虚假数字人脸内容生成与检测技术

蔺琛皓[1),2)] 沈 超[1),2)] 邓静怡[1),2)] 胡鹏斌[1),2)]
王 骞[3)] 马仕清[4)] 李 琦[5)] 管晓宏[1)2)]

[1)](西安交通大学电子与信息学部网络空间安全学院 西安 710049)
[2)](智能网络与网络安全教育部重点实验室(西安交通大学) 西安 710049)
[3)](武汉大学网络安全学院 武汉 430072)
[4)](罗格斯大学计算机学院 新泽西州 08854 美国)
[5)](清华大学网络科学与网络空间研究院 北京 100084)

摘 要 近年来,以深度学习算法为代表的人工智能技术在安防视频监控、个人隐私保护、自动驾驶等领域广泛应用.尤其在人脸识别等领域,深度学习方法显示出超越人类感知及辨别的能力,为人类的日常生活带来了诸多便利.然而,利用人工智能生成、对抗、伪造等技术产生的虚假数字人脸给个人隐私安全、社会安全乃至国家安全等方面带来了诸多风险和挑战.本文通过回顾虚假数字人脸内容生成与检测的相关研究工作,揭示其对国民、国家安全造成的潜在威胁.具体来说,本文首先介绍虚假数字人脸内容的攻击对象及攻击类型,从两种攻击对象—人工智能系统及人类感知系统,两大攻击类型—人脸对抗样本及人脸深度篡改,归纳、分析相应的生成、攻击及检测、防御技术.最后,本文讨论和展望虚假数字人脸内容生成与检测技术未来的研究方向和发展趋势.

关键词 虚假数字内容;人脸对抗样本;人脸深度篡改;人工智能安全;隐私保护
中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2023.00469

Digitally Forged Face Content Creation and Detection

LIN Chen-Hao[1)2)] SHEN Chao[1)2)] DENG Jing-Yi[1)2)] HU Peng-Bin[1)2)]
WANG Qian[3)] MA Shi-Qing[4)] LI Qi[5)] GUAN Xiao-Hong[1)2)]

[1)](*School of Cyber Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049)
[2)](*Key Laboratory for Intelligent Networks and Network Security (Xi'an Jiaotong University), Xi'an* 710049)
[3)](*School of Cyber Science and Engineering, Wuhan University, Wuhan* 430072)
[4)](*Department of Computer Science, Rutgers University, New Jersey* 08854 *USA*)
[5)](*Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing* 100084)

Abstract In recent years, artificial intelligence technologies, represented by deep learning algorithms, have been widely applied in many fields including smart video surveillance, privacy protection, autonomous driving and so on. Particularly in the field of face recognition, deep learning based methods have shown the ability of surpassing human perception and brought great conveniences to our daily lives. However, digitally forged face generated by adversarial and deepfake

收稿日期:2021-05-19;在线发布日期:2021-12-09.本课题得到科技创新2030—"新一代人工智能"重大项目(No.2020AAA0107702)、国家自然科学基金重大项目(62006181, U20A20177, 61822309, 61703301, U21B2018)、陕西重点研发计划项目(2021ZDLGY01-02)资助.
蔺琛皓,博士,研究员,博士生导师,主要研究领域为人工智能安全、对抗机器学习、深度伪造、智能身份认证.E-mail:linchenhao@xjtu.edu.cn.**沈 超(通信作者)**,博士,教授,博士生导师,CCF专业会员,主要研究领域为可信人工智能、人工智能安全和信息物理电力系统安全.E-mail:chaoshen@mail.xjtu.edu.cn.**邓静怡**,博士研究生,主要研究领域为数字图像伪造检测.**胡鹏斌**,硕士研究生,主要研究领域为数字图像伪造生成.**王 骞**,博士,教授,博士生导师,主要研究领域为人工智能安全、云计算安全与隐私、无线系统安全、应用密码学.**马仕清**,博士,助理教授,主要研究领域为系统与软件安全、对抗性机器学习、软件工程.**李 琦**,博士,副教授,博士生导师,主要研究领域为互联网和云安全、移动安全、机器学习与安全、大数据安全、区块链与安全.**管晓宏**,博士,教授,博士生导师,中国科学院院士,主要研究领域为电力与制造系统的优化调度、网络安全.

