DF-Platter: Multi-Face Heterogeneous Deepfake Dataset

Kartik Narayan*, Harsh Agarwal*, Kartik Thakral*, Surbhi Mittal*, Mayank Vatsa, and Richa Singh
IIT Jodhpur, India
{narayan.2, agarwal.10, thakral.1, mittal.5, mvatsa, richa}@iitj.ac.in

Abstract

Deepfake detection is gaining significant importance in the research community. While most of the research efforts are focused towards high-quality images and videos with controlled appearance of individuals, deepfake generation algorithms now have the capability to generate deepfakes with low-resolution, occlusion, and manipulation of multiple subjects. In this research, we emulate the real-world scenario of deepfake generation and propose the DF-Platter dataset, which contains (i) both low-resolution and high-resolution deepfakes generated using multiple generation techniques and (ii) single-subject and multiple-subject deepfakes, with face images of Indian ethnicity. Faces in the dataset are annotated for various attributes such as gender, age, skin tone, and occlusion. The dataset is prepared in 116 days with continuous usage of 32 GPUs accounting to 1,800 GB cumulative memory. With over 500 GBs in size, the dataset contains a total of 133,260 videos encompassing three sets. To the best of our knowledge, this is one of the largest datasets containing vast variability and multiple challenges. We also provide benchmark results under multiple evaluation settings using popular and state-of-the-art deepfake detection models, for c0 images and videos along with c23 and c40 compression variants. The results demonstrate a significant performance reduction in the deepfake detection task on low-resolution deepfakes. Furthermore, existing techniques yield declined detection accuracy on multiple-subject deepfakes. It is our assertion that this database will improve the state-of-the-art by extending the capabilities of deepfake detection algorithms to real-world scenarios. The database is available at: http://iab-rubric.org/df-platter-database.

1. Introduction

With the advent of diverse deep learning architectures, significant breakthrough have been made in the field of image/video forgery. This has led to an incredible rise in the amount of fake multimedia content being generated due to increased accessibility and less training requirements. Not only has the amount of such media risen, but the sophistication of such content has also improved drastically, making it indistinguishable from real videos. While most deepfakes are used for entertainment purposes like parody films and filters in apps, they can also be used to illicitly defame someone, spread misinformation or propaganda, or conduct fraud. In 2020 Delhi state elections in India, a deepfake video of a popular political figure was created [34] and according to some estimates, the deepfake was disseminated to about 15 million people in the state [13]. Given the abuse of deepfakes and their possible impact, the necessity for better and robust deepfake detection methods is unavoidable.

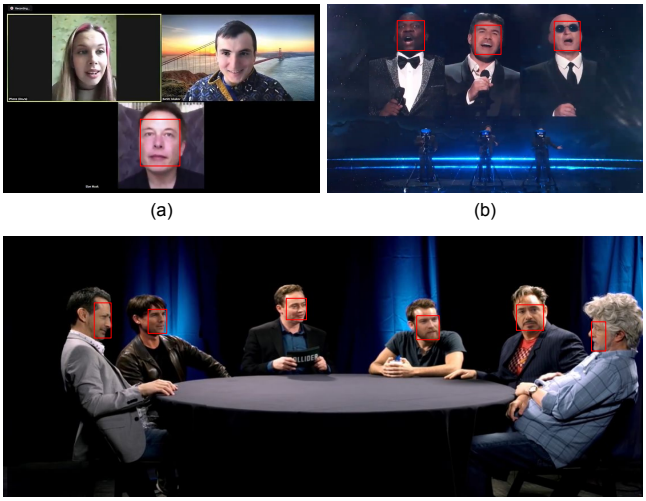

Figure 1. Samples showcasing multi-face deepfakes circulated on social media. (a) A zoom call with a deepfake of Elon Musk [8] (b) Real-time deepfake generation at America's Got Talent [9] (c) Deepfake round-table with multiple deepfake subjects [33].

Designing a dependable deepfake detection system requires availability of comprehensive deepfake datasets for training. Table 1 summarizes the key characteristics of the publicly available deepfake datasets. Most of the datasets

*Equal contribution by student authors.

Table 1. Quantitative comparison of DF-Platter with existing Deepfake datasets.

Dataset	Real Videos	Fake Videos	Total Videos	Total Subjects	Real Source	Multiple faces per image/video	Face Occlusion	Generation Techniques	Low Resolution [1]	Annotations [2]
FF++ [30]	1,000	4,000	5,000	N/A	YouTube	✗	✗	4	✗	✗
Celeb-DF [21]	590	5,639	6,229	59	YouTube	✗	✗	1	✗	✗
UADFV [36]	49	49	98	49	YouTube	✗	✗	1	✗	✗
DFDC [6]	23,654	104,500	128,154	960	Self-Recording	✗	✗	8	✗	✗
DeepfakeTIMIT [15]	640	320	960	32	VidTIMIT	✗	✗	2	✓	✗
DF-W [29]	N/A	1,869	1,869	N/A	YouTube & Bilibili	✗	✗	4	✗	✗
KoDF [16]	62,166	175,776	237,942	403	Self-Recording	✗	✗	6	✗	✗
WildDeepfake [39]	707	707	1,414	N/A	Internet	✗	✗	N/A	✗	✗
OpenForensics [17]	45,473*	70,325*	115,325*	N/A	Google Open Images	✓	✓	1	✗	✗
DeePhy [26]	100	5,040	5,140	N/A	YouTube	✗	✓	3	✗	✓
DF-Platter (ours)	**764**	**132,496**	**133,260**	**454**	**YouTube**	✓	✓	**3**	✓	✓

[1] Low resolution means the dataset contains low-resolution deepfakes generated using low-resolution videos and not by down-sampling.
[2] The dataset provides annotations such as skin tone, facial attributes and face occlusion.
* The number of images have been reported since the dataset contains only images.

contain high-resolution images with single faces in the image, while some of them contain deepfakes generated through multiple generation techniques with multiple levels of compression. In the online era, where most content is shared over the web and social media channels, the videos and images shared are of low-resolution to provide transmission efficiency. There are increasing instances of deepfake videos in unconstrained settings, for instance, occlusions on face (such as a pair of spectacles, hat, cap, turban, or hijab) and multiple faces with pose variations. While there have been several works [11, 22, 25, 37, 38] related to deepfake detection, we empirically observe that state-of-the-art detection techniques fail to detect such deepfakes. This demonstrates the need to enhance the deepfake detection technology to address such upcoming challenges.

