2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Deepfake: Classifiers, Fairness, and Demographically Robust Algorithm

Akshay Agarwal[1] and Nalini Ratha[2]
[1]IISER Bhopal, India and [2]University at Buffalo, USA
[1]akagarwal@iiserb.ac.in, [2]nratha@buffalo.edu

Abstract— **Deepfake detection research has seen tremendous success and has achieved remarkably high performance on a few existing datasets. However, the significant drawback of the existing works is the generalizability of the detection algorithms under cross-datasets and cross-attack/manipulation settings. On top of that, another critical bottleneck of deepfake detection literature is the understanding of the fairness quotient of these algorithms. One big reason for such a less explored domain is the unavailability of deepfake datasets covering multiple ethnicities and genders with proper annotations. For example, the popular deepfake detection datasets such as FaceForensics++ and Celeb-DF are highly biased toward Caucasian ethnicity. Recently, a multi-ethnicity multi-modal dataset namely FakeAVCeleb has been released which can fulfill this gap. Henceforth by utilizing the potential of this dataset, *we have performed the fairness study of deepfake detection algorithms*. For that, several image classifiers are selected which range from *deep convolutional neural networks to handcrafted image feature extraction to vision transformers*. The experiments performed using such a wide variety of classifiers reveal that the deepfake detectors are not fair and can detect one ethnicity with high accuracy but fail miserably on others. For instance, the performance of one of the popular deepfake detection networks namely XceptionNet shows a reduction of more than 30% when dealing with different ethnicities and genders. Not only ethnicity or gender but also the type of classifiers have a huge impact on the performance. We assert that the proposed study can help in building a fair, robust, and accurate deepfake classifier utilizing insightful findings that can help in the selection of an effective and robust backbone architecture.**

I. Introduction

With the availability of computing devices and software, the generation of deepfake videos becomes a couple of minutes task [3], [7], [8], [32], [45]. These deepfake videos are heavily used in several malicious practices such as pornography harassment [21] and looting millions of dollars of money [41]. We recently witnessed a surge of misinformation through these deepfake videos including during the time of the Ukraine and Russia war. Deepfake videos are also heavily used to achieve unwanted advantages in several national elections including in India, South Korea, and the USA. It shows that the impact of deepfake videos is not limited to any particular demography but is affecting the population of the entire world [14], [25] and can fulfill 'any' mischievous purpose. Therefore, the effective detection of fake videos including deepfake is extremely critical [2], [20].

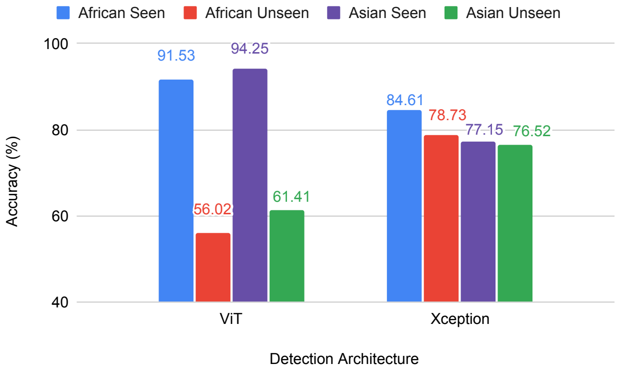

Fig. 1. Sensitivity of two deepfake detection architectures when they evaluated in seen (same gender train-test) and unseen (cross-gender train-test) domains deepfake images. The biasness is not even due to out-of-distribution in training-testing images (unseen) but also in seen training-testing as well. For example, even under-seen setting, the accuracy of XceptionNet is 7.46% lower for Asians than Africans.

The impact is widespread and equal to different ethnicities and genders; interestingly, still, the majority of the deepfake datasets contain images/videos of Caucasian ethnicities ignoring the negative impact on other demographic entities. Therefore, it became extremely difficult to protect other ethnicities by building a universally robust deepfake detector. Henceforth, by looking at ethnicity-agnostic impact, recently researchers have started developing deepfake datasets of other ethnicities such as Indian [34], [35] and Korean [27]. While these datasets cover not-so-popular deepfake ethnicities they come with single ethnicity only; hence, ethnicity-agnostic detection is not fairly possible. The prime reason to study the fairness in deepfake detection can be understood from the fact that machine learning classifiers are found biased toward demographic information including ethnicity and gender [4], [36], [43]. Therefore, it is critical to ensure that the deepfake detection architectures are *bias-free and fair on different demographic entities*. Recently, Trinh et al. [48] have performed a fairness study of three deepfake detection algorithms namely MesoNet, Face-Xray, and XceptionNet. The study has been performed using the FaceForensics++ dataset and the authors have revealed that the detectors trained do not possess gender bias. However, it is observed that the dataset used consists of highly imbalanced ethnicity and gender-related videos. For instance, more than 60% of the videos in the dataset belong to Caucasian ethnicity. Furthermore, the authors have not used sophisticated deepfake videos, but have used the 68 landmark strategy to generate the blended images. Another study [50] utilizing such an imbalanced dataset not only for training but also in an evaluation led to misleading findings as shown in the recent paper [4]. One more recent work tackles the issue of gender bias

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

only through a novel dataset [38]. However, the authors have not tackled other critical demographic factors, which might lead to shallow findings. Therefore, we can safely assume the absence of a comprehensive and fair study to analyze the bias and fairness issues in deepfake detection algorithms covering ethnicity, gender, and type of classifier. *In this paper, we aim to fill this gap by presenting a comprehensive and fair study by utilizing deepfake datasets balanced across different demographic modalities and covering a variety of classifiers that are popular in several computer vision tasks.*

