# Vulnerability assessment and detection of Deepfake videos

Pavel Korshunov and Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland
{pavel.korshunov,sebastien.marcel}@idiap.ch

## Abstract

*It is becoming increasingly easy to automatically replace a face of one person in a video with the face of another person by using a pre-trained generative adversarial network (GAN). Recent public scandals, e.g., the faces of celebrities being swapped onto pornographic videos, call for automated ways to detect these Deepfake videos. To help developing such methods, in this paper, we present the first publicly available set of Deepfake videos generated from videos of VidTIMIT database. We used open source software based on GANs to create the Deepfakes, and we emphasize that training and blending parameters can significantly impact the quality of the resulted videos. To demonstrate this impact, we generated videos with low and high visual quality (320 videos each) using differently tuned parameter sets. We showed that the state of the art face recognition systems based on VGG and Facenet neural networks are vulnerable to Deepfake videos, with 85.62% and 95.00% false acceptance rates (on high quality versions) respectively, which means methods for detecting Deepfake videos are necessary. By considering several baseline approaches, we found the best performing method based on visual quality metrics, which is often used in presentation attack detection domain, to lead to 8.97% equal error rate on high quality Deepfakes. Our experiments demonstrate that GAN-generated Deepfake videos are challenging for both face recognition systems and existing detection methods, and the further development of face swapping technology will make it even more so.*

## 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 [2]. 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 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. Therefore, the development of efficient tools that can automatically detect these videos with swapped faces is of a paramount importance.

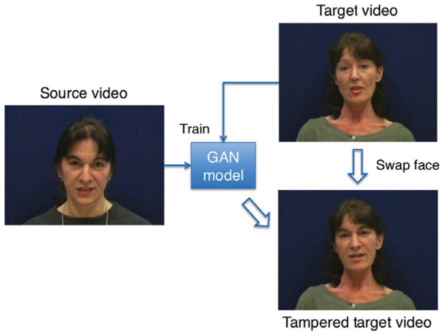


Figure 1: Process of generating Deepfake videos.

Until recently, most of the research was focusing on advancing the face swapping technology [8, 10, 15, 17]. However, responding to the public demand to detect such face swapping, researchers are starting to work on databases and detection methods, including image and video data [18] generated with a previous generation of face swapping approach Face2Face [21] or videos collected using Snapchat[3] application [1].

In this paper, we present the first publicly available database (with a permissible license) of videos where faces are swapped using the open source GAN-based approach[4] (see Figure 1 for illustration), which is developed from the original autoencoder-based Deepfake algorithm[1]. We man-

[1] Open source: https://github.com/deepfakes/faceswap
[2] BBC (Feb 3, 2018): http://www.bbc.com/news/technology-42912529
[3] https://www.snapchat.com/
[4] https://github.com/shaoanlu/faceswap-GAN

978-1-7281-3640-0/19/$31.00 ©2019 IEEE

ually selected 16 similar looking pairs of people from publicly available VidTIMIT database[5]. For each of 32 subjects, we trained two different models (see Figure 2 for examples), referred to in the paper as the low quality (LQ) model, with $64 \times 64$ input/output size, and the high quality (HQ) model, with $128 \times 128$ size. Since there are 10 videos per person in VidTIMIT database, we generated 320 videos corresponding to each version, resulting in total 620 videos with faces swapped. For the audio, we kept the original audio track of each video, i.e., no manipulation was done to the audio channel.

It is important to understand how much of a threat Deepfake videos are to face recognition systems. Because if these systems are not fooled by Deepfakes, creating a separate system for detecting Deepfakes would not be necessary. To assess the vulnerability of face recognition to Deepfake videos, we evaluate two state of the art systems: based on VGG [16] and Facenet[6] [19] neural networks, on both untampered videos and videos with faces swapped.

For detection of the Deepfakes, we first used an audio-visual approach that detects inconsistency between visual lip movements and speech in audio [9]. It allows us to understand how well the generated Deepfakes can mimic mouth movement and whether the lips are synchronized with the speech. We also applied several baseline methods from presentation attack detection domain, by treating Deepfake videos as digital presentation attacks [1], including simple principal component analysis (PCA) and linear discriminant analysis (LDA) approaches, and the approach based on image quality metrics (IQM) and support vector machine (SVM) [7, 22].

