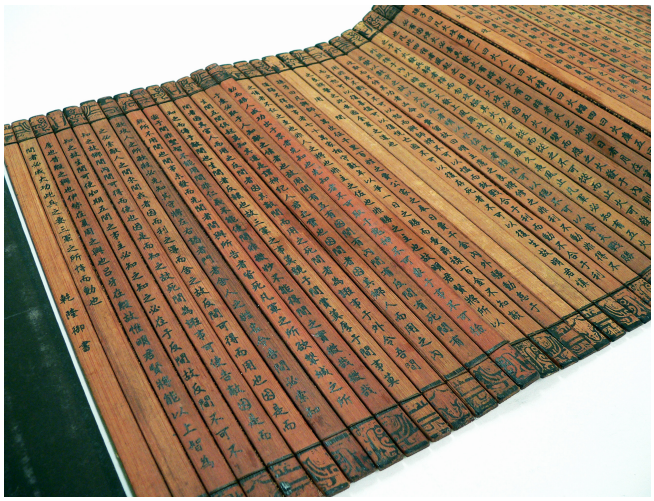
May 2021, Issue Code NL-2022-1-C

DDD (Digital Data Deception) Technology Watch Newsletter: Chinese Addendum

Table of Contents



Source: `https://www.flickr.com/photos/bluefootedbooby/370460130/`

“兵者，诡道也。故能而示之不能，用而示之不用，近而示之远，远而示之近。利而诱之，乱而取之，实而备之，强而避之，怒而挠之，卑而骄之，佚而劳之，亲而离之。攻其无备，出其不意。此兵家之胜，不可先传也。”

— 孙武:《孙子兵法·始计篇》

(The above is the original Chinese version of the English quotation, shown on the cover page of the newsletter's main issue covering English papers.)

Editors: Shujun Li, Sanjay Bhattacherjee, Enes Altuncu, and Virginia Franqueira
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 recent progress on research and the current use of the technology by non-researchers. For this issue, we continue to include a Chinese addendum covering selected research papers published in Chinese. For the use of deepfake by non-researchers, we chose to search for relevant Chinese web pages mentioning deepfake, using the Google web search engine.

The research papers covered in this Chinese addendum were identified mainly via keyword-based searches into the English scientific database Scopus (`https://www.scopus.com/`) and a similar Chinese database – the China Online Journals (COJ, 中国学术期刊数据库, `https://c.wanfangdata.com.cn/periodical`, which is part of Wanfang Data – 万方数据知识服务平台 – provided by Wangfang Data Co., Ltd.). Considering the term "deepfake" and the corresponding Chinese term "深度伪造" are well-established and widely used, we decided to mainly use their two keywords and their variants for our search queries. A similar term "deep forgeries" was also considered because some researchers used it instead of "deepfake". On Scopus, the following single search query was used:

```
deepfake* OR deep-fake* OR "deep fake*" OR "deep forger*"
```

On COJ, two separate search queries with a single word were used, one with the English keyword "deepfake" and the other with the corresponding Chinese term "深度伪造". In total 8 papers were selected, including 2 survey papers, 1 paper on deepfake generation, 3 papers on deepfake detection, and 2 papers on deepfake-related crime and legal issues. A 2020 special issue on deepfake in the Journal of Cyber Security (JoCS, 《信息安全学报》) was identified from the searches (`http://jcs.iie.ac.cn/xxaqxb/ch/reader/issue_list.aspx?year_id=2020&quarter_id=2`). Some papers from this JoCS special issue have been covered before ([4, 5, 10] of NL-2021-5-C), and one paper is less about deepfake but more on adversarial AI. In total, we selected 3 papers from this JoCS special issue. An ad hoc search into the JoCS website was also conducted using the Chinese keyword for "deepfake" ("深度伪造"), which led to the selection of an additional paper.

The Google search for Chinese web pages mentioning deepfake focused more on existing software tools and online services and their use in real-world applications. In total, 8 separate search queries were conducted, by combining the English and Chinese words for deepfake and four Chinese words "软件", "工具", "平台" and "服务" that mean "software", "tool", "platform" and "service", respectively. Web pages returned were screened and inspected manually to extract relevant tools and services, until Google Search showed the following message "In order to show you the most relevant results, we have omitted some entries very similar to the ... already displayed." Some separate Google search queries were also conducted to search for official web pages of some tools and services, based on names and keywords mentioned on some of the returned web pages.

As in previous issues, for each paper covered we provide an objective summary of the research and our more subjective editorial comments. We also paid attention to datasets used and source code releases, but did not observe any. This consolidated the previous observation that Chinese researchers seem less active in releasing their source code and data in public.

Following the practice of the previous issue, the References section lists only the 8 research papers selected for this special issue. We embed relevant URLs and other references cited in the main text.

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

List of Acronyms

- ASV: Automatic Speaker Verification
- ASVspoof: Automatic Speaker Verification Spoofing and Countermeasures Challenge
- BLSTM: Bidirectional LSTM
- CNN: Convolutional Neural Network
- DBN: Deep Belief Network
- DTW: Dynamic Time Warping
- GMM: Gaussian Mixture Models
- HMM: Hidden Markov Models
- HOG: Histogram of Oriented Gradient
- LBP: Local Binary Pattern
- LSTM: Long and Short-Time Memory
- RMOAVIS: Regulation for Managing Online Audio and Video Information Services (《网络音视频信息服务管理规定》) of China
- RNN: Recurrent Neural Network
- VC: Voice Conversion

Survey Papers

Introduction

We identified a number of survey papers on deepfake in the Chinese research literature, indicating an increased level of general interests among Chinese researchers on this topic in recent years. We decided to include two survey papers in this section, one covering deepfake generation and detection in general, and the other one focusing on a narrower area of speech forgery and detection. Note that in a previous issue we covered one general survey paper on deepfake ([11] of NL-2021-3-C).

Deepfake Generation and Detection – General Survey

Li (李旭嵘) et al. [1] conducted a comprehensive survey on deepfake, covering the following areas: 1) deepfake generation methods for facial images (face swapping and fake facial expressions) and speech forgeries, 2) public datasets for deepfake research (facial videos and speech spoofing), 3) deepfake detection methods, and 4) deepfake and adversarial AI. The authors' discussions on deepfake generation are relatively brief, however, they did a more detailed review of existing off-the-shelf tools (see Table 1 in the paper). The paper covers 12 deepfake datasets (see Table 2 in the paper), but the coverage on speech forgeries is relatively thin (just 2 datasets). The paper covers deepfake detection in much greater detail, and splits reviewed methods into the following six categories: i) traditional image forensic methods based on signal processing techniques in spatial and spectral domains, ii) analysis of biometric characteristics, iii) detection of visual inconsistencies caused by manipulation to original real facial images, iv) analysis of characteristics of the deepfake GAN (generative adversarial network) used, v) data-driven learning at the frame level and for the whole video, vi) detection of deep speech forgeries. The authors gave a brief comparison of the five categories of visual deepfakes, as shown in Table 1. They looked at the relationships between deepfake and adversarial AI from two angles: a) deepfake generation based on adversarial samples; b) robustness of deepfake detectors against adversarial samples. After reviewing related work, the authors also gave their opinions on different types of risks about deepfake (reputational damage of people, negative impacts on face recognition based systems, negative impact on digital evidence processing, negative impact on economic activities), and major technical difficulties related to (lossy) compression, diverse resolutions, unknown deepfake generation methods, and complicated adversarial AI scenarios. The authors also gave suggestions for future research, including more generalisable and robust detectors, more proactive and preventative methods, and joint detection of audio-visual deepfakes. They also recommended more work in three other broader areas: establishing a wider research community, deepfake-aware legislation, and training journalists about deepfakes.

