软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(2):496−518 [doi: 10.13328/j.cnki.jos.006140]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

深度伪造与检测技术综述*

李旭嵘[1,2], 纪守领[1], 吴春明[1,3], 刘振广[4], 邓水光[1], 程 鹏[5], 杨 珉[6], 孔祥维[7]

[1](浙江大学 计算机科学与技术学院,浙江 杭州 310007)

[2](阿里巴巴,浙江 杭州 311121)

[3](之江实验室,浙江 杭州 310000)

[4](浙江工商大学 计算机与信息工程学院,浙江 杭州 310018)

[5](浙江大学 控制科学与工程学院,浙江 杭州 310007)

[6](复旦大学 计算机科学技术学院,上海 201203)

[7](浙江大学 管理学院,浙江 杭州 310007)

通讯作者: 纪守领, Email: sji@zju.edu.cn

摘 要: 深度学习在计算机视觉领域取得了重大成功,超越了众多传统的方法.然而近年来,深度学习技术被滥用在假视频的制作上,使得以 Deepfakes 为代表的伪造视频在网络上泛滥成灾.这种深度伪造技术通过篡改或替换原始视频的人脸信息,并合成虚假的语音来制作色情电影、虚假新闻、政治谣言等.为了消除此类伪造技术带来的负面影响,众多学者对假视频的鉴别进行了深入的研究,并提出一系列的检测方法来帮助机构或社区去识别此类伪造视频.尽管如此,目前的检测技术仍然存在依赖特定分布数据、特定压缩率等诸多的局限性,远远落后于假视频的生成技术.并且不同学者解决问题的角度不同,使用的数据集和评价指标均不统一.迄今为止,学术界对深度伪造与检测技术仍缺乏统一的认识,深度伪造和检测技术研究的体系架构尚不明确.回顾了深度伪造与检测技术的发展,并对现有研究工作进行了系统的总结和科学的归类.最后讨论了深度伪造技术蔓延带来的社会风险,分析了检测技术的诸多局限性,并探讨了检测技术面临的挑战和潜在研究方向,旨在为后续学者进一步推动深度伪造检测技术的发展和部署提供指导.

关键词: 深度学习;深度伪造;假视频;取证;检测技术

中图法分类号: TP309

中文引用格式: 李旭嵘,纪守领,吴春明,刘振广,邓水光,程鹏,杨珉,孔祥维.深度伪造与检测技术综述.软件学报,2021,32(2):496−518. http://www.jos.org.cn/1000-9825/6140.htm

英文引用格式: Li XR, Ji SL, Wu CM, Liu ZG, Deng SG, Cheng P, Yang M, Kong XW. Survey on deepfakes and detection techniques. Ruan Jian Xue Bao/Journal of Software, 2021,32(2):496−518 (in Chinese). http://www.jos.org.cn/1000-9825/6140.htm

Survey on Deepfakes and Detection Techniques

LI Xu-Rong[1,2], JI Shou-Ling[1], WU Chun-Ming[1,3], LIU Zhen-Guang[4], DENG Shui-Guang[1], CHENG Peng[5], YANG Min[6], KONG Xiang-Wei[7]

[1](College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China)

* 基金项目: 国家重点研发计划(2018YFB0804102, 2020YFB1804705); 浙江省自然科学基金(LR19F020003); 浙江省重点研发计划(2019C01055, 2020C01021); 国家自然科学基金(61772466, U1936215, U1836202); 前沿科技创新专项(2019QY(Y)0205)

Foundation item: National Key Research and Development Program of China (2018YFB0804102, 2020YFB1804705); Zhejiang Provincial Natural Science Foundation (LR19F020003); Zhejiang Provincial Key Research and Development Program (2019C01055, 2020C01021); National Natural Science Foundation of China (61772466, U1936215, U1836202); Frontier Science and Technology Innovation Project (2019QY(Y)0205)

收稿时间: 2020-05-07; 修改时间: 2020-06-22; 采用时间: 2020-08-27; jos 在线出版时间: 2020-09-10

[2](Alibaba Group, Hangzhou 311121, China)
[3](Zhejiang Lab, Hangzhou 310000, China)
[4](College of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China)
[5](College of Control Science and Engineering, Zhejiang University, Hangzhou 310007, China)
[6](College of Computer Science, Fudan University, Shanghai 201203, China)
[7](College of Management, Zhejiang University, Hangzhou 310007, China)

Abstract: Deep learning has achieved great success in the field of computer vision, surpassing many traditional methods. However, in recent years, deep learning technology has been abused in the production of fake videos, making fake videos represented by Deepfakes flooding on the Internet. This technique produces pornographic movies, fake news, political rumors by tampering or replacing the face information of the original videos and synthesizes fake speech. In order to eliminate the negative effects brought by such forgery technologies, many researchers have conducted in-depth research on the identification of fake videos and proposed a series of detection methods to help institutions or communities to identify such fake videos. Nevertheless, the current detection technology still has many limitations such as specific distribution data, specific compression ratio, and so on, far behind the generation technology of fake video. In addition, different researchers handle the problem from different angles. The data sets and evaluation indicators used are not uniform. So far, the academic community still lacks a unified understanding of deep forgery and detection technology. The architecture of deep forgery and detection technology research is not clear. In this review, the development of deep forgery and detection technologies are reviewed. Besides, existing research works are systematically summarize and scientifically classified. Finally, the social risks posed by the spread of Deepfakes technology are discussed, the limitations of detection technology are analyzed, and the challenges and potential research directions of detection technology are discussed, aiming to provide guidance for follow-up researchers to further promote the development and deployment of Deepfakes detection technology.

Key words: deep learning; Deepfakes; fake video; forensics; detection technique

近年来,以 Deepfakes[1]为代表的换脸技术开始在网络兴起.此类技术可将视频中的人脸替换成目标人物,从而制作出目标人物做特定动作的假视频.随着深度学习技术的发展,自动编码器、生成对抗网络等技术逐渐被应用到深度伪造中.由于Deepfakes技术只需要少量的人脸照片便可以实现视频换脸,一些恶意用户利用互联网上可获取的数据,生成众多的假视频并应用在灰色地带,如将色情电影的女主角替换成女明星,给政客、公司高管等有影响力的人伪造一些视频内容,从而达到误导舆论、赢得选取、操纵股价等目的.这些虚假视频内容极其逼真,在制作的同时往往伴随着音频的篡改,使得互联网用户几乎无法鉴别.如果这些深度伪造的内容作为新闻素材被制作传播,这会损害新闻机构的声誉和公众对媒体的信心.更深层次的,当遇到案件侦查和事故取证时,如果缺乏对 Deepfakes 类虚假影像资料的鉴别,将对司法体系产生巨大的挑战.尽管深度伪造技术有其积极的一面,如“复活”一些去世的人进行影视创作,以及 Zao APP[2]提供大众换脸娱乐服务等,但是目前负面影响远远大于正面,拥有鉴别此类深度伪造视频的能力变得尤为重要.

为了尽量减少深度伪造技术带来的影响,消除虚假视频的传播,学术界和工业界开始探索不同的深度伪造检测技术.相继有学者构造数据集,展开对 Deepfakes 检测的多角度研究.脸书公司也联合微软一起举办全球 Deepfakes 检测竞赛[3]以推动检测技术的发展.然而这些 Deepfakes 检测工作各有侧重,存在众多局限性.针对本领域的综述工作还比较缺乏,只有针对早期图像篡改工作的一些总结[4,5],亟需对现有工作进行系统的整理和科学的总结、归类,以促进该领域的研究.

本文第 1 节介绍深度伪造的各种相关技术.第 2 节列举出当下深度伪造研究的数据集.第 3 节对现有的深度伪造检测技术进行系统的总结和归类.第 4 节讨论深度伪造生成和检测技术的双面对抗性.第 5 节总结面临的挑战和未来可行的研究方向.最后,第 6 节对全文的工作进行总结.

1 深度伪造生成技术

现有的深度伪造图像主要是指脸部的篡改,而脸部篡改伪造主要分为两大类:一类是换脸伪造,通过交换两张图像的人脸达到人身份修改的目的,其技术从传统的 3D 重建方法发展到现在以生成对抗网络为基础的深度伪造;另一类是脸部表情属性伪造,迁移指定表情等动作到目标图像而不修改目标人脸标志,达到伪造表情或者

特定动作目的,其技术也从基于 3D 的图形学方法演变到最新的深度学习方法.此外,制作深度伪造素材时通常还包含了语音的伪造,使得欺骗效果更佳.本节将对这些伪造生成技术进行概述,其中重点关注深度伪造技术,并总结了一些开源的生成工具.

1.1 换脸伪造技术

1.1.1 基于图形学的伪造

在过去 10 多年里,基于图形学的人脸篡改技术一直被研究者所关注,Zollhofer 等人[6]综述了当前比较主流的 3D 模型重建追踪等技术.FaceSwap[7]是基于图形学的换脸方法,首先获取人脸关键点,然后通过 3D 模型对获取到的人脸关键点位置进行渲染,不断缩小目标形状和关键点定位间的差异,最后将渲染模型的图像进行混合,并利用色彩校正技术获取最终的图像.Kevin 等人[8]提出了在视频里自动换脸的 3D 方法,不需要大量的手动操作和硬件采集,只需要一个单相机视频,通过用 3D 多线性模型追踪视频中的人脸,并用相应的 3D 形状将源人脸仿射到目标人脸.Pablo 等人[9]用类似的 3D 方法来替换目标视频中演员的人脸,而保留原始的表情.Pablo 等人[10]还设计了一个系统,通过高质量的 3D 人脸捕捉技术,改变人脸从而匹配嘴巴的动作.Nirkin 等人[11]用分割的思路促进换脸,通过网络分割出来的人脸估计 3D 人脸形状,最后融合源和目标这两个对齐的 3D 人脸形状.

1.1.2 基于学习的伪造

尽管基于图形学的脸部篡改方法研究了多年,但是时间开销大、门槛高、成本大,使得这项技术很难普及.随着深度学习技术的飞速发展,研究者们开始关注深度学习在人脸篡改上的应用[12].Deepfakes[1]是网络上较早开源的基于深度学习的换脸项目,如图 1 所示,训练两个自动编码器,两个编码器共享权重参数,使得两个解码器学会重建人脸的能力.训练结束后,在换脸阶段,交换两个解码器,从而使得换脸效果达成.这只需要具备原人物和目标人物的人脸图片即可训练,大大降低了使用门槛.但是也需要一定的训练技巧,否则生成器的生成质量无法保障.鉴于此,研究者们开始关注 GAN[13]技术的融合,Faceswap-GAN[14]就是增加了 GAN 技术的 Deepfakes,引入判别器的对抗损失函数,在生成的时候判别生成图像和原图的相似度,使得生成的图像质量有大幅度提高,另外引入了感知损失函数增加眼珠的转动效果.GAN 技术的加入使得换脸更加逼真自然,也一定程度增加了深度伪造技术的流行度.

