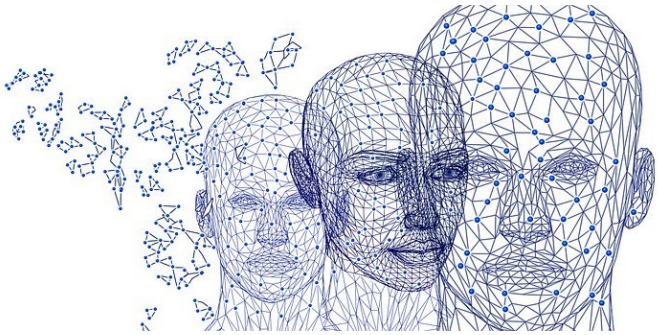
May 2021, Issue Code NL-2022-1

DDD (Digital Data Deception) Technology Watch Newsletter

Table of Contents



"*All warfare is based on deception. Hence, when we are able to attack, we must seem unable; when using our forces, we must appear inactive; when we are near, we must make the enemy believe we are far away; when far away, we must make him believe we are near.*"

— Sun Tzu, *The Art of War*

Editors: Enes Altuncu, Virginia Franqueira, Sanjay Bhattacherjee and Shujun Li
Affiliation: Institute of Cyber Security for Society (iCSS), University of Kent, UK
Contact Us: ddd-newsletter@kent.ac.uk

Editorial

This sixth issue of the Digital Data Deception (DDD) Technology Watch Newsletter is dedicated to *deepfake technology* in terms of state-of-the-art and state-of-practice; it is organised into four sections.

The first section covers technical developments of deepfake and reviews 12 articles published in 2020 or 2021. These articles are related to the generation, detection, and prevention of deepfake. The second section covers deepfake from the perspective of Psychology. It summarises four articles relating to the psychological and behavioural effect of deepfake technology. Articles for Section 1 were selected based on a systematic literature review process. As such, searches were performed on Scopus with keywords: *deepfake* and *deep-fake* (which includes *deep fake*). Then, inclusion/exclusion criteria were applied to the search results. Finally, a number of articles were selected according to the quality of their venues. Articles for Section 2 were selected using a venue-based approach.

Sections 3 and 4 cover the state-of-practice of deepfake technology. The third section results from searches on three main sources (Google, GitHub, and Google Play Store), and provides an overview of the most popular off-the-shelf tools and apps for generation and detection of deepfake videos, images or audio. The last section reviews example incidents that used deepfake, and imminent deepfake risks or opportunities in specific sectors.

We hope you enjoy reading this issue. Feedback is always welcome, and should be directed to ddd-newsletter@kent.ac.uk.

List of Acronyms

- AEI-NET: Adaptive Embedding Integration Network
- API: Application Programming Interface
- ASVSpoof: Automatic Speaker Verification Spoofing
- CNN: Convolutional Neural Network
- ConvLSTM: Convolutional Long Short-Term Memory
- CW-L_2: Carlini and Wagner L_2
- CycleGAN: Cycle-Consistent Adversarial Networks
- DCGAN: Deep Convolutional Generative Adversarial Network
- DEA-Net: Dynamic Enhanced Authentication Network
- DFDC: DeepFake Detection Challenge
- DIP: Deep Image Prior
- DNN: Deep Neural Network
- FFHQ: Flickr-Faces-HQ
- FGSM: Fast Gradient Sign Method
- FWA: Face Warp Artifacts
- GAN: Generative Adversarial Networks
- GRID: Global Research Identifier Database
- GPU: Graphics Processing Unit
- GTEA: Georgia Tech Egocentric Activity
- HEAR-NET: Heuristic Error Acknowledging Refinement Network
- IP-GAN: Identity and Pose Disentangled Generative Adversarial Networks
- LSTM: Long Short-Term Memory
- PoA: Proof-of-Authentication
- ResNet: Residual Neural Network
- RL: Reinforcement Learning
- VGG: Visual Geometry Group

Deepfake Generation, Detection & Prevention

Introduction

Recent advances in deep learning, especially Generative Adversarial Networks (GANs) have enabled the generation of realistic fake images, videos and audios. This has exposed online users to misinformation that, in some cases, is now more convincing because of deepfake content. It is therefore important to develop stronger defences to identify and mitigate deepfake content efficiently.

Deepfake Generation & Datasets

Gandhi and Jain [6] used adversarial perturbations to enhance deepfake images and fool common deepfake detectors. The authors generated adversarial perturbations using the Fast Gradient Sign Method (FGSM) and the Carlini and Wagner L_2 norm attack ($CW-L_2$) in both blackbox and whitebox settings. Figure 1 shows the workflow of the proposed attack with an example. The proposed technique decreased the accuracy of detectors from over 95% to less than 27%. To defend against adversarial perturbations, the authors explored two improvements to deepfake detectors; Lipschitz regularization and Deep Image Prior (DIP). While the former increases robustness to input perturbations by constraining the gradient of the detector with respect to the input, the latter removes perturbations using a generative Convolutional Neural Network (CNN) in an unsupervised manner. As a result, regularisation improved the detection of perturbed deepfakes on average, including a 10% accuracy boost in the blackbox case. The DIP defence achieved 95% accuracy on perturbed deepfakes that fooled the original detector while retaining 98% accuracy in other cases on a 100 image subsample. For the experiments, the authors used the CelebA dataset containing real celebrity face images, and generated fake images by using Few-Shot Face Translation GAN. In addition, they tested VGG-16 (Visual Geometry Group) and ResNet-18 (Residual Neural Network) architectures on their dataset.