techniques poses huge risks and challenges for individual privacy, social security and even national security. This paper reviews previous work on the creation and detection of digitally forged face content, revealing potential risks to individual privacy, social security and national security. Specifically, we firstly introduce the attack targets and attack types of digitally forged face content. Secondly, we summarize and analyze digitally forged face content creation, attack, detection and defense technologies in terms of two attack targets: artificial intelligence system and human perception system, and in terms of two attack types: adversarial face example and deep face manipulation. Finally, the directions of future research on digitally forged face content creation and detection are discussed.

Keywords digitally forged content; adversarial face example; deep face manipulation; artificial intelligence security; privacy protection

1 引 言

近年来,以深度学习算法[1-3]为代表的人工智能技术[4-5]不断发展创新,并在计算机视觉[6-9]、语音识别[10-13]、自然语言处理[14-17]等多个领域不断取得突破及成功.基于人工智能技术的智能系统也在近些年被广泛部署应用于日常生活、生产场景中,如智能视频监控场景[18]、自动驾驶场景[19]、智慧医疗场景[20]等等,其算法认知及判断能力甚至超越人类的感知、辨别能力.尤其在人脸识别[21-24]的相关领域,2014 年,采用深度学习的人脸识别技术在识别精度上达到 98.52%[25],首次超越人类专家,促进了人脸相关的检测、识别、生成技术[26-27]的深入研究和发展,相关的电子身份认证、人脸门禁、人脸监控追踪等智能系统,为日常生活、公共安全等方面带来诸多便利.据市场调研报告①预测,到 2024 年,人工智能人脸识别技术应用的市场份额将达到近 70 亿美元.除此之外,人工智能技术给网络安全防御方面注入了新的活力.例如采用人工智能技术,智能防火墙通过分析和评测邮件的基本信息,可实现高精度的垃圾邮件过滤,并对数据包进行特征提取与分析提高了防火墙的防御性能.面对典型的网络攻击手段如 dos 攻击、网络蠕虫、恶意代码以及木马后门等,基于神经网络的人工智能技术利用有限的资源,有效地提高了攻击的检测能力.目前,人工智能技术已被广泛应用于网络系统安全及网络内容安全等领域.

然而,人工智能技术是一把双刃剑,它也可以被用作威胁网络空间安全的武器.基于人工智能生成对抗网络(GAN)[28]及自编码器(autoencoder)[29]等算法的深度生成、对抗、伪造技术[30-32]在近来受到广泛关注,利用该技术构造虚假的数字内容对人工智能系统或人类感知系统进行攻击的事件频有发生,给日常生活及社会稳定带来了较大隐患.如文献[33]中,作者通过对抗样本(adversarial examples)生成技术构造对抗扰动标记,贴在“停止”标志(Stop Sign)上,造成自动驾驶系统中的智能图像识别模块将原先的“停止”标志识别为“限速”等其他标志,这将严重威胁公共交通安全,带来严重的后果.利用深度伪造(deepfake)等技术,进行虚假数字人脸图像、语音、视频内容生成及克隆②的事件也频繁发生,给受害者造成了巨大的经济损失,甚至导致了严重的威胁商业乃至政治安全的事件.

作为敏感的个人生物识别特征,人脸生物特征及智能人脸识别技术得到了广泛研究.与此同时,虚假数字人脸内容生成技术也引起了研究者的注意,近年来虚假数字人脸内容也常被用来攻击人工智能系统或人类感知系统.如多伦多大学的研究人员[34]通过对抗训练的方式,可造成人脸检测系统及识别系统出错,从而导致 99.5%人脸无法被正确识别,造成智能人脸识别系统瘫痪.以色列一家公司利用深度伪造换脸技术合成并发布了一段虚假视频,视频中伪造的脸书(Facebook)首席执行官扎克伯格就技术垄断问题发表了不利言论,成功地欺骗了人类

① Facial Recognition Market by Component (Software, Tools, and Services). https://www.researchandmarkets.com/reports/4791675/facial-recognition-market-by-component-softwar

② Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

的感知,造成了巨大且恶劣的国际影响①.除此以外,基于深度伪造换脸技术的应用程序如 FakeAPP②、ZAO③等,在一段时间内被人们广泛使用生成虚假人脸,该类软件存在严重的安全隐患并对个人隐私信息具有潜在威胁.

