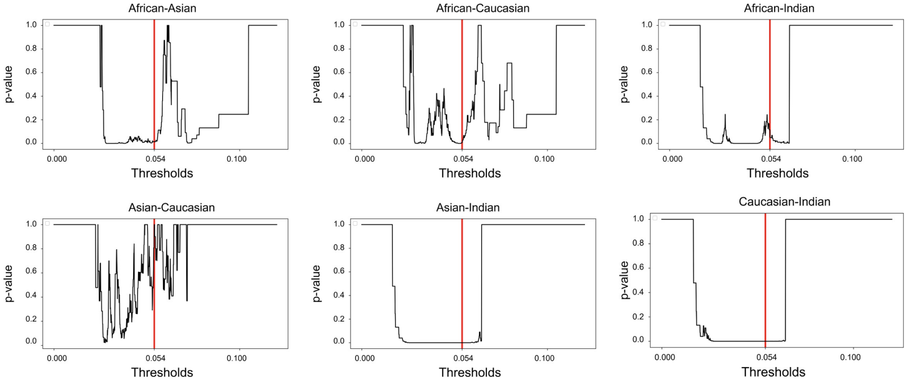

Fig. 2. For each pair of races, graphs of the p-value as a function of the threshold.

4.2 Statistical Analysis of the Scalar Responses

For an insight in the behaviour of the graphs in Fig. 2, we analyse the classifier's scalar responses on the premise that a complex behaviour of their density fun ctions, will induce complex bias behaviour. Table 2 summarises the statistics computed on the responses of each race: mean, standard deviation, and Hartigan's Dip value [11]. Figure 3 shows plots of histograms and density functions for each pair of races.

Table 2. Response means, st. dev., and Hartigan's dip values for each race in SiW.

	Af	As	Ca	In
μ	0.0418	0.0446	0.0438	0.0355
$s.d.$	0.0142	0.0109	0.0122	0.0096
dip	0.0299	0.0233	0.0366	0.0324

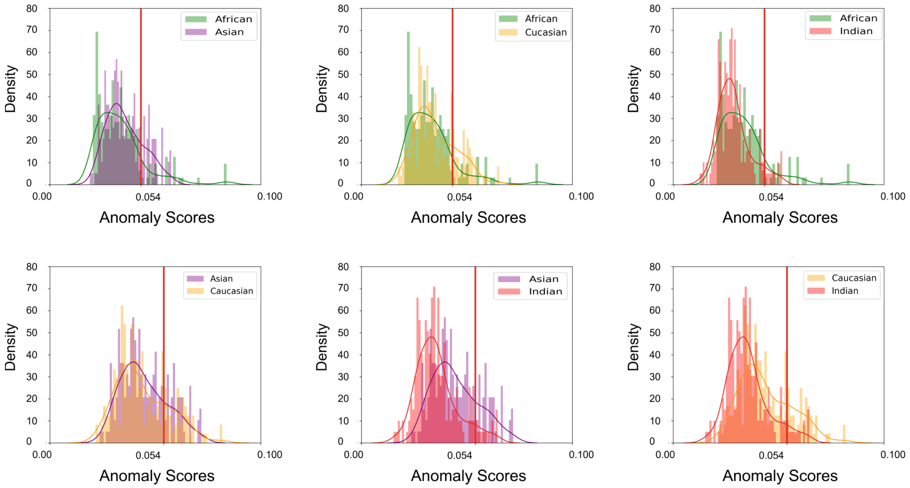

Fig. 3. For each pair of races in SiW, histograms and density functions of the responses.

We tested for statistically different mean responses with the Mann-Whitney U test, as the Shapiro-Wilk test rejected the normality hypothesis. Table 3 shows for each pair of races p-values for the hypothesis that randomly selected responses from the two populations have different values. We note that, for example, the p-value of the Asian and Indian pair is very low, and the large range of high bias thresholds in the corresponding U-shaped diagram in Fig. 2 is due to a statistically significant higher mean response on Asians compared to Indians.

Table 3. p-values of the Mann-Whitney U test on each pair of races.

Af-As	Af-Ca	Af-In	As-Ca	As-In	Ca-In
0.0001	0.0078	0.0000	0.1560	0.0000	0.0000

In contrast, the mean response difference between Asians and Caucasians is not statistically significant. Thus, the bias we can observe in the corresponding diagram in Fig. 2, which for small threshold values on the left-hand side of the diagram is statistically significant, is due to different standard deviations.

We checked for bimodality using Hartigan's Dip Test with 50 bins. For the 200 samples we have from a race, a statistical significance of 95% corresponds to a critical value of 0.037. We notice that all Dip values are below the significance threshold, and thus, all populations should be considered unimodal. In particular, that means that some very high responses on African people should be treated as outliers. We note that against all the other three races, these outliers create a second, or third region of high bias thresholds, in which regions samples from the African population are treated less favourably.

