Thinking Racial Bias in Fair Forgery Detection: Models, Datasets and Evaluations

Decheng Liu[1*], **Zongqi Wang**[2*], **Chunlei Peng**[1†], **Nannan Wang**[1], **Ruimin Hu**[1], **Xinbo Gao**[3]

[1]Xidian University, Xi'an, China
[2]Tsinghua University, Beijing, China
[3]Chongqing University of Posts and Telecommunications, Chongqing, China
dchliu@xidian.edu.cn, zq-wang24@mails.tsinghua.edu.cn, clpeng@xidian.edu.cn, nnwang@xidian.edu.cn, hrm1964@163.com, gaoxb@cqupt.edu.cn

Abstract

Due to the successful development of deep image generation technology, forgery detection plays a more important role in social and economic security. Racial bias has not been explored thoroughly in the deep forgery detection field. In the paper, we first contribute a dedicated dataset called the Fair Forgery Detection (FairFD) dataset, where we prove the racial bias of public state-of-the-art (SOTA) methods. Different from existing forgery detection datasets, the self-constructed FairFD dataset contains a balanced racial ratio and diverse forgery generation images with the largest-scale subjects. Additionally, we identify the problems with naive fairness metrics when benchmarking forgery detection models. To comprehensively evaluate fairness, we design novel metrics including Approach Averaged Metric and Utility Regularized Metric, which can avoid deceptive results. We also present an effective and robust post-processing technique, Bias Pruning with Fair Activations (BPFA), which improves fairness without requiring retraining or weight updates. Extensive experiments conducted with 12 representative forgery detection models demonstrate the value of the proposed dataset and the reasonability of the designed fairness metrics. By applying the BPFA to the existing fairest detector, we achieve a new SOTA. Furthermore, we conduct more in-depth analyses to offer more insights to inspire researchers in the community.

Introduction

Face forgery refers to the creation of fake images or videos of a person's face using conventional techniques or deep learning methods. These forgeries can be used to spread misinformation, commit fraud, or even blackmail people. There are numerous methods are proposed for detecting face forgery (Rossler et al. 2019; Afchar et al. 2018; Wang et al. 2020; Tan and Le 2019; Nguyen, Yamagishi, and Echizen 2019; Li and Lyu 1811; Li et al. 2020a; Dang et al. 2020; Ni et al. 2022; Cao et al. 2022; Yan et al. 2023a; Qian et al. 2020; Liu et al. 2021; Luo et al. 2021). Although an increasing number of advanced face forgery detection technologies are being developed, the racial fairness of these detectors is consistently overlooked by researchers (Masood et al. 2023). Detectors with severe racial bias can lead to significant social impact. These detectors might disproportionately label faces from a particular racial group as fake, thereby indicating discrimination towards this particular racial group. Therefore, when a detector is ready for deployment, evaluating and analyzing its fairness is a crucial process. However, although there is extensive available research about the fairness in machine learning to draw upon (Agarwal et al. 2018; Agarwal, Dudík, and Wu 2019; Hardt, Price, and Srebro 2016; Tian et al. 2024; Shen et al. 2024), evaluating the fairness in face forgery detection systems remains difficult. This is due to several distinct differences between face forgery detection and other deep learning tasks.

This work aims to fill the gap in research on racial fairness in face forgery detection by proposing an accurate, comprehensive and credible fairness evaluation system. To achieve this goal, we analyze the shortcomings of existing evaluation components (i.e. dataset and metric), and our corresponding solutions. **(1) Dataset.** Existing face forgery detection datasets have a limited number of subjects. We find performance fluctuations significantly across subjects, so individual fairness may overshadow group fairness, which will make the evaluation results inaccurate. It also can be found that different forgery approaches have different fairness levels, limited forgery approaches will lead to a non-comprehensive result. Otherwise, undefined ethnicity (faces from two ethnicities are swapped) will lead to an inaccurate result. **(2) Metric.** We also propose two issues (Bias Offset and Aggregation Distortion) that will cause deceptive results. Bias Offset arises because existing fairness metrics typically use overall average accuracy for calculations instead of assessing each forgery method separately. This way may obscure some biases as different forgery methods may have different privileged races. Aggregation Distortion arises because detectors often show significant utility variations across different forgery techniques. Even if two forgery methods exhibit the same bias, detectors with lower utility can be more unfair. Treating each forgery method as equal will lead to unreliable results.

To tackle these problems, we firstly introduce the Fair Forgery Detection (FairFD) dataset for racial bias evaluation, which contains the largest scale subjects, and incorporates diverse forgery approaches including *Face Swapping:* FaceSwap (Kowalski 2016), SimSwap (Chen

[*]These authors contributed equally.
[†]Corresponding author.

