Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics

Yuezun Li¹, Xin Yang¹, Pu Sun², Honggang Qi² and Siwei Lyu¹
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 disinformation 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 Challenge¹. 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 FaceForensics++ dataset (FF-DF) [40]², the Google DeepFake detection dataset (DFD) [15], and the FaceBook DeepFake 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-DF³, 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 ethnic 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 DeepFake detection methods. This is the most comprehensive performance evaluation of DeepFake detection methods to

¹FaceForensics++ contains other types of fake videos. We consider only the DeepFake videos.
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 \textit{tool of choice} to create DeepFake videos in scale, with several independent open-source implementations, \textit{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 \((E)\) 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), \textit{i.e.,} the encoder and the decoder. The encoder \((E)\) converts the input target’s face to a vector known as the \textit{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 (\(\ell_1\) 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 \textit{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 DeepFake 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
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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Real Video</th>
<th># DeepFake Video</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>UADFV</td>
<td>49</td>
<td>17.3k</td>
<td>2018.11</td>
</tr>
<tr>
<td>DF-TIMIT-LQ</td>
<td>320</td>
<td>34.0k</td>
<td>2018.12</td>
</tr>
<tr>
<td>DF-TIMIT-HQ</td>
<td>320</td>
<td>34.0k</td>
<td>2018.12</td>
</tr>
<tr>
<td>FF-DF</td>
<td>1,000</td>
<td>599.9k</td>
<td>2019.01</td>
</tr>
<tr>
<td>DFD</td>
<td>363</td>
<td>315.4k</td>
<td>2019.09</td>
</tr>
<tr>
<td>DFDC</td>
<td>1,131</td>
<td>488.4k</td>
<td>2019.10</td>
</tr>
</tbody>
</table>

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

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\(^3\). 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

\(^3\)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.
the synthesized face to $256 \times 256$ 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.

**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 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,
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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>UADFV</th>
<th>DF-TIMIT</th>
<th>FF-DD</th>
<th>DFD</th>
<th>DFDC</th>
<th>Celeb-DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask-SSIM</td>
<td>0.82</td>
<td>0.80</td>
<td>0.80</td>
<td>0.81</td>
<td>0.88</td>
<td>0.84</td>
</tr>
</tbody>
</table>

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 MesoInception4. 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-
Table 3. Summary of compared DeepFake detection methods. See texts for more details.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Model Type</th>
<th>Training Dataset</th>
<th>Repositories</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>HeadPose [53]</td>
<td>SVM</td>
<td>UADFV [53]</td>
<td><a href="https://bitbucket.org/ericyang3171/HeadPose_forensics/">https://bitbucket.org/ericyang3171/HeadPose_forensics/</a></td>
<td>2018.11</td>
</tr>
<tr>
<td>VA-LogReg [53]</td>
<td>Logistic Regression</td>
<td>FaceForensics++</td>
<td><a href="http://github.com/ondyari/FaceForensics">http://github.com/ondyari/FaceForensics</a></td>
<td>2019.01</td>
</tr>
</tbody>
</table>

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.

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LQ</td>
<td>HQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-stream [54]</td>
<td>85.1</td>
<td>83.5</td>
<td>73.5</td>
<td>70.1</td>
<td>52.8</td>
<td>61.4</td>
</tr>
<tr>
<td>Meso4 [6]</td>
<td>84.3</td>
<td>87.8</td>
<td>68.4</td>
<td>84.7</td>
<td>76.0</td>
<td>75.3</td>
</tr>
<tr>
<td>Mesolnception4</td>
<td>82.1</td>
<td>80.4</td>
<td>62.7</td>
<td>83.0</td>
<td>75.9</td>
<td>73.2</td>
</tr>
<tr>
<td>HeadPose [53]</td>
<td>89.0</td>
<td>55.1</td>
<td>53.2</td>
<td>47.3</td>
<td>56.1</td>
<td>55.9</td>
</tr>
<tr>
<td>FWA [28]</td>
<td>97.4</td>
<td>99.9</td>
<td>93.2</td>
<td>80.1</td>
<td>74.3</td>
<td>72.7</td>
</tr>
<tr>
<td>VA-MLP [33]</td>
<td>70.2</td>
<td>61.4</td>
<td>62.1</td>
<td>66.4</td>
<td>69.1</td>
<td>61.9</td>
</tr>
<tr>
<td>VA-LogReg</td>
<td>54.0</td>
<td>77.0</td>
<td>77.3</td>
<td>78.0</td>
<td>77.2</td>
<td>66.2</td>
</tr>
<tr>
<td>Xception-raw [40]</td>
<td>80.4</td>
<td>56.7</td>
<td>54.0</td>
<td>99.7</td>
<td>53.9</td>
<td>49.9</td>
</tr>
<tr>
<td>Xception-c23</td>
<td>91.2</td>
<td>95.9</td>
<td>94.4</td>
<td>99.7</td>
<td>85.9</td>
<td>72.2</td>
</tr>
<tr>
<td>Xception-c40</td>
<td>83.6</td>
<td>75.8</td>
<td>70.5</td>
<td>95.5</td>
<td>65.8</td>
<td>69.7</td>
</tr>
<tr>
<td>Multi-task [34]</td>
<td>65.8</td>
<td>62.2</td>
<td>55.3</td>
<td>76.3</td>
<td>54.1</td>
<td>53.6</td>
</tr>
<tr>
<td>Capsule [36]</td>
<td>61.3</td>
<td>78.4</td>
<td>74.4</td>
<td>96.6</td>
<td>64.0</td>
<td>53.3</td>
</tr>
<tr>
<td>DSP-FWA</td>
<td>97.7</td>
<td>99.9</td>
<td>99.7</td>
<td>93.0</td>
<td>81.1</td>
<td>75.5</td>
</tr>
</tbody>
</table>

Figure 9. ROC curves of six state-of-the-art detection methods (FWA, Meso4, Mesolnception4, Xception-c23, Xception-40 and DSP-FWA) on four largest datasets (FF-DF, DFD, DFDC and Celeb-DF).

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

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>c23</th>
<th>c40</th>
</tr>
</thead>
<tbody>
<tr>
<td>FWA</td>
<td>56.9</td>
<td>54.6</td>
<td>52.2</td>
</tr>
<tr>
<td>Xception-c23</td>
<td>65.3</td>
<td>65.5</td>
<td>52.5</td>
</tr>
<tr>
<td>Xception-c40</td>
<td>65.5</td>
<td>65.4</td>
<td>59.4</td>
</tr>
<tr>
<td>DSP-FWA</td>
<td>64.6</td>
<td>57.7</td>
<td>47.2</td>
</tr>
</tbody>
</table>

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.

Acknowledgement. This material is based upon work supported by NSF under Grant No (IIS-1816227). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.
References