Editorial Comments

Although Li (李旭嵘) et al. attempted to conduct a more comprehensive survey, the pa-

Table 1: Comparison of the five types of visual deepfake detection methods summarised in [1].

Category	Key Features	Weaknesses
i	Mature techniques, explainable features	Image deepfakes only, lossy compression could introduce issues
ii	Focusing on local biometric features	High error rate for compressed videos, some features not always available, low accuracy
iii	Local features, more effective for low-resolution deepfakes	Less generalisable, low accuracy
iv	Very specific features (GAN characteristics)	High dependency on data and knowledge of the generation method, less generalisable
v	Big data, richer features, high accuracy	High dependency on good datasets, higher impact of unknown data types and compression

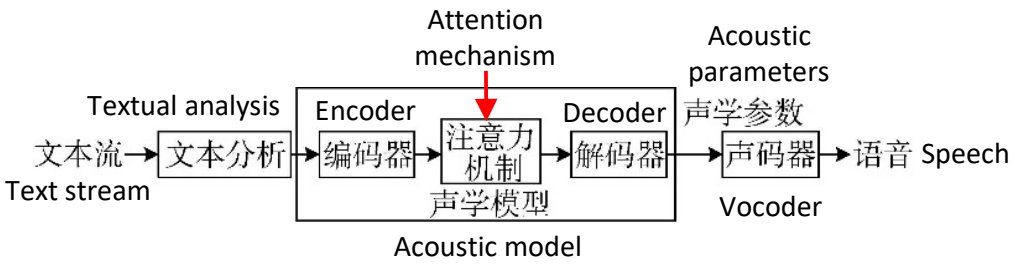


Figure 1: The end-to-end speech forgery framework in [5].

per's coverage is skewed heavily towards deepfake facial images and videos. The coverage of deep speech forgeries is relatively thin and other types of visual deepfakes are not covered. The paper does not cover the deepfake terminology (e.g., the origin of the term and other alternative terms) and general architectures of deepfake generation and detection methods. There is also little discussion on how different deepfake detection methods compare with each other quantitatively.

Speech Forgery and Detection – More Focused Survey

Tao (陶建华) et al. [5] did a more focused survey covering speech forgeries and detection, including those methods based on deep learning. They classified speech forgeries into four different types: 1) speaker style (voiceprint) forgeries, where the aim is to create forged speeches simulating the (learned) style of a target speaker; 2) vocal timbre forgeries, where the aim is to create high-quality forged speeches of a target speaker, e.g., based on an end-to-end framework shown in Figure 1; 3) voice prosody forgeries, where the aim is to simulate a target speaker's personalised features such as general attitude and attention while speaking; 4) speech simulation, where the aim is to manipulate some attributes of a given speech signal towards a target speaker's style, without manipulating the content and background of the input speech signal. For all the four types of speech forgeries, earlier methods are mostly based on traditional techniques, such as DTW (dynamic time warping), HMM (hidden Markov models) and GMM (Gaussian mixture models), but recently, deep learning models, such as DBN (Deep Belief Network), LSTM (long and short-time memory), RNN (recurrent neural network) and GAN, have been more widely used to outperform traditional methods. Regarding speech forgery detection, the authors classified relevant methods into four classes: 1) those detecting parametric speech synthesis methods; 2) those detecting waveform-based speech synthesis; 3) those detecting speech simulation; and 4) those based on deep learning. Based on the review of related work, the authors also discussed major challenges and future research trends for speech forgeries (multi-style forgeries, low-cost forgeries, and robust forgeries) and for speech forgery detection (more datasets, better generalisability, and more explainable and highly robust detectors).

Editorial Comments

This paper covers speech forgery generation more than detection. While listing construction of more datasets as a major future research direction, the paper does not cover existing datasets. It mentions ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures Challenge) 2017 and 2019 competitions (`https://www.asvspoof.org/`), but does not give a detailed description of the competitions and datasets used. Note that this competition started in 2015 and runs biannually, and has become a major benchmarking event for speech forgery detection.

Deepfake Generation

Introduction

For this category we identified only one paper. We speculate that many researchers may have considered positive applications of deepfake and therefore tried to avoid using the more negative-leaning term "deepfake" in their papers on deepfake generation. Other alternative terms they may have used include "deep synthesis" and "photo-realistic". We will consider expanding the search keywords in future, if we decide to cover the topic of deepfake generation again. Note that this terminological issue is limited to deepfake generation, since the other topic of deepfake detection naturally refers to more malicious applications of such techniques.

Deep Speech Forgeries

Miao (苗晓孔) et al. [4] proposed an enhanced voice conversion (VC) method for generating harder-to-detect deep speech forgeries, based on more accurate conversion of the Mel cepstrum. The enhancement is achieved via 1) the use of a bidirectional long and short-time memory (BLSTM) network with improved depth residuals, 2) an optimised loss function, and 3) the introduction of the global mean filter to filter out the cepstrum clutter. The overall architecture of the proposed method is shown in Figure 2, in which the shaded boxes indicate major changes proposed. As a result, the authors managed to achieve an increased similarity of the deepfake speech and the original speech, without reducing the subjective quality perceived by human ears. They tested the performance of the proposed method against two automatic speaker verification (ASV) systems (based on i-vector and x-vector, respectively) to see if they could still recognise converted speeches. The experiments were based on the CMU ARCTIC speech synthesis datasets (`http://www.festvox.org/cmu_arctic/`), under the "male to female" setting. The results showed that the proposed method defeated both ASV systems perfectly