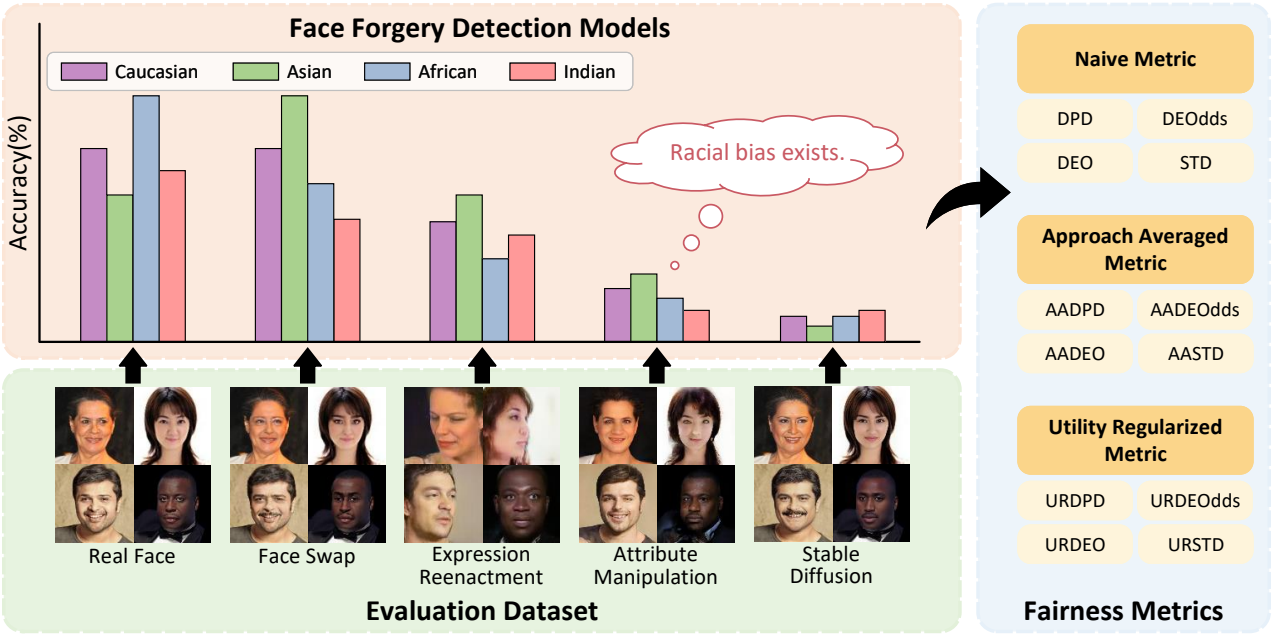

Figure 1: Workflow of fairness evaluation in forgery detection. We first construct an evaluation dataset containing a large number of subjects, diverse forgery approaches, and racial balance. Subsequently, we obtain the test results of the forgery detector on each race and forgery method. Finally, we comprehensively evaluate the detector using three sets of 12 fairness metrics in total.

et al. 2020), *Expression Reenactment:* FastReen (Zakharov et al. 2020), DualReen (Hsu, Tsai, and Wu 2022), *Face Editing:* MaskGAN (Lee et al. 2020), StarGAN (Choi et al. 2018), StyleGAN (Karras, Laine, and Aila 2019), *Diffusion-Based:* SDSwap (Xen 2023), DCFace (Kim et al. 2023), Face2Diffusion (Shiohara and Yamasaki 2024) and *Transformer-Based:* FSRT (Rochow, Schwarz, and Behnke 2024). And the self-constructed FairFD dataset does not have any undefined ethnicity annotations. For the specific metric, we address the mentioned two issues by introducing the Approach Averaged Metric, which calculates fairness separately for each forgery approach and then aggregates them, and the Utility Regularized Metric, which uses the utility to regularize the fairness.

We propose BPFA (Bias Pruning with Fair Activations), a novel pruning approach that uses an innovative metric to identify and prune weights with minimal impact on utility but significant contributions to bias (e.g., racial bias). BPFA enhances fairness without compromising utility and requires no retraining, making it a plug-and-play post-processing method applicable to any detector, including those optimized with existing fairness strategies, to further improve fairness performance. The workflow of fairness evaluation is illustrated in Figure 1. Sufficient experimental results prove that the proposed BPFA is an efficient, plug-and-play and robust pruning scheme, outperforming other baseline pruning methods by a significant margin.

The key contributions are summarized as follows:

- To our knowledge, it is the early exploration to introduce a comprehensive racial bias evaluation benchmark for forgery detection, providing a large-scale dataset, fairness metrics, and evaluation protocols. We newly introduce the Fair Forgery Detection (FairFD) dataset for racial bias in forgery detection evaluation, which contains the largest scale subjects, race-balanced ratio and incorporates diverse forgery approaches.
- We identify the bias offset and aggregation distortion problems with naive fairness metrics. Following, we design two metrics to address the mentioned issues. Extensive experimental results demonstrate the limited fairness of existing SOTA methods and validate the value of our proposed metric.
- We introduce Bias Pruning with Fair Activations (BPFA) to enhance forgery detector fairness. Extensive experiments highlight BPFA's efficiency, plug-and-play capability, and robustness. Integrating BPFA with the fairest detector achieves SOTA racial fairness performance. Our in-depth analyses provide valuable insights to advance the field.

Related Work

Fairness in Face Forgery Detection

Fairness Algorithm. Fairness in face forgery detection is a relatively novel topic. DAG-FDD (Ju et al. 2024) is first proposed to address fairness without demographic information by setting a probability threshold for minority groups to ensure low error rates for all groups meeting this threshold. DAW-FDD (Ju et al. 2024) utilizes demographic information to design losses to ensure similar performance across specified groups. PFGDFD (Lin et al. 2024) improves fairness by using disentanglement loss to separate demographic

Dataset	Race Rate					Race Balance	Undefined Ethnicity	Subject Number	App-roach	Real Img Number	Fake Img Number
	Caucasian	Asian	Indian	African	Others						
FF++ (2019)	∼43.9%	∼16.8%	∼3.2%	∼3.8%	∼32.3%	✘	*Yes*	∼1000	1	73k	266k
UADFV (2019)	97.96%	2.04%	0	0	0	✘	*Yes*	49	1	241	252
CelebDF-v2 (2020b)	88.10%	5.10%	0	6.80%	0	✘	*Yes*	59	1	225k	2,116k
DFDC (2020)	-	-	-	-	-	✘	*Yes*	960	8	488k	1,783k
DF-1.0 (2020)	∼25%	∼25%	∼25%	∼25%	0	✓	*Yes*	100	1	total 17,600k	
ForgeryNet (2021)	-	-	-	-	-	✘	*Yes*	5400	15	1438k	1457k
FairFD(ours)	∼25%	∼25%	∼25%	∼25%	0	✓	*No*	11430	11	52k	572k

Table 1: Face Forgery Detection Dataset Comparison. Our dataset is race-balanced, with no undefined races, the maximized number of subjects and forgery approaches exhibit diversity.

and forgery features.

