DeepFake the menace: mitigating the negative impacts of AI-generated content

Siwei Lyu
University at Buffalo, Buffalo, New York, USA

Organizational Cybersecurity Journal: Practice, Process and People

Received 11 October 2022
Revised 17 August 2023
Accepted 19 October 2023

Abstract

Purpose – Recent years have witnessed an unexpected and astonishing rise of AI-generated (AIGC), thanks to the rapid advancement of technology and the omnipresence of social media. AIGCs created to mislead are more commonly known as DeepFakes, which erode our trust in online information and have already caused real damage. Thus, countermeasures must be developed to limit the negative impacts of AIGC. This position paper aims to provide a conceptual analysis of the impact of DeepFakes considering the production cost and overview counter technologies to fight DeepFakes. We will also discuss future perspectives of AIGC and their counter technology.
Design/methodology/approach – We summarize recent developments in generative AI and AIGC, as well as technical developments to mitigate the harmful impacts of DeepFakes. We also provide an analysis of the cost-effect tradeoff of DeepFakes.
Research limitations/implications – The mitigation of DeepFakes call for multi-disciplinary research across the traditional disciplinary boundaries.
Practical implications – Government and business sectors need to work together to provide sustainable solutions to the DeepFake problem.
Social implications – The research and development in counter-technologies and other mitigation measures of DeepFakes are important components for the health of future information ecosystem and democracy.
Originality/value – Unlike existing reviews in this topic, our position paper focuses on the insights and perspective of this vexing sociotechnical problem of our time, providing a more global picture of the solutions landscape.

Keywords Generative AI, DeepFakes, Multimedia forensics

Paper type Conceptual paper

Introduction

Since late 2017, *DeepFake* [1] has become a buzzword frequently featured in the news and media. The term is a portmanteau of *deep* learning and *fake* media, and the namesake refers to the multimedia (texts, audio, images, and videos) created using generative AI models that often rely on deep neural network models.

Fabrication and manipulation of digital media are not new phenomena (Farid, 2012). However, the process with the pre-AI tools is usually lengthy, costly, and technically

The author is supported by funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123, the Center for Identification Technology Research and the National Science Foundation under Grant No. SaTC-2153112 and IIS-2008532. The views, opinions and findings expressed are the author's opinions and should. They should not be interpreted as representing the official views or policies of the Department of Defense, the National Science Foundation or the US Government.

Organizational Cybersecurity Journal: Practice, Process and People
Emerald Publishing Limited
e-ISSN: 2635-0289
p-ISSN: 2635-0270
DOI 10.1108/OCJ-08-2022-0014

demanding for ordinary users—the current widespread of DeepFakes results from the "democratization" of powerful generative AI technologies. The generative AI systems can train on vast amounts of unlabeled data, and these models are potent enough to generate convincing media. As a result, the technical threshold has been significantly lowered, making it more accessible and cheaper for users to generate DeepFakes in large quantities and better quality. With social media's rapid and broad reach, DeepFakes can now spread wide and fast.

The generative AI technology has many beneficial uses. Examples include immersive communication (e.g. Apple Vision Pro to be released in late 2023) and faster video streaming (e.g. Nvidia Maxine), reducing cost and effort in the movie and advertisement industry, and rehabilitation efforts for stroke victims and individuals with hearing impairment.

On the other hand, the DeepFakes can be weaponized and pose significant threats across multiple societal dimensions, including personal security, democratic processes, financial sectors, and the integrity of digital media. Personal security risks emerge as DeepFake can be used to fabricate convincing yet false representations of individuals, leading to reputation damage and psychological distress for anyone with accessible personal images or videos.

DeepFakes can escalate the scale and danger of online fraud and disinformation when used for deception to threaten our *cognitive security*. Similar to threats to our physical security or cybersecurity, which aims to break into our physical infrastructure or cyber systems, threats to cognitive security target our perceptual system and decision-making process. In particular, by creating illusions of an individual's presence and activities that did not occur in reality, DeepFakes can influence our opinions or decisions. For instance, a fake video showing a politician engaged in an inappropriate activity or maybe enough to sway an election. A voice call from a CEO requesting an employee to wire transfer funds to an offshore bank account could lead to actual financial losses to the company. A fake social media post with an AI-generated image of an emerging crisis could send the stock market awry. Using a synthesized realistic human face as the profile photo for a fake social platform account can significantly increase the effect of deception. An online predator can masquerade as a family member or a friend in a video chat to lure unaware victims. In addition, the training of generative AI models may use personal or copy-righted data, a gray area challenging the current laws.

Although few DeepFakes can cause long-lasting effects, they are effective in creating short-term chaos and confusion, and polarizing opinions when strong confirmation biases exist. The more fundamental impact of DeepFakes is the erosion of our trust in digital media — the fact that digital media can be synthesized or manipulated with AI makes it possible to challenge the authenticity of all digital media, particularly those conflicting with a particular agenda, a phenomenon often known as the *liar's dividend* or *plausible deniability* (Citron and Chesney, 2019). We have already seen a few recent cases of DeepFakes with real-world impact. GAN-generated face images were used as the profile photos for fake accounts on social platforms such as Twitter, Facebook, Instagram, and Linkedin. In 2020 alone, there have been 4,000 fake accounts found on these social platforms. Using such realistic face images as profile photos significantly increases the deceptiveness of those fake accounts. Another recent incident is that a scammer successfully used a synthesized voice using AI algorithms to impersonate the CEO of a UK company and misled an employee to wire transfer a substantial amount of money to the scammer's bank account (WSJ, 2019). In addition, reports show that hackers use DeepFakes to falsify biometric data to gain access to essential information systems (Verge, 2022). A DeepFake video of the Ukrainian President calling for Ukrainian troops to surrender began circulating on social media and Ukrainian news websites before being debunked and removed. As a recent case, on May 22nd, 2023, an AI-generated image showing a supposed explosion at the Pentagon building circulated on Twitter. Minutes later, the S&P 500 index dropped by 0.26%, showing such disinformation's impact [2].

The mounting concerns over the DeepFakes have spawned increasing interest in counter technologies, with substantial support from government and private companies [3]. On October 30th, 2023, the US Government issued an executive order on AI with a focus on the generative AI technology, to provide guarding rails for potential misuse including spreading falsified information and breaching privacy. This position paper aims to provide a high-level overview of the countermeasures to DeepFakes as a harmful impact of generative AI. While there have been numerous articles and surveys on the technical aspects of DeepFakes and counter technologies, e.g. Juefei-Xu *et al.* (2022), we focus on the insights and perspective of this vexing sociotechnical problem of our time, trying to provide a more global picture of the solutions landscape. The remaining of the position paper will be organized as follows. In Section 2, we will go over existing forms of DeepFakes and their underlying technology. Section 3 analyzes DeepFakes based on the production cost and perceptual impact tradeoff. Section 4 overviews current countermeasures to DeepFakes. Section 5 concludes the article with predictions on the future of DeepFakes and their countermeasures and some recommendations.

