Celeb-DF: A Large-scale Challenging 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 DeepFakes*, is an emerging problem threatening the trustworthiness of online information. The need to develop and evaluate DeepFake detection algorithms calls for large-scale datasets. However, current DeepFake datasets suffer from low visual quality and do not resemble DeepFake videos circulated on the Internet. We present a new large-scale challenging DeepFake video dataset,* Celeb-DF*, which contains* 5,639 *high-quality DeepFake videos of celebrities generated using improved synthesis process. We conduct a comprehensive evaluation of DeepFake detection methods and datasets to demonstrate the escalated level of challenges posed by Celeb-DF.*

1. Introduction

A recent twist to the disconcerting problem of online dis-information is falsified videos created by AI technologies, in particular, deep neural networks (DNNs). Although fabrication and manipulation of digital images and videos are not new [16], the use of DNNs has made the process to create convincing fake videos increasingly easier and faster.

One particular type of DNN-based fake videos, commonly known as *DeepFakes*, has recently drawn much attention. In a DeepFake video, the faces of a *target* individual are replaced by the faces of a *donor* individual synthesized by DNN models, retaining the target's facial expressions and head poses. Since faces are intrinsically associated with identity, well-crafted DeepFakes can create illusions of a person's presence and activities that do not occur in reality, which can lead to serious political, social, financial, and legal consequences [11].

With the escalated concerns over the DeepFakes, there is a surge of interest in developing DeepFakes detection methods recently [6, 17, 27, 53, 33, 28, 41, 40, 35, 34, 36], with an upcoming dedicated global *DeepFake Detection Challenge*[1]. The availability of large-scale datasets of DeepFake

videos is an enabling factor in the development of DeepFake detection method. To date, we have the UADFV dataset [53], the DeepFake-TIMIT dataset (DF-TIMIT) [25], the FaceForenscics++ dataset (FF-DF) [40][2], the Google Deep-Fake detection dataset (DFD) [15], and the FaceBook Deep-Fake detection challenge (DFDC) dataset [14].

However, a closer look at the DeepFake videos in existing datasets reveals stark contrasts in visual quality to the actual DeepFake videos circulated on the Internet. Several common visual artifacts that can be found in these datasets are highlighted in Fig.1, including low-quality synthesized faces, visible splicing boundaries, color mismatch, visible parts of the original face, and inconsistent synthesized face orientations. These artifacts are likely the result of imperfect steps of the synthesis method and the lack of curating of the synthesized videos before included in the datasets. Moreover, DeepFake videos with such low visual qualities can hardly be convincing, and are unlikely to have real impact. Correspondingly, high detection performance on these dataset may not bear strong relevance when the detection methods are deployed *in the wild*.

In this work, we present a new large-scale and challenging DeepFake video dataset, *Celeb-DF*[3], for the development and evaluation of DeepFake detection algorithms. There are in total 5,639 DeepFake videos, corresponding more than 2 million frames, in the Celeb-DF dataset. The real source videos are based on publicly available `YouTube` video clips of 59 celebrities of diverse genders, ages, and ethic groups. The DeepFake videos are generated using an improved DeepFake synthesis method. As a result, the overall visual quality of the synthesized DeepFake videos in Celeb-DF is greatly improved when compared to existing datasets, with significantly fewer notable visual artifacts, see Fig.2. Based on the Celeb-DF dataset and other existing datasets, we conduct an evaluation of current Deep-Fake detection methods. This is the most comprehensive performance evaluation of DeepFake detection methods to

---

Figure 1. *Visual artifacts of DeepFake videos in existing datasets. Note some common types of visual artifacts in these video frames, including low-quality synthesized faces (row 1 col 1, row 3 col 2, row 5 col 3), visible splicing boundaries (row 3 col 1, row 4 col 2, row 5 col 2), color mismatch (row 5 col 1), visible parts of the original face (row 1 col 1, row 2 col 1, row 4 col 3), and inconsistent synthesized face orientations (row 3 col 3). This figure is best viewed in color.*

date. The results show that Celeb-DF is challenging to most of the existing detection methods, even though many Deep-Fake detection methods are shown to achieve high, sometimes near perfect, accuracy on previous datasets.

