Celeb-DF: A New Dataset for DeepFake Forensics

Yuezun Li[1], Xin Yang[1], Pu Sun[2], Honggang Qi[2] and Siwei Lyu[1]
[1] University at Albany, State University of New York, USA
[2] University of Chinese Academy of Sciences, China

Abstract

AI-synthesized face swapping videos, commonly known as the DeepFakes, have become an emerging problem recently. Correspondingly, there is an increasing interest in developing algorithms that can detect them. However, existing dataset of DeepFake videos suffer from low visual quality and abundant artifacts that do not reflect the reality of DeepFake videos circulated on the Internet. In this work, we present a new DeepFake dataset, Celeb-DF, for the development and evaluation of DeepFake detection algorithms. The Celeb-DF dataset is generated using a refined synthesis algorithm that reduces the visual artifacts observed in existing datasets. Based on the Celeb-DF dataset, we also benchmark existing DeepFake detection algorithms.

1. Introduction

A recent twist to the disconcerting problem of online disinformation are the *DeepFakes* – falsified images or videos with human faces created by AI algorithms, most notably, deep neural networks and generative adversarial networks [8]. DeepFakes are more often referred to a specific type of synthesized videos, in which the faces of a *target* individual is replaced by the synthesized faces of *donor* that retains the target's facial expressions and head poses. While technologies to fabricate or manipulate faces are not new, the AI-powered algorithms that generate DeepFakes are making it increasingly easier for users to create fake videos of higher visual qualities. Many such tools (*e.g.*, `FakeApp`, `faceswap-GAN` [2], `faceswap` [3] and `DeepFaceLab` [1]) are publicly available as open-source software. Since faces are intrinsically associated with an individual's identity, well-crafted DeepFakes can create illusions of a person's presence and activities that do not occur in reality, which thereby can lead to serious political, social, financial and legal consequences [5].

With the escalated concerns over the DeepFakes, recently there is a surge of research efforts devoting to methods that can detect DeepFakes [4, 9, 12, 22, 15, 13, 19, 16]. The effectiveness of DeepFake detection methods are usually evaluated on large scale datasets containing real and synthesized videos generated with DeepFake tools. Evaluation results show that many of the existing DeepFake detection algorithms achieve high, sometimes near perfect, detection accuracy on these datasets. Does this mean that we have effective detection methods and do not need to worry about DeepFakes?

However, a closer look at the synthesized videos in these datasets in Figure 1 suggests that more caution to be taken when interpreting the current performance of DeepFake detection methods. Even with a casual glance, we can notice some obvious artifacts in these videos such as the low resolution, visible boundaries and color difference with the surrounding area of the synthesized face. When viewed as a video, temporal flickering is also very common. Compare these videos with those actually circulated on the Internet, the difference in visual quality is drastic. It is hard to believe that synthesized videos with such abundant visual artifacts as in these datasets will pose challenge to human viewers if they are put online. Therefore, high accuracy on these dataset may not bear relevance to the actual performance when these detection methods are deployed *in the wild*.

It is for this reason that we construct a new DeepFake dataset, Celeb-DF, for the development and evaluation of DeepFake detection algorithms. The Celeb-DF dataset is generated using a refined synthesis algorithm that reduces the visual artifacts observed in existing datasets. Based on the Celeb-DF dataset, we also benchmark existing DeepFake detection algorithms. The dataset is located at `http://www.cs.albany.edu/~lsw/celeb-deepfakeforensics.html`.

2. Previous DeepFake Datasets

To date, there are several existing DeepFake video datasets that are widely used for the purpose of training and evaluating performance of DeepFake detection methods, namely, UADFV [22], DeepFake-TIMIT [11], and the DeepFake subset in FaceForenscics++ [18][1] (FF++ / DF).

The UADFV dataset [22] consists of 98 videos, with

[1]This dataset supersedes the FaceForensics dataset [17].