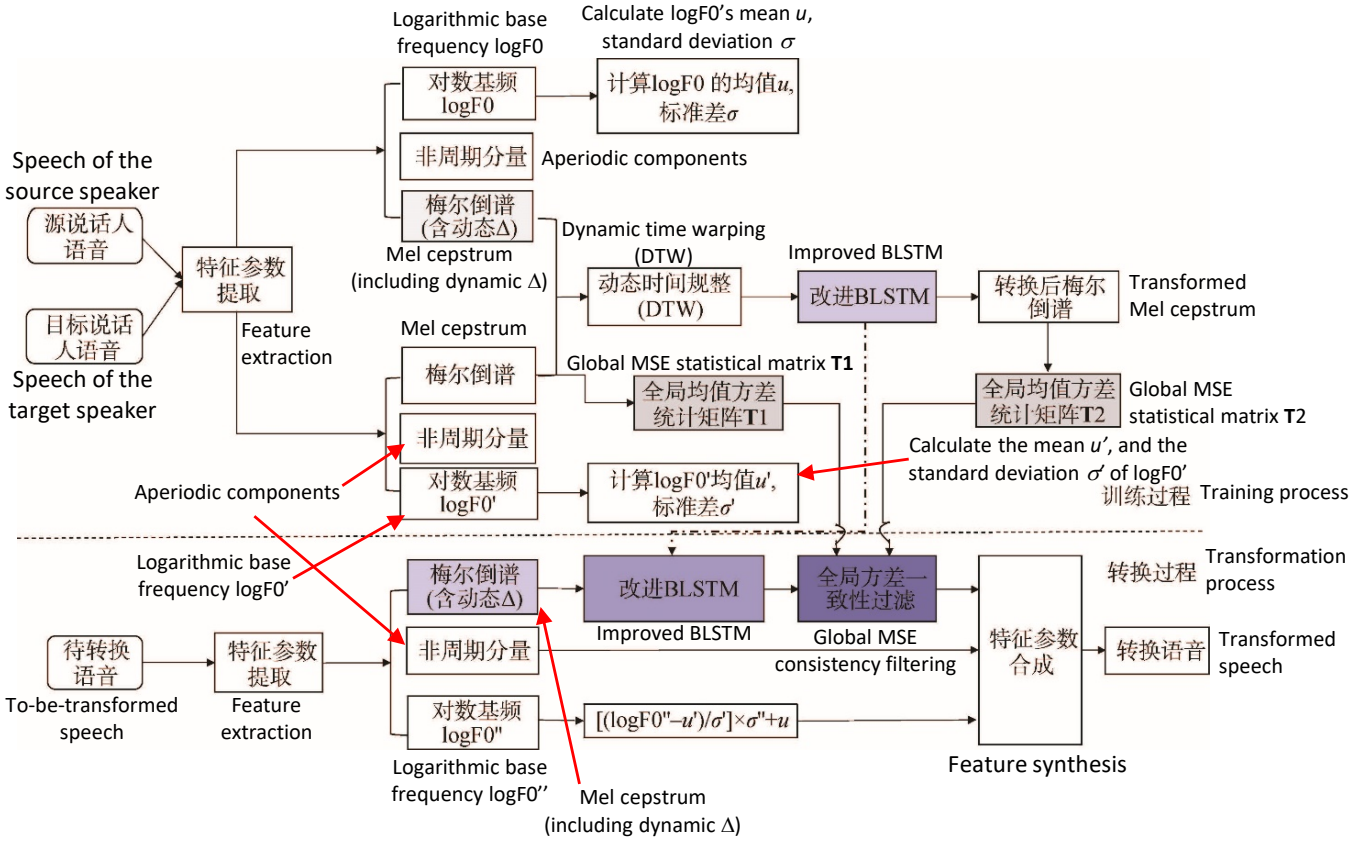


Figure 2: The architecture of the proposed deep speech forgery method in [4].

(i.e., with a 100% success rate). The authors compared their method with two state-of-the-art VC methods, one based on Gaussian mixture models (GMM, https://doi.org/10.1109/ICASSP.2015.7178896), and another based on BLSTM (without the proposed improvement). They showed that their method could better preserve both objective and subjective quality of the converted speeches.

Editorial Comments

Miao (苗晓孔) et al. did not clearly explain the rationale for the selection of the two specific state-of-the-art VC methods for comparison. The GMM-based method was ranked second in the Voice Conversion Challenge (VCC) 2018 (http://www.vc-challenge.org/vcc2018/), so not the best-performing method. For the BLSTM-based method, the authors referred to the VCC 2018 challenge report, which does not explicitly mention such a method. The Related Work section of [4] refers to a 2012 paper for the use of BLSTM as a general method (https://doi.org/10.1109/ICASSP.2012.6288834), which may be the actual method the authors meant for the BLSTM-based method they chose for comparison.

The authors did not release their source code, so the reported results cannot be easily validated.

Introduction

Our literature search returned a number of papers focusing on deepfake detection. All papers are about detection of deepfake facial images or videos. This narrow focus is not totally surprising given the fact that face swapping and transformation remain the most popular deepfake applications. This also echoes the narrow focus of the survey paper [1] on facial images and videos for visual deepfakes. In future, we will consider covering other sub-topics of deepfake detection, which will likely require the use of a richer set of keywords and more scientific databases, to have a wider coverage of the research literature.

Of those papers identified, two were covered in a previous issue ([4, 10] of NL-2021-3-C). We decided to cover three papers, two on automatic deepfake detection, and the third on human-based deepfake detection for forensic analysis.

Automatic Deepfake Detection

Zhang (张怡暄) et al. [8] proposed two different methods for detecting deepfake facial videos, based on the observation that the inter-frame differences of the facial region in deepfake videos tend to be higher than those in real videos. Both methods are based on the general framework shown in Figure 3, but with different feature sets. The first method is a more traditional one, based on two sets of manually defined features: LBP (local binary pattern) and HOG (histogram of oriented gradient). The second method uses a range of deep learning models to automatically extract features: a Siamese network working with LightCNN, Inception net or ResNet. To show the performance of the proposed framework and the two different feature extraction methods, the authors conducted experiments with two different datasets: FaceForensics++ (`http://kaldir.vc.in.tum.de/faceforensics_benchmark/`) and TIMIT (VidTIMIT `https://conradsanderson.id.au/vidtimit/` and DeepfakeTIMIT `https://www.idiap.ch/en/dataset/deepfaketimit`). The experimental results showed that the deep learning based method significantly outperformed the first method with the performance figures ranging between 91-99% for the FaceForensics++ dataset and between 90-100% for the TIMIT dataset, at the video detection level. The authors also conducted a comprehensive set of experiments to test different configurations of the deep learning based method, showing consistent high performance. They also compared the efficiency of the two methods, and observed that the deep learning based method is much faster for both training and testing phases. The results indicate that the deep learning based method has overwhelming better performance in all areas. The authors also compared the proposed method with 10 state-of-the-art methods, showing either a better or comparable performance.

Editorial Comments

This paper proposes a relatively straightforward idea and a general framework, which

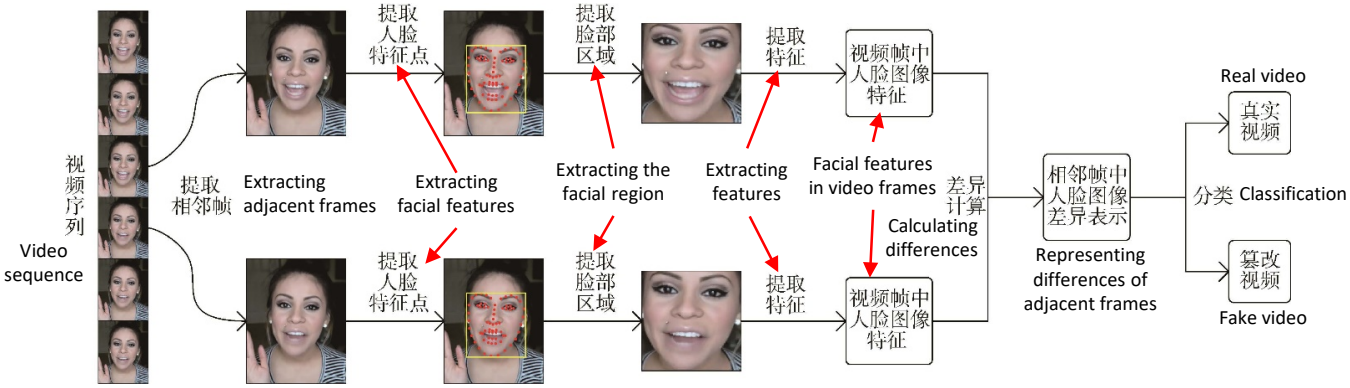


Figure 3: The framework for detecting deepfake facial videos based on inter-frame differences proposed in [8].