2. Backgrounds

2.1. DeepFake Video Generation

Although in recent years there have been many sophisticated algorithms for generating realistic synthetic face videos [9, 13, 46, 51, 26, 47, 37, 20, 23, 10, 21, 50], most of these have not been in mainstream as open-source software tools that anyone can use. It is a much simpler method based on the work of neural image style transfer [29] that becomes the *tool of choice* to create DeepFake videos in scale, with several independent open-source implementations, *e.g.*, FakeApp [5], DFaker [2], faceswap-GAN [3], faceswap [4], and DeepFaceLab [1]. We refer to this method as the *basic DeepFake maker*, and it is underneath many DeepFake videos circulated on the Internet or in the existing datasets.

The overall pipeline of the basic DeepFake maker is shown in Fig.3 (left). From an input video, faces of the target are detected, from which facial landmarks are further extracted. The landmarks are used to align the faces to a standard configuration [22]. The aligned faces are then cropped and fed to an auto-encoder [24] to synthesize faces of the donor with the same facial expressions as the original target's faces.

The auto-encoder is usually formed by two convolutional neural networks (CNNs), *i.e.*, the *encoder* and the *decoder*. The encoder E converts the input target's face to a vector known as the *code*. To ensure the encoder capture identity-independent attributes such as facial expressions, there is one single encoder regardless the identities of the subjects. On the other hand, each identity has a dedicated decoder D_i, which generates a face of the corresponding subject from the code. The encoder and decoder are trained in tandem using uncorresponded face sets of multiple subjects in an unsupervised manner, Fig.3 (right). Specifically, an encoder-decoder pair is formed alternatively using E and D_i for input face of each subject, and optimize their parameters to minimize the reconstruction errors (ℓ_1 difference between the input and reconstructed faces). The parameter update is performed with the back-propagation until convergence.

The synthesized faces are then warped back to the configuration of the original target's faces and trimmed with a *mask* from the facial landmarks. The last step involves smoothing the boundaries between the synthesized regions and the original video frames. The whole process is automatic and runs with little manual intervention.

2.2. DeepFake Detection Methods

Since DeepFakes become a global phenomenon, there has been an increasing interest in DeepFake detection methods. Most of the current DeepFake detection methods use data-driven deep neural networks (DNNs) as backbone.

Figure 2. *Example frames from the Celeb-DF dataset. Left column is the frame of real videos and right five columns are corresponding DeepFake frames generated using different donor subject.*
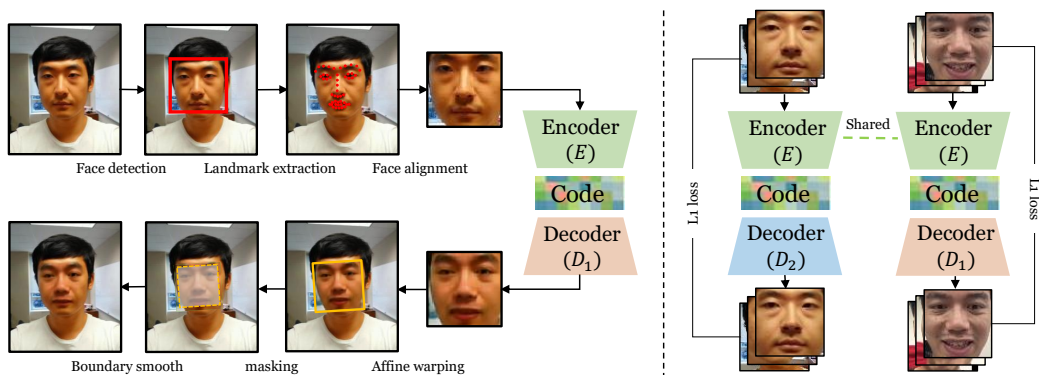

Figure 3. *Synthesis (left) and training (right) of the basic DeepFake maker algorithm. See texts for more details.*

Since synthesized faces are spliced into the original video frames, state-of-the-art DNN splicing detection methods, *e.g.*, [54, 55, 30, 8], can be applied. There have also been algorithms dedicated to the detection of Deep-Fake videos that fall into three categories. Methods in the first category are based on inconsistencies exhibited in the **physical/physiological** aspects in the DeepFake videos. The method in work of [27] exploits the observation that many DeepFake videos lack reasonable eye blinking due to the use of online portraits as training data, which usually do not have closed eyes for aesthetic reasons. Incoherent head poses in DeepFake videos are utilized in [53] to

Dataset	# Real		# DeepFake		Release Date
	Video	Frame	Video	Frame	
UADFV	49	17.3k	49	17.3k	2018.11
DF-TIMIT-LQ	320*	34.0k	320	34.0k	2018.12
DF-TIMIT-HQ			320	34.0k	
FF-DF	1,000	509.9k	1,000	509.9k	2019.01
DFD	363	315.4k	3,068	2,242.7k	2019.09
DFDC	1,131	488.4k	4,113	1,783.3k	2019.10
Celeb-DF	590	225.4k	**5,639**	2,116.8k	2019.11

