Chapter 12
DeepFake Detection

Siwei Lyu

One particular disconcerting form of disinformation are the impersonating audios/videos backed by advanced AI technologies, in particular, deep neural networks (DNNs). These media forgeries are commonly known as the DeepFakes. The AI-based tools are making it easier and faster than ever to create compelling fakes that are challenging to spot. While there are interesting and creative applications of this technology, it can be weaponized to cause negative consequences. In this chapter, we survey the state-of-the-art DeepFake detection methods. We introduce the technical challenges in DeepFake detection and how researchers formulate solutions to tackle this problem. We discuss the pros and cons, as well as the potential pitfalls and drawbacks of each types of the solutions. Notwithstanding this progress, there are a number of critical problems that are yet to be resolved for existing DeepFake detection methods. We will also highlight a few of these challenges and discuss the research opportunities in this direction.

12.1 Introduction

Falsified images and videos created by AI algorithms, more commonly known as *DeepFakes*, are a recent twist to the disconcerting problem of online disinformation. Although fabrication and manipulation of digital images and videos are not new (Farid 2012), the rapid developments of deep neural networks (DNNs) in recent years have made the process of creating convincing fake images/videos increasingly easier and faster. DeepFake videos first caught the public's attention in late 2017, when

S. Lyu (✉)
Department of Computer Science and Engineering, University at Buffalo,
State University of New York, Buffalo, NY 14260, USA
e-mail: siweilyu@buffalo.edu

H. T. Sencar et al. (eds.), *Multimedia Forensics*, Advances in Computer Vision and Pattern
Recognition, https://doi.org/10.1007/978-981-16-7621-5_12

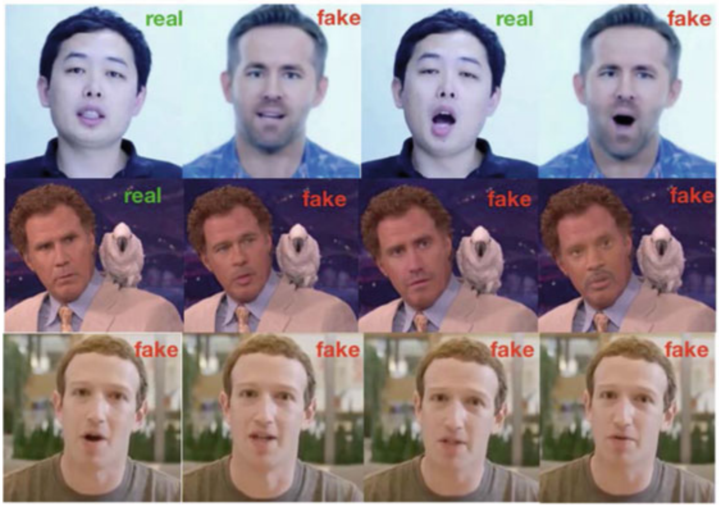

Fig. 12.1 Examples of DeepFake videos: (top) Head puppetry, (middle) face swapping, and (bottom) lip syncing

a Reddit account with the same name began posting synthetic pornographic videos generated using a DNN-based face-swapping algorithm. Subsequently, technologies that make DeepFakes have been mainstreamed through readily available software freely available on GitHub.[1] There are also emerging online services[2] and start-up companies also commercialized tools that that can generate DeepFake videos on demand.[3]

Currently, there are three major types of DeepFake videos.

- Head puppetry entails synthesizing a video of a target person's whole head and upper shoulder using a video of a source person's head, so the synthesized target appears to behave the same way as the source.
- Face swapping involves generating a video of the target with the faces replaced by synthesized faces of the source while keeping the same facial expressions.
- Lip syncing is to create a falsified video by only manipulating the lip region so that the target appears to speak something that s/he does not speak in reality.

Figure 12.1 shows some example frames of each type of DeepFake videos aforementioned.

While there are interesting and creative applications of DeepFake videos, due to the strong association of faces to the identity of an individual, they can also be

[1] E.g., `FakeApp` (FakeApp 2020), `DFaker` (DFaker github 2019), `faceswap-GAN` (faceswap-GAN github 2019), `faceswap` (faceswap github 2019), and `DeepFaceLab` (DeepFaceLab github 2020).

[2] E.g., https://deepfakesweb.com.

[3] E.g., Synthesia (https://www.synthesia.io/) and Canny AI https://www.cannyai.com/.

weaponized. Well-crafted DeepFake videos 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 (Chesney and Citron 2019). The potential threats range from revenge pornographic videos of a victim whose face is synthesized and spliced in, to realistically looking videos of state leaders seeming to make inflammatory comments they never actually made, a high-level executive commenting about her company's performance to influence the global stock market, or an online sex predator masquerades visually as a family member or a friend in a video chat.

Since the first known case of DeepFake videos were reported in December 2017,[4] the mounting concerns over the negative impacts of DeepFakes have spawned an increasing interest in DeepFake detection in the Multimedia Forensics research community, and the first dedicated DeepFake detection algorithm was developed by my group in June 2018 (Li et al. 2018). Subsequently, there are avid developments with many DeepFake detection methods developed in the past few years, e.g., Li et al. (2018, 2019a, b), Li and Lyu (2019) Yang et al. (2019), Matern et al. (2019), Ciftci et al. (2020), Afchar et al. (2018), Güera and Delp (2018a, b), McCloskey and Albright (2018), Sabir et al. (2019), Rössler et al. (2019), Nguyen et al. (2019a, b, c), Nataraj et al. (2019), Xuan et al. (2019), Jeon et al. (2020), Mo et al. (2018), Liu et al. (2020), Fernando et al. (2019), Shruti et al. (2020), Koopman et al. (2018), Amerini et al. (2019), Chen and Yang (2020), Wang et al. (2019a), Guarnera et al. (2020), Durall et al. (2020), Frank et al. (2020), Ciftci and Demir (2019), Chen and Yang (2020), Stehouwer et al. (2019), Bonettini et al. (2020), Khalid and Woo (2020). Correspondingly, there have also been several large-scale benchmark datasets to evaluate DeepFake detection performance (Yang et al. 2019; Korshunov and Marcel 2018; Rössler et al. 2019; Dufour et al. 2019; Dolhansky et al. 2019; Ciftci and Demir 2019; Jiang et al. 2020; Wang et al. 2019b; Khodabakhsh et al. 2018; Stehouwer et al. 2019). DeepFake detection has also been supported by government funding agencies and private companies alike.[5]

