Analyzing Fairness in Deepfake Detection With Massively Annotated Databases

Ying Xu, *Graduate Student Member, IEEE*, Philipp Terhörst, *Member, IEEE*, Marius Pedersen, *Member, IEEE*, and Kiran Raja, *Senior Member, IEEE*

***Abstract*—In recent years, image and video manipulations with Deepfake have become a severe concern for security and society. Many detection models and datasets have been proposed to detect Deepfake data reliably. However, there is an increased concern that these models and training databases might be biased and, thus, cause Deepfake detectors to fail. In this work, we investigate factors causing biased detection in public Deepfake datasets by (a) creating large-scale demographic and non-demographic attribute annotations with 47 different attributes for five popular Deepfake datasets and (b) comprehensively analysing attributes resulting in AI-bias of three state-of-the-art Deepfake detection backbone models on these datasets. The analysis shows how various attributes influence a large variety of distinctive attributes (from over 65M labels) on the detection performance which includes demographic (age, gender, ethnicity) and non-demographic (hair, skin, accessories, etc.) attributes. The results examined datasets show limited diversity and, more importantly, show that the utilised Deepfake detection backbone models are strongly affected by investigated attributes making them not fair across attributes. The Deepfake detection backbone methods trained on such imbalanced/biased datasets result in incorrect detection results leading to generalisability, fairness, and security issues. Our findings and annotated datasets will guide future research to evaluate and mitigate bias in Deepfake detection techniques. The annotated datasets and the corresponding code are publicly available. The code link is: https://github.com/xuyingzhongguo/DeepFakeAnnotations.**

***Index Terms*—Deepfake, deepfake detection, databases, bias, fairness, image manipulation, video manipulation.**

I. Introduction

DEEPFAKE refers to a deep learning-based technique that is able to create fake videos/images by swapping the face of a person with the face of another. An example of a Deepfake would be a video of someone convincingly speaking a language they have never learned, or even impersonating a public figure so seamlessly that it's nearly impossible to distinguish from reality. Deepfake has become a great concern for security and society [1] due to the harmful usage of such fake content, such as fake news, fake pornography, or financial fraud. Moreover, the availability of large-scale public face datasets and the development of strong generative artificial intelligence (AI), and especially deep learning techniques, such as Autoencoder or Generative Adversarial Networks (GAN) [2], [3] have strongly increased the realism of Deepfake. Various open-source and mobile applications [4], [5] further allow to create highly realistic Deepfake videos or images without any expert knowledge and thus, make it possible for everyone to automatically manipulate images of videos with Deepfake technology.

Consequently, many works have developed detection methods capable of detecting such face manipulations [6]. Previous studies, however, pointed out some bias issues with these detection methods for different factors such as age, gender, and ethnicity [7], [8], [9]. The main reasons for bias in such AI models are believed to originate from unbalanced training databases [7], [8], [9]. Biased decisions from detection approaches significantly impact both security and society if, for example, images from a certain group of people are constantly scrutinised as Deepfake.

A. Societal and Technological Aspects

In 2018, the disappearance of Gabon's President Ali Bongo led to public unrest and speculation of his assassination, culminating in a government-released DeepFake New Year's broadcast that inadvertently provoked a military mutiny due to its unnatural appearance [10]. Reliable detection across different demographic groups in such situations can prevent inadvertent consequences. Despite the availability of DeepFake detection algorithms, there is a growing concern that these algorithms as with other machine learning algorithms, misclassify authentic images from specific ethnic and demographic groups as DeepFakes.

Such misclassification for certain groups can have unintended and significant societal and political consequences [11], [12]. To address this issue, it is essential to thoroughly investigate and analyse the factors related to bias in DeepFake detection. With a comprehensive understanding of these factors, the deployment of such technology can be justified and known limitations can be disclosed. Therefore, our primary motivation is to thoroughly investigate and analyse the factors responsible for bias in DeepFake detection.

This paper specifically focuses on analysing these factors and does not propose or introduce any new DeepFake detection methods. Our objective is to highlight the key aspects to be

Manuscript received 3 April 2023; revised 20 September 2023; accepted 2 February 2024. Date of publication 16 February 2024; date of current version 16 May 2024. *(Ying Xu and Philipp Terhörst contributed equally to this work.) (Corresponding author: Ying Xu.)*

The authors are with the Institutt for datateknologi og informatikk, Norges teknisk-naturvitenskapelige universitet, 2802 Gjøvik, Norway (e-mail: ying.xu@ntnu.no).

This article has supplementary downloadable material available at https://doi.org/10.1109/TTS.2024.3365421, provided by the authors.

Digital Object Identifier 10.1109/TTS.2024.3365421

addressed before implementing a DeepFake detection algorithm in operational contexts. By doing so, we aim to provide valuable insights to inform the development and deployment of more fair and accurate DeepFake detection systems.

B. Contributions

The present study highlights the necessity for annotated datasets and balanced performance metrics to assess the impact of biased datasets to determine the efficacy of detection models. In this regard, this work makes two significant contributions by analyzing factors that lead to perceived bias in Deepfake detection.

1) We provide massive and diverse annotations for five widely-used Deepfake detection datasets. Existing Deepfake detection datasets contain none or only sparse annotations restricted to demographic attributes, as shown in Table II. This work provides over 65.3M labels using 47 different attributes for five popular Deepfake detection datasets (Celeb-DF [13], DeepFakeDetection (DFD) [14], FaceForensics++ (FF++) [15], DeeperForensics-1.0 (DF-1.0) [16] and Deepfake Detection Challenge Dataset (DFDC) [17]).
2) We comprehensively analyse detection bias in three state-of-the-art Deepfake detection backbone models with respect to various demographic and non-demographic attributes regarding to four current Deepfake datasets. Previous investigations restricted their analysis to a maximum of four demographic attributes on a single dataset. Contrarily, we analyse detection bias on a much larger scale of distinctive attributes on four widely-used Deepfake datasets.[1]

For the first contribution, five annotated datasets are created in the direction of earlier work using the MAAD-Face principle [18]. By computing a reliability score from the predictions of the MAAD classifier, we consider high-confidence predictions for labelling process to ensure a high annotation correctness. While the annotations from previous works at most contain demographic information like age, gender, and ethnicity, the annotations in this work are highly diverse and include attributes such as hair-color and -style, skin, face geometry, mouth, noise, and various accessories. We assert that these rich annotations will allow future works to evaluate the role of each attribute and use it to train better detection models that can mitigate bias issues.

The second contribution of our work is a detailed analysis of detection bias in Deepfake detection approaches by comparing the differential outcomes of three state-of-the-art Deepfake backbone networks (EfficientNetB0 [19], Xception [20], and Capsule-Forensics-v2 [21]) on four of the proposed Deepfake annotation datasets with respect to 31 demographic and non-demographic attributes.[2]

The results indicate that the investigated datasets are highly imbalanced leading to highly biased detection backbone models when trained on such databases for a large variety of demographic and non-demographic attributes.

The observed bias in the detection backbone models can further explain the low generalisability of current Deepfake detectors [22], [23] across different attributes. Interestingly, the effect of the imbalanced attributes often differs in detection performance if the attribute is observed on a pristine (nonfake) image or a Deepfake. The results indicate that the detection backbone models learn several questionable factors that require a deeper investigation. For example, a person smiling or wearing a hat is strongly detected as a real person despite being a Deepfake image. Depending on the application, these factors can lead to biases and, subsequently, strong fairness issues when a fake video of a smiling woman is detected as a real one. Conversely, a biased detection backbone model deciding a manipulated video as an unaltered video may lead to security implications. A complete list of such findings from our work along with the recommendations for future work is provided in Section VI.

II. Related Work

A. Deepfake Detection

There are two main approaches that are used to detect manipulated media. One focuses on the spatial features extracted from frames of a video. The other utilises temporal features among frames to capture falsified clues.

1) *Spatial features:* Most of the early efforts to detect Deepfake have been made using spatial features extracted from video frames. Researchers have been working on detecting artifacts using unnatural facial features [24], blending traces [25], CNN-generated/GAN-generated fingerprints [26], [27]. Some studies have also been conducted in the frequency domain in order to detect artificial image contents [28], [29].
2) *Temporal features:* Instead of individual frames, temporal features across frames have also been used recently, for example, unsynchronised color [30], [31], and phoney heartbeats appearing on faces [32], [33] and inconsistent facial information [34], [35], [36], [37].