Deepfake Dataset. We summarize the information of existing datasets in Table 1. We provide the proportions of each race, along with whether the datasets are race-balanced. Additionally, we present whether undefined ethnicity faces are included(,i.e., faces from one race are replaced with another race. We also supply the number of subjects, the number of forgery approaches, and the total number of frames. DAG(W)-FDD (Ju et al. 2024) and PFGDFD (Lin et al. 2024) directly use several of the datasets mentioned above as test data. Our work reveals inherent limitations when using these datasets for evaluation. There is currently no suitable dataset to evaluate the fairness of forgery detection. We also give details of widely used face forgery datasets in the section "Face Forgery Detection Datasets" in *Supp*.

Fairness Metric. To evaluate racial fairness, what we require is group fairness metrics. There are various group fairness metrics, and the selection of a metric depends on the application context. In this work, we consider the commonly used metrics including DPD (Agarwal et al. 2018; Agarwal, Dudík, and Wu 2019), DEOdds (Agarwal et al. 2018), DEO (Hardt, Price, and Srebro 2016), STD (Wang et al. 2019; Robinson et al. 2020; Gong, Liu, and Jain 2020; Yu et al. 2020; Wang et al. 2023) and our proposed novel fairness metrics. The definitions of these metrics can be found in the section "Existing Fairness Metrics" in *Supp*.

FairFD Dataset

Limitations of Current Datasets

Limited Number of Subjects. The construction process of the existing face forgery detection datasets involves collecting videos, subsequently creating forgeries, and then extracting frames. As a result, there are typically a small number of subjects and each subject has a large number of frames in these datasets. The limited number of subjects makes it challenging to draw meaningful comparisons across groups. Furthermore, we conducted the verification experiment to analyze and prove it in the section "Number of Subjects is Limited" in *Supp*.

Lack of Diversity of Forgery Approaches. In Table 1, only DFDC and ForgeryNet employ a variety of forgery techniques. However, we find that different forgery methods have different fairness levels. We validate this point in the subsequent Figure 4. Thus, we should strive to diversify forgery methods as much as possible, enabling a more comprehensive evaluation of the system's fairness.

Undefined Attribute Annotation. For these identity-replaced forgery approaches, there is a possibility of faces from one ethnicity being replaced with those from another. In related work (Xu et al. 2022), this phenomenon is referred to as "undefined attribute annotation." Undefined attributes can also significantly lower the quality of the evaluation of racial fairness, which is ignored in existing face forgery detection datasets.

FairFD Description

Considering the mentioned limitations in existing datasets, we introduce our dataset, FairFD, aiming to address these shortcomings. FairFD endeavors to overcome previous challenges and provide a more accurate, reliable and comprehensive benchmark for evaluating fairness in face forgery detection. Representative examples of ours are presented in Figure 1. The overview of FairFD can be shown in the Table 1. Subsequently, we delve into several pivotal facets of our dataset.

Our dataset is an image-level dataset, and for each image, there are 11 kinds of corresponding forgery images, i.e., *Face Swapping:* FaceSwap (2016), SimSwap (2020), *Expression Reenactment:* FastReen (2020), DualReen (2022), *Face Editing:* StarGAN (2018), StyleGAN (2019), MaskGAN (2020), *Diffusion-Based:* SDSwap (2023), DCFace (2023), Face2Diffusion (2024) and *Transformer-Based:* FSRT (2024). In addition to the forgery approach label, our approach also includes labels for four ethnicities (i.e., Caucasian, Asian, African, and Indian). Each ethnicity contains approximately 3000 subjects.

Source Data Collection and Forgery Process

To align with our requirements, which include having racial labels, and containing a sufficient number of subjects, we use the RFW (Wang et al. 2019) dataset as pristine images. The RFW dataset comprises face images with four racial labels (,i.e., Caucasian, Asian, African, and Indian), containing approximately 3000 subjects for each racial group, with a roughly equal distribution. Each subject has approximately $3 \sim 7$ images. All images in the RFW dataset have a resolution of 400×400 pixels. Besides, the images are carefully selected to maintain similar distributions in terms of

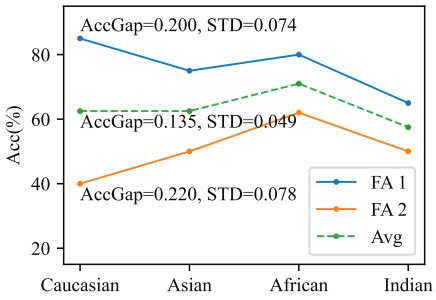

Figure 2: A face forgery detection model exhibits different biases for different forgery approaches.

age, gender, yaw angle, and pitch angle. See details in "Detailed Distribution of FairFD" in *Supp*. To reduce the human resources, we use RFW as the source data.