Table 1. *Basic information of various DeepFake video datasets.* ∗: *the original videos in DF-TIMIT are from Vid-TIMIT dataset.*

expose DeepFake videos. In [7], the idiosyncratic behavioral patterns of a particular individual are captured by the time series of facial landmarks extracted from real videos are used to spot DeepFake videos. The second category of DeepFake detection algorithms (*e.g.*, [33, 28]) use **signal-level** artifacts introduced during the synthesis process such as those described in the Introduction. The third category of DeepFake detection methods (*e.g.*, [6, 17, 35, 36]) are **data-driven**, which directly employ various types of DNNs trained on real and DeepFake videos, not relying on any specific artifact.

2.3. Existing DeepFake Datasets

DeepFake detection methods require training data and need to be evaluated. As such, there is an increasing need for large-scale DeepFake video datasets. Table 1 lists the current DeepFake datasets.

UADFV: The UADFV dataset [53] contains 49 real `YouTube` and 49 DeepFake videos. The DeepFake videos are generated using the DNN model with `FakeAPP` [5].

DF-TIMIT: The DeepFake-TIMIT dataset [25] includes 640 DeepFake videos generated with `faceswap-GAN` [3] and based on the Vid-TIMIT dataset [43]. The videos are divided into two equal-sized subsets: DF-TIMIT-LQ and DF-TIMIT-HQ, with synthesized faces of size 64×64 and 128×128 pixels, respectively.

FF-DF: The FaceForensics++ dataset [40] includes a subset of DeepFakes videos, which has $1,000$ real `YouTube` videos and the same number of synthetic videos generated using `faceswap` [4].

DFD: The Google/Jigsaw DeepFake detection dataset [15] has $3,068$ DeepFake videos generated based on 363 original videos of 28 consented individuals of various genders, ages and ethnic groups. The details of the synthesis algorithm are not disclosed, but it is likely to be an improved implementation of the basic DeepFake maker algorithm.

DFDC: The Facebook DeepFake detection challenge dataset [14] is part of the DeepFake detection challenge, which has $4,113$ DeepFake videos created based on $1,131$ original videos of 66 consented individuals of various genders, ages and ethnic groups[4]. This dataset is created us-

ing two different synthesis algorithms, but the details of the synthesis algorithm are not disclosed.

Based on release time and synthesis algorithms, we categorize UADFV, DF-TIMIT, and FF-DF as the *first generation* of DeepFake datasets, while DFD, DFDC, and the proposed Celeb-DF datasets are the *second generation*. In general, the second generation datasets improve in both quantity and quality over the first generation.

3. The Celeb-DF Dataset

Although the current DeepFake datasets have sufficient number of videos, as discussed in the Introduction and demonstrate in Fig.1, DeepFake videos in these datasets have various visual artifacts that easily distinguish them from the real videos. To provide more relevant data to evaluate and support the future development DeepFake detection methods, we construct the Celeb-DF dataset. A comparison of the Celeb-DF dataset with other existing DeepFake datasets is summarized in Table 1.

3.1. Basic Information

The Celeb-DF dataset is comprised of 590 real videos and $5,639$ DeepFake videos (corresponding to over two million video frames). The average length of all videos is approximate 13 seconds with the standard frame rate of 30 frame-per-second. The real videos are chosen from publicly available `YouTube` videos, corresponding to interviews of 59 celebrities with a diverse distribution in their genders, ages, and ethnic groups[5]. 56.8% subjects in the real videos are male, and 43.2% are female. 8.5% are of age 60 and above, 30.5% are between 50 - 60, 26.6% are 40s, 28.0% are 30s, and 6.4% are younger than 30. 5.1% are Asians, 6.8% are African Americans and 88.1% are Caucasians. In addition, the real videos exhibit large range of changes in aspects such as the subjects' face sizes (in pixels), orientations, lighting conditions, and backgrounds. The DeepFake videos are generated by swapping faces for each pair of the 59 subjects. The final videos are in MPEG4.0 format.

3.2. Synthesis Method

The DeepFake videos in Celeb-DF are generated using an improved DeepFake synthesis algorithm, which is key to the improved visual quality as shown in Fig.2. Specifically, the basic DeepFake maker algorithm is refined in several aspects targeting the following specific visual artifacts observed in existing datasets.