Editorial Comments

Although the DIP method proposed by Gandhi and Jain [6] showed promising results, the time it took seemed to be a limitation (around 30 minutes per image on a NVIDIA Tesla K80 Graphics Processing Unit (GPU)).

=

Fernandes and Jha [5] utilised a state-of-the-art reinforcement learning (RL)-based texture patch attack to fool the FaceForensics++ deepfake detection system. They also validated the efficacy of using DeepExplainer to obtain attributions of deepfake videos, and heart rate detection from the face, for differentiating real faces from adversarial attacks. For the experiments, two new deepfake datasets were constructed by collecting donor and original subject videos from YouTube and COHFACE and using Deepfakes Web (`https://deepfakesweb.com/`). Apart from these constructed datasets, the DeepfakeTIMIT dataset, which was obtained from the VidTIMIT dataset, was also used. The proposed attack achieved 84%, 86% and 97% accuracy rates for COHFACE, DeepfakeTIMIT and the YouTube datasets, respectively.

Li et al. [12] presented a new large-scale deepfake video dataset, Celeb-DF (`http://www.cs.albany.edu/~lsw/celeb-deepfakeforensics.html`), containing 5,639 high-quality deepfake videos of celebrities generated using an improved synthesis process. The real source videos were based on publicly available YouTube video clips of 59 celebrities of diverse genders, ages, and ethnic groups. The authors aimed to improve the overall visual quality of the deepfake videos in Celeb-DF, compared to the existing deepfake video datasets. These improvements included generating higher resolutions, reducing colour mismatch between faces, improving mask generation accuracy and reducing temporal flickering. The constructed dataset was compared with the existing datasets, including UADFV, DeepfakeTIMIT, FF-DF, DFD and DeepFake Detection Challenge (DFDC), in terms of visual quality and performance. While Celeb-DF provided better visual quality, it also resulted in more accuracy than the other datasets when evaluated with nine publicly available deepfake detection methods.

Deepfake Detection

Li et al. [11] proposed a face swapping frame-

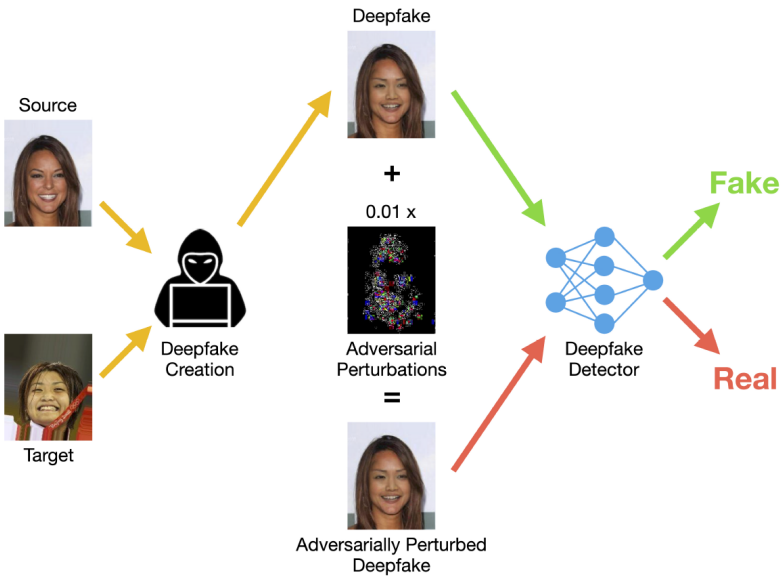


Figure 1: An example of deepfake generation by using adversarial perturbations from the study of Gandhi and Jain [6].

work, *FaceShifter*, for high fidelity and occlusion-aware face swapping. FaceShifter involves two stages. In the first stage, a GAN-based network, named *Adaptive Embedding Integration Network* (AEI-NET), has been designed to implement a thorough and adaptive integration of target attributes. For the second stage, the authors designed a *Heuristic Error Acknowledging Refinement Network* (HEAR-NET) to refine the results of AEI-NET in a self-supervised manner without manual annotations. Figure 2 shows some example face images generated by FaceShifter. The proposed framework has been trained by using several datasets such as CelebA-HQ, Flickr-Faces-HQ (FFHQ), VGGFace, EgoHands and GTEA (Georgia Tech Egocentric Activity) Hand2K, and evaluated on the FaceForensics++ dataset. The authors compared their method with existing methods including FaceSwap, DeepFakes and Identity and Pose Disentangled Generative Adversarial Networks (IP-GAN), qualitatively and quantitatively. The experiments showed that the proposed method outperformed existing methods. Finally, the authors examined the performance of two face forgery detection algorithms, FF++ and Face X-Ray, on the images generated by FaceShifter. The results showed that Face X-Ray had promising detection performance on the generated images.