Most works have focused on developing Deepfake detection solutions tailored to available datasets. However, these solutions can be imbalanced, leading to bias and low generalisability across different demographic factors. We analyse four Deepfake detection approaches to demonstrate the biased performances for different demographic factors.

B. Deepfake Datasets

Table II shows seven popular Deepfake datasets that are popularly used for the development and evaluation of reliable Deepfake detection backbone models. DeepfakeTIMIT [38] and FFW [39] were published in 2018, followed by DFD [14] and FF++ [15], [40] in 2019. DFDC [17], Celeb-DF [13] and DF-1.0 [16] were released in 2020. Over the years, the size of the datasets has increased in terms of manipulations and

[1] For the analysis, we do not consider DF-1.0 data as the detection methods did not produce enough errors (Details in Table XIV, Table XV, and Table XVI) on this dataset to analyse biased behaviours.

[2] For experiments, we neglected attributes that are not frequently occurring to avoid wrong conclusions caused by limited testing data. Details can be found in the Appendix and Table IV.

TABLE I
COMPARISON OF PREVIOUS BIAS INVESTIGATIONS IN DEEPFAKE DETECTION - THIS WORK PROVIDES A MORE COMPREHENSIVE BIAS ANALYSIS INVOLVING MORE DATASETS UP TO 4 AND MORE INVESTIGATED ATTRIBUTES REACHING THE QUANTITY OF 47

	Attributes	Models	Datasets
Hazirbas *et al.* [7]	4	2	1
Loc and Yan [8]	2	**3**	1
Pu *et al.* [9]	2	1	1
This work	**47**	**3**	**4**

the total number of images/videos. However, there are limited efforts to create more balanced datasets for gender and ethnicity. Both Celeb-DF and DF-1.0 maintain parity between males and females. Celeb-DF has a more extensive range of ages, while DF-1.0 holds balanced skin types. Despite these efforts, only a few databases provide additional annotations that could be utilised for developing Deepfake detection algorithms or testing these for influences of demographic factors. In contrast to previous works, we provide high-quality annotations for five popular databases for 47 demographic and non-demographic attributes. We hope to enable the development and evaluation of balanced and less-biased Deepfake detectors.

C. Analysing Bias in Deepfake Detection

Internal representations of neural network models preserve attribute-related information of the training data even if it is not directly needed for the model objective [42], [43]. These encoded attribute patterns are reported to lead to biased performance in AI models [44]. Although there are many works on studying fairness in AI [44], [45], [46], only a few works analyse biases in the Deepfake detection field. Hazirbas et al. [7] measured the robustness of Deepfake detection backbone models across four primary dimensions: age, gender, apparent skin type, and lighting. They analysed the top five winners of the Deepfake Detection Challenge [47], [48], [49], [50], [51] for these attributes and concluded that all methods are biased towards lighter skin tones and fail in subjects with darker skin. Trinh and Liu [8] measured the predictive performance of popular Deepfake detectors, MesoInception-4 [52], Xception [20] and Face X-Ray [25] on racially balanced datasets for gender and race. Significant disparities were found in predictive performances across races and large representation bias in widely used FF++ [15]. Pu et al. [9] used a subset of the Face2Face dataset in FF++ and investigated MesoInception-4 to verify the existence of gender bias. Studying bias in Deepfake detection so far is limited to a few demographic factors such as gender and ethnicity. In contrast, this work analyses bias of three state-of-the-art Deepfake detection methods on four widely-used Deepfake datasets considering 31 demographic and non-demographic attributes as shown in Table I. With this work, we provide up to 47 attribute annotations on 4 popular Deepfake datasets. This work makes it possible to study the bias problem in a more comprehensive and reliable manner.

III. METHODOLOGY

To analyse different attributes, we first create large-scale annotations of 47 demographic and non-demographic attributes for five Deepfake detection databases. Following this, we conduct a comprehensive bias analysis of the state-of-the-art Deepfake detection methods on these annotated databases. In the following section, the process for creating the large-scale annotations is described, and methodology for measuring bias is presented.

A. Annotating Deepfake Databases

We utilize MAAD-Face classifier [18] trained on LFW [53] and CelebA [54] as source databases to implement a novel annotation-transfer technique that transfers the attribute annotations from several source databases to target databases. This approach prioritizes annotations with high confidence predictions, thereby enhancing annotation correctness and minimizing potential biases. We annotate five current Deepfake detection databases (DFD [14], FF++ [15], DFDC [17], Celeb-DF [13], and DF-1.0 [16]) in this work.

In the annotation process, each image is assigned with one of three possible labels for an attribute, positive (1), negative (-1), or undefined (0). A positive annotation for attribute a of an image means that the face in the image has attribute a. For instance, a face with a positive annotation for 'Young' represents a face of an young individual. In contrast, a negative annotation for attribute a of an image means that the face in the image does not possess attribute a. e further enforce a confidence-driven threshold to assert if an attribute cannot be classified. The confidence score is calculated based on the reliability measure from [55] and aims at preventing error-prone annotations. Specifically, if the classifier produces a decision for an attribute with a confidence below 90%, we annotate the attribute as undefined (0). We apply this methodology on five Deepfake detection datasets (DFD, FF++, DFDC, Celeb-DF, DF-1.0), resulting in the annotation datasets A-DFD (4.7M labels), A-FF++ (8.5M labels), A-DFDC (4.6M labels), A-Celeb-DF (9.2M labels) and A-DF-1.0 (38.3M labels), shown in Table II. These provide annotations for 47 attributes including information on demographics, skin, hair, beard, face geometry, mouth, nose, and accessories.

B. Evaluating Bias in Imbalanced Data

In this study, we assess the bias of a detection backbone model to an attribute a by comparing its performance when the attribute is present versus absent. However, there is a potential issue of an unbalanced distribution of positive and negative labelled testing samples during the experiments. To avoid inaccurate results caused by this imbalance, we introduce a corrected performance measure using control groups of positive and negative samples. We adopt the methods of creating two control groups for each attribute a by randomly selecting N samples from the testing data from [44], where N is the number of samples with or without attribute a. By doing so, we ensure that each control group has the same number of samples as their counterparts in the real data, thus making the

TABLE II

OVERVIEW OF POPULAR DEEPFAKE DATASETS AND THE PROPOSED ANNOTATIONS DATABASES - WHILE PREVIOUS DATABASES LACK DIVERSE ANNOTATIONS, THE FIVE PROPOSED ANNOTATION DATABASES CLOSE THIS GAP AND PROVIDE THE RESOURCES NEEDED TO COMPREHENSIVELY ANALYSE AND MITIGATE BIAS IN DEEPFAKE DETECTION BACKBONE MODELS

			Number of videos		Number of frames		Annotated
	Dataset	Identities	Pristine	Forged	Pristine	Forged	Attributes
Previous works	DeepfakeTIMIT [38]	32	320	620	34.0k	68.0k	-
	FFW [39]	150	-	150	-	53k	-
	DeepFakeDetection (DFD) [14]	28	363	3,068	315.4k	2.2M	-
	FaceForensics++ (FF++) [15], [40]	1000	1,000	5,000	300k	1.5M	-
	Deepfake Detection Challenge Dataset (DFDC) [17]	960	23,654	104,500	7M	31M	-
	Celeb-DF [13]	59	590	5639	225.4k	2.1M	-
	DeeperForensics-1.0 (DF-1.0) [16]	100	50,000	10,000	2.9M	14.7M	1
	KoDF [41]	403	62,166	175,776	135M	65.9M	2
This work	**A-DFD**	28	363	3068	10.8k	89.6k	**47**
	A-FF++	1000	1000	5000	29.8k	149.1k	**47**
	A-DFDC	960	23,654	104,500	54.5k	52.6k	**47**
	A-Celeb-DF	59	590	5639	26.3k	166.5k	**47**
	A-DF-1.0	100	50000	10000	870.3k	321.5k	**47**

positive and negative control groups independent of individual sample properties and drawn from the same distribution.

Comparing the classification performance of the positive and negative control groups for an attribute a allows us to measure the effect of data imbalance on performance. If the performances of the negative and positive control groups are similar, the distribution of testing data does not significantly impact the performance. Contrarily, if the performances of the negative and positive control groups are dissimilar, the unbalanced testing data affects the classification performance. To measure the bias effect of an attribute a on the performance, we adopt the *relative performance (RP)* measure from [44]

$$RP_{type}(a) = 1 - \frac{err^{(+)}_{type}(a)}{err^{(-)}_{type}(a)}, \tag{1}$$

with $type = \{data, control\}$. $RP_{type}(a)$ measures the performance differences for an attribute a based on the error rates for a positive $err^{(+)}_{type}(a)$ and a negative $err^{(-)}_{type}(a)$ group. If the error rates are the same, $RP(a) = 0$ and, thus, attribute a does not affect performance. Positive RP values refer to lower error rates for the positive class (samples with this attribute). Contrarily, negative RP values refer to lower error rates for the negative class.