Low resolution of synthesized faces: The basic DeepFake maker algorithm generate low-resolution faces (typically 64×64 or 128×128 pixels). We improve the resolution of

---

[4]The full set of DFDC has not been released at the time of CVPR submission, and information is based on the first round release in [14].

[5]We choose celebrities' faces as they are more familiar to the viewers so that any visual artifacts can be more readily identified. Furthermore, celebrities are anecdotally the main targets of DeepFake videos.

64 × 64   128 × 128   256 × 256

Figure 4. *Comparison of DeepFake frames with different sizes of the synthesized faces. Note the improved smoothness of the* 256 × 256 *synthesized face, which is used in Celeb-DF. This figure is best viewed in color.*

the synthesized face to 256 × 256 pixels. This is achieved by using encoder and decoder models with more layers and increased dimensions. We determine the structure empirically for a balance between increased training time and better synthesis result. The higher resolution of the synthesized faces are of better visual quality and less affected by resizing and rotation operations in accommodating the input target faces, Fig.4.

Color mismatch: Color mismatch between the synthesized donor's face with the original target's face in Celeb-DF is significantly reduced by training data augmentation and post processing. Specifically, in each training epoch, we randomly perturb the colors of the training faces, which forces the DNNs to synthesize an image containing the same color pattern with input image. We also apply a color transfer algorithm [38] between the synthesized donor face and the input target face. Fig.5 shows an example of synthesized face without (left) and with (right) color correction.
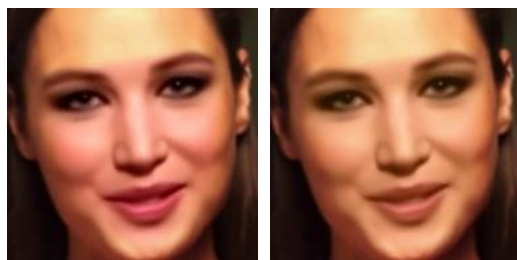

Figure 5. *DeepFake frames using synthesized face without (left) and with (right) color correction. Note the reduced color mismatch between the synthesized face region and the other part of the face. Synthesis method with color correction is used to generate Celeb-DF. This figure is best viewed in color.*

Inaccurate face masks: In previous datasets, the face masks are either rectangular, which may not completely cover the facial parts in the original video frame, or the convex hull of landmarks on eyebrow and lower lip, which
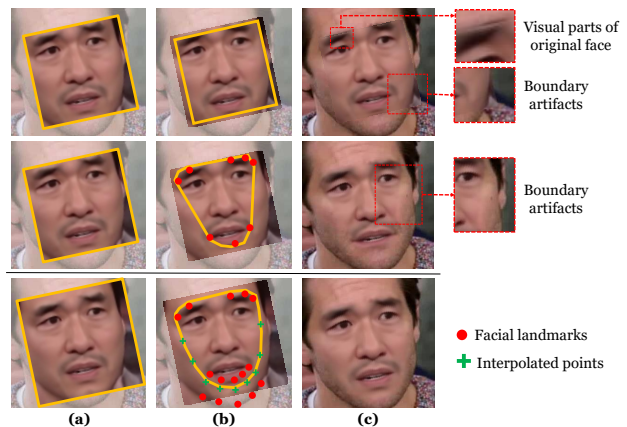

Figure 6. *Mask generation in existing datasets (Top two rows) and Celeb-DF (3rd row). (a) warped synthesized face overlaying the target's face. (b) mask generation. (c) final synthesis result.*

leaves the boundaries of the mask visible. We improve the mask generation step for Celeb-DF. We first synthesize a face with more surrounding context, so as to completely cover the original facial parts after warping. We then create a smoothness mask based on the landmarks on eyebrow and interpolated points on cheeks and between lower lip and chin. The difference in mask generation used in existing datasets and Celeb-DF is highlighted in Fig.6 with an example.

Temporal flickering: We reduce temporal flickering of synthetic faces in the DeepFake videos by incorporating temporal correlations among the detected face landmarks. Specifically, the temporal sequence of the face landmarks are filtered using a Kalman smoothing algorithm to reduce imprecise variations of landmarks in each frame.

3.3. Visual Quality

The refinements to the synthesis algorithm improve the visual qualities of the DeepFake videos in the Celeb-DF dataset, as demonstrated in Fig.2. We would like have a more quantitative evaluation of the improvement in visual quality of the DeepFake videos in Celeb-DF and compare with the previous DeepFake datasets. Ideally, a reference-free face image quality metric is the best choice for this purpose. However, unfortunately, to date there is no such metric that is agreed upon and widely adopted.