To allow researchers to verify, reproduce, and extend our work, we provide the database coined DeepfakeTIMIT of Deepfake videos[7], face recognition and Deepfake detection systems with corresponding scores as an open source Python package[8].

Therefore, this paper has the following main contributions:

- Publicly available database of low and high quality sets of videos from VidTIMIT database with swapped faces using GAN-based approach;
- Vulnerability analysis of VGG and Facenet based face recognition systems;
- Evaluation of several detection methods of Deepfakes, including lip-syncing approach and image quality metrics with SVM method;

## 2. Related work

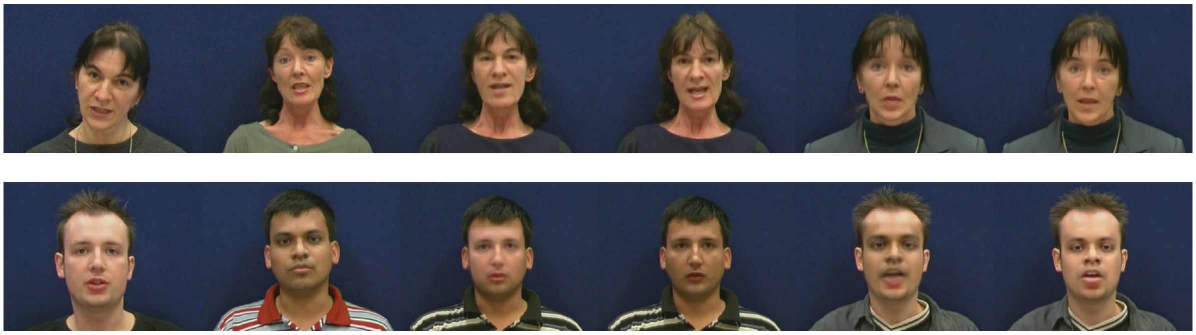
One of the first works on face swapping is by Bitouk *et al.* [4], where the authors searched in a database for a face similar in appearance to the input face and then focused on perfecting the blending of the found face into the input image. The main motivation for this work was de-identification of an input face and its privacy preservation. Hence, the approach did not allow for a seamless swapping of any two given faces. Until the latest era of neural networks, most of the techniques for face swapping or facial reenacment were based on similarity searchers between faces or face patches in target and source video and various blending techniques [3, 23, 6, 13, 15].

The first approach that used a generative adversarial network to train a model between pre-selected two faces was proposed by Korshunova *et al.* in 2017 [10]. 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 [20]. Right after these publication became public, they attracted a lot of publicity. Open source approaches replicating these techniques started to appear, which resulted in the Deepfake phenomena.

The rapid spread of Deepfakes and the ease of generating such videos are calling for a reliable detection method. So far, however, there are only few publications focusing on detecting GAN-generated videos with swapped faces and very little data for evaluation and benchmarking is publicly available. For instance, Zhang *et al.* [25] proposed the method based on speeded up robust features (SURF) descriptors and SVM classifier. The authors evaluated this approach on a set of images where the face of one person was replaced with a face of another by applying color correction and smoothing techniques based on Gasussian blurring, which means the facial expressions of the input faces were not preserved. Another method based on LBP-like features with SVM classifier was proposed by Agarwal *et al.* [1] and evaluated on the videos collected by the authors with Snapchat[3] phone application. Snapchat uses active 3D model to swap faces in real time, so the resulted videos are not really Deepfakes, but it is still a widely used tool and database of such videos, if it will ever become public (the authors promised to release it but have not done so at the moment of publication), it can be interesting to research community.

Rössler *et al.* [18] presented the most comprehensive database of non-Deepfake swapped faces ($500'000$ images from more than $1000$ videos) to date. The authors also benchmarked the state of the art forgery classification and segmentation methods. The authors used Face2Face [21] tool to generate the database, which is based on expression transformation using 3D facial model and a pre-computed database of mouth interiors. One of the latest

[5] http://conradsanderson.id.au/vidtimit/
[6] https://github.com/davidsandberg/facenet
[7] https://www.idiap.ch/dataset/deepfaketimit
[8] Source code: https://gitlab.idiap.ch/bob/bob.report.deepfakes



(g) Original 1 (h) Original 2 (i) LQ swap 1 → 2 (j) HQ swap 1 → 2 (k) LQ swap 2 → 1 (l) HQ swap 2 → 1

Figure 2: Screenshot of the original videos from VidTIMIT database and low (LQ) and high quality (HQ) Deepfake videos.

approaches [12] proposed to use blinking detection as the means to distinguish swapped faces in Deepfake videos. The authors generated 49 videos (not publicly available) and argued that the proposed eye blinking detection was effective in detecting Deepfake videos.