49 real videos from YouTube and 49 synthesized videos, which are made using the FakeAPP implementation of the DeepFake generation algorithm. The DeepFake-TIMIT dataset [11] consists of 620 DeepFake videos of 32 subjects from the Vid-TIMIT dataset [20]). The synthesis is relatively easier as the subjects in the original videos have mostly frontal faces and the background is monocolor. In DeepFake-TIMIT, each subject has 20 DeepFake videos, where 10 videos are generated by the model with 64×64 output size (DeepFake-TIMIT-LQ) and the other 10 videos are generated by ones with 128×128 output size (DeepFake-TIMIT-HQ). The synthesized videos are generated using faceswap-GAN [2]. The FaceForensics++ [18] is a large scale dataset which contains five subsets: FaceSwap, DeepFake, DeepFakeDetection, Face2Face and NeuralTextures. In this paper we compare the DeepFake videos in DeepFake (FF++ / DF) and DeepFakeDetection (FF++ / DFD) subsets. Specifically, DF subset has $1,000$ real videos downloaded from Internet and are converted to DeepFake videos using faceswap [3]. The DFD subset is provided by Google & JigSaw [7] and hosted in FF++, which contains over $3,000$ DeepFake videos created based on original videos of 28 consented individuals.

Although these datasets provide sufficient number of synthesized videos, the synthesized videos typically exhibit low visual quality with many visible artifacts (Fig. 1). These are in stark contrast with the synthesized videos circulated on the Internet. As such, it is not convincing if a detection algorithm achieves high accuracy on these datasets, as such low quality synthesized videos can be easily distinguished from real videos and can hardly cause any real impact.

3. The Celeb-DF Dataset

To better evaluate existing DeepFake detection methods and support development of more effective detection methods, we construct the DeepFake Forensics (Celeb-DF) dataset that contains synthesized videos of better visual quality. Specifically, the Celeb-DF dataset includes 408 real videos and 795 synthesized videos generated with DeepFake. The average length of videos in the Celeb-DF dataset is 13 seconds and they all have the standard 30FPS frame rate.

3.1. Real Videos

The real videos are collected from YouTube, which are split into two sets. The first set contains 158 videos of 13 celebrities including different gender and skin color. The second set includes 250 videos, where each video contains a different subject. We name the two sets as *Celeb-real* and *YouTube-real* respectively. Several sample frames are shown in Fig. 2. The usage of each set is described in next section.

3.2. Synthesized videos

We generate our own set of synthesized videos *(Celeb-synthesis)* from the *Celeb-real* videos using a refined version of the DeepFake generation algorithm, in order to remove or reduce visual artifacts observed in previous DeepFake datasets. The details of these refinements will be presented in next section subsequently.

For each real video of a particular subject from the *Celeb-real* set, we select multiple donors of the other subjects to create synthesized videos using the synthesis model trained between this target subject and each of the donor subjects.

3.3. Train and Test Split

A subset of the Celeb-DF dataset is reserved for evaluating purpose while the remaining are for training. For testing set, we select 61 videos from DeepFake videos and corresponding 8 videos from the *Celeb-real* set as real videos. To balance the amount of real and DeepFake videos for evaluation, we select another 30 videos from *YouTube-real* set as real videos.

3.4. Refined Video Synthesis

The improved visual quality of the synthesized videos in the Celeb-DF dataset is attributed to several refinements to the original DeepFake synthesis algorithm.

- Low resolution of synthesized faces: Many of the visual artifacts such as the over-smooth skin and lack of skin details are due to the low resolution of the synthesized faces – the original DeepFake algorithm generate faces of 64×64 pixels, which need to be resized to accommodate the different sizes of the original target's faces. To improve the resolution of the synthesized face and thus reduce the necessity of resizing, we enhance the original auto-encoder structure by adding extra convolutional layers to the decoder, and our method is capable of generating synthesized faces of 256×256 pixels.

- Color inconsistency: Color mismatch between the synthesized faces and the original faces is another prominent visual artifact in the existing DeepFake video dataset. This is mostly caused by the difference in color and illumination between the target and donor's faces used in training. To alleviate this problem, we apply an augmentation procedure to the training faces, by randomly change the brightness, contrast, color distortion and sharpness of input image with a random scale from range $[0.4, 1.6]$ during training. This improves the diversity of the training data and effectively reduces color mismatch in the Celeb-DF dataset.