Pu et al. [17] presented *NoiseScope*, a blind deepfake image detection algorithm, for discovering GAN images among real images without utilising the GAN images beforehand for training. NoiseScope extracts any available model fingerprints or patterns that identify a GAN and utilises the extracted fingerprint to detect deepfake images. The authors evaluated NoiseScope on 11 GAN image datasets covering StyleGAN, BigGAN, PGGAN and CycleGAN models, and reported that the proposed algorithm achieved an F1 score of over 90% when detecting GAN images under different settings. They also analysed NoiseScope against a number of countermeasures such as (1) compressing images, (2) denoising using the defender's denoising filter, (3) blurring and fingerprint spoofing, and (4) recommended recovery schemes.

Editorial Comments

The source code of the work done by Pu et al. [17] is publicly available at `https://github.com/jmpu/NoiseScope`.

Nguyen et al. [16] introduced a deep 3-dimensional CNN model which can extract spatial and temporal features in a short sequence of consecutive frames to detect deepfake videos. The authors evaluated the proposed model on FaceForensics++ and VidTIMIT datasets. Experiments showed that the proposed model achieved over 94% detection accuracy for low-quality videos and over 99% for high-quality videos on both datasets. The authors reported that these results outperformed other exist-

Figure 2: Face swapping results generated by the model proposed by Li et al. [11] under various conditions.

ing detection methods.

Chintha et al. [3] introduced a convolutional, bidirectional, recurrent architecture for detecting deepfake videos, and developed an analogous architecture for detecting deepfake audio, i.e., audio spoof. The first architecture was inspired by the XceptionNet architecture and recurrent processing used in ConvLSTM (Convolutional Long Short-Term Memory) and FaceNetLSTM. The authors evaluated the deepfake video detection architecture on the FaceForensics++ and Celeb-DF datasets, and showed that it outperformed existing methods. The audio spoof detection architecture was tested on the Automatic Speaker Verification Spoofing (ASVSpoof) 2019 Challenge dataset. The experiments demonstrated that the proposed models were outperformed by the existing ensemble models although they were the best among non-ensemble models.

Editorial Comments

The experimental results of the study conducted by Chintha et al. [3] seem to be too promising. Considering the 100% accuracy values, the proposed architecture seemed to suffer from overfitting, especially on the deepfake video detection side.

Kong et al. [10] proposed a cross-modality based methodology that leveraged both face and audio information, to retrieve the hidden face behind the deepfake content. The authors relied on the assumption that the faked face preserves the surrounding face region and rich texture information, and useful clues such as age, gender and ethnicity could be obtained from the audio information. The proposed method was evaluated quantitatively and qualitatively on the VoxCeleb2 and CelebDF2 datasets, and YouTube, which was used to obtain authentic videos and audios. The experiments showed promising face reconstruction performance in terms of reconstruction quality, as well as identity and face attribute inference accuracy.

Editorial Comments

Kong et al. [10] reported that their study was the first to predict the authentic face, from all available information in the deepfake content. The proposed study has the potential to be used in several applications, such as video copyright infringement detection.

Li et al. [13] developed an online open-source platform, called *DeepFake-o-meter*, which integrates various deepfake detection methods, and provides a user interface for potential deepfake videos to be submitted for detection. The platform contains 11 state-of-the-art deepfake detection methods, including Xception, Face Warp Artifacts (FWA) and MesoNet, and supports multiple methods running simultaneously. The authors also provided an Application Programming Interface (API) to wrap individual algorithms and run on a third-party remote server.

Deepfake Prevention

Ma et al. [14] proposed a Deep Neural Network (DNN) structure, *Dynamic Enhanced Authentication Network* (DEA-Net), to extract robust lip features for visual speaker authentication, against deepfake videos. They introduced two network units:

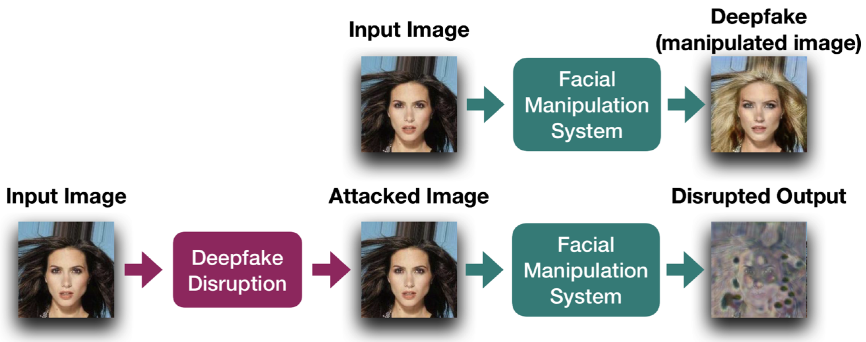


Figure 3: An illustration of deepfake disruption proposed by Ruiz et al. [18].

the Difference block and the Dynamic Response block, which enables the system to consider the user's unique talking habits. While the former decreases the impact of the static lip information in the final feature by calculating the inter-frame differences, the latter extracts pixel-level dynamic features with a global receptive field. The authors used the Global Research Identifier Database (GRID) dataset, containing talking videos of 33 speakers, for evaluating the proposed structure. They also generated fake talking videos with the Deepfakes-faceswap (`github.com/deepfakes/faceswap`) tool for the same purpose. The experiments showed that DEA-Net achieved better authentication performance compared to the existing visual speaker authentication methods.