A climax of these efforts is the first DeepFake Detection Challenge from late 2019 to early 2020 (Deepfake detection challenge 2021). Overall, the DeepFake Detection Challenge corresponds to the state-of-the-art, with the winning solutions being a tour de force of advanced DNNs (an average precision of 82.56% by the top performer). These provide us effective tools to expose DeepFakes that are automated and mass produced by AI algorithms. However, we need to be cautious in reading these results. Although the organizers have made their best effort to simulate situations that DeepFake videos are deployed in real life, there is still a significant discrepancy between the performance on the evaluation dataset and a more real dataset—when tested on unseen videos, the top performer's accuracy reduced to 65.18%. In addition, all solu-

[4] https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn.

[5] Detection of DeepFakes is one of the goals of the DARPA MediFor (2016–2020), which also sponsored the NIST MFC 2018 and 2020 Synthetic Data Detection Challenge (NIST MFC 2018), and SemaFor (2020–2024) programs. The Global DeepFake Detection Challenge (Deepfake detection challenge 2021) in 2020 was sponsored by leading tech companies including Facebook, Amazon, and Microsoft.

tions are based on clever designs of DNNs and data augmentations, but provide little insight beyond the "black box"-type classification algorithms. Furthermore, these detection results may not completely reflect the actual detection performance of the algorithm on a single DeepFake video, especially ones that have been manually processed and perfected after being generated from the AI algorithms. Such "crafted" DeepFake videos are more likely to cause real damages, and careful manual post processing can reduce or remove artifacts that the detection algorithms predicate on.

In this chapter, we survey the state-of-the-art DeepFake detection methods. We introduce the technical challenges in DeepFake detection and how researchers formulate solutions to tackle this problem. We discuss the pros and cons, as well as the potential pitfalls and drawbacks of each types of the solutions. We also provide an overview of research efforts for DeepFake detection and a systematic comparison of existing datasets, methods, and performances. Notwithstanding this progress, there are a number of critical problems that are yet to be resolved for existing DeepFake detection methods. We will also highlight a few of these challenges and discuss the research opportunities in this direction.

12.2 DeepFake Video Generation

Although in recent years there have been many sophisticated algorithms for generating realistic synthetic face videos (Bitouk et al. 2008; Dale et al. 2011; Suwajanakorn et al. 2015, 2017; Thies et al. 2016; Korshunova et al. 2017; Pham et al. 2018; Karras et al. 2018a, 2019; Kim et al. 2018; Chan et al. 2019), 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 (Liu et al. 2017) that becomes the *tool of choice* to create DeepFake videos in scale, with several independent open-source implementations. 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. 12.2 (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 (Kazemi and Sullivan 2014). The aligned faces are then cropped and fed to an auto-encoder (Kingma and Welling 2014) 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 convoluntional 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. 12.2 (right). Specifically, an encoder-

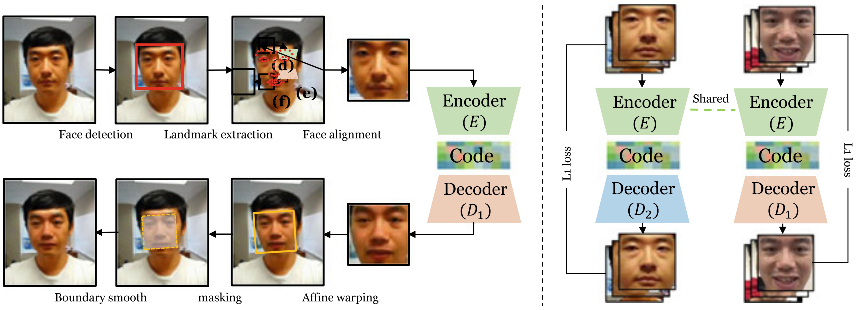

Fig. 12.2 Synthesis (left) and training (right) of the basic DeepFake maker algorithm. See texts for more details

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.

12.3 Current DeepFake Detection Methods

In this section, we provide a brief overview of current DeepFake detection methods. As this is an actively researched area with the literature growing rapidly, we cannot promise comprehensive coverage of all existing methods. Furthermore, as the focus here is on DeepFake videos, we exclude detection methods for GAN-generated images. Our objective is to summarize the existing detection methods at a more abstract level, pointing out some common characteristics and challenges that can benefit future developments of more effective detection methods.

12.3.1 General Principles

As detection of DeepFakes is a problem in Digital Media Forensics, it complies with three general principles of Digital Media Forensics.

- **Principle 1***: Manipulation operations leave traces in the falsified media.*

Digitally manipulated or synthesized media (including DeepFakes) are created by a process other than a capturing device recording an event actually occurs in the physical world. This fundamental difference in the creation process will be reflected in the resulting falsified media, albeit revealed with different scales depending on the amount of changes to the original media. Based on this principle, we can conclude that DeepFakes are detectable.

- **Principle 2**: *A single forensic measure can be circumvented.*
  Any forensic detection method is based on differentiating the real and falsified media on certain characteristics in the media signals. However, the same characteristics can be exploited to evade the very forensic detections, either by hiding the traces or disrupting them. This principle is the premise for anti-forensic measures for DeepFake detections.
- **Principle 3**: *There is an intention behind every falsified media.*
  A falsified media in the wild is made for a reason. It could be a satire or a prank but also a malicious attack to the victim's reputation and credibility. Understanding the motivation behind the DeepFake can provide richer information to the detection of DeepFakes, as well as to prevent and mitigate the damages.