To correct this bias in the relative performance measure originating from the unbalanced testing data, we propose the *corrected relative performance* (CRP)

$$CRP(a) = RP_{data}(a) - RP_{control}(a) \tag{2}$$

which describes the difference between the relative performance of the real data RP_{data} and the relative performance of the control groups $RP_{control}$. The CRP measure simplifies to

$$CRP(a) = \frac{err^{(+)}_{control}(a)}{err^{(-)}_{control}(a)} - \frac{err^{(+)}_{data}(a)}{err^{(-)}_{data}(a)}, \tag{3}$$

and aims at removing the influence of the testing data distribution from the performance measure. If biased performance comes only from unbalanced test data, RP_{data} and $RP_{control}$ will be equal, and thus the corrected relative performance CRP will be zero. We use the $CRP(a)$ to measure the influence of the presence of attribute a on the performance and thus, to measure bias independently of the testing data parity.

IV. EXPERIMENTAL SETUP

A. Database and Considered Attributes

For the experiments, we choose five widely-used Deepfake detection datasets, DFD [14], FF++ [15], [40], DFDC [17], Celeb-DF [13] and DF-1.0 [16]. Details for the different databases are provided in Table II. 30 frames are extracted from the first 300 frames of each video using a 10-frame interval. The faces are detected and aligned using MTCNN [56] for each of the frames. To ensure that enough data is available for analysing bias originating from specific attributes, we ignore attributes where a minimum of 100 positive or negative labelled images are unavailable. Out of the 47 attributes available in the annotated databases, only 31 were included in the bias analysis due to such a curation process. The specific details of this process can be found in Appendix Table IV.

B. Deepfake Detection Backbone Models

For the experiments, we choose three well used Deepfake detection backbone models, EfficientNetB0 [19], Xception [20], and Capsule-Forensics-v2 [21]. These three networks have been used frequently as backbone networks in the Deepfake detection [25], [29], [57], [58], [59], [60]. Therefore, we consider it reasonable to use them for the bias analysis. Furthermore, we have trained and evaluated the three backbone networks with a subject-exclusive train/dev/test for all the attributes. Due to the lack of a standardised protocol for all datasets, we spilt the datasets with a 60%/20%/20% proportion for train/val/test respectively.

- *Xception* uses depth-wise separable convolutions to reduce the computational cost of traditional convolutions while maintaining high accuracy. This is achieved by performing spatial and channel convolutions separately, allowing for more efficient image feature processing. It is a highly effective deep learning architecture for image recognition tasks that require high accuracy and computational efficiency.

- *EfficientNet* is a model scaling method that uses a simple yet highly effective compound coefficient to scale up CNNs in a more structured manner, balancing the network's depth, width, and resolution to optimize its performance on a given resource budget. The architecture includes several novel features, including a mobile inverted bottleneck block, squeeze-and-excitation optimisation, and stochastic depth regularisation, further improving its performance. In our paper, we select the most lightweight version of EfficientNet, EfficientNetB0, to showcase its effectiveness.
- *Capsule-Forensics-v2* uses capsules to extract facial features and their spatial relationships from the input image to detect discrepancies. It incorporates a novel loss function that encourages disentangled representations, improving forgery detection accuracy. The model has demonstrated high effectiveness in detecting image manipulations, including copy-move, splicing, and face morphing.

C. Evaluation Metrics

Previous work on Deepfake detection has reported its results based on a simple accuracy measure [8], [9]. However, dealing with unbalanced testing data is the norm, and a simple accuracy measure is vulnerable to this. We further notice many attributes being unbalanced in terms of the positive/negative labels from Figure 1. We, therefore, make use of a balanced accuracy measure, which computes the arithmetic mean of the sensitivity and specificity and is more robust to unbalanced data [61]. More precisely, we report the performances in terms of error rates (1-balanced accuracy) since this work investigates bias issues driven by inaccurate predictions.

V. Results

This section presents our findings on the presence of bias in Deepfake detection datasets utilising our proposed annotations. We analysed the relationship among various variables regarding RP-vs-CRP and PDRP-vs-DDRP and used 24 plots to visualize these relationships. As we discuss these findings, it is crucial to keep in mind the concepts of causality and correlation in research and statistics. Causality refers to the relationship between cause and effect, while correlation measures the degree to which two or more variables are associated. It is important to note that correlation does not necessarily imply causality. Therefore, our results will focus on explaining correlation, and we will leave an in-depth exploration of causality for future studies.

A. Analysing Database Annotations

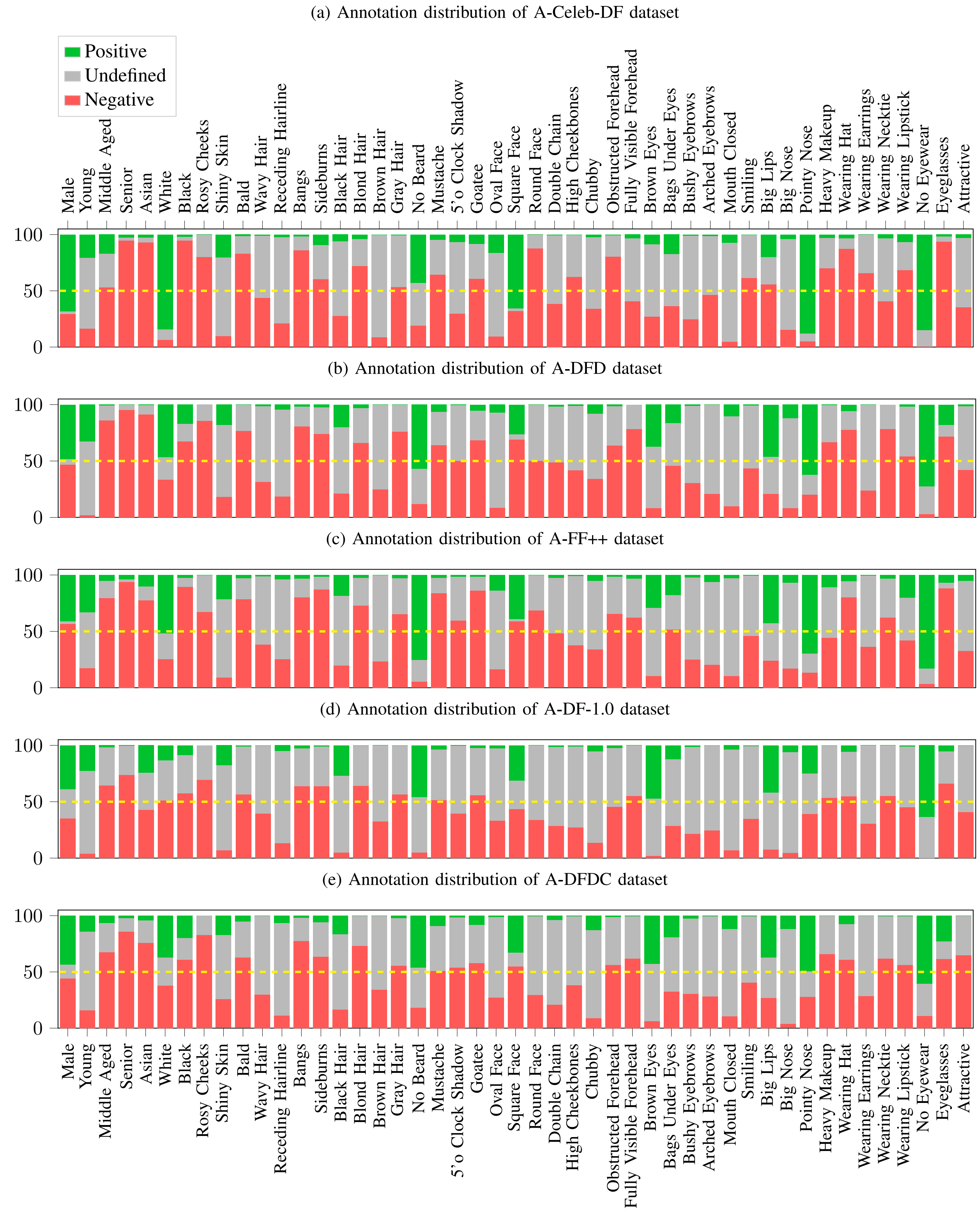
1) Attribute Statistics: Figure 1 shows the annotation distribution of the five annotated Deepfake detection datasets. For each attribute, green indicates the percentage of positive annotations, red indicates the percentage of negative annotations, and grey represents the percentage of images that have an undefined annotation for the attribute. According to the data given by Celeb-DF [13], this dataset contains male of 56.8% and female 43.2%, and in Figure 1(a), the percentage of male (positive) is 70.15% and the percentage of female (negative) is 29.85%. The reason of the increased gap between gender asymmetry is that Celeb-DF only generates Deepfake videos using the same gender, so the number of differences between male and female are enlarged among the synthesised videos. We notice that most databases are quite balanced for the gender attribute.