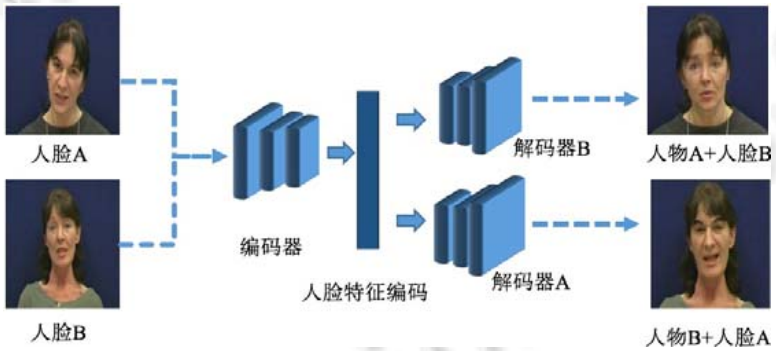

Fig.1 Framework for Deepfakes generation[1]

图 1 Depefakes 生成框架[1]

Korshunova 等人[15]将换脸问题视为风格迁移问题,训练一个卷积神经网络,从非结构化的图片中学习这种外观,并设计内容损失和风格损失函数来保障生成高质量真实度的人脸图像.这些人脸转换还是依赖于大量的源和目标人物的人脸图片训练,泛化性不强.Yuval 等人[16]基于 GAN 技术提出了一个主体无关的人脸替换和重建方法,通过引入特定域感知损失、重建损失和对抗损失,可以应用于成对的人脸,不需要在大量人脸上训练.除换脸外,GAN 技术还被广泛用于生产虚拟的人脸和篡改人脸属性.如 StarGAN[17]、Stackgan[18]、PGAN[19]等一系列 GAN 技术可以生成虚假的人脸,Grigory 等人[20]利用 conditional-GAN[21]技术改变人的年龄,Rui 等人[22]利

用GAN生成不同的人脸视角而保持全局的结构和局部细节.GAN技术的发展使得人脸的生成和属性篡改都越来越真实,这也给人脸伪造的滥用留下了空间.

1.2 表情伪造技术

表情伪造是指不改变人脸的属性,迁移其他人脸图像的表情到目标人脸,从而达到目标人物做指定表情的目的.Thies 等人[23]基于一个消费级的 RGB-D 相机,重建、追踪源和目标演员的 3D 模型并最后融合,从而进行实时的表情迁移.另外,Thies 等人[24]提出了 Face2Face,通过利用 3D 重建技术和图像渲染技术,能够在商业视频流中进行人脸移动表情的修改.Head on[25]通过修改视角和姿态独立的纹理实现视频级的渲染方法,从而实现完整的人重建方法,包括表情眼睛、头部移动等.Kim 等人[26]利用含有时空架构的生成网络将合成的渲染图转换成真实图,并能迁移头部表情等动作.尽管现有的图形学方法可以较好地合成或重建图像,但是严重依赖于高质量的 3D 内容.Thies 等人[27]提出了延迟神经渲染的框架,与渲染网络一起优化神经纹理而生成合成的图像,此方法可以在不完美的3D内容上操作.Suwajanakorn等人[28]利用循环神经网络建立语音到嘴型动作的映射,可以匹配输入的语音合成嘴型指定纹理动作.此外,还有针对人物特写镜头中的图像合成[29,30]、基于 2D 仿射的源演员表情匹配[31]、基于网络编码空间的属性修改的表情迁移[32]等相继被研究者提出,不同场景的表情伪造技术日益成熟.

1.3 语音伪造技术

语音伪造也叫做语音版 Deepfakes,利用 AI 技术合成虚假语音.通常有文本到语音合成(text-to-speech synthesis,简称 TTS)和语音转换(voice conversion)两种形式:文本到语音合成主要完成指定文本的语音信息输出,而语音转换是指转换人的音色到目标音色.这些语音的合成不仅可以欺骗人的听觉,还可以欺骗一些自动语音认证系统.早期的语音合成主要依赖隐马尔科夫模型和高斯混合模型,而随着深度学习技术的发展,语音合成和转化技术的质量有了大幅度提高.来自谷歌的 Oord 等人提出了 WaveNet[33],这是第一个端到端的语音合成器,一种基于音频生成模型,能够产生于人相似的音频.相似的文本到语音合成系统有 Deep voice[34]和 Tacotron[35],均在原始语音材料上训练,速度比 WaveNet 更快.随后,百度对 Deep voice 进行了扩展,提出了 Deep voice2[36],通过使用低维度可训练的说话者编码来增强文本到语音的转换,使得单个模型能生成不同的声音.Ping 等人提出的 Deep voice3[37]进一步改进了之前的 Deep voice 系列,Deep voice3 是一个基于注意力机制的全卷积 TTS 系统,通过设计字符到频谱图的结构,能够实现完全并行的计算,在不降低合成性能的情况下,速度更加快.Santiago 等人[38]则利用 GAN 技术对语音的噪音进行过滤,提高了生成语音的质量.Chris 等人[39]提出了无监督音频合成模型,能够从小规模语音库中学习生成可理解的词汇.语音合成技术愈发成熟,且与视频中的换脸伪造往往同时出现,使得鉴别的难度更大.

1.4 开源工具与商业软件

随着对深度伪造生成技术的深入研究,网络上逐渐出现了众多开源软件和商业应用.已有文献[40]做了部分总结,但是不够全面.本文对其进行扩充和比较,结果见表 1,主要分为人脸伪造和语音伪造.其中,人脸伪造主要分为两类:一类是以 Faceswap 为代表的在 GitHub 网站上开源的伪造项目,此类项目均是对原始项目进行改进,或在新的深度学习框架下实现;另一类是商业化的 APP,如 Zao[1]、FakeAPP[41]、FaceApp[42]等提供换脸、修改表情或者人类属性等功能.网上开源软件需要使用者对深度学习相关知识比较熟悉,需要使用者拥有一定数量的人物图像并在 GPU 上进行训练,训练的稳定结果取决于使用者的专业水平.而商业化软件的使用门槛很低,只需使用者上传一张图像就可以实现伪造目的.其中,FakeAPP 需要用户安装在有 GPU 的电脑上使用.总的来说,开源软件使用复杂,适合专业人士,并对生成效果进行改造;而商业软件适合大部分普通非专业用户,但是生成效果也取决于开发软件的公司或组织.语音合成伪造已逐渐成熟,被大多数云服务厂商开发为接口服务向大众开发,这里选取有代表性的软件展示.这些软件的流行和传播使得深度伪造变得更加低门槛、大众化,也进一步加剧了恶意用户带来的负面影响.

Table 1 Summary of Deepfakes tools

表 1 深度伪造工具汇总

人脸伪造	功能及特点	使用者要求,GPU	素材	支持换脸
FaceSwap[7]	采用 3D 图形学	掌握基本的图形学指示,要 GPU	大量人脸	1 对 1
Deepfakes[1]	采用自动编码器	掌握深度学习专业知识,要 GPU	大量人脸	1 对 1
Faceswap-GAN[14]	在 Faceswap 项目的基础上增加 GAN 的判别器,并做了后期融合处理.	掌握深度学习专业知识,要 GPU	大量人脸照片	1 对 1
DeepfaceLab[43]	对 Faceswap 项目的模型进行扩充,对人脸模型进行扩充	掌握深度学习专业知识,要 GPU	大量人脸照片	1 对 1
DFaker[44]	使用 DSSIM loss 函数	掌握深度学习专业知识,要 GPU	大量人脸照片	1 对 1
DeepFake-tf[45]	同 Dfakeer 项目,使用 tensorflow 实现	掌握深度学习专业知识,要 GPU	大量人脸照片	1 对 1
Faceswap-Deepfake-Pytorch[46]	原理同 Faceswap 项目,使用 Pytorch 实现	掌握深度学习专业知识,要 GPU	大量人脸照片	1 对 1
Zao[1]	提供指定的影视模板换脸,只需要一张目标人脸即可换脸	无门槛,不需要 GPU	1 张人脸照片	1 对多
FakeAPP[41]	Windows 上安装的软件,原理同 Faceswap	无门槛,需要 GPU	大量人脸照片	1 对 1
Faceapp[42]	人脸编辑器,可以换脸,换表情,编辑人脸属性	无门槛,不需要 GPU	1 张人脸照片	1 对多
语音伪造	功能及特点	使用者要求,GPU	素材	语音转换
Deep-voicev-conversion[47]	只需要目标说话者的音波素材,即可转换成特定目标人物的声音	掌握深度学习专业知识,要 GPU	大量声波文件	多对 1
MelNet[48]	基于频谱图的端到端语音生成	掌握深度学习专业知识,要 GPU	大量音频文件	多对多

2 深度伪造数据集

随着深度伪造的泛滥,研究人员开始了针对这些伪造视频、图像和语音的研究,逐渐有新的数据集被开源以促进此领域的研究.数据集的质量和规模对深度伪造领域的研究尤为重要,学术界和工业界均开源了部分数据集以促进该领域的研究.本节将逐一介绍这些数据集(见表 2).

Table2 Open source dataset of the Deepfake

表 2 深度伪造开源数据集

数据集	篡改类型	描述	假:真(比例)	大小	获取源
UADFV[49]	FakeAPP	早期视频数据,量小	1:1.00	98 视频	Youtube
FaceForensics (FF)[50]	Face2Face	FaceForensics++的前身,只有一种篡改类型	1:1.00	2 008 视频	Youtube
FaceForensics++ (FF++)[51]	Deepfakes FaceSwap Face2face Neuraltexture	每一类篡改视频均被 C0, C23,C40 这 3 种参数压缩	1:1.00	5 000 视频	Youtube
Deepfake-TIMIT[52]	faceswap-GAN	GAN 版本 Deepfakes 换脸.有高清和低清两个版本	1:0.5	640 视频,高清和低清视频各 320 个	VidTIMIT[53]
Mesonet data[54]	Unknown	网络搜集的不同渠道的 Deepfake 换脸图片	Unknown	2W (图片)	Youtube
Celeb-DF[55]	Deepfakes	针对过去伪造视频的质量差、不稳定等缺点进行改进,效果更好	1:0.51	1 203 视频	Youtube
Deepfake-Detection[56] (DFD)	Deepfakes	363 个不同场景下的原视频,然后进行换脸.篡改视频均 C0, C23,C40 这 3 种参数压缩	1:0.12	363 原始视频,3 068 个篡改视频	演员拍摄
DFDC preview dataset[57]	Unknown	Deepfakes 竞赛的预赛数据	1:0.28	5 214 视频	演员拍摄

Table 2 Open source dataset of the Deepfake (Continued)

表 2 深度伪造开源数据集(续)

数据集	篡改类型	描述	假:真(比例)	大小	获取源
DFDC[58]	Unknown	Deepfakes 竞赛的正式全部数据	1:0.19	119 154 视频	演员拍摄
DeeperForensics-1.0[59]	DeepFake Variational Auto-Encoder	改进的生成方式	5:1	60 000 视频 1 760 万帧	演员拍摄
ASVspoof 2015 database[60]	synthetic and converted speech	106 speakers	14:1	16 651 段原始音频, 246 500 段合成转换视频	人说话片段
ASVspoof 2019 database[61]	synthetic and converted speech replayed speech	107 speakers	Unknown	训练集:15 928 原视频, 117 996 合成转换视频, 测试集未知	人说话片段

2.1 深度伪造视频数据集

- UADFV:此数据集素材取自 YouTube,分别有 49 个真实视频和 49 个合成视频,合成视频由 FakeAPP[41] 生成,每个视频的平均长度是大约 11s.然而,作为早期深度伪造研究的数据集之一,视频分辨率较低、生成质量差,有较明显的换脸痕迹,数量规模过于少,篡改类型比较单一.
- FaceForensics(FF):早期的大规模深度伪造数据集之一,素材来源于 Youtube8M[62],选取该数据集中标签为人脸、新闻播报员、新闻联播的视频以及 YouTube 上有类似标签的视频共 1 004 个,所有选取的视频分辨率大于 480p.除此之外,作者用人脸检测器抽取视频中的人脸序列,确保所选视频连续 300 帧中含有人脸,并手动过滤掉人脸遮挡过多的视频以确保视频质量.最后,采用 Face2Face 的换表情的方法构造 1 004 个假视频.此数据集视频规模大、源视频人脸质量高,但是篡改痕迹明显,篡改方式单一.
- FaceForensics++(FF++):目前较大规模、种类最多的深度伪造数据集之一.素材与 FaceForensics 相似,取自 YouTube 的 1 000 个视频.在筛选素材的过程中,同样用人脸检测器进行检测,确保连续帧含有人脸,并手动过滤掉人脸遮挡过多的视频以确保视频质量.在这个数据集中,作者共采用 4 种类型的人脸篡改来制作假视频.
  - Deepfakes:采用基于自动编码器的 Deepfakes 方法实现,训练一对一的生成模型,可以实现一对一的换脸.
  - Face2Face:采用 Face2Face 方法实现.
  - FaceSwap:采用 FaceSwap 方法实现,基于 3D 图像的方法.
  - Neural Textures:利用延迟神经渲染网络优化纹理的方法实现.