面对虚假数字人脸内容的生成与攻击所带来的威胁,研究人员也提出了一些针对性的检测及防御策略[35,36],预防及降低虚假人脸攻击造成的严重损失.除此之外,多个国家及机构也出台了相应的政策法规以防止虚假数字人脸内容被滥用.如美国两党议员提出了《2018 年恶意伪造禁令法案》及《2019 年深度伪造报告法案》④⑤,用于阻止深度伪造内容可能带来的风险及构成的威胁.美国社交论坛 Reddit 关闭了 deepfake 等虚假数字内容相关的讨论⑥.在 2019 年,由全国信息安全标准化技术委员会大数据安全标准特别小组发布的"人工智能安全标准化白皮书"⑦中也分析讨论了虚假数字人脸内容的安全隐患,并指出需要通过法律条例加强此方面技术的管控和限制.

由此可见,针对虚假数字人脸内容的深入研究是重要且迫切的需求.虽然现有的研究中提出了诸多虚假数字人脸内容生成及检测方法,但是虚假数字人脸内容的类型众多,生成及攻击的形式、方法多样,相应的检测、防御策略也种类多样且各有侧重,生成及检测方法中也存在诸多问题,相关的研究尚未形成完整的技术体系.因此,围绕虚假数字人脸内容的生成与检测,我们亟需对现有的研究工作进行科学的归纳、分析及讨论,以便发现现有研究中的不足并为后续从事相关领域的研究人员提供方向性的指导.

现有的人工智能安全文献综述[37-39]多围绕人工智能技术的通用安全问题展开介绍,而缺乏针对数字人脸内容安全问题的详细介绍.如文献[37]总结了通用图像的对抗样本的生成技术,但未涉及到人脸对抗样本相关的内容.虽然针对人脸深度伪造技术有少量的文献综述[40,41],但现有文章多集中于对人脸换脸技术的总结,鲜有文献针对人脸内容涉及的安全问题进行全面的介绍.本文专注于数字人脸内容技术中全面的安全问题,从对抗样本及深度篡改两个方面,对现有的攻击与防御技术及篡改生成与检测技术进行细致介绍,并对现有研究中的不足之处进行总结分析,旨在推动虚假数字人脸内容生成与检测技术的进一步发展,并为保障人工智能人脸相关技术的安全应用提供指导和参考.

本文针对虚假数字人脸内容生成与检测的研究进展进行梳理、归纳、分析及讨论.第 1 章首先对虚假数字人脸内容进行定义及分类,并对相关的生成与检测算法及评价方式进行分类和概述.然后,第 2 章与第 3 章分别对干扰人工智能及人类感知系统的人脸对抗样本及人脸深度篡改的生成及检测相关研究进行归纳总结,并讨论现有研究的局限性与不足.最后,第 4 章对虚假数字人脸内容生成与检测技术面临的挑战及未来的研究方向进行讨论和展望.

2　虚假数字人脸内容的类型

虚假数字人脸内容涵盖的技术种类众多、生成与检测方法繁多、数据集及评测指标多样,因此,本文按图 1 所示的综述框架对虚假数字人脸内容进行展开介绍.如图 1 所示,本文将虚假数字人脸内容按攻击对象分为干扰人工智能系统及干扰人类感知系统的攻击.更进一步的,将其按攻击类型分为人脸对抗样本及人脸深度篡改两大类,进而分别对其生成与检测技术进行介绍,并从实现算法、使用数据集、评测指标及现存问题等多个方面进行归纳总结和展望.

本节首先对虚假数字人脸内容进行定义及分类,并从攻击对象及攻击类型的不同对虚假数字人脸内容进行分类及简要归纳介绍.

