Chapter 5

The Threat of Deepfakes to Computer and Human Visions

Pavel Korshunov and Sébastien Marcel

Abstract Deepfake videos, where a person's face is automatically swapped with a face of someone else, are becoming easier to generate with more realistic results. The concern for the impact of the widespread deepfake videos on the societal trust in video recordings is growing. In this chapter, we demonstrate how dangerous deepfakes are for both human and computer visions by showing how well these videos can fool face recognition algorithms and naïve human subjects. We also show how well the state-of-the-art deepfake detection algorithms can detect deepfakes and whether they can outperform humans.

5.1 Introduction

Recent advances in automated video and audio editing tools, generative adversarial networks (GANs), and social media allow creation and fast dissemination of high-quality tampered video content. Such content already led to appearance of deliberate misinformation, coined “fake news,” which is impacting political landscapes of several countries [3]. A recent surge of videos, often obscene, in which a face can be swapped with someone else's using neural networks, so-called Deepfakes,¹ are of a great public concern.² Accessible open-source software and apps for such face swapping (see Fig. 5.1 for illustration of the process) lead to large amounts of synthetically generated deepfake videos appearing in social media and news, posing a significant technical challenge for detection and filtering of such content. Some of the latest

¹ Open source: <https://github.com/deepfakes/faceswap>.

² BBC (Feb 3, 2018): <http://www.bbc.com/news/technology-42912529>.

P. Korshunov (✉) · S. Marcel
Idiap Research Institute, Martigny, Switzerland
e-mail: pavel.korshunov@idiap.ch

S. Marcel
e-mail: sebastien.marcel@idiap.ch

Fig. 5.1 Process of generating deepfake videos

approaches to detect deepfakes demonstrate encouraging accuracy, especially if they are trained and evaluated on the same datasets [17, 21].

At this stage, the research on deepfakes is still a relatively immature field, however the main research questions are already clear:

1. How to increase the amount of data with different types of deepfakes?
2. Can deepfakes fool automated face recognition?
3. Can deepfakes fool human visual system?
4. Can deepfakes be effectively detected?

In this chapter, we cover all the above research questions by (i) extending the pool of available deepfake datasets, (ii) demonstrating vulnerability of face recognition to deepfakes, (iii) presenting the results of subjective assessment of human ability to detect deepfakes, and (iv) showing the abilities and challenges of state-of-the-art deepfake detection approaches.

5.2 Related Work

The first approach that used a generative adversarial network to train a model between pre-selected two faces was proposed by Korshunova et al. [12]. Another related work with even a more ambitious idea was to use long short-term memory (LSTM)-based architecture to synthesize a mouth feature solely from an audio speech [24]. Right after these publications became public, they attracted a lot of publicity. Open-source approaches replicating these techniques started to appear, which resulted in the Deepfake phenomena.

Fig. 5.2 Examples of deepfakes (faces cropped from videos) in different databases

Many databases with deepfake videos (see examples in Fig. 5.2) were created to help develop and train deepfake detection methods. One of the first freely available databases was based on VidTIMIT [10], followed by the FaceForensics database, which contained deepfakes generated from 1'000 YouTube videos [20] and which later was extended with a larger set of high-resolution videos provided by Google and Jigsaw [21]. Another recently proposed 5'000 videos-large database of deepfakes generated from YouTube videos is Celeb-DF [14]. But the most extensive and the largest database to date with more than 100 K videos (80% of which are deepfakes) is the dataset from Facebook [5], which was available for download to the participants in the recent Deepfake Detection Challenge hosted by Kaggle.³

These datasets were generated using either the popular open-source code,⁴ e.g., DeepfakeTIMIT [10], FaceForensics [20], and Celeb-DF [14], or the latest methods implemented by Google and Facebook for creating deepfakes (see Fig. 5.2 for the examples of different deepfakes). This availability of large deepfake video databases allowed researchers to train and test detection approaches based on very deep neural networks, such as Xception [21], capsules networks [18], and EfficientNet [17], which were shown to outperform the methods based on shallow CNNs, facial physical characteristics [2, 13, 26, 27], or distortion features [1, 28].

5.3 Databases and Methods

Table 5.1 summarizes the databases of deepfake videos that we have used in the experiments presented in this chapter. The database by Google and Jigsaw and DF-Mobio database were split into three approximately equal in size subsets, for training, validation, and testing. The authors of Celeb-DF dataset predefined file lists for training and testing subsets but there was no validation set provided. DeepfakeTIMIT

³ <https://www.kaggle.com/c/deepfake-detection-challenge>.