However, no public Deepfake video database where GAN-based approach was applied is available. Hence, it is unclear whether the above methods would be effective in detecting such faces. In fact, the Deepfakes that we have generated can effectively mimic the facial expressions, mouth movements, and blinking, so the current detection approaches need to be evaluated on such videos. However, it is practically impossible to evaluate the methods proposed in [18] and [12] as their implementations are not yet available.

## 3. Deepfake database

As the original data, we took video from VidTIMIT database[5]. The database contains 10 videos for each of 43 subjects, which were shot in controlled environment with people facing camera and reciting predetermined short phrases. From these 43 subject, we manually selected 16 pairs in such a way that subjects in the same pair have similar prominent visual features, e.g., mustaches or hair styles. Using GAN-based face-swapping algorithm based on the available code[4], for each pair, we generated videos with swapped faces from subject one to subject two and visa versa (see Figure 2 for the video screenshots).

For each pair of subjects, we have trained two different GAN models and generated two versions of the videos:

1. The low quality (LQ) model has input and output image (facial regions only) of size $64 \times 64$. About 200 frames from the videos of each subject were used for training and the frames were extracted at 4 fps from the original videos. The training was done for $10'000$ iterations and took about 4 hours per model on Tesla P40 GPU.
2. The high quality (HQ) model has input/output image size of $128 \times 128$. About 400 frames extracted at 8 fps from videos were used for training, which was done for $20'000$ iterations (about 12 hours on Tesla P40 GPU).

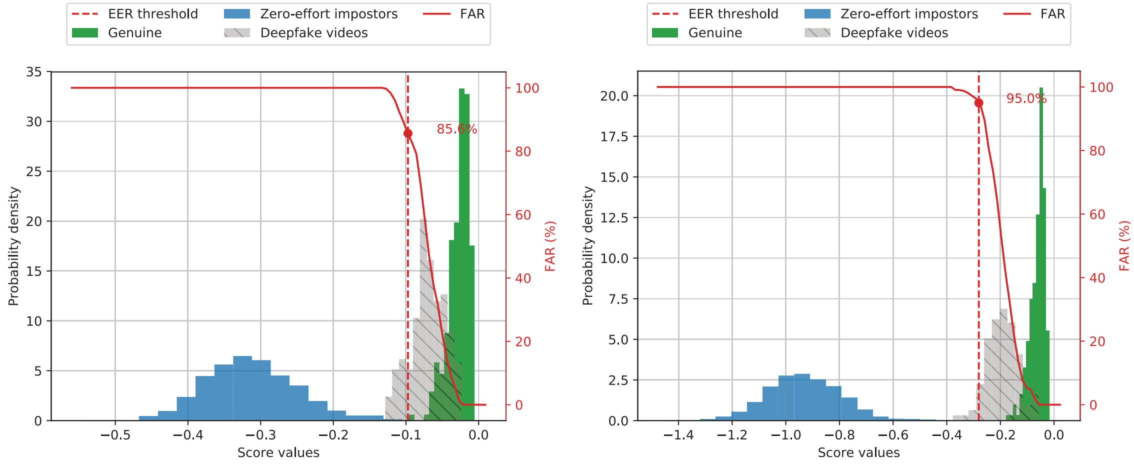
Also, different blending techniques were used when generating Deepfake videos using different models. With LQ model, for each frame from an input video, generator of the GAN model was applied on the face region to generate the fake counterpart. Then a facial mask was detected using a CNN-based face segmentation algorithm proposed in [15]. Using this mask, the generated fake face was blended with the face in the target video. For HQ model, the blending was done based on facial landmarks (detected with publicly available MTCNN model [24]) alignment between generated fake face and the original face in the target video. Finally, histogram normalization was applied to the blended result to adjust for the lighting conditions, which makes the result more realistic (see Figure 2).

### 3.1. Evaluation protocol

When evaluating vulnerability of face recognition, for the *licit* non-tampered scenario, we used the original VidTIMIT[5] videos for the 32 subjects for which we have generated corresponding Deepfake videos. In this scenario, we used 2 videos of the subject for enrollment and the other 8 videos as probes, for which we computed the verification scores.