Existing datasets traditionally comprise single-subject deepfakes generated using a single-generation technique [14,17,21,36]. However, developing a deepfake video with multiple forged subjects is also possible. Recently, developers at Collider [33] published a deepfake with multiple fake faces in a single frame. The video titled "Deepfake Roundtable" envisions a discussion involving deepfakes of 5 celebrities. A state-of-the-art model trained on the FaceForensics++ dataset is unable to identify the deepfake faces in the video. Some of these examples are shown in Figure 1. The recently published OpenForensics dataset [17] contains deepfakes with multiple faces and occlusion; however, it contains only one generation technique and does not contain any annotations for skin tone and age of subjects. Further, the OpenForensics dataset is a segmentation-based dataset while the DF-Platter dataset is a new detection dataset with multiple generation techniques and low-resolution variations.

Contributions: This research proposes a novel deepfake detection dataset titled the *DF-Platter dataset* to promote the capabilities of deepfake detection for the upcoming challenges. The dataset contains a total of 133,260 videos encompassing different sets. The subjects in the real videos are annotated for various attributes such as gender, age, skin tone, and occlusion. The video samples comprise of occluded deepfakes, low-resolution (LR) deepfakes, and multi-face (multiple-subject) deepfakes. Following are the key characteristics of the research work:

- The dataset utilizes low-resolution videos for creating deepfakes. While existing datasets synthetically interpolate deepfakes from high-resolution videos, we generate low-resolution deepfakes by utilizing low-resolution videos. This improves the visual quality of low-resolution deepfakes.
- The dataset contains multi-face (multiple-subject) deepfake sets using multiple generation techniques where each face in the video frame is annotated as real or fake. We also use three metrics for thorough evaluation on multi-face deepfakes.
- The dataset provides a gender-balanced distribution of deepfakes with subjects of Indian ethnicity and is annotated on various attributes like gender, age, skin tone, and occlusion.

2. The DF-Platter Dataset

In this work, we introduce a large-scale deepfake dataset termed as DF-Platter. This dataset contains a total of 133,260 videos having an approximate duration of 20 seconds each (estimated total of 30.67 days). It is the second largest dataset in terms of the total number of videos, only behind KoDF [16]. The dataset contains deepfake videos curated and generated at high-resolution (HR) as well as low-resolution (LR). It comprises of three sets: Set A, Set B, and Set C. Set A contains single-subject deepfakes. For generating single-subject deepfakes, there is a source video and a target video containing one subject each. The background in the target video is preserved while the face in the target video is swapped with the face in the source video.

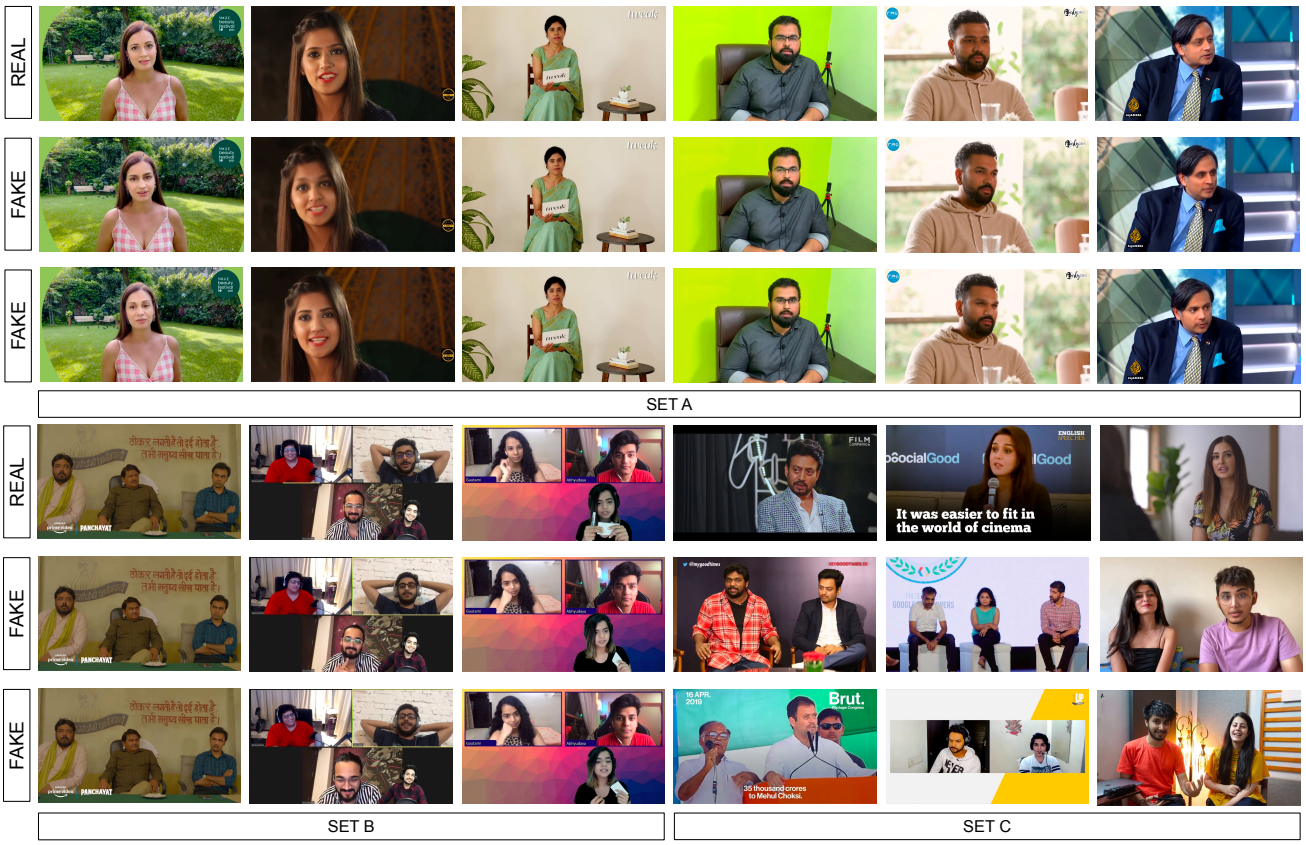

Figure 2. Samples from the DF-Platter dataset in Set A: Occluded and low-resolution deepfakes, Set B: Multi-face intra-deepfakes, and Set C: Multi-face deepfakes with celebrities as the target.