We have utilized several deepfake detection algorithms ranging from handcrafted features to standard deep learning architectures to modern vision transformers. The experiments are conducted using balanced datasets of four different ethnicities and two genders. The extensive set of experiments reveals interesting and surprising outcomes concerning the biasness of deepfake detectors towards different ethnicities in the case when they are present or absent at the time of training. We believe the presence of such a detailed fairness study on deepfake detectors can help in building a bias-free algorithm or at least in a careful selection of the backbone architecture that reflects low bias. Fig. 1 gives a quick overview of why the fairness study of deepfake detectors is essential. We have showcased the sensitivity of two deep computer vision architectures namely vision transformer (ViT) [19] and a pure CNN namely XceptionNet [16] on deepfake images of two ethnicities. As expected both the classifiers are found highly vulnerable against the out-of-distribution ethnicity. However, as mentioned, the type of classifier and ethnicity also have an impact on the detection performance. For example, when the African deepfake videos are aimed to be detected, the performance of ViT is 91.53% which is 2.72% lower than the value obtained when defending deepfake on Asian. The above example shows that even if the detector is trained on individual ethnicity their performance can be drastically different. Another surprising finding, the type of classifier can also show such sensitivity. For example, the performance of Xception: a pure CNN architecture shows at least 6.92% lower accuracy than ViT even when the same distribution deepfake videos are utilized for training-testing. The above findings show that if a study covers a single dimension of fairness, it can lead to either misleading or shallow findings. To tackle the need for an unbiased deepfake detector, utilizing the findings of this research, in the end, *we have proposed a novel fair deepfake detector* by combining the decision of transformer model and handcrafted feature-based deepfake detectors.

II. Related Work

Swapping of faces/images has a long history and is claimed to have existed for more than a decade now [5], [11], [12]. Several research works have showcased the potential of different toolboxes ranging from mobile applications to internet websites to deep neural architecture to develop novel and large-scale face swap datasets [8], [7], [17], [30], [42]. Due to the significant threat of these face swaps and deepfake videos, several detection algorithms are also proposed in the literature. The deepfake and face swap video detection algorithms can be broadly divided into handcrafted plus machine learning classifiers and data-driven deep neural network algorithms. In one of the early works towards face swap detection, Agarwal et al. [7] proposed a novel feature engineering algorithm to highlight the subtle moiré patterns in the face swap videos. Other image feature algorithms used for deepfake detection are: eye color and missing reflections [33], 3D head poses [51], facial movements [10], and image artifacts [40], [52]. Zhao et al. [54] and Nirkin et al. [39] have proposed source image features and face contextual information extraction networks for deepfake detection. The deepfake detection network proposed by Zhao et al. [53] uses the multi-attention convolution network consisting of spatial attention heads and textural feature enhancement block. Agarwal et al. [1] have proposed the generalized convolutional network architecture utilizing two branches consisting of raw images and transformed images and introduced the cross-stitch connections to transfer knowledge among layers of two branches. Zhou et al. [55] have utilized both audio and video discrepancies for the detection of deepfake videos. DSP-FWA [31], Face X-ray [29], and PCL + I2G [54] proposed the boundaries in the facial regions which possibly exist due to the swapping of two faces. The details of the existing deepfake generation along with its countermeasures can also be found in the comprehensive survey papers [43], [37], [47].

As mentioned earlier the detection of deepfake videos through the development of sophisticated machine learning algorithms has received significant attention; however, no comprehensive study tackles the issue of fairness or bias. As seen machine learning algorithms whether working on image data text data or other input are found highly biased towards different demographic individuals [13], [15], [46]. Therefore, ignoring the issue of fairness in deepfake detection research can be problematic in building an effective defense algorithm. The proposed research takes a step toward that goal by analyzing the fairness issue of multiple detection algorithms and providing several insights on choosing the right deepfake detection architecture. The proposed study is impactful in the sense that the classifiers used for evaluation are popular not only in deepfake detection but also for several computer vision tasks. Therefore, the understanding of their shortcomings including in the form of fairness can help other computer vision domains as well.

III. Deepfake Detection Architectures, Dataset, and Protocols

In this section, we describe the architectures used for the fairness study in deepfake detection. Later, the dataset consisting of deepfake videos of multiple ethnicities and genders is described. In the end, the experimental protocols covering several interesting and challenging scenarios exploiting the full potential of the dataset are mentioned.

A. Detection Architectures

In this research, we have studied the fairness of several deepfake detection architectures which can be broadly categorized into three broad classes: (i) pure convolutional neural network (CNN), (ii) attention network, and (iii) handcrafted image features. Pure CNNs are when they are free of any attention or dynamic mechanism, e.g., squeeze-and-excitation, multi-head self-attention, or dynamic weights. Further, we have evaluated the fairness of two recent state-of-the-art generalized deepfake detectors namely MCX-API [49] and ID-unaware [18].