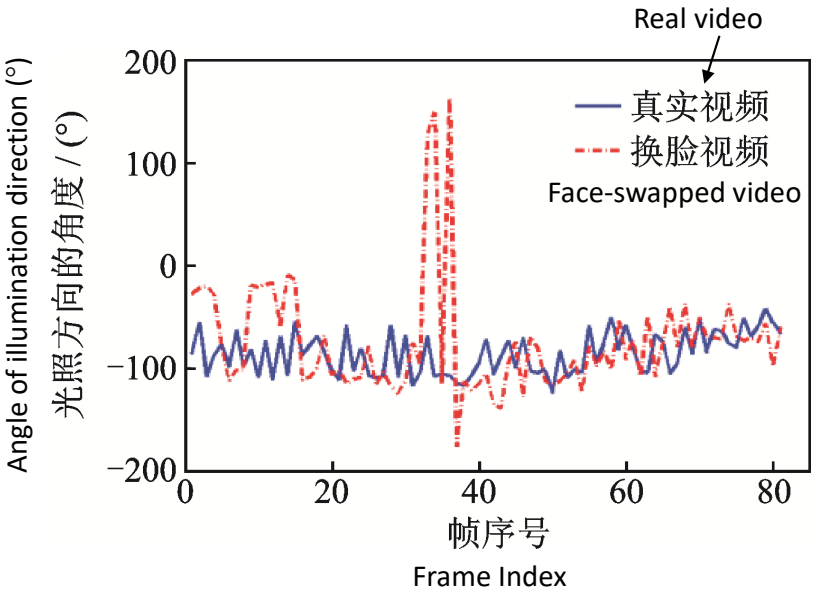


Figure 4: Comparison of angular changes of the illumination direction over time, for a real video and its corresponding deepfake version, as evidence supporting the detection method proposed in [2].

seems very extendable. The design and the experiments conducted look very comprehensive. The performance figures are very promising and also relatively stable across different settings. Although no source code is provided, the general framework proposed should be relatively easy to follow for independent implementations.

Li (李纪成) et al. [2] proposed to detect deepfake facial videos based on observable inter-frame inconsistencies of illumination. The idea is based on the assumption that even very advanced deepfake models cannot precisely model the complicated lighting conditions of the external environment, thus leaving traces of inconsistent illumination across all frames of the whole video. The method uses the Lambert illumination model to calculate the 2-D illumination direction for each frame, and checks if the angular change of the direction is smooth enough to detect deepfake videos. Figure 4 shows a comparison of the angular changes of the illumination direction for a real video and its deepfake version, from which we can see the deepfake video has more abrupt changes at times than real videos. The experiments were conducted with the same two datasets as in [8] (FaceForensics++ and TIMIT). The results showed that the proposed method performed better on the FaceForensics++ dataset (96.6%) than on the TIMIT dataset (85.3% for high-quality videos and 88.6% for low-quality videos). The performance was also compared with three state-of-the-art methods in the literature, and the proposed method had the best performance.

Editorial Comments

When the framework proposed in [8] is compared with the work reported in [2], the former seems less effective. The experimental results in [8] are also less comprehensive, so further validation will be needed. If the better performance of the general framework in [2] is genuine, it may be explained qualitatively by the fact that the illumination directional changes across frames could actually be covered by some features extracted from inter-frame differences.

Human-based Deepfake Detection

Wang (王怡) and Yang (杨洪臣) [6] conducted a conceptual analysis on the different clues a forensic examiner can follow to detect deepfake facial videos. They considered five different categories of forensic analysis: metadata, traces of manipulation (eye blinking, shadows, skin colour, facial shape, positions of key facial features, unnatural facial features, and inter-frame inconsistencies of facial features),

violation from photographic principles (view depth and illumination), overall analysis (e.g., joint audio-visual analysis), and context-based analysis considering the background information of the case under study. This paper does not provide experimental results or case studies, so it may be seen as a pseudo-survey.

Editorial Comments

This is not primarily a technical paper – all discussions are around human-conducted forensic analysis (with assistance of relevant digital forensic tools). Some of the methods discussed may help inform human-machine teaming based deepfake detection.

Deepfake Use by Non-Researchers

Introduction

As mentioned in the editorial on Page 1 of this issue, we used Google Search with a series of keywords-based search queries to identify relevant Chinese pages mentioning deepfake. Most web pages returned were news reports, general reviews or commentaries on deepfake. Some web pages, including some news reports, are about published research papers. Both deepfake generation and detection tools and services are mentioned in returned web pages. These two types of tools and services are reviewed in the following two subsections. In the last subsection, we give some editorial comments for the whole section.

Please note that our Google search results may not have covered all deepfake-related software tools and online services used by Chinese users, but they offer a representative subset.

Deepfake Generation Tools and Services

As we expected, many Chinese web pages mentioned deepfake generation tools developed by non-Chinese researchers, developers or companies, especially those freely available on mobile app markets and open-source repositories (e.g., GitHub). Notable examples include DeepFaceLab (`https://github.com/iperov/DeepFaceLab`), Faceswap (`https://faceswap.dev/`), Avatarify (`https://avatarify.ai/`), Fakeapp (`https://www.fakeapp.org/`), FaceApp (`https://www.faceapp.com/`), Reface (`https://reface.app/`), MyVoiceYourFace (`https://myvoiceyourface.com/`), MyFakeApp (`https://bitbucket.org/radeksissues/myfakeapp`), Botika (`https://botika.io/`), Synthesia AI video generator (`https://www.synthesia.io/`), MyHeritage Deep Nostalgia (`https://www.myheritage.com/deep-nostalgia`), DeepArtEffect (`https://www.deeparteffects.com/`), and DeepFake-o-meter (`http://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/`).

For deepfake generation tools produced by Chinese researchers and developers, and those developed for Chinese users, ZAO (`https://apps.apple.com/cn/app/zao/id1465199127`) is the tool mostly discussed. ZAO is a mobile app operated by the Changsha Shenduronghe Network Technology Co., Ltd (长沙深度融合网络科技有限公司), which is a company associated with the popular Chinese online social media app Momo (陌陌, `http://www.immomo.com/`). Like many other deepfake apps, ZAO leverages face swapping deepfake technologies to enable a number of deepfake-based features, such as generating deepfake videos by replacing a famous actor's face with the user's own face in a TV show or movie, or allowing the user to try different hairstyles and costumes. More about this tool can be found on its Chinese Wikipedia page (`https://zh.wikipedia.org/wiki/ZAO`).

(a) The input image (b) The output image

Figure 5: An example of the face swapping feature of the Meitu AI Open Platform, created by us using a public domain image (`https://pixabay.com/photos/woman-headscarf-face-girl-lady-918776/`) as the input.