Sets B and C contain videos with multiple subjects and we create multi-face deepfakes where the face of more than one subject in the video is manipulated. Set B consists of *intra-deepfakes* where faces of one or more subjects within a particular video are swapped whereas Set C consists of *multi-face deepfakes* where faces in the videos (source) are manipulated to look like celebrity faces (target). The dataset is generated using FSGAN, FaceShifter and FaceSwap techniques to diversify the dataset in terms of the generation techniques. The dataset consists of subjects belonging to Indian ethnicity and is richly annotated in attributes such as resolution, gender, age, skin tone, and facial occlusion. While most publicly available datasets have an imbalance in terms of different attributes such as gender, skin tone, and age [24, 35], the DF-Platter dataset is balanced across resolution and gender. Further, all videos in the dataset are provided at two additional compression levels- c23 and c40.

2.1. Dataset Statistics

The dataset statistics for the DF-Platter dataset are presented in Table 2. The DF-Platter dataset comprises of 764 real videos encompassing LR as well as HR videos with 454 different subjects. Many of the existing deepfake datasets consist of source videos filmed in an extremely controlled environment with limited variation in expressions, poses, background, and illumination conditions [6, 16]. To closely mimic real-world scenarios, the videos in our dataset are collected in the wild, specifically from YouTube, with diversity in gender, orientation, skin tone (measured in Fitzpatrick scale [7]), size of face (in pixels), lighting conditions, background, and in the presence of occlusion. Occlusion occurs when hands, hair, spectacles, or any other object blocks part of the source or target face. The videos

Table 2. Summarizing the details of the DF-Platter dataset.

Sets	**Resolution**		**Compression**			**Protocol**	
	Low	High	c0	c23	c40	Train	Test
Set A	65,649	65,649	✓	✓	✓	✓	✓
Set B	500	500	✓	✓	✓	✗	✓
Set C	481	481	✓	✓	✓	✗	✓
Total	**66,630**	**66,630**	✓	✓	✓	-	-

are downloaded at 720p resolution for HR videos and at 360p resolution for LR videos. For Set A, a total of 602 real videos are used for the generation of deepfakes. These videos have nearly equal distribution in gender and resolution, i.e., 151 videos for the male gender and 150 videos for the female gender. All videos are collected for both low and high resolutions. The time duration of each video is approximately 20 seconds. For Set B, we employ 100 real videos for deepfake generation, with an equal split of low-resolution and high-resolution videos. These videos have multiple subjects within one frame of a video. Set C is generated using 62 real videos. These deepfakes are generated with celebrity faces pasted over multiple target subjects in each frame. The dataset contains all sets at three compression levels- c0, c23, and c40, where c0 encompasses the default compression of the videos when downloaded from the YouTube platform.

2.2. Dataset Generation Techniques

The deepfakes are generated using three state-of-the-art synthesis methods- FSGAN [28], FaceSwap [2] and FaceShifter [19]. Set A contains deepfakes generated using FSGAN and FaceShifter, whereas Sets B and C consist of deepfakes generated using all three generation techniques. The details of the methods employed are described below:

FSGAN [28] is capable of face-swapping as well as reenactment. It restores the missing attributes of the reenacted face and blends the whole face with the target. It captures the identity of the source subject and recreates it to suit the target subject's pose, expression, and angle. We fine-tune the FSGAN model on the real videos as suggested by the authors for producing more realistic results [3]. The FSGAN architecture is adopted because of its efficiency and good generation of occluded samples.

FaceSwap [2] is a popular, computationally expensive open-source deepfake generation software employed to swap faces across videos and images. It comprises of an encoder-decoder-based architecture with a common encoder and two different decoders for source and target faces. It is one of the generation techniques used in the FaceForensics++ [30] dataset.

FaceShifter [19] is a two-stage framework to generate face-swapped videos in an occlusion-aware manner with high fidelity. In contrast to previous face-swapping algorithms, FaceShifter tries to thoroughly blend facial features by transferring localized feature maps between facial regions. For our dataset, we utilize publicly available pretrained weights [1]. This method is adopted because of its occlusion-aware nature. The GAN-based framework of FaceShifter guarantees less training time and high realism in generated fake videos.

2.3. Dataset Organization and Description

In this section, we discuss the organization of the dataset. An equal distribution amongst gender and resolution is maintained across all sets. The dataset comprises 16,337 fake videos per generative model (FSGAN, FaceShifter), per gender (male, female), and per resolution (low-resolution, high-resolution) for Set A. Similarly, for Set B and Set C, there are 150 fake videos per model (FSGAN, FaceShifter, FaceSwap), per gender (male, female), per resolution (low-resolution, high-resolution). Manual filtering of sets is performed to ensure good-quality deepfakes. The dataset has been organized into different subsets on the basis of resolution, and compression. Notations 'HR' and 'LR' indicate high-resolution and low-resolution, respectively. 'c0', 'c23' and 'c40' represent the three levels of compression namely no compression, medium compression and hard compression, respectively. We generate LR deepfakes by using low-resolution videos instead of downsampling HR deepfakes. Samples from the dataset are shown in Figure 2. The entire dataset has been organized into three sets on the basis of distinct properties:

Set A: It consists of 130,696 single-subject deepfake videos synthesized using two-generation techniques, FSGAN and FaceShifter. The set consists of annotations for skin tone, facial occlusion, and the apparent age of each subject. We include a variety of facial occlusions such as beard, spectacles, cap/turban, hair as present in an uncontrolled environment. The set is also gender-balanced, comprising 150 female and 151 male subjects. The distribution of attributes for real subjects in Set A is shown in Figure 3.

Set B: It consists of a total of 900 intra-deepfake videos and 100 real videos. The fake videos are synthesized using the three generation techniques (FSGAN, FaceSwap and FaceShifter). In each real video, a minimum of 2 and a maximum of 5 subjects are present, out of which a minimum of 2 and a maximum of 3 faces are swapped during the generation of fake videos.