- Pure CNNs: In this research, we have used five widely used image classification architectures namely VGG16 (VGG) [24], XceptionNet (XNet) [16], MobileNet (MNet) [22], DensNet-121 (DNet) [24], and InceptionNet (INet) [44]. These architectures are not only in terms of the number of layers but also in the type of connections they use. For example, VGG consists of 16 layers; whereas, DenseNet is a significantly deeper architecture with 121 layers. XceptionNet is the extreme case of an InceptionNet. The InceptionNet applies multiple convolutional layers in parallel and concatenates the responses in the end. The variation in the network style and scale ensures the robustness of the proposed finding toward the fairness of deepfake detection. Further, being popular architectures, these architectures are heavily used for deepfake detection and general-purpose vision tasks either directly or as a backbone architecture [42], [1], [9]. Therefore, the evaluation of these networks can help future research in picking the fair architecture to build a robust deepfake detector. The networks are pre-trained on ImageNet and the last 40% of the layers are finetuned for deepfake detection. The weights are optimized for 50 epochs using an Adam optimizer with an initial learning rate set to 0.0001.
- Vision Transformers: Vision transformers (ViT) are recent computer vision architectures that have shown tremendous success in image and video classification. These transformers replace the convolution operations and use the popular attention mechanism by dividing the images into several non-overlapping patches. Several attention blocks are added sequentially similar to convolution blocks to learn a deep transformer model. In the very first vision transformer paper [19], the images are divided into 16×16 blocks and along with positional embedding of these blocks are fed into the self-attention blocks. The architecture is not only computationally heavy but also requires large-scale datasets for pre-training. Apart from that, Lee et al. [28] pointed out that self-attention layers in original ViT lack locality inductive bias and hence need a large amount of training data. The authors have proposed an advanced version of ViT by including a few more layers to overcome this limitation. In this research, we have used both these variants and referred to as ViT-O [19] and ViT-A [28]. These ViT models are trained from scratch using an adaptive learning rate of initial value of 0.001 and weight decay of 0.0001. The networks are trained for 50 epochs to minimize the sparse categorical cross-entropy loss using an Adam optimizer.
- Image Features and Classification: While deep learning architectures have shown tremendous success in image classification, one can not ignore the potential of image feature extraction algorithms and traditional image classifiers. Recently, in one such attempt, Agarwal et al. [7], [8] have proposed a novel image features extraction algorithm namely weighted local magnitude pattern ('WLMP') to detect face-manipulated images including deepfakes. The feature extraction coupled with a support vector machine classifier (SVM), i.e., WLMP + SVM demonstrated the state-of-the-art performance surpassing several deep learning architectures and helped in developing a green and responsible machine learning algorithm. We have followed the implementation detail used in the work by Agarwal et al. [7].

TABLE I
NUMBER OF REAL AND DEEPFAKE IMAGES USED IN THE RESEARCH BELONGING TO DIFFERENT GENDERS AND ETHNICITIES.

Class	Gender	African	Asian	American	European
Fake	Male	6974	7982	8643	8303
	Female	7120	8071	8752	8521
Real	Male	6570	7268	8228	8803
	Female	6555	7698	7948	8034

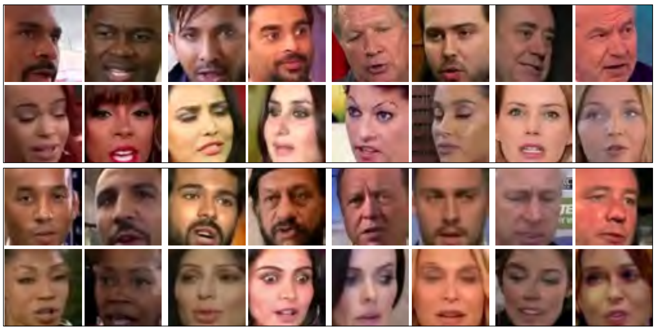

Fig. 2. Real and deepfake images of different ethnicities and genders in the dataset. The variation in images among ethnicities reflects the challenges in the detection and why the fairness study is important using such wide variation images.

B. Dataset

In the literature, several deepfake detection datasets are proposed; however, the majority of the datasets are biased towards Caucasian ethnicity and hence unfit for the fairness study. Recently, a novel dataset namely FakeAVCeleb[1] [26] is proposed which consists of deepfake videos of multiple ethnicities. Due to this property, in this research, we have used this dataset of four ethnicities: African, Asian (South), Caucasian (American), and Caucasian (European). On all

[1] https://github.com/DASH-Lab/FakeAVCeleb

TABLE II

SEEN SETTING AND GENDER AGNOSTIC DEEPFAKE DETECTION ACCURACIES (%) USING PURE CNN ARCHITECTURES. MNET, DNET, INET, AND XNET ARE MOBILENET, DENSENET-121, INCEPTIONNET-V3, AND XCEPTIONNET ARCHITECTURES, RESPECTIVELY. M AND F REPRESENT THE MALE AND FEMALE GENDER, RESPECTIVELY.

	Test	Train	MNet	VGG	DNet	INet	XNet
American	M	M	58.89	**99.86**	52.02	54.25	83.76
	M	F	52.02	**84.16**	52.02	52.16	68.47
	F	M	60.37	**81.54**	53.89	58.02	78.99
	F	F	53.93	**99.48**	53.89	54.09	63.57
European	M	M	47.88	**99.69**	47.57	51.66	60.36
	M	F	47.97	**80.63**	47.52	58.79	53.90
	F	M	52.35	**84.98**	52.39	48.52	70.23
	F	F	52.56	**99.61**	52.40	66.93	55.66
Asian	M	M	56.82	**99.51**	53.81	59.24	91.54
	M	F	58.67	**80.75**	53.71	58.01	75.93
	F	M	53.63	**79.98**	51.94	52.44	77.12
	F	F	56.81	**99.48**	52.19	63.45	62.69
African	M	M	58.92	**99.51**	52.47	52.80	91.99
	M	F	62.99	**86.31**	51.56	64.36	76.28
	F	M	57.86	**86.29**	53.39	54.10	84.19
	F	F	65.57	**99.17**	54.03	61.76	77.23

four ethnicities, deepfake and real videos of both male and female genders are presented in the datasets. In this research, we have used more than 125k real and deepfake images for the experimental purpose of performing an extensive fairness study. Individual images about ethnicities and genders are given in Table I. A few samples shown in Fig. 2 reflect a few possible variations that create an out-of-distribution scenario and make the detection tasks challenging. These changes in the testing images can be expected in the real-world unconstrained setting and might be one potential reason for detectors trained on one ethnicity-biased dataset perform poorly on other ethnicities.

C. Protocols

In this research, three sets of experiments are performed reflecting the following conditions: (i) seen distribution training testing, (ii) gender agnostic detection, and (iii) ethnicity agnostic detection. In the first condition, the images used in the training and testing belong to the same ethnicity and gender. While this setting is essential, it does not represent real-world scenarios. Therefore, gender agnostic protocol is developed where training and testing images correspond to different genders but might be coming from the same ethnicity. In the ethnicity-agnostic setting, the images that come for evaluation might be of a different ethnicity than the ones used for training. To perform these sets of experiments, first, the images of each type are divided into 40% training and 60% testing. The same testing set has been used across experiments whether belongs to seen or unseen settings. The datasets are divided in a way that training and testing images belong to different subjects to avoid any identity bias.