Editorial Comments

Ma et al. [14] assumed that the attacker has limited information about the user (a number of face photos, etc.) and cannot obtain the user's unique talking habit/style when pronouncing specific words/phrases.

Ki Chan et al. [9] presented a theoretical framework, utilising a permissioned, decentralised blockchain (Hyperledger Fabric 2.0) to keep track of historical immutable records of online content, to prevent deepfakes. In this framework, online content is hashed, then deep-encoded with a triplet CNN-LSTM network which generates its discriminative features. It is then merged with hashed descriptive image or video captioning. The output generated from these steps is then stored in a permissioned blockchain as a Proof-of-Authentication (PoA). Using blockchain in the proposed framework enables users to trace media content back to its origin. It also enables artists to lay claim to their original work and to grant permission for amendments to content.

Editorial Comments

The study of Ki Chan et al. [9] involves a theoretical framework, the authors reported that they will perform experiments and complete the implementation as future work. They also reported that the proposed system has limitations, including that it requires authentic media content upon reception, and allows at most a 100MB transaction payload size.

Ruiz et al. [18] proposed a class transferable, adversarial attack against image translation systems, to prevent deepfakes from being generated. This was achieved by disrupting the resulting output image as shown in Figure 3. The proposed attack is class transferable, indicating that the attacker does not need to know the conditioning attribute (smile, eye closing etc.). The authors also explored the defences of image translation systems and introduced adversarial training for GAN, in order to alleviate disruptions. Lastly, they presented a spread-spectrum adversarial attack so that deepfakers were unable to defend against disruption by blurring the image. For the evaluation of the proposed attack, a number of image translation architectures including GANimation, StarGAN, pix2pixHD and CycleGAN were tested on the CelebA and Cityscapes datasets. The results showed that the proposed attack was successful against the given architectures.

Psychology & Deepfake

Introduction

This section includes works that investigate the impact of deepfake videos on human psychology and the defense mechanisms that could be effective against them.

Psychological Effect of Deepfake

Murphy and Flynn [15] built upon previous research and showed that while some kinds of deepfake videos may distort memory drastically, others may have only as much impact as any other form of false information (image, text, etc.). Memory distortion can happen for individuals viewing deepfakes regarding themselves. In their first experiment, participants were presented with (1) text-based fake news stories, (2) text-based fake news with a photograph, and (3) text-based fake news with a deepfake video. The deepfake videos did not increase false memory any more than the other media without video. However, participants found the videos "convincing, dangerous, and unethical". There were also indications that deepfake videos may alter memory for public events. Their experiments also confirmed that better quality deepfake videos were more convincing to participants. Limitations of the study, as reported by the authors, included the small number of videos used in the experiments, and the single exposure of the videos to the participants (internet users are likely to encounter many videos, multiple times).

Editorial Comments

One aspect of exposure to deepfake videos that is particularly challenging to guard against is when they appear mixed with undoctored or real videos. Designing experiments for such studies will also be challenging.

Vaccari and Chadwick [21] showed that observers of deepfake videos were more likely to experience heightened uncertainty, than be misled by the videos. They found that such feelings of uncertainty reduced the individual's trust in news on social media. The authors were attempting to understand how deepfakes deceive their viewers. They concluded that even though the viewers may not have been fooled by deepfakes, the overall outcome was their loss of faith in information content within the online space. They distinguished between the result of viewing deepfakes as *ambivalence* and not *uncertainty*. In other words, deepfake viewers have to choose between conflicting opinions within themselves regarding the content. When further information is provided, it increases their internal conflict. This is distinctly different from *uncertainty* that the viewer may be able to resolve with further information. The authors concluded that deepfake videos may lead to the general population becoming cynical and ambivalent. In turn this may aggravate online practices in democratic spaces.

Editorial Comments

As reported by Vaccari and Chadwick [21], the limitations of their work included the use of an online questionnaire, as opposed to field studies or a combination. They also had limited resources to create the deepfake videos, consequently these were of low quality. These areas could be improved in future studies.

Behavioural Effect of Deepfake

Ahmed [1] claimed to be one of the first to explore people's behaviour in sharing deepfake videos. Surveys were conducted in the US and Singapore to form an empirical study on how participant's behaviour, regarding sharing of deepfake videos online, related to their cognitive ability, political interest and the size of their social network. Results indicated that users who are interested in politics, also have a tendency to share deepfake videos, with the perception that they will gain socially and politically. The author asserts that cognitive skills are not directly correlated to education, but that deepfake sharing behaviour is. People with high cognitive skills were less likely to share deepfake videos. The author also states that there is no direct relationship between the size of an individual's online social network and their deepfake video sharing tendencies. However, those with a large network, as well as political interests, were significantly more inclined to share deepfake videos.