Set C: This set is similar to Set B, focusing specifically on the Indian celebrities as source faces in the deepfake videos. We utilize real videos used in Set B and swap them with single-subject celebrity faces. The set contains 62 real and 900 deepfake videos. As in Set B, the videos contain a minimum of 2 and a maximum of 5 subjects in the videos, out of which 2 to 3 faces are swapped out in each video.

Size and Format: The DF-Platter dataset is around 417 GB in its raw form. It contains a total of 133,260 videos wherein each video is approximately 20 seconds in duration. The videos are made available in MPEG4.0 format with high-resolution as well as its corresponding low-resolution. All videos have a frame rate of 25 fps. The dataset consistently contains the same videos across resolution, compression and the generation technique utilized. For compression at levels c23 and c40, H.264 video compression is utilized.

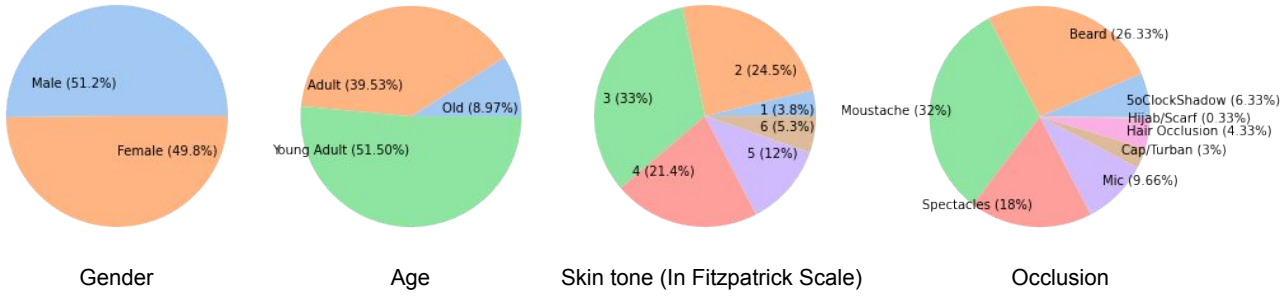

Figure 3. Distribution of real subjects of the DF-Platter dataset across multiple attributes. The dataset is gender-balanced, and the subjects are majorly young adults. The subjects are distributed in the skin tone types as per the Fitzpatrick scale [7]. The rightmost pie chart shows the occlusion types and their distribution across the dataset.

Annotation and Diversity: DF-Platter has subjects of Indian ethnicity and is annotated with attributes- gender, resolution, occlusion, and skin tone. Gender has been annotated in two classes- Male or Female. The skin tone is annotated on a scale of 1 to 6 for each subject using the Fitzpatrick scale [7]. The skin tone annotations are generated automatically by utilizing the method introduced in Groh et al. [10] and then verified by a human annotator. The apparent age attribute is classified into three classes- Young Adult (subjects with evident age between 18 to 30 years), Adult (subjects with evident age between 30 to 55 years), and Old (subjects with evident age above 55 years). 51.33% subjects are classified as "Young Adult", 42% subjects as "Adult" and 6.66% subjects as "Old". Annotations relating to facial occlusion are annotated into eight broad categories: 5'o'clock shadow, Beard, Moustache, Spectacles, Shades, Microphone, Cap/Turban/Hijab/Scarf, and Hair Occlusion. These attributes are binary in nature. Moustache and Beard are the most common type of occlusion in males with around 90% of subjects having them. Figure 3 summarizes the different types of annotations with their distribution.

2.4. Visual Quality Assessment

To evaluate the visual quality of the proposed dataset, we use the BRISQUE [23] quality metric for all (HR,c0) sets, as shown in Figure 4. On a scale of 0 (best) - 100 (worst), the average BRISQUE score for the dataset is 43.25. The set-wise BRISQUE scores for Set A (Train), Set A (Test), Set B, and Set C are 42.69, 43.80, 52.46 and 51.66 respectively. The BRISQUE scores for FaceForensics++, CelebDF, DFDC, and OpenForensics are approximated from Le et al. [17]. These scores highlight that the proposed dataset is of high quality and is, therefore, challenging with multiple covariates. We also perform a user study with 28 participants having prior experience in the computer vision domain. A total of 200 samples from the dataset (gender balanced) were randomly selected to perform the study. Each participant was asked to classify the sample as real or fake, along with their confidence level out of 5. We observed that the deepfakes in our dataset are hard for humans to detect, with an overall detection accuracy of 59.94% and an average classification confidence of 3.9.

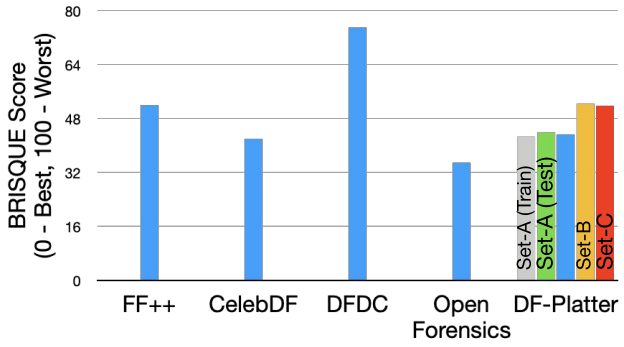

Figure 4. Comparison of BRISQUE score with existing deepfake detection datasets. The blue colored bars signify the scores for the entire dataset (Best viewed in color).

2.5. Computational Setup

The real videos for the DF-Platter dataset are collected from YouTube. The videos are generated using FaceSwap [2], FaceShifter [19], and FSGAN [28] through their publicly available GitHub repositories. For FaceSwap, each video was generated after 8 hours of training on sixteen Nvidia A100 GPUs of 80GB memory each and twelve Nvidia V100 GPUs of 32GB memory each. Similarly, deepfake videos using FaceShifter were generated using pre-trained weights with default parameters on three Nvidia RTX 3090 GPUs of 24 GB memory each. Further, for gen-

Table 3. The dataset protocol used for training, validation, and testing in the DF-Platter dataset (HR, c0). The numbers will approximately be the same for other resolutions and compressions.