IV. RESULTS AND ANALYSIS

In this section, we provide the analysis observed through the extensive experiments performed using multiple deepfake detection algorithms on multiple traditional and generalized training-testing conditions. First, we present the analysis obtained from the pure CNN architectures; followed by the analysis of computational heavy vision transformers and computationally efficient WLMP algorithm. In the end, an analysis has been performed to understand which ethnicity or gender is effective if the testing set is completely unseen. In this case, we have used the benchmark FaceForensics++ [42] as an unseen and open-set evaluation dataset. *We assert that understanding through these experiments can help not only in boosting the deepfake detection accuracy but also ensure the fair behavior (bias-free) of the algorithm.*

A. Pure CNN Analysis

The analysis can be performed based on the following terms: (i) impact of gender, (ii) impact of classifier, and (iii) impact of ethnicity. The above analysis points can be further broken down into seen gender and unseen gender evaluation. The results of the pure CNNs in terms of seen gender and unseen gender are reported in Table II. In terms of CNNs, it is observed that the majority of the image classifiers whether evaluated in seen or unseen gender training testing settings show poor deepfake detection accuracy. The VGG network is found the most effective and yields almost perfect ($\sim$ 100%) detection accuracy in seen gender deepfake detection across each ethnicity. XceptionNet (XNet) is found the second best among all the classifiers used. Interestingly, in the deepfake detection literature [42], [1], XceptionNet is heavily explored in comparison to other networks, which we found is not the best architecture to use. In another surprising observation, DenseNet (DNet) in almost all the cases yields an accuracy close to random accuracy (50%). The two best-performing networks are found highly effective in detecting male deepfakes as compared to female deepfakes. While the sensitivity of the VGG is not significant between genders; the XNet is found highly biased towards the male class. The above analysis is observed when in training and testing images of the same gender are used. In the cross-gender setting, the XcepionNet shows contrasting performance. The detector trained on the male deepfake images is found less effective in detecting female deepfakes as compared to the deepfake detectors trained on females and evaluated on males. For example, when the XNet is trained on the American ethnicity images, in the seen gender evaluation setting, it yields 83.76% accuracy on male images which is 20.19% better than the accuracy on female images. Whereas, in the cross-gender setting, the accuracy on the female images is 10.52% higher than on the male images. Other networks such as MobileNet (MNet) and DenseNet (DNet) are found more effective in detecting female deepfake images than male images. Among all the ethnicities, Caucasian (European) ethnicity is found challenging to defend; whereas, the deepfake images of African ethnicity are found easy to detect. All five CNNs perform consistently better on African

TABLE III

ETHNICITY AGNOSTIC DEEPFAKE DETECTION ACCURACIES (%). MNET, DNET, INET, AND XNET ARE MOBILENET, DENSENET-121, INCEPTIONNET-V3, AND XCEPTIONNET ARCHITECTURES, RESPECTIVELY. M AND F REPRESENT THE MALE AND FEMALE GENDER, RESPECTIVELY. THE ACRONYMS FOR THE ETHNICITIES AFRICAN, AMERICAN, ASIAN, AND EUROPEAN ARE AFR, AMR, ASI, AND EURO, RESPECTIVELY.

Test	Train	MNet	VGG	DNet	INet	XNet
Amr - M	Euro - M	52.49	**87.81**	52.02	52.34	65.44
	Afr - M	57.36	**87.59**	52.02	53.21	74.37
	Asi - M	49.89	**84.33**	52.02	52.99	81.68
Amr - F	Euro - F	53.88	**86.48**	53.89	64.27	56.97
	Afr - F	65.30	**83.74**	54.03	58.41	64.63
	Asi - F	56.64	**85.97**	53.41	57.28	66.78
Euro - M	Amr - M	57.14	**83.81**	47.52	50.47	76.85
	Afr - M	55.95	**83.10**	47.50	48.58	77.10
	Asi - M	51.16	**82.79**	47.52	51.68	76.67
Euro - F	Amr - F	52.40	**84.86**	52.40	52.82	67.97
	Afr - F	63.68	**83.83**	52.43	59.06	68.27
	Asi - F	52.72	**86.29**	51.22	60.02	70.56
Asi - M	Euro - M	53.82	**79.88**	53.78	48.88	69.24
	Afr - M	61.66	**81.41**	53.81	55.12	77.69
	Amr - M	63.13	77.95	53.78	54.32	**78.72**
Asi - F	Euro - F	51.97	**79.61**	51.94	65.38	53.35
	Afr - F	63.56	**79.33**	51.98	56.73	63.88
	Amr - F	51.94	**79.94**	51.94	53.2	65.44
Afr - M	Euro - M	52.11	**86.20**	52.40	52.03	74.01
	Amr - M	62.69	**86.97**	52.44	52.97	83.75
	Asi - M	49.06	**84.79**	52.44	48.37	77.28
Afr - F	Euro - F	53.88	**85.15**	53.35	62.84	56.65
	Amr - F	53.35	**82.56**	53.35	53.78	70.03
	Asi - F	52.96	**85.13**	51.37	56.52	73.76

deepfake images as compared to other ethnicities irrespective of the training-testing condition. We want to highlight that ethnicity raises a significant challenge and can result in drastically different observations. For example, in a majority of the cross-gender cases across CNNs, when the detectors are trained on females they yield better performance on Caucasians (whether European or American) but on Asians and Africans, training on males surpasses the performance obtained using training on females.