To achieve the goal of diversity, we choose various approaches and techniques. We classify the forgery methods into face swap, expression reenactment, attribute manipulation and advanced forgery methods(stable diffusion and transformer). See details in the section "Classification of Forgery Approaches" in *Supp*. We reimplement these methods and apply them to the source data. Details configuration and process of forgery crafting can be found in the section "Forgery Crafting Process" in *Supp*.

The Proposed Evaluation Metrics

Even though we have obtained a reliable dataset for evaluating deepfake detection's fairness, we still can not get a credible evaluation result due to existing fairness metrics having two flaws. Firstly, the *bias offset* may lead to an underestimation of racial bias. Secondly, *aggregation distortion* can result in biased evaluation outcomes favoring specific forgery approaches. Below, we introduce the two flaws and their corresponding solutions respectively.

We make corrections to four widely used metrics: DPD (Agarwal et al. 2018), DEOdds (Agarwal et al. 2018), DEO (Hardt, Price, and Srebro 2016), and STD (Wang et al. 2019). For clarity, we leverage DPD to introduce our new metric in the following discussions as an example. The following is the definition of DPD:

$$DPD = \max_{s,s' \in \mathbb{S}, s \neq s'} \left| P\left(\hat{Y} \mid S = s\right) - P\left(\hat{Y} \mid S = s'\right) \right|, \tag{1}$$

where $\hat{Y}$ is the predicted labels. $\mathbb{S}$ represents the set of sensitive attributes, $s \in \mathbb{S}$ and $\mathbb{S} = \{\text{Caucasian}, \text{Asian}, \text{Indian}, \text{African}\}$.

Bias Offset Problem

Bias offset refers to bias that will be partially obscured due to the calculation process of existing fairness metrics. Existing fairness metrics do not calculate separately for each forgery method instead of the final averaged performance scores. However, this way may obscure certain biases, which we call bias offset. Taking an example, in Figure 2, face forgery detectors may exhibit different biases for various forgery approaches. We calculate AccGap (the maximum differences in accuracy) and STD (standard deviation) for each forgery approach. In this example, both Forgery Approach 1 (FA1) and Forgery Approach 2 (FA2) exhibit an AccGap greater than 0.2 and an STD greater than 0.07. However, for Forgery Approach 1(FA1), the performance of Caucasians is better than Asians, while for FA2, the performance of Asians is better than Caucasians. In this situation, when calculating the fairness score using the final averaged performance scores, bias is to some extent offset, resulting in a smaller bias score. AccGap is less than 0.2, and the STD is less than 0.07 calculated using average accuracy. We refer to this phenomenon as *bias offset*. A more reliable way is to calculate fairness metrics separately for each forgery method and average them. We call this novel strategy as *Approach Averaged Metric*.

$$\text{AADPD} = \frac{1}{|\mathbb{F}|} \sum_{f \in \mathbb{F}} \max_{s,s' \in \mathbb{S}, s \neq s'} \left| P\left(\hat{Y} \mid S = s, F = f\right) - P\left(\hat{Y} \mid S = s', F = f\right) \right|, \tag{2}$$

where $\mathbb{F}$ donates real face and forgery approaches. $f \in \mathbb{F}$ and $\mathbb{F}$ is the set of forgery methods.

Aggregation Distortion Problem

Various forgery approaches not only exhibit different fairness situations but also demonstrate distinct levels of performance. For example, in Figure. 3 and Figure. 4, this is clearly evident. We identify the aggregation distortion problem where even if we calculate fairness scores separately for each forgery method and average them together, the averaged result can achieve a distorted fairness score due to the performance difference.

Directly averaging fairness scores when employing common fairness metrics might lead to misleading conclusions. For instance, consider two approaches with the accuracy of 20% and 80% respectively. We assume that both approaches yield a bias of 10% if we employ DEO as the fairness metric. Then, we calculate a simple average, which is also 10%. This would lead us to focus solely on the absolute differences in error rates without taking into account the variations in baselines. If the racial biases are both calculated to be 10%, the forgery method with only a 20% accuracy would evidently be much more unfair. This oversimplified average fails to capture the substantial disparity in performance between the two methods. To address this issue, we propose a fixed version. For each forgery approach, we have:

$$\text{URDPD} = \frac{1}{|\mathbb{F}|} \sum_{f \in \mathbb{F}} \max_{s,s' \in \mathbb{S}, s \neq s'} \left| P\left(\hat{Y} \mid S = s, F = f\right) - P\left(\hat{Y} \mid S = s', F = f\right) \right| \; / \; ACC_{F=f}, \tag{3}$$

where $ACC_{F=f}$ calculates the accuracy of given different forgery approaches.

After applying Eq. 3 to each forgery method, we calculate their results and then obtain the final fairness score by averaging them. We refer to this approach as *Utility Regularized Metric*. This nuanced method acknowledges the significance of each forgery approach, providing a more accurate and insightful evaluation of fairness in the context of the diverse fairness and performance landscape. In summary, we recommend not relying on a single fairness metric but rather considering a combination of multiple metrics to collectively reflect the fairness of a detector.

Bias Pruning with Fair Activations

In this section, we present the Bias Pruning with Fair Activations (BPFA) approach, which develops a novel pruning metric combining weights and the fairness of activations to determine weight importance. Then, we prune those with the lowest pruning scores based on a predefined pruning rate by layer. Here we utilize an unstructured pruning strategy. Noting that the proposed BPFA can be directly extended to process other biases except for racial bias.