5 Bias Analysis on RFW

Here, we apply the same analysis on a test set from the RFW database, consisting of 200 images from each race. This time the race labels are part of the database, rather than being annotated by us. As RFW database is not a specialised face anti-spoofing database, we do not have imposter images and thus we do not have empirically established operating thresholds, as for example the ones corresponding to EER values. Instead, in our diagrams we indicate thresholds corresponding to bona fide error rates of: 1%, 2%, 5%, 10%, 20%.

In Fig. 4, for each race pair, we plot the p-values of the chi-squared test as a function of the threshold. We observe behaviours similar to those in Sect. 4.

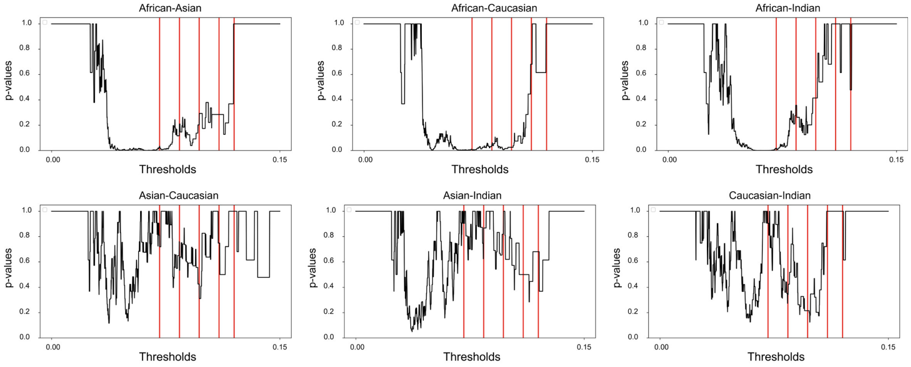

Fig. 4. For each pair of races, graphs of the p-value as a function of the threshold.

Table 4 shows the means, standard deviations and dip values for each race, and Table 5 shows the p-values of the Mann-Whitney U test for each race pair. We note in Table 4 that the Hartigan's test detects a bimodality in the responses on Indian people, having a dip value of 0.055, above the significance threshold of 0.037. This can also be verified by visual inspection of the corresponding histograms and density functions, shown in Figs. 5 for race pairs. We also note that this bimodality can be detected in the behaviour of the corresponding graphs of the p-values of the chi-squared test. Indeed, in the three graphs in Fig. 4 corresponding to Indian people, we can detect two distinct regions of higher bias, even though the second one does not reach the level of statistical significance.

Table 4. Response means, st. dev., and Hartigan's dip values for each race in SiW.

	Af	As	Ca	In
μ	0.0509	0.0579	0.0569	0.0579
$s.d.$	0.0175	0.0220	0.0223	0.0220
dip	0.0114	0.0225	0.0149	0.0550

Table 5. p-values of the Mann-Whitney U test on each pair of races.

Af-As	Af-Ca	Af-In	As-Ca	As-In	Ca-In
0.0004	0.0058	0.0062	0.2509	0.2743	0.4805

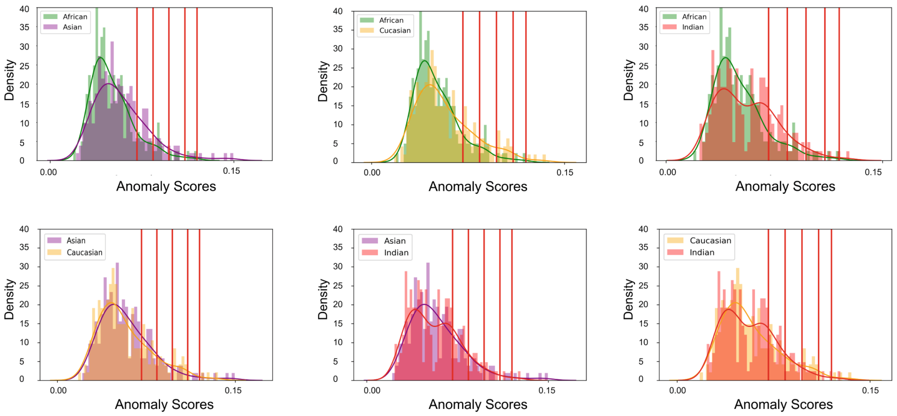

Fig. 5. For each race pair in RFW, histograms and density functions of the responses.

6 Conclusion

We conducted an empirical study of race bias in face anti-spoofing with the following characteristics: we analysed the bona fide error; the classifier's binary outcomes and scalar responses were both analysed for bias; the threshold determining the classifier's operating point was considered a variable.