Sets	Training			
	Real Frames	Fake frames	Total frames	Videos
Set A	307,221	323,120	630,341	32,824
Sets	**Testing**			
	Real Frames	Fake frames	Total frames	Videos
Set A	85,135	330,360	415,495	32,825
Set B	500	4,500	5000	500
Set C	310	4,500	4,810	481

erating deepfake videos using FSGAN, the re-enactment generator of FSGAN was fine-tuned for each source video. The inferencing was performed using twelve Nvidia V100 GPUs of 32GB memory each and an Nvidia DGX A40 GPU of 48 GB memory. The dataset generation was completed in over 116 days with parallel usage of the above-mentioned GPUs. The benchmark experiments of the dataset are performed on a Nvidia DGX station with four V100 GPUs in a multi-GPU fashion.

2.6. Implementation Details

In this section, we provide the implementation details for reproducibility of the benchmarking experiments. We use the DSFD detector [18] to extract faces from the frames of each video. For all the protocols, the models are trained for 30 epochs with early stopping and the model with the best validation accuracy is selected. Each experiment is performed by training the same model three times with different training and validation splits. The performance of the three models is averaged across the test set and reported. We use the Adam Optimizer with an initial learning rate of 0.0001. A batch size of 256 is used for distributed training.

3. Experimental Setup

In this section, we describe the designed protocol for training and testing on the DF-Platter dataset, followed by the deepfake detection algorithms and evaluation metrics used for benchmarking. The proposed dataset aims to address the following research questions:

RQ1: Can we detect occluded deepfakes?

RQ2: Can we detect multi-face deepfakes in videos?

RQ3: Can we detect low-resolution and compressed deepfakes on the web and on social media channels?

3.1. Evaluation Protocol

DF-Platter comprises of three sets. Set A in the dataset is used for training as well as evaluation whereas Sets B and C are evaluation sets only. Since the dataset comprises of videos, we conduct experiments by extracting frames from videos. We extract 10 frames from each fake video and all the frames from each real video. This is done in order to mitigate the imbalance between the number of real and fake videos in the dataset (refer Table 2). The details of the training and testing sets are summarized in Table 3.

Protocol 1 - Occluded Deepfakes: This protocol employs Set A as the test set. It is divided in a subject-disjoint manner such that there are 130 subjects in the training set, 10 subjects in the validation set, and 150 in the testing set. In terms of the number of frames, this set contains 677,980 frames in the training split, 119,320 frames in the validation split, and 840,408 frames in the testing split. There is a significant difference between the two classes - "Real" and "Fake" in the number of videos (the real:fake ratio is nearly 1:4) which leads to a skew in the dataset. To counter this, we repeat the real videos while training different architectures so as to get a nearly equal number of real and fake samples. The state-of-the-art models are then tested to detect occluded deepfakes. The results are provided in three compression settings - c0, c23, and c40 and help gauge the quality of deepfakes in the dataset with existing datasets.

Protocol 2 - Multi-Face Deepfakes: This protocol employs Set B and Set C as the test sets. Set B consists of a total of 500 videos. In terms of frames, it consists of 500 real frames and 4500 fake frames. Each frame can have one or more real (or fake) faces. A video (or frame) is considered fake if at least one face is manipulated. Similar to Set B, Set C is also an evaluation set. It consists of 481 videos, out of which 31 are real and 450 are fake. The models are tested for performance on Sets B and C, which have multiple subjects using a variety of metrics.

Protocol 3 - Cross-Resolution and Cross-Compression: We perform experiments to analyze the cross-resolution and cross-compression performance of existing deepfake detectors in real-world settings where the deepfakes are shared on the web and social media. In the cross-resolution experiment, the models are trained on (c0, HR) samples, and tested on (c0, LR) samples. The models in the cross-compression experiment are trained on (c23, HR) samples, and tested on (c40, HR) samples. In both experiments, the samples for training are taken from Set A and tested on all three sets.

3.2. DeepFake Detection Methods

We employ six state-of-the-art deepfake detection models to benchmark the dataset, for both single-subject as well as multiple-subject deepfake detection.

MesoNet [4] takes a mesoscopic approach for detecting facial forgery. Both its variants, Meso-4 and MesoInception-4 are utilized for benchmarking.

FWA [20] is a CNN-based detection technique that focuses on the artifacts stemming from affine and other transforms applied during deepfake generation.

XceptionNet [5] is based on depthwise separable convo-

Table 4. All the methods are trained and tested in the same setting of compression and resolution for this experiment. We report the Accuracy(%) and AUC for Set A *(Protocol 1)*, and FaceWA(%), FaceAUC, FLA(%) and VLA(%) for Sets B and C *(Protocol 2)*.