Pruning Metric. Consider a convolutional layer weights W of shape $(C_{out}, C_{in}, S^h_{Ker}, S^w_{Ker})$, where C_{out} represents the number of output filters, and each filter has dimension $(C_{in}, S^h_{Ker}, S^w_{Ker})$. For one data sample, the output of this layer is denoted as X with shape $(C_{out}, S^h_{out}, S^w_{out})$. The L2 norm by filter of X is represented as $\|X\|_2 \in \mathbb{R}^{C_{out}}$. We compute the average L2 norm across all samples from a specific race $s \in \mathbb{S}$, denoted by $Z^s = \|X\|^s_2$. For each filter, we then calculate the standard deviation of these norms across all races, which serves as the bias for that filter:

$$BIAS_i = std\left(\{Z^s_i\}_{s\in\mathbb{S}}\right), \tag{4}$$

where $std(\cdot)$ denotes the standard deviation. This bias measures the variability of outputs across different races. The pruning score (PS) for each weight W_{ijkm} in the convolutional layer at the position (i, j, k, m) is then calculated by combining the weight with respect to the computed bias:

$$PS_{ijkm} = \frac{|W_{ijkm}|}{BIAS_i}. \tag{5}$$

By comparing the pruning scores, we can identify and potentially remove weights that have the least impact on model utility but the greatest impact on bias. This allows us to reduce bias and improve fairness without sacrificing performance. Note that while our method is illustrated using convolutional layers as an example, it can be easily extended to linear layers.

Benchmark Experiments

Experimental Setup

Dataset. We use FF++ (c23) as our training set. Specifically, for each video, we select 32 frames, crop the facial region, and finally resize it to 256 × 256. We utilize the preprocessed data provided by (Yan et al. 2023b), which has already undergone the aforementioned operations. Our proposed new dataset serves as the testing set. As our dataset inherently consists of face images with backgrounds and bodies removed, there is no need for additional face cropping. Subsequently, we resize the images to 256 × 256 for inference. Note that we still provide the original dataset with a resolution of 400 × 400 for scenarios requiring higher resolution.

Algorithms. We summarize the face forgery detection algorithms in the section "Face Forgery Detection Algorithms Categories" in *Supp*. For a comprehensive and fair analysis, we select several representative algorithms. For spatial-based detectors, we choose Xception (Rossler et al. 2019), RECCE (Cao et al. 2022), UCF (Yan et al. 2023a), Capsule (Nguyen, Yamagishi, and Echizen 2019), FFD (Dang et al. 2020) and CORE (Ni et al. 2022). For frequency-based detectors, we select F3Net (Qian et al. 2020), SPSL (Liu et al. 2021) and SRM (Luo et al. 2021). For fairness-enhanced detectors, we select DAG (Ju et al. 2024)(Xception as base model), DAW (Ju et al. 2024)(Xception as base model) and PFGDFD (Lin et al. 2024)(UCF as base model). In detail, these models are trained with the Adam optimization algorithm with a learning rate of 0.0002 and an epoch number of 10. The batch size is 32. And data augmentation methods including image compression, horizontal flip and rotation are applied. However, when applying these data augmentation methods to DAG, we find that its fairness level significantly deteriorated. For a fair comparison, we report below the results using data augmentation. Meanwhile, the results without data augmentation are presented in the section "Results without Data Augmentation" in *Supp*.

Benchmarking Fairness of Face Forgery Detectors

Benchmark Results. The benchmark results (shown in Table 2) present a comprehensive evaluation of 12 face forgery detectors using various fairness metrics. We highlight the four fairest detectors using different colors. Based on the results, we draw the following significant observations: *(1) Current face forgery detectors all exhibit a high degree of racial bias.* The DPD metric of SPSL achieves 0.0203, showing the smallest racial bias, with a 2.03% difference in fake classification probability between the most advantaged and disadvantaged groups. The AADPD metric averages 5.56% across forgery methods. In contrast, the least fair detector, UCF, shows a 17.65% difference, highlighting the need to address racial bias in face forgery detection models. *(2) Current face forgery detectors have racial bias variation.* Comparing the least fair detector UCF with the most fair detector SPSL, the former's URDPD is 4.76 times that of the latter, URDEOdds is 3.58 times, URDEO 4.98 times, and URSTD is 4.48 times, showing significant gap in racial bias. Other detectors also exhibit varying degrees of racial bias. Furthermore, we observe that three frequency-based detectors, SPSL, F3Net, and SRM, consistently demonstrate a smaller racial bias across all fairness metrics. We conduct an in-depth investigation into this in the section "Analyses and Discussions" in *Supp*.

Detailed Utility Results. To present more detailed results, we present the AUC for each detector, each race, and each forgery method in Figure 3. Results show that different forgery methods exhibit varying levels of utility. This validates the advantage of Utility Regularized Metric.