What is DeepFake?

From its original narrow meaning of face-swap videos created using deep neural network models, DeepFake is nowadays broadly used to refer to any digital media created or edited by generative AI algorithms. The rapid developments of DeepFakes are enabled by the ready availability of four key elements, namely.

(1) *Data and dissemination channel:* The fast growth and maturity of the Internet and social media since the 1990s provide a massive source of digital media data that can be used to train powerful generative AI models. They also offer fast and broad dissemination channels for DeepFakes.

(2) *Computation power:* Training deep neural network models requires high computing power and storage space that was nonexistent two decades ago. The rise of generative AI and the underlying deep learning technology can only be possible with the fast development of parallel high-performance computing hardware such as graphical processing units (GPUs) and advanced storage technology such as solid-state disks and flash memory.

(3) *Advancement in the generative AI models*, including the variational auto-encoder models (VAEs), *generative adversarial networks* (GANs), the diffusion models, and the large-language models. We briefly overview each type of model here:

- *The VAE model* (Kingma and Welling, 2014) consists of two DNNs, an encoder, and a decoder, trained using the target and the donor's faces. The encoder retains the target's facial expressions, and head poses while the decoder combines these with the target's identity. The auto-encoder is trained on the faces of the subjects whose faces will be swapped. The two issues share the same encoder, and they have different decoders. We then form two other encoder-decoder pairs. Training proceeds by adjusting the networks to minimize the difference between the input face images and the reconstructions, typically measured in L1 or L2 losses. This arrangement ensures that the encoder can capture common characteristics between the images of the two subjects. At the same time, the decoders can retain the individualities of each issue.
- *The GAN model* (Goodfellow *et al.*, 2014) consists of a pair of deep neural networks called the generator and discriminator. The generator creates a face from random noise. At the same time, the discriminator performs a binary classification

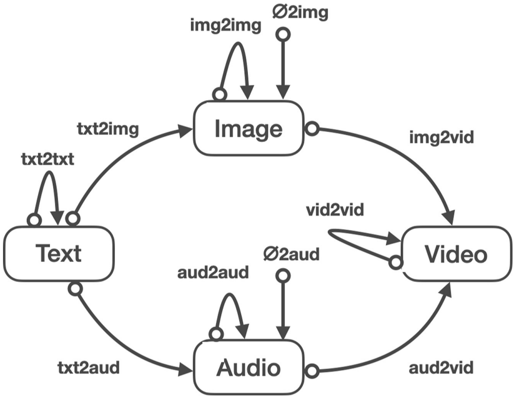
between the real images and the fake images synthesized by the generator—the training proceeds as a competition between the generator and the discriminator. The generator aims to create more realistic images to beat the discriminator, while the discriminator tries to be more accurate in its classification. The training ends when the two DNNs reach an equilibrium. The inputs of these networks are random vectors, while the outputs are high-quality fake face images. The classic examples are DCGAN (Radford *et al.*, 2015), WGAN (Arjovsky *et al.*, 2017), PGGAN (Karras *et al.*, 2017), and StyleGAN (Karras *et al.*, 2019). GAN images are becoming much easier to make (e.g. a user can obtain high-resolution human faces created by the most recent StyleGAN model from thispersondoesnotexist.com).

- The diffusion models work under a different mechanism. The process is iterative: the input image is smeared with additive noise, and then a "denoiser" in the form of a deep neural network is used to remove the noise and recover the original image. This noise-adding-removal process is stacked into several steps until the final output becomes indistinguishable from pure random noise. Then to generate a new image, in a similar way as in the GAN model, we start with a random noise sample and use the denoising process to create an image. The forward process of adding consecutive noises resembles a physical diffusion process (e.g. the dispersion of heat in a medium), hence the model's name. An additional feature of the diffusion model is its ability to incorporate text inputs (known as *prompts*) to create images per the user's description. Diffusion models have become mainstream image synthesis models since 2021 with commercial systems such as Midjourney and Stable Diffusion.
- Since 2020, there has been a surge in the popularity of Large Language Models (LLMs) in creating human-level texts using a variant of deep neural networks known as the *transformers*. Unlike typical deep neural networks, a transformer is a sequential neural network model that predicts a sample's new component following specific order (e.g. the sentence structure in texts) based on the components created. The transformer also uses the attention mechanism, which understands the relative importance of the existing component in predicting the new component. Transformer-based LLMs such as the GPT family (GPT1.0, GPT2.0, ChatGPT, GPT4.0) can create highly convincing texts that are difficult to distinguish from human written texts.

(4) *Open source software and web-based tools:* Most generative AI tools currently have code or implementations on open-source platforms like GitHub.com or HaggingFace.com. This has dramatically facilitated generative AI research's dissemination, reproduction, and augmentation. Furthermore, many generative AI tools, notable examples including ChatGPT, DALL-E2, Stable Diffusion, and Midjourney, have user-friendly, web-based interfaces, which further obviate the need for users' knowledge of programming, machine learning, and underlying computer systems.

Regarding the input/output type of media, current DeepFakes can also be summarized into ten major categories [4]. The relations between different types of DeepFake synthesis modalities are given in Figure 1.

(1) *∅2img* (input: null, output: image): This corresponds to the synthesis of images of objects (e.g. faces, vehicles, buildings) or scenes (indoor or outdoor) from random noises. The underlying models are GANs and diffusion models, which can synthesize images, and there is typically no user input or control over the synthesized image.



Source(s): Figure by author

Figure 1. Major types of DeepFake synthesis

(2) *∅2aud* (input: null, output: audio): This is the process of synthesizing audio signals of specific types (e.g. music, animal sounds) or scenes (e.g. crowded restaurants, tropical forest) from random noises. Commonly used models for audio synthesis are also GANs and diffusion models. As in the case of image synthesis, there is typically no user input or control over the synthesized results.

(3) *txt2txt* (input: text, output: text): This is the process of generating longer texts based on input text prompts from the users, and the models behind this task are sequential models trained on large text corpus (i.e. the large language models).

(4) *img2img* (input: image(s), output: image): This process creates new images based on the input image(s). It subsumes tasks such as image style transfer — rendering an image in the style of another image, face image editing — modification of facial attributes such as hair color, baldness, and smile and retouching complex characteristics like gender, age, etc., and image retouching, restoration, and super-resolution — filling missing or damaged image details.

(5) *txt2aud* (input: text, output: audio): This task is more commonly called text-to-speech (TTS), which converts the input text to corresponding audio. Many TTS systems can create a model of someone's voice, which can read the text in the same manner, intonation, and cadence as the target person. Others, such as Modulate.ai and lyrebird.ai, allow users to choose a voice of any age and gender rather than emulate a specific target.