Instead, we follow the face in-painting work [45] and use the Mask-SSIM score [32] as a referenced quantitative metric of visual quality of synthesized DeepFake video frames. Mask-SSIM corresponds to the SSIM score [52] between the head regions (including face and hair) of the DeepFake video frame and the corresponding original video frame, *i.e.*, the head region of the original target is the reference for visual quality evaluation. As such, low Mask-SSIM score may be due to inferior visual quality as well as changes of the identity from the target to the donor. On the other hand,

| Datasets | UADFV | DF-TIMIT | | FF-DF | DFD | DFDC | **Celeb-DF** |
		LQ	HQ				
Mask -SSIM	0.82	0.80	0.80	0.81	0.88	0.84	**0.92**

Table 2. *Average Mask-SSIM scores of different DeepFake datasets. Computing Mask-SSIM requires exact corresponding pairs of DeepFake synthesized frames and original video frames, which is not the case for DFD and DFDC. For these two datasets, we calculate the Mask-SSIM on videos that we have exact correspondences,* i.e., 311 *videos in DFD and* 2, 025 *videos in DFDC.*

since we only compare frames from DeepFake videos, the errors caused by identity changes are biased in a similar fashion to all compared datasets. Therefore, the numerical values of Mask-SSIM may not be meaningful to evaluate the absolute visual quality of the synthesized faces, but the difference between Mask-SSIM reflects the difference in visual quality.

The Mask-SSIM score takes value in the range of $[0, 1]$ with higher value corresponding to better image quality. Table 2 shows the average Mask-SSIM scores for all compared datasets, with Celeb-DF having the highest scores. This confirms the visual observation that Celeb-DF has improved visual quality, as shown in Fig.2.

4. Evaluating DeepFake Detection Methods

Using Celeb-DF and other existing DeepFake datasets, we perform the most comprehensive performance evaluation of DeepFake detection to date, with the largest number of DeepFake detection methods and datasets considered. There are two purposes of this evaluation. First, using the average detection performance as an indicator of the challenge levels of various DeepFake datasets, we further compare Celeb-DF with existing DeepFake datasets. Furthermore, we survey the performance of the current DeepFake detection methods on a large diversity of DeepFake videos, in particular, the high-quality ones in Celeb-DF.

4.1. Compared DeepFake Detection Methods

We consider nine DeepFake detection methods in our experiments. Because of the need to run each method on the Celeb-DF dataset, we choose only those that have code and the corresponding DNN-model publicly available or obtained from the authors directly.

- **Two-stream** [54] uses a two-stream CNN to achieve state-of-the-art performance in general-purpose image forgery detection. The underlying CNN is the GoogLeNet InceptionV3 model [48] trained on the SwapMe dataset [54]. We use it as a baseline to compare other dedicated DeepFake detection methods.
- **MesoNet** [6] is a CNN-based DeepFake detection method targeting on the mesoscopic properties of images. The model is trained on unpublished DeepFake datasets collected by the authors. We evaluate two variants of MesoNet, namely, *Meso4* and *MesoIncep-*

tion4. Meso4 uses conventional convolutional layers, while MesoInception4 is based on the more sophisticated Inception modules [49].
- **HeadPose** [53] detects DeepFake videos using the inconsistencies in the head poses of the synthesized videos, based on a SVM model on estimated 3D head orientations from each video. The SVM model in this method is trained on the UADFV dataset.
- **FWA** [28] detects DeepFake videos using a ResNet-50 [19] to expose the face warping artifacts introduced by the resizing and interpolation operations in the basic DeepFake maker algorithm. This model is trained on self-collected face images.
- **VA** [33] is a recent DeepFake detection method based on capturing visual artifacts in the eyes, teeth and facial contours of the synthesized faces. There are two variants of this method: VA-MLP is based on a multilayer feedforward neural network classifier, and VA-LogReg uses a simpler logistic regression model. These models are trained on unpublished dataset, of which real images are cropped from CelebA dataset [31] and the DeepFake videos are from YouTube.
- **Xception** [40] corresponds to a DeepFake detection method based on the XceptionNet model [12] trained on the FaceForensics++ dataset. There are three variants of Xception, namely, *Xception-raw*, *Xception-c23* and *Xception-c40*: *Xception-raw* are trained on raw videos, while *Xception-c23* and *Xception-c40* are trained on H.264 videos with medium (23) and high degrees (40) of compression, respectively.
- **Multi-task** [34] is another recent DeepFake detection method that uses a CNN model to simultaneously detect manipulated images and segment manipulated areas as a multi-task learning problem. This model is trained on the FaceForensics dataset [39].
- **Capsule** [36] uses capsule structures [42] based on a VGG19 [44] network as the backbone architecture for DeepFake classification. This model is trained on the FaceForensics++ dataset.
- **DSP-FWA** is a recently further improved method based on FWA, which includes a spatial pyramid pooling (SPP) module [18] to better handle the variations in the resolutions of the original target faces. This method is trained on self-collected face images.