However, there is a big imbalance with respect to skin colour, especially in Celeb-DF where people with white skin tones occupy the vast majority of this dataset. The big gap between numbers corresponds to the number disparity of Celeb-DF [13] (5.1% Asian, 6.8% African, and 88.1% Caucasian) which clarifies the high accuracy of the MAAD-classifier. The DFDC dataset appears to have a noticeable under-representation of individuals of Asian descent. Furthermore, there is a prevalence of white individuals in both the DFD and FF++ datasets. The variations in skin color distribution across different datasets may result in biases in the Deepfake detection system.

To conclude, it is clearly visible that the investigated Deepfake detection databases (DFD, FF++, DF-1.0, and DFDC) are strongly imbalanced with respect to most analysed attributes. Future work should consider creating balanced datasets to prevent any potential biases in Deepfake detection algorithms when such datasets are used for training.

2) Attribute Correlations: We present 20 most positive and negative pairwise attribute correlations in Figure 2 to understand the quality of the labels and potential biases in the attribute space. For instance, we notice in Figure 2(a) that the attributes of Mustache and Goatee are highly correlating with each other. A high correlation also occurs between Heavy Makeup and Wearing Lipstick. This is easy to understand as the former attributes are mainly associated with males, and the latter ones mainly are with females. In contrast, Mustache and Goatee are negatively correlated to Heavy Makeup, and Wearing Lipsticks. Similar patterns are observable across different attribute correlations. The presence of negative correlations, such as the inverse relationship between *No Beard* and *Mustache*, as well as *Goatee*, highlights the quality of the annotations. It still should be noted that some correlations might also origin from the MAAD-classifier. Most of these correlations can be explained with background knowledge of the databases. For instance, the Celeb-DF dataset contains mainly images of celebrities in which these are presenting themselves to the camera. Therefore, a high correlation between Heavy Makeup and Wearing Lipstick is observed which may not necessarily represent real-world Deepfake of non-celebrities.

To conclude, our investigation has identified attribute pairs within the databases that exhibit strong correlations. It is imperative that future studies utilising these datasets and annotations consider these attribute correlations to avoid any misinterpretations that may result in biases. By acknowledging



(a) Annotation distribution of A-Celeb-DF dataset

(b) Annotation distribution of A-DFD dataset

(c) Annotation distribution of A-FF++ dataset

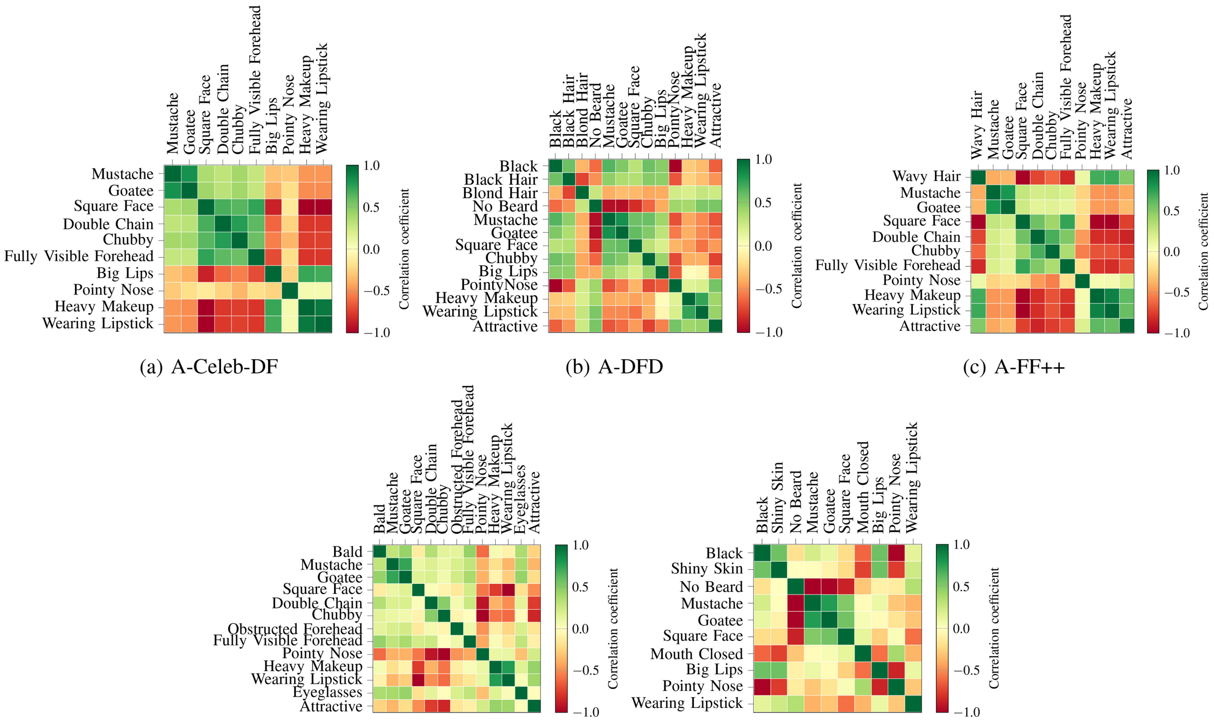
(d) Annotation distribution of A-DF-1.0 dataset

(e) Annotation distribution of A-DFDC dataset

Fig. 1. **Annotation distribution of the annotated Deepfake detection datasets** - The distributions of the proposed dataset annotations are shown with the y-axis presenting percentage. For each attribute, green indicates the percentage of positive annotations, red indicates the percentage of negatively annotations, and grey represents the percentage of images that have an undefined annotation for the attribute. The distributions show that these databases are highly unbalanced concerning these attributes.

and accounting for these correlations, we can enhance the accuracy and fairness of any analysis or application of these databases.

3) Annotation Correctness: To evaluate the effectiveness of the proposed annotations, we have adopted Table III from the MAAD-classifier [18]. This table verifies the accuracy of the



(a) A-Celeb-DF

(b) A-DFD

(c) A-FF++

(d) A-DF-1.0

(e) A-DFDC

Fig. 2. **Attribute annotation correlations of the Deepfake detection databases** - The 20 most positive and negative (Pearson) correlations are shown for each of the five databases. Green indicates positive correlations, while red indicates negative correlations. For working with these databases, the highly-correlating attributes should be considered to prevent misinterpretations.

MAAD-classifier for the attributes utilised in our study. This table originates from [18] and shows the attribute correctness of the classifier with respect to three human evaluators. For each attribute, 100 images with and 100 images without this attribute were chosen randomly and shown to the evaluators to determine the true attribute label for each image. If the evaluators disagreed on an attribute, majority voting was used to decide on a label. Then, the accuracy, precision, and recall of the classifier predictions are calculated based on the ground truth provided by the human evaluators. The results are shown in Table III. For most attributes, the classifier agrees with human evaluators, resulting in an average accuracy of 92%, precision of 90%, and recall of 94%. Compared to similar facial annotation databases, such as LFW [53] (72% accuracy, 61% precision, 84% recall) and CelebA [54] (85% accuracy, 83% precision, 89% recall) [18], the proposed annotations are of high correctness.

The annotations provided in this work are of higher quality than the annotations provided for previous databases and we assert them to be suitable for analysing bias in Deepfake detection. Future works can make use of these attributes for developing and analysing bias-mitigating approaches in Deepfake detection.