As mentioned above, another important covariate in the dataset is ethnicity, and as we have seen it has a significant impact on deepfake detection performance. In this setting, we have kept the gender fixed while the ethnicity variable is changing between training and testing images. For example, if the detector is trained on American males, it has been tested on the male images of remaining ethnicities, i.e., African, Asian, and European. The quantitative analysis of ethnicity agnostic detection experiments is reported in Table III. Similar to seen and unseen gender results, in an ethnicity-agnostic setting, the VGG architecture performs the best and XNet performs second best. The training on Asian ('Asi') ethnicity is found least effective in detecting the other ethnicities across all the networks. Whereas, the detectors trained on Caucasian ethnicities whether American or European, are found most effective in handling unseen ethnicities. We believe such broad analysis reflecting the impact of ethnicity, gender, and type of classifier ensures that in the future either we use the most robust demographic deepfake videos/images or utilize the combination of robust classifiers and ethnicities.

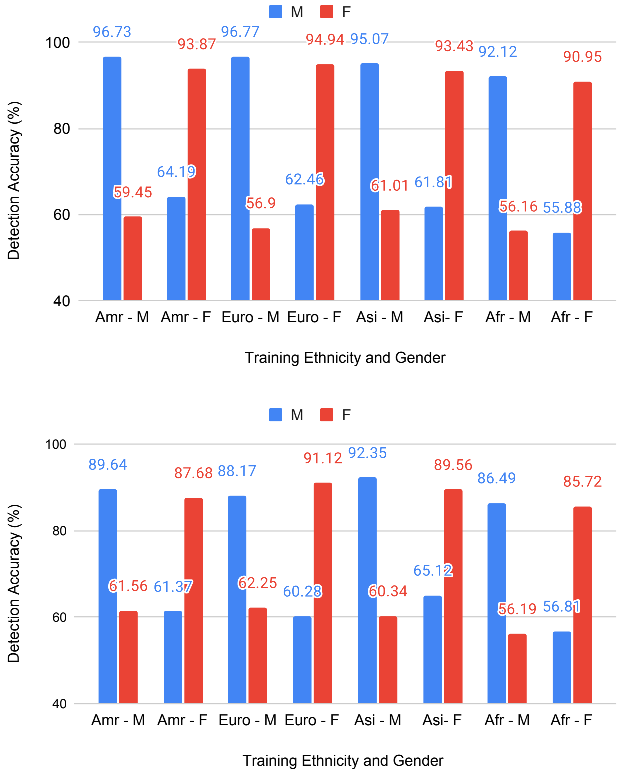

Fig. 3. Seen and unseen gender setting deepfake detection accuracies (%) using ViT-O (top) and ViT-A (bottom) architectures. M and F represent the male and female gender, respectively.

TABLE IV

SEEN AND UNSEEN ETHNICITY AND GENDER DEEPFAKE DETECTION ACCURACY (%) USING VIT-O. THE EXPERIMENTS REVEAL THAT BOTH ETHNICITY AND GENDER ARE IMPORTANT COVARIATES.

Train	Test	M	F	M	F
		Seen Gender		Cross Gender	
African	African	**92.12**	**90.95**	55.88	56.16
	Asian	55.60	58.08	57.78	53.44
Asian	African	56.14	57.85	53.03	62.00
	Asian	**95.07**	**93.43**	61.81	61.01
American	American	**96.73**	**93.87**	64.19	59.45
	European	63.06	61.81	60.90	55.76
European	American	58.57	64.36	60.21	63.09
	European	**96.77**	**94.94**	62.45	56.90

B. Vision Transformers Analysis

Similar to a comprehensive analysis of pure CNNs; the fairness of vision transformers (ViT) versions namely ViT-O and ViT-A are evaluated under several training testing conditions. In the first condition, seen ethnicity is seen gender experiments are performed; in the second category, the gender variable is kept fixed while the ethnicity variable is unseen. In the end, an analysis of the dual generalizability category where both variables are kept different in testing images from the training images is presented. The ViTs are found more effective than pure CNNs except compared to VGG.

The ViT-A [28] is an advanced version of ViT-O as the authors claimed that they have utilized two more layers to reduce the locality inductive bias issue. However, in terms of deepfake detection performance, we have observed that ViT-O outperforms ViT-A by a significant margin across experiments. While under cross gender setting the performance gap is low; in terms of seen gender testing, the gap can go up to 8.6%. Further, both the ViTs are found sensitive to the gender and ethnicity variable and show a significant drop in the detection accuracy. European ethnicity which was least effective when the pure CNNs were used for deepfake detection, shows high detection performance under ViT-O. The performance seen in gender training testing is found highest followed by Americans. We want to highlight that both these ethnicities are defined under Caucasian in the dataset. Interestingly, pure CNNs are found most effective in handling African entities, but ViT is found least effective. It again verifies our assertion that not only demographic entities but also the configuration of classifiers lead to a major bias. Under cross gender setting, the detection of male deepfakes in American ('Amr') yields the highest accuracy; whereas, the female deepfake of Asian ('Asi') ethnicity yields the best accuracy. The results corresponding to seen and unseen gender of the same ethnicity are reported in Fig. 3.

The fairness issue becomes more prevalent when the ethnicity variable changes or both ethnicity and gender variables are kept different in the testing images. The results of these settings are reported in Table IV. Interestingly, in a few scenarios under cross-gender experiments, if cross-ethnicity is used for evaluation, the detection accuracy is found better compared to the same ethnicity setting. For example, when the ViT-O is trained on Africans and tested on Asians, the performance of males under cross-gender (57.78%) is found better than the same ethnicity (55.88%) cross-gender setting. Regarding females, this observation can be seen when European images are used for training and American images are used for evaluation. Otherwise, in the majority of the cases, the performance in cross-ethnicity and cross-gender is found lower than in the same ethnicity and cross-gender, i.e., where at least one factor is unchanged.