  其中,Deepfakes 与 FaceSwap 属于换脸伪造,Face2Face 与 Neural Textures 属于换表情伪造.4 种类型均在 1 000 个原始视频上生成对应的 1 000 个假视频,并对真假视频均做了 H.264 codec 压缩方式中的 C0、C23、C40 这 3 种压缩水平的压缩.另外,数据集中还提供了对应人脸篡改位置的 mask.然而这些篡改的质量不是很高,人眼能明显观察出篡改痕迹,修改的轮廓很明显;同时,在合成的假视频中还存在人脸闪烁现象.
- Deepfake-TIMIT:由 Faceswap-GAN 方法生成,是第一个 GAN 版本的 Deepfakes 数据集.源数据是在 VidTIMIT 中选取的 32 个人(16 对相似的人)两两相互替换组成的视频,每个人有 10 个动作视频,生成的假视频有高清(128×128)和低清(64×64)两个版本,共有 640 段假视频.生成质量比 FaceForensics++要好,但是视频分辨率不高,在脸部边界处有少量痕迹.
- Mesonet data:早期深度伪造研究数据集,数据量较小,由 YouTube 渠道搜集的网络爱好者制作的伪造视频与图像.
- Celeb-DF:针对 UADFV、FaceForensics++、Deepfake-TIMIT 等数据集的一些缺陷,如图片分辨率不高、合成的视频质量差、篡改痕迹粗糙、视频人脸闪烁感过多等缺陷,对 Deepfakes 生成方法进行改进,增

大生成图像的大小,并在训练阶段增加色调亮度、对比度等,以减小篡改区域与周边区域的不一致性.此外,使用更加精准的人脸关键点定位信息减轻人脸闪烁现象.数据集由从 YouTube 渠道下载的 408 个原始视频和生成的 795 假视频组成,视频的平均长度是 13s,帧率是 30.

- DeepfakeDetection(DFD):为了填充深度伪造数据的多样性,谷歌公司征集 28 个演员拍摄了 363 个原始视频,并将这些视频截取成一个个场景不同的片段,最后对这些片段进行相互换脸,生成 3 068 个假视频.同样,此数据集也提供了 H.264 codec 压缩方式中的 C0,C23,C40 这 3 种压缩水平的压缩版本.
- DFDC preview Dataset:为了推进深度伪造领域的研究,Facebook 举办了 The Deepfake Detection Challenge,在比赛前夕公开了预赛数据集,由 5 214 个视频组成,真假比例 1:0.28,原始视频均由 66 个演员拍摄而成,假视频有两种篡改方式,大量的替换在相似人脸之间进行,如皮肤颜色、头发、眼睛等.每个视频均是 15s 左右的小片段.
- DFDC:The Deepfake Detection Challenge 的正式数据集,共有 119 196 个视频,真假视频比例约为 1:5.原始视频均由演员拍摄,视频长度约为 10s.视频分辨率跨度很大,视频场景涵盖了多种复杂场景,如黑人黑背景、侧脸、走动、强光、多人等.
- DeeperForensics-1.0:为了应对深度伪造研究数据量少的问题,南洋理工大学和商汤科技推出了大规模深度伪造数据集.研究人员从 26 个国家收集了 100 名演员的面部数据,演员在 9 种灯光条件下转头做各种表情,并使用 FaceForensics++中的 1 000 个原始视频作为目标视频,其中,100 个演员的脸中的每一个都被交换为 10 个目标.他们故意以 35 种不同的方式扭曲每个视频,以模拟现实情况,从而最终数据集包含 50 000 个未修改的视频和 10 000 个修改的视频.

以上深度伪造数据集的示例如图 2 所示.

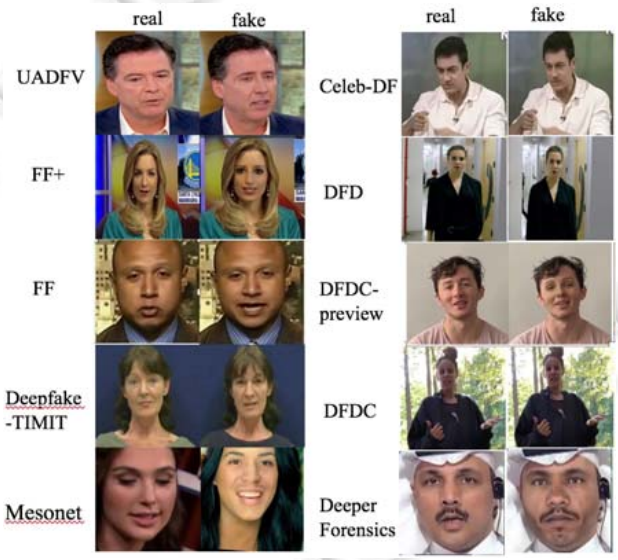

Fig.2 Exmaples of Deepfakes datasets

图 2 深度伪造数据集示例

2.2 深度伪造语音数据集

- ASVspoof 2015 database

为了应对语音合成欺骗的攻击威胁,2015 年举办了第 1 届自动说话人认证竞赛.该竞赛上开放了第一个大规模伪造语音数据集,以期发现多样的防御应对策略.数据集由 10 种不同的语音合成和语音转换欺骗算法生成,包含原始的和欺骗的语音数据.原始语音是由 106 个人(45 男与 61 女)说话记录构成,这些记录没有噪音影响.其中,训练集由 3 750 个原始话语片段和 12 625 个欺骗话语片段组成,验证集由 3 497 个原始话语片段和 49 875

个欺骗话语片段组成,测试集由 9 404 个原始话语片段和 184 000 个欺骗话语片段组成.

- ASVspoof 2019 database

2019 年,自动说话人认证竞赛包含了所有语音欺骗类型的攻击,如语音合成、语音转换、语音重放等.将攻击分类为两种场景:第 1 种场景是逻辑访问,即直接将欺骗攻击的语音注入到自动说话人认证系统,这些语音由最新的语音合成和语音转换技术生成;另一种是物理访问场景,语音数据由麦克风等设备捕捉到,再经一些专业设备重放.这些语音数据由 107 个人(46 男与 61 女)说话组成,其中,训练集、验证集、和测试集分别由 20,10,48 个人的语音数据构成.测试集中的攻击类型与训练验证集中均不相同.

3 深度伪造检测技术

随着深度伪造技术的发展,互联网上充斥着大量包含伪造人脸和语音的虚假视频,Deepfakes 类技术的滥用带来巨大的负面影响,如损坏他人名誉、伪造证据、传播谣言,影响政客形象干涉选举等.这也吸引了一批研究者对深度伪造检测技术的重视.本节将综述现有的一些代表性检测工作,其中,前 5 小节重点介绍研究较多的深度伪造视频检测,第 6 小节概述伪造语音的检测工作,并在第 7 小节对这些工作进行总结.

3.1 基于传统图像取证的方法

传统的图像取证初始主要是基于传统的信号处理方法,大多数依赖于特定篡改的证据,利用图像的频域特征和统计特征进行区分,如局部噪音分析、图像质量评估、设备指纹、光照等,解决复制-移动[63]、拼接[64]、移除这些图像篡改问题.而深度伪造视频本质也是一系列伪造合成的图片合成,因此可以将此类方法应用到深度伪造检测.Lukas 等人[65]提出了数字图像的相机设备指纹光响应不均匀性(PRNU),Chierchia 等人[66]进一步利用光响应不均匀性检测小的篡改图像.Jessica 等人[67]通过组装噪声分量模型提出了数字图像的隐写特征,随后,噪声特征被广泛运用在图像取证领域.此外,还存在诸多基于信号处理的取证方法,如利用 JPEG 压缩分析篡改痕迹[68]、向 JPEG 压缩的图像中添加噪声提升检测性能[69,70]、利用局部噪音方差分析拼接痕迹[71]、利用色彩过滤矩阵(color filter array,简称 CFA)模型[72]进行篡改定位等.然而随着人工智能技术的发展,基于卷积神经网络的深度学习技术在诸多任务上均超过了传统方法,取证方法逐渐融合了机器学习方法特别是深度学习技术.此类方法检测成功率高,不依赖特定类型的篡改痕迹,比传统的信号处理方法鲁棒性更好.Cozzolino 等人[73]设计了一个孪生网络,在来自不同相机的图像块上训练来提取图片的噪音指纹,从而实现检测.Zhou 等人[74]提出了基于双流的 Faster R-CNN 网络,其中,RGB 流主要从 RGB 图像中输入提取特征,从而发现强烈对比差异与不自然的篡改痕迹;而噪音流利用噪声特征发现篡改区域与源区域的噪声不一致性.最后,融合两条流的特征进行学习两个模态空间的信息.利用深度学习技术提取关键取证特征的工作也被不断探究[75].Liu 等人[76]提出一个新的深度融合网络通过追踪边界来定位篡改区域.Minyoung 等人[77]通过训练照片所包含的相机 EXIF 源数据指纹信息来区分图片是否被拼接.Xiaodong 等人[78]根据全局与局部块的特征不一致性学习一个半-全局网络实现拼接定位.Cozzolino 等人[79]提出使用卷积神经网络来学习基于残差的特征,此类特征可以有效提升取证检测和定位的性能.Chen 等人[80]则利用神经网络学习自然模糊和人为模糊带来的光直方图不一致性.Zhou 等人[81]将隐写噪声特征和卷积网络学习边界特征结合,提出了一个双流神经网络的方法.具体是用一个脸分类流训练一个 GoogleNet[82]检测篡改的人工痕迹,利用捕捉的局部噪音特征和拍照特征训练一个基于块的三元组(triplet)网络,用这两条流的得分,综合判断是否图像被篡改.这是因为基于同一张图像的隐藏特征是相似的,距离小;不同图像的块之间的隐藏特征距离大,用三元组训练出块的距离编码后,用一个 SVM 分类得到概率分数.

尽管基于取证的技术很成熟,但是在应对新的深度伪造视频时仍存在很多短板,因为此类伪造视频通常会被不同的后处理,如不同的压缩方式、不同的压缩率、不同的放缩合成.针对图片级的取证技术更多关注局部的异常特征,仍然应对乏力,很容易被绕过,并不能直接应用到日益升级的深度伪造视频检测上.