B. Analysing Bias in Deepfake Detection

To understand the bias in Deepfake detection, we will first study the general detection performance in presence of several potentially imbalanced attributes and secondly, analysing the detection performance in presence of these attributes separately on pristine and fake data. We exclude DF-1.0 dataset as the detection methods did not produce high enough errors necessary to analyse biased behaviours. The detailed results are shown in Appendices Table XIV, Table XV, and Table XVI due to page limits.

1) Investigating General Bias Issues: This section analyses the general bias issues in Deepfake detection based on RP-vs-CRP plots as shown in Figure 3. In these plots, the relative performance (RP) for each attribute is shown with respect to the corrected relative performance (CRP). As mentioned in Section III-A, *RP* describes the ratio of the performance for images with a certain attribute versus the performance without this attribute. Consequently, $RP(a) = -100\%$ for an attribute a means that the error is twice as high if the image has this attribute than without it. Since the testing data is imbalanced for many attributes, the CRP was introduced in Section III-A to remove the influence of data imbalance. Consequently, attributes that lie in the top area (green) of the RP-vs-CRP plots indicate an increased detection performance and, contrarily, attributes that lie in the bottom (red) indicate increased detection errors. Moreover, each plot contains a bisectrix line where the attributes close to this line are less affected by imbalanced testing data than attributes away from it.

The RP-vs-CRP plots in Figure 3 are shown for three models on four Deepfake detection databases. The plots show

TABLE III

Annotation Correctness Study - Annotation Correctness of the Utilized Annotation Generator Is Compared With the Annotations of Three Human Evaluators [18]. Compared to Similar Large-Scale Facial Annotation Classifiers Used for Databases, Such as LFW [53] (72% Annotation Accuracy) and CelebA [54] (85% Annotation Accuracy), the Proposed Annotations Are of High Correctness [18] (92% Annotation Accuracy)

Category	Attribute	Accuracy	Precision	Recall
Demographics	Male	0.99	0.98	1.00
	Young	0.99	1.00	0.98
	Asian	0.90	0.88	0.92
	White	0.89	1.00	0.82
	Black	0.94	0.90	0.98
Skin	Shiny Skin	0.77	0.84	0.74
Hair	Bald	0.96	0.92	1.00
	Wavy Hair	0.99	1.00	0.98
	Receding Hairline	0.77	0.54	1.00
	Bangs	0.98	0.96	1.00
	Black Hair	0.98	0.96	1.00
	Blond Hair	1.00	1.00	1.00
Beard	No Beard	0.98	1.00	0.96
	Mustache	0.98	0.98	0.98
	Goatee	0.95	0.90	1.00
Face Geometry	Oval Face	0.81	0.90	0.76
	Square Face	0.80	0.78	0.81
	Double Chin	0.94	0.88	1.00
	Chubby	0.94	0.88	1.00
	Obstructed Forehead	0.91	0.94	0.89
	Fully Visible Forehead	0.80	0.75	1.00
Mouth	Mouth Closed	0.84	0.80	0.87
	Smiling	0.95	1.00	0.91
	Big Lips	0.70	0.50	0.83
Nose	Big Nose	0.97	0.98	0.96
	Pointy Nose	0.88	0.88	0.88
Accessories	Heavy Makeup	0.98	0.98	0.98
	Wearing Hat	0.92	0.84	1.00
	Wearing Lipstick	0.95	0.90	1.00
	No Eyewear	0.98	0.98	0.98
	Eyeglasses	0.90	0.80	1.00
Other	Attractive	1.00	1.00	1.00
		0.92	0.90	0.94

strong influences of most of the investigated attributes on the performance, indicating strongly biased Deepfake detectors. For instance, the analysis of EfficientNetB0 on Celeb-DF shows that having a big nose/big lips/or being black or chubby leads to more than twice the detection errors compared to images without these attributes. This shows serious fairness issues of these models when these are applied in real-world applications for specific category of people. In general, most attributes can be observed as strong factors leading to unfair performance differences in DeepFake detection.

Moreover, we have observed that training Deepfake detection backbone models on various datasets results in significant variations in the influence of attributes on detection performance. For instance, the misclassification of the pattern *Obstructed Forehead* is observed in the Celeb-DF and DFDC datasets. This finding suggests that both the selection of Deepfake detection backbone networks and the choice of datasets may significantly impact bias in the system.

To conclude, the experimental results demonstrate that the analysed Deepfake detection backbone models are strongly biased against a variety of demographic and non-demographic attributes. The variation of the biased performances across the models and databases indicates that this bias originates from several sources such as unbalanced training data, the utilised network, and their training process. The observed attribute-related variation in performances shows a strong need for mitigating bias in Deepfake detection models.

2) Investigating Bias Issues in Pristine and Fake Data: To investigate the bias issues in DeepFake detection in more detail, we conduct another analysis for pristine and fake data individually in this section. The results of this analysis is shown in Figure 4, for three Deepfake detection backbone models on four databases. The pristine data relative performance (*PDRP*) refers to the *CRP* that is only evaluated on pristine data and, analogous, the Deepfake data relative performance (*DDRP*) refers to the *CRP* that is calculated on fake data only. For an attribute a, a negative *CRP* on the pristine data means that people having this attribute are more likely to be falsely detected as fakes than people without these attributes. A negative *CRP* on fake data means that fake images that are generated with such an attribute are less likely to be detected as fake and thus, demonstrate weak points that attackers are likely to exploit. Each plot also contains bisectrix line where attributes that lie close to this line have a similar affect on pristine data than on Deepfake. Attributes placed above this line have a higher *CRP* on the pristine than on the Deepfake. Conversely, attributes below this line have a higher *CRP* on Deepfake than on the pristine data.

The results clearly show that the effect of the investigated attributes on the detection performance strongly differ between pristine and fake data since most attributes lie far away from the bisectrix line. Analysing the four quadrants of the plots shows that in most cases the attributes are distributed in all four. Attributes in quadrant I (top right), indicate that the attributes have the same positive effect on the performance, while attribute in quadrant III (bottom left), have the same negative effect on the detection performance. Observed performance in these areas indicates a similar biased effect on the decision. Attributes in quadrant II (top left) and IV (bottom right) show the opposite effect on the detection performance on pristine and Deepfake data. Consequently, for attributes in these areas, the model learnt the critical assumptions that the presence of the attribute is an indicator for Deepfake detection decision. For instance, the analysis of EfficientNetB0 on Celeb-DF for attribute wearing hat, shows a positive *DDPR* $\approx 100\%$ and a negative *PDPR* $\approx -75\%$. Consequently, if a person in a Deepfake image is wearing a hat the model comes twice as often to the right decision than if the person is not wearing a hat. Conversely, if a real person with hat is analysed by the model it leads to nearly twice as many errors as without a hat. The model in this case has seemingly learnt the presence of a hat as a strong indicator for the Deepfake data. Such observations point to questionable assumptions learned by the network and can result in biased detection performance.

In general, the observations reveal similar trends and patterns corresponding to the investigation from Section V-B1. The biased performances for the different attributes vary across