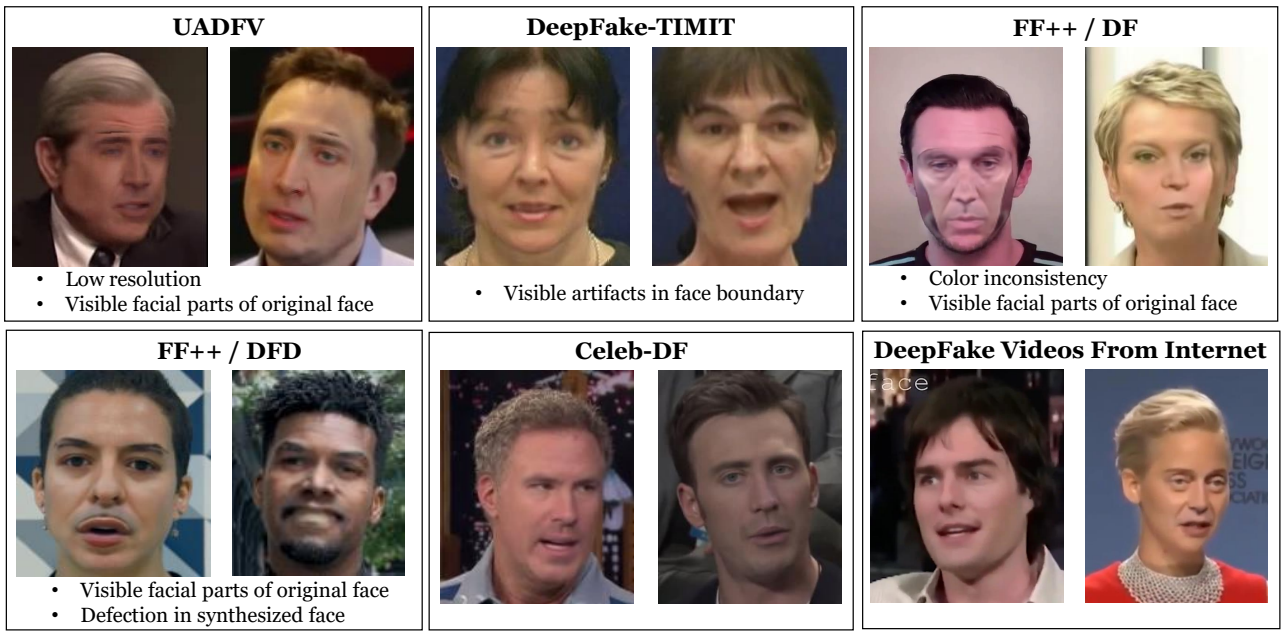

Figure 1. *Visual comparison of synthesized videos in three existing datasets,* i.e.*, UADFV [22], high quality subset of DeepFake-TIMIT [11], DeepFake subset (FF++ / DF) and DeepFakeDetection subset (FF++ / DFD) in FaceForensics++ [18], the Celeb-DF dataset and DeepFake videos recently found on the Internet. We note that videos in the existing datasets exhibit strong visual artifacts that are not present in the synthesized videos from the Internet, while videos from Celeb-DF have much improved visual quality. This figure is best viewed in color.*

- Visible facial parts of the original face: This is mostly caused by the improper choice of the mask regions where the synthesized faces are spliced into the original target's head. Previous algorithms typically use a tight mask as the bounding box of the detected face landmarks. This cannot adapt to various shapes and characteristics of the targets. Instead, we enlarge the initial mask by a factor and then interpolate to get a smoother and non-rectangular mask that better fit the target faces.

- Temporal flickering. The input target's faces in each frame of the original video have temporal consistent differences, but inaccurate face landmark locations, imprecise facial part configurations in the synthesized faces, and color mismatching artifacts introduce inconsistent temporal changes present as flickering in videos from the previous DeepFake datasets. Temporal flickering has been significantly reduced in the Celeb-DF dataset as we have improved these components.

More examples of synthesized faces in Celeb-DF dataset are shown in Fig. 2 and the video demo can be viewed here https://youtu.be/vLTiluewGQY

4. Evaluating DeepFake Detection Methods

Based on the Celeb-DF dataset, we evaluate the performance of several recent DeepFake detection methods that have code publicly available. As a comparison, we also run the same methods on the three previous DeepFake datasets: UADFV, DeepFake-TIMIT and FF++ / DF[2]. Specifically, we include the following DeepFake detection methods in our experiments.