Trained & Tested on	Models	Set A		Set B				Set C			
		Accuracy	AUC	FaceWA	FaceAUC	FLA	VLA	FaceWA	FaceAUC	FLA	VLA
c0, HR	MesoNet [4]	84.90 ± 2.25	0.67 ± 0.08	78.46 ± 2.92	0.57 ± 0.01	58.13 ± 3.87	62.6 ± 5.60	78.57 ± 2.42	0.69 ± 0.03	57.90 ± 4.02	58.76 ± 6.96
	Meso-Inception [4]	86.62 ± 0.40	0.70 ± 0.01	79.92 ± 1.95	0.58 ± 0.00	60.89 ± 2.01	65.41 ± 1.48	79.68 ± 1.58	0.69 ± 0.01	60.81 ± 2.51	62.60 ± 2.18
	FWA [20]	82.47 ± 1.50	0.59 ± 0.04	84.83 ± 2.35	0.55 ± 0.01	71.98 ± 7.01	79.90 ± 8.34	83.71 ± 1.04	0.64 ± 0.06	66.75 ± 1.35	78.72 ± 7.79
	Xception [5]	84.76 ± 0.77	0.64 ± 0.02	86.02 ± 1.60	0.56 ± 0.01	74.07 ± 4.29	80.41 ± 5.10	86.00 ± 0.74	0.71 ± 0.03	71.36 ± 1.41	78.12 ± 4.94
	DSP-FWA [20]	91.92 ± 0.57	0.81 ± 0.01	81.59 ± 1.60	0.42 ± 0.26	62.29 ± 2.95	65.41 ± 4.11	83.08 ± 0.46	0.77 ± 0.02	65.52 ± 0.97	64.16 ± 3.26
	Capsule [27]	92.70 ± 1.92	0.83 ± 0.05	84.09 ± 2.57	0.64 ± 0.03	66.91 ± 5.48	70.96 ± 7.53	85.02 ± 2.91	0.81 ± 0.01	69.01 ± 4.75	67.04 ± 7.61
c23, HR	MesoNet [4]	85.05 ± 1.21	0.65 ± 0.03	82.21 ± 0.47	0.57 ± 0.00	64.37 ± 1.36	70.07 ± 2.37	82.35 ± 1.34	0.69 ± 0.05	64.16 ± 2.98	68.59 ± 2.41
	Meso-Inception [4]	86.49 ± 0.30	0.68 ± 0.02	83.33 ± 0.42	0.58 ± 0.01	68.92 ± 0.53	75.09 ± 0.73	82.24 ± 1.11	0.67 ± 0.02	66.14 ± 1.83	71.47 ± 1.21
	FWA [20]	83.93 ± 0.63	0.58 ± 0.02	83.49 ± 0.75	0.53 ± 0.00	68.65 ± 1.66	75.54 ± 2.02	82.83 ± 0.51	0.61 ± 0.03	65.03 ± 0.75	76.79 ± 3.26
	Xception [5]	84.22 ± 0.84	0.59 ± 0.03	86.05 ± 1.75	0.55 ± 0.01	74.01 ± 4.70	81.30 ± 6.10	84.20 ± 1.76	0.65 ± 0.02	68.72 ± 3.01	78.12 ± 5.77
	DSP-FWA [20]	87.44 ± 4.51	0.68 ± 0.13	85.27 ± 4.27	0.56 ± 0.04	72.47 ± 13.15	78.57 ± 16.05	85.73 ± 1.90	0.69 ± 0.14	70.85 ± 4.19	78.20 ± 16.60
	Capsule [27]	90.09 ± 0.83	0.74 ± 0.02	83.71 ± 2.45	0.60 ± 0.01	66.27 ± 5.83	71.10 ± 7.06	83.78 ± 1.90	0.74 ± 0.02	66.55 ± 4.40	68.07 ± 6.99
c40, HR	MesoNet [4]	82.98 ± 0.28	0.59 ± 0.01	80.20 ± 3.74	0.54 ± 0.01	61.62 ± 7.82	67.04 ± 9.38	79.74 ± 3.05	0.61 ± 0.02	58.53 ± 4.75	66.44 ± 8.18
	Meso-Inception [4]	84.21 ± 0.41	0.63 ± 0.01	81.04 ± 1.58	0.57 ± 0.02	61.92 ± 3.53	67.41 ± 4.62	79.55 ± 1.25	0.64 ± 0.01	57.49 ± 2.25	61.05 ± 3.25
	FWA [20]	82.61 ± 0.18	0.55 ± 0.01	84.80 ± 1.47	0.53 ± 0.01	71.67 ± 4.00	79.23 ± 4.94	83.59 ± 0.44	0.60 ± 0.01	65.74 ± 1.27	78.49 ± 1.73
	Xception [5]	82.64 ± 0.06	0.55 ± 0.00	87.47 ± 0.30	0.53 ± 0.01	78.65 ± 1.13	86.62 ± 1.51	85.47 ± 0.78	0.59 ± 0.02	69.27 ± 1.92	84.11 ± 2.72
	DSP-FWA [20]	85.14 ± 2.24	0.63 ± 0.05	82.93 ± 0.84	0.56 ± 0.00	66.45 ± 1.88	73.39 ± 2.39	81.89 ± 0.55	0.63 ± 0.04	62.39 ± 1.23	70.66 ± 1.86
	Capsule [27]	87.40 ± 0.18	0.68 ± 0.00	83.90 ± 0.90	0.60 ± 0.01	68.09 ± 1.35	74.20 ± 1.86	83.40 ± 0.80	0.68 ± 0.02	65.39 ± 1.39	71.25 ± 4.07

lutions and skip-connections as in ResNet [12]. It comprises depth-wise convolution followed by point-wise convolution.

DSP-FWA [20] is an improvement over the aforementioned FWA that utilizes a dual spatial pyramid strategy.

Capsule [27] is a deep neural network that is able to compensate for the information lost during pooling operations by utilizing spatial information. It is a VGG19 [32] based architecture consisting of various capsules [31] and employs a dynamic routing algorithm to calculate the agreement between extracted features.

3.3. Evaluation Metrics

We utilize the following evaluation metrics to evaluate the performance of different algorithms on different subsets of the DF-Platter dataset.

Set A: We report the frame-level accuracy (Accuracy) and ROC-AUC scores on the frame-level (AUC). Each frame is used for computation and classified as fake or real.

Set B and Set C: For Sets B and C, accuracy is computed under three settings: face-level, frame-level, and video-level. At the face-level (FaceWA), the predictions corresponding to each face are used for computation. We also report the face-level ROC-AUC score (FaceAUC). At the frame-level (FLA), the frame is considered to be correctly classified only if the predictions corresponding to all the faces in the frame are correct. At the video-level, we combine the predictions obtained for the frames of a particular video, and if more than 50% frames are classified as fake (or, real), we classify the video as fake (or, real).

4. Results and Analysis

This section summarizes the benchmark results obtained using the state-of-the-art deepfake detection models mentioned in Section 3.2, when trained and evaluated on the DF-Platter dataset. The evaluation is performed on the three protocols described in Section 3.1.

Protocol 1 - Occluded Deepfakes: The results of deepfake detection for high-resolution videos from Set A over different compressions are summarized in Table 4. At (c0, HR) setting, all architectures are seen to perform well on classical single-subject deepfake videos. However, a significant drop in performance is observed for c23 and c40 compressed videos. Capsule outperforms all other network architectures, achieving mean AUC scores of 0.83 for c0, 0.74 for c23, and 0.68 for (c40, HR) videos, followed by DSP-FWA. The AUC scores obtained across different models are in the range of 0.55 to 0.83, indicating that the existing deepfake detectors fall short of detecting deepfakes with facial occlusions. This also reflects the high quality of deepfakes in the DF-Platter dataset and a scope for improvement in deepfake detection.