(6) *aud2aud* (input: audio, output: audio): This task is more commonly referred to as voice conversion (VC), which transfers the input audio of one person's voice into another.

(7) *vid2vid* (input: video, output: video): This task is also known as video rewrite, and it is the original media synthesis technique that made DeepFakes widely known. The most prominent example of a vid2vid task is face replacement (or face-swapping), which involves generating an image of someone's face (the source) and carefully "stitching" it onto that of another person (the target). The target's identity is concealed, with the focal point being the source. Face replacement is often created using the auto-encoder model (AE). The approach of synthesizing face-swap videos has been mainstreamed through open-source software implementations on GitHub,

e.g. FakeApp (FaceApp, 2021), DFaker, and face swap-GAN (Lu, 2018), face swap (FaceSwap, 2016), and DeepFaceLab, Zao (2022). Another widely used vid2vid method is face re-enactment (or face puppetry), which entails manipulating the features of a target's face, including the movement of their mouth, eyebrows, eyes, and head tilting. Re-enactment does not aim to replace identities but rather to contort a person's expressions so they appear to be saying something they are not. Face puppetry videos can be made with AE-based models, similar to face-swap videos. In this case, the difference is that we will use the whole head and upper shoulder regions as input. Face puppetry is implemented in several commercialized apps, including Pinscreen (2021) and Face2face (Thies *et al.*, 2016).

(8) *txt2img* (input: text, output: image): This process creates images representing the information in user-provided text prompts. Unlike the unconditional image synthesis, txt2img gives users more control over the content of the generated content. Several text-prompted image synthesis platforms exist, including OpenAI's DALL-E2, Google's ImageGen, Stable Diffusion, and Midjourney.

(9) *aud2vid* (input: audio, output: video): This subsumes the task of lip-synching, which generates the lips of a person conditioned on the input audio, to create a manipulated video in which he/she speaks according to the input audio. A state-of-the-art lip-syncing method is wav2lip (Prajwal *et al.*, 2020).

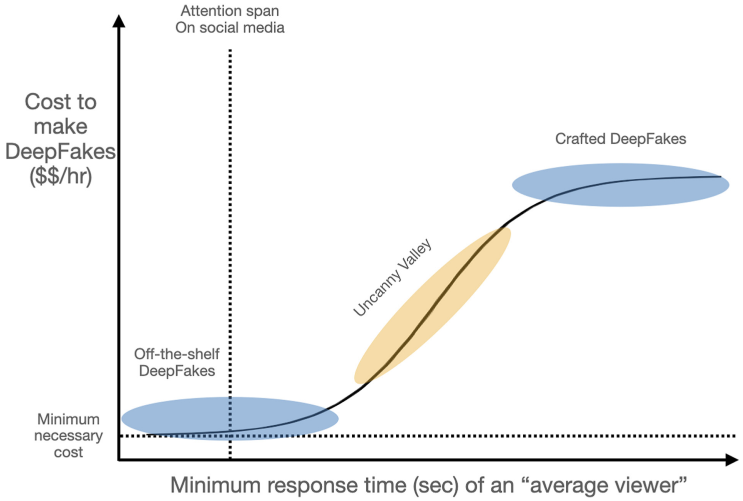
(10) *img2vid* (input: image, output: video): This includes the task of face animation, which transfers the target video's facial expressions and head motions to the "animate" input image. Face animation fulfills the same function as face-swap or face puppetry, but it can make synthesized videos of faces using only a single input image. An easy-to-use app, Avatarify (2022), has been developed to provide face animation functions on mobile phones.

A cost-perception analysis of DeepFakes

Given the ever-increasing levels of sophistication of the models, the realism of the deepfakes, and the easiness of access to the tools, the public media on generative AI models that create DeepFakes often portray a future of dooming dystopia or even an apocalypse, leading to outcries of limiting the generative AI technology from the media and the scientific community.

These reactions are predicated on two assumptions: that deepfakes can be cheaply made and that humans cannot tell them apart from the real media. However, other than some sporadic works, the two hypotheses have not been carefully examined. For one thing, we know that the cost of making generative AI models is hefty. For instance, training the OpenAI ChatGPT model, which has 175 billion parameters, will cost more than $12 million for three weeks [5], and the cost includes the utility (electricity and water for cooling), facility (space), equipment (computing servers or cloud charges), personnel (payment to workers who collect data and operate the models), etc. In addition, an aware and dedicated human can spot a synthesized medium given sufficient time and context. What we need to take into consideration when making predictions about generative AI and deepfakes are the more complex cost and effect factors.

Here we analyze the interaction between the cost of making DeepFakes and the impact on human perception and cognition. This is not an actual data analysis, as such data have been systematically collected in the literature, which would be a good future research topic. We base our analysis on a gloss estimation from our experience, and our primary purpose is to demonstrate the general relation. In addition, to simplify the analysis, we need to consider the effect of other information sources that can verify media authenticity. In Figure 2, the x-axis is



Source(s): Figure by author

Figure 2.
A cost-perception analysis of DeepFakes

the *minimum response time (MRT)*, a term in cognitive psychology and neuroscience that quantifies how quickly an individual perceives, processes, and reacts to a specific stimulus. Here it is used to quantify how long it takes an average human to identify a synthetic media in an *unaware* state, i.e. not consciously looking for a DeepFake. This is the case for most online users when they use social media. This is a crude simplification of all the complex factors, such as the user's understanding, educational background, demographics, and other cognitive and perceptive differences. It is, therefore, a different value for different subjects. For explanation, we assume an average user and use this to demonstrate the overall trend. MRT corresponds to a measure of difficulty or challenges for a subject to spot deepfake without prior knowledge of their presence. One bane of the ever-growing social media is that our attention span (i.e. the time spent examining a media) is limited and getting shorter (Kies, 2018). The *y*-axis corresponds to the hypothetical *dollar amount per hour* for making a DeepFake. This value amortizes all the costs involved in the process of making deepfakes.

We hypothesize the relation between cost and perception of DeepFakes as the curve shown in Figure 2 [6], emphasizing again that this is not based on actual data and our purpose is to illustrate the general relation. The overall trend is a shape of a sigmoid curve (S-curve) with three segments. The initial relatively flat segment corresponds to low-cost mass production of DeepFakes with off-the-shelf tools and minimal post-processing. The quality of the DeepFakes is often good enough to evade an unaware or occupied user who only spends a short time inspecting them, for instance, when the user is browsing social media posts. The middle range of the curve is a sharp increment of cost that can be regarded as the "uncanny valley" [7] for DeepFakes. This corresponds to the scenario when the user is more likely to be aware of DeepFakes, hence will spend more time examining the media. The available generative AI tools, albeit easy to use, often struggle to make consistently high-quality syntheses, resulting in conspicuous artifacts in the synthetic media that a conscious user can quickly spot. The last segment of the curve is another relatively flat region but with higher costs in production. This corresponds to the situation when the effort in production goes beyond a certain threshold — using specially designed tools, careful choice of the original media, and extensive post-processing to remove or conceal artifacts, the resulting "crafted" DeepFakes will become difficult to detect for an average user within a reasonable time range.