12.3.2 Categorization Based on Methodology

We categorize existing DeepFake detection methods into three types. The first two work by seeking specific artifacts in DeepFake videos that differ them from the real videos (Principle 1 of Sect. 12.3.1).

12.3.2.1 Signal Feature-Based Methods

The **signal feature-based methods** (e.g., Afchar et al. 2018; Güera and Delp 2018b; McCloskey and Albright 2018; Li and Lyu 2019) look for abnormalities at the signal level, treating the videos as a sequence of frames $f(x, y, t)$, and a synchronized audio signal $a(t)$ if the audio track exists. Such abnormalities are often caused by various steps in the processing pipeline of DeepFake generation.

For instance, our recent work (Li and Lyu 2019) exploits the signal artifacts introduced by the resizing and interpolation operations during the post-processing of DeepFake generation. Similarly, the work in Li et al. (2019a) focuses on the boundary when the synthesized face regions are blended into the original frame. In Frank et al. (2020), the authors observe that synthesized faces often exhibit abnormalities in high frequencies, due to the up-sampling operation. Based on this observation, a simple detection method in the frequency domain is proposed. In Durall et al. (2020), the detection method uses a classical frequency domain analysis followed by an SVM classifier. The work in Guarnera et al. (2020) uses the EM algorithm to extract a set of local features in the frequency domain that are used to distinguish frames of DeepFakes from those of the real videos.

The main advantage of signal feature-based methods is that the abnormalities are usually fundamental to generation process, and fixing them may require significant changes to the underlying DNN model or post-processing steps. On the other hand, the reliance on signal features also means that signal feature-based DeepFake detection methods are susceptible to disturbance to the signals, such as interpolation (rotation, up-sizing), down-sampling (down-sizing), additive noise, blurring, and compression.

12.3.2.2 Physical/Physiological-Based Methods

The **physical/physiological-based methods** (e.g., Li et al. 2018; Yang et al. 2019; Matern et al. 2019; Ciftci et al. 2020; Hu et al. 2020) expose DeepFake videos based on their violations to the fundamental laws of physics or human physiology. The DNN model synthesizing human faces do not have direct knowledge about physiological traits of human faces or the physical laws of the surrounding environment, such information may be incorporated into the model indirectly (and inefficiently) through training data. This lack of knowledge can lead to detectable traits that are also intuitive to humans. For instance, the first dedicated DeepFake detection method (Li et al. 2018) works by detecting the inconsistency or lack of realistic eye blinking in DeepFake videos. This is due to the training data for the DeepFake generation model, which are often obtained from the Internet as the portrait images of a subject. As the portrait images are dominantly those with the subject's eyes open, the DNN generation models trained using them cannot reproduce closed eyes for a realistic eye blinking. The work of Yang et al. (2019) use the physical inconsistencies in the head poses of the DeepFake videos due to splicing the synthesized face region into the original frames. The work of Matern et al. (2019) summarizes various appearance artifacts that can be observed in DeepFake videos, such as the inconsistent eye colors, missing reflections, and fuzzy details in the eye and teeth areas. The authors then propose a set of simple features for for DeepFake detection. The work in Ciftci et al. (2020) introduces a DeepFake detection method based on biological signals extracted from facial regions that are invisible to human eyes in portrait videos, such as the slight color changes due to the blood flow caused by heart beats. These biological signals are used to reveal the spatial coherence and temporal consistency.

The main advantage of physical/physiological-based methods is its superior intuitiveness and explainability. This is especially important when the detection results are used by practitioners such as journalists. On the other hand, physical/physiological-based detection methods are limited by the effectiveness and robustness of the underlying Computer Vision algorithms, and their strong reliance on the semantic cues also limits their applicability.

12.3.2.3 Data-Driven Methods

The third category of detection methods are **data-driven methods** (e.g., Sabir et al. 2019; Rössler et al. 2019; Nguyen et al. 2019a, b, c; Nataraj et al. 2019; Xuan et al. 2019; Wang et al. 2019b; Jeon et al. 2020; Mo et al. 2018; Liu et al. 2020; Li et al. 2019a, b; Güera and Delp 2018a; Fernando et al. 2019; Shruti et al. 2020; Koopman et al. 2018; Amerini et al. 2019; Chen and Yang 2020; Wang et al. 2019a; Guarnera et al. 2020; Durall et al. 2020; Frank et al. 2020; Ciftci and Demir 2019; Chen and Yang 2020; Stehouwer et al. 2019; Bonettini et al. 2020; Khalid and Woo 2020), which do not target at specific features, but use videos labeled as real or DeepFake to train ML models (oftentimes DNNs) that can differentiate DeepFake videos from the real ones. Note that feature-based methods can also use DNN classifiers, so what makes the data-driven methods different is that the cues for classifying the two types of videos are implicit and found by the ML models. As such, the success of data-driven methods largely depends on the quality and diversity of training data, as well as the design of the ML models.

Early data-driven DeepFake detection methods reuse standard DNN models (e.g., VGG-Net Do et al. 2018, XceptionNet Rössler et al. 2019) designed for other Computer Vision tasks such as object detection and recognition. There also exist methods that use more specific or novel network architectures. MesoNet (Afchar et al. 2018) is an early detection method that uses a customized DNN model to detect DeepFakes, and interprets the detection results by correlating them with some mesoscopic level image features. The DeepFake detection method of Liu et al. (2020) uses the Gram-Net to capture differences in global image texture, which has the additional “Gram blocks” in the conventional CNN models that can calculate the Gram matrices at different levels of the network. The detection methods in Nguyen et al. (2019c, b) use capsule networks and show that it can achieve similar performance but with fewer parameters than the CNN models.

