DF-Platter: Multi-Face Heterogeneous Deepfake Dataset

Kartik Narayan*, Harsh Agarwal*, Kartik Thakral*, Surbhi Mittal*, Mayank Vatsa, Richa Singh

IIT Jodhpur

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

Current Challenges in Deepfake Detection: Existing datasets lack low-resolution deepfakes, multi-face deepfakes, phylogenetic deepfakes, and lack of facial-attribute annotations.

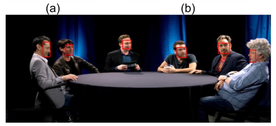
We present the DF-Platter dataset with the following key contributions-

(a) **Low-Resolution Deepfakes:**
Deepfakes are generated using low-resolution videos as they are better suited for training in real-world scenarios.

(b) **Multi-Ethnic Deepfakes:**
Existing datasets lack subjects with multiple ethnicities. Our dataset caters for subjects with Indian ethnicity.

(c) **Multi-Face Deepfakes:**
The proposed dataset provides sophisticated deepfakes with multiple manipulated faces within a single frame.

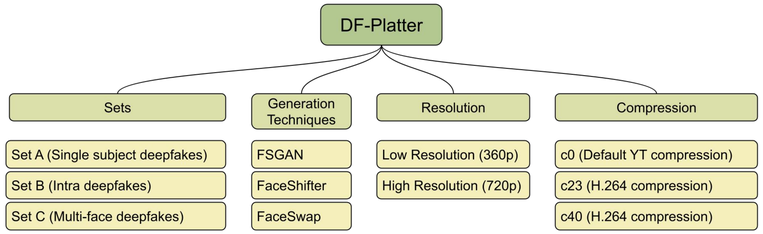
(a) (b)



(c)

Samples showcasing various challenges in deepfakes.

DF-Platter Dataset

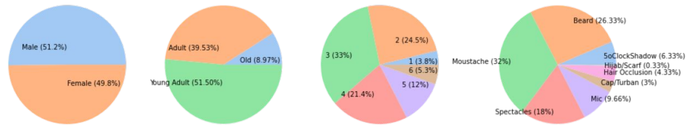


Dataset link - http://www.iab-rubric.org/df-platter-database

DF-Platter is a gender-balanced video deepfake dataset with subjects of Indian ethnicity.

	Real Videos	Fake Videos	Total Videos	Total Subjects	Real Source	Multiple faces per image/video	Face Occlusion	Generation Techniques	Resolution	Annotations
Statistics	2,292	397,488	399,780	454	YouTube	✓	✓	3	✓	✓

Video characteristics:
• 132,496 deepfake videos (c0) • 764 real videos (c0) • ~ 20 sec/video
• MPEG4.0 format • 25 FPS • H.264 compression used • ~ 400GB



1. Gender: Male, Female 2. Age: Young Adult, Adult, Old 3. Skin tone: Fitzpatrick scale (1 to 6)
4. Occlusion: Moustache, 5 o'Clock Shadow, Beard, Hair Occlusion, Spectacles, Mic, Cap/Turban, Hijab/Scarf

Dataset Details

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**	✓	✓	✓	–	–

Splits of the DF-Platter dataset

Set A: Single-face deepfakes: [Total videos - 130,696]
Consists of single-subject deepfakes. The face in target video is swapped with the face in source video. The background of target video is preserved. This set contains deepfake generated using FSGAN and FaceShifter

Set B: Multi-face deepfakes/Intra-deepfakes: [Total videos - 900]
Contains video with multiple subjects. Faces of multiple subjects (Min-2, Max-5) within a video are swapped with each other to create deepfakes. This set contains deepfake generated using FSGAN, FaceSwap, and FaceShifter.

Set C: Multi-face celebrity deepfakes/Inter-deepfakes:[Total videos - 900]
Contains video with multiple subjects. Faces of multiple subjects (Min-2, Max-5) within a video are swapped with a celebrity face to create deepfakes. This set contains deepfake generated using FSGAN, FaceSwap, and FaceShifter.

SET A

SET B SET C

Protocols and Benchmarking Results

Protocols

1. Protocol 1: Deepfake Detection
Training using Set-A, we perform deepfake detection. Set-A is used for evaluation in this protocol, achieving an average accuracy of 87.22%.
2. Protocol 2: Multi-face Deepfakes
We employ Set B & Set C as the test sets which contain deepfakes with multiple subjects. We report face-level, frame-level and video-level performance of detectors.
3. Protocol 3 : Cross-Resolution and Cross-Compression
This protocol analyzes the performance of deepfake detectors in real-world settings where the deepfake samples comprise of different compression and resolution.
   - In cross-resolution, we train on (c0,HR) and test on (c0,LR) samples.
   - In cross-compression, we train on (c23,HR) and test on (c40,HR) samples.

Evaluation Metrics

For all protocols, we report classification accuracy (Accuracy/Face-wise Accuracy/Frame-level Accuracy/Video-level Accuracy) and ROC-AUC scores (AUC/Face-wise AUC). We calculate these under three settings:

1. **Face-level (FaceWA/FaceAUC)**: Each face is used for computation and classified as fake (or, real).
2. **Frame-level (Accuracy/FLA/AUC)**: A frame is considered to be correctly classified only if the predictions corresponding to all the faces in the frame are correct.
3. **Video-level (VLA)**: If more than 50% frames are classified as fake (or real), we classify the video as fake (or real).

Trained & Tested On	Models	Set A		Set B				Set C			
		Accuracy	AUC	FaceWA	FaceAUC	FLA	VLA	FaceWA	FaceAUC	FLA	VLA
c0, HR	MesoNet	84.90 ± 2.25	0.67 ± 0.08	78.46 ± 2.92	0.57 ± 0.01	58.13 ± 3.87	62.60 ± 5.60	78.57 ± 2.42	0.69 ± 0.03	57.90 ± 4.02	58.76 ± 6.96
	Meso-Inception	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	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	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	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	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
Trained - Tested On	**Models**	**Set A**		**Set B**				**Set C**			
		Accuracy	AUC	FaceWA	FaceAUC	FLA	VLA	FaceWA	FaceAUC	FLA	VLA
c0, HR - c0, LR	MesoNet	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	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	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	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	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	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
Trained - Tested On	**Models**	**Set A**		**Set B**				**Set C**			
		Accuracy	AUC	FaceWA	FaceAUC	FLA	VLA	FaceWA	FaceAUC	FLA	VLA
c23, HR - c40, HR	MesoNet	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	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	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	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	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	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

Observations

- Deepfake detectors perform poorly on DF-platter dataset which reflects its high quality with an average BRISQUE score of 43.25 (in top 3 amongst existing deepfake datasets).
- The performance of existing detectors on Set B and Set C of DF-Platter dataset shows it's challenging nature and the limitation of existing detectors to handle multi-face deepfakes.
- Cross-resolution and Cross-compression results show that existing deepfake detectors struggle to handle real-world media that are in different resolutions and compressions.

Acknowledgement: *This research is supported by a grant from the Ministry of Home Affairs, Government of India.*