PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

TrueMedia.org

The Tug-of-War Between Deepfake Generation and Detection

Hannah Lee **Changyeon Lee** **Kevin Farhat** **Lin Qiu** **Steve Geluso** **Aerin Kim** **Oren Etzioni**

Summary

- Multimodal generative models are rapidly evolving, leading to a surge in generation of realistic video, audio, and images
- We present an up-to-date **survey paper**, examining the dual landscape of **deepfake video generation** and **detection**
- We discuss the evolution of **deepfake datasets**, and advocate for a proactive approach in the “tug-of-war” between deepfake creators and detectors

Common Deepfake Datasets

Dataset	Modality	Identities	Real Samples	Generated Samples	Generation Methods	Year
CNNDetect	Image	/	72,400	72,400	Multiple CNNs	2020
CIFAKE	Image	/	60,000	60,000	Diffusion	2024
FoR	Audio	/	>111,000	>87,000	Text to Speech	2019
ASVspoof (LA)	Audio	107	12,483	108,978	Text to Speech, Voice Conversion	2019
H-Voice	Audio	/	3,268	3,404	Multiple	2020
WaveFake	Audio	/	/	117,985	Multiple	2021
In-the-Wild	Audio	58	20.7 hours	17.2 hours	Multiple	2022
EmoFake	Audio	10	17,500	36,400	Multiple EVC Models	2022
SceneFake	Audio	107	19,838	64,642	Multiple	2022
DEEP-VOICE	Audio	8	62 min 22 sec	62 min 22 sec	RVC Model	2023
ADD	Audio	/	243,194	273,874	Multiple	2023
DeepfakeTIMIT	Video	32	320	640	GAN (face swap)	2018
FaceForensics++	Video	/	1,000	4,000	Multiple	2019
Celeb-DF	Video	59	590	5,639	GAN (face swap)	2019
WildDeepfake	Video	707	3,805	3,509	Multiple	2020
DFDC	Video	960	23,654	104,500	Multiple	2020
DeeperForensics-1.0	Video	100	50,000	10,000	DF-VAE (face swap)	2020
AV-Deepfake1M	Video	2,068	286,721	860,039	Multiple	2023

Deepfake Generation

Deepfake video media generation consists of **visual** and **audio** content. Common generation processes for both modalities include:

- **Face Swap** (GANs)
- **Reenactment** (Mimicking expressions)
- **Diffusion** (Image/video generation)
- **Audio-Driven Facial Animation** (lip sync)
- **Text-to-Speech Generation**
- **Voice Conversion** (modifying voice)
- **Emotion and Scene Fakes**
- **Partial Audio Fakes**

Deepfake Detection

Deepfake video detection is broken down into three main categories: (1) fake image detection for individual frames; (2) fake audio detection; and (3) video detection, which may utilize both images and audio as well as temporal data.

Images	• Visual artifact detection (blending, warping, etc.) • Detecting GAN-generated images • Detecting diffusion-generated images
Audio	• Feature extraction using STFT spectrograms or Mel-frequency Cepstral coefficients (MFCC) • CNN, RNN, and transformer-based models
Video	• Frame-by-frame analysis for inconsistencies • Identifying anomalies in physiological features • Audio-visual analysis for dissonance between modalities

Detection Competitions

- Deepfake detection competitions and challenges serve as one method to promote the development of novel, robust datasets as well as new detection technologies
- Meta’s Deepfake Detection Challenge (DFDC) [1] and the ASVspoof Challenge [2] serve as examples of large-scale challenges

Detection Challenges

- **Data scarcity and bias**: lack of comprehensive and diverse datasets results in detection models that fail to generalize
- **Evolving generation techniques**: new techniques evade specific detection mechanisms, and increased realism makes human verification more difficult

Findings and Future Directions

- A need for creation and maintenance of robust, diverse, representative, and public deepfake datasets
- Focused efforts should be on representation and consent to minimize concerns surrounding non-consensual deepfake creation and identity misuse
- Foundation model capabilities should be harnessed for deepfake detection

[1] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. The Deepfake Detection Challenge (DFDC) Dataset, 2020.
[2] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., Juvela, L., Alku, P., Peng, Y.-H., Hwang, H.-T., Tsao, Y., Wang, H.-M., Maguer, S. L., Becker, M., Henderson, F., Clark, R., Zhang, Y., Wang, Q., Jia, Y., Onuma, K., Mushika, K., Kaneda, T., Jiang, Y., Liu, L.-J., Wu, Y.-C., Huang, W.-C., Toda, T., Tanaka, K., Kameoka, H., Steiner, I., Matrouf, D., Bonastre, J.-F., Govender, A., Ronanki, S., Zhang, J.-X., and Ling, Z.-H. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, 2020c.