It should be mentioned that, to date, the state-of-the-art performance in DeepFake detection is obtained with the data-drive methods based on large-scale DeepFake datasets and innovative model design and training procedures. For instance, the top performer of the DeepFake Detection Challenge[6] is based on the use of an ensemble DNN models that entails seven different EfficientNet models Tan and Le (2019). The runner-up method[7] also uses ensemble models that includes EfficientNet and XceptionNet models but also introduce a new data augmentation methods (WSDAN) to anticipate the different configurations of faces.

[6] https://github.com/selimsef/dfdc_deepfake_challenge.

[7] https://github.com/cuihaoleo/kaggle-dfdc.

12.3.3 Categorization Based on Input Types

Another way to categorize existing DeepFake detection methods can be obtained by the type of the inputs. Most existing DeepFake detection methods are based on binary classification at the frame level, i.e., determining the likelihood of an individual frame as real or of DeepFake. Although simple and easy to implement, there are two issues related with the frame-based detection methods. First, the temporal consistency among frames are not explicitly considered, as (i) many DeepFake videos exhibit temporal artifacts and (ii) real or DeepFake frames tend to appear in continuous intervals. Second, it necessitates an extra step when video-level integrity score is needed: we have to aggregate the scores over individual frames to compute such a score using certain types of aggregation rules, common choices of which include the average, the maximum, or the average of top range. Temporal methods (e.g., Sabir et al. 2019; Amerini et al. 2019; Koopman et al. 2018), on the other hand, take the whole frame sequences as input, and use the temporal correlation between the frames as an intrinsic feature. Temporal methods often uses sequential models such as RNNs as basic model structure, and directly output the videos level prediction. The audio-visual detection methods also use the audio track as an input, and detect DeepFakes based on the asynchrony between the audios and frames.

12.3.4 Categorization Based on Output Types

Regardless of the underlying methodology or types of input, the majority of existing DeepFake detection methods formulate the problem as binary classification, which returns a label for each input video that signifies if the input is a real or a DeepFake. Often, the predicted labels are accompanied with a confidence score, a real value in [0, 1], which may be interpreted as the probability of the input to belong to one of the classes (i.e., real or DeepFake). A few methods extend this to a multi-class classification problem, where the labels also reflect different types of DeepFake generation models. There are also methods that solve the location problem (e.g., Li et al. 2019b; Huang et al. 2020), which further identifies the spatial area (in the form of bounding boxes or masked out region) and time interval of DeepFake treatments.

12.3.5 The DeepFake-o-Meter Platform

Unfortunately, the plethora of DeepFake detection algorithms were not taken full advantage of. On the one hand, differences in training datasets, hardware, and learning architectures across research publications make rigorous comparisons of different detection algorithms challenging. On the other hand, the cumbersome process of downloading, configuring, and installing of individual detection algorithms deny

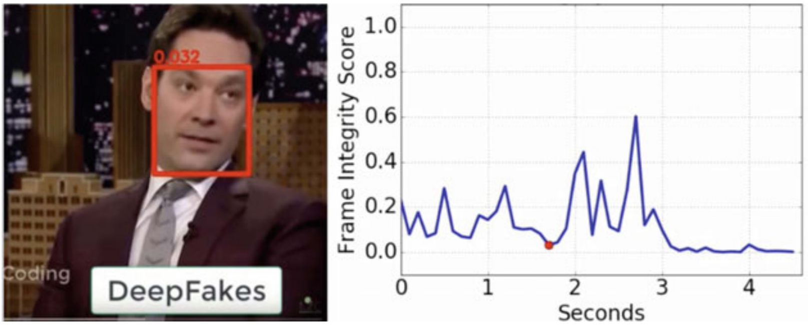

Fig. 12.3 Detection results on DeepFake-o-meter of a state-of-the-art DeepFake detection method (Li and Lyu 2019) over a fake video on youtube.com. The lower integrity score (range in [0, 1]) suggests a video frame more likely to be generated using DeepFake algorithms

the access of the state-of-the-art DeepFake detection methods to most users. To this end, we have developed `DeepFake-o-meter` http://zinc.cse.buffalo.edu/ubmdfl/deepfake-o-meter/, an open platform for DeepFake detections. For developers of DeepFake detection algorithms, it provides an API architecture to wrap individual algorithms and run on a third-party remote server. For researchers, it is an evaluation/benchmarking platform to compare multiple algorithms on the same input. For users, it provides a convenient portal to use multiple state-of-the-art detection algorithms. Currently, we have incorporated 11 state-of-the-art DeepFake image and video detection methods. A sample analysis result is shown in Fig. 12.3.

12.3.6 Datasets

The availability of large-scale datasets of DeepFake videos is an enabling factor to the development of DeepFake detection methods. The first DeepFake dataset UADFV (Yang et al. 2019) only has 49 DeepFake videos with visible artifacts when it was released in June 2018. Subsequently, more DeepFake datasets are proposed with increasing quantities and qualities. Several examples of synthesized DeepFake video frames are shown in Fig. 12.4.

- **UADFV** (Yang et al. 2019): This dataset contains 49 real videos downloaded from Youtube, which were used to create 49 DeepFake videos.
- **DeepfakeTIMIT** (Korshunov and Marcel 2018): The original videos in this dataset is from the VidTIMIT database, from which a total of 640 DeepFake videos were generated.
- **FaceForensics++** (Rössler et al. 2019): FaceForensics++ is a forensics dataset consisting of 1,000 original video sequences that have been manipulated with four automated face manipulation methods: DeepFakes, Face2Face, FaceSwap and NeuralTextures.

