500 Evaluation Set Curation

Jan 15th, 2024

This curation compiles 500 test cases from various public deepfake benchmarks, enabling a comprehensive assessment of existing deepfake detection methods.

[Table 1]: Comparison to the Source Benchmarks

Source Dataset	with Audio	Included in the Dataset?						
		(Face) Low Res	(Face) High Res	Adversarial Condition	Dynamic Subjects	Multiple Faces	Method Diversity	Diffusion Generation
Celeb-DF-v2	N	❌	✔️	❌	❌	❌	❌	❌
DFDC	Y	✔️	🔸	✔️	🔸	❌	✔️	❌
Google DFD	N	❌	✔️	🔸	✔️	❌	❌	❌
FaceForensics++	N	✔️	❌	❌	❌	❌	✔️	❌
AV-DeepFake1M	Y	✔️	❌	❌	❌	❌	✔️	❌
OpenForensics	N	❌	✔️	✔️	❌	✔️	❌	❌
DF-Platter	N	✔️	✔️	❌	✔️	✔️	🔸	❌
Curated (Ours)	Both	✔️	✔️	✔️	✔️	✔️	✔️✔️[1]	✔️

🔸 : *Exist in the dataset, but not sufficient to contribute to the evaluation*

---

1. Why we curated this way.

- **(Leakage Prevention)** For each benchmark, subsets are chosen only from the '*valid*' or *'test*' splits to avoid data leakage when evaluating the deepfake detection model.

---

[1] The number of generation techniques used for each dataset is listed below. It's important to note that each dataset undergoes multiple generation stages. Therefore, even the same method (e.g., FaceShifter) might be applied differently across datasets:

Celeb-DF: 1
DFDC: 8
GoogleDFD: Unknown
FaceForensics++: 5
AV-DeepFake1M: 1
DF-Platter: 3

Our curation includes all these methods, plus three diffusion-based ones. To highlight this, I used two check marks.

- **(Filtered Quality)** Fake samples with irregular boundaries or discontinuous optical flow are excluded. *(This elevates the challenge of our internal dataset by weeding out samples with apparent flaws.)*

- **(Evaluation Diversity)** Choose core samples with **unique features** from each source dataset for a diversified evaluation.

With the above principles, curation strategy consists of repeating the following 2-steps.

[Step A] Form a Local Set

- Identify unique evaluation attributes (e.g. low resolution face, adversarial condition, multiple faces, etc.) that each source dataset contributes to the curation.
- Conduct stratified sampling across different segments of these evaluation attributes. For each attribute, select the same amount of samples from each segment. (e.g., "generation method" attribute: segment1 - *method1*, segment2 - *method2* …)

[Step B] Merge into the Global Set[2]

- Merge these individual local sets to form a comprehensive global collection.
- Discard samples with redundant or duplicate evaluation characteristics.

---

2. Subset Selection from Benchmarks[3]

[2] Top-Down vs Bottom-up
The top-down approach makes it easy to evenly extract attributes from the entire data, but a downside is that it requires metadata for an overview of the entire data.

(Top-Down Approach):
- Involved random sampling from various datasets with an emphasis on attribute balance.
- Review of datasets indicated a general lack of proper metadata, with most providing only binary real/fake labels.
- An attempt was made to label 20-30 attributes per video (e.g., Scene Consistency, Subject Number, Light Condition, Quality, Ethnicity), However, manual recording of these attributes for video data was larger-scope work. Filling the metadata for a specific candidate sample requires over 2 minutes.

(Bottom-Up Approach) - Applied in our curation:
- Selected distinct attributes from each dataset that enhanced the diversity of evaluation.
- Aimed to sample these attributes to achieve as much balance as possible, forming a unified set.
- Employed a sequential process, moving from Dataset A to B to C, and so on.
- Continuously added samples with new characteristics to the Curation Set, not yet represented in the Unified Set.
- This method resembles a Greedy approach, focusing on progressively building a comprehensive and diverse dataset.

[3] Consideration on the Attribute Balance

Each video sample contains a complex array of attributes, making it nearly impossible to perfectly balance them. This challenge mirrors that in Detection/Segmentation datasets, where balancing information across multiple classes in a single scene is impractical.

This part explains how we select certain examples from the original data sets to create our final collection. The included images are all fake examples.

(1) Celeb-DF-v2:

- Selected 98 (Real: 48, Fake: 48) samples[4]
- Interview scenes from 48 unique subjects where camera & subjects are static.
- The face parts are further refined to have high resolution and smooth boundaries.