⁴ <https://www.kaggle.com/c/deepfake-detection-challenge/discussion/121313>.

Table 5.1 Databases of deepfakes

Database	Number of swapped identities	Original videos	Deepfakes
DeepfakeTIMIT ⁵	32	320	640
DF-Mobio database	72	31 950	14 546
FaceForensics	1000	1000	1000
from Google and Jigsaw	approx. 150	360	3068
from Facebook	not large	23 849	95 396
Celeb-DF	1711	590	5639

was also split only into two subsets: training and testing, due to its small size. From the Facebook dataset, we manually selected 120 videos, which we used in the subjective evaluation.

5.3.1 *DeepfakeTIMIT*

The DeepfakeTIMIT⁵ is one of the first databases of deepfakes that we have generated by using videos from a VidTIMIT database, which contains short video clips of 43 subjects shot in a controlled environment when they are facing camera and reciting predetermined short phrases. Deepfakes were generated using open-source code⁶ for 16 pairs of subjects selected based on how similar their visual appearance is, including mustaches or hair styles.

DeepfakeTIMIT contains two types of deepfakes (see examples in Fig. 5.3), the lower quality (LQ) fakes where a GAN model was trained to generate 64×64 size images and a higher quality (HQ), where GAN was trained to generate 128×128 images. The generated faces were placed in the target video using automated blending techniques that relied on histogram normalization and selective masking with Gaussian blur.

5.3.2 *DF-Mobio*

DF-Mobio⁷ dataset is also generated by us and is one of the largest databases available with almost 15 K deepfake and 31 K real videos (see Table 5.1 for the comparison with other databases). Original videos are taken from Mobio database [15], which contains videos of a single person talking to the camera recorded with a phone or a

⁵ <https://www.idiap.ch/dataset/deepfaketimit>.

⁶ <https://github.com/shaoanlu/faceswap-GAN>.

⁷ <https://www.idiap.ch/dataset/dfmobio>.

Fig. 5.3 Screenshot of the original videos from VidTIMIT database and low- (LQ) and high-quality (HQ) DeepfakeTIMIT videos

Fig. 5.4 Screenshots of the original videos and a deepfake swap from DF-Mobio database

laptop. The scenario simulates the participation in a virtual meeting over Zoom or Skype.

The original Mobio database contains 31 K videos from 152 subjects but deepfakes were generated only for manually pre-selected 72 pairs of people with similar hairstyles, facial features, facial hair, and eyewear. Using GAN-based face-swapping algorithm based on the available code⁶, for each pair, we generated videos with swapped faces from subject one to subject two and visa versa (see Fig. 5.4 for video screenshot examples).

The GAN model for face swapping was trained on face size input of 256×256 pixels. The training images were generated from laptop-recorded videos at 8 fps, resulting in more than 2 K faces for each subject, the training was done for 40 K iterations (about 24 hours on Tesla P80 GPU). The availability of this database to public is pending a publication.

5.3.3 Google and Jigsaw

To make this dataset, Google and Jigsaw [23] (see Table 5.1 for the comparison with other databases) worked with paid and consenting actors to record hundreds of videos. Using publicly available deepfake generation methods, Google then generated about

3 K of deepfakes from these videos. The resulting videos, real and fake, comprise the contribution, which was created to directly support deepfake detection efforts. As part of the FaceForensics++ [21] benchmark, this dataset is now available, free to the research community, for developing synthetic video detection methods.

5.3.4 Facebook

For construction of Facebook database, a data collection campaign [6] (see Table 5.1 for the comparison with other databases) has been carried out where participating actors have entered into an agreement to the use and manipulation of their likenesses in the creation of the dataset. Diversity in several axes (gender, skin tone, age, etc.) has been considered and actors recorded videos with arbitrary backgrounds thus bringing visual variability. A number of face swaps were computed across subjects with similar appearances, where each appearance was inferred from facial attributes (skin tone, facial hair, glasses, etc.). After a given pairwise model was trained on two identities, each identity was swapped onto other's videos.

For our experiments, we have manually looked through many videos of Facebook database and pre-selected 60 deepfake videos, split into five categories depending of how fake they look to an expert eye, with the corresponding 60 original videos (see examples in Fig. 5.5).