This curve also illustrates the battlefield frontier between DeepFake making and countermeasures. The advancements of generative AI technology are to “lower and flatten” this curve, i.e. reducing cost and producing DeepFakes less noticeable or detectable. The countermeasures to DeepFakes, on the other hand, aim to heighten the curve. In the case of non-technical countermeasures, we can view various governmental policies and regulations and industrial content moderation efforts [8] as increasing the cost of producing DeepFakes. The United States and the European Union have recently taken a major step toward regulation of generative AI technologies, such as ChatGPT, Midjourney, and Stable Diffusion, by implementing comprehensive measures for their development, deployment, and utilization. In the case of already released powerful generative models, if regulators require that all GPUs/CPUs can only execute generative models that create digital content with watermarks, all manufacturers would need to update their settings and produce such cards in the future. Consequently, pre-existing models without watermarks would become inoperative, making all AI-generated digital content identifiable. It provides an additional layer of protection instead of totally replying to deep fake detectors. On the other hand, enhancing user awareness and building resilience through improved user education on generative AI and synthetic media, especially for vulnerable groups such as teenagers and seniors, help to reduce the response time and are the keys to mitigating the negative social impacts of DeepFakes. The same is true for technical countermeasures — detection facilitates the users to spot DeepFakes while the obstruction and tracing methods increase the production costs of DeepFakes. We will overview these counter technologies in the sequel.

Counter technologies: detection

DeepFake forensics, focusing on counter technologies of DeepFakes, has become an active research area in response to the concerns around DeepFakes. The current efforts in DeepFake Forensics heavily tilt towards detection, formulating as a binary classification problem. So far, we have detection methods for all types of DeepFakes, as described in the previous section. There is a vast literature on DeepFake detection methods, and our purpose here is not to survey existing works thoroughly. More technical and detailed overviews of existing DeepFake detection methods can be found in, e.g. Juefei-Xu *et al.* (2022).

In practice, DeepFake detection methods are used for two primary purposes: triage and evidence. A DeepFake detector used for *triage* screens out a smaller subset of likely DeepFakes from a large number of images, audio, and videos for closer scrutiny. For triage detectors, the primary issue is detection accuracy, minimal human intervention, and run-time efficiency. As such, they are often implemented with data-driven methods, directly employing machine learning models trained on real media and DeepFakes to classify them. Data-driven triage detection is effective for low-cost DeepFakes (initial flat region of the S-shaped curve in Figure 2). However, they often lack explainability. If a data-driven detection method tells us that the result is 70% of confidence that the input is a DeepFake, it often sheds little light on the reasons or evidence supporting the result.

A DeepFake detector for evidence is required to have high explainability, i.e. the results must be accompanied by supporting evidence that is explainable and understandable to humans. Because of this requirement, DeepFake detection for evidence is often cue-based, exposing DeepFakes based on their lack, inconsistency, or violations of physical or physiological characteristics of the event that was supposed to be captured by the media in question. Cue-based evidence detection targets more carefully crafted DeepFakes with production costs corresponding to the middle and final segments of the S-shaped curve in Figure 2. We also point out that the triage classifier and the evidence classifier can be used in the same workflow, where the former is first used to narrow a large amount of media down to a smaller manageable subset of likely DeepFakes. The evidence classifier is applied to each case for a more detailed analysis.

Data-drive triage detection methods: The data-driven strategies apply various classification models, the majority of which are deep neural networks, to the problem of DeepFake detection. Most popular deep neural networks for other computer vision problems have been used, and data-driven methods achieve state-of-the-art performance on benchmark datasets. Many DNN-based DeepFake detection methods are developed for GAN-generated images or AE-generated face swap videos. For the former case, a common approach is to take the whole image as input and employ various forms of DNNs for classification, e.g. (Do *et al.*, 2018; Mo *et al.*, 2018; Chen *et al.*, 2021). Some recent methods, e.g. Wang *et al.* (2020), further use frequency-domain features based on the observations that GAN-synthesized images exhibit statistical differences from the original photographic images in the high-frequency range. DNN-based detection methods for face-swap videos follow the general intuition that each frame in the DeepFake video is created by blending the AE-synthesized faces into the face region in the original video frame. Many methods construct classifiers based on the difference between the interior and the exterior parts relative to the face regions. Other approaches focus on the boundary where blending is applied and expose the traces of blending operation as a trace of manipulation.

Early generations of the DNN-based DeepFake detection methods, e.g. (Afchar *et al.*, 2018; Güera and Delp, 2018; Li and Lyu, 2019; Nguyen *et al.*, 2019), focused on using various types of DNN architectures used in other Computer Vision tasks. This includes CNNs, RNNs, Capsule Networks, Vision Transformers, etc. However, the increasing model complexity entails larger datasets with higher qualities, which are complex and time-consuming. Subsequently, more current detection methods, e.g. (Sun *et al.*, 2021; Dong *et al.*, 2022; Chen *et al.*, 2022), explore novel approaches for training data synthesis and augmentation, i.e. to apply various transforms to existing training data to simulate the generation process. This is particularly true for detecting face-swap videos, where augmented training data can be obtained by randomly grafting faces into authentic images or frames.

The data augmentation approach has two advantages. The first is the reduced reliance on training data — many recent methods only are genuine real photos or videos and apply generation simulations to create negative examples. The second is the flexibility to inject information about the generation process. As the training data are made in the model training cycles, the algorithm developer can choose simulation operations close to those used in the generation process. The difficulty, however, is how to design appropriate augmentation methods that can lead to more effective detection methods. The other trend that combines with the data augmentation approach is self-supervised learning, where new training data are created by occluding or degrading the input. The algorithm then tries to recreate the original information and the classification task. The data augmentation and self-supervised learning approach have significantly improved the reliance on large-scale datasets for the DNN-based DeepFake detection methods.

Data-driven detection methods have achieved considerable progress recently, but their performance in the wild may still have much room for improvement [9]. Data-driven detection can often be hindered by post-processing, such as compression, performed by social platforms to reduce file size and save bandwidth. In addition, deliberate anti-forensic attacks can mislead the detection method to make classification errors by hiding traces of DeepFakes. It has been shown that DNN classifiers are vulnerable to deliberate adversarial attacks, which aim to mislead the classifier to make classification errors on a perturbed input.