Fairness Metric		Spatial-based						Frequency-based			Fairness-enhanced		
		Xception	RECCE	UCF	Capsule	FFD	CORE	F3Net	SPSL	SRM	DAG	DAW	PFGDFD
Naive Metric	DPD↓	0.1810	0.1338	0.1765	0.0969	0.1099	0.0951	**0.0674**	**0.0203**	0.0990	0.1723	***0.0513***	***0.0805***
	DEOdds↓	0.1666	0.1264	0.1495	0.0902	0.1005	0.0798	***0.0763***	**0.0304**	**0.0714**	0.2288	***0.0593***	0.1396
	DEO↓	0.2088	0.1548	0.2014	0.1118	0.1242	0.1084	**0.0801**	**0.0215**	0.1090	0.2105	***0.0611***	***0.1032***
	STD↓	0.0647	0.0474	0.0631	0.0343	0.0398	0.0342	**0.0265**	**0.0080**	0.0355	0.0636	***0.0195***	***0.0328***
Approach Averaged Metric	AADPD↓	0.2024	0.1572	0.2175	0.1323	0.1552	**0.1147**	***0.1158***	**0.0556**	0.1413	0.2201	***0.0735***	0.1393
	AADEOdds↓	0.1669	0.1302	0.1630	0.1034	0.1196	**0.0858**	0.0961	**0.0481**	***0.0925***	0.2324	***0.0662***	0.1560
	AADEO↓	0.2095	0.1626	0.2284	0.1381	0.1623	***0.1205***	**0.1197**	**0.0571**	0.1511	0.2177	***0.0749***	0.1360
	AASTD↓	0.0750	0.0578	0.0809	0.0493	0.0576	***0.0449***	**0.0448**	**0.0219**	0.0531	0.0834	***0.0283***	0.0530
Utility Regularized Metric	URDPD↓	0.1357	0.1118	0.1523	0.0808	0.1037	**0.0803**	***0.0806***	**0.0320**	0.0904	0.1474	***0.0555***	0.0881
	URDEOdds↓	0.1057	0.0852	0.1069	0.0639	0.0763	**0.0567**	0.0625	**0.0299**	***0.0584***	0.1445	***0.0440***	0.0986
	URDEO↓	0.1417	0.1171	0.1614	**0.0842**	0.1092	***0.0850***	**0.0842**	**0.0324**	0.0968	0.1480	***0.0578***	0.0860
	URSTD↓	0.0501	0.0410	0.0565	**0.0301**	0.0384	0.0313	***0.0312***	**0.0126**	0.0339	0.0559	***0.0214***	0.0335
Utility	AUC↑	***0.6911***	0.6897	***0.7214***	0.6815	**0.7304**	0.6864	0.6564	0.6763	**0.7102**	0.6672	0.6604	0.6302

Table 2: Bias evaluation on FairFD for 12 face forgery detectors using Naive Metrics, Approach Averaged Metrics, Utility Regularized Metrics. For each row, the best values are **underlined and bolded**, followed by the second-best values which are ***underlined, bolded, and italicized***, the third-best values are **bolded**, and the fourth-best values are ***bolded and italicized***.

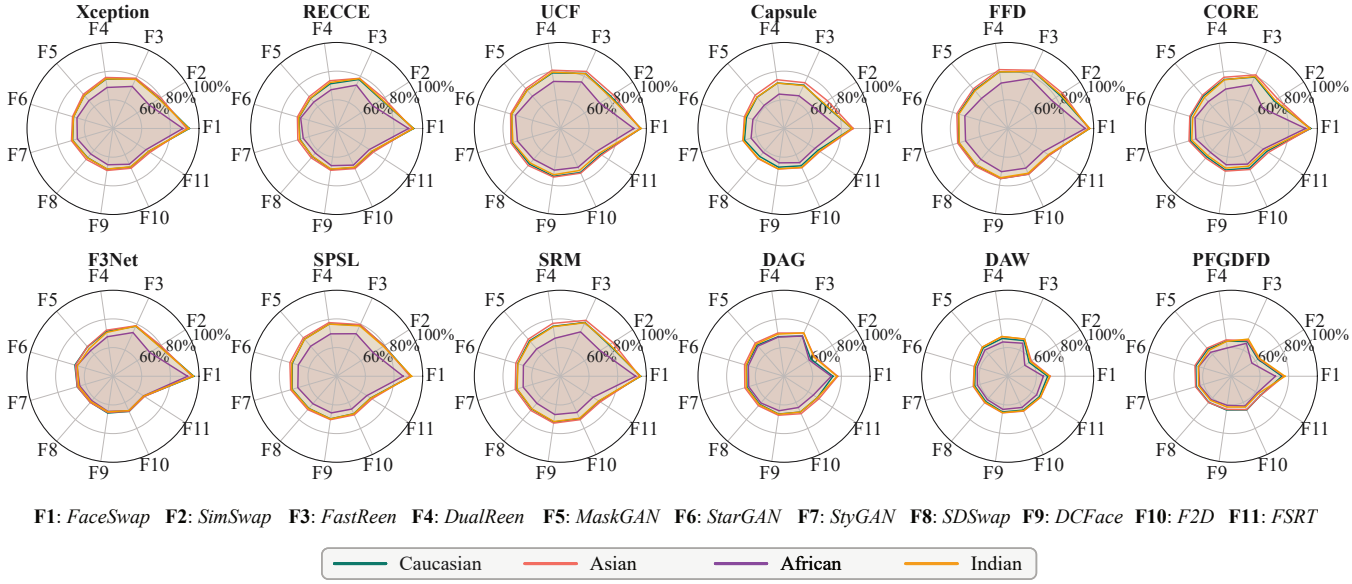

Figure 3: Detailed utility (AUC) for diverse races, forgery approaches, detectors.

Detailed Fairness Results. We present the standard deviation (STD) of the ACC for the four races in Figure 4 for each forgery method(including Real Face). Our findings reveal that different forgery methods exhibit varying levels of fairness, and different detectors rank the fairness of these forgery methods differently. This validates the advantage of the proposed Approach Averaged Metric.

Evaluating BPFA

Baseline Algorithm We select two baseline methods for comparison. The first WEIG uses only the absolute values of the weights as the pruning score, and the second RoBA uses only the reciprocal of the bias of the activations. More details about these baselines are shown in *Supp*. For all three methods, we prune the parameters with the lowest pruning scores. These pruning baselines can be directly applied in existing SOTA forgery detection models.