We use this manually selected subset of the videos in the subjective evaluations aimed to study the level of difficulty human subjects have in recognizing different types of deepfakes. We also use the same videos to evaluate deepfake detection systems and compare their performance with the human subjects.

Fig. 5.5 Cropped faces from different categories of deepfake videos of Facebook database (top row) and the corresponding original versions (bottom row)

5.3.5 *Celeb-DF*

Celeb-DF (v2) [14] dataset contains real and deepfake synthesized videos having similar visual quality on par with those circulated online. The Celeb-DF (v2) dataset is greatly extended from the previous Celeb-DF (v1), which only contained 795 deepfake videos.

The v2 of the database contains more than 5 K deepfakes and nearly 2 K real videos, which 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 [14], which essentially is an extension of methods available online¹, similar to the one used to generate both FaceForensics and DF-Mobio databases. The authors of the Celeb-DF database claim that their modified algorithm improves the overall visual quality of the synthesized deepfakes when compared to existing datasets. The authors also state that Celeb-DF is challenging to most of the existing detection methods, even though many deepfake detection methods are shown to achieve high, sometimes near perfect, accuracy on previous datasets. No consent was obtained for the videos, because the data is from the celebrities of the YouTube videos.

5.4 Evaluation Protocols

In this section, we explain how we evaluate face recognition and deepfake detection systems and what kind of objective metrics we compute.

5.4.1 *Measuring Vulnerability*

We use DeepfakeTIMIT database to evaluate vulnerability of face recognition. For the licit non-tampered scenario, the original VidTIMIT videos for the 32 subjects for which we have generated corresponding deepfake videos. In this scenario, we used two videos of the subject for enrollment and the other eight videos as probes, for which we computed the verification scores.

Using the scores, for each possible threshold θ , we compute commonly used metrics for evaluation of classification systems: false accept rate (FAR), which is the same as false match rate (FMR) and false reject rate (FRR), which is the same as false non-match rate (FNMR).⁸ These rates are generally defined as follows:

⁸ Strictly speaking, FRR and FNMR are not the same but are equivalent in our evaluations, since we are not concerned with such preprocessing errors like failure to enroll.

$$\begin{aligned}
FAR(\theta) &= \frac{|\{h_{neg} \mid h_{neg} \geq \theta\}|}{|\{h_{neg}\}|} \\
FRR(\theta) &= \frac{|\{h_{pos} \mid h_{pos} < \theta\}|}{|\{h_{pos}\}|},
\end{aligned} \tag{5.1}$$

where h_{pos} is a score for original genuine samples and h_{neg} is a score for the tampered samples.

Threshold at which these FAR and FRR are equal leads to an equal error rate (EER), which is commonly used as a single value metric of the system performance.

To evaluate vulnerability of face recognition to deepfakes, in *tampered* scenario, we use deepfake videos (10 for each of 32 subjects) as probes and compute the corresponding scores using the enrollment model from the licit scenario. To understand if face recognition perceives deepfakes to be similar to the genuine original videos, we report the FAR metric computed using EER threshold θ from licit scenario. If FAR value for deepfake tampered videos is significantly higher than the one computed in licit scenario, it means the face recognition system cannot distinguish tampered videos from originals and is therefore vulnerable to deepfakes.

5.4.2 Measuring Detection

We consider deepfake detection as a binary classification problem and evaluate the ability of detection approaches to distinguish original videos from deepfake videos. All videos in DF-Mobio and Google datasets were proportionally split into training, validation, and test subsets. For Celeb-DF database, only training and test subset were provided, so the test set was used in place of validation when necessary. Similarly, DeepfakeTIMIT was also split into tow training and test subsets due to its smaller size.

The result of a deepfake detection is a set of probabilistic scores where the values close to zero correspond to deepfakes and those that are close to one correspond to genuine videos.

We define the threshold θ_{far} on the validation set to correspond to the FAR value of 10%, which means 10% of fake videos are allowed to be misclassified as genuine. Using this threshold θ_{far} on the scores of the test set will result in test FAR and FRR values. As a single value metric, we can then use the half total error rate (HTER) defined as

$$HTER(\theta_{far}) = \frac{FAR_{test} + FRR_{test}}{2}. \tag{5.2}$$

In addition to reporting FAR, FRR, and HTER values for the scores of the test set, we also report the area under the curve (AUC) metric, which is a popular metric for evaluation of classification system and is often used in the deepfake detection literature.