- **<u>Unique Eval Attribute - Clean & static unique subjects with high resolution.</u>**
    - **Select 1 video per “Unique Subject Identity”**

(2) Deep Fake Detection Challenge (DFDC by Meta)

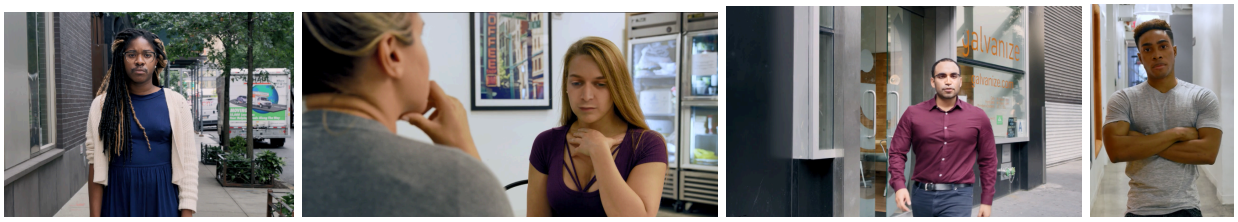
Existing Deepfake datasets are highly skewed in their consideration of attributes for curation, where even the latest datasets only focus mainly on balancing sexuality and ethnicity of the subject.
Our curated data is more balanced in attribute terms than earlier datasets but lacks metadata for quantitative comparison against existing benchmarks.

[4] 1:1 ratio of Real and Fake helps prevent any unwanted correlation between real/fake labels and dataset-specific modalities.
One might consider using fewer real and more fake samples for each dataset and evaluating them using balanced accuracy or weighted F1-scores. However, this approach is only valid under the assumption that there is no significant correlation between the real/fake label and specific dataset attributes. As the figures indicate, each dataset has very distinct visual characteristics. If this balance is not maintained, measuring balanced accuracy remains valid only for individual subsets.

- Selected 158 (Real: 79, Fake: 79) samples.
- Have diverse environments that potentially make detection harder.
    - (e.g., Light condition, Minor ethnicity, Blurred scene, Noise, Camera view…)

- **Unique Eval Attribute - Diverse Conditions.**
    - **Select samples with unusual "Environment & Situation"**
    - (Each selected sample has at least one unusual condition other than "A single subject with clean face view under indoor studio conversation situation")

(3) Google Deep Fake Detection (Google DFD)



- Selected 28 (Real: 14, Fake: 14) samples.
- Have high-resolution diverse subject actions with facial expressions.
- **Unique Eval Attribute - Dynamic Subject Actions.**
    - **For each unique subject, select 1 specific action.**
    - (e.g. Action types: "walk_down_hall_angry" , "exit_phone_room" …)

(4) FaceForensics++

- Selected 100 (Real: 50, Fake: 50) samples.
- Fake news videos generated by 5 different deepfake methods.
- **Unique Eval Attribute - Diverse Deepfake Methods**
    - **Select 10 samples per each deepfake method.**
    - Deepfake Methods: "deepfake", "face2face", "faceshifter", "faceswap", and "neuraltextures" — The above examples are from each method.

(5) AV-Deepfake1M

- Selected 20 (Real: 10, Fake: 10) samples.
- [5]Closed-up faces with *partially* modified fake videos.
    - Unlike other datasets where the entire video is real or fake, this dataset's fake videos contain specific fake segments throughout their duration.

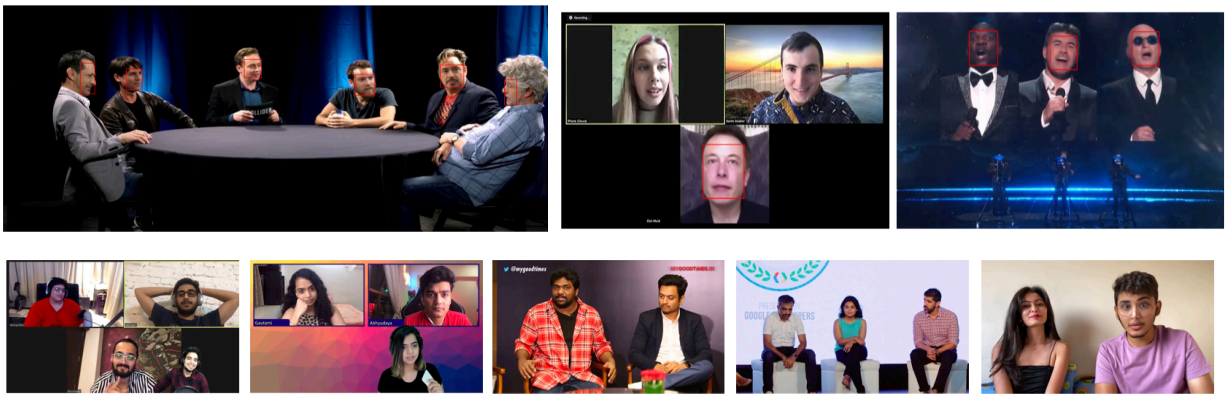
- **Unique Eval Attribute - Partial Fake Segments**
    - **Select 10 samples each with a unique subject identity.**
    - Only samples with partially fake segments are selected, as they contribute to the evaluation of partially modified fake samples.