Results Analysis The experimental results under optimal pruning rates for each detector and method are shown in Table 3. It can be found that the proposed method BPFA consistently outperforms all baseline methods, achieving superior fairness without compromising utility. Although WEIG generally preserves good utility and enhances fairness, its improvements in fairness are not as significant as those achieved by BPFA. On the other hand, RoBA exhibits highly unstable performance. It results in improving fairness but at

Method		Naive Metric↓				Approach Averaged Metric↓				Utility Regularized Metric↓				Utility↑	
		DPD	DEOdds	DEO	STD	MA DPD	MA DEOdds	MA DEO	MA STD	UR DPD	UR DEOdds	UR DEO	UR STD	AUC	ACC
SPSL	Original	0.0203	0.0304	0.0215	0.0080	0.0556	0.0481	0.0571	0.0219	0.0320	0.0299	0.0324	0.0126	0.6763	0.7618
	WEIG	0.0183	0.0258	0.0201	**0.0072**	0.0564	0.0451	0.0586	0.0219	0.0324	0.0277	0.0334	0.0126	0.6769	0.7615
	RoBA	0.1128	0.1598	0.1395	0.0445	0.1462	0.1616	0.1432	0.0583	0.0893	0.1024	0.0867	0.0356	0.6331	0.7037
	BPFA	**0.0181**	**0.0209**	**0.0200**	**0.0072**	**0.0473**	**0.0357**	**0.0496**	**0.0182**	**0.0265**	**0.0218**	**0.0275**	**0.0102**	**0.6862**	**0.8055**
FFD	Original	0.1099	0.1005	0.1242	0.0398	0.1552	0.1196	0.1623	0.0576	0.1037	0.0763	0.1092	0.0384	0.7304	0.5751
	WEIG	0.1098	0.1003	0.1240	0.0398	0.1550	0.1194	0.1621	0.0576	0.1035	0.0761	0.1090	0.0384	0.7304	0.5751
	RoBA	-	-	-	-	-	-	-	-	-	-	-	-	0.5967	-
	BPFA	**0.1096**	**0.0999**	**0.1237**	**0.0397**	**0.1546**	**0.1189**	**0.1617**	**0.0574**	**0.1032**	**0.0758**	**0.1087**	**0.0382**	**0.7305**	**0.5760**
PFG-DFD	Original	0.0805	0.1396	0.1032	0.0328	0.1393	0.1560	0.1360	0.0530	0.0881	0.0986	0.0860	0.0335	0.6302	0.6019
	WEIG	0.0789	0.1340	0.1012	0.0319	0.1349	0.1494	0.1320	0.0513	0.0853	**0.0944**	0.0835	0.0324	0.6298	0.6021
	RoBA	-	-	-	-	-	-	-	-	-	-	-	-	0.5468	-
	BPFA	**0.0594**	**0.1337**	**0.0796**	**0.0238**	**0.1079**	**0.1442**	**0.1006**	**0.0411**	**0.0644**	0.0969	**0.0578**	**0.0245**	**0.6445**	**0.7415**

Table 3: Experiments with different fairness pruning methods. We highlight the best method for each metric in **bold**. And we use '-' to indicate methods that cause severe performance degradation, rendering the detector unusable even setting a pruning rate as low as 0.1%.

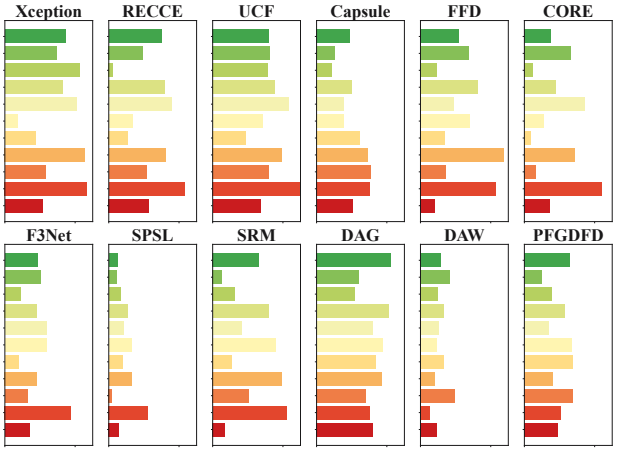

Figure 4: Fairness (STD) for different forgery approaches. Forgery approaches, listed from top to bottom, are FaceSwap, SimSwap, FastReen, DualReen, MaskGAN, StarGAN, StyGAN, SDSwap, DCFace, F2D, and FSRT.

the cost of significantly reduced utility, which can render the model nearly unusable in the forgery detection task. These results prove the superior performance of BPFA in enhancing both utility and fairness for forgery detection. *It is encouraging to find that SPSL+BPFA achieves the new state-of-the-art performance*. The only parameter in our method is the pruning rate. The ablation study for pruning rate is detailed in three tables in the section "Ablation Study on Pruning Rate" in *Supp*. Our findings indicate that different detectors and different methods have significantly different optimal pruning rates (which correspond to the least decrease in utility (or even improvement) while achieving the best average value across the 12 fairness metrics.

Analyses and Discussions

Here we conduct deeper analyses and give some insights: (1) We train detectors using balanced training data, finding that while racial bias can be reduced, the cost of collecting balanced data is substantial; (2) We set the optimal classification threshold for each race and then use the resulting accuracy values to calculate fairness. The conclusions show that this method is highly cost-effective and significantly improves both utility and fairness simultaneously; (3) We analyze that frequency-based detectors exhibit superior fairness performance because of not utilizing race-sensitive information, e.g. color. More analyses and details are shown in *Supp*.