5.5 Vulnerability of Face Recognition

As examples of face recognition systems, we used publicly available pre-trained VGG [19] and Facenet [22] architectures. We used the *fc7* and *bottleneck* layers of these networks, respectively, as features and used cosine distance as a classifier. For a given test face, the confidence score of whether it belongs to a pre-enrolled model of a person is the cosine distance between the average feature vector, i.e., model, and the feature vector of a test face. Both of these systems are state-of-the-art recognition systems with VGG of 98.95% [19] and Facenet of 99.63% [22] accuracies on labeled faces in the wild (LFW) dataset.

We conducted the vulnerability analysis of VGG- and Facenet-based face recognition systems on low-quality (LQ) and high-quality (HQ) face swaps in Deepfake-TIMIT database. The results are presented in Table 5.2. In a licit scenario when only original non-tampered videos are present, both systems performed very well, with EER value of 0.03% for VGG and 0.00% for Facenet-based system. Using the EER threshold from licit scenario, we computed FAR value for the scenario when deepfake videos are used as probes. In this case, for VGG the FAR is 88.75% on LQ deepfakes and 85.62% on HQ deepfakes, and for Facenet the FAR is 94.38% and 95.00% on LQ and HQ deepfakes, respectively. To illustrate this vulnerability, we plot the score histograms for high-quality deepfake videos in Fig. 5.6. The histograms show a considerable overlap between deepfake and genuine scores with clear separation from the zero-effort impostor scores (the probes from licit scenario).

From the results, it is clear that both VGG- and Facenet-based systems cannot effectively distinguish GAN-generated and swapped faces from the original ones. The fact that more advanced Facenet system is more vulnerable is also consistent with the findings about presentation attacks [16].

5.6 Subjective Assessment of Human Vision

Since the resulted videos produced by automated deepfake generation algorithms vary drastically visually, depending on many factors (training data, the quality of the video for manipulation, and the algorithm itself), we cannot label all deepfakes into

Table 5.2 Vulnerability analysis of VGG and Facenet-based face recognition (FR) systems on low-quality (LQ) and high-quality (HQ) deepfakes in DeepfakeTIMIT database. EER value (Test set) is computed in a licit scenario without deepfakes. Using the corresponding EER threshold, FAR value (Test set) is computed for the scenario when deepfake videos are used as probes

Dataset version	VGG-based FR		Facenet-based FR	
	EER (%)	FAR (%)	EER (%)	FAR (%)
LQ deepfake	0.03	88.75	0.00	94.38
HQ deepfake	0.03	85.62	0.00	95.00

(a) VGG-based face recognition

(b) FaceNet-based face recognition

Fig. 5.6 Histograms showing the vulnerability of VGG- and Facenet-based face recognition to high-quality face swapping on low- and high-quality Deepfakes

Fig. 5.7 Screenshot of one step of subjective evaluation (the video is courtesy of Facebook database)

one visual category. Therefore, we have manually looked through many videos of Facebook database³ and pre-selected 60 deepfake videos, split into five categories depending on how clearly fake they look, with the corresponding 60 original videos (see examples in Fig. 5.5).

The evaluation was conducted using QualityCrowd 2 framework [9] designed for crowdsourcing-based evaluations (Fig. 5.7 shows a screenshot of a typical evaluation step). This framework allows us to make sure subjects watch each video fully at least once and are not able to skip any question. Prior to the evaluation itself, a display brightness test was performed using a method similar to that described in [8]. Since deepfake detection algorithms typically evaluate only the face regions cropped using a face detector, to have a comparable scenario, we have also shown to the human subjects cropped face regions next to the original video (see Fig. 5.7).

Each of the 60 naïve subjects who participated in the evaluation had to answer the question after watching a given video: “Is face of the person in the video real or fake?” with the following options: “Fake,” “real,” and “I do not know.” Prior to the evaluation, the explanation of the test was given to the subjects with several test video examples of different fake categories and real videos. The 120 were also split in random batches of 40 each to reduce the total evaluation time for one subject, so the average time per one evaluation was about 16 minutes, which is consistent with the standard recommendations.

Due to privacy concerns, we did not collect any personal information from our subjects such as age or gender. Also, the licensing conditions of Facebook database³ restricted the evaluation to the premises of Idiap Research Institute, which signed the license agreement not to distribute data outside. Therefore, the subjects consisted of

PhD students, scientists, administration, and management of Idiap. Hence, the age can be estimated to be between 20 and 65 years old and the gender distribution to be of a typical scientific community.