3.2 基于生理信号特征的方法

生成的伪造视频往往忽略人的真实生理特征,无法做到在整体上与真人一致,因此,基于生理信号的特征不

断被研究者挖掘.Yang 等人[83]认为 Deepfakes 创造的是分离的合成脸区域,这样在计算 3D 头部姿态评估的时候就会引入错误.因为 Deepfakes 是交换中心脸区域的脸,脸外围关键点的位置仍保持不变,中心和外围位置的关键点坐标不匹配,会导致 3D 头部姿态评估的不一致,故用中心区域的关键点计算一个头方向向量,整个脸计算的头方向向量,衡量这两个向量之间的差异.针对视频计算所有帧的头部姿态差异,最后训练一个支持向量机(SVM)分类器来学习这种差异.Yang 等人[84]同时发现,GAN 网络生成的假人脸在关键点位置分布上与真实人脸不尽相同,尽管生成的假人脸在脸部细节上与真人相似,但是自然性和连贯性还是与真人有很大的不同之处,通过将关键点归一化的位置坐标作为特征喂入 SVM 分类器进行学习.Li 等人[85]发现,正常人的眨眼频率和时间都有一定的范围,而 Deepfakes 伪造视频的人基本没有眨眼现象,或者频率跟正常视频有较大差别,这可能是伪造视频在生成时没有丰富多样的眨眼素材导致的.因此,作者将 CNN 和循环神经网络联合一起,设计了长期循环卷积网络来识别视频中的状态是否闭眼,从而最终判断是否是伪造的假视频.Ciftci 等人[86]从脸部抽取 3 块区域来测量光电容积脉搏波信号,并将信号转换为一致性和连贯性特征,最后使用 SVM 对特征进行二分类.类似的,Fernandes 等人[87]利用心率生物信号来区分伪造视频,先通过血流造成的脸部皮肤颜色变化、前额的平均光密度、欧拉影像变化等 3 种方法来提取心率,然后采用神经常微分方程模型训练,最后测试 Deepfakes 视频时,主要依据正常视频与异常视频的心率分布不同.

基于生理信号特征的检测方法大部分利用深度伪造技术的局限性,但是随着生成技术的改进,如眨眼数据、头部转动、眼球转动等的加入,使得此类方法失效.此外基于脉搏、心率等生物信号的方法会因为伪造视频的压缩等处理而准确度大大降低.

3.3 基于图像篡改痕迹的方法

深度伪造图像受限于早期深度网络的生成技术,在生成的人脸在细节上存在很多不足.因此,有研究者对此展开了探索.Li 等人[88]认为 Deepfakes 算法生成的图像分辨率有限,之后需要被转换到匹配替换的脸,这使得 Deepfakes 的视频中留下更多可以辨别的人工痕迹,这个可以被深度神经网络有效地捕捉.作者人工构造了大量的负样本,如将要替换的人脸进行高斯模糊、旋转等操作后放缩到源位置,这个扭曲的人脸人工痕迹就保存了,最后使用 Resnet50[89]网络区分这些伪造视频或图像.同标记视觉人工痕迹篡改视频类似,Matern 等人[49]利用真假脸的不一致性来区分,如:

(1) 全局不一致性:新的人脸的生成,图像的数据点插值是随机的,并不是很有意义,这会导致的全局眼睛的左、右颜色不一致,鼻子的左、右色彩等.

(2) 光照不一致性:篡改区域和正常区域对光照的反射不一样,如眼睛区域,Deepfakes 生成的视频大多丢失这个眼睛反射细节.

(3) 几何位置不一致:细节位置缺失,如牙齿,只有一些白色斑点,这个细节没有建模.通过对这些特定区域(牙齿、眼睛等)提取的特征向量训练多层感知机进行分类.

尽管基于篡改痕迹的方法在一些数据集上表现良好的检测能力,但是这些数据集大多是早期的生成器生成的,随着生成技术的提升,高分辨率和更多细节处理的伪造图像不断出现,同时容易受到一些对抗措施的影响,如加噪、压缩、放缩,会使得这类方法的检测能力大大减弱.

3.4 基于GAN图像特征的方法

由于当前的深度伪造视频大部分借助了 GAN[13]技术,因此研究 GAN 生成技术的特点也成为了检测伪造图像的方法之一.研究[90,91]发现:GAN 生成技术改变了图像的像素和色度空间统计特征,通过对特征共生矩阵的学习来区分生成图像的差异.Xuan 等人[92]使用图像预处理,如滤波、噪音等预处理方法破坏 GAN 图像低级别的生成缺陷,迫使模型学习高级别的固有的线索.Scott 等人[93]发现:GAN 生成器的中间值通常通过归一化来限制输出,这一定程度上也会限制饱和像素的频率.此外,生成器在多通道使用的权重与真实相机的光敏感度有很大不同,通过对这两个指标进行量化提取分类特征.也有相关研究尝试用 GAN 指纹[94,95]来区分伪造,不同的 GAN 生成的图片在中间分类层具有唯一的特征,可以作为 GAN 生成器的辨别指纹.

Wang 等人[96]提出了 FakeSpotter,利用神经元监控的方法来进行分类,原理如图 3 所示.使用神经元覆盖的方法观察真假图像经过人脸识别器中的神经元激活变化情况,用 SVM 去学习神经元激活的差异,而假脸在神经元覆盖的行为上表示相似.

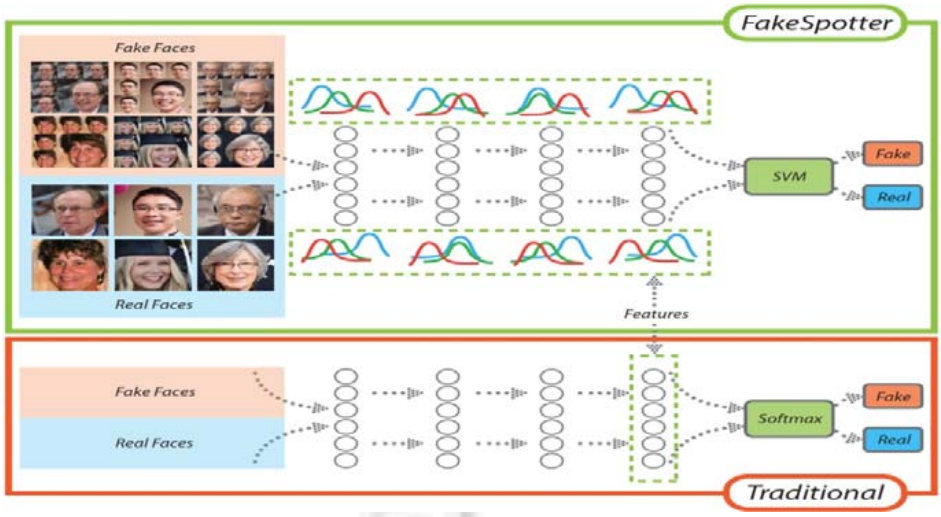

Fig.3 Using neuron coverage method to track fake face features[96]

图 3 利用神经元覆盖方法追踪假脸特征[96]

此类基于 GAN 特征的方法会依赖 GAN 的结构,使得特征分类器在已有的生成器行为上过拟合,而无法处理未知的生成器,泛化能力很差.研究不同 GAN 结构生成伪造图像的共同特点,依然是一个研究难题.

3.5 基于数据驱动的方法

新的伪造生成算法和数据量的规模都在不断增加,使得研究者开始关注用基于数据驱动的方式来学习这些 Deepfakes.基于数据驱动的学习方法主要分为两大类:一类是图片级,将视频处理成帧,设计不同的网络结构,对帧进行判别,实现帧级的识别,最终对视频的帧进行综合决策;另一类视频级,利用循环神经网络学习帧序列的时序特征对一个视频进行整体判断.

3.5.1 基于图片级学习的方法

Afchar 等人[54]设计了多个小的卷积模块来捕捉篡改图像的微观特征,Rossler 等人[51]利用 Xception[97]架构对视频的全帧和人脸分别训练.结果显示,基于人脸训练的模型效果远远好于全帧模型.同时,实验结果显示:在面对高度压缩的图片时,模型的训练难度会上升且检测率会下降.其中,利用人脸关键点信息提升性能的结论也被 Songsri-in 等人[98]实验证实.Nguyen 等人[99]设计了胶囊网络来判别造假的图片或视频,通过抽取人脸,用 VGG-19[100]提取特征编码,然后输入胶囊网络进行分类.Mo 等人[101]增加高通滤波和背景作为 CNN 输入,对检测结果有提升.Durall 等人[102]通过离散傅里叶变换提取特征学习,显示了很好的效果.Ding 等人[103]利用迁移学习,使用 Resnet18 进行调优;同时对于这些部署的关键系统,对每个预测提供一个不确定水平,如每个神经网层络输出值差异.现有的神经网络能够快速地过拟合特定的篡改痕迹,学习到的 features 有高度的区分性,但是缺乏迁移性.Cozzolino 等人[104]设计了一个新的基于自动编码器的神经网络结构,能够学习在不同的扰动域下的编码能力,只需要在一个数据集上训练,在另一个数据集上获取小规模进行调优,就能达到很好的效果.在此基础上, Nguyen 等人[105]设计了 Y 型解码器,在分类的同时融入分割和重建损失,通过分割辅助分类效果.此外,一些针对现有神经网络结构的修改也被研究:Hsu 等人[106,107]采用对比损失寻找不同生成器生成的图像的特征,后面再连接一个分类器进行分类;Dang 等人[108]设计了特定的 CGFace 网路,专门检测计算机生成的人脸;Bayar 等人[109]设计了受限制的卷积层学习特定的篡改特征;Stehouwer 等人[110]通过在主干网络增加注意力机制来聚焦

篡改区域;Rahmouni 等人[111]加入了计算统计数据的全局池化层.Li 等人[112]则设计了基于图片块的双流网路框架,如图 4 所示,一条流学习人脸块的微观特征,另一条流学习人脸和背景区域的差异性.通过多任务学习,能够较好地提升模型的泛化能力.

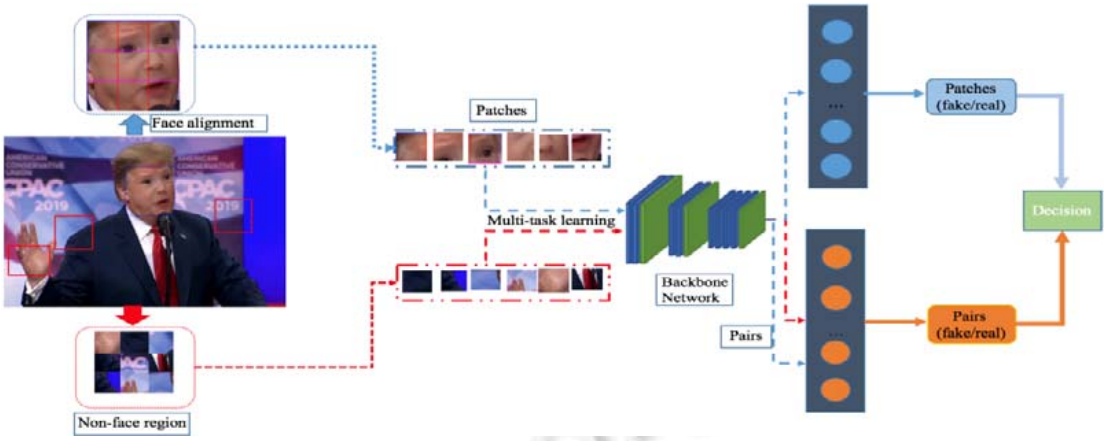

Fig.4 Multi-task forgery classification framework based on image patches[112]

图 4 基于图像块的多任务伪造分类框架[112]

基于图片级的学习方法是现有研究较多的方向之一,借助深度学习强大的学习能力和日益大幅增长的数据集,学习篡改图片的特点可行且高效.此类方法不仅可以判断单帧图像的真伪,还可以利用组合策略检测视频帧,应用范围较广,但是也存在很多局限性,学习到的模型大多数依赖相同的数据分布,在面对未知篡改类型时很乏力[113,114];同时,对高度压缩的视频帧检测能力会大幅下降.此外,如果视频中的篡改人脸非常少,这对基于图片级方法的综合决策策略提出了挑战.

3.5.2 基于视频级学习的方法

Agarwal 等人[115]发现:作为个体,他们有不一致的面部表情和移动,通过追踪面部和头部移动然后抽取特定动作集合的存在和强度,脸部肌肉的移动可以编码成动作单元,再利用皮尔森系数对特征之间的相关性进行扩充,最后在扩充后的特征集合上建立一个新的单分类 SVM 来区分各类造假视频.然而实验结果显示:虽然 AUC 达到 0.9 以上,但是召回普遍不高,实用性较差.