Our main finding is that the behaviour of race bias depends on several characteristics of the response distributions: different means or different variances between two demographics; bimodality or existence of outliers in a certain demographic. The implication is that race bias is cannot always be attributed to different mean responses, a misconception sometimes reinforced by the fact that in statistics, colloquially, the term bias is often used to describe the component of the error corresponding to the difference in means. As a practical implication of our findings, we note that methods for automatically choosing low bias thresholds should not assume a unique solution to the problem.

In our future work, we would like to conduct a theoretical study of bias, assuming, for example, that the responses follow log-normal distributions.

References

1. Abduh, L., Ivrissimtzis, I.: Race bias analysis of bona fide errors in face anti-spoofing. arXiv:2210.05366 (2022)
2. Alshareef, N., Yuan, X., Roy, K., Atay, M.: A study of gender bias in face presentation attack and its mitigation. Future Internet **13**(9), 234 (2021)
3. Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 289–295 (2019)
4. Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face anti-spoofing using patch and depth-based CNNs. In: Proceedings of IJCB, pp. 319–328. IEEE (2017)
5. Cai, R., Li, H., Wang, S., Chen, C., Kot, A.C.: DRL-FAS: a novel framework based on deep reinforcement learning for face anti-spoofing. IEEE TIFS **16**, 937–951 (2020)
6. Cavazos, J.G., Phillips, P.J., Castillo, C.D., O'Toole, A.J.: Accuracy comparison across face recognition algorithms: where are we on measuring race bias? IEEE TBBIS **3**(1), 101–111 (2020)
7. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: Demographic bias in presentation attack detection of iris recognition systems. In: Proceedings of EUSIPCO, pp. 835–839. IEEE (2021)
8. de Freitas Pereira, T., Marcel, S.: Fairness in biometrics: a figure of merit to assess biometric verification systems. IEEE TBBIS **4**(1), 19–29 (2021)
9. Garcia, R.V., Wandzik, L., Grabner, L., Krueger, J.: The harms of demographic bias in deep face recognition research. In: Proceedings of ICB, pp. 1–6. IEEE (2019)
10. Glüge, S., Amirian, M., Flumini, D., Stadelmann, T.: How (not) to measure bias in face recognition networks. In: Schilling, F.-P., Stadelmann, T. (eds.) ANNPR 2020. LNCS (LNAI), vol. 12294, pp. 125–137. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58309-5_10
11. Hartigan, J.A., Hartigan, P.M.: The dip test of unimodality. Ann. Stat. 70–84 (1985)

12. ID R&D whitepaper: Mitigating Demographic Bias in Facial Presentation Attack Detection (2022). https://idrnd.ai/mitigating-bias-in-biometric-facial-liveness-detection
13. Liu, A., et al.: Cross-ethnicity face anti-spoofing recognition challenge: a review (2020)
14. Liu, A., et al.: Face anti-spoofing via adversarial cross-modality translation. IEEE TIFS **16**, 2759–2772 (2021)
15. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: Proceedings of CVPR (2018)
16. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**(1), 50–60 (1947)
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021)
18. Phillips, P.J., Jiang, F., Narvekar, A., Ayyad, J., O'Toole, A.J.: An other-race effect for face recognition algorithms. ACM Trans. Appl. Percept. **8**(2), 1–11 (2011)
19. Serna, I., Peña, A., Morales, A., Fierrez, J.: Insidebias: measuring bias in deep networks and application to face gender biometrics. In: Proceedings of ICPR, pp. 3720–3727. IEEE (2021)
20. Van Den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
21. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: reducing racial bias by information maximization adaptation network. In: Proceedings of ICCV, pp. 692–702. IEEE (2019)
22. Wang, Z., et al.: Deep spatial gradient and temporal depth learning for face anti-spoofing. In: Proceedings of CVPR, pp. 5042–5051 (2020)
23. Wang, Z., Wang, Q., Deng, W., Guo, G.: Face anti-spoofing using transformers with relation-aware mechanism. IEEE TBBIS **4**(3), 439–450 (2022)
24. Wu, H., Zeng, D., Hu, Y., Shi, H., Mei, T.: Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning. IEEE TCSVT **32**(7), 4626–4638 (2022)
25. Yu, Z., Li, X., Shi, J., Xia, Z., Zhao, G.: Revisiting pixel-wise supervision for face anti-spoofing. IEEE TBBIS **3**(3), 285–295 (2021)
26. Yu, Z., et al.: Auto-FAS: searching lightweight networks for face anti-spoofing. In: Proceedings of ICASSP, pp. 996–1000. IEEE (2020)
27. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: NAS-FAS: static-dynamic central difference network search for face anti-spoofing. IEEE TPAMI **43**(9), 3005–3023 (2020)
28. Yu, Z., et al.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of CVPR (2020)
29. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
30. Zhang, K.-Y., et al.: Face anti-spoofing via disentangled representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12364, pp. 641–657. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_38
31. Zhang, Y., et al.: CelebA-spoof: large-scale face anti-spoofing dataset with rich annotations. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 70–85. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_5