Unlike laboratory-based subjective experiments where all subjects can be observed by operators and its test environment can be controlled, the major shortcoming of the crowdsourcing-based subjective experiments is the inability to supervise participants behavior and to restrict their test conditions. When using crowdsourcing for evaluation, there is a risk of including untrusted data into analysis due to the wrong test conditions or unreliable behavior of some subjects who try to submit low-quality work in order to reduce their effort. For this reason, unreliable workers detection is an inevitable process in crowdsourcing-based subjective experiments. There are several methods for identifying the “trustworthiness” of the subject but since our evaluation was conducted within premises of a scientific institute, we only used so-called “honeypot” method [8, 11] to filter out scores from people who did not pay attention at all. Honeypot is a very easy question that refers to the video the subject just watched in the previous steps, e.g., “what was visible in the previous video?” with obvious answers that test if a person even looked at the video. Using this question, we filtered out the scores from five people from our final results, hence we ended up with 18.66 answers on average for each video, which is the number of subjects commonly considered in subjective evaluations.

5.6.1 Subjective Evaluation Results

For each deepfake or original video, we computed the percentage of answers that were “certain and correct,” when people selected “Real” for an original or “Fake” for a deepfake, “certain and incorrect” (selected “Real” for a deepfake or “Fake” for an original) and “uncertain,” when the selection was “I do not know.” We have averaged those percentages across videos in each category to obtain the final percentages, which are shown in Fig. 5.8a. From the figure, we can note that the pre-selected deepfake categories, on average, reflect the difficulty level of recognizing them. The interesting result is the low number of uncertain answers, which means people tend to be sure when it comes to judging the realism of a video. And it also means people can be easily spoofed by a good quality deepfake video, since only in 24.5% cases “well done” deepfake videos are perceived as fakes, even though these subjects already knew they are looking for fakes. In the scenario, when such deepfake would be distributed to an unsuspected audience (e.g., via social media), we can expect the number of people noticing it to be significantly lower. Also, it is interesting to note that even videos from “easy” category were not as easy to spot (71.1% correct answers) compared to the original videos (82.2%). Overall, we can see that people are better at recognizing very obvious examples of deepfakes or real unaltered videos.

To check whether the difference between videos from the five deepfake categories is statistically significant based on the subjective scores, we performed ANOVA test with the corresponding box plot shown in Fig. 5.8b. The scores were computed

Fig. 5.8 Subjective answers and median values with error bars from ANOVA test for different deepfake categories

(a) Subjective answers

(b) ANOVA test

for each video (and per category when applicable) by averaging the answers from all corresponding observers. For each correct answer, the score is 1 and for both wrong and uncertain answers the score is 0. Please note that the red lines in Fig. 5.8b correspond to median values, not average, which what we plotted in Fig. 5.8a. The *p-value* of ANOVA test is below $4.7e - 11$, which means the deepfake categories are significantly different on average. However, Fig. 5.8b shows that “easy,” “moderate,” and “difficult” categories have large scores variations and overlap, which means some of the videos from these categories are perceived similarly. It means some of the deepfake videos could be moved to another category. This observation is

Fig. 5.9 Average scores with confidence intervals for each video in every video category

also supported by Fig. 5.9 which plots the average scores with confidence intervals (computed using student's t-distribution [7]) for each video in the deepfake category (12 videos each) and originals (60 videos).

5.7 Evaluation of Deepfake Detection Algorithms

For the example of machine vision, we took two state-of-the-art algorithms: based on Xception model [4] and EfficientNet variant B4 [17] shown to be performing very well on different deepfake datasets and benchmarks [21]. We pre-trained these models for 20 epochs each on the Google and Jigsaw dataset database [21] and Celeb-DF [14] to demonstrate the impact of different training conditions on the evaluation results. If evaluated on the test sets of the same databases they were trained on, both Xception and EfficientNet classifiers demonstrate a great performance as shown in Table 5.3. We can see that the area under the curve (AUC), which is the common metric used to compare the performance of deepfake detection algorithms, is almost at 100% in all cases.