From the scores, for each possible threshold $\theta$, we computed commonly used metrics for evaluation of classification systems: false acceptance rate (FAR) and false reject rate (FRR). 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*



(a) VGG-based face recognition

(b) FaceNet-based face recognition

Figure 3: Histograms showing the vulnerability of VGG and Facenet based face recognition to high quality face-swapping on low and high quality Deepfakes.

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 $\theta$ 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.

When evaluating Deepfake detection, we consider it as a binary classification problem and evaluate the ability of detection approaches to distinguish original videos from Deepfake videos. All videos in the dataset, including genuine and tampered parts, were split into training (*Train*) and evaluation (*Test*) subsets. To avoid bias during training and testing, we arranged that the same subject would not appear in both sets. We did not introduce a development set, which is typically used to tune hyper parameters such as threshold, because the dataset is not large enough. Therefore, for Deepfake detection system, we report the EER and the FRR (using the threshold when $FAR = 10\%$) values on the *Test* set.

## 4. Analysis of deepfake videos

In this section, we evaluate the vulnerability of face VGG [16] and Facenet[6] [19] based recognition systems to videos with swapped faces and test several baseline detection systems.

### 4.1. Vulnerability of face recognition

We used publicly available pre-trained VGG and Facenet architectures for face recognition. 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 features vector of a test face. Both of these systems are state of the art recognition systems with VGG of $98.95\%$ [16] and Facenet of $99.63\%$ [19] 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 VidTIMIT[5] database. The results are presented in Table 1. 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 Figure 3. 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 [14].

Table 1: 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** | **VGG-based FR** | | **Facenet-based FR** | |
|---|---|---|---|---|
| **version** | **EER (%)** | **FAR (%)** | **EER (%)** | **FAR (%)** |
| LQ Deepfake | 0.03 | 88.75 | 0.00 | 94.38 |
| HQ Deepfake | 0.03 | 85.62 | 0.00 | 95.00 |

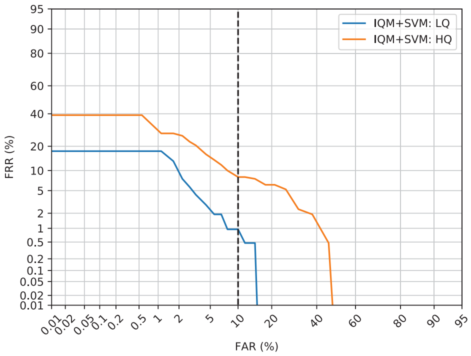


Figure 4: Performance of IQM+SVM detection on low (LQ) and high quality (HQ) Deepfakes.

### 4.2. Detection of Deepfake videos

We considered several baseline Deepfake detection systems, including system that uses audio-visual data to detect inconsistencies between lip movements and audio speech, as well as, several variations of purely image based systems.

The goal of the lip-sync based detection system is to distinguish genuine video, where lip movement and speech are synchronized, from tampered video, where lip movements and audio, which may not necessarily be speech, are not synchronized. The stages of such system include feature extraction from video and audio modalities, processing these features, and then, a two-class classifier trained to separate tampered videos from genuine. In this system, we used MFCCs as audio features [11] and distances between mouth landmarks as visual features (inspired by [20]). PCA is applied to the joint audio-visual features to reduce the dimensionality of the blocks of features and long short-term memory (LSTM) [5] network is trained to separate tampered and non-tampered videos as proposed in [9].

As image based systems, we implemented the following:

- *Pixels+PCA+LDA*: use raw faces as features with PCA-LDA classifier, with 99% retained variance resulting in 446 dimensions of transform matrix.

Table 2: Baseline detection systems for low (LQ) and high quality (HQ) Deepfake videos of VidTIMIT database. EER and FRR when FAR equal to 10% are computed on Test set.

| **Database** | **Detection system** | **EER (%)** | **FRR@FAR10% (%)** |
|---|---|---|---|
| LQ Deepfake | LSTM lip-sync [9] | 41.8 | 81.67 |
| | Pixels+PCA+LDA | 39.48 | 78.10 |
| | IQM+PCA+LDA | 20.52 | 66.67 |
| | IQM+SVM | 3.33 | 0.95 |
| HQ Deepfake | IQM+SVM | 8.97 | 9.05 |

- *IQM+PCA+LDA*: IQM features with PCA-LDA classifier with 95% retained variance resulting in 2 dimensions of transform matrix.