Amerini 等人[116]探索帧间光流的不同,采用 VGG16 学习光流的差异并进行分类,因为光流是连续帧间的运动差异计算的,自然拍摄和伪造的视频之间的运动差异很大.

Guera 等人[117]考虑用循环神经网络处理深度伪造的序列数据,因为多个相机视角,光照条件的不同,不同的视频压缩率使得生成器很难产生实际真实的在不同条件下的脸,这个会导致交换的脸在剩下的场景下看起来不一致.此外,因为生成器没办法意识到皮肤或者其他场景信息,所以新脸和剩下帧之间的融合性差,不同帧场景间的光源会引起大多数脸部闪烁现象,这个可以被时序网络较好地捕捉到.

整体框架如图 5 所示,分为两阶段分析器,一个 CNN 抽取帧内 feature,输入一个测试序列,CNN 获取一个每一帧的特征集合,然后将这些多个连续的帧特征集串联传输到 LSTM 分析,并产生一个概率估计.

相似地,Sabir 等人[118]采用双向时序网络和人脸对齐结合的方法学习伪造序列,结果显示,基于关键点的人脸对齐与 Bidrectional-recurrent-denset 对视频的篡改检测最佳.

基于视频级的学习方法可以学习到视频的时序特征,如前后帧的不一致、人脸区域的不稳定等一些篡改视频均会出现的缺陷,泛化性较好;同时,也能检测到视频中的少量篡改.但是基于时序特征的检测依然对视频的预处理很敏感,如视频压缩、背景光线的变化等,也无法判断单帧的真伪.

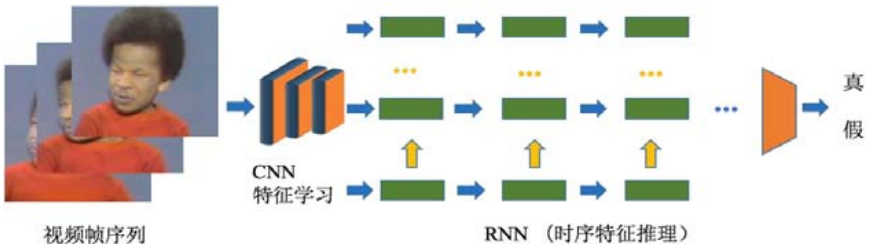

Fig.5 Frame sequences are learned by recurrent neural networks and convolutional neural networks

图 5 循环神经网络和卷积神经网络学习帧序列

3.6 深度伪造语音检测

随着合成伪造语音技术的发展,对伪造语音的检测工作也在兴起.尤其是 2019 年自动说话人语音认证竞赛(ASVspoof2019)的举办,产出了一些针对性的语音欺骗工作.初始伪造语音检测主要是传统的信号处理方法,研究者尝试对不同低水平的频谱特征进行建模,如 Todisco 等人[119]提出的常量 Q 倒谱系数(constant-Q cepstral coefficients,简称 CQCC)、Wu 等人提出的归一化的余弦相位和修改的群延迟[120,121],在一些音频处理技术上有效,但是在 ASVspoof2019 数据集上泛化性很差.有研究[122]针对 ASVspoof2019 数据集进行了数十种声学特征分析,结果显示,这些声学特征均不能在未知类型欺骗攻击有很好的泛化性.随后,基于深度学习的检测方法逐渐被研究者所关注.Zeinal 等人[123]利用 CQT 特征[124]和功率谱图特征进行学习,并分别使用网络混合、VGG 与 light CNN、VGG 与 Sincnet 应对物理访问和逻辑访问场景的攻击.目前,语音欺骗系统检测的最大问题是泛化能力,Alejandro 等人[125]提出了基于光卷积门的循环神经网络来同时抽取帧级的浅层特征和序列依赖的深层特征,检测率在 ASVspoof2019 数据集上显示有很大提升.Chen 等人[126]通过随机掩去相邻的频率频道、加入背景噪音和混合噪声提高检测系统的泛化性.

伪造语音的检测从传统信号处理方法发展到深度学习方法,在应对语音欺骗领域取得了一定的成果,但是现有方法还是依赖特定攻击类型,对未知类型攻击检测的泛化性提升还有很大的空间.

3.7 检测技术总结

前述研究工作在提出的同时,大多在开源数据集上进行了评测,本文将主流的深度伪造检测算法在公开数据集上的检测表现总结见表 3.所有数据均由论文的实验整理而得,大多数是深度伪造视频检测的工作.其中,主要评估指标有准确率(*Acc*)、ROC 曲线面积(*AUC*)、等错误率(*EER*);Raw、HQ、LQ 分别代表原生态、高清和低清;DF/F2F/FS/NT 分别是 FF+中 4 种篡改类型的缩写.

Table 3 Performance evaluation of representative methods on major test sets

表 3 代表性方法在主要测试集上的性能评估

研究工作	模型	特点	数据集	性能:*Acc*%,*AUC*%
Jessica 等人[67]	SVM	高通图像的隐写特征	FF++(DF/F2F/FS/NT) Raw HQ LQ	*Acc*% 99.03 99.1 98.27 99.88 77.12 74.68 79.51 76.94 65.58 57.55 60.58 60.69
Cozzolino 等人[79]	CNN	残差特征的学习	FF++(DF/F2F/FS/NT) Raw HQ LQ	*Acc*% 98.83 98.56 98.89 99.88 81.78 85.32 85.69 80.60 68.26 59.38 62.08 62.42
Afchar 等人[54]	CNN	微观特征的学习	FF++ Raw(DF/F2F/FS/NT) HQ(DF/F2F/FS/NT) LQ(DF/F2F/FS/NT) Mesonet Data UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ Cele-DF	*Acc*% 99.59 99.61 99.14 99.36 98.85 98.36 98.23 94.5 94.28 91.56 93.7 82.11 *Acc*=98.4% *AUC*=84.3% *AUC*=87.8% *AUC*=68.4% *AUC*=53.6%

Table 3 Performance evaluation of representative methods on major test sets (Continued)

表 3 代表性方法在主要测试集上的性能评估(续)

研究工作	模型	特点	数据集	性能:*Acc*%,*AUC*%
Rossler 等人[51]	Xception	对整帧的人脸区域学习	FF++(DF/F2F/FS/NT) Raw HQ LQ UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ Cele-DF DFDC preview	*Acc*% 99.59 99.61 99.14 99.36 98.85 98.36 98.23 94.5 94.28 91.56 93.7 82.11 *AUC*% 80.4 54.0 56.7 38.7 *Precision*=93% *recall*=8.4%
Nguyen 等人[99]	CNN+胶囊网络	胶囊网络分类	FF++/F2F-raw FF++/F2F-HQ FF++/F2F-LQ	99.33 98 83.33
Cozzolino 等人[104]	Autoencoder	分类和分割双任务	FF++(HQ) F2F FS	*Acc*% 94.47 72.57
Nguyen 等人[105]	Autoencoder	分类和分割、重建融合	UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ FF++/DF Cele-DF	*AUC*=65.8% *AUC*=55.3% *AUC*=62.2% *AUC*=76.3% *AUC*=36.5%
Agarwal 等人[115]	SVM	动作单元编码	Own (FaceSwap,HQ)	*AUC*=96.3%
Guera 等人[117]	CNN+RNN	图片的时序信息	Own	*Acc*=97.1%
Sabir 等人[118]	CNN+Bi-LSTM	图片的时序信息	FF++/LQ DF/F2F/FS	*AUC* 96.9%94.4%96.3%
Zhou 等人[81]	CNN+SVM	人脸和隐写特征结合	UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ FF+/DF Celeb-DF	*AUC*=85.1% *AUC*=73.5% *AUC*=83.5% *AUC*=70.1% *AUC*=55.7%
Li 等人[88]	CNN	学习人脸边框篡改遗留痕迹	UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ FF+/DF Celeb-DF	*AUC*=97.4 *AUC*=93.2 *AUC*=99.9 *AUC*=79.2 *AUC*=53.8
Matern 等人[49]	Logistic Regression MLP	学习篡改痕迹的细节缺失	UADFV DeepfakeTIMIT-LQ DeepfakeTIMIT-HQ FF++/F2F FF++/DF Celeb-DF	*AUC*=70.2% *AUC*=77.0% *AUC*=77.3% *AUC*=86.6% *AUC*=78.0% *AUC*=48.8%
Yang 等人[83]	SVM	头部姿态评估	UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ FF+/DF Celeb-DF	*AUC*=89.0% *AUC*=53.2% *AUC*=55.1% *AUC*=47.3% *AUC*=54.8%
Korshunov 等人[52]	PCA+RNN PCA+LDA	图像质量,声频校对	DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ)	*EER*=3.3% *EER*=8.9%
Bayar 等人[109]	–	–	FF++(DF/F2F/FS/NT) Raw HQ LQ	*Acc*% 99.28 98.79 98.98 98.78 90.18 94.93 93.14 86.04 80.95 77.30 76.83 72.38
Stehouwer 等人[110]	CNN+Attention	增加注意力机制	DFFD	*AUC*=99.4%,*EER*=3.1%
Chen 等人[126]	Deep Residual Network+Frequency Masking	大边际距离损失函数	ASVspoof2019	*LA*:*EER*=4.04% *PA*:
Alejandro 等人[125]	LightCNN+RNN	混合光卷积和门递归单元	ASVspoof2015 ASVspoof2019	*EER*=0.69% *LA*:*EER*=6.28% *PA*:=2.23%
Li 等人[127]	Butterfly Unit Multi-Task	多特征融合 多任务学习	ASVspoof2019	*LA*:*EER*=7.63% *PA*:*EER*=0.96%
Zeinali 等人[123]	Light CNN VGG,SincNet	多网络融合	ASVspoof2019	*LA*:*EER*=8.01% *PA*:*EER*=1.51%

此外,如前文所述,深度伪造视频检测归纳为 5 大类的检测算法适用于不同的场景,也在不断的推进发展中,但是都存在一定的局限性,各有优劣,总结见表 4.

Table 4 Advantages and disadvantages of various detection methods are summarized

表 4 各类检测方法优劣总结

方法	特点	缺陷
基于图像取证的方法	技术成熟,特征可解释	主要面向图像,压缩等预处理会加大提取难度
基于生理信号的方法	捕获特定的生理特征, 关注图像的局部信息	在压缩的视频里特征提取误差大 一些特征在新技术中被隐藏.准确度不高
基于图像篡改 痕迹的方法	学习局部信息,针对 粗糙的 Deepfakes 有效	通用性不强, 精准度不高
基于 GAN 图像 特征的方法	聚焦 GAN 指纹信息	数据依赖性强,依赖 生成算法,通用性不好
基于数据驱动的方法	数据量大、可学习 信息多,准确度高	依赖同分布数据集,未知 类型以及压缩对性能影响大

4 深度伪造的对抗性研究

4.1 深度伪造生成的对抗性

基于深度伪造生成的人脸能够修改人的身份属性,还可以操控人脸做不同的表情,这使得依赖人脸识别的应用存在着重大威胁.而针对人脸识别的对抗性攻击一直层出不穷.Goswami 等人[128]研究发现:对人脸图片的遮挡和加噪等操作,能够一定程度欺骗人脸检测器 VGGface[129]和 Openface[130].文献[131,132]利用查询优化的方式对人脸图片进行加噪,以此来绕过人脸识别引擎.Song 等人[133]使用注意力机制和生成对抗网络生成指定语义信息的假人脸,使得人脸识别器误判.Majumdar 等人[134]研究发现:对人脸部分区域的修改和变形,可以让人脸识别器有很高的误识率.人脸识别系统的脆弱性,使得基于深度伪造的 Deepfakes 类技术更容易攻击成功.Korshunov 等人[52,135]测试了基于 VGGnet[100]和 FaceNet[136]的人脸检测器的安全性,通过输入生成的 Deepfakes 视频,发现这两类人脸检测器分别有 85.62%和 95.00%的错误接受率,说明人脸检测器分辨不出深度伪造人脸和源人脸.