1. *Two-stream* DNN-based forgery detection method [23], which is a general image forgery detection method that are used for DeepFake detection. We use the code provided by the authors that is based on the GoogLeNet InceptionV3 model [21] and trained on the SwapMe dataset [23].

2. *MesoNet* [4] is a CNN-based DeepFake detection method. We use the published code from (https://github.com/DariusAf/MesoNet) and evaluate two CNN architectures namely *Meso4* and *MesoInception4* with provided checkpoints, which are trained on an internal DeepFake datasets collected by the authors.

3. *HeadPose* [22] detects DeepFake videos based on the inconsistencies in the head poses of the synthesized videos. The SVM model used in this method is trained on the UADFV dataset.

[2]Since FF++ / DFD dataset was released very recently, we have not completed the benchmarking process on it. This will be added to the revision of this report.

Methods	UADFV [22]	DeepFake-TIMIT [11]		FF++ / DF [18]	Celeb-DF
		LQ	HQ		
Two-stream [23]	85.1	83.5	73.5	70.1	**55.7**
Meso4 [4]	84.3	87.8	68.4	84.7	53.6
MesoInception4	82.1	80.4	62.7	83.0	49.6
HeadPose [22]	89.0	55.1	53.2	47.3	54.8
FWA [13]	**97.4**	**99.9**	**93.2**	79.2	53.8
VA-MLP [15]	70.2	61.4	62.1	66.4	48.8
VA-LogReg	54.0	77.0	77.3	78.0	46.9
Multi-task [16]	65.8	62.2	55.3	76.3	36.5
Xception [18]	80.4	56.7	54.0	**99.7**	38.7

Table 1. *AUC (%) performance of each method on different datasets. See text for details.*

4. *Face Warping Artifacts (FWA)* [13] detects DeepFake videos by exposing the artifacts caused by the post-processing operations in DeepFake video generation. This method crafts negative training samples directly from real images using basic image processing operations. The real images are collected from Internet. The code is obtained from (https://github.com/danmohaha/CVPRW2019_Face_Artifacts), and we use the CNN model based on ResNet-50 [10] that was pre-trained and provided by the authors.

5. *Visual Artifacts (VA)* [15] detects DeepFake videos based on visual features in eyes, teeth and facial contours. We use the code obtained from (https://github.com/FalkoMatern/Exploiting-Visual-Artifacts), which includes two variants, one based on a neural network classifier as MLP and the other uses logistic regression model as LogReg. Both models are trained on internal collected dataset, where real images are cropped from CelebA dataset [14] while the DeepFake videos are from YouTube.

6. *Multi-task* [16] detects and segments the manipulated face areas using a CNN model. We use the code obtained from (https://github.com/nii-yamagishilab/ClassNSeg) with the provided checkpoint trained on FaceForensics dataset [17].

7. *Xception* [18] detects DeepFake videos by training the Xception model [6] on FaceForensics++ dataset. We used the code obtained from (https://github.com/ondyari/FaceForensics).

Many of these methods use a deep neural network but not all methods have the code for training the network. We do not re-train these models on the corresponding data set but just use the released model, with the default parameters of these methods as suggested by the corresponding authors. To increase robustness to numerical imprecision, the output of each classifier is rounded to five digit after the decimal point, *i.e.*, 10^{-5}. The performance of each method is measured using Area Under the Receiver Operating Characteristics (AUC). The value range of AUC is in $[0, 1]$, and higher value denotes better detection performance.