Another class of software tools and online services related to deepfake are for photo and video editing. Typical examples are Meitu XiuXiu or MeituPic (美图秀秀, `https://pc.meitu.com/`) and the Meitu AI Open Platform (美图 AI 开放平台, `https://ai.meitu.com/`) developed by Meitu Inc. (美图公司). The original purpose of such tools and services is to "beautify" imperfect photos or to create highly personalised photos and videos with special effects. Some special effects are deepfake-based, e.g., the Meitu AI Open Platform has a face swapping feature (`https://ai.meitu.com/algorithm/imageEditing/facefuse?t=1622475696499`, see Figure 8 for an example). Another very similar online service is Tencent Cloud Shentu, which has two deepfake-based features – Face Transformation

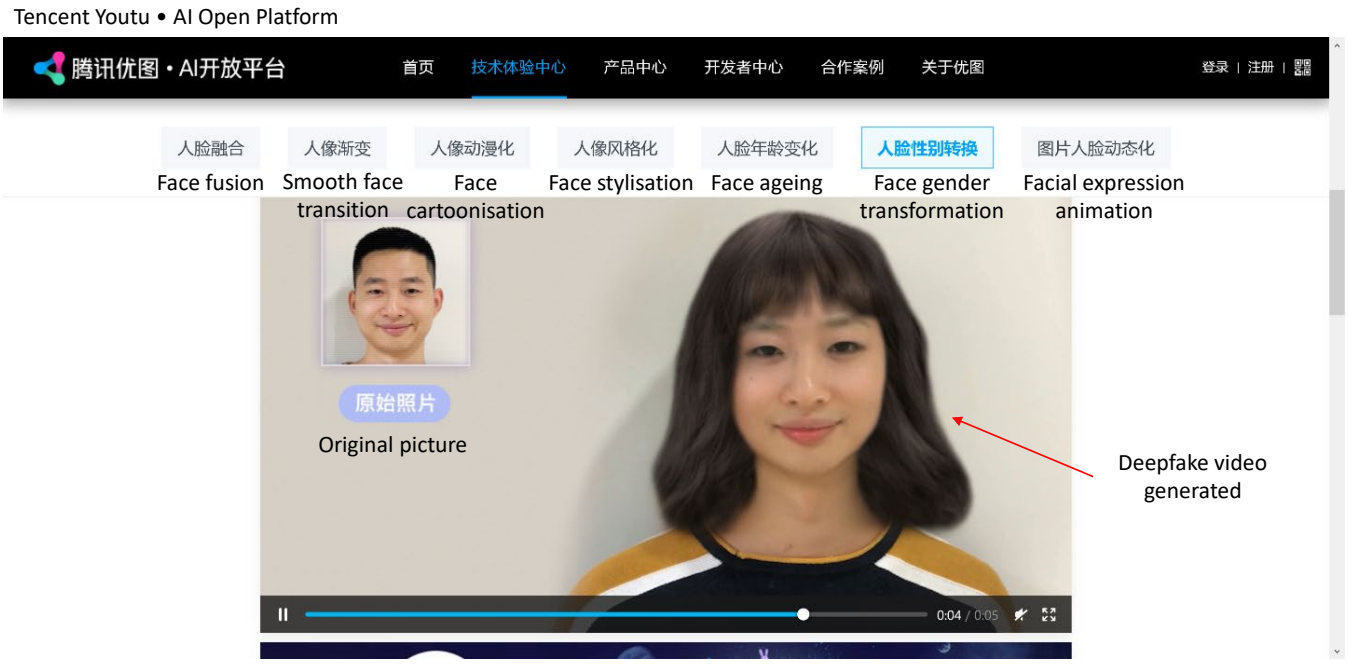


Figure 6: A screenshot of the web page of the Tencent Youtu Open AI Platform. The page shows an example of the face gender transformation, and the second-level navigation menu shows other deepfake-related special effects the platform supports, most of which seem to be based on deepfake techniques.

(腾讯云神图·人像变换, `https://cloud.tencent.com/product/ft`) and Face Fusion (腾讯云神图·人脸融合, `https://cloud.tencent.com/product/facefusion`), which allow generation of photo-realistic but fake facial images. In addition to the online services on Tencent Cloud, Tencent also provides a separate Tencent Youtu AI Open Platform (腾讯优图·AI 开放平台, `https://open.youtu.qq.com/`) for developers to use. Tencent Cloud Shentu services seem to be based on techniques from the Tencent Youtu AI Open Platform. Figure 6 shows a screenshot of the open platform, with an example of deepfake-based face gender transformation.

Another application of deepfake in China is cloud-based services for generating synthesised graduation photos (云毕业照). Such services generate synthetic graduation photos based on uploaded facial images of the user and their classmates and teachers. They have become popular among students due to the COVID-19 pandemic, which has prevented many students from having a proper graduation ceremony and physical opportunities of taking graduation photos in the real world. One such service was provided by Tencent Cloud (腾讯云, `https://cloud.tencent.com/developer/article/1650422`) in the summer of 2020.

Deepfake Detection Tools and Services

Similar to the case of deepfake generation, for deepfake detection, most web pages we inspected referred to software tools and online services developed by non-Chinese researchers, developers or companies. Notable examples include the Google-led Project Assembler (`https://projectassembler.org/`), Microsoft Video Authenticator (`https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/`), Sensity Detection API (`https://sensity.ai/api-2/`), and Deepware Scanner (`https://scanner.deepware.ai/`).

A number of Chinese online services are also mentioned in some web pages we inspected. These included Baidu Security Deepfake Face Detection platform (百度安全深度换脸检测平台, `https://anquan.baidu.com/product/deepfake`), Baidu AI Open Platform's Synthetic Face Detection tool (百度 AI 开放平台合成图检测, `https://ai.baidu.com/tech/face/spoofing`), Tecent Cloud Anti-Deepfake (腾讯云换脸甄别, `https://cloud.tencent.com/product/atdf`), and RealAI's DeepReal Deepfake Content Detection platform (DeepReal 深度伪造内容检测平台, `https://deepfakes.real-ai.cn/`). Some web pages also mentioned deepfake detec-

Figure 7: The cover page of the 2020 Tencent AI White Paper (source: `https://tech.sina.com.cn/roll/2020-07-14/doc-iivhvpwx5201226.shtml`). The white paper was jointly produced by Tencent Research Institute (腾讯研究院, `https://www.tisi.org/`) and Tencent Youtu (腾讯优图, `https://open.youtu.qq.com`). Tencent Research Institute is the research arm of Tencent, covering all areas of research activities including AI. Tencent Youtu is the brand name of the Tencent Computer Vision R&D Center (腾讯计算机视觉研发中心), which was known as Tencent Youtu Lab (腾讯优图实验室) before 2018.

tion features in NetEase Yidun (网易易盾, `https://dun.163.com/`), but we did not find an explicit mention of such features on the system's official web site. Most such systems are paid services but some offer free trials (e.g., the services provided by Tencent and Baidu).

Editorial Comments

One major observation for deepfake generation is that most software tools and online services have a legitimate purpose, e.g., entertainment, fashion, productivity enhancement, family life, and digital arts. It is therefore not surprising to see that many Chinese companies are very active in producing deepfake generation tools and services. Considering many positive applications of deepfake techniques, the 2020 Tencent AI White Paper (《腾讯人工智能白皮书: 泛在智能》, `https://tech.sina.com.cn/roll/2020-07-14/doc-iivhvpwx5201226.shtml`, see Figure 7 for its cover page) suggests the use of the more neutral term "deep synthesis" to replace the negative-leaning term "deepfake" (see Section 4.3 of the white paper). This suggestion is not surprising for a company like Tencent, which has a lot of interest in the entertainment sector, e.g., Tencent Pictures (腾讯影业, `https://en.wikipedia.org/wiki/Tencent_Pictures`) and Tencent Animation and Comics (腾讯动漫, `https://en.wikipedia.org/wiki/Tencent_Animation_and_Comics`), in addition to its main business in the ICT sector.