A concise summary of the underlying model, source code, and training datasets of the DeepFake detection methods considered in our experiments is given in Table 3.

4.2. Experimental Settings

We evaluate the overall detection performance using the area under ROC curve (AUC) score at the frame level for all key frames. There are several reasons for this choice. First, all compared methods analyze individual frames (usu-

Methods	Model Type	Training Dataset	Repositories	Release Date
Two-stream [54]	GoogLeNet InceptionV3 [48]	SwapMe [54]	Unpublished code provided by the authors	2018.03
MesoNet [6]	Designed CNN	Unpublished	https://github.com/DariusAf/MesoNet	2018.09
HeadPose [53]	SVM	UADFV [53]	https://bitbucket.org/ericyang3721/headpose_forensic/	2018.11
FWA [28]	ResNet-50 [19]	Unpublished	https://github.com/danmohaha/CVPRW2019_Face_Artifacts	2018.11
VA-MLP [33]	Designed CNN	Unpublished	https://github.com/FalkoMatern/Exploiting-Visual-Artifacts	2019.01
VA-LogReg [33]	Logistic Regression Model			
Xception [40]	XceptionNet [12]	FaceForensics++ [40]	https://github.com/ondyari/FaceForensics	2019.01
Multi-task [34]	Designed CNN	FaceForensics [39]	https://github.com/nii-yamagishilab/ClassNSeg	2019.06
Capsule [36]	Designed CapsuleNet [42]	FaceForensics++	https://github.com/nii-yamagishilab/Capsule-Forensics-v2	2019.10
DSP-FWA	SPPNet [18]	Unpublished	https://github.com/danmohaha/DSP-FWA	2019.11

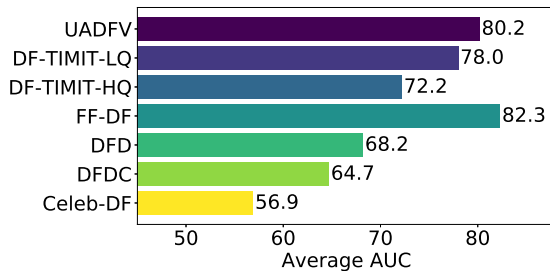Table 3. *Summary of compared DeepFake detection methods. See texts for more details.*

Figure 7. *Average AUC performance of all detection methods on each dataset.*

Figure 8. *Average AUC performance of each detection method on all evaluated datasets.*

ally key frames of a video) and output a classification score for each frame. Using frame-level AUC thus avoids differences caused by different approaches to aggregating frame-level scores for each video. Second, using frame level AUC score obviates the necessity of calibrating the classification outputs of these methods across different datasets. To increase robustness to numerical imprecision, the classification scores are rounded to five digits after the decimal point, *i.e.*, with a precision of 10^{-5}. As the videos are compressed, we perform evaluations only on the key frames.

we compare performance of each detection method using the inference code and the published pre-trained models. This is because most of these methods do not have published code for training the machine learning models. As such, we could not practically re-train these models on all datasets we considered. We use the default parameters provided with each compared detection method.

4.3. Results and Analysis

In Table 4 we list individual frame-level AUC scores of all compared DeepFake detection methods over all datasets including Celeb-DF, and Fig.9 shows the frame-level ROC curves of several top detection methods on several datasets.