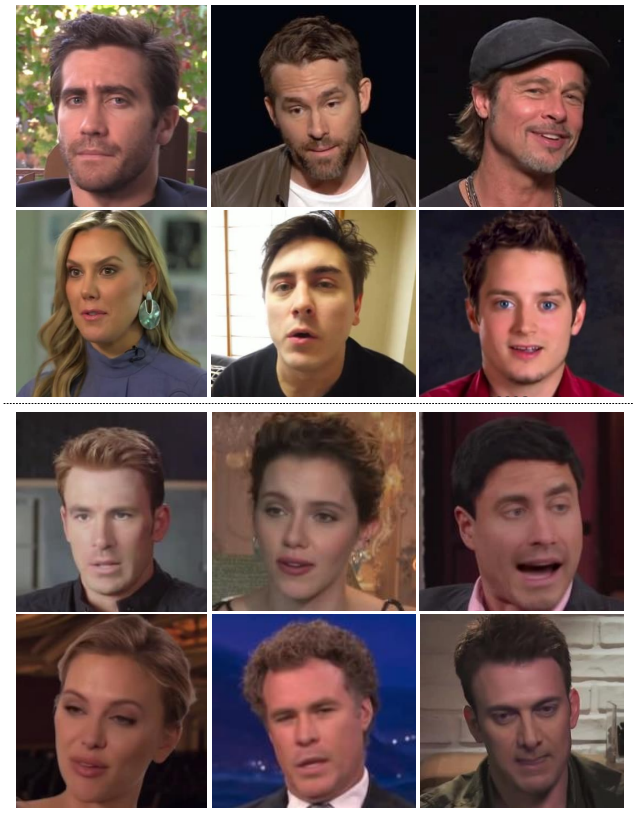

Figure 2. *Sample frames from real (top) and DeepFake (bottom) videos in Celeb-DF dataset.*

4.1. Analysis

Table 1 summarizes the performance of all compared methods on these datasets. There are a few points that we would like to highlight in these experimental results. First, note that although individual methods can achieve high accuracy in terms of AUC scores on the three previous Deep-

Fake datasets, the overall performance drops drastically on synthesized videos in the Celeb-DF dataset. This may be explained as that some detection methods detect DeepFake videos based on the artifacts such as low resolution, color mismatch and visible boundaries. When such visual artifacts are removed for the videos in the Celeb-DF dataset, the performance of these detection methods is reduced. It could also be attributed to the fact that many of these methods were not trained on the Celeb-DF dataset and thus do not generalize to the synthetic videos.

In terms of individual DeepFake detection methods, *Two-stream*, *MesoNet* and *FWA* are trained using a dataset collected by authors, which exhibit promising detection performance on all previous datasets. *VA* is also trained on a dataset collected by authors from `YouTube`. Note the *VA* method will opt-out images it can not handle, *e.g.*, 76% images are opt out. Thus the detection performance of this method is slightly higher than other methods. *HeadPose* is trained on UADFV dataset, which shows favorable performance on UADFV but is greatly degraded on other datasets. *Multi-task* is more effective on the FF++ / DF dataset compared to other datasets. *Xception* is trained on FaceForensics++ dataset, so its performance on the FF++ / DF dataset is close to perfect, yet the performance on other datasets reduces significantly. But for Celeb-DF dataset, we can observe all the methods do not perform well on it.

5. Conclusion

In this work, we present a new large scale dataset for AI-synthesized face swapping videos (commonly known as the DeepFakes), *DeepFake Forensics* (Celeb-DF) dataset, for the development and evaluation of detection algorithms. Our motivation to construct the Celeb-DF dataset is due to the low visual qualities of existing evaluation datasets, which do not reflect the synthesized videos circulating on the Internet. The Celeb-DF dataset brings closer this gap in visual quality and is expected to lead to more relevant evaluation of detection algorithms. We describe the dataset in detail and also provide benchmarking evaluations for existing detection algorithms.

As future works, we would like to extend the Celeb-DF dataset in the following aspects. A natural extension is to further enlarge the dataset. The synthesis algorithm can be improved to further reduce the visual artifacts. In particular, we plan to include synthesized videos that not only include synthetic face part but also the whole upper body. Such will be more challenging cases for detection algorithms as there will be no processing artifacts intrinsically due to the face swapping process.

References

[1] Deepfacelab github. https://github.com/iperov/DeepFaceLab, Accessed Sep 4, 2019.

[2] faceswap-gan github. https://github.com/shaoanlu/faceswap-GAN, Accessed Sep 4, 2019.

[3] faceswap github. https://github.com/deepfakes/faceswap, Accessed Sep 4, 2019.

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[5] Robert Chesney and Danielle Keats Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21*, 2018.

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[7] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google & jigsaw.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[9] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *AVSS*, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

[12] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[13] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[15] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019.

[16] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019.

[17] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

[18] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForen-

sics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

[19] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent-convolution approach to deepfake detection-state-of-art results on faceforensics++. *arXiv preprint arXiv:1905.00582*, 2019.

[20] Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Biometrics*, 2009.

[21] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[22] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[23] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839, 2017.