When looking at the content of the Chinese web page mentioning deepfake software tools and online services used by non-researchers, we can see a

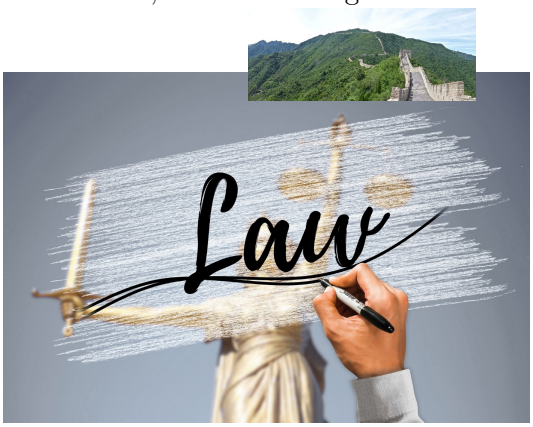
clear trend of rapid commercialisation. Many Chinese companies active in the ICT sector (e.g., Tencent and Baidu) are regularly publishing research papers on deepfake generation and detection, and they have regularly stated that their latest research results are being actively incorporated into their products and services. In addition, some university researchers have transferred their research quickly through university spin-offs, e.g., RealAI's DeepReal platform is developed by a university spin-off of the Institute for Artificial Intelligence at the Tsinghua University (清华大学人工智能研究院, `http://ai.tsinghua.edu.cn/`). This rapid commercialisation trend is not unique for Chinese companies, but a common phenomenon in the ICT sector across the world.

An interesting trend in commercialisation of deepfake is that many major ICT companies in China have decided to welcome the open source ideology. A good example is Tencent, which set up its Open Source Management Office (开源管理办公室) in 2018 (`https://kuaibao.qq.com/s/20190627A0140C00`). The open source projects Tencent has released include some from Tencent Youtu (`https://www.infoq.cn/article/wo0a9qtwivtgwwjatsfg`), e.g., a number of open frameworks and libraries that can support development of deepfake-related applications. Most such open source projects are released on GitHub, which is the most actively used open source repository among AI researchers and developers. The repository of Tencent on GitHub is available at `https://github.com/Tencent`, and that of Tencent Youtu at `https://github.com/TencentYoutuResearch`. One deepfake-related supporting library from Tencent Youtu is FaceAttribute-FAN (`https://github.com/TencentYoutuResearch/FaceAttribute-FAN`), which can be used to detect 40 different facial attributes as a pre-processing step of face swapping or face fusion.

Deepfake-related Crime and Legal Issues

Introduction

Originally, we hoped to identify some Chinese research papers on psychological aspects of deepfake. However, our research queries did not return any psychology-related papers. Instead, a number of papers on deepfake-related crime and legal issues were returned. We expected that this topic would be of interest to our readers, so we decided to switch to this topic and selected two representative papers. Before delving into descriptions of the two papers, we first introduce a recently passed, deepfake-oriented regulation in China, as useful background.



Collectively, the two papers indicate the need to refine current criminal law legislation in China and operational guidelines to better manage deepfake-related crime, without unnecessarily harming legitimate use of the deepfake technology. While the discussions focused on Chinese law, we believe that the general principles also hold for other jurisdictions. Given the fact that the UK has not introduced any deepfake-specific legislation, the recommendations given in the two papers can help inform law makers in the UK.

A notable missing point in both papers is the legitimate and defensive use of deepfake, e.g., a user creates a deepfake profile image to avoid being targeted by criminals on online social networks. This can make it even harder to define the legality and criminality of the use of deepfake.

Deepfake-related Legislation in China

Following the precedence in the US, where the Malicious Deep Fake Prohibition Act (`https://www.congress.gov/bill/115th-congress/senate-bill/3805`) and the DEEP FAKES Accountability Act (`https://www.congress.gov/bill/116th-congress/house-bill/3230`) were introduced at the federal level in 2018 and 2019, the Cyberspace Administration of China (CAC), China's Ministry of Culture and Tourism and National Radio and Television Administration jointly defined the Regulation for Managing Online Audio and Video Information Services (RMOAVIS, 《网络音视频信息服务管理规定》, `http://www.cac.gov.cn/2019-11/29/c_1576561820967678.htm`), which became effective from 1st January 2020. This regulation was defined with potential abuse and misuse of deepfake in mind, and several articles explicitly refer to generation and spreading of deepfake-based false information:

- Article 10 requires service providers to conduct a security evaluation for deepfake-based services;
- Article 11 bans the use of deepfake for creation of fake news and requires service providers to clearly mark deepfake-generated media;
- Article 12 requires service providers to deploy deepfake detection techniques and to remove deepfake-based false information once detected;
- Article 13 requires service providers to establish mechanisms to debunk deepfake-based rumour;
- Articles 12 and 13 both require service providers to report deepfake-based false information to relevant authorities.

Article 8 of this regulation also asks service providers to follow China's Network Security Act to verify real identities of their users (who create user-generated content). The regulation does not explain what this real-identity verification mechanism is for, but it can clearly help deter malicious use of deepfake images and videos and facilitate investigation of such media. While this regulation has deepfake in mind, its text does not actually use the term "deepfake", but "deep learning". This is aligned with the call in the 2020 Tencent AI White Paper to not use the term "deepfake".

Two Selected Papers

Li (李腾) [3] presented his view on how China's criminal law should be applied to address challenges and risks caused by deepfake. He first discussed three aspects: 1) the necessity of applying criminal law to address deepfake's harm to individuals and the society as a whole; 2) the appropriateness of applying criminal law by arguing the easy access of deepfake tools and the difficulties of detecting deepfake; and 3) the feasibility of applying criminal law by looking at related legislation in other countries (mainly the USA). For the third aspect, the author also argued that the RMOAVIS defined in 2019 does not include sufficient details to be very practical, e.g., it is unclear if user-generated deepfake videos that do not represent or spread fake news or rumour should be clearly marked or banned, and the RMOAVIS does not clearly define obligations of actual content creators (users on many online services) or how non-compliance can be managed. Based on the discussions, the author further proposed to consider duality of deepfake (i.e., legitimate/benign and illegitimate/malicious application) at the macro level and preventative measures at the micro level. Finally, the author proposed a concrete "personal data protection + application management + platform monitoring" approach, for applying criminal law to deepfake-related crime and online harms.