Conclusion

This paper early explores a comprehensive racial bias evaluation benchmark for forgery detection, which provides a newly self-construct dataset, fairness metric and unified protocols. We identify numerous disadvantages in existing datasets and fairness metrics, then propose a novel dataset FairFD dataset and two sets of fairness metrics to address these mentioned issues. Besides, we also propose a novel Bias Pruning with Fair Activations algorithm to improve the fairness performance without an extra training process. Emphatically, we evaluate the fairness of multiple existing face forgery detectors. The results indicate the racial bias in current detectors is generally high and prove the advantages of our proposed BPFA. Further analyses reveal some interesting insights into the emergence of racial bias. We hope the proposed benchmark can inspire more researchers to develop the field. In the future, we will explore a unified fairness metric for diverse biases in more kinds of datasets, and construct the video-level forgery detection datasets for more real applications. See social impact in *Supp*.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62306227, Grant 62276198, Grant U22A2035, Grants U22A2096, Grant 62441601 and Grant 62036007; in part by the Fundamental Research Funds for the Central Universities under Grant ZYTS24142, Grant QTZX23083 and Grant QTZX23042; in part by the Key Research and Development Program of Shaanxi (Program No. 2023-YBGY-231); in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001; in part by the Guangxi Natural Science Foundation Program under Grant 2021GXNSFDA075011; in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11; in part by Open Research Project of Key Laboratory of Artificial Intelligence Ministry of Education under Grant AI202401, in part by the Nanning Scientific Research and Technological Development Project 20231042; in part by the '111 Center' (B16037).

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, 1–7. IEEE.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Agarwal, A.; Dudík, M.; and Wu, Z. S. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 120–129. PMLR.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4122.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2003–2011.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 5781–5790.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Gong, S.; Liu, X.; and Jain, A. K. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 330–347. Springer.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

He, Y.; Gan, B.; Chen, S.; Zhou, Y.; Yin, G.; Song, L.; Sheng, L.; Shao, J.; and Liu, Z. 2021. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4360–4369.

Hsu, G.-S.; Tsai, C.-H.; and Wu, H.-Y. 2022. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 642–650.

Jiang, L.; Li, R.; Wu, W.; Qian, C.; and Loy, C. C. 2020. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2889–2898.

Ju, Y.; Hu, S.; Jia, S.; Chen, G. H.; and Lyu, S. 2024. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4655–4665.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kim, M.; Liu, F.; Jain, A.; and Liu, X. 2023. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 12715–12725.

Kowalski, M. 2016. Faceswap. https://github.com/MarekKowalski/FaceSwap.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5549–5558.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020a. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5001–5010.

Li, Y.; and Lyu, S. 1811. Exposing deepfake videos by detecting face warping artifacts. arXiv 2018. *arXiv preprint arXiv:1811.00656*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3207–3216.

Lin, L.; He, X.; Ju, Y.; Wang, X.; Ding, F.; and Hu, S. 2024. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16815–16825.

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 772–781.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16317–16326.

Masood, M.; Nawaz, M.; Malik, K. M.; Javed, A.; Irtaza, A.; and Malik, H. 2023. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4): 3974–4026.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2307–2311. IEEE.

Ni, Y.; Meng, D.; Yu, C.; Quan, C.; Ren, D.; and Zhao, Y. 2022. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12–21.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, 86–103. Springer.

Robinson, J. P.; Livitz, G.; Henon, Y.; Qin, C.; Fu, Y.; and Timoner, S. 2020. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–1.

Rochow, A.; Schwarz, M.; and Behnke, S. 2024. FSRT: Facial Scene Representation Transformer for Face Reenactment from Factorized Appearance Head-pose and Facial Expression Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7716–7726.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Shen, X.; Du, C.; Pang, T.; Lin, M.; Wong, Y.; and Kankanhalli, M. 2024. Finetuning Text-to-Image Diffusion Models for Fairness. In *International Conference on Learning Representations (ICLR)*.

Shiohara, K.; and Yamasaki, T. 2024. Face2Diffusion for Fast and Editable Face Personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6850–6859.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tian, Y.; Shi, M.; Luo, Y.; Kouhana, A.; Elze, T.; and Wang, M. 2024. Harvard FairSeg: A Large-Scale Medical Image Segmentation Dataset for Fairness Learning Using Segment Anything Model with Fair Error-Bound Scaling. In *International Conference on Learning Representations (ICLR)*.

Wang, F.-E.; Wang, C.-Y.; Sun, M.; and Lai, S.-H. 2023. Mixfairface: Towards ultimate fairness via mixfair adapter in face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14531–14538.

Wang, M.; Deng, W.; Hu, J.; Tao, X.; and Huang, Y. 2019. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, 692–702.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.

Xen, T. 2023. DiffusionFaceswap. https://github.com/glucauze/sd-webui-faceswaplab.

Xu, Y.; Terhörst, P.; Raja, K.; and Pedersen, M. 2022. A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. *arXiv preprint arXiv:2208.05845*.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023a. UCF: Uncovering Common Features for Generalizable Deepfake Detection. *arXiv preprint arXiv:2304.13949*.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023b. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. *arXiv preprint arXiv:2307.01426*.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8261–8265. IEEE.

Yu, J.; Hao, X.; Xie, H.; and Yu, Y. 2020. Fair face recognition using data balancing, enhancement and fusion. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 492–505. Springer.

Zakharov, E.; Ivakhnenko, A.; Shysheya, A.; and Lempitsky, V. 2020. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 524–540. Springer.