C. WLMP Analysis

In contrast to previous analyses which come from deep learning and computationally heavy architectures, in this section, a comprehensive analysis of the hand-crafted image feature algorithm is presented. The experiments are performed under the cross-ethnicity scenario where the gender variable is kept fixed. The results of deepfake detection using WLMP are reported in Fig. 4. The WLMP being the shallow classifier is found less effective in handling the different ethnicities. The performance achieved by the WLMP algorithm is close to 70% on the majority of the ethnicities in the seen gender setting except for Asian ethnicity. However, a similar trend of vulnerability is observed under cross-gender settings. Interestingly Caucasian ethnicities presented a diverse phenomenon. For example, when European males are used for training and tested on American males, the accuracy is 3.7% better than testing on European males. A similar boost in detection accuracy on females can be seen when American ethnicity is used for training and testing has been performed on Europeans.

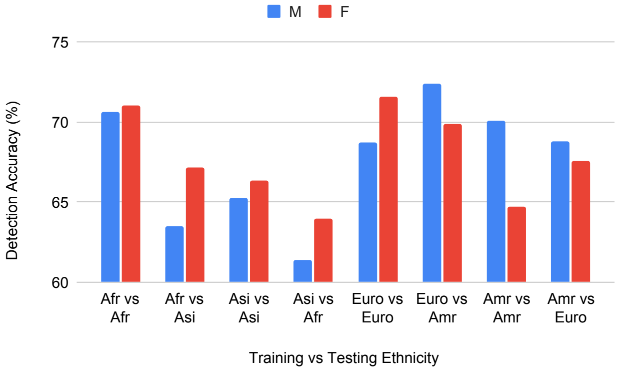

Fig. 4. Deepfake detection performance (%) of WLMP under seen and unseen ethnicity scenarios. The experiments are performed by keeping the gender variable fixed (seen) in training and testing images.

While the capacity (accuracy) of the WLMP detector is low; it is found more generalizable as compared to computationally heavy deep architectures including ViTs, INet, DNet, MNet, and a few places with XNet as well. For example, when African images are used for training and evaluation has been done on Asians, WLMP yields 7.88% and 9.07% better accuracy than ViT-O on male and female images, respectively. Therefore, we assert that if the deployed platform is sure that the testing images are going to come from unseen distribution, the use of WLMP might be preferred as compared to ViT and pure CNNs. *For example, if the detectors are trained on synthetic images or the ethnicity variable is known, WLMP might be a better choice to detect deepfakes*. It further demands the rethinking of the effective feature engineering process in a computationally efficient manner which is truly discriminating between real and deepfake images.

D. Analysis of Existing Algorithms

Similar to the observation noticed above through the experiments on multiple deepfake detectors including ViT and pure CNNs, the recent existing algorithms yield a high bias towards different ethnicities. For example, MCX-API [49] which focuses more on facial landmarks such as the eye and mouth is found highly effective in handling European ethnicity but yields poor performance/generalizability on Asian and African ethnicities even when these ethnicities are seen at the time of training. While the ID-unaware algorithm [18] shows a lower amount of biasness under a single unseen variable setting, it is found ineffective when ethnicity and gender variables are kept unseen at the time of testing.

TABLE V

Analyzing the impact of individual ethnicity and gender through cross dataset testing. Both ViTs, best-performing pure CNN architectures (VGG and XNet), and WLMP are trained on an individual subset of the FakeAVCeleb dataset and tested on the C23 subset of the FaceForensics++ (FF++) dataset.

Ethnicity	Gender	ViT-O	ViT-A	VGG	XNet	WLMP
Asian	Male	48.86	50.07	**65.21**	61.46	54.50
	Female	52.32	49.25	**64.14**	63.50	51.11
African	Male	52.21	50.61	60.21	**65.75**	57.93
	Female	49.89	49.07	60.46	**64.25**	60.25
American	Male	51.64	51.21	59.00	**67.18**	50.00
	Female	52.14	52.24	**68.25**	59.75	56.04
European	Male	51.14	50.64	**67.25**	63.04	50.25
	Female	52.68	52.64	**64.61**	52.43	54.78
Average		51.36	50.71	**63.64**	62.17	54.35

E. Effect of Ethnicity and Gender under Cross Dataset Deepfake Detection

In the final analysis, we have studied the impact of individual ethnicity and gender present in the FakeAVCeleb dataset [26] when an entirely unseen dataset comes for testing, i.e., FaceForensics++ [42]. In this, we have evaluated the ViTs, pure CNNs, and WLMP for fairness study and understand which ethnicity, gender, and classifier are robust in handling this extreme distribution shift in the testing dataset. Irrespective of ethnicity and gender, the detection accuracy of ViTs is found close to 50% even sometimes lower than that as well. It shows that the ViTs are not robust in handling such unseen variations. The prime reason might be that ViT requires a large amount of data for pre-training to achieve state-of-the-art results; however, due to a lack of datasets and limited computation we have used original [19] and optimized version [28] trained on the training subset of the dataset only. As shown in Table V, except for Caucasian males, the WLMP outperformed the ViT by a significant margin and demonstrated better robustness. The pure CNNs namely VGG and Xception show approximately similar performance even though the VGG shows significantly better performance under different settings on the FakeAVCeleb dataset than XNet.

F. Fusion of Deepfake Detectors: A Proposed Potential Solution for a Fair Deepfake Detection

As shown earlier the WLMP shows lower performance in seen distribution training-testing; however, found more generalized than ViTs and performs comparatively to the pure CNNs. Henceforth, the one quick solution to achieve the generalized deepfake detectors is to utilize the strength of WLMP and combine it with the best-performing CNN, i.e., VGG. Recently, Agarwal et al. [6] demonstrate whether simple fusion of multiple CNNs can boost deepfake detection performance. Inspired by this and to combine complementary decisions of classifiers, we have utilized the WLMP and VGG trained on African and Asian ethnicity and fused them at the decision level to test on the FF++ dataset [42]. The prime reason for the selection of these two ethnicities is to avoid an ethnicity bias issue as the FF++ dataset is highly dominated by the Caucasian ethnicity. The decision probabilities of these classifiers are fused with equal weight. The results reported in Table VI show promising improvement in the generalization accuracy of the fused model. Due to being complementary, the fusion of WLMP and CNN (VGG) shows higher improvement as compared to a fusion of the same type of models (VGG and XNet).