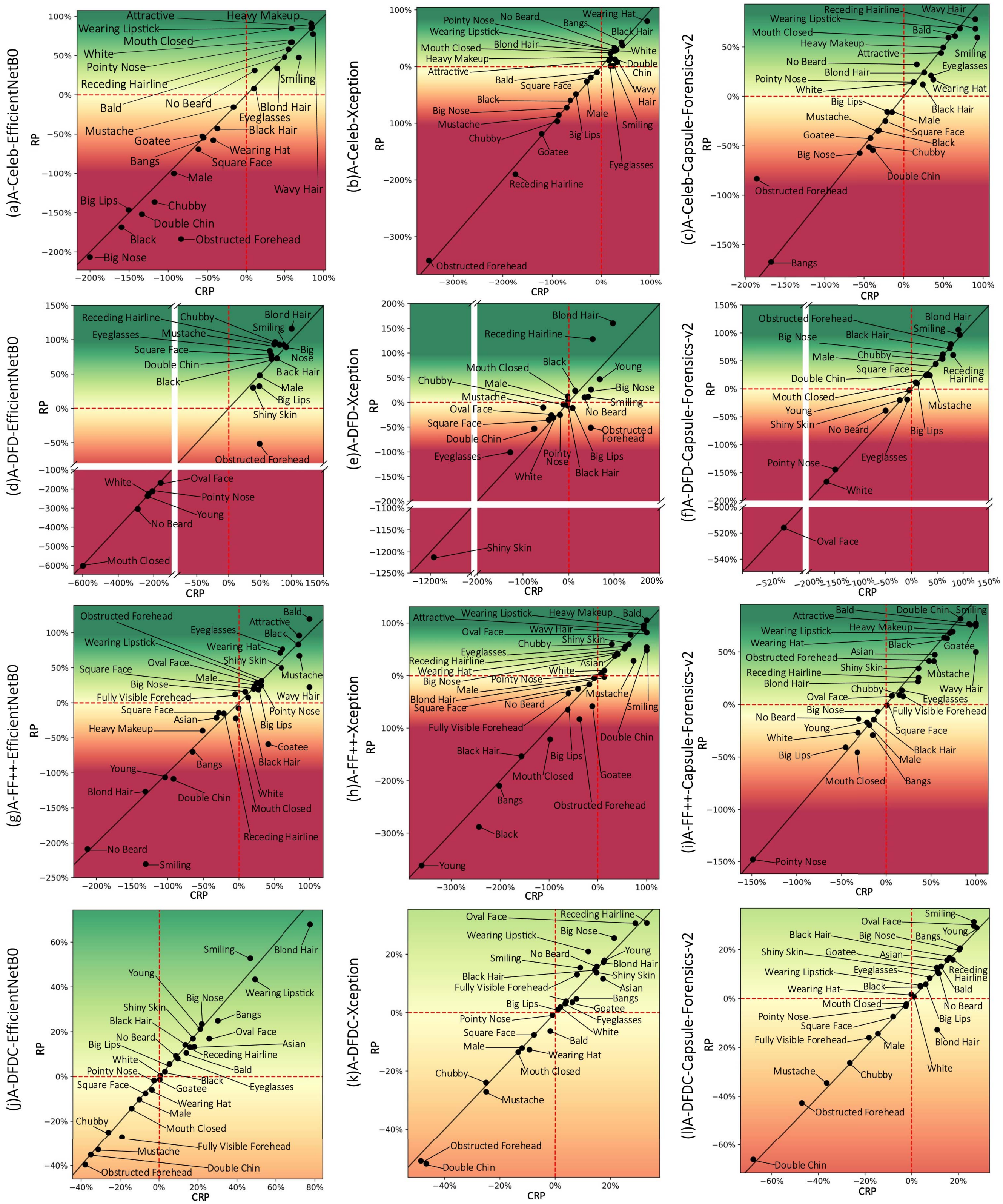


Fig. 3. **Bias analysis** - The relative performance *RP* is reported with respect to the corrected relative performance *CRP* using three Deepfake detection backbone models, EfficientNetB0 [19], Xception [20], and Capsule-Forensics-v2 [21] on four annotated databases, A-Celeb-DF, A-DFD, A-FF++, and A-DFDC. Many attributes strongly influence the detection performance.

the utilised models and training databases. If a Deepfake person has a goatee, a big nose, is chubby, male, or black, the probability that the model (EfficientNetB0 on Celeb-DF) comes to a wrong decision is doubled compared to persons without these attributes. The detection models therefore show strong biases leading to fairness issues in real-life applications

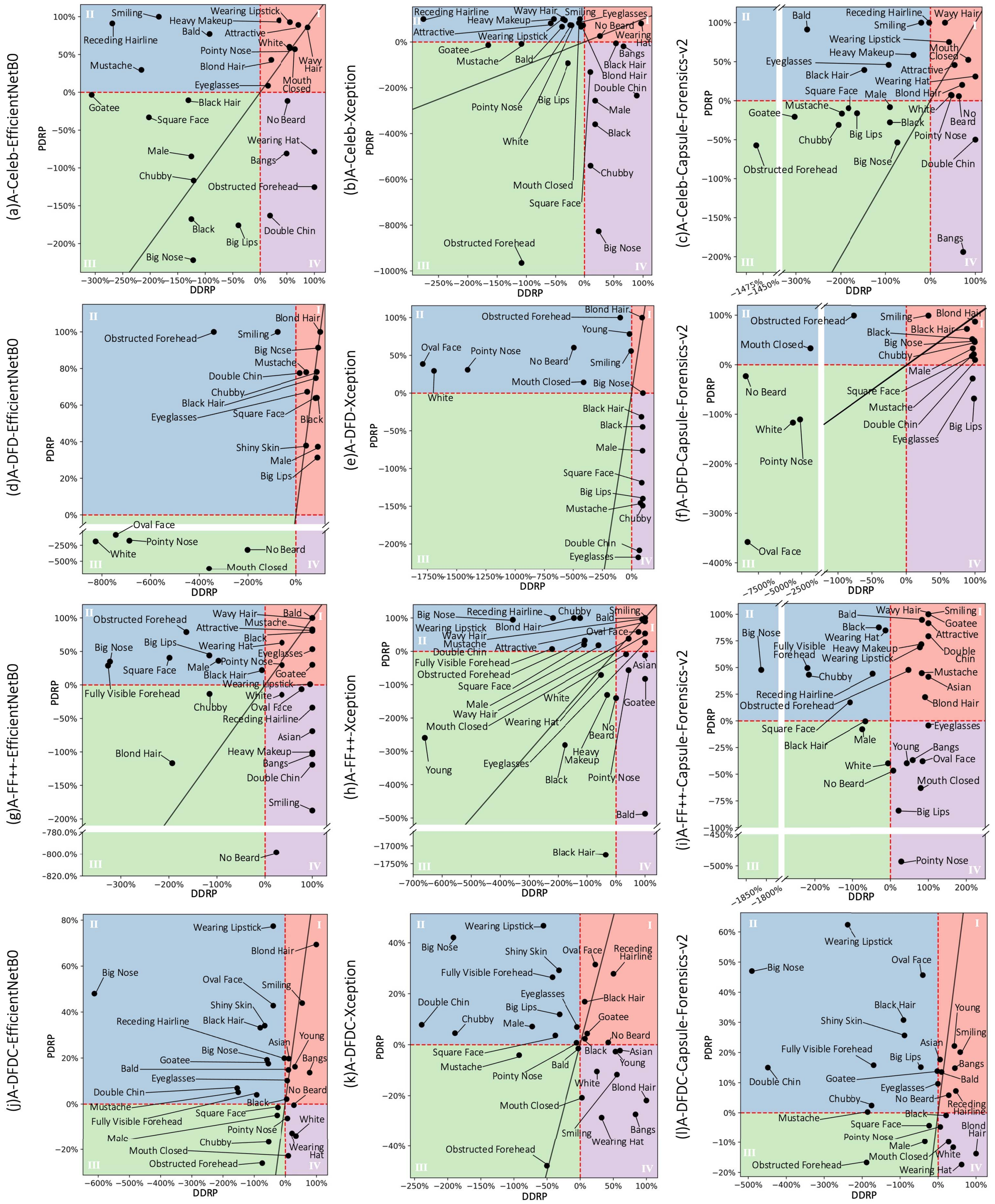


Fig. 4. **Bias analysis on Pristine Data and Deepfake Data** - The *CRP* on the pristine data (PDRP) is reported with respect to the *CRP* on Deepfake data (DDRP) using three Deepfake detection backbone models, EfficientNetB0 [19], Xception [20], and Capsule-Forensics-v2 [21] on our four annotated databases, A-Celeb-DF, A-DFD, A-FF++, and A-DFDC. Many of the attributes strongly influence the detection performance.

if deployed without considering the attribute distribution. It should also be kept in mind that this analysis limits its investigation to the influence of single attributes on the detection performance. The analysis of multiple attributes can be asserted to lead to an exponential increase in its bias effects. However, this aspect is not considered in this work.

To conclude, the impact of biased performance for the analysed attributes on detection accuracy varies significantly between pristine and fake data for several attributes. The results suggest that the models learn several questionable assumptions that the presence of a certain attribute, such as if the person is smiling or wears a hat, is an strong indicator for Deepfake detection decision. Lastly, the investigated Deepfake detection backbone models have demonstrated unfair behavior, with a significant increase in the probability of making incorrect decisions when presented with specific attributes such as having a big nose or belonging to a particular gender or race. This bias in current Deepfake detectors affects their accuracy and limits their generalisability. In other words, these biases may cause the Deepfake detectors to perform well on certain datasets or scenarios, but may fail to perform effectively in others, especially those where such attributes are different or not present. Therefore, addressing these biases and improving the generalisability of Deepfake detectors is crucial to ensure their robustness and reliability in real-world applications.