Cue-based evidence detection methods: Although potent, the DeepFake synthesis models also have limitations in representing the more semantic aspects of the physical world. Indeed, these have been the primary cues that human viewers use to detect DeepFakes, such as the different colors of the eyes, halo effects near the hair, and blurry backgrounds in GAN-generated faces. Such semantic cues are human-understandable and can only be fixed by improving the generation model. If we can build detection methods based on such signals, the procedures will be more robust to adversarial attacks and afford intuitive interpretations. Along these lines, there are several existing works to detect GAN-generated faces. For instance, in Yang *et al.* (2019a), it was noticed that the early generation of GAN-synthesized faces tends to be asymmetric. This is because the earlier GAN models were more capable of generating facial parts but might not constrain the forged face to have a natural configuration. A simple classification method based on facial landmarks highlighting such asymmetry in faces is then developed to detect such artifacts. The work of Matern *et al.* (2019) focused on the relics of the synthesized facial parts. One of their method's exciting cues is the difference in color and resolution of the two synthesized eyes. In two more recent works (Hu *et al.*, 2021; Guo *et al.*, 2022), the authors noticed that GAN-synthesized faces exhibit particular irises artifacts. In Hu *et al.* (2021), it was seen that there are inconsistent corneal specular highlights between the two synthesized eyes (Figure 2). That is to say, for a natural face, the two corneal specular highlights are usually similar, while for the GAN synthesized face, they are different, as if the two eyes are looking at the other scenes. In Guo *et al.* (2022), the authors further noticed that the pupil shapes are often irregular, as the pupils of healthy adults have a circular shape (Figure 2). Such inconsistencies can be captured using automatic computer vision algorithms and used as tell-tale signs of GAN-synthesized faces.

One of our earlier works detecting face-swap videos (Li *et al.*, 2018) is also based on the simple physiological observation that faces do not blink naturally in those videos. The authors further explained the lack of flashing in generated face-swap videos. The training data used to create these early face-swap videos were portrait images obtained from the Internet using the image search provided by Google or other search engines. These portrait images carry an implicit bias due to the choice of the photographer, i.e. they mostly correspond to the subject with open eyes. The model inherited a bias in training data and reproduced the synthesized faces in these videos as an artifact or imperfection. This lack of realistic blinking can be exposed with an algorithm dedicated to detecting the opening/close of the eyes. When the blinking frequency is below the physiological data, about once in between 6–12 s, it is reasonable to doubt if the faces in the video are real live humans.

Another physical cue detects face-swap videos (Yang *et al.*, 2019b). For a real video, it is easy to understand that when the head moves in 3D space, the face, as part of the head, will move together. The synthesized faces are spliced onto the original video frames for face-swap videos. In particular, the synthesized face does not follow the 3D constraint; it is transformed in 2D to simulate the 3D motion of the head as in the original real video. Such differences between 2D and 3D can create artifacts in the face-swap videos, especially when the subject is not looking directly at the camera. The subject's face looks strange and does not seem to move with the head. This inconsistency can be used to expose fake videos. In particular, we can use 3D face alignment algorithms to estimate the 3D direction of the face orientation. In this case, all we need to do is compare the face orientations estimated using the central part of the case and estimate using the whole face. If this is a video of a natural person, we should see the two orientations to be very close. On the other hand, if the video is a DeepFake, the two orientations can have significant differences. Indeed this is what we see in actual data when we measure the differences in the orientations using their cosine distance, and we can observe more significant differences in the fake videos.

For synthesized audios, the work of AlBadawy *et al.* (2019) provides an approach based on inconsistencies in local phases. When we speak in a conversation, we make sound waves, and the sounds created by the speaker will reach the ear of the listener or a receiving device such as a microphone. However, the sound waves usually reach the destination via multiple paths. For instance, consider the indoor case; the listener's ears or the microphone can directly receive the sound waves, all bounced back from the surrounding walls and ceilings.

This means that the sound waves for the same voice will typically reach their destinations at slightly different times because of the lengths of the paths they use. They are also attenuated more for longer routes. The result is that the sound heard in our ears is a mixture of the same sound wave with slightly different amplitudes and phases. The difference in grades is caused by the other arrival times of the same sound wave signal. However, our auditory system usually does not perceive this difference in local phases. Physiological studies have shown that our ears are not particularly sensitive to differences in local phases.

The reason is simple if we want to pick up the message in the voice, the auditory system needs to focus on the local amplitudes of the sound wave, not the phases. Amplitudes can be captured by second-order signal representations such as spectrograms, but they are insensitive to local phases. We can use higher-order signal statistics to capture the local phases in sound audio. One such higher-order statistic is known as the bi-spectra. In bi-spectral analysis, we first transform the time-windowed audio signals into the frequency domain using Fourier transform. Conventional spectrogram looks at the correlation between a pair of frequency components, while in bi-spectral analysis, we look for mutual dependencies among triplets of frequency components. The resulting bi-spectrum, or its normalized version known as bi-coherence, is a complex quantity. The differences in the generation process may be exposed when we look at them through the bi-spectral analysis. Indeed this is the case. Here are the bi-spectral analysis results of actual human voice samples (right) and several different types of synthetic voice samples. Because bispectra are complex-valued, we offer both the amplitude and the phases. These bi-spectral analyses show the significant differences between these samples with authentic human voices. We collect the top four order cumulants (mean, variance, skewness, and kurtosis) considering symmetry in bi-coherence from the amplitudes and phases. This leads to an 8-dimensional feature vector. Then we feed statistical features extracted from this bi-spectra and train a simple ensemble classifier based on SVM. We visualize the parts in this two-dimensional graph. We see that human voices, highlighted with the region with the ellipse, can be easily separated from the AI-synthesized voices.

The explainability of the cue-based evidence detection method could be a double-edged sword, as the makes can be easily understood by DeepFake makes and incorporated in the generative AI model, for instance, by augmenting training data or designing dedicated modules to reproduce the cues. Another limitation of cue-based detection methods is that they are often only applicable under certain conditions (e.g. detecting GAN-generated faces using iris reflection requires a high-resolution area of the eye region). This makes the cue-based methods more narrowly scoped when compared with the data-driven detection methods.

DeepFake datasets: Developing DNN-based detection methods require large-scale datasets for model training and performance evaluation. The availability of large-scale datasets of various types of DeepFakes is an enabling factor and a reflector of the rapid development of DeepFake Forensics. To date, several such large-scale datasets exist for face-swap videos. The first dedicated DeepFake dataset, UADFV (Yang *et al.*, 2019a, b), only had 49 DeepFake face-swap videos with visible artifacts when released in June 2018. Subsequently, more DeepFake datasets are proposed with increasing quantities and qualities (Li *et al.*, 2020; Rossler *et al.*, 2019). For instance, the Celeb-DF dataset (Li *et al.*, 2020) is a State-of-the-art, high-quality, large-scale face-swapping DeepFake video dataset. There are 5,500+ face-swapping DeepFake videos with more than two million frames. More than 2000 downloads since November 2019. It is one of the most used datasets for DeepFake forensics research, its size, and quality. The DeepFake detection is usually evaluated on large benchmark datasets. Most current datasets are for DeepFake images and videos, but DeepFake audio datasets have also received more attention (Todisco *et al.*, 2019).