- *IQM+SVM*: IQM features with SVM classifier, each video has an averaged score from 20 frames.

The systems based on image quality measures (IQM) are borrowed from the domain of presentation (including replay attacks) attack detection, where such systems have shown good performance [7, 22]. As IQM feature vector, we used 129 measures of image quality, which include such measures like signal to noise ratio, specularity, bluriness, etc., by combining the features from [7] and [22].

The results for all detection systems are presented in Table 2. Figure 4 shows the detection error tradeoff (DET) curves for the best performing IQM+SVM system applied to two different versions of Deepfake videos. The results demonstrate that first, lip-syncing based algorithm is not able to detect face swapping, as GANs are able to generate facial expressions with high quality that can match audio speech. Therefore, currently, only image based approaches are capable to effectively detect Deepfake videos. Second, the IQM+SVM system has a reasonably high accuracy of detecting Deepfake videos, although videos generated with HQ model pose a more serious challenge. It means that a more advanced techniques for face swapping will be even more challenging to detect.

## 5. Conclusion

In this paper, we presented the first publicly available database of 620 Deepfake videos for 16 pairs of subjects from VidTIMIT database. We generated two versions of the videos for each subject: based on low quality 64 × 64 GAN model and higher quality 128 × 128 model. We also demonstrated that 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 evaluated several baseline face swap detection algorithms and found that lip-sync based approach fails to detect mismatches between lip movement and speech. The techniques based on image quality measures with SVM classifier can detect HQ Deepfake videos with $8.97\%$ equal error rate.

However, the continued advancements in development of face swapping techniques will result in more challenging Deepfake videos, which will be harder to detect by the existing algorithms. Therefore, new databases and new more generic detection methods need to be developed in the future. Possibly, a new arms race between Deepfake methods and detection algorithms has begun.

## Acknowledgements

This research was sponsored by Hasler Foundation's VERIFAKE project and the United States Air Force, Air Force Research Laboratory (AFRL) and the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-16-C-0170. The views, opinions and/or findings expressed are those of the author and are not representing the official views or policies of AFRL or DARPA.

## References

[1] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 659–665, Oct 2017.

[2] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[3] N. M. Arar, N. K. Bekmezci, F. Güney, and H. K. Ekenel. Real-time face swapping in video sequences: Magic mirror. In *IEEE Signal Processing and Communications Applications Conference (SIU)*, pages 825–828, April 2011.

[4] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):39:1–39:8, Aug. 2008.

[5] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453, July 2017.

[6] T. M. den Uyl, H. E. Tasli, P. Ivan, and M. Snijdewind. Who do you want to be? real-time face swap. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, May 2015.

[7] J. Galbally and S. Marcel. Face anti-spoofing based on general image quality assessment. In *International Conference on Pattern Recognition*, pages 1173–1178, Aug 2014.

[8] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, July 2017.

[9] P. Korshunov and S. Marcel. Speaker inconsistency detection in tampered video. In *European Signal Processing Conference (EUSIPCO)*, pages 2375–2379, Sept. 2018.

[10] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705, Oct 2017.

[11] N. Le and J.-M. Odobez. Learning multimodal temporal representation for dubbing detection in broadcast media. In *ACM Multimedia Conference (MM'16)*, pages 202–206, New York, USA, 2016.

[12] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv.org*, 2018.

[13] S. Mahajan, L. Chen, and T. Tsai. SwapItUp: A face swap application for privacy protection. In *IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pages 46–50, March 2017.

[14] A. Mohammadi, S. Bhattacharjee, and S. Marcel. Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. *IET Biometrics*, 7(1):15–26, 2018.

[15] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni. On face segmentation, face swapping, and face perception. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 98–105, May 2018.

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[17] H. X. Pham, Y. Wang, and V. Pavlovic. Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. *arXiv.org*, 2018.

[18] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv.org*, 2018.

[19] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

[20] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, July 2017.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, June 2016.

[22] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, April 2015.

[23] Z. Xingjie, J. Song, and J. Park. The image blending method for face swapping. In *IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 95–98, Sept 2014.

[24] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[25] Y. Zhang, L. Zheng, and V. L. L. Thing. Automated face swapping and its detection. In *IEEE International Conference on Signal and Image Processing (ICSIP)*, pages 15–19, Aug 2017.