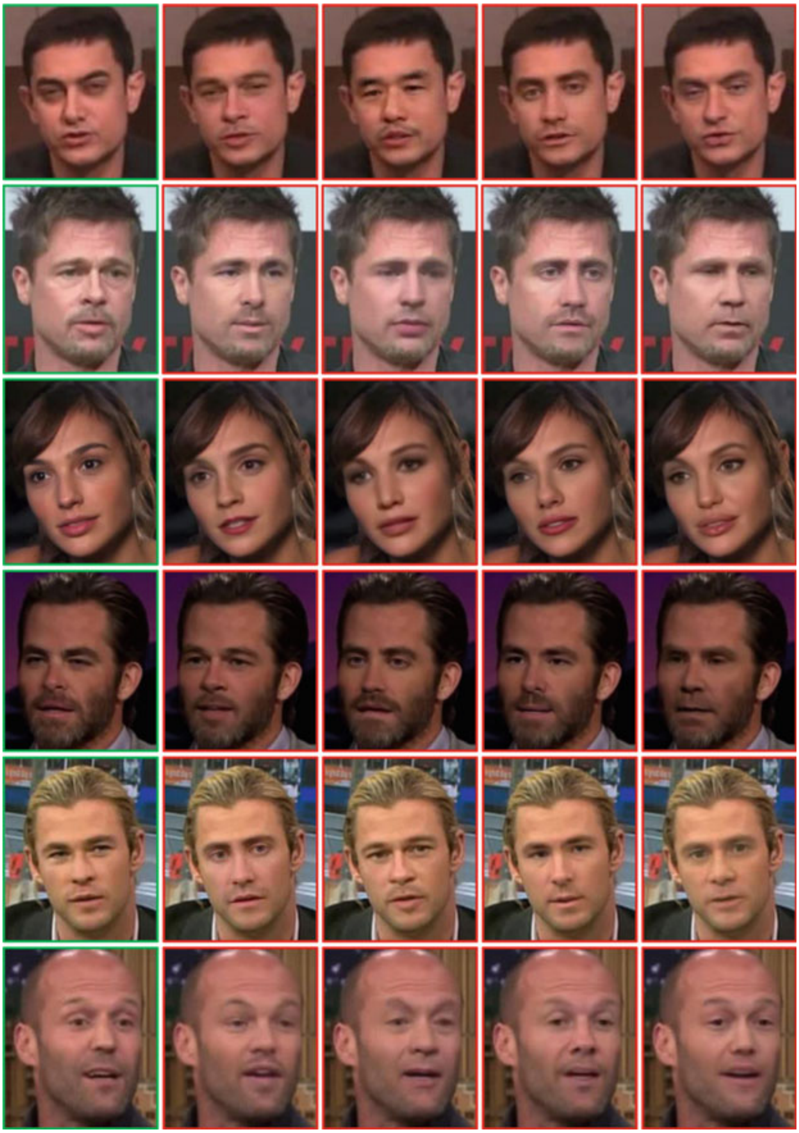

Fig. 12.4 Example frames of DeepFake videos from the Celeb-DF dataset (Li et al. 2020). The left most column corresponds to frames from the original videos, and the other columns are frame from DeepFake videos swapped with synthesized faces

- **DFD**: The Google/Jigsaw DeepFake detection dataset has 3,068 DeepFake videos generated based on 363 original videos of 28 consented individuals of various genders, ages, and ethnic groups.
- **DFDC** (Dolhansky et al. 2019): The Facebook DeepFake Detection Challenge (DFDC) Dataset 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.
- **Celeb-DF** (Li et al. 2020): The Celeb-DF dataset contains real and DeepFake synthesized videos having similar visual quality on par with those circulated online. It includes 590 original videos collected from YouTube with subjects of different ages, ethic groups and genders, and 5,639 corresponding DeepFake videos.
- **DeeperForensics-1.0** (Jiang et al. 2020): DeeperForensics-1.0 consists of 10,000 DeepFake videos generated by a new DNN-based face swapping framework.
- **DFFD** (Stehouwer et al. 2019): The DFFD dataset contains 3,000 videos of four types of digital manipulations as identity swap, expression, swap, attribute manipulation, and entire synthesized faces.

12.3.7 Challenges

Albeit impressive progress has been made in the performance of detection of DeepFake videos, there are several concerns over the current detection methods that suggest caution.

Performance Evaluation. Currently, the problem of detecting DeepFake videos is commonly formulated, solved, and evaluated as a binary classification problem, where each video is categorized as real or a DeepFake. Such dichotomy is easy to set up in controlled experiments, where we develop and test DeepFake detection algorithms using videos that are either pristine or made with DeepFake generation algorithms. However, the picture is murkier when the detection method is deployed in real world. For instance, videos can be fabricated or manipulated in ways other than DeepFakes, so not being detected as a DeepFake video does not necessarily suggest the video is a real one. Also, a DeepFake video may be subject to other types of manipulations and a single label may not comprehensively reflect such. Furthermore, in a video with multiple subjects' faces only one or a few are generated with DeepFake for a fraction of the frames. So the binary classification scheme needs to be extended to multi-class, multi-label, and local classification/detection to fully handle the complexities of real-world media forgeries.

Explainability of Detection Results. Current DeepFake detection methods are mostly designed to perform batch analysis over a large collection videos. However, when the detection methods are used in the field by journalists or law enforcement, we usually need only to analyze a small number of videos. Numerical score corresponding to the likelihood of a video being generated using a synthesis algorithm is not as useful to the practitioners if it is not corroborated with proper reasoning of the score. In such scenarios, it is very typical to request a justification for the numerical score for the analysis to be acceptable for publishing or used in court. However, many data-driven Deepfake detection methods, especially those based on the use of deep neural networks, usually lack explainability due to the black box nature of the DNN models.

Generalization Across Datasets. As a DNN-based DeepFake detection method needs to be trained on a specific training dataset, a lack of generalization has been observed for DeepFake videos created using models not represented in the training dataset. This is different from overfitting, as the learned model may perform well on testing videos that are created with the same model but not used in training. The problem is caused by "domain shifting" when the trained detection method is applied to DeepFake videos generated using different generation models. A simple solution is to enlarge the training set to represent more diverse generation models, but a more flexible approach is needed to scale up to previously unseen models.