4.2 深度伪造检测的对抗性

深度伪造检测算法大部分均采用了神经网络技术,而神经网络本身存在着对抗样本攻击[137−139].对抗样本攻击是一种对模型输入进行扰动,从而使模型产生误判的技术.这使得深度伪造技术在生成的时候可以隐藏自身的一些特征从而绕过检测,因此对检测算法进行对抗性评估也十分必要.Wang 等人[140]研究发现:不同的 GAN 生成的伪造图像都留下特定的指纹特征,虽然依赖于指纹特征训练的检测器泛化能力不好,但是对训练数据进行预处理,如增加 JPEG 压缩、模糊等操作,大大提高模型的泛化性能,同时在检测时对图片进行后处理,可以增加模型的鲁棒性.但是 Neves 等人[141]设计了一个自动编码器能够将合成的伪造图像移除指纹等信息,让现有的伪造检测系统失效.Brockschmidt 等人[113]对深度伪造检测器(Xception[51]、Mesonet[54])进行了对抗性评估,作者采用 6 个伪造数据集对检测器的可靠性进行探测,结果显示:在同分布的数据集上,检测器均能达到非常高的检测率;但是在未知篡改类型数据集上,只有特征重合程度高的数据集之间迁移性较好,否则检测效果非常差.Marra 等人[142]则模拟了篡改图片在社交网络的场景中的检测,结果显示,现有的检测器在现实网络对抗环境下(未知压缩和未知类型等)表现很差.Zhang 等人[143]寻找 GAN 的共有痕迹,提高检测器的鲁棒性.现有的检测器对数据依赖强,泛化性不够,Du 等人[144]利用局部性感知的自动编码器实现造检测,使得模型聚焦篡改区域,通用性更强.Huang 等人[145]则借鉴了对抗样本的思想,对这些基于神经网络的检测器进行对抗性攻击,设计了单个对抗攻击和通用对抗攻击两种方式,使得检测器的篡改分类和定位失效.尽管现在已经存在众多的检测器,在一些数据集上表现很好,但是攻击者依然可以完善生成方法,隐藏一些标志性特征从而绕过检测器,这是一个长期的攻防博弈过程.

5 总结与展望

5.1 技术风险

深度伪造技术的发展给社会带来了巨大的负面影响,从社会国家领导人到普通的互联网公民,都有被此类技术侵害的可能性[146].对深度伪造技术带来的技术风险如下.

(1) 舆论负面影响:如色情电影的制作、政治家的谣言传播,会严重损害个人名誉.

(2) 对人脸认证的影响:目前大多依赖活体检测来识别视频攻击,如果在没有活体检测的应用场景以及活体功能失效的场景,如端劫持,对换脸的人与本人的识别面临挑战.

(3) 对视频人脸识别系统的影响:通过追踪视频人脸并识别的技术面对挑战,换脸的视频与真人的视频分辨不出来.

(4) 影响司法体系:由于缺乏完全可靠的鉴别深度伪造数据的能力,法院需要重新审视图片或者视频证据的效力.

(5) 影响经济活动:名人的假视频能让股市瞬间暴跌.

而这些风险后面还隐藏着国家治安稳定、伦理道德、经济发展、信任危机等更深层次的社会问题,亟需研究更有效的应对措施.

5.2 研究难点

从深度伪造技术诞生至今,有不少的研究工作展开对伪造图像或视频进行检测,但是依然没有完美的解决方案[40],在检测领域依然存在着诸多研究难点问题.

(1) 压缩方式的不同、压缩率的不同:视频不同于图片,在上传到网站时会做不同的压缩方式处理;同样,视频在线下制作时也可以做不同的后处理裁剪压缩,这会使得很多篡改特征模糊甚至消失.制作者甚至可以对视频中的部分帧进行压缩处理,人为地增加检测难度.此外,不同的压缩方式和压缩率下的数据分布也有很大不同,这也意味着基于学习的方法会很容易在已有的训练集上过拟合.现有的检测方法还无法有效地检测未知压缩的视频,大多是在训练集中扩充压缩的数据,增加模型的决策边界以此来应对压缩[51].此类方法本质还是基于同分布压缩的假设.

(2) 视频分辨率的不同:互联网上的视频质量和大小各异,不同的视频有着不同的分辨率,人脸大小跨度从几百像素到百万像素级别.如果统一放缩到指定大小处理,会丧失部分特征,在一定程度上影响着检测器的特征提取,这就需要检测算法从根本上考虑不同尺度特征的融合.

(3) 篡改算法未知:生成算法层出不穷,不同的生成算法篡改的侧重点不同,所具有的特征也不尽相同.基于学习的方法虽然能快速捕捉到训练集中的人脸篡改特征,但是大多是拟合已有的生成器特征,对未知的篡改类型不鲁棒.现有的应对方法大多是将新的生成算法数据集加入到训练集[51,112],以此来提高跨生成算法之间的检测率.如何设计鲁棒性强、泛化性能高的检测算法,依然是难点.

(4) 一些复杂的对抗场景:真实网络世界中的视频远远比公开数据集的复杂度要高的多,而且存在较强的对抗性.一些在实验数据上表现很好的模型,在面对真实网络伪造数据集时可能束手无策.如多人脸的视频如何无误地检测、针对只有部分帧部分区域篡改的视频如何区分、视频里过强或过暗的光线对人脸检测的影响如何评估等,人脸生成伪造者在制作的同时也会考虑加入这些对抗性场景,以此来降低检测效果,这些复杂场景对伪造检测算法带来巨大的挑战.

5.3 未来研究方向

虽然针对伪造图像或语音的检测已经取得了一部分研究成果,但目前该领域的研究依然存在诸多关键问题尚待解决.同时,一些新的生成技术的发展成熟,会让此类深度伪造的鉴别工作越来越困难.针对以上的难点和问题,我们可以考虑从多角度多层次来探索深度伪造检测未来可行的方向.

(1) 研究泛化性好的检测算法:已有的检测方法容易依赖特定的数据集和生成算法,泛化能力很弱.这往

往是由于训练数据的单一同分布所致.仅仅粗暴地对数据直接学习并不能满足多样的伪造类型,需要探索尽可能多的深度伪造类型,寻找其中的共性特征,如生成器的指纹[94,95]、不同伪造数据中人脸与嘴唇一致性差异等.通过对共性特征的学习,使得检测模型能够适用于更多的深度伪造类型.

(2) 研究鲁棒性强的检测算法:论文中展现的检测算法大多在单一的场景下测试,而现实世界中常常面对压缩、噪音等复杂情况,使得检测算法不鲁棒.可以在训练阶段和测试阶段对数据进行压缩、放缩等预处理,探索不同预处理对检测算法鲁棒性的影响.同时,还可以将对抗样本技术应用到检测模型的鲁棒性提升上,探索检测模型在对抗样本攻击下的缺陷,进而可以利用对抗环境下生产的对抗样本对模型进行对抗训练以增加模型的鲁棒性.此外,已有的数据集大多数都为单人脸的真伪鉴别,检测模型缺乏应对视频中多人脸的复杂场景.如何在保证准确率的同时对视频中多人脸的篡改进行判断,是一个具有挑战性的课题.

(3) 研究主动防御算法:现有的检测算法总是依赖已发现的深度伪造类型,对未知类型的伪造数据检测很被动,这使得检测算法总是落后于生成技术.可以从两个角度进行主动防御:第 1 种思路是利用对抗样本技术对上传到互联网上的媒体数据注入对抗噪音,如注入对抗人脸检测的噪音,使得人脸检测技术在预处理人脸数据时检测错位或失败,从而使得依赖人脸检测技术的深度伪造换脸技术不再精准,导致换脸异常或失败;第 2 种思路是控制视频传播的源头,对互联网上的视频进行溯源,研究视频网站上的视频追踪技术,如 Hasan 等人[147]尝试用区块链技术对互联网上的视频进行追踪.

(4) 研究深度伪造图像和伪造语音的融合检测技术:现有针对深度伪造的检测技术基本只关注了一个单一的伪造领域,而伪造的多媒体数据通过图像和语音结合能达到更逼真的效果.因此,对伪造数据进行图像语音多模态的检测是一个有意义的方向.如,Facebook 举办的深度伪造检测竞赛[58]已经增加了同时篡改音频和图像的数据类型.这种伪造类型将会越来越普遍,带来的负面影响也会更大.针对此类伪造的检测研究也给单模态(图像或语音)伪造的检测提供了思路.

(5) 建立研究性社区:现有的研究资源没有得到很好地共享,缺乏如全球研究者认可维护的研究性网站.对现有的研究数据集共享,需要建立统一的社区,集中现有零散的数据资源,让研究者们能更好地利用已有的资源和成果.现有网站[56]在开源部分数据集,但是力度不够,需要集成大规模数据集并公开现有最好的评估指标,定期举办学术研讨会和比赛,增加研究者们对深度伪造检测领域的关注度.

(6) 进行司法立法:深度伪造的检测仅仅依靠技术手段可能不能完美地解决问题,因为生成与检测是一个永恒博弈的过程,仅依靠一门检测技术来杜绝深度伪造现象不太现实,需要社会建立完整的法律体系,对恶意制作或传播的互联网用户进行一定的惩戒.如美国加州[148]已经立法,禁止制作政治人物的伪造视频,同时也明确规定了制作色情伪造人物视频属于违法行为.中国的互联网信息办公室也发行了《网络信息内容生态治理规定》[149],其中明确规定,网络信息内容服务使用者、内容生产者和内容服务平台不得利用深度学习、虚拟现实等新技术新应用从事法律、行政法规禁止的活动.尽管已出台了相关法律抑制深度伪造的滥用,但是此类法律还不健全,如何区分伪造视频是属于娱乐和恶性传播等在法律边界游走的现象,还需要相关部门建立更加完整细致的法律体系.

(7) 培训新闻工作者:法律和技术检测能一定程度增加恶意伪造传播的代价,但是给社会带来的负面影响无法挽回,这需要在视频传播的源头进行控制,如一些社交媒体,特别是主流媒体承担着大量的视频图像的传播任务,需要对这些新闻工作者进行专业培训,培养鉴别一些假视频的能力,从源头减少伪造视频的传播,降低负面影响.同时,对本身制作视频的新闻工作者,要明确在视频上打上是否伪造的标签,以减少新闻媒体的误导能力.

6 结束语

随着深度学习技术的发展,深度伪造技术会不断完善,生成更加逼真难以鉴别的视频和语音数据.这对深度伪造的检测提出了巨大的挑战.尽管已存在有一些针对深度伪造检测的工作,但是都依赖特定的数据集或者场

景,依然存在许多关键的科学问题尚待解决.为了理清现有研究的进展,明确未来研究方向,本文从生成技术、研究数据集、主流检测方法进行总结,回顾了大量极具影响力的研究成果,并对相关研究进行了科学的分类、总结和分析.同时,本文指出了深度伪造检测领域当前面临的挑战,探讨了未来可行的研究方向,旨在为推动深度伪造检测领域的进一步发展和应用提供指导和参考.

References:

[1] Deepfakes. 2019. https://github.com/deepfakes/faceswap

[2] Zao app. 2019. https://zao-app.com/

[3] Deepfake detection challenge. 2020. https://www.kaggle.com/c/deepfake-detection-challenge

[4] Girish N, Nandini C. A review on digital video forgery detection techniques in cyber forensics. Science, Technology and Development, 2019,3(6):235−239.