Editorial Comments

The duality of deepfake means more than just the two applications of the technology (legitimate/benign and illegitimate/malicious). It is also about the complicated interactions between the two sides, e.g., a deepfake mobile app may be created for a legitimate purpose, but is later misused by malicious parties for illegitimate purposes. Some applications by nature have a mixed flavour of being legitimate and illegitimate. For instance, the DeepNude app had sparked questions on moral and ethical responsibilities of the devel-

2020 年 10 月
第 10 期(总第 286 期)

中 州 学 刊
Academic Journal of Zhongzhou

Oct., 2020
No.10

【法学研究】

“深度伪造”技术的刑法规制体系构建 *

李　腾

摘　要:作为人工智能应用的重点领域,“深度伪造”技术借助于“深度学习”技术和“生成式对抗网络”模型,能对影音图像进行高精度合成。刑法规制“深度伪造”技术的必要性在于该技术易被滥用;正当性源于对该技术的自主性、便捷性、逼真性等特征及其隐患的审视;可行性基于域外有对该技术进行规制的立法经验。我国刑法规制“深度伪造”技术应秉持以下立场:以客观、中立的态度审视该技术的风险与收益,合理控制刑法介入的深度,以发挥刑法的预防机能为目标。具体可采取三条规制路径:通过强化个人生物识别信息的刑法保护,从前端防范“深度伪造”技术被滥用;通过增设“身份冒用罪”,强化个人身份的刑法保护,规范“深度伪造”技术的合理使用;通过“通知—删除”“发现—标识”“发现—删除”义务的履行,强化网络平台的刑事责任,预防“深度伪造”的危害后果传播。

关键词:“深度伪造”;社会危害;刑法介入;应然立场;规制路径

中图分类号:D924.1　　文献标识码:A　　文章编号:1003-0751(2020)10-0053-10

一、“深度伪造”技术带来实践挑战

2019 年 8 月底,一款名为 ZAO 的智能换脸软件爆红网络,该软件依托“深度伪造”(deep fake)技术,用户只需在该软件的应用程序中注册并上传自己的头像照片,即可将视频中人物的头像更换为自己的。该款智能应用程序的推广,使得“换脸”对公众而言不再是难以企及的高端科技,任何人都可以享受“换脸”带来的娱乐,甚至体验一次成为经典影视剧主角的快感。但是,仅仅数天后,该款应用程序设计团队便因用户隐私协议不规范、存在数据泄露风险等网络安全问题而被工信部约谈,随即该款应用程序被微信屏蔽访问。[①]智能换脸应用程序的“天折”隐含全社会对“深度伪造”技术被滥用的担忧。

在美国,“深度伪造”技术走红同样引发诸多道德和法律问题。早在 2017 年 11 月,美国社交新闻网站红迪网(Reddit)的一位用户便通过“深度伪造”技术,将电影《神奇女侠》中女主角加尔·加多(Gal Gadot)的脸嫁接到色情电影中,并将该视频上传至网络,引发了极大轰动。2018 年 1 月,“深度伪造”技术的应用程序正式上线,换脸的对象从公众人物迅速扩展至亲友、同学、同事之间,加剧了合成视频在网络的传播,红迪网不得不关闭“深度伪造”技术讨论区以减少该技术被滥用的可能性。同年,加拿大蒙特利尔大学三名博士联合创办的一家名为“琴鸟”(Lyrebird)的公司更是开发出了“语音合成”软件,只要对目标人物的声音进行 1 分钟以上的录音,再将录音“喂”给该软件进行学习,就能得到一个特别的密钥,利用这个密钥就可以用目标人物的声音伪造出任何想说的话。[②]由此,经过“深度伪造”技术处理的视频、声音成为具有高度欺骗性的影音资料,使人们不仅陷入“耳听为虚”的窘境,更陷入“眼见不为实”的认知混乱,而换脸技术与语音合成技术的叠加适用更增加了合成视频的辨别难度,也加剧了各界对“深度伪造”技术被滥用的担忧。

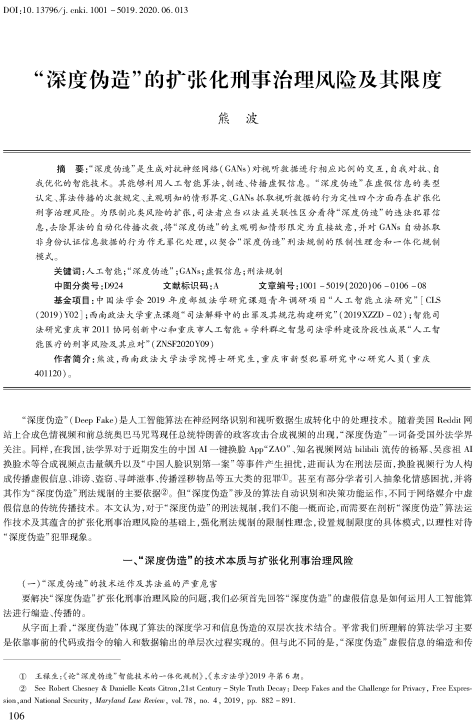
作为人工智能技术发展的新兴领域,“深度伪造”技术确实能增加社交的互动性、娱乐性,这种功效与国家大力扶持人工智能产业发展的初衷是一致

收稿日期:2020-07-09

*基金项目:国家社会科学基金一般项目“民营企业的刑事法律风险及刑法保护研究”(19BFX072)。

作者简介:李腾,男,中国政法大学刑事司法学院博士后研究人员(北京　100088),荷兰鹿特丹大学访问学者。

53



DOI:10.13796/j.cnki.1001-5019.2020.06.013

“深度伪造”的扩张化刑事治理风险及其限度

熊　波

摘　要:“深度伪造”是生成对抗神经网络(GANs)对视听数据进行相应比例的交互、自我对抗、自我优化的智能技术。其能够利用人工智能算法,制造、传播虚假信息。“深度伪造”在虚假信息的类型认定、算法传播的次数规定、主观明知的情形界定、GANs 抓取视听数据的行为定性四个方面存在扩张化刑事治理风险。为限制此类风险的扩张,司法者应当以法益关联性区分肯待“深度伪造”的违法积极信息,去除算法的自动化传播次数,将“深度伪造”的主观明知情形限定为直接故意,并对 GANs 自动抓取非身份认证信息数据的行为作无罪化处理,以契合“深度伪造”刑法规制的限制性理念和一体化规制模式。

关键词:人工智能;“深度伪造”;GANs;虚假信息;刑法规制

中图分类号:D924　　文献标识码:A　　文章编号:1001-5019(2020)06-0106-08

基金项目:中国法学会 2019 年度部级法学研究课题青年调研项目“人工智能立法研究”[CLS(2019)Y02];西南政法大学重点课题“司法解释中的出罪及其规范构建研究”(2019XZZD-02);智能司法研究重庆市 2011 协同创新中心和重庆市人工智能+学科群之智慧司法学科建设阶段性成果“人工智能医疗的刑事风险及其应对”(ZNSF2020Y09)

作者简介:熊波,西南政法大学法学院博士研究生,重庆市新型犯罪研究中心研究人员(重庆　401120)。

“深度伪造”(Deep Fake)是人工智能算法在神经网络识别和视听数据生成转化中的处理技术。随着美国 Reddit 网站上合成色情视频和前总统奥巴马咒骂现任总统特朗普的政客攻击合成视频的出现,“深度伪造”一词备受国外法学界关注。同样,在我国,法学界对于近期发生的中国 AI 一键换脸 App“ZAO”、知名视频网站 bilibili 流传的杨幂、吴彦祖 AI 换脸术等合成视频点击量飙升以及“中国人脸识别第一案”等事件产生担忧,进而认为在刑法层面,换脸视频行为人构成传播虚假信息、诽谤、盗窃、寻衅滋事、传播淫秽物品等五大类的犯罪①。甚至有部分学者引入抽象化情感困扰,并将其作为“深度伪造”刑法规制的主要依据②。但“深度伪造”涉及的算法自动识别和决策功能运作,不同于网络媒介中虚假信息的传统传播技术。本文认为,对于“深度伪造”的刑法规制,我们不能一概而论,而需要在剖析“深度伪造”算法运作技术及其蕴含的扩张化刑事治理风险的基础上,强化刑法规制的限制性理念,设置规制限度的具体模式,以理性对待“深度伪造”犯罪现象。

一、“深度伪造”的技术本质与扩张化刑事治理风险

(一)“深度伪造”的技术运作及其法益的严重危害

要解决“深度伪造”扩张化刑事治理风险的问题,我们必须首先回答“深度伪造”的虚假信息是如何运用人工智能算法进行编造、传播的。

从字面上看,“深度伪造”体现了算法的深度学习和信息伪造的双层次技术结合。平常我们所理解的算法学习主要是依靠事前的代码或指令的输入和数据输出的单层次过程实现的。但与此不同的是,“深度伪造”虚假信息的编造和传

① 王福生:《论“深度伪造”智能技术的一体化规制》,《东方法学》2019 年第 6 期。

② See Robert Chesney & Danielle Keats Citron, 21st Century-Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security, *Maryland Law Review*, vol. 78, no. 4, 2019, pp. 882-891.