We evaluated these models on the 120 videos we used in the subjective test. Since these videos come from Facebook database, they can be considered as unseen data, which is still an obstacle for many DNN classifiers, as they do not generalize well on the unseen data the fact also highlighted in the recent Facebook Deepfake

Table 5.3 Area under the curve (AUC) value on the test sets of Google and Celeb-DF databases of Xception and EfficientNet models

Model	Trained on	AUC (%) on Test set
Xception	Google database	100.00
Xception	Celeb-DF database	100.0
EfficientNet	Google database	99.99
EfficientNet	Celeb-DF database	100.0

Fig. 5.10 The detection accuracy (the threshold corresponds to FAR 10% on development set of the respective database) for each video category from subjective test by Xception and EfficientNet models pre-trained on Google and Celeb-DF databases

Detection Challenge [25]. To compute performance accuracy, we need to select threshold. We chose the threshold corresponding to the false accept rate (FAR) of 10% selected on the development set of the respective database. We selected threshold based on FAR value as oppose to equal error rate (EER) commonly used in biometrics, because many practical deepfake detection or anti-spoofing systems have a low bound requirement on FAR value. In our case, FAR of 10% is quite generous.

Figure 5.10 demonstrates the evaluation results of pre-trained Xception and EfficientNet models on the videos from the subject test averaged for each deepfake category and originals (when using threshold corresponding to $FAR = 10\%$). In the figure, blue bar corresponds to the percent of correctly detected videos in the given category, and the orange bar corresponds to the percent of incorrectly detected. The results for algorithms are very different from the results of the subjective test (see Fig. 5.8a for the evaluation results by human subjects). The accuracy of the algorithms has no correlation to the visual appearance of deepfakes. The algorithms “see” these

Fig. 5.11 ROC curves with the corresponding AUC value of Xception and Efficient models pre-trained on Google and Celeb-DF databases evaluated on all the videos from subjective test

videos very differently from how humans perceive the same videos. To a human observer, the result may even appear random. We can even notice that all algorithms struggle the most with the deepfake videos that were easy for human subjects. It is evident that the choice of threshold and the training data have major impact on the evaluation accuracy. However, when selecting a deepfake detection system to use in practical scenario, one cannot assume an algorithm’s perception will have any relation to the way we think the videos look like.

If we remove the choice of the threshold and the pre-selected video categories and simply evaluate the models on the 120 videos from the subjective tests, the receiver operating characteristic (ROC) curve and the corresponding AUC values are presented in Fig. 5.11. From this figure, we can note that ROC curves look “normal,” as typical curves for classifiers that do not generalize well on unseen data, especially taking into account excellent performance on the test sets shown in Table 5.3. Figure 5.11 also shows that human subjects were more accurate at assessing this set of videos since the corresponding ROC curve is consistently higher with the highest AUC value of 87.47%.

5.8 Conclusion

In this chapter, we presented several publicly available databases of deepfake videos, including two DeepfakeTIMIT and DF-Mobio generated and provided by us. We demonstrated that the state-of-the-art VGG and Facenet-based face recognition algorithms are vulnerable to the Deepfake videos and fail to distinguish such videos from the original ones with up to 95.00% equal error rate.

We also conducted a subjective evaluation on 120 different videos (60 deepfakes and 60 originals) manually pre-selected from the Facebook database, which demonstrated that people are confused by good quality deepfakes in 75.5% of cases.

On the other hand, the evaluated state-of-the-art deepfake detection algorithms (based on Xception and EfficientNets (B4 variant) neural networks pre-trained on Google or Celeb-DF datasets) show very different perception of deepfakes compared to human subjects. The algorithms struggle to detect many videos that look obviously fake to humans, while some of the algorithms (depending on the training data and the selected threshold) can accurately detect videos that are difficult for people. The experiments also demonstrate that deepfake detection algorithm struggle to generalize to unknown set of videos, for which they were not trained for.

The continued advancements in development of face swapping techniques will result in more challenging deepfake, which will be even harder to detect by the existing algorithms. Therefore, new databases and approaches that can better generalize on unseen and realistic deepfakes need to be developed in the future. The arms race between deepfake generation and detection methods is in full swing.