[5] Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S. Deep learning for Deepfakes creation and detection. arXiv preprint arXiv:1909.11573, 2019.

[6] Zollhöfer M, Thies J, Garrido P, Bradley D, Beeler T, Perez P, Stamminger M, Niessner M, Theobalt C. State of the art on monocular 3D face reconstruction, tracking, and applications. Computer Graphics Forum, 2018,37(2):523−550.

[7] FaceSwap. 2019. https://github.com/MarekKowalski/FaceSwap/

[8] Dale K, Sunkavalli K, Johnson MK, Vlasic D, Matusik W, Pfister H. Video face replacement. In: Proc. of the SIGGRAPH Asia Conf. 2011. 1−10.

[9] Garrido P, Valgaerts L, Rehmsen O, Thormae T, Perez P, Theobalt C. Automatic face reenactment. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 4217−4224.

[10] Garrido P, Valgaerts L, Sarmadi H, Steiner I, Varanasi K, Perez P, Theobalt C. VDub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. Computer Graphics Forum, 2015,34(2):193−204.

[11] Nirkin Y, Masi I, Tuan AT, Hassner T, Medioni G. On face segmentation, face swapping, and face perception. In: Proc. of the 13th IEEE Int'l Conf. on Automatic Face and Gesture Recognition (FG 2018). IEEE, 2018. 98−105.

[12] Lu Z, Li Z, Cao J, He R, Sun Z. Recent progress of face image synthesis. In: Proc. of the 4th IAPR Asian Conf. on Pattern Recognition (ACPR). IEEE, 2017. 7−12.

[13] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courvile A, Bengio Y. Generative adversarial nets. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 2672−2680.

[14] Faceswap-GAN. 2019. https://github.com/shaoanlu/faceswap-GAN

[15] Korshunova I, Shi W, Dambre J, Theis L. Fast face-swap using convolutional neural networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 3677−3685.

[16] Nirkin Y, Keller Y, Hassner T. FSGAN: Subject agnostic face swapping and reenactment. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 7184−7193.

[17] Choi Y, Choi M, Kim M, Ha J, Kin S, Choo J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 8789–8797.

[18] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Netaxas D. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,41(8):1947−1962.

[19] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: Proc. of the 6th Int'l Conf. on Learning Representations (ICLR). 2018.

[20] Antipov G, Baccouche M, Dugelay JL. Face aging with conditional generative adversarial networks. In: Proc. of the IEEE Int'l Conf. on Image Processing (ICIP). IEEE, 2017. 2089−2093.

[21] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[22] Huang R, Zhang S, Li T, He R. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2439−2448.

[23] Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M, Theobalt C. Real-time expression transfer for facial reenactment. ACM Trans. on Graphics (TOG), 2015,34(6):Article No.183.

[24] Thies J, Zollhofer M, Stamminger M, Theobalt C, Niebner M. Face2face: Real-time face capture and reenactment of RGB videos. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2387−2395.

[25] Thies J, Zollhöfer M, Theobalt C, Stamminger M, Niessner M. Headon: Real-time reenactment of human portrait videos. ACM Trans. on Graphics (TOG), 2018,37(4):1−13.

[26] Kim H, Garrido P, Tewari A, Xu W, Thies J, Niessner M, Perez P, Richardt C, Zollhofer M, Theobalt C. Deep video portraits. ACM Trans. on Graphics (TOG), 2018,37(4):1−14.

[27] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures. ACM Trans. on Graphics (TOG), 2019,38(4):1−12.

[28] Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing Obama: Learning lip sync from audio. ACM Trans. on Graphics (TOG), 2017,36(4):1−13.

[29] Zakharov E, Shysheya A, Burkov E, Lempitsky V. Few-shot adversarial learning of realistic neural talking head models. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 9459−9468.

[30] Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman D, Genova K, Jin Z, Theobalt C, Agrawala M. Text-based editing of talking-head video. ACM Trans. on Graphics (TOG), 2019,38(4):1−14.

[31] Averbuch-Elor H, Cohen-Or D, Kopf J, Cohen M. Bringing portraits to life. ACM Trans. on Graphics (TOG), 2017,36(6):Article No.196.

[32] Lample G, Zeghidour N, Usunier N, Bordes A, Denoyer L, Ranzato M. Fader networks: Manipulating images by sliding attributes. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 5967−5976.

[33] Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior AW, Kavukcuoglu K. Wavenet: A generative model for raw audio. In: Proc. of the 9th Speech Synthesis Workshop. 2016.

[34] Arik S, Chrzanowski M, Coates A, Diamos G, Kang Y, Li X, Miller J, Ng A, Raiman J, Sengupta S, Shoeybi M. Deep voice: Real-time neural text-to-speech. In: Proc. of the 34th Int'l Conf. on Machine Learning. 2017. 195−204.

[35] Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss R, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le Q, Agiomyrgiannakis Y, Clark B, Saurous R. Tacotron: Towards end-to-end speech synthesis. In: Proc. of the Interspeech 2017, 18th Annual Conf. of the Int'l Speech Communication Association. 2017. 4006−4010.

[36] Arik S, Diamos G, Gibiansky A, Miller J, Peng K, Ping W, Raiman J, Zhou Y. Deep voice 2: Multi-speaker neural text-to-speech. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 2962−2970.

[37] Ping W, Peng K, Gibiansky A, Arik S, Kannan A, Narang S. Deep voice 3: 2000-speaker neural text-to-speech. In: Proc. of the ICLR. 2018. 214−217.

[38] Pascual S, Bonafonte A, Serra J. SEGAN: Speech enhancement generative adversarial network. In: Proc. of the Interspeech 2017, 18th Annual Conf. of the Int'l Speech Communication Association. 2017. 3642−3646.

[39] Donahue C, McAuley J, Puckette M. Adversarial audio synthesis. In: Proc. of the 7th Int'l Conf. on Learning Representations (ICLR). 2019.

[40] Li XR, Yu K. A Deepfakes detection technique based on two-stream network. Journal of Cyber Security, 2020,5(2):84−91 (in Chinese with English abstract).

[41] FakeApp. 2019. https://www.deepfakescn.com

[42] Faceapp. 2019. https://www.faceapp.com/

[43] DeepFaceLab. 2019. https://github.com/iperov/DeepFaceLab

[44] Dfaker. 2019. https://github.com/dfaker/df

[45] DeepFake-tf. 2019. https://github.com/StromWine/DeepFake-tf

[46] Faceswap-Deepfake-Pytorch. 2019. https://github.com/Oldpan/Faceswap-Deepfake-Pytorch

[47] Deep-voice-conversion. 2020. https://github.com/andabi/deep-voice-conversion

[48] MelNet. 2020. https://sjvasquez.github.io/blog/melnet/

[49] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose Deepfakes and face manipulations. In: Proc. of the IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019. 83−92.

[50] Rössler A, Cozzolino D, Verdoliva L, Christian R, Justus T, Matthias N. Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179, 2018.

[51] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. Faceforensics++: Learning to detect manipulated facial images. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 1–11.

[52] Korshunov P, Marcel S. Deepfakes: A new threat to face recognition? Assessment and detection. arXiv preprint arXiv:1812.08685, 2018.

[53] VidTIMIT. 2019. http://conradsanderson.id.au/vidtimit/

[54] Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: A compact facial video forgery detection network. In: Proc. of the IEEE Int'l Workshop on Information Forensics and Security (WIFS). IEEE, 2018. 1–7.

[55] Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: A new dataset for Deepfake forensics. arXiv preprint arXiv:1909.12962, 2019.

[56] DeepfakeDetection. 2019. https://github.com/ondyari/FaceForensics

[57] Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer C. The Deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854, 2019.

[58] DFDC. 2020. https://www.kaggle.com/c/deepfake-detection-challenge/data

[59] Jiang L, Li R, Wu W, Qian C, Loy C. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2020. 2886–2895.

[60] ASVspoof 2015 database. 2020. https://datashare.is.ed.ac.uk/handle/10283/853

[61] ASVspoof 2019 database. 2020. https://datashare.is.ed.ac.uk/handle/10283/3336

[62] Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S. Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675, 2016.

[63] Amerini I, Ballan L, Caldelli R, Bimbo AD, Serra G. A sift-based forensic method for copy-move attack detection and transformation recovery. IEEE Trans. on Information Forensics and Security, 2011,6(3):1099–1110.

[64] De Carvalho TJ, Riess C, Angelopoulou E, Pedrini H, Rocha A. Exposing digital image forgeries by illumination color classification. IEEE Trans. on Information Forensics and Security, 2013,8(7):1182–1194.

[65] Lukáš J, Fridrich J, Goljan M. Detecting digital image forgeries using sensor pattern noise. In: Proc. of the Security, Steganography, and Watermarking of Multimedia Contents VIII, Vol.6072. Int'l Society for Optics and Photonics, 2006.

[66] Chierchia G, Parrilli S, Poggi G, Verdoliva L, Sansone C. PRNU-based detection of small-size image forgeries. In: Proc. of the 17th Int'l Conf. on Digital Signal Processing (DSP). IEEE, 2011. 1–6.

[67] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2012, 7(3):868–882.

[68] Wang W, Dong J, Tan T. Exploring DCT coefficient quantization effects for local tampering detection. IEEE Trans. on Information Forensics and Security, 2014,9(10):1653–1666.

[69] Nataraj L, Sarkar A, Manjunath BS. Adding gaussian noise to "denoise" JPEG for detecting image resizing. In: Proc. of the 16th IEEE Int'l Conf. on Image Processing (ICIP). IEEE, 2009. 1493–1496.

[70] Bianchi T, De Rosa A, Piva A. Improved DCT coefficient analysis for forgery localization in JPEG images. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011. 2444–2447.

[71] Pan X, Zhang X, Lyu S. Exposing image splicing with inconsistent local noise variances. In: Proc. of the IEEE Int'l Conf. on Computational Photography (ICCP). IEEE, 2012. 1–10.

[72] Ferrara P, Bianchi T, De Rosa A, Piva A. Image forgery localization via fine-grained analysis of CFA artifacts. IEEE Trans. on Information Forensics and Security, 2012,7(5):1566–1577.

[73] Cozzolino D, Verdoliva L. Noiseprint: A CNN-based camera model fingerprint. IEEE Trans. on Information Forensics and Security, 2019,15:144–159.

[74] Zhou P, Han X, Morariu VI, Davis LS. Learning rich features for image manipulation detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1053–1061.

[75] Rao Y, Ni J. A deep learning approach to detection of splicing and copy-move forgeries in images. In: Proc. of the IEEE Int'l Workshop on Information Forensics and Security (WIFS). IEEE, 2016. 1–6.

[76] Liu B, Pun CM. Deep fusion network for splicing forgery localization. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 237−251.

[77] Huh M, Liu A, Owens A, Efros A. Fighting fake news: Image splice detection via learned self-consistency. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 101−117.

[78] Cun X, Pun CM. Image splicing localization via semi-global network and fully connected conditional random fields. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 252−266.

[79] Cozzolino D, Poggi G, Verdoliva L. Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. In: Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security. 2017. 159−164.

[80] Chen C, McCloskey S, Yu J. Focus manipulation detection via photometric histogram analysis. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1674−1682.

[81] Zhou P, Han X, Morariu VI, Davis LS. Two-stream neural networks for tampered face detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017. 1831−1839.

[82] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1−9.

[83] Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019. 8261−8265.

[84] Yang X, Li Y, Qi H, Lyu S. Exposing GAN-synthesized faces using landmark locations. In: Proc. of the ACM Workshop on Information Hiding and Multimedia Security. 2019. 113−118.