Social Media Laundering. A large fraction of online videos are now spread through social networks, e.g., FaceBook, Instagram, and Twitter. To save network bandwidth and also to protect the users' privacy, these videos are usually striped off meta-data, down-sized, and then heavy compressed before they are uploaded to the social platforms. These operations, commonly known as *social media laundering*, are detrimental to recover traces of underlying manipulation, and at the same time increase the false positive detections, i.e., classifying a real video as a DeepFake. So far, most data-driven DeepFake detection methods that use signal-level features are much affected by social media laundering. A practical measure to improve the robustness of DeepFake detection methods to social media laundering is to actively incorporate simulations of such effects in training data, and also enhance evaluation datasets to include performance on social media laundered videos, both real and synthesized.

Anti-forensics. With the increasing effectiveness of DeepFake detection methods, we also anticipate developments of corresponding anti-forensic measures, which take advantage of the vulnerabilities of current DeepFake detection methods to conceal revealing traces of DeepFake videos. The data-driven DNN-based DeepFake detection methods are particularly susceptible to anti-forensic attacks due to the known vulnerability of general deep neural network classification models (see Principle 2 of Sect. 12.3.1). Indeed, recent years have witnessed a rapid development of anti-forensic methods based on adversarial attacks to DNN models targeting DeepFake detectors (Huang et al. 2020; Carlini and Farid 2020; Gandhi and Jain 2020; Neekhara 2020). Anti-forensic measures can also be developed in the other aspect, to disguise a real video as a DeepFake video by adding simulated signal level features used by current detection algorithms, a situation we term as *fake DeepFake*. Further, DeepFake detection methods must improve to handle such intentional and adversarial attacks.

12.4 Future Directions

Besides continuing improving to solve the aforementioned limitations, we also envision a few important directions of DeepFake detection methods that will receive more attention in the coming years.

Other Forms of DeepFake Videos. Although face swapping is currently the most widely known form of DeepFake videos, it is by no means the most effective. In particular, for the purpose of impersonating someone, face swapping DeepFake videos have several limitations. Psychological studies (Sinha et al. 2009) show that human face recognition largely relied on information gleaned from face shape and hairstyle. As such, to create convincing impersonating effect, the person whose face is to be replaced (the target) has to have similar face shape and hairstyle to the person whose face is used for swapping (the donor). Second, as the synthesized faces need to be spliced into the original video frame, the inconsistencies between the synthesized region and the rest of the original frame can be severe and difficult to conceal. In these respects, other forms of DeepFake videos, namely, head puppetry and lip-syncing, are more effective and thus should become the focus of subsequent research in DeepFake detection. Methods studying whole face synthesis or reenactment have experienced fast development in recent years. Although there have not been as many easy-to-use and free open-source software tools generating these types of DeepFake videos as for the face-swapping videos, the continuing sophistication of the generation algorithms will change the situation in the near future. Because the synthesized region is different from face swapping DeepFake videos (the whole face in the former and lip area in the latter), detection methods designed based on artifacts specific to face swapping are unlikely to be effective for these videos. Correspondingly, we should develop detection methods that are effective to these types of DeepFake videos.

Audio DeepFakes. AI-based impersonation are not limited to imagery, recent AI-synthesized content-generation are leading to the creation of highly realistic audios (Ping et al. 2017; Gu and Kang 2018; AlBadawy and Lyu 2020). Using synthesized audios of the impersonating target can significantly make the DeepFake videos more convincing and compounds its negative impact. As audio signals are 1D signals and have very different nature from images and videos, different methods need to be developed to specifically target such forgeries. This problem has drawn attention in the speech processing community recently with part of the most recent Global ASVspoofing Challenge (https://www.asvspoof.org/). Dedicated to AI-driven voice conversion detection, and a few dedicated methods for audio DeepFake detection, e.g., AlBadawy et al. (2019), have also shown up recently. In the coming years, we expect more developments in these areas, in particular, those can leverage features in both visual and audio features of the fake videos.

Intent Inference. Even though the potential negative impacts of DeepFake videos are tremendous, in reality, the majority of DeepFake videos are not created not with a malicious intent. Many DeepFake videos currently circulated online are of a pranksome, humorous, or satirical nature. As such, it is important to expose the underlying intent of a DeepFake in the context of legal or journalistic investigation (Principle 3 in Sect. 12.3.1). Inferring intention may require more semantic and contextual understanding of the content, few forensic methods are designed to answer this question, but this is certainly a direction that future forensic methods will focus on.

Human Performance. Although the potential negative impacts of online DeepFake videos are widely recognized, currently there is a lack of formal and quantitative study of the perceptual and psychological factors underlying their deceptiveness. Interesting questions such as if there exist an *uncanny valley*[8] for DeepFake videos, what is the *just noticeable difference* between high-quality DeepFake videos and real videos to human eyes, or what type/aspects of DeepFake videos are more effective in deceiving the viewers, have yet to be answered. To pursue these questions, it calls for close collaboration among researchers in digital media forensics and in perceptual and social psychology. There is no doubt that such studies are invaluable to research in detection techniques as well as a better understanding of the social impact that DeepFake videos can cause.

Protection measures. However, given the speed and reach of the propagation of online media, even the currently best forensic techniques will largely operate in a postmortem fashion, applicable only after AI synthesized fake face images or videos emerge. We aim to develop *proactive* approaches to protect individuals from becoming the victims of such attacks, which complement to the forensic tools.

One such method we have recently studied (Li et al. 2019c; Sun et al. 2020) is to add specially designed patterns known as the *adversarial perturbations* that are imperceptible to human eyes but can result in detection failures. The rationale is as follows. High-quality AI face synthesis models need large number of, typically in the range of thousands, sometimes even millions, training face images collected using automatic face detection methods, i.e., the *face sets*. Adversarial perturbations "pollute" a face set to have few actual faces and many non-faces with low or no utility as training data for AI face synthesis models. The proposed adversarial perturbation generation method can be implemented as a service of photo/video sharing platforms before a user's personal images/videos are uploaded or as a standalone tool that the user can use, to process the images and videos before they are uploaded online.