VI. Key Findings & Recommendations for Future Works

In the following, we summarise our key findings from our bias investigation of three Deepfake detection backbone models in four databases with respect to 31 demographic and non-demographic attributes:

- *Deepfake detection databases and strong attribute imbalance* - The investigated Deepfake detection databases (Celeb-DF, DFD, FF++, and DFDC) lack diversity for most analysed attributes. Future works should aim to provide more unbiased, balanced, and diverse datasets to prevent the development of potentially biased Deepfake detection algorithms.
- *Strongly correlating attribute pairs in current Deepfake detection databases* - Future works using these databases (or our annotations) should take into account that some attributes show strong pairwise correlations to prevent misinterpretations in their results.
- *Deepfake detection backbone networks and demographic/non-demographic attributes* - The results demonstrate that the analysed Deepfake detection backbone models are strongly biased for a variety of demographic and non-demographic attributes. The variation of the biased performances across the models and databases indicates bias possibly originating from several sources such as imbalanced training data, the utilised network, and their training process. Low generalizability of current Deepfake detection methods can also be attributed to these omnipresent biases or imbalanced attributes. We expect bias-mitigating Deepfake detection solutions in future work can also improve the generalizability.
- *Bias due to imbalance in attributes for pristine and Deepfake data* - For many of the investigated attributes, the biased performance similarly affects the pristine and Deepfake data. However, also the strong opposite behaviour was observed for many attributes leading the models to learn potentiality wrong patterns.
- *Deepfake detection backbone models and questionable assumptions* - The results suggest that the models tend to learn questionable assumptions where the presence of a certain attribute, such as if the person is smiling or wears a hat, is a strong indicator for Deepfake. Although this could have originated due to training data distribution, our analysis is limited and indicates it as a potential topic in future works to enhance the reliability of these systems.
- *Deepfake detection backbone models and societal security* - The presence of a certain attribute in a Deepfake image resulted in an increased error rate, several times higher than for a Deepfake without this attribute. Attackers can likely exploit these issues to increase their chances of overcoming Deepfake detection if unaddressed. On the other hand, the strong performance differences based on the presence of an attribute show a strong unfairness of these models. Future works therefore should focus on mitigating bias problems for Deepfake detection for the sake of security and society.

Based on the key observations of the three backbone networks analysed, there appears to be a significant research gap in developing Deepfake detection methods suitable for real-world applications. However, further analysis of additional methods may be necessary to make a more definitive statement. Our analysis points to a need for more diverse and richly annotated databases for training and testing, as well as developing bias-mitigating Deepfake detection approaches.

A. Limitations of Our Analysis

While our study reveals bias issues in Deepfake detection datasets and AI-based detectors, it is essential to note the difference between correlation and causation in our analysis. Our results demonstrate strong correlations between attributes and biased performance in detection, but they do not necessarily establish causation. Bias in datasets arises from various complex factors, including data collection methodologies, historical biases, and societal contexts. While we provide valuable insights into the presence of bias, further research is needed to ascertain the causative factors responsible for these biases. This understanding is crucial for designing effective strategies to mitigate bias in Deepfake detection and to develop more equitable and reliable detectors. Further, our work does not analyse all available state-of-the-art detection approaches leaving an open question on architecture dependence and fairness factors.

VII. Conclusion

In this work, we provided large-scale annotations for five popular Deepfake detection datasets and used these to comprehensively analyse bias in Deepfake detection. While existing Deepfake detection databases are only sparsely annotated, we closed this gap by making over 65.3M annotations of 47 different attributes for five Deepfake detection datasets publicly available. Based on these datasets, we comprehensively analyse bias-causing factors in Deepfake detection purely from an attribute perspective. The results indicated that both the datasets as well as the state-of-the-art AI-based Deepfake detectors trained on this data, demonstrate strong bias issues for many demographic and non-demographic attributes. Depending on the use case, the biased performance can result in serious societal fairness and security problems. Moreover, imbalanced attributes in these datasets can further lead to generalisation problems across different attributes in current Deepfake detection algorithms. Our findings from the study and proposed publicly-available annotations are expected to help future works to effectively evaluate and mitigate bias issues in Deepfake detection and thus, to develop reliable Deepfake detectors.

Acknowledgment

Parts of this work was carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Programme.

Statement of Ethical Use of Datasets

We affirm that all data utilized in this study from public databases adhere to ethical guidelines, and no special ethical clearance is applicable for publicly available information.

References

[1] R. Tolosana, R. Vera-Rodríguez, J. Fiérrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2020.06.014

[2] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. 27th Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2672–2680. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[3] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015. [Online]. Available: https://doi.org/10.1016/j.neunet.2014.09.003

[4] "Reface: face swapp videos." Accessed: Oct. 2, 2022. [Online]. Available: https://hey.reface.ai/

[5] "Faceswap: Open source multi-platform deepfakes software." Accessed: Oct. 2, 2022. [Online]. Available: https://faceswap.dev/

[6] T. Zhang, "Deepfake generation and detection, a survey," *Multim. Tools Appl.*, vol. 81, no. 5, pp. 6259–6276, 2022. [Online]. Available: https://doi.org/10.1007/s11042-021-11733-y

[7] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. Canton-Ferrer, "Towards measuring fairness in AI: The casual conversations dataset," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 4, no. 3, pp. 324–332, Jul. 2022. [Online]. Available: https://doi.org/10.1109/TBIOM.2021.3132237

[8] L. Trinh and Y. Liu, "An examination of fairness of AI models for deepfake detection," in *Proc. 30th Int. Joint Conf. Artif. Intell., (IJCAI)*, Montreal, QC, Canada, 2021, pp. 567–574. [Online]. Available: https://doi.org/10.24963/ijcai.2021/79

[9] M. Pu, M. Y. Kuan, N. T. Lim, C. Y. Chong, and M. K. Lim, "Fairness evaluation in deepfake detection models using metamorphic testing," in *Proc. 7th Int. Workshop Metamorph. Test.*, 2022, pp. 7–14. [Online]. Available: https://doi.org/10.1145/3524846.3527337

[10] A. Zhadan (Cybernews, Vilnius, Lithuania). *Power of Deepfakes: Three Times the World Fell for Dangerous Fakes*. 2022. Accessed: Jul. 2, 2022. [Online]. Available: https://cybernews.com/editorial/power-of-deepfakes-three-times-the-world-fell-for-dangerous-fakes/

[11] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2020, pp. 145–151.

[12] R. A. Waelen, "The struggle for recognition in the age of facial recognition technology," *AI and Ethics*, vol. 3, no. 1, pp. 215–222, 2023.

[13] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Seattle, WA, USA, 2020, pp. 3204–3213. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.html

[14] "Contributing data to deepfake detection research." Sep. 2019. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[15] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., (ICCV)*, Seoul, South Korea, 2019, pp. 1–11. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00009

[16] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Seattle, WA, USA, 2020, pp. 2886–2895. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Jiang_DeeperForensics-1.0_A_Large-Scale_Dataset_for_Real-World_Face_Forgery_Detection_CVPR_2020_paper.html

[17] B. Dolhansky et al., "The deepfake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[18] P. Terhörst, D. Fährmann, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Maad-face: A massively annotated attribute dataset for face images," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 3942–3957, 2021. [Online]. Available: https://doi.org/10.1109/TIFS.2021.3096120

[19] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn., (ICML)*, Long Beach, CA, USA, 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 1800–1807. [Online]. Available: https://doi.org/10.1109/CVPR.2017.195

[21] H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," 2019, *arXiv:1910.12467*.

[22] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Vancouver, BC, Canada, 2023, pp. 943–952. [Online]. Available: https://doi.org/10.1109/CVPRW59228.2023.00101

[23] Y. Xu, K. B. Raja, and M. Pedersen, "Supervised contrastive learning for generalizable and explainable deepfakes detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops, (WACV)*, Waikoloa, HI, USA, 2022, pp. 379–389. [Online]. Available: https://doi.org/10.1109/WACVW54805.2022.00044

[24] Y. Li, M. Chang, and S. Lyu, "In Ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur., (WIFS)*, Hong Kong, China, 2018, pp. 1–7. [Online]. Available: https://doi.org/10.1109/WIFS.2018.8630787

[25] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Seattle, WA, USA, 2020, pp. 5000–5009. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.html

[26] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Seattle, WA, USA, 2020, pp. 2841–2850. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w39/Guarnera_DeepFake_Detection_by_Analyzing_Convolutional_Traces_CVPRW_2020_paper.html

[27] L. Chai, D. Bau, S. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. 16th Eur. Conf., Comput. Vis. (ECCV)*, Glasgow, U.K., 2020, pp. 103–120. [Online]. Available: https://doi.org/10.1007/978-3-030-58574-7_7

[28] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," 2019, *arXiv:1911.00686*.