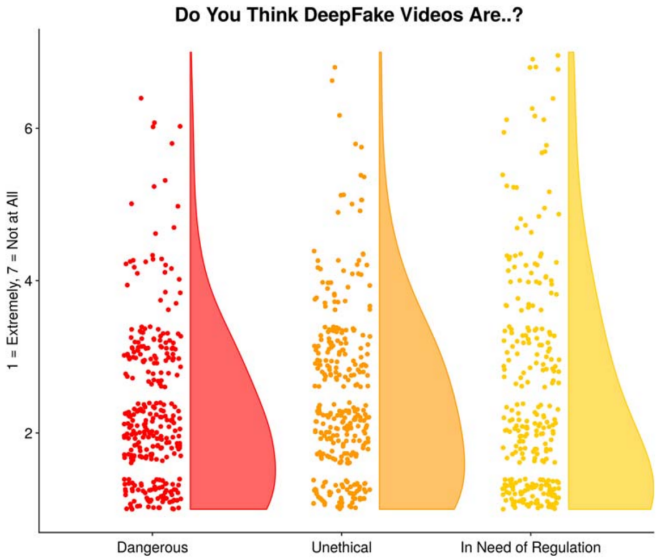


Figure 4: Murphy and Flynn [15] provided this raincloud plot illustrating participant ratings of (1) *whether deepfake videos are dangerous*, (2) *whether the creation of deepfake videos is unethical*, and (3) *whether deepfake videos are in need of regulation* (on a scale of 1 "Extremely" to 7 "Not at All").

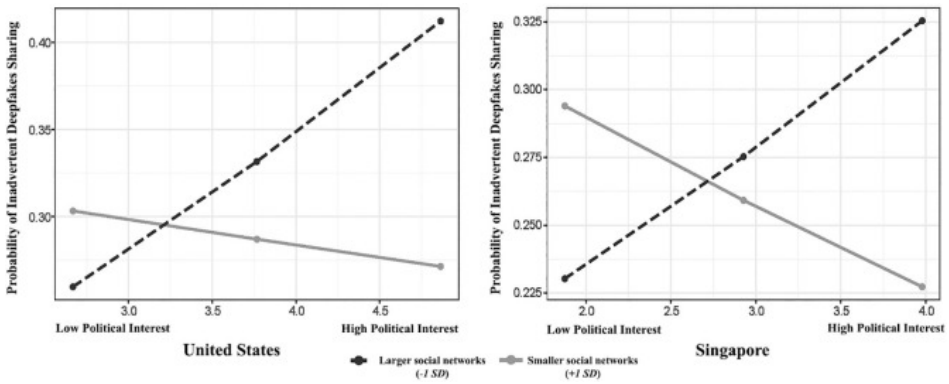


Figure 5: Graphs from Ahmed [1] showing that people with high political interest and large online social networks are more likely to share deepfake videos.

Editorial Comments

A key point from the findings of Ahmed [1] is that "the ability to distinguish the false nature of deepfakes from real information lies in the differences in the cognitive skills of online users. Perhaps, along with digital media literacy, a greater emphasis on the development of the cognitive abilities of online users is required to restrict the spread of deepfakes."

Tahir et al. [20] discussed the importance of correctly flagging deepfake videos and increasing awareness. The differences in cognition and perception of individuals makes flagging of deepfake videos challenging. Hence, as a starting point, the authors performed a preliminary survey with videos generated using three different deepfake generation algorithms and evaluated them in a baseline study with 95 individuals. During the survey, GazeCloud [8] was simultaneously used to track the gaze of participants. This gazing data was used to determine the regions of deepfake videos that the user looked at to identify their authenticity. These were used to create "a rich collection of frame coordinates containing regions of relevance given a particular type of video". The authors conducted a comparative study of deepfake detection capabilities of humans and machines by identifying regions of videos that were used for detection and how it compared with the eye gaze dataset of

the human participants. They found that the higher the quality of the deepfake videos, the more difficult they were to detect. An interesting outcome of this research was a 10-minute long deepfake detection training programme for raising awareness about detecting deepfakes. The authors presumed that the less literate people are, the more likely they are to fall for deepfakes and as such, a further study was conducted to judge the efficacy of their training programme. They report that the training was successful in increasing the accuracy of detecting deepfakes among participants by around 33%.

Editorial Comments

Tahir et al. [20] conducted a comprehensive "end-to-end" study on deepfake detection awareness. They started by building separate datasets for human capability and behaviour regarding deepfake detection and compared the performance of humans with deepfake detection algorithms. This comparison helped them devise a short, but effective, training programme to improve human awareness regarding deepfake detection.

Readily Available Deepfake Technology

Introduction

This section surveys off-the-shelf open-source, freeware and commercial software for generation or detection of deepfake content.

Deepfake Generation

One of the most popular tools for deepfake generation is *DeepFakes Web* (`https://deepfakesweb.com`). This is a web-based commercial tool used for producing fake videos. Two well-known open-source alternatives are *DeepFaceLab* (`https://github.com/iperov/DeepFaceLab`) and *FaceSwap* (`https://faceswap.dev/`) with the latter also providing deepfake image generation. Additionally, *High Resolution Face Swap* (`https://github.com/jinfagang/faceswap_pytorch`) and *DCGAN (Deep Convolutional Generative Adversarial Network) Face Generator* (`https://github.com/gsurma/face_generator`) are other examples of open-source tools for generating deepfake images. *Instagram-DeepFake-Bot* (`https://github.com/dome272/Instagram-DeepFake-Bot`) is also worth considering as it is a more contextual implementation. It is an open-source Instagram bot that produces deepfake images for Instagram users by using First Order Model for Image Animation [19]. Another example of a deepfake image generation tool is *Deep Art Effects* (`https://www.deeparteffects.com/`), released as both desktop and mobile application; it also provides an API for developers.