DeepFake detection platforms: Although we now have many effective detection methods, there is a gap for users to benefit from these state-of-the-art methods. As code for these detection methods scatters on online code-sharing platforms like GitHub, a user wishing to analyze one image or video must find the code, download it to her computer, set up the environment, and compile it, before using it for analysis. There is a need to close this last-mile gap between the users and the detection algorithms.

DeepFake-o-meter (Li *et al.*, 2021) is an open-source, online, and user-friendly online platform for third-party DeepFake detection algorithms. For users, it provides a convenient service to analyze DeepFake media with multiple state-of-the-art detection algorithms, with secure and private delivery of the analysis result. For developers of DeepFake detection algorithms, it provides an API architecture to wrap individual algorithms and run them on a remote machine. For researchers of digital media forensics, it is an evaluation/benchmarking platform to compare the performance of multiple algorithms on the same input. The DeepFake-o-meter can be accessed at https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/.

Counter technologies: beyond detection

More than merely detecting DeepFakes is required to combat the negative impacts of DeepFakes. Exposing a DeepFake is only the first step in the investigation, and we need more information about the origin, means, and intent behind the observed DeepFakes. These are information not provided by a classifier. Furthermore, DeepFake detectors primarily operate postmortem, applicable after DeepFakes emerge. This gap in time between the emergence and circulation of DeepFakes and their eventual exposure may lead to damage. Due to these reasons, recent years have also seen active developments of DeepFake Forensic methods beyond detection.

(1) Model attribution: The specific means by which the DeepFakes are created are helpful in the forensic investigation of a DeepFake attack. For instance, knowing that the synthesis model is a copy of a publicly available tool can narrow down the list of downloaded users. If the synthesis model is a slightly modified version of a known model, it may indicate that the DeepFake maker has enough skill to modify an existing tool. If the synthesis model is novel and unrelated to the existing tools, it may indicate that the DeepFake maker has significant resources and skills. The problem of DeepFake generation model attribution is to infer the type and details of the generation model using DeepFakes created from it. Many model attribution methods use multi-class classification to differentiate DeepFakes made with different generative AI models. More recent approaches, e.g. Jia *et al.* (2022), look for features that can differentiate different generative models for detection.

(2) Active Forensics of DeepFakes: Recent tools (e.g. Avatarify and DeepFaceLive) create real-time face-swap or face reenactment DeepFakes. The real-time DeepFakes pose new challenges to existing detection methods that often struggle to achieve the levels of accuracy needed to be incorporated into a practical video-conferencing application and run in real-time. As such, new approaches of active DeepFake forensics, which combines liveliness detection in biometrics to identify if the digital synthesis is a live presentation of an individual. For instance, the work (Gerstner and Farid, 2022) exploits the unique constrained environment afforded by a video-conferencing call to detect real-time DeepFakes by varying the lighting pattern on the screen and extracting the same lighting variation from the attendant's face. A similar idea is proposed to expose real-time face-swap DeepFakes using corneal reflection from real-time captured iris images (Guo *et al.*, 2023). Injecting "traces" to the would-be training data is a different approach. The traces are unnoticeable to humans and can be extracted later from the DeepFakes created using models trained on the tainted data and used as reliable and definite evidence of DeepFakes (Sun *et al.*, 2021).

(3) Proactive obstruction of DeepFakes: Unlike detection methods, the preemptive approach directly obstructs the training or generation of deepfakes, either leading to failures in the generation or stalled training process. One idea is to poison the would-be training data by adding specially designed noises. Training with the poisoned data will lead to dysfunctional models that create low-quality DeepFakes. The work of Sun *et al.* (2021) obstructs the DeepFake generation by attacking a vital step of the generation pipeline, facial landmark extraction. The method generates adversarial perturbations to disrupt the facial landmark extraction, such that the DeepFake models cannot locate the face to swap.

(4) Authentication of real media: The flip side of exposing DeepFake is to verify the authenticity of untouched real media. Two major approaches for authenticating real media are watermarking and controlled capture. The former is to embed invisible signals known as digital watermarks to authentic media, which can be later extracted for verification. Controlled capture does not add signals to the real media. However, it extracts statistical features (the signatures) that can identify the real media and store them in a secure location, e.g. using blockchain technology and on a secure cloud-based server. The stored signatures can be compared later with those extracted from a media to authenticate. The authentication of real media has been supported by the Content Authenticity Initiative (CAI, https://contentauthenticity.org/) and the Coalition for Content Provenance and Authenticity (C2PA, https://c2pa.org/), both initiated by Adobe and adopted by many major tech companies and news outlets.

Future perspectives

Looking into the future, we will continue to see accelerated development of deepfake technology with newer forms, a higher level of visual realism, fewer artifacts, and an increasing level of automation to facilitate production at scale. In the not-too-distant future, ordinary users may have access to more easy-to-use and ready-made tools to manipulate media as they use Photoshop to edit images today. Furthermore, creating DeepFake targeting an individual can also become a service for users who do not access the computation resources. We will likely reach a tipping point where the production of fake content outpaces our ability to detect it. This could have several implications. With this advancement, we anticipate an increase in the volume and quality of fake content on social media and the Internet. The speed of generation and distribution of this content may accelerate, potentially flooding our information ecosystem. Furthermore, we might see more comprehensive disinformation campaigns that combine multimedia and multi-site illusions. Deceptive narratives could be reinforced across texts, images, videos, and audio for more compelling storytelling. Additionally, we can expect more refined and subtle forms of manipulation targeting individuals, businesses, and government agencies. Rather than aiming for broad-scale influence, attackers may focus on less-known targets, choose to time strategically, and employ more comprehensive forms of media to sway public opinion subtly. In other words, the attacks might become more granular, specific, and potentially more effective.

The continuous improvement of synthesis models put DeepFake making and forensics in a cat-and-mouse game. As such, we must continue developing DeepFake Forensic methods that are more effective, efficient, robust, and explainable, focusing on identifying inconsistencies or inaccuracies, understanding their impact, and discerning the intent behind DeepFakes. New methods also need to resolve zero-day attacks when a new approach, form, or model of media synthesis is not strongly related to previously known cases. In addition, even the distinction between real and DeepFake may become less relevant when the majority are synthetic (and harmless) in the future with the broader use of generative AI and synthetic media, for instance, high-quality video compression and virtual and realistic avatars in the meta-verse. This means that the very problem of DeepFake detection may become irrelevant. A more critical question may be to ensure the authorized usage of generative AI models. In addition, it is also helpful to test the hypothesis presented in the S-shaped curve about the cost and perceptual effect of DeepFakes, through careful user study and quantification.

There are also multiple non-technical issues that we need to consider in the future development of DeepFake counter technologies.

(5) The current policies and regulations need to be more precise in delineating innocent and malicious use of synthetic media as the generative AI technology and synthetic media are dual-used, unless in extreme cases of misuse (e.g. generating targeted pornography videos to defame an individual), banning or restricting the technology may lead to unwanted side effects.