[29] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. 16th Eur. Conf., Comput. Vis. (ECCV)*, Glasgow, U.K, 2020, pp. 86–103. [Online]. Available: https://doi.org/10.1007/978-3-030-58610-2_6

[30] S. McCloskey and M. Albright, "Detecting GAN-generated imagery using saturation cues," in *Proc. IEEE Int. Conf. Image Process., (ICIP)*, Taipei, Taiwan, 2019, pp. 4584–4588. [Online]. Available: https://doi.org/10.1109/ICIP.2019.8803661

[31] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," *Signal Process.*, vol. 174, Sep. 2020, Art. no. 107616.

[32] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.

[33] H. Qi et al., "Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms," in *Proc. 28th ACM Int. Conf. Multimedia, (MM 20)*, Seattle, WA, USA, 2020, pp. 4318–4327. [Online]. Available: https://doi.org/10.1145/3394171.3413707

[34] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, Brighton, U.K., 2019, pp. 8261–8265. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8683164

[35] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (CVPR)*, 2021, pp. 5039–5049. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Haliassos_Lips_Dont_Lie_A_Generalisable_and_Robust_Approach_To_Face_CVPR_2021_paper.html

[36] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., (ICCV)*, Montreal, QC, Canada, 2021, pp. 15024–15034. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01477

[37] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., (ICCV)*, Montreal, QC, Canada, 2021, pp. 15088–15097. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01483

[38] P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[39] A. Khodabakhsh, R. Raghavendra, K. B. Raja, P. S. Wasnik, and C. Busch, "Fake face detection methods: Can they be generalized?" in *Proc. Int. Conf. Biometr. Spec. Interest Group, (BIOSIG)*, Darmstadt, Germany, 2018, pp. 1–6. [Online]. Available: https://doi.org/10.23919/BIOSIG.2018.8553251

[40] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Seattle, WA, USA, 2020, pp. 5073–5082. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Advancing_High_Fidelity_Identity_Swapping_for_Forgery_Detection_CVPR_2020_paper.html

[41] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "KoDF: A large-scale Korean deepfake detection dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., (ICCV)*, Montreal, QC, Canada, 2021, pp. 10724–10733. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01057

[42] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "Beyond identity: What information is stored in biometric face templates?" in *Proc. IEEE Int. Joint Conf. Biometr., (IJCB)*, Houston, TX, USA, 2020, pp. 1–10. [Online]. Available: https://doi.org/10.1109/IJCB48548.2020.9304874

[43] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "On soft-biometric information stored in biometric face embeddings," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 4, pp. 519–534, Oct. 2021. [Online]. Available: https://doi.org/10.1109/TBIOM.2021.3093920

[44] P. Terhörst et al., "A comprehensive study on face recognition biases beyond demographics," *IEEE Trans. Technol. Soc.*, vol. 3, no. 1, pp. 16–30, Mar. 2022.

[45] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surveys*, vol. 54, no. 6, pp. 1–35, 2022. [Online]. Available: https://doi.org/10.1145/3457607

[46] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 25–34, Aug. 2021. [Online]. Available: https://doi.org/10.1109/MIS.2020.3000681

[47] S. Seferbekov. "Selimsef/dfdc_deepfake_challenge." 2020. [Online]. Available: https://github.com/selimsef/dfdc_deepfake_challenge

[48] W. Zhou, H. Zhao, and H. Cui, "Cuihaoleo/kaggle-dfdc." Accessed: Jun. 14, 2020. [Online]. Available: https://github.com/cuihaoleo/kaggle-dfdc

[49] "NTech-Lab/deepfake-detection-challenge." Accessed: Jun. 8, 2020. [Online]. Available: https://github.com/NTech-Lab/deepfake-detection-challenge

[50] "Siyu-C/RobustForensics." Accessed: Jun. 11, 2020. [Online]. Available: https://github.com/Siyu-C/RobustForensics

[51] I. Pan and J. Howard. "Jphdotam/DFDC." Accessed: Jun. 7, 2020. [Online]. Available: https://github.com/jphdotam/DFDC/

[52] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security, (WIFS)*, Hong Kong, China, 2018, pp. 1–7. [Online]. Available: https://doi.org/10.1109/WIFS.2018.8630761

[53] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life' Images, Detect., Align., Recognit.*, 2008, pp. 1–11.

[54] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis., (ICCV)*, Santiago, Chile, 2015, pp. 3730–3738. [Online]. Available: https://doi.org/10.1109/ICCV.2015.425

[55] P. Terhörst et al., "Reliable age and gender estimation from face images: Stating the confidence of model predictions," in *Proc. 10th IEEE Int. Conf. Biometr. Theory, Appl. Syst., (BTAS)*, Tampa, FL, USA, 2019, pp. 1–8. [Online]. Available: https://doi.org/10.1109/BTAS46853.2019.9185975

[56] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016. [Online]. Available: https://doi.org/10.1109/LSP.2016.2603342

[57] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (CVPR)*, 2021, pp. 2185–2194. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.html

[58] Y. Xu, K. B. Raja, L. Verdoliva, and M. Pedersen, "Learning pairwise interaction for generalizable deepfake detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops, (WACV)*, Waikoloa, HI, USA, 2023, pp. 1–11. [Online]. Available: https://doi.org/10.1109/WACVW58289.2023.00074

[59] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (CVPR)*, 2021, pp. 16317–16326. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Generalizing_Face_Forgery_Detection_With_High-Frequency_Features_CVPR_2021_paper.html

[60] D. M. Montserrat et al., "Deepfakes detection with automatic face weighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Seattle, WA, USA, 2020, pp. 2851–2859. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w39/Montserrat_Deepfakes_Detection_With_Automatic_Face_Weighting_CVPRW_2020_paper.html

[61] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit., (ICPR)*, Istanbul, Turkey, 2010, pp. 3121–3124. [Online]. Available: https://doi.org/10.1109/ICPR.2010.764

Ying Xu (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from Shanghai University, China, in 2015, and the M.Sc. degree in applied computer science from the Norwegian University of Science and Technology, Norway, in 2021, where she is currently pursuing the Ph.D. degree, focusing on Deepfake detection.

Philipp Terhörst (Member, IEEE) received the Master of Science degree in physics from the Technical University of Darmstadt in 2017, and the Ph.D. degree in computer science for his work on "Mitigating Soft-Biometric Driven Bias and Privacy Concerns in Face Recognition Systems." in 2021. Since 2017, he has been working with the Smart Living and Biometric Technologies Department, Fraunhofer Institute for Computer Graphics Research (IGD) as a Research Scientist and as a Ph.D. student with the Technical University of Darmstadt. His areas of specialization include topics in machine learning as well as biometric face recognition with a focus on quality assessment, privacy, and fairness. He is the author of several publications in conferences and journals, such as CVPR and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and regularly works as a Reviewer for, e.g., IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, PR, BTAS, and *Integrative and Comparative Biology*. For his scientific work, he received several awards, such as the 2020 EAB Biometrics Industry Award from the European Association for Biometrics for his dissertation or the IJCB 2020 Qualcomm PC Chairs Choice Best Student Paper Award. He furthermore participated in the 'Software Campus' Program, a management program of the German Federal Ministry of Education and Research (BMBF).

Marius Pedersen (Member, IEEE) received the B.Sc. degree in computer engineering and the media technology from Gjøvik University College, Norway, in 2006 and 2007, respectively, and the Ph.D. degree in color imaging from the University of Oslo, Norway, in 2011, sponsored by Océ. He is a Professor with the Department of Computer Science, NTNU in Gjøvik, Norway. He is also the Director of the Norwegian Colour and Visual Computing Laboratory. His work is centered on subjective and objective image quality.

Kiran Raja (Senior Member, IEEE) received the Ph.D. degree in computer Science from the Norwegian University of Science and Technology, Norway, in 2016, where he is Faculty Member with the Department of Computer Science. He was/is participating in EU projects SOTAMD, iMARS, and other national projects. His main research interests include statistical pattern recognition, image processing, and machine learning with applications to biometrics, security, and privacy protection. He is a member of the European Association of Biometrics, chairs the Academic Special Interest Group at EAB, and the Section Chair of IEEE Norway.