Several mobile applications have been released for deepfake generation. Some of these include *FaceApp* (`https://www.faceapp.com/`) (for Android/iOS), *Reface App* (`https://hey.reface.ai/`) (for Android/iOS), Avatarify (`https://avatarify.ai`) (for iOS only) and *Celebrity Face Morph* (`https://play.google.com/store/apps/details?id=com.zmobileapps.celebrityfacemorph`) (for Android only).

Deepfake Detection

Sensity (`https://sensity.ai`) is a promising free web platform for both deepfake image and video detection. It also provides advanced monitoring capabilities and a detection API for developers. An open-source alternative to Sensity is *Deepware* (`https://deepware.ai`), which has been released as a web-based tool as well as an Android application for detecting deepfake videos. Other open-source projects include a prize winning solution for DFDC (`https://github.com/selimsef/dfdc_deepfake_challenge`) that implements a detection tool for deepfake videos, and *DeepFake Audio Detection* (`https://github.com/dessa-oss/fake-voice-detection`), built with the Foundations Atlas platform, to detect deepfake audio. Lastly, *Spot the Deepfake* (`https://www.spotdeepfakes.org/en-US`) is an educational online service that aims to increase user awareness in identifying deepfakes.

Editorial Comments

The deepfake software mentioned in this section were selected by searching on Google, GitHub and Google Play Store. The searches were performed by using the keywords *deepfake* and *deep-fake* (the latter includes *deep fake*). From the search results on Google, only the first page was considered. In addition, only the GitHub repositories having at least 100 stars were included. For mobile applications, only those that had more than 100K downloads were included. Finally, software known or identified by the team members via ad-hoc methods were also included.

Deepfake in the Real World

Introduction

Incidents of deepfake in practice have steadily increased since November 2017 when it first emerged outside the academic domain in a Reddit post, most likely as a demonstration of the technology's potential to swap female celebrity faces [7]. Leveraging from the availability of off-the-shelf tools, as discussed in the previous section, the technology since then has turned into a commodity for criminals. A threat analysis report from *Recorded Future* indicated that forums in the Dark Web contain offerings of deepfake services, ranging from "training lessons on face replacement in videos" for US$20 to audio and video deepfake services according to "buyer's requirement" [7]. The report also mentions requirements such as "deepfake services to construct and design fraudulent bank cards, signatures, documents, persons (images), and card numbers that are not detectable via Google or Yandex searches" with indication of willingness to pay at least US$1500.

This section consolidates a sample of 14 web and academic sources either describing deepfake incidents in the wild, or analysing imminent deepfake risks or opportunities in specific sectors. These have been grouped into six categories.

Politics

Deepfake has been used extensively to spread fake news. However, the power of deepfake also serves the opposite purpose, i.e., to cast doubt on true evidence (e.g., videos and images). This is especially convenient for authoritarian regimes:

`https://www.wired.com/story/opinion-authoritarian-regimes-could-exploit-cries-of-deepfake/`

Deepfake can be a powerful instrument to influence outcomes both at corporate and nation-state levels. For example, Twitter profiles using deepfake images were used to try to influence labor union voting at a US-based Amazon warehouse:

`https://www.technologyreview.com/2021/03/31/1021487/deepfake-amazon-workers-are-sowing-confusion-on-twitter-thats-not-the-problem/`

The FBI has also warned of an imminent threat of foreign actors making use of deepfake technology to fabricate non-existing personas that appear "authentic to online users" to launch influence campaigns:

`https://www.lawfareblog.com/fbi-warns-deepfakes-will-be-used-increasingly-foreign-influence-operations`

Another possibility of exerting influence is via impersonation using deepfake. This situation has recently occurred when EU leaders (including UK parliamentarians) took part in Zoom video calls with a Russian opposition figure impersonated using deepfake video and audio:

`https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition`

Financial Services

Bateman [2] assessed ten scenarios which build on top of existing deepfake technology and would negatively impact the financial system. Figure 6 provides a synthesis of seven of those scenarios.

Banks have realised the threat of deepfake already as they are starting to use biometric checks in their identification and authentication systems to overcome deepfake based client impersonation:

`https://www.techregister.co.uk/banks-work-with-fintechs-to-counter-deepfake-fraud/`

Geography

Zhao et al. [22] have empirically explored the possibility of applying deepfake technology to satellite images to add non-existent landscape features (e.g., rivers, open land, and buildings). The authors used an existing GAN model, i.e., Cycle-Consistent Adversarial Networks (CycleGAN), and a dataset of satellite images of three cities to train the model. They then tested different detection models using a testing dataset containing 8,064 satellite images. Results achieved a good rate of detection (95%). However, this was a proof-of-concept exercise. The authors' key argument is the doubt it raises about trustworthiness and the potential implications of inaccurate satellite images, e.g., in terms of national security.