2.1　虚假数字人脸内容的攻击对象

虚假数字人脸内容是通过人工智能技术对自然人脸内容修改、生成、构造而成,它是指通过对人脸数据添加扰动、噪声或篡改、交换、伪造人脸,从而生成使人工智能系统或人类感知系统难以辨别或判断出错的虚假数字人脸内容.

本文根据虚假数字人脸内容的攻击对象的不同,即干扰的不同感知系统将其分为以下两类:

① Deepfake video of Facebook CEO Mark Zuckerberg posted on Instagram. https://www.cnet.com/news/deepfake-video-of-facebook-ceo-mark-zuckerberg-posted-on-instagram/

② FakeApp. https://www.malavida.com/en/soft/fakeapp

③ ZAO. https://apkproz.com/app/zao

④ Malicious Deep Fake Prohibition Act of 2018. https://www.congress.gov/bill/115th-congress/senate-bill/3805

⑤ Deepfake Report Act of 2019. https://www.congress.gov/bill/116th-congress/senate-bill/2065

⑥ Deepfake has been banned. https://www.reddit.com/r/SFWdeepfakes/comments/7vy36n/rdeepfakes_has_been_banned/

⑦ Artificial Intelligence security standardization white paper. http://www.cesi.cn/images/editor/20191101/20191101115151443.pdf

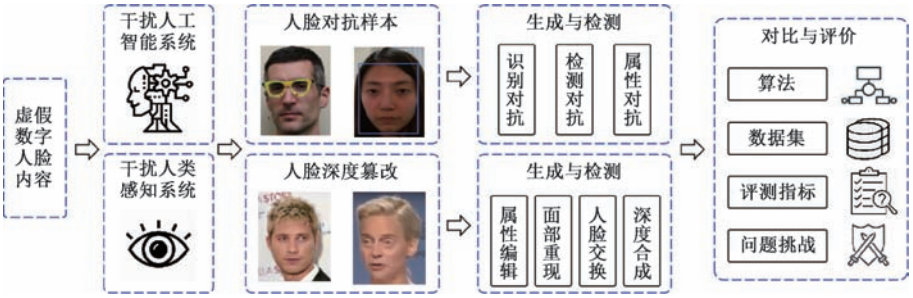

图 1 虚假数字人脸内容综述框架

干扰人工智能系统的人脸内容攻击. 通过对真实的人脸数据添加人类难以感知或不影响人类主观感知判断的扰动,从而造成以深度学习算法为核心的人工智能人脸检测、人脸识别等系统判断出错.

干扰人类感知系统的人脸内容攻击. 通过生成对抗网络[28]、自编码网络[29]、深度伪造①等技术,实现例如编辑人脸属性内容[42]、交换人脸内容[41]、重现人脸内容[43]等,从而产生人眼难以辨别的虚假数字人脸,造成人类感知系统判断出错,经过深度篡改的人脸同样可以影响人脸识别等人工智能系统的判断.

两种类型的虚假数字人脸内容在呈现形式及攻击对象上有明显的不同. 虚假人脸对抗样本不改变或不影响原始人脸数据的呈现内容,以造成人工智能人脸识别等系统判断出错为目标;虚假人脸深度篡改直接修改、交换或伪造原始人脸数据的呈现内容,主要以造成人类感知系统判断出错为目标.

2.2 虚假数字人脸内容的攻击类型

根据攻击类型的不同,本节对两种虚假数字人脸内容展开进一步细分并分别进行攻击类型的介绍.

人脸对抗样本常采用对抗样本(adversarial examples)生成技术[29]构造产生,该技术最初应用于通用图像并对图像分类、识别系统进行攻击[32],之后被用于生成虚假人脸内容,并攻击人脸相关的智能系统,包括人脸识别系统[44]、人脸检测系统[34]、人脸属性识别系统[45]等.