12.5 Conclusion and Outlook

We predict that several future technological developments will further improve the visual quality and generation efficiency of the fake videos. Firstly, one critical disadvantage of the current DeepFake generation methods are that they cannot produce good details such as skin and facial hairs. This is due to the loss of information in the encoding step of generation. However, this can be improved by incorporating GAN models (Goodfellow et al. 2014) which have demonstrated performance in recov-

[8] The uncanny valley in this context refers to the phenomenon whereby a DeepFake generated face bearing a near-identical resemblance to a human being arouses a sense of unease or revulsion in the viewers.

ering facial details in recent works (Karras et al. 2018b, 2019, 2020). Secondly, the synthesized videos can be more realistic if they are accompanied with realistic voices, which combines video and audio synthesis together in one tool.

In the face of this, the overall running efficiency, detection accuracy, and more importantly, false positive rate, have to be improved for wide practical adoption. The detection methods also need to be more robust to real-life post-processing steps, social media laundering, and counter-forensic technologies. There is a perpetual competition of technology, know-hows, and skills between the forgery makers and digital media forensic researchers. The future will reckon the predictions we make in this work.

References

Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: WIFS

AlBadawy E, Lyu S (2020) Voice conversion using speech-to-speech neuro-style transfer. In: Interspeech, Shanghai, China

AlBadawy E, Lyu S, Farid H (2019) Detecting ai-synthesized speech using bispectral analysis. In: Workshop on media forensics (in conjunction with CVPR), Long Beach, CA, United States

Amerini I, Galteri L, Caldelli R, Del Bimbo A (2019) Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE international conference on computer vision workshops, pp 0-0

Bitouk D, Kumar N, Dhillon S, Belhumeur P, Nayar SK (2008) Face swapping: automatically replacing faces in photographs. ACM Trans Graph (TOG)

Bonettini N, Cannas ED, Mandelli S, Bondi L, Bestagini P, Tubaro S (2020) Video face manipulation detection through ensemble of cnns. arXiv:2004.07676

Carlini N, Farid H (2020) Evading deepfake-image detectors with white- and black-box attacks. arXiv:2004.00622

Chan C, Ginosar S, Zhou T, Efros AA (2019) Everybody dance now. In: ICCV

Chen Z, Yang H (2020) Manipulated face detector: joint spatial and frequency domain attention network. arXiv:2005.02958

Chesney R, Citron DK (2019) Deep Fakes: a looming challenge for privacy, democracy, and national security. In: 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

Ciftci UA, Demir I (2019) Fakecatcher: detection of synthetic portrait videos using biological signals. arXiv:1901.02212

Ciftci UA, Demir I, Yin L (2020) How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. In: IEEE/IAPR international joint conference on biometrics (IJCB)

Dale K, Sunkavalli K, Johnson MK, Vlasic D, Matusik W, Pfister H (2011) Video face replacement. ACM Trans Graph (TOG)

DeepFaceLab github. https://github.com/iperov/DeepFaceLab. Accessed 4 July 2020

Deepfake detection challenge. https://deepfakedetectionchallenge.ai

DFaker github. https://github.com/dfaker/df. Accessed 4 Nov 2019

Do N-T, Na I-S, Kim S-H (2018) Forensics face detection from gans using convolutional neural network

Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The deepfake detection challenge (DFDC) preview dataset. arXiv:1910.08854

Dufour N, Gully A, Karlsson P, Vorbyov AV, Leung T, Childs J, Bregler C (2019) Deepfakes detection dataset by google & jigsaw

Durall R, Keuper M, Keuper J (2020) Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. arXiv:2003.01826

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

faceswap-GAN github. https://github.com/shaoanlu/faceswap-GAN. Accessed 4 Nov 2019

FakeApp. https://www.malavida.com/en/soft/fakeapp/. Accessed 4 July 2020

Farid H (2012) Digital image forensics. MIT Press, Cambridge

Fernando T, Fookes C, Denman S, Sridharan S (2019) Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks. Computer vision and pattern recognition. arXiv:1911.07844

Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T (2020) Leveraging frequency analysis for deep fake image recognition. arXiv:2003.08685

Gandhi A, Jain S (2020) Adversarial perturbations fool deepfake detectors. arXiv:2003.10596

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: NeurIPS

Guarnera L, Battiato S, Giudice O (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops

Güera D, Delp EJ (2018a) Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–6

Güera D, Delp EJ (2018b) Deepfake video detection using recurrent neural networks. In: AVSS

Gu Y, Yongguo K (2018) Multi-task WaveNet: a multi-task generative model for statistical parametric speech synthesis without fundamental frequency conditions. In: Interspeech, Hyderabad, India

Huang Y, Juefeixu F, Wang R, Xie X, Ma L, Li J, Miao W, Liu Y, Pu G (2020) Fakelocator: robust localization of gan-based face manipulations via semantic segmentation networks with bells and whistles. Computer vision and pattern recognition. arXiv:2001.09598

Hu S, Li Y, Lyu S (2009) Exposing GAN-generated faces using inconsistent corneal specular highlights. arXiv:11924:2020

Jeon H, Bang Y, Woo SS (2020) Fdftnet: facing off fake images using fake detection fine-tuning network. Computer vision and pattern recognition. arXiv:2001.01265

Jiang L, Wu W, Li R, Qian C, Loy CC (2020) Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection. arXiv:2001.03024

Karras T, Aila T, Laine S, Lehtinen J (2018a) Progressive growing of GANs for improved quality, stability, and variation. In: ICLR