Comparing different datasets, in Fig.7, we show the average frame-level AUC scores of all compared detection methods on each dataset. Celeb-DF is in general the most challenging to the current detection methods, and their overall performance on Celeb-DF is lowest across all datasets. These results are consistent with the differences in visual quality. Note many current detection methods predicate on visual artifacts such as low resolution and color mismatch, which are improved in synthesis algorithm for the Celeb-DF dataset. Furthermore, the difficulty level for detection
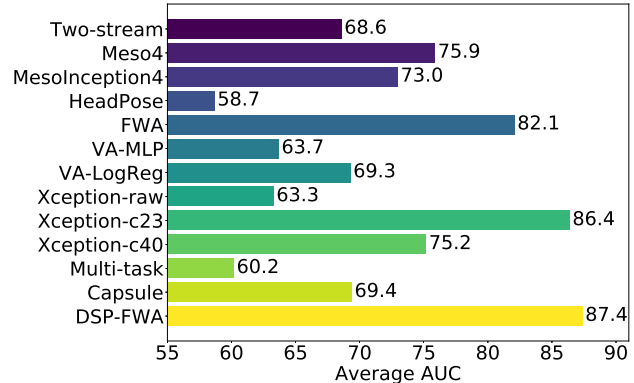
is clearly higher for the second generation datasets (DFD, DFDC, and Celeb-DF, with average AUC scores lower than 70%), while some detection methods achieve near perfect detection on the first generation datasets (UADFV, DF-TIMIT, and FF-DF, with average AUC scores around 80%).

In term of individual detection methods, Fig.8 shows the comparison of average AUC score of each detection method on all DeepFake datasets. These results show that detection has also made progress with the most recent DSP-FWA method achieves the overall top performance (87.4%).

As online videos are usually recompressed to different formats (MPEG4.0 and H264) and in different qualities during the process of uploading and redistribution, it is also important to evaluate the robustness of detection performance with regards to video compression. Table 5 shows the average frame-level AUC scores of four state-of-the-art DeepFake detection methods on original MPEG4.0 videos, and medium (23), and high (40) degrees of H.264 compressed videos of Celeb-DF, respectively. The results show that the performance of each method is reduced along with the compression degree increased. In particular, the performance of FWA and DSP-FWA degrades significantly on recompressed video, while the performance of Xception-c23 and Xception-c40 is not significantly affected. This is expected because the latter methods were trained on compressed H.264 videos such that they are more robust in this setting.

Methods↓ Datasets→	UADFV [53]	DF-TIMIT [25]		FF-DF [40]	DFD [15]	DFDC [14]	**Celeb-DF**
		LQ	HQ				
Two-stream [54]	85.1	83.5	73.5	70.1	52.8	61.4	53.8
Meso4 [6]	84.3	87.8	68.4	84.7	76.0	75.3	54.8
MesoInception4	82.1	80.4	62.7	83.0	75.9	73.2	53.6
HeadPose [53]	89.0	55.1	53.2	47.3	56.1	55.9	54.6
FWA [28]	97.4	99.9	93.2	80.1	74.3	72.7	56.9
VA-MLP [33]	70.2	61.4	62.1	66.4	69.1	61.9	55.0
VA-LogReg	54.0	77.0	77.3	78.0	77.2	66.2	55.1
Xception-raw [40]	80.4	56.7	54.0	**99.7**	53.9	49.9	48.2
Xception-c23	91.2	95.9	94.4	99.7	**85.9**	72.2	65.3
Xception-c40	83.6	75.8	70.5	95.5	65.8	69.7	**65.5**
Multi-task [34]	65.8	62.2	55.3	76.3	54.1	53.6	54.3
Capsule [36]	61.3	78.4	74.4	96.6	64.0	53.3	57.5
DSP-FWA	**97.7**	**99.9**	**99.7**	93.0	81.1	**75.5**	64.6

Table 4. *Frame-level AUC scores (%) of various methods on compared datasets. Bold faces correspond to the top performance.*
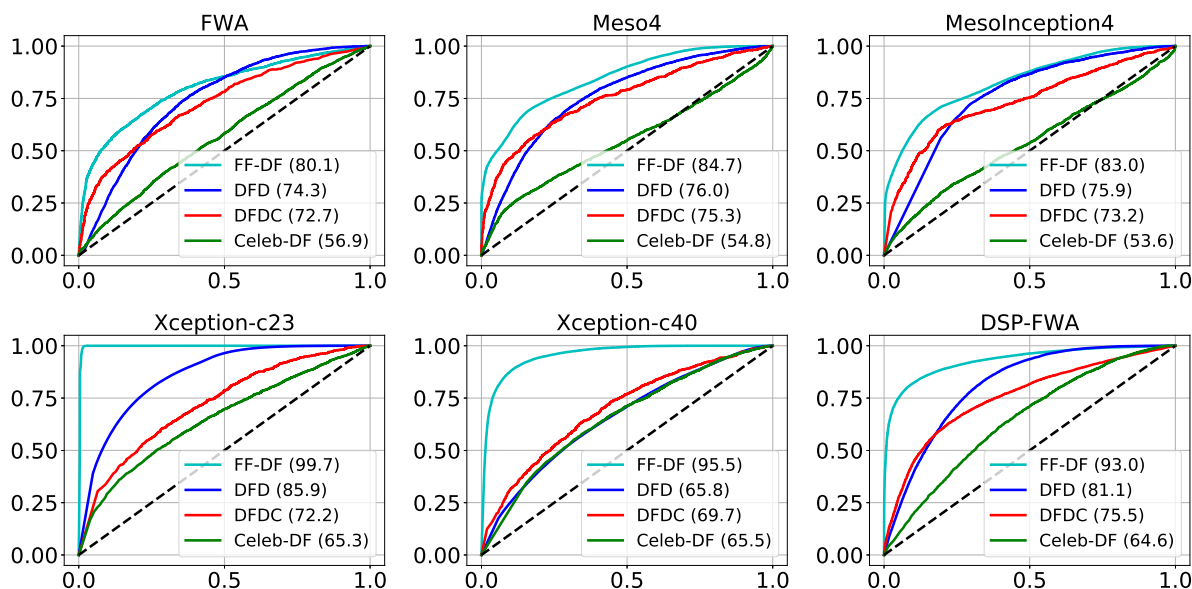