(6) Openforensics

- Selected 60 (Real: 30, Fake: 30) samples.
- Multiple faces in a single frame scene under diverse circumstances.
    - Having a single time frame, the detection models cannot use the optical flow discrepancy between different time frames during their evaluation.
- **Unique Eval Attribute - Multiple Face & Single Time Frame**
    - **Select 30 samples with multiple faces in the same scene.**
    - The samples are selected from segments: 2-faces, 3-faces, many-faces.
    - Age & Ethnicity are balanced during sampling.

(7) DF-Platter - (*Working in Process, don't have an access yet.*)

[5] AV-Deepfake1M is created for audio-visual deepfake detection, with the primary focus on synchronizing audio and visual elements. For video manipulation, it employs only a single video manipulation method, Talk2Lip, resulting in quite trivial and similar-looking videos. A few samples should be sufficient for evaluating segments with partial fakes.



- Selected 60 (Real: 30, Fake: 30) samples.
- This dataset contains two types of multiple subject deepfake samples:
  - (Intra): Multiple subjects, face swap within the same video.
  - (Multi-Face): Multiple subjects, face swap from source to target (celebrities)
- **Unique Eval Attribute - Multiple Faces in the Same Scene**
  - **Select 30 samples each with multiple subjects (More than 2).**
  - selected 15 from “Intra” and 15 from “Multi-Face”

(7) (Custom) Diffusion-based Generation

- Selected an additional 16 fake samples.
- Generated by *Diffusion-based Methods* either from text prompt or a single image
- Unlike other *face swapping* or *face reenactment* methods, those fake samples do not have corresponding original video.
- **Unique Eval Attribute - Diffusion-based Generation**
  - **Collected 16 different unique samples from the multiple web sources.**

[Table2]: Curation Subsets Summary

Source Dataset	Contribution to Comprehensive Evaluation	File Index in Curation
Celeb-DF-v2 [1]	High-resolution clean and static faces	0-48 (49 Samples)
DFDC [2]	Diverse environmental and subject conditions	79 - 127 (79 Samples)
Google DFD [3]	Dynamic subject actions	128 - 155 (28 Samples)
FaceForensics++ [4]	Diverse deepfake synthesis methods	156 - 205 (50 Samples)
AV-Deepfake1M [5]	Partially fake samples	206 - 220 (14 Samples)
Openforensics[6]	Multiple unique subject faces in a single frame	221 - 250 (30 Samples)
DeepFlatter [6]	Multiple unique subject faces in the same scene	221 - 250 (30 Samples)
Diffusion-based [7-10]	Video diffusion methods (only for fake samples)	251 - 262 (11 Samples)

[1]. Li, Yuezun, et al. "Celeb-df: A large-scale challenging dataset for deepfake forensics." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020. [Link]

[2]. Dolhansky, Brian, et al. "The deepfake detection challenge (dfdc) dataset." *arXiv preprint arXiv:2006.07397* (2020).

[3]. Dufou Nick and Jigsaw Andrew. Contributing Data to Deepfake Detection Research, 2019.

[4].Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[5].Cai, Zhixi, et al. "AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset." *arXiv preprint arXiv:2311.15308* (2023).

[6]. Le, Trung-Nghia, et al. "Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[7]. Feng, Mengyang, et al. "DreaMoving: A Human Video Generation Framework based on Diffusion Models." *arXiv e-prints* (2023): arXiv-2312.

[8]. Zhang, Wenxuan, et al. "SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[9]. Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[10]. Stypułkowski, Michał, et al. "Diffused heads: Diffusion models beat gans on talking-face generation." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.