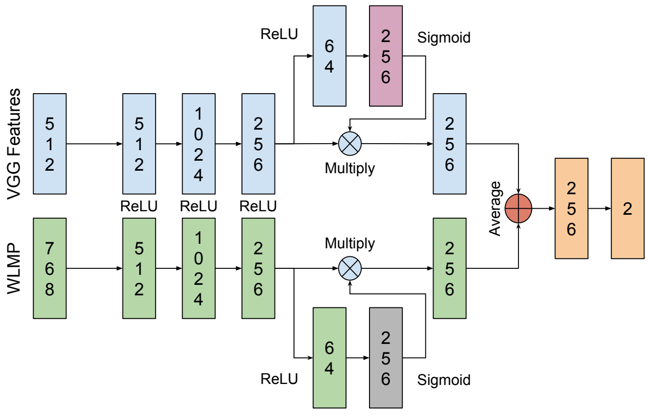

Fig. 5. Proposed attention guided generalized deepfake detection network by combining CNN and WLMP (handcrafted feature engineering).

Thrilled from the improvement in the deepfake detection accuracy in cross-dataset settings, "*we have developed novel attention-guided deep neural network architecture*". The proposed architecture consists of two branches containing features from VGG and WLMP as input. Inspired from the squeeze and excitation block which excite the convolution map along the channel dimension [23]. In this architecture, we have added attention by squeezing the dimension of the feature layer and enlarging it to excite the feature by multiplying the previous layer features. Once both the features are excited in this way, the average features are computed and are passed to the classification layer consisting of two neurons. The complete architecture is shown in Fig. 5. When the architecture is trained on the female images of Asian (South) ethnicity, the accuracy on the FF++ images improved to $\mathbf{69.96}\%$ from $\mathbf{64.14}\%$ of the best-performing feature component (VGG). The approach shows an improvement of 2.10% from the decision fusion of the VGG and WLMP. Further, if American females are used for training, the detection accuracy can improve again and the best detection accuracy of $\mathbf{71.46}\%$ can also be achieved. We want to highlight that while this is preliminary work; however, *it is the first-ever work utilizing an attention mechanism to fuse deep learning and non-learning features* to build a robust and fair deepfake detector.

G. Relevance in Current Era of Threat and ML

Fairness and bias of machine learning are not limited to any particular community and have significant concern among multiple communities whether it is biometrics or computer vision. On top of that, the image classifiers used in

TABLE VI

RESULTS OF THE FUSION OF DEEPFAKE DETECTORS TO ADVANCE GENERALIZED DEEPFAKE DETECTION UNDER CROSS-DATASET SETTING. V+X, X+W, AND V+W REPRESENT THE COMBINATION OF VGG+XNET, XNET+WLMP, AND VGG+WLMP, RESPECTIVELY. THE RESULTS ARE REPORTED ON FF++ WHERE THE NETWORKS ARE TRAINED ON THE ASIAN SUBSET OF FAKEAVCELEB.

Ethnicity	Gender	ViT	VGG	XNet	WLMP	V+X	V+W	X+W
Asian	Male	48.86	65.21	61.46	54.50	65.36	**66.18**	65.61
	Female	52.32	64.14	63.50	51.11	65.68	**67.86**	56.93
Average		50.59	64.67	62.48	52.80	65.52	**<u>67.02</u>**	61.27

this research are the benchmark architectures, and henceforth awareness of their sensitivity against any particular demographic can have a huge impact on the future development of an algorithm. Further, the impact of deepfake is not limited to any specific demographic and is a global concern. Therefore, we can assert that the proposed research can make a global impact in securing mankind in general by disseminating the knowledge that which ethnicity or gender needs more security attention.

V. CONCLUSION

Deepfake videos have created tremendous havoc, especially in the current social media era, where blind trust in digital media is hard. The impact of these videos is not limited to any gender or ethnicity; the prime reason can be seen from the fact that the Internet does not know any boundaries. These deepfake videos are heavily used for several mischievous activities such as blackmail, theft of money, harassment, and political gain. By looking at this serious impact, in the literature, several deepfake detection algorithms (machine learning) are presented. However, as we have seen machine learning algorithms reflect biased behavior when it comes to different ethnicity or gender. However, deepfake detection has not addressed this issue due to the lack of annotated large-scale multi-ethnicity and multi-gender datasets. *In this research, utilizing the recently proposed dataset, we have conducted a comprehensive fairness study of several deepfake detection algorithms*. The detection algorithms grouped into pure CNN, vision transformers, and handcrafted image features reveal the potential impact of ethnicity, gender, and classifier type on detection accuracy. It is found that few ethnicities are hard to defend while others are easy; further, the training on specific ethnicity/gender can give better generalization capability as well. We assert that the presence of such extensive analysis can help in building a fair deepfake detection architecture to protect each ethnicity and gender. We also proposed a novel attention-guided deepfake detection algorithm that outperforms several deep neural networks including vision transformers in an open-set evaluation setting. In the future, we will utilize the understanding obtained from this research to further develop a robust, fair, and trustworthy deepfake detection algorithm.

REFERENCES

[1] A. Agarwal, A. Agarwal, S. Sinha, M. Vatsa, and R. Singh. MD-CSDNetwork: Multi-domain cross stitched network for deepfake detection. In *IEEE F&G*, pages 1–8, 2021.

[2] A. Agarwal, A. Noore, M. Vatsa, and R. Singh. Generalized contact lens iris presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):373–385, 2022.