Figure 9. *ROC curves of six state-of-the-art detection methods (FWA, Meso4, MesoInception4, Xception-c23, Xception-40 and DSP-FWA) on four largest datasets (FF-DF, DFD, DFDC and Celeb-DF).*

	Original	c23	c40
FWA	56.9	54.6	52.2
Xception-c23	65.3	65.5	52.5
Xception-c40	65.5	65.4	59.4
DSP-FWA	64.6	57.7	47.2

Table 5. *AUC performance of four top detection methods on original, medium (23) and high (40) degrees of H.264 compressed Celeb-DF respectively.*

5. Conclusion

We present a new challenging large-scale dataset for the development and evaluation of DeepFake detection methods. The Celeb-DF dataset reduces the gap in visual quality of DeepFake datasets and the actual DeepFake videos circulated online. Based on the Celeb-DF dataset, we perform a comprehensive performance evaluation of current DeepFake detection methods, and show that there is still much room for improvement.

For future works, the foremost task is to enlarge the Celeb-DF dataset and improve the visual quality of the synthesized videos. This entails improving the running efficiency and model structure of the current synthesis algorithm. Furthermore, while the forgers can improve the visual quality in general, they may also adopt *anti-forensic* techniques, which aim to hide traces of DeepFake synthesis on which the detection methods predicate. Anticipating such counter-measures at the forgers' disposal, we aim to incorporate anti-forensic techniques in Celeb-DF.

References

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

[2] DFaker github. https://github.com/dfaker/df, Accessed Nov 4, 2019.

[3] faceswap-GAN github. https://github.com/shaoanlu/faceswap-GAN, Accessed Nov 4, 2019.

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

[5] FakeApp. https://www.malavida.com/en/soft/fakeapp/, Acessed Nov 4, 2019.

[6] 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.

[7] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[8] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder-decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing (TIP)*, 2019.

[9] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. *ACM Transactions on Graphics (TOG)*, 2008.

[10] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019.

[11] 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.

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

[13] Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. *ACM Transactions on Graphics (TOG)*, 2011.

[14] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

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

[16] Hany Farid. *Digital Image Forensics*. MIT Press, 2012.

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

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2015.

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

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[22] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[23] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep Video Portraits. *ACM Transactions on Graphics 2018 (TOG)*, 2018.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

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

[26] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *ICCV*, 2017.

[27] 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.

[28] 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.

[29] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017.

[30] Yaqi Liu, Qingxiao Guan, Xianfeng Zhao, and Yun Cao. Image forgery localization based on multi-scale convolutional neural networks. In *ACM Workshop on Information Hiding and Multimedia Security (IHMMSec)*, 2018.

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

[32] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017.

[33] 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.

[34] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2019.

[35] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[36] 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.

[37] Hai X Pham, Yuting Wang, and Vladimir Pavlovic. Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. *arXiv preprint arXiv:1803.07716*, 2018.

[38] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 2001.

[39] 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.

[40] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

[41] 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.

[42] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NeurIPS*, 2017.

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

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

[45] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *CVPR*, 2018.

[46] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. What makes tom hanks look like tom hanks. In *ICCV*, 2015.

[47] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 2017.

[48] 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.

[49] 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.

[50] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. In *SIGGRAPH*, 2019.

[51] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, June 2016.

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 2004.

[53] 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.

[54] 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)*, 2017.

[55] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *CVPR*, 2018.