Karras T, Aila T, Laine S, Lehtinen J (2018b) Progressive growing of GANs for improved quality, stability, and variation. In: International conference on learning representations (ICLR)

Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: CVPR

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8110–8119

Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: CVPR

Khalid H, Woo SS (2020) Oc-fakedect: classifying deepfakes using one-class variational autoencoder. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops

Khodabakhsh A, Ramachandra R, Raja KB, Wasnik P, Busch C (2018) Fake face detection methods: can they be generalized?, pp 1–6

Kim H, Garrido P, Tewari A, Xu W, Thies J, Nießner N, Pérez P, Richardt C, Zollhöfer M, Theobalt C (2018) Deep video portraits. ACM Trans Graph (TOG)

Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: ICLR

Koopman M, Rodriguez AM, Geradts Z (2018) Detection of deepfake video manipulation. In: The 20th Irish machine vision and image processing conference (IMVIP), pp 133–136

Korshunova I, Shi W, Dambre J, Theis L (2017) Fast face-swap using convolutional neural networks. In: ICCV

Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? Assessment and detection. arXiv:1812.08685

Li Y, Lyu S (2019) Exposing deepfake videos by detecting face warping artifacts. In: IEEE conference on computer vision and pattern recognition workshops (CVPRW)

Li Y, Chang M-C, Lyu S (2018) In Ictu Oculi: exposing AI generated fake face videos by detecting eye blinking. In: IEEE international workshop on information forensics and security (WIFS)

Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B (2019a) Face x-ray for more general face forgery detection. arXiv:1912.13458

Li J, Shen T, Zhang W, Ren H, Zeng D, Mei T (2019b) Zooming into face forensics: a pixel-level analysis. Computer vision and pattern recognition. arXiv:1912.05790

Li Y, Yang X, Wu B, Lyu S (2019c) Hiding faces in plain sight: disrupting ai face synthesis with adversarial perturbations. arXiv:1906.09288

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 patten recognition (CVPR), Seattle, WA, United States

Liu M-Y, Breuel T, Kautz J (2017) Unsupervised image-to-image translation networks. In: NeurIPS

Liu Z, Qi X, Jia J, Torr P (2020) Global texture enhancement for fake face detection in the wild. arXiv:2002.00133

Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE winter applications of computer vision workshops (WACVW)

McCloskey S, Albright M (2018) Detecting gan-generated imagery using color cues. arXiv:1812.08247

Mo H, Chen B, Luo W (2018) Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM workshop on information hiding and multimedia security, pp 43–47

Nataraj L, Mohammed TM, Manjunath BS, Chandrasekaran S, Flenner A, Bappy JH, Roy-Chowdhury AK (2019) Detecting gan generated fake images using co-occurrence matrices. Electron Imag (2019)5:532–1

Neekhara P (2020) Adversarial deepfakes: evaluating vulnerability of deepfake detectors to adversarial examples. arXiv:2002.12749

Nguyen HH, Fang F, Yamagishi J, Echizen I (2019a) Multi-task learning for detecting and segmenting manipulated facial images and videos. In: IEEE international conference on biometrics: theory, applications and systems (BTAS)

Nguyen HH, Yamagishi J, Echizen I (2019b) Capsule-forensics: using capsule networks to detect forged images and videos. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2307–2311

Nguyen HH, Yamagishi J, Echizen I (2019c) Use of a capsule network to detect fake images and videos. arXiv:1910.12467

NIST MFC 2018 Synthetic Data Detection Challenge. https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2019-0

Pham HX, Wang Y, Pavlovic V (2018) Generative adversarial talking head: bringing portraits to life with a weakly supervised neural network. arXiv:1803.07716

Ping W, Peng K, Gibiansky A, Arik SO, Kannan A, Narang S, Raiman J, Miller J (2017) Deep voice 3: 2000-speaker neural text-to-speech. arXiv:1710.07654

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) FaceForensics++: learning to detect manipulated facial images. In: ICCV

Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3:1

Shruti Agarwal HF, El-Gaaly T, Lim S-N (2020) Detecting deep-fake videos from appearance and behavior shruti. arXiv:2004.14491

Sinha P, Balas B, Ostrovsky Y, Russell R (2009) Face recognition by humans: 20 results all computer vision researchers should know about. https://www.cs.utexas.edu/users/grauman/courses/spring2007/395T/papers/sinha_20results.pdf

Stehouwer J, Dang H, Liu F, Liu X, Jain AK (2019) On the detection of digital face manipulation. Computer vision and pattern recognition. arXiv:1910.01717

Sun P, Li Y, Qi H, Lyu S (2020) Landmark breaker: obstructing deepfake by disturbing landmark extraction. In: IEEE workshop on information forensics and security (WIFS), New York, NY, United States

Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2015) What makes tom hanks look like tom hanks. In: ICCV

Suwajanakorn S, Seitz SM, Kemelmachershlizerman I (2017) Synthesizing obama: learning lip sync from audio. ACM Trans Graph 36(4):95

Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, vol 97 of Proceedings of machine learning research, Long Beach, California, USA, 09–15 Jun 2019. PMLR, pp 6105–6114

Thies J, Zollhofer M, Stamminger M, Theobalt C, Niessner M (2016) Face2Face: real-time face capture and reenactment of rgb videos. In: IEEE conference on computer vision and pattern recognition (CVPR)

Wang R, Juefeixu F, Ma L, Xie X, Huang Y, Wang J, Liu Y (2019a) Fakespotter: a simple yet robust baseline for spotting ai-synthesized fake faces. Cryptography cryptography and security. arXiv:1909.06122

Wang S, Wang O, Zhang R, Owens A, Efros AA (2019b) Cnn-generated images are surprisingly easy to spot... for now. Computer vision and pattern recognition. arXiv:1912.11035

Xuan X, Peng B, Wang W, Dong J (2019) On the generalization of gan image forensics. In: Chinese conference on biometric recognition. Springer, pp 134–141

Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: ICASSP