(6) The open-source nature of many DeepFake forensic tools can be exploited to improve synthesis and anti-forensics, a phenomenon often known as the "detection dilemma." Thus, we need to find better ways to disseminate such forensic tools to balance the users' benefits and reduce abuses by malicious players.

(7) Furthermore, we must also pay more attention to the intervention procedure after exposure to a DeepFake. In the recent work of Shan *et al.* (2022), in which the authors show that the straightforward approach of labeling an exposed AI-synthesized video (commonly known as a DeepFake) may undermine mitigation efforts by not taking socio-psychological considerations into account — users are drawn by curiosity to watch a "DeepFake," which conversely increases the attention and spread of the labeled fake content.

(8) Corporations should incorporate synthetic media as part of the cybersecurity training to accommodate the increasing risks (frauds, scams, and attacks) using DeepFakes alongside other cognitive and cyber-security threats such as social engineering.

(9) More investment and effort should also be invested to increase the awareness and resilience of the general public about the risks of generative AI and synthetic media.

(10) As a socio-technical problem by nature, the research on mitigating the harmful effects of DeepFakes calls for collaborations across traditional disciplinary boundaries. AI researchers must be willing to work with humanists, social scientists, psychologists, and communication and education researchers to work out a holistic solution with technical methods supported by a deeper understanding of the root cause and effective delivery of analysis results.

(11) Last but not least, compared with making DeepFakes, exposing DeepFakes is not deemed profitable, with few successful commercial endeavors. Thus, market forces must be used to create a sustainable model for DeepFake forensic services.

Notes

1. Its origin was the Reddit account name that shared AI-generated pornographic videos with transplant celebrity faces in late 2017.

2. Source: https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/

3. Notable examples include the DARPA MediFor and SemaFor programs, the NIST 2018, 2020 and 2021 Synthetic Data Detection Challenge, and the DeepFake Detection Challenge (https://deepfakedetectionchallenge.ai) in 2020 sponsored by Facebook, Microsoft, Amazon and Partnership in AI.

4. There are other less known types of media synthesis (e.g. text-prompt video generation, txt2vid) that are currently under active development and expected to join the mainstream in the coming years.

5. https://www.forbes.com/sites/johnkoetsier/2023/02/10/chatgpt-burns-millions-every-day-can-computer-scientists-make-ai-one-million-times-more-efficient/?sh=6550517d6944

6. We consider modifications that lead to cognitive impacts, i.e. can affect the user's opinion or decision-making. Therefore, cost will be the dominating factor affecting the perception. This excludes trivial modifications (e.g. changing the value of a single pixel in an image) that cannot be detected by the user.

7. The term "uncanny valley" is originally used to describe the relationship between the human-like appearance of a robotic object and the emotional response it evokes.

8. In the USA, Congress has passed several bills to regulate the abuse of DeepFakes, including the Malicious Deep Fake Prohibition Act of 2019, the DEEP FAKES Accountability Act, the Deepfake Report Act, and the IOGAN Act of 2019. More recently, President Biden has requested watermarking synthetic contents made with tools provided by major tech-companies [source]. Similar legislative efforts have also been made in the EU and China. Major social platforms (e.g. Twitter and TikTok) have followed suit to specify policies to control DeepFakes.

9. The best-performing algorithm in the DeepFake Detection Challenge 2020 achieved a detection accuracy of about 85% on the synthesized data, but such performance sharply drops to 65% on data in the test dataset (Dolhansky *et al.*, 2022).

References

Afshar, D., Nozick, V., Yamagishi, J and Echizen, I. (2018), "Mesonet: a compact facial video forgery detection network", WIFS, doi: 10.1109/wifs.2018.8630761.

AlBadawy, E., Lyu, S. and Farid, H. (2019), "Detecting ai-synthesized speech using bispectral analysis", *Workshop on Media Forensics (in conjunction with CVPR)*, Long Beach, CA, United States, 2019.

Arjovsky, M., Chintala, S., Bottou, L. (2017), "Wasserstein GAN", *arXiv preprint arXiv:170107875*.

Avatarify (2022), available at: https://avatarify.ai/

Chen, B., Liu, X. and Zheng, Y. (2021), "A robust GAN-generated face detection method based on dual-color spaces and an improved Xception", TCSVT, 2021.

Chen, L., Zhang, Y., Song, Y., Liu, L. and Wang, J (2022), "Self-supervised learning of adversarial example: towards good generalizations for deepfake detection", *arXiv Preprint arXiv:2203*, 12208, doi: 10.1109/cvpr52688.2022.01815.

Citron, D.K. and Chesney, R. (2019), "Deep fakes: a looming challenge for privacy, democracy, and national security", *California Law Review*, Vol. 107, p. 1753 (In press).

Do, N.-T., Na, I.-S. and Kim, S.-H. (2018), "Forensics face detection from GANs using convolutional neural network", ISITC, 2018.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C. (2022), "*The deepfake detection challenge (dfdc) dataset", arXiv preprint, arXiv:2006.07397.*

Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F. and Guo, B. (2022), "Protecting celebrities with identity consistency transformer", *arXiv Preprint arXiv: 2203.01318.*

FaceApp (2021), "FaceApp", available at: https://faceapp.com/app

FaceSwap (2016), "FaceSwap", available at: https://github.com/deepfakes/faceswap

Farid, H. (2012), *Digital Image Forensics*, MIT Press (In press).

Gerstner, C.R. and Farid, H. (2022), "Detecting real-time deep-fake videos using active illumination," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2022*, Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), "Generative adversarial networks", *Proceedings of Advances of Neural Information Processing Systems* (In press).

Güera, D. and Delp, E.J. (2018), "Deepfake video detection using recurrent neural networks", AVSS, doi: 10.1109/avss.2018.8639163.

Guo, H., Hu, S., Wang, X., Chang, M.-C. and Lyu, S. (2022), "Eyes tell all: irregular pupil shapes reveal GAN-generated faces", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore.

Guo, H., Wang, X. and Lyu, S. (2023), "Detection of real-time deepfakes in video conferencing with active probing and corneal reflection", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rohdes Island, Greek.

Hu, S., Li, Y. and Lyu, S. (2021), "Exposing GAN-generated faces using inconsistent corneal specular highlights", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada.

Jia, S., Li, X. and Lyu, S. (2022), "Model attribution of face-swap deepfake videos", *IEEE Conference on Image Processing (ICIP)*, Bordeau, France.

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q. and Liu, Y. (2022), "Countering malicious DeepFakes: survey, battleground, and horizon", *International Journal of Computer Vision*, Vol. 130 No. 7, pp. 1678-1734, doi: 10.1007/s11263-022-01606-8.