106

(a) Li (李腾)'s work [3]　　(b) Xiong (熊波)'s work [7]

Figure 8: The first page of the two papers covered in this section [3, 7].

oper (https://www.vice.com/en/article/qv7agw/deepnude-app-that-undresses-photos-of-women-takes-it-offline) and its use of such an app (particularly for fake revenge porn purposes) may not be legal in many jurisdictions.
While this paper does not cover any technological aspects, the proposed "personal data protection + application management + platform monitoring" approach can inspire researchers and technologists to develop useful software tools, for monitoring the use of deepfake techniques to provide protection to victims, e.g., automatically detect fake revenge porn posted on social media and inform victims and law enforcement agencies.

Xiong (熊波) [7] looked at the risks and limitations of enlarged criminality when applying criminal law to manage deepfake-related offences. He first discussed a number of reasons why there are risks of enlarged criminality: 1) simply treating deepfake as illegal information, as defined in criminal law, does not take into consideration the fact that deepfake's harm is often more limited than other criminal activities; 2) automated visits to and spreading of deepfake should be excluded for defining the threshold of criminality; 3) the definition of deepfake distributors' intent may be too broad and vague; and 4) automated collection of data should be considered more flexibly. The author suggested that we should follow more restrictive principles when applying criminal law to deepfake-related crime, in order to protect freedom of speech, to adhere to criminal law's proportionality principle and technical characteristics of deepfake generation methods, and to avoid harming legitimate online resource sharing. Finally, based on the four reasons of enlarged criminality, the authors proposed four corresponding operational guidelines: 1) applying analysis of legal benefits to selectively consider different types of deepfake information; 2) excluding automated visits to and sharing of deepfake information; 3) restricting the intent of deepfake distributors to direct intent; and 4) avoiding criminalising automated collection of non-personal data for deepfake generation.

Editorial Comments

For the third reason of the extended criminality ("the definition of deepfake distributors' intent may be too broad and vague"), Xiong (熊波) focused on the intent of deepfake distributors only. While this is an interesting area to consider, the intent of deepfake *creators* should also be covered. Although one may argue that deepfake creators always have a direct intent (since they know the created data is fake), their intent is not necessarily illegitimate or malicious. As an example, most users of MyHeritage Deep Nostalgia (https://www.myheritage.com/deep-nostalgia) should have a very legitimate reason to use the online service, but created deepfake videos could be distributed by malicious parties for illegitimate purposes.

References

[1] Xurong Li (李旭嵘), Shouling Ji (纪守领), Chunming Wu (吴春明), Zhenguang Liu (刘振广), Shuiguang Deng (邓水光), Peng Cheng (程鹏), Min Yang (杨珉), and Xiangwei Kong (孔祥维). 2021. Survey on Deepfakes and Detection Techniques / 深度伪造与检测技术综述. *Journal of Software* /《软件学报》 32, 2 (2021), 496–518. http://www.jos.org.cn/1000-9825/6140.htm

[2] Jicheng Li (李纪成), Beibei Liu (刘琲贝), Yongjian Hu (胡永健), Yufei Wang (王宇飞), Guangjun Liao (廖广军), and Guangyao Liu (刘光尧). 2020. Deepfake Video Detection Based on Consistency of Illumination Direction / 基于光照方向一致性的换脸视频检测. *Journal of Nanjing University of Aeronautics and Astronautics* /《南京航空航天大学学报》 52, 5 (2020), 760–767. https://jnuaa.nuaa.edu.cn/njhkht/article/html/202005012

[3] Teng Li (李腾). 2020. Construction of Criminal Law Regulation System of "Deep Forgery" Technology / "深度伪造"技术的刑法规制体系构建. *Academic Journal of Zhongzhou* /《中州学刊》 42, 10 (2020), 53–62. http://www.zzxk1979.com:8085/CN/Y2020/V42/I10/53

[4] Xiaokong Miao (苗晓孔), Meng Sun (孙蒙), Xiongwei Zhang (张雄伟), Jiakang Li (李嘉康), and Xingyu Zhang (张星昱). 2020. Deep Speech Forgery Based on Parameter Transformation and Threat Assessment to Voiceprint Authentication / 基于参数转换的语音深度伪造及其对声纹认证的威胁评估. *Journal of Cyber Security* /《信息安全学报》 5, 6 (2020), 53–59. http://jcs.iie.ac.cn/xxaqxben/ch/reader/view_abstract.aspx?file_no=20200605&flag=1

[5] Jianhua Tao (陶建华), Ruibo Fu (傅睿博), Jiangyan Yi (易江燕), Chenglong Wang (王成龙), and Tao Wang (汪涛). 2020. Development and Challenge of Speech Forgery and Detection / 语音伪造与鉴伪的发展与挑战. *Journal of Cyber Security* /《信息安全学报》 5, 2 (2020), 28–38. http://jcs.iie.ac.cn/xxaqxb/ch/reader/view_abstract.aspx?file_no=20200204&flag=1

[6] Yi Wang (王怡) and Hongchen Yang (杨洪臣). 2021. Analysis of Fake Videos Generated with AI-manipulated Replacement of Face Image / 一种 AI 换脸方法生成的伪造视频分析. *Forensic Science and Technology* /《刑事技术》 46, 1 (2021), 16–22. http://www.xsjs-cifs.com/article/2021/1008-3650-46-1-16.html

[7] Bo Xiong (熊波). 2020. Risks and Limitations of Enlarged Criminality Management of Deepfakes / "深度伪造"的扩张化刑事治理风险及其限度. *Journal of Anhui University (Philosophy and Social Science Edition)* /《安徽大学学报（哲学社会科学版）》 44, 6 (2020), 106–113. http://ahdxzsb.paperopen.com/oa/DArticle.aspx?type=view&id=202000613

[8] Yixuan Zhang (张怡暄), Gen Li (李根), Yun Cao (曹纭), and Xianfeng Zhao (赵险峰). 2020. A Method for Detecting Human-face-tampered Videos based on Interframe Difference / 基于帧间差异的人脸篡改视频检测方法. *Journal of Cyber Security* /《信息安全学报》 5, 2 (2020), 49–72. http://jcs.iie.ac.cn/xxaqxben/ch/reader/view_abstract.aspx?file_no=20200206&flag=1