Protocol 2 - Multi-Face Deepfakes: The results for the task of multi-face deepfake detection of (c0, HR) videos is shown in Table 4 block (c0, HR) for Sets B and C. In comparison to Set A, these are more challenging sets which is also confirmed by the frame-level accuracies achieved by different models on them. We also observe consistently low FLA and VLA performance by all the deepfake detectors on Set B and Set C due to the strict nature of correct classification. For Set B, the face-wise ROC-AUC score is just above 0.5 for most models which indicates a near random classification performance. We also test state-of-the-art deepfake detectors on Set C where we observe XceptionNet provides mean VLA of 78.12%, 78.12%, and 84.11% on c0, c23, and c40 videos. Table 4 block (c0, HR) for Set C shows the results for the task of multi-face deepfake detection for (c0, HR) videos. Similar to Set B, a decrease in the frame-level accuracy is observed when compared to Set A. Though, the performance of these models is comparable to Set B for (c0, HR) videos, there is a significant dip in performance for the

Table 5. All the methods are trained on (c0, HR) and tested on (c0, LR) videos *(Protocol 3: Cross-resolution)*.

Trained on	Models	Set A		Set B				Set C			
		Accuracy	AUC	FaceWA	FaceAUC	FLA	VLA	FaceWA	FaceAUC	FLA	VLA
c0, HR	MesoNet [4]	82.00 ± 0.20	0.58 ± 0.02	81.89 ± 2.34	0.53 ± 0.01	66.07 ± 4.79	71.62 ± 6.79	81.06 ± 2.91	0.58 ± 0.01	61.10 ± 5.12	73.17 ± 8.81
	Meso-Inception [4]	82.12 ± 0.46	0.57 ± 0.01	84.25 ± 1.43	0.54 ± 0.01	71.13 ± 3.11	77.61 ± 3.49	82.76 ± 0.77	0.56 ± 0.01	63.64 ± 1.11	77.68 ± 1.94
	FWA [20]	80.80 ± 0.27	0.53 ± 0.01	85.62 ± 2.38	0.51 ± 0.00	75.43 ± 5.80	84.85 ± 6.78	84.41 ± 1.48	0.55 ± 0.02	66.79 ± 2.69	85.29 ± 5.61
	Xception [5]	81.12 ± 0.46	0.53 ± 0.02	88.05 ± 1.38	0.53 ± 0.01	80.79 ± 3.97	89.95 ± 4.90	86.33 ± 1.19	0.56 ± 0.02	70.46 ± 2.44	89.65 ± 5.28
	DSP-FWA [20]	84.11 ± 1.21	0.61 ± 0.03	82.71 ± 2.24	0.54 ± 0.02	67.38 ± 4.84	72.95 ± 6.19	81.77 ± 1.61	0.58 ± 0.04	62.03 ± 3.52	73.39 ± 2.51
	Capsule [27]	85.37 ± 1.50	0.64 ± 0.04	82.79 ± 1.51	0.56 ± 0.01	67.61 ± 3.26	74.06 ± 4.11	82.08 ± 1.78	0.59 ± 0.04	63.29 ± 3.28	74.13 ± 3.24

Table 6. All the methods are trained on (c23, HR) videos and tested on (c40, HR) videos *(Protocol 3: Cross-compression)*.

Trained on	Models	Set A		Set B				Set C			
		Accuracy	AUC	FaceWA	FaceAUC	FLA	VLA	FaceWA	FaceAUC	FLA	VLA
c23, HR	MesoNet [4]	82.77 ± 0.59	0.59 ± 0.02	84.21 ± 0.48	0.55 ± 0.01	69.48 ± 1.04	75.31 ± 1.15	83.51 ± 0.94	0.64 ± 0.03	65.12 ± 2.30	75.09 ± 2.29
	Meso-Inception [4]	83.34 ± 0.18	0.61 ± 0.01	84.75 ± 0.19	0.56 ± 0.01	72.09 ± 0.73	78.49 ± 1.55	83.25 ± 0.84	0.63 ± 0.03	65.66 ± 2.03	76.42 ± 2.10
	FWA [20]	82.49 ± 0.39	0.55 ± 0.01	84.91 ± 1.46	0.52 ± 0.01	72.01 ± 3.79	80.12 ± 4.34	83.56 ± 1.36	0.58 ± 0.01	64.43 ± 2.55	80.05 ± 4.39
	Xception [5]	82.02 ± 0.33	0.54 ± 0.02	87.21 ± 1.43	0.54 ± 0.01	77.35 ± 4.19	85.81 ± 5.72	84.85 ± 1.89	0.59 ± 0.01	67.71 ± 3.68	82.85 ± 6.39
	DSP-FWA [20]	83.76 ± 1.79	0.61 ± 0.08	85.41 ± 4.44	0.54 ± 0.03	73.43 ± 12.97	79.67 ± 15.96	84.73 ± 2.33	0.63 ± 0.10	67.73 ± 5.06	78.94 ± 15.90
	Capsule [27]	85.48 ± 0.60	0.66 ± 0.02	80.06 ± 3.43	0.55 ± 0.01	60.55 ± 7.17	65.56 ± 7.83	79.41 ± 2.61	0.64 ± 0.01	58.59 ± 4.38	63.86 ± 6.41

low-resolution test set. These results are available in the supplementary file.

Protocol 3 - Cross-Resolution and Cross-Compression:
(a) Cross-Resolution: Table 5 demonstrates the performance of various architectures when trained on HR videos and tested on LR videos. It is observed that the performance of these architectures drops significantly in the LR test set. The artifacts induced by the generative models in low-resolution videos are different than those induced in high-resolution videos. The models have not seen these artifacts in HR videos during training and thus perform poorly on LR videos. For instance, Capsule shows a deterioration of 0.19 in AUC score on Set A, 0.08 in FaceAUC on Set B, and 0.22 in FaceAUC on Set C.
(b) Cross-Compression: Table 6 showcases the results for cross-compression experiments for all three sets. The detection models are trained on c23 compression and tested on c40 compression. From 6, it can be inferred that as the compression increases, the performance of the models shows a slight dip. In particular, models trained and tested on the same level of compression perform better than those tested on different compressions (refer Table 4). The FaceAUC scores for Set B remain marginally above 0.5 which denotes that the performance is close to random classification. For Set C, we observe that Capsule gives a FWA value of 83.40% when trained and tested on c40, which is ∼4% greater than when tested in a HR cross-compression setting. We also evaluate the performance of these networks on the low-resolution test sets and observe a dip in accuracy for all networks except MesoNet and FWA indicating their robustness to resolution. These results have been provided in the supplementary file.