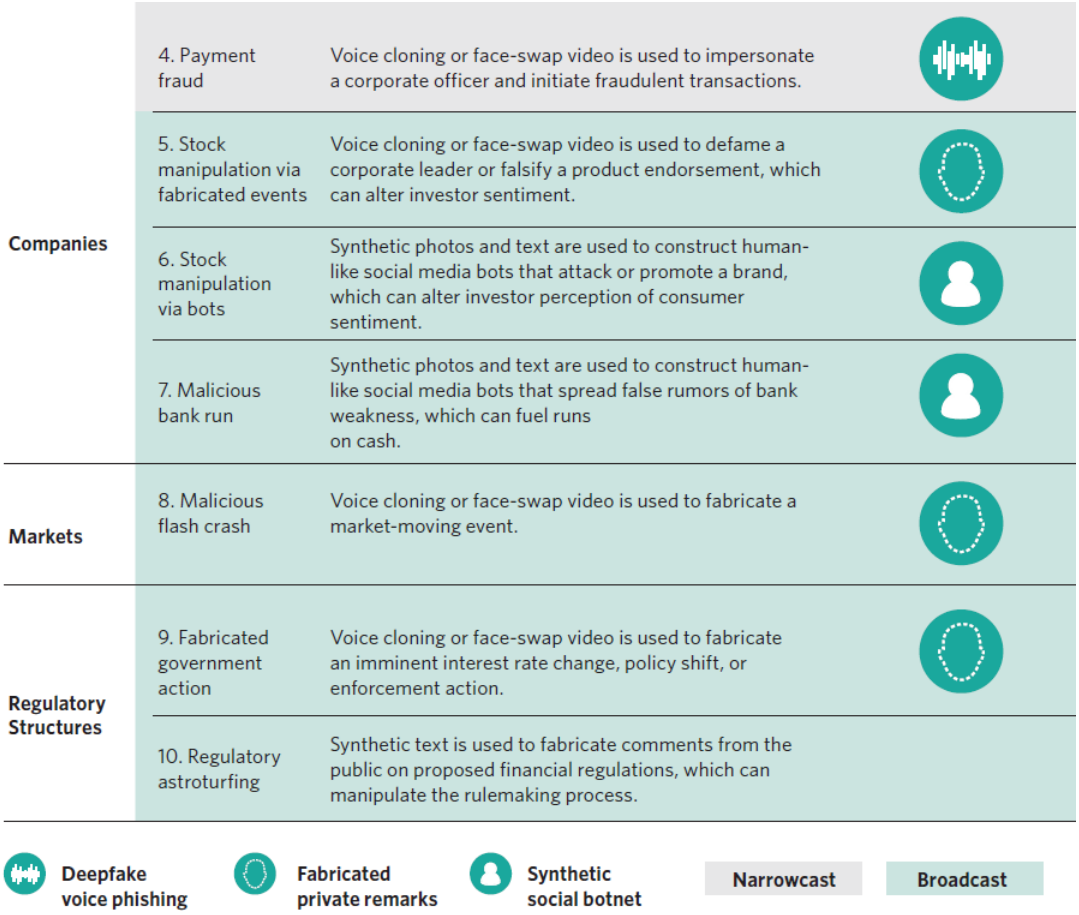


Figure 6: Threat scenarios, discussed by Bateman [2], which exploit deepfake technology in the financial system context. The first column represents the target, the second column indicates the scenario, the third column provides the role of deepfake, and the last column indicates the type of deepfake exploit – explained by the icons provided in the lower part of the figure.

Editorial Comments

Zhao et al. [22] make their code and data available at `https://figshare.com/s/eeedcd150e759ef4353c`

Damage to Reputation

Another target application of deepfake technology is the damage to reputation at corporate, government, public institution, or personal levels. The following URL described two examples centred around personal cases, which illustrate the extent of damage which may be inflicted:

`https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/`

Cybercrime

A recent Europol report [4] identified seven conceivable scenarios for the near future which exploit deepfake technologies. Some are already reality. These scenarios are: disinformation campaigns, securities fraud, extortion, online crimes against children, obstruction of justice, cryptojacking, and illicit markets. Fraud is another criminal activity which benefits from deepfake. For example, fraudsters used deepfake technology to overcome liveness detection checks in the Chinese Taxation face recognition system to issue tax invoices:

`https://findbiometrics.com/fraudsters-use-deepfake-biometrics-hack-chinas-taxation-system-040103/`

Arts, Marketing & Training

The potential of deepfake technology is not all negative. It has been embraced for very different purposes in a number of domains. One such domain is Arts. For example, deepfake has been used to generate music, complete with lyrics, in a variety of genres and artist styles. However, it remains unclear if they lead to copyright infringement:

`https://www.theguardian.com/music/2020/nov/09/deepfake-pop-music-artificial-intelligence-ai-frank-sinatra`

Deepfake has also been used in more unusual legitimate situations. The following example shows the use of deepfake to solve the problem of the absence of a self-isolating cast member on a TV programme:

`https://www.france24.com/en/tv-shows/encore/20201210-face-swap-france-s-top-soap-uses-deepfake-technology-for-self-isolating-actress`

Training and marketing campaigns are now taking advantage of deepfake technology to create videos, audio and images to an extent that might revolutionise practices. For example, deepfake has been used to achieve language coverage (i.e., inexpensive translation to different languages), diversity, ethnicity and minority representation (i.e., without the need to resource individuals). The following news articles illustrate the potential of deepfake for these domains:

`https://www.globalgoals.org/news/so-what-are-you-waiting-for-your-voice-is-powerful`

`https://www.wired.com/story/covid-drives-real-businesses-deepfake-technology/`

Editorial Comments

A powerful aspect of deepfake that makes it useful for legitimate and criminal purposes is the level of personalisation and flexibility provided by the technology for fulfilling "user requirements". This intrinsic advantage echoes the offerings observed in the Dark Web [7], discussed at the beginning of this section.