[85] Li Y, Chang MC, Lyu S. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In: Proc. of the IEEE Int'l Workshop on Information Forensics and Security (WIFS). IEEE, 2018. 1−7.

[86] Ciftci UA, Demir I. FakeCatcher: Detection of synthetic portrait videos using biological signals. arXiv preprint arXiv:1901.02212, 2019.

[87] Fernandes S, Raj S, Ortiz E, Vintila I, Salter M, Urosevic G, Jha S. Predicting heart rate variations of Deepfake videos using neural ODE. In: Proc. of the IEEE Int'l Conf. on Computer Vision Workshops. 2019. 1721−1729.

[88] Li Y, Lyu S. Exposing Deepfake videos by detecting face warping artifacts. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019. 46−52.

[89] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770−778.

[90] Nataraj L, Mohammed TM, Chandrasekaran S, Flenner A, Bappy JH, Roy-Chowdhury AK, Manjunath BS. Detecting GAN generated fake images using co-occurrence matrices. Electronic Imaging, 2019,2019(5):532-1−532-7.

[91] Li H, Li B, Tan S, Huang J. Identification of deep network generated images using disparities in color components. arXiv preprint arXiv:1808.07276, 2018.

[92] Xuan X, Peng B, Wang W, Dong J. On the generalization of GAN image forensics. In: Proc. of the Chinese Conf. on Biometric Recognition. Cham: Springer-Verlag, 2019. 134−141.

[93] McCloskey S, Albright M. Detecting GAN-generated imagery using color cues. arXiv preprint arXiv:1812.08247, 2018.

[94] Marra F, Gragnaniello D, Verdoliva L, Poggi G. Do GANs leave artificial fingerprints? In: Proc. of the IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2019. 506−511.

[95] Yu N, Davis LS, Fritz M. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 7556−7566.

[96] Wang R, Ma L, Juefei-Xu F, Xie X, Wang J, Liu Y. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. In: Proc. of the 29th Int'l Joint Conf. on Artifical Intelligence (IJCAI). 2020. 3444−3451.

[97] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1251−1258.

[98] Songsri-in K, Zafeiriou S. Complement face forensic detection and localization with faciallandmarks. arXiv preprint arXiv:1910.05455, 2019.

[99] Nguyen HH, Yamagishi J, Echizen I. Capsule-forensics: Using capsule networks to detect forged images and videos. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP 2019). IEEE, 2019. 2307−2311.

[100] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations (ICLR). 2015.

[101] Mo H, Chen B, Luo W. Fake faces identification via convolutional neural network. In: Proc. of the 6th ACM Workshop on Information Hiding and Multimedia Security. 2018. 43−47.

[102] Durall R, Keuper M, Pfreundt FJ, Keuper J. Unmasking DeepFakes with simple features. arXiv preprint arXiv:1911.00686, 2019.

[103] Ding X, Raziei Z, Larson EC, Olinick EV, Krueger PS, Hahsler M. Swapped face detection using deep learning and subjective assessment. EURASIP Journal on Information Security, 2020(2020):Article No.6.

[104] Cozzolino D, Thies J, Rössler A, Riess C, Niebner M, Verdoliva L. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510, 2018.

[105] Nguyen HH, Fang F, Yamagishi J, Echizen I. Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876, 2019.

[106] Hsu CC, Lee CY, Zhuang YX. Learning to detect fake face images in the wild. In: Proc. of the Int'l Symp. on Computer, Consumer and Control (IS3C). IEEE, 2018. 388−391.

[107] Hsu CC, Zhuang YX, Lee CY. Deep fake image detection based on pairwise learning. Applied Sciences, 2020,10(1):Article No.370.

[108] Dang LM, Hassan SI, Im S, Lee J, Lee S, Moon H. Deep learning based computer generated face identification using convolutional neural network. Applied Sciences, 2018,8(12):Article No.2610.

[109] Bayar B, Stamm MC. A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proc. of the 4th ACM Workshop on Information Hiding and Multimedia Security. 2016. 5−10.

[110] Dang H, Liu F, Stehouwer J, Liu X, Jain A. On the detection of digital face manipulation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2020. 5780−5789.

[111] Rahmouni N, Nozick V, Yamagishi J, Echizen I. Distinguishing computer graphics from natural images using convolution neural networks. In: Proc. of the IEEE Workshop on Information Forensics and Security (WIFS). IEEE, 2017. 1−6.

[112] Li X, Yu K, Ji S, Wang Y, Wu C, Xue H. Fighting against Deepfake: Patch&Pair convolutional neural networks (PPCNN). In: Proc. of the Companion Web Conf. 2020. 2020. 88−89.

[113] Brockschmidt J, Shang J, Wu J. On the generality of facial forgery detection. In: Proc. of the IEEE 16th Int'l Conf. on Mobile Ad Hoc and Sensor Systems Workshops (MASSW). IEEE, 2019. 43−47.

[114] Sohrawardi SJ, Chintha A, Thai B, Seng S, Hickerson A, Ptucha R, Wright M. Poster: Towards robust open-world detection of Deepfakes. In: Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. 2019. 2613−2615.

[115] Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H. Protecting world leaders against deep fakes. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2019. 38−45.

[116] Amerini I, Galteri L, Caldelli R, Bimbo AD. Deepfake video detection through optical flow based CNN. In: Proc. of the IEEE Int'l Conf. on Computer Vision Workshops. 2019. 1205−1207.

[117] Güera D, Delp EJ. Deepfake video detection using recurrent neural networks. In: Proc. of the 15th IEEE Int'l Conf. on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018. 1−6.

[118] Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. arXiv preprint arXiv:1905.00582, 2019.

[119] Todisco M, Delgado H, Evans NWD. A new feature for automatic speaker verification anti-spoofing: Constant *Q* cepstral coefficients. In: Proc. of the Odyssey. 2016. 283−290.

[120] Wu Z, Kinnunen T, Chng ES, Li H, Ambikairajah E. A study on spoofing attack in state-of-the-art speaker verification: The telephone speech case. In: Proc. of the Asia Pacific Signal and Information Processing Association Annual Summit and Conf. IEEE, 2012. 1−5.

[121] Wu Z, Chng ES, Li H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Proc. of the 13th Annual Conf. of the Int'l Speech Communication Association. 2012. 1700−1703.

[122] Das RK, Yang J, Li H. Long range acoustic and deep features perspective on ASVspoof 2019. In: Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019. 1018−1025.

[123] Zeinali H, Stafylakis T, Athanasopoulou G, Rohdin J, Gkinis I, Burget L, Cernocky JH. Detecting spoofing attacks using VGG and SincNet: BUT-Omilia submission to ASVspoof 2019 challenge. In: Proc. of the 20th Annual Conf. of the Int'l Speech Communication Association. 2019. 1073−1077.

[124] Schörkhuber C, Klapuri A. Constant-*Q* transform toolbox for music processing. In: Proc. of the 7th Sound and Music Computing Conf. Barcelona, 2010. 3−64.

[125] Gomez-Alanis A, Peinado AM, Gonzalez JA, Gomez AM. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In: Proc. of the Interspeech 2019. 2019. 1068−1072.

[126] Chen T, Kumar A, Nagarsheth P, Sivaraman G, Khoury E. Generalization of audio Deepfake detection. In: Proc. of the Odyssey 2020 Speaker and Language Recognition Workshop. 2020. 132−137.

[127] Li R, Zhao M, Li Z, Li L, Hong Q. Anti-spoofing speaker verification system with multi-feature integration and multi-task learning. In: Proc. of the Interspeech. 2019. 1048−1052.

[128] Goswami G, Ratha N, Agarwal A, Singh R, Vatsa M. Unravelling robustness of deep learning based face recognition against adversarial attacks. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018. 6829−6836.

[129] Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. In: Proc. of the British Machine Vision Conf. (BMVC). BMVA Press, 2015. 41.1−41.12.

[130] Baltrušaitis T, Robinson P, Morency LP. Openface: An open source facial behavior analysis toolkit. In: Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV). IEEE, 2016. 1−10.

[131] Li X, Ji S, Han M, Ji J, Ren Z, Liu Y, Wu C. Adversarial examples versus cloud-based detectors: A black-box empirical study. arXiv preprint arXiv:1901.01223, 2019.

[132] Dong Y, Su H, Wu B, Li Z, Liu W, Zhang T, Zhu J. Efficient decision-based black-box adversarial attacks on face recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 7714−7722.

[133] Song Q, Wu Y, Yang L. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. arXiv preprint arXiv:1811.12026, 2018.

[134] Majumdar P, Agarwal A, Singh R, Vatsa M. Evading face recognition via partial tampering of faces. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2019. 11−20.

[135] Korshunov P, Marcel S. Vulnerability of face recognition to deep morphing. arXiv preprint arXiv:1910.01933, 2019.

[136] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 815−823.

[137] Szegedy C, Zaremba W, Sutskever I, Bruna J. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Leaning Representations (ICLR). 2014.

[138] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Leaning Representations (ICLR). 2015.

[139] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. In: Proc. of the 5th Int'l Conf. on Leaning Representations (ICLR) Workshop. 2017.

[140] Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-generated images are surprisingly easy to spot for now. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2020. 8692−8701.

[141] Neves JC, Tolosana R, Vera-Rodriguez R, Vera-Rodriguez R, Lopes V, Proena H, Fierrez J. Ganprintr: Improved fakes and evaluation of the state-of-the-art in face manipulation detection. IEEE Journal of Selected Topics in Signal Processing, 2020,14(5): 1038−1048.

[142] Marra F, Gragnaniello D, Cozzolino D, Verdoliva L. Detection of GAN-generated fake images over social networks. In: Proc. of the IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2018. 384−389.

[143] Zhang X, Karaman S, Chang SF. Detecting and simulating artifacts in GAN fake images. In: Proc. of the IEEE Int'l Workshop on Information Forensics and Security (WIFS). 2019. 1−6.

[144] Du M, Pentyala S, Li Y, Hu X. Towards generalizable forgery detection with locality-aware autoencoder. arXiv preprint arXiv: 1909.05999, 2019.

[145] Huang R, Fang F, Nguyen HH, Yamagishi J, Echizen I. Security of facial forensics models against adversarial attacks. arXiv preprint arXiv:1911.00660, 2019.

[146] Hall HK. Deepfake videos: When seeing isn't believing. Catholic University Journal of Law and Technology, 2018,27(1):Article No.51.

[147] Hasan HR, Salah K. Combating deepfake videos using blockchain and smart contracts. IEEE Access, 2019,7:41596–41606.

[148] The law of California to Deepfake. 2019. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

[149] Regulations of China Internet Information Office on the control of online content. 2020 (in Chinese). http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm

附中文参考文献:

[40] 李旭嵘,于鲲.一种基于双流网络的 Deepfakes 检测技术.信息安全学报,2020,5(2):84–91.

[149] 中国互联网信息办关于网络内容管控的规定. 2020. http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm

李旭嵘(1992－),男,博士,主要研究领域为人工智能安全,对抗学习.

邓水光(1979－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为服务计算,边缘计算.

纪守领(1986－),男,博士,研究员,博士生导师,CCF 专业会员,主要研究领域为人工智能与安全,数据驱动安全,IoT 安全,软件与系统安全,大数据分析.

程鹏(1982－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为控制系统安全,物联网,数据安全.

吴春明(1967－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为网络体系结构,可重构网络与虚拟化,软件定义网络,网络空间内生安全.

杨珉(1979－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为智能系统安全.

刘振广(1988－),男,博士,研究员,CCF 专业会员,主要研究领域为视频图像处理,多媒体,区块链.

孔祥维(1963－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能安全和可解释,非结构数据分析,跨媒体检索和哈希,数据驱动的决策.