Karras, T., Aila, T., Laine, S., Lehtinen, J (2017), "Progressive growing of GANs for improved quality, stability, and variation", *arXiv preprint arXiv:171010196*.

Karras, T., Laine, S. and Aila, T. (2019), "A style-based generator architecture for generative adversarial networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401-4410, doi: 10.1109/cvpr.2019.00453.

Kies, S.C. (2018), "Social media impact on attention span", *Journal of Management and Engineering Integration*, Vol. 11 No. 1, pp. 20-27.

Kingma, D.P. and Welling, M. (2014), "Auto-encoding variational bayes in ICLR".

Li, Y. and Lyu, S. (2019), "Exposing deepfake videos by detecting face warping artifacts", *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Li, Y., Chang, M.-C. and Lyu, S. (2018), "In Ictu Oculi: exposing AI created fake videos by detecting eye blinking", *IEEE Workshop on Information Forensics and Security (WIFS)*, Hong Kong, doi: 10.1109/wifs.2018.8630787.

Li, Y., Sun, P., Qi, H. and Lyu, S. (2020), "Celeb-DF: a large-scale challenging dataset for DeepFake forensics", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Seattle et al.)*, doi: 10.1109/cvpr42600.2020.00327.

Li, Y., Zhang, C., Sun, P., Ke, L., Ju, Y., Qi, H. and Lyu, S. (2021), "Deepfake-o-meter: an open platform for deepfake detection", *International Conference on Systematic Approaches to Digital Forensic Engineering (SADFE)*. doi: 10.1109/spw53761.2021.00047.

Lu, S.A. (2018), "FaceSwap-GAN", available at: https://github.com/shaoanlu/faceswap-GAN

Matern, F., Riess, C. and Stamminger, M. (2019), "Exploiting visual artifacts to expose deepfakes and face manipulations", *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019, pp. 83-92, doi: 10.1109/wacvw.2019.00020.

Mo, H., Chen, B. and Luo, W. (2018), "Fake faces identification via a convolutional neural network", *ACM IH&MMSEC, 2018*.

Nguyen, H., Yamagishi, J. and Echizen, I. (2019), "Capsule-forensics: using capsule networks to detect forged images and videos", *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 2307-2311.

Pinscreen (2021), "Pinscreen AI-driven virtual avatars", available at: http://www.pinscreen.com/

Prajwal, K.R. and Mukhopadhyay, R., and Namboodiri, V.P. and Jawahar, C.V. (2020), "A lip sync expert are all you need for speech to lip generation in the wild, published at ACM multimedia 2020".

Radford, A., Metz, L., Chintala, S. (2015), "Unsupervised representation learning with deep convolutional generative adversarial networks", *arXiv preprint arXiv:151106434*.

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

Shan, J., Li, X. and Lyu, S. (2022), "Model attribution of face-swap DeepFake videos", *IEEE Conference on Image Processing (ICIP)*, Bordeaux, France, doi: 10.1109/icip46576.2022.9897972.

Sun, Z., Han, Y., Hua, Z., Ruan, N. and Jia, W. (2021), "Improving the efficiency and robustness of deepfakes detection through precise geometric features", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/cvpr46437.2021.00361.

Thies, J, Zollhofer, M., Stamminger, M., Theobalt, C. and Nießner, M. (2016), "Face2face: real-time face capture and reenactment of rgb videos", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387-2395, doi: 10.1109/cvpr.2016.262.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Deutsch, A., Yamagishi, J., Evans, N., Kinnunen, T. and Lee, K.A. (2019), "ASVSpoof 2019: future horizons in spoofed and fake audio detection", Interspeech 2019. doi: 10.21437/interspeech.2019-2249.

Verge (2022), available at: https://www.theverge.com/2022/5/18/23092964/deepfake-attack-facial-recognition-liveness-test-banks-sensity-report (accessed 20th May 2022).

Wang, S.-Yu, Wang, O., Zhang, R., Owens, A. and Efros, A.A. (2020), "CNN-generated images are surprisingly easy to spot. for now", *CVPR*, doi: 10.1109/cvpr42600.2020.00872.

WSJ (2019), "Wall street journal", available at: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Yang, X., Li, Y., Qi, H. and Lyu, S. (2019a), "Exposing GAN-synthesized faces using landmark locations", *International Workshop on Information Hiding and Multimedia Security*, Paris, France, 2019, doi: 10.1145/3335203.3335724.

Yang, X., Li, Y., and Lyu, S. (2019b), "Exposing deep fakes using inconsistent head poses in ICASSP".

Zao (2022), "Apple app store", available at: https://apps.apple.com/cn/app/zao/id1465199127

Further reading

AlBadawy, E.A. and Lyu, S. (2020), "Voice conversion using speech-to-speech neuro-style transfer", *Proceedings Interspeech*. doi: 10.21437/interspeech.2020-3056.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S. and Choo, J. (2018), "Stargate: unified generative adversarial networks for multi-domain image-to-image translation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8789-8797.

He, Z., Zuo, W., Kan, M., Shan, S. and Chen, X. (2019), "Attgan: facial attribute editing by only changing what you want", *IEEE Transactions on Image Processing*, Vol. 28 No. 11, pp. 5464-5478, doi: 10.1109/tip.2019.2916751.

Mohammadi, S.H. and Kain, A. (2014), "Voice conversion using deep neural networks with speaker-independent pre-training", *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp. 19-23.

Ping, W., Peng, K. and Chen, J. (2018), "Clarinet: parallel wave generation in end-to-end text-to-speech", *arXiv Preprint arXiv:1807.07281*.

Ren, C.Hu, Tan, Xu, Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2020), "FastSpeech 2: fast and high-quality end-to-end text-to-speech", *arXiv preprint arXiv:2006.04558*.

Wang, Z., Liu, Y. and Shan, L. (2021), "CE-Tacotron2: end-to-end emotional speech synthesis", *2021 60th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pp. 48-52.

About the author

Siwei Lyu is Empire Innovation Professor at the Department of Computer Science and Engineering, the Founding Director of UB Media Forensic Lab (UB MDFL) and the Founding Co-Director of the Multidisciplinary Center of Information Integrity at the University at Buffalo, State University of New York. Dr Lyu received his Ph.D. degree in Computer Science from Dartmouth College in 2005, his M.S. degree in Computer Science in 2000 and his B.S. degree in Information Science in 1997, both from Peking University, China. Dr Lyu's research interests include digital media forensics, computer vision and machine learning. Dr Lyu has published over 170 refereed journal and conference papers. He is the recipient of the IEEE Signal Processing Society Best Paper Award (2011), the National Science Foundation CAREER Award (2010), SUNY Albany's Presidential Award for Excellence in Research and Creative Activities (2017), SUNY Chancellor's Award for Excellence in Research and Creative Activities (2018) and Google Faculty Research Award (2019). He is Fellow of IEEE and IAPR. Siwei Lyu can be contacted at: siweilyu@buffalo.edu