References

1. Agarwal A, Singh R, Vatsa M, Noore A (2017) Swapped! digital face presentation attack detection via weighted local magnitude pattern. In: IEEE international joint conference on biometrics (IJCB), pp 659–665. <https://doi.org/10.1109/BTAS.2017.8272754>
2. Agarwal S, El-Gaaly T, Farid H, Lim S (2020) Detecting deep-fake videos from appearance and behavior. [arXiv:2004.14491](https://arxiv.org/abs/2004.14491)
3. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *J Econ Perspect* 31(2):211–236. <https://doi.org/10.1257/jep.31.2.211>
4. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1800–1807
5. Dolhansky B, Bitton J, Pflaum B, Russ Howes JL, Wang M, Ferrer CC (2020) The deepfake detection challenge dataset. [arXiv:2006.07397](https://arxiv.org/abs/2006.07397)
6. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The deepfake detection challenge (dfdc) preview dataset. In: *ArXiv*
7. Hanhart P, Rerabek M, Korshunov P, Ebrahimi T (2013) Subjective evaluation of HEVC intra coding for still image compression
8. Hossfeld T, Keimel C, Hirth M, Gardlo B, Habigt J, Diepold K, Tran-Gia P (2014) Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Trans Multimed* 16(2):541–558
9. Keimel C, Habigt J, Horsch C, Diepold K (2012) Qualitycrowd—a framework for crowd-based quality evaluation. <https://doi.org/10.1109/PCS.2012.6213338>
10. Korshunov P, Marcel S (2019) Vulnerability assessment and detection of Deepfake videos. In: *International conference on biometrics (ICB 2019)*. Crete, Greece
11. Korshunov P, Nemoto H, Skodras A, Ebrahimi T (2014) Crowdsourcing-based evaluation of privacy in HDR images. In: *Optics, photonics, and digital technologies for multimedia applications III*, vol 9138. SPIE, pp 1 – 11
12. Korshunova I, Shi W, Dambre J, Theis L (2017) Fast face-swap using convolutional neural networks. In: *IEEE international conference on computer vision (ICCV)*, pp 3697–3705. <https://doi.org/10.1109/ICCV.2017.397>

13. Li Y, Chang MC, Lyu S (2018) In ictu oculi: exposing ai generated fake face videos by detecting eye blinking. [arXiv:1806.02877](https://arxiv.org/abs/1806.02877)
14. Li Y, Sun P, Qi H, Lyu S (2020) Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: IEEE conference on computer vision and pattern recognition (CVPR). Seattle, WA, United States
15. McCool C, Marcel S, Hadid A, Pietikäinen M, Matejka P, Cernocký J, Poh N, Kittler J, Larcher A, Lévy C, Matrouf D, Bonastre J, Tresadern P, Cootes T (2012) Bi-modal person recognition on a mobile phone: using mobile phone data. In: 2012 IEEE international conference on multimedia and expo workshops, pp 635–640
16. Mohammadi A, Bhattacharjee S, Marcel S (2018) Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. *IET Biom* 7(1):15–26. <https://doi.org/10.1049/iet-bmt.2017.0079>
17. Montserrat DM, Hao H, Yarlagadda SK, Baireddy S, Shao R, Horvath J, Bartusiak E, Yang J, Güera D, Zhu F, Delp EJ (2020) Deepfakes detection with automatic face weighting. In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 2851–2859
18. Nguyen H, Yamagishi J, Echizen I (2019) Capsule-forensics: using Capsule networks to detect forged images and videos. In: IEEE international conference on acoustics, speech and signal processing (2019)
19. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: *BMVC* (2015)
20. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2018) Faceforensics: a large-scale video dataset for forgery detection in human faces. [arXiv:1803.09179](https://arxiv.org/abs/1803.09179)
21. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) FaceForensics++: learning to detect manipulated facial images. In: International conference on computer vision (ICCV)
22. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: a unified embedding for face recognition and clustering. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 815–823. <https://doi.org/10.1109/CVPR.2015.7298682>
23. Stanton D, Karlsson P, Vorobyov A, Leung T, Childs J, Sud A, Bregler C (2019) Contributing data to deepfake detection research. In: *Blogpost*
24. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing obama: learning lip sync from audio. *ACM Trans Graph* 36(4):95:1–95:13. <https://doi.org/10.1145/3072959.3073640>
25. Tolosana R, Romero-Tapiador S, Fierrez J, Vera-Rodriguez R (2020) Deepfakes evolution: analysis of facial regions and fake detection performance
26. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head pose. In: IEEE international conference on acoustics, speech and signal processing
27. Yang X, Li Y, Qi H, Lyu S (2019) Exposing GAN-synthesized faces using landmark locations. In: *ACM workshop on information hiding and multimedia security*, pp 113–118
28. Zhang Y, Zheng L, Thing VLL (2017) Automated face swapping and its detection. In: IEEE international conference on signal and image processing (ICSIP), pp 15–19. <https://doi.org/10.1109/SIPROCESS.2017.8124497>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.