[3] A. Agarwal and N. Ratha. Manipulating faces for identity theft via morphing and deepfake: Digital privacy. In *Handbook of Statistics*, volume 48, pages 223–241. Elsevier, 2023.

[4] A. Agarwal, N. Ratha, A. Noore, R. Singh, and M. Vatsa. Misclassifications of contact lens iris pad algorithms: Is it gender bias or environmental conditions? In *IEEE/CVF WACV*, pages 961–970, 2023.

[5] A. Agarwal, N. Ratha, M. Vatsa, and R. Singh. Crafting adversarial perturbations via transformed image component swapping. *IEEE Transactions on Image Processing*, 31:7338–7349, 2022.

[6] A. Agarwal and N. K. Ratha. Deepfake catcher: Can a simple fusion be effective and outperform complex dnns? In *IEEE/CVF CVPRW*, 2024.

[7] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE IJCB*, pages 659–665, 2017.

[8] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Magnet: Detecting digital presentation attacks on face recognition. *Frontiers in AI*, 4, 2021.

[9] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74, 2021.

[10] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE F&G*, pages 59–66, 2018.

[11] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, pages 1–8. 2008.

[12] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676, 2004.

[13] R. Bogacz, S. M. McClure, J. Li, J. D. Cohen, and P. R. Montague. Short-term memory traces for action bias in human reinforcement learning. *Brain research*, 1153:111–121, 2007.

[14] S. Cahlan. How misinformation helped spark an attempted coup in gabon. *The Washington Post*, 2020.

[15] K.-W. Chang, V. Prabhakaran, and V. Ordonez. Bias and fairness in natural language processing. In *EMNLP-IJCNLP*, 2019.

[16] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE CVPR*, pages 1251–1258, 2017.

[17] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[18] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] M. Gupta, V. Singh, A. Agarwal, M. Vatsa, and R. Singh. Generalized iris presentation attack detection algorithm under cross-database settings. In *International Conference on Pattern Recognition*, pages 5318–5325, 2021.

[21] K. Hao. An ai app that "undressed" women shows how deepfakes harm the most vulnerable, 2019.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[23] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE CVPR*, pages 7132–7141, 2018.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE CVPR*, pages 4700–4708, 2017.

[25] D. Ingram. A face-swapping app takes off in china, making ai-powered deepfakes for everyone. *NBC*, 2019.

[26] H. Khalid, S. Tariq, M. Kim, and S. S. Woo. Fakeavceleb: a novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[27] P. Kwon, J. You, G. Nam, S. Park, and G. Chae. Kodf: A large-scale korean deepfake detection dataset. In *ICCV*, pages 10744–10753, 2021.

[28] S. H. Lee, S. Lee, and B. C. Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

[29] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *IEEE CVPR*, 2020.

[30] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE WIFS*, pages 1–7, 2018.

[31] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF CVPR*, pages 3207–3216, 2020.

[32] P. Majumdar, A. Agarwal, M. Vatsa, and R. Singh. Facial retouching and alteration detection. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 367–387. Springer International Publishing Cham, 2022.

[33] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE WACVW*, pages 83–92, 2019.

[34] A. Mehra, A. Agarwal, M. Vatsa, and R. Singh. Detection of digital manipulation in facial images (student abstract). In *AAAI*, 2021.

[35] A. Mehra, A. Agarwal, M. Vatsa, and R. Singh. Motion magnified 3-d residual-in-dense network for deepfake detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):39–52, 2023.

[36] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM CSUR*, 54(6):1–35, 2021.

[37] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM CSUR*, 54(1):1–41, 2021.

[38] A. V. Nadimpalli and A. Rattani. Gbdf: gender balanced deepfake dataset towards fair deepfake detection. *arXiv preprint arXiv:2207.10246*, 2022.

[39] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE TPAMI*, 2021.

[40] R. Raghavendra, K. Raja, S. Venkatesh, and C. Busch. Face morphing versus face averaging: Vulnerability and detection. In *IEEE IJCB*, pages 555–563, 2017.

[41] L. Ropek. Bank robbers in the middle east reportedly 'cloned' someone's voice to assist with $35 million heist. https://gizmodo.com/bank-robbers-in-the-middle-east-reportedly-cloned-someo-1847863805, 2021.

[42] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF ICCV*, pages 1–11, 2019.

[43] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *AAAI*, volume 34, pages 13583–13589, 2020.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, pages 2818–2826, 2016.

[45] E. Team. Give me 5 minutes, i'll give you a deepfake! https://towardsai.net/p/l/give-me-5-minutes-ill-give-you-a-deepfake, 2021.

[46] Y. Tian, Z. Zhong, V. Ordonez, G. Kaiser, and B. Ray. Testing DNN image classifiers for confusion & bias errors. In *ACM/IEEE ICSE*, pages 1122–1134, 2020.

[47] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

[48] L. Trinh and Y. Liu. An examination of fairness of ai models for deepfake detection. *arXiv preprint arXiv:2105.00558*, 2021.

[49] Y. Xu, K. Raja, L. Verdoliva, and M. Pedersen. Learning pairwise interaction for generalizable deepfake detection. In *IEEE/CVF WACV*, pages 672–682, 2023.

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

[51] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE ICASSP*, pages 8261–8265, 2019.

[52] L.-B. Zhang, F. Peng, and M. Long. Face morphing detection using fourier spectrum of sensor pattern noise. In *IEEE ICME*, pages 1–6, 2018.

[53] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. Multi-attentional deepfake detection. In *IEEE/CVF CVPR*, pages 2185–2194, 2021.

[54] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. Learning self-consistency for deepfake detection. In *IEEE/CVF ICCV*, pages 15023–15033, 2021.

[55] Y. Zhou and S.-N. Lim. Joint audio-visual deepfake detection. In *IEEE/CVF ICCV*, pages 14800–14809, 2021.