5. Conclusion and Broader Impact

For aiding researchers in developing robust and generalized deepfake detection methods, we curate a novel large-scale DF-Platter dataset. It introduces the novel concept of intra-deepfakes, and generates low-resolution deepfakes from low-resolution videos instead of downsampling high-resolution videos. The DF-Platter dataset also contains occluded deepfakes to make the problem of deepfake detection more challenging. The dataset is balanced across gender and resolution and provides annotations for age, gender, resolution, occlusion, and skin tone. The benchmark results provided for various state-of-the-art deepfake detectors demonstrate that there is still a large scope of improvement for deepfake detection. We anticipate that this novel dataset will introduce new avenues in deepfake detection research and serve as a building block for further exploration.

Ethics Statement

The collection and generation of the DF-Platter dataset is approved by the Institutional Ethics Review Committee. The dataset will be provided only to academic institutions for research purposes. Further, the collection and generation of the proposed dataset are in accordance with the *YouTube's fair use policy*[1] since (1) we use the material in a transformative way and for non-commercial purposes, (2) we only use a small portion (∼20s) of each YouTube video in our dataset, and (3) we do not use the collected videos to cause harm to the copyright owner's ability to profit from their original work in any way.

Acknowledgement

This research is supported by a grant from the Ministry of Home Affairs, Government of India. Thakral is partially supported by the PMRF Fellowship. Mittal is partially supported by the UGC-Net JRF Fellowship and IBM fellowship. Vatsa is partially supported through the Swarnajayanti Fellowship.

[1] http://bitly.ws/CaoJ

References

[1] FaceShifter. https://github.com/richarduuz/Research_Project/tree/master/ModelC. [Accessed: 06-June-2022]. 4

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap/. [Accessed: 27-April-2022]. 4, 5

[3] FSGAN. https://github.com/YuvalNirkin/fsgan. [Accessed: 06-June-2022]. 4

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE WIFS*, pages 1–7, 2018. 6, 7, 8

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE/CVF CVPR*, pages 1251–1258, 2017. 6, 7, 8

[6] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *CoRR*, abs/2006.07397, 2020. 2, 3

[7] Thomas B Fitzpatrick. Soleil et peau. *J Med Esthet*, 2:33–34, 1975. 3, 5

[8] Fanatical Futurist. DeepFake Elon Musk bombs a Zoom call. https://www.youtube.com/watch?v=JiJKXCkWH3w. [Accessed: 2022-06-04]. 1

[9] Got Talent Global. ELVIS AND THE AMERICA'S GOT TALENT JUDGES SING?! All Auditions and Performances from Metaphysic! https://www.youtube.com/watch?v=JiJKXCkWH3w. [Accessed: 2022-06-04]. 1

[10] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *IEEE/CVF CVPR*, pages 1820–1828, 2021. 5

[11] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *IEEE/CVF CVPRw*, June 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF CVPR*, pages 770–778, 2016. 7

[13] Charlotte Jee. Deepfakes enter Indian Election Campaigns. https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/, 2020. [Accessed: 06-June-2022]. 1

[14] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *IEEE/CVF CVPR*, pages 2889–2898, 2020. 2

[15] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 2

[16] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *IEEE/CVF ICCVs*, pages 10744–10753, 2021. 2, 3

[17] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *IEEE/CVF ICCV*, pages 10117–10127, October 2021. 2, 5

[18] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *IEEE/CVF CVPR*, pages 5060–5069, 2019. 6

[19] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *CoRR*, abs/1912.13457, 2019. 4, 5

[20] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 6, 7, 8

[21] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *CoRR*, abs/1909.12962, 2019. 2

[22] Aman Mehra, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Motion magnified 3-d residual-in-dense network for deepfake detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):39–52, 2022. 2

[23] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 5

[24] Aakash Varma Nadimpalli and Ajita Rattani. GBDF: Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection. *https://arxiv.org/abs/2207.10246*, 2022. 3

[25] Kartik Narayan, Harsh Agarwal, Surbhi Mittal, Kartik Thakral, Suman Kundu, Mayank Vatsa, and Richa Singh. Desi: Deepfake source identifier for social media. In *IEEE/CVF CVPRw*, pages 2858–2867, 2022. 2

[26] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Deephy: On deepfake phylogeny. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2022. 2

[27] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019. 7, 8

[28] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *IEEE/CVF ICCV*, pages 7184–7193, 2019. 4, 5

[29] Jiameng Pu, Neal Mangaokar, Lauren Kelly, Parantapa Bhattacharya, Kavya Sundaram, Mobin Javed, Bolun Wang, and Bimal Viswanath. Deepfake videos in the wild: Analysis and detection. In *Proceedings of the Web Conference 2021*, pages 981–992, 2021. 2

[30] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF ICCV*, pages 1–11, 2019. 2, 4

[31] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017. 7

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[33] Collider Video. Deepfake Roundtable with George Lucas, Tom Cruise, Robert Downey Jr. and More. https://collider.com/deepfake-roundtable-george-lucas-tom-cruise-robert-downey-jr. [Accessed: 2022-06-04]. 1, 2

[34] John Xavier. Deepfakes enter Indian Election Campaigns. https://www.thehindu.com/news/national/deepfakes-enter-indian-election-campaigns/article61628550.ece, 2020. [Accessed: 06-June-2022]. 1

[35] Ying Xu, Philipp Terhörst, Kiran Raja, and Marius Pedersen. A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. *https://arxiv.org/abs/2208.05845*, 2022. 3

[36] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *CoRR*, abs/1811.00661, 2018. 2

[37] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *IEEE/CVF CVPR*, pages 2185–2194, June 2021. 2

[38] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *IEEE/CVF ICCV*, pages 14800–14809, October 2021. 2

[39] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 2