References

[1] Saifuddin Ahmed. 2021. Who Inadvertently Shares Deepfakes? Analyzing the Role of Political Interest, Cognitive Ability, and Social Network Size. *Telematics and Informatics* 57 (2021), 10. https://doi.org/10.1016/j.tele.2020.101508

[2] Jon Bateman. 2020. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Technical Report. Carnegie Endowment for International Peace. https://carnegieendowment.org/files/Bateman_FinCyber_Deepfakes_final.pdf

[3] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. 2020. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 1024–1037. https://doi.org/10.1109/JSTSP.2020.2999185

[4] Vincenzo Ciancaglini, Craig Gibson, David Sancho, Odhran McCarthy, Maria Eira, Philipp Amann, and Aglika Klayn. 2020. *Malicious Uses and Abuses of Artificial Intelligence*. Technical Report. Europol's European Cybercrime Centre (EC3). https://www.europol.europa.eu/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf

[5] Steven Lawrence Fernandes and Sumit Kumar Jha. 2020. Adversarial Attack on Deepfake Detection Using RL Based Texture Patches. In *European Conference on Computer Vision*. Springer, 220–235. https://doi.org/10.1007/978-3-030-66415-2_14

[6] Apurva Gandhi and Shomik Jain. 2020. Adversarial Perturbations Fool Deepfake Detectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207034

[7] Insikt Group. 2021. *The Business of Fraud: Deepfakes, Fraud's Next Frontier*. Technical Report. Recorded Future. https://go.recordedfuture.com/hubfs/reports/cta-2021-0429.pdf

[8] Yoshio Ishiguro and Jun Rekimoto. 2012. GazeCloud: A Thumbnail Extraction Method Using Gaze Log Data for Video Life-Log. In *2012 16th International Symposium on Wearable Computers*. IEEE, 72–75. https://doi.org/10.1109/ISWC.2012.32

[9] Christopher Chun Ki Chan, Vimal Kumar, Steven Delaney, and Munkhjargal Gochoo. 2020. Combating Deepfakes: Multi-LSTM and Blockchain as Proof of Authenticity for Digital Media. In *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*. IEEE, 55–62. https://doi.org/10.1109/AI4G50087.2020.9311067

[10] Chenqi Kong, Baoliang Chen, Wenhan Yang, Haoliang Li, Peilin Chen, and Shiqi Wang. 2021. Appearance Matters, So Does Audio: Revealing the Hidden Face via Cross-Modality Transfer. *IEEE Transactions on Circuits and Systems for Video Technology* (2021), 1–14. https://doi.org/10.1109/TCSVT.2021.3057457

[11] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing High Fidelity Identity Swapping for Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10. https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Advancing_High_Fidelity_Identity_Swapping_for_Forgery_Detection_CVPR_2020_paper.pdf

[12] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10. https://openaccess.thecvf.com/content_CVPR_

2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf

[13] Yuezun Li, Cong Zhang, Pu Sun, Honggang Qi, and Siwei Lyu. 2021. DeepFake-o-meter: An Open Platform for DeepFake Detection. arXiv:2103.02018 [cs.CV] https://arxiv.org/abs/2103.02018

[14] Jun Ma, Shilin Wang, Aixin Zhang, and Alan Wee-Chung Liew. 2020. Feature Extraction For Visual Speaker Authentication Against Computer-Generated Video Attacks. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1326–1330. https://doi.org/10.1109/ICIP40778.2020.9190976

[15] Gillian Murphy and Emma Flynn. 2021. Deepfake False Memories. *Memory* (2021), 1—13. https://doi.org/10.1080/09658211.2021.1919715

[16] Xuan Hau Nguyen, Thai Son Tran, Van Thinh Le, Kim Duy Nguyen, and Dinh-Tu Truong. 2021. Learning Spatio-Temporal Features to Detect Manipulated Facial Videos Created by the Deepfake Techniques. *Forensic Science International: Digital Investigation* 36 (2021), 8. https://doi.org/10.1016/j.fsidi.2021.301108

[17] Jiameng Pu, Neal Mangaokar, Bolun Wang, Chandan K Reddy, and Bimal Viswanath. 2020. NoiseScope: Detecting Deepfake Images in a Blind Setting. In *Annual Computer Security Applications Conference*. ACM, 913—927. https://doi.org/10.1145/3427228.3427285

[18] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. 2020. Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems. In *European Conference on Computer Vision*. Springer, 236–251. https://doi.org/10.1007/978-3-030-66823-5_14

[19] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*, Vol. 32. 11. https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf

[20] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. 2021. Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Article 174, 16 pages. https://doi.org/10.1145/3411764.3445699

[21] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (2020), 1–13. https://doi.org/10.1177/2056305120903408

[22] Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun, and Chengbin Deng. 2021. Deep Fake Geography? When Geospatial Data Encounter Artificial Intelligence. *Cartography and Geographic Information Science* (2021), 15. https://doi.org/10.1080/15230406.2021.1910075