(1)人脸识别(face recognition)系统是通过人脸来识别来确定用户身份. 人脸识别对抗样本是通过对人脸图像添加人类不易感知的扰动或不影响主观判断的对抗补丁生成人脸识别对抗样本,造成深度学习人脸识别模型的输出出错或使其人脸识别置信率大幅降低,无法正确识别人脸所属身份.

(2)人脸检测(face detection)系统是用于检测人脸存在并捕捉人脸. 人脸检测对抗样本,面向人脸检测器或深度学习通用目标检测模型,通过引入对抗扰动或补丁至人脸图像生成人脸检测对抗样本,造成检测器的人脸检测框出错或使检测框的置信率大幅度降低.

(3)人脸属性识别(facial attribute recognition)系统是对某一人脸属性进行识别. 人脸属性对抗样本. 针对人脸图像引入特定的对抗扰动或补丁生成人脸属性对抗样本,从而造成人类属性包括性别、种族、年龄、表情等多种识别模型的判断出错,无法正确判断人脸所属用户的性别、种族等属性.

图 2 展示了现有研究中常见的可造成人脸识别系统出错的对抗样本. 如图 2 中(b-d)所示,通过添加特定设计且不影响人类主观感知判断的眼镜配饰,人脸对抗样本可造成智能人脸识别系统判断出错.

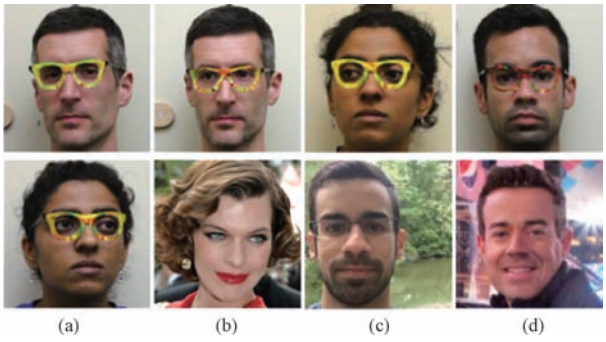

图 2 人脸对抗样本示意图((a)为人脸对抗样本,(b)～(d)为人脸对抗样本及对应的误判人脸[46])

在人脸对抗样本中,从原理上讲,人脸识别对抗样本与人脸属性对抗样本类似,都是面向图像识别/分类任务产生的攻击,其攻击目的是使目标人脸无法被识别或目标人脸属性无法被正确识别. 而人脸检测对抗攻击则是面向图像检测任务产生的攻击,其攻击目的是使经过对抗扰动的目标人脸无法被检测. 从技术成熟度上讲,目前围绕人脸识别对抗样本

① Deepfakes github. https://github.com/deepfakes/faceswap

的研究最为广泛,近年来围绕人脸检测对抗样本的研究也逐步兴起,而关于人脸属性对抗样本的研究较少,但其方法可借鉴通用图像对抗技术快速实现.现有的智能人脸识别系统多采用先进行人脸检测再进行人脸识别的流程,因此如果能成功对人脸检测环节实现攻击,则可以造成整个人脸识别系统的失效.

人脸深度篡改通常借鉴通用图像编辑、生成、伪造等方法而实现,主要攻击类型包括对人脸属性的内容编辑篡改[42,47],对人脸面部内容的重现、生成[43,48,49]、深度伪造人脸交换[41,50]及深度合成人脸[26]等.

(1)人脸属性内容编辑(face attribute manipulation)是指通过操纵面部属性,对包括眼镜、性别、头发、胡须、年龄、表情等单个或多个属性进行编辑修改从而生成新的人脸,但同时保留人脸的其他细节信息不变.

(2)人脸面部内容重现(face reenactment).该技术也称为面部表情或面部行为迁移,指通过将源人脸的表情、行为迁移到目标人脸上,实现目标人脸重现源人脸的表情或行为.较新的研究可以在没有源人脸数据的情况下,基于一段语音的学习,直接生成目标人脸的说话行为重现.

(3)深度伪造人脸交换.指深度伪造(deepfake)中针对人脸的伪造、交换(face swap)技术.旨在将视频或图片中的源人脸完全替换为目标人脸,但仍然保持源人脸的表情、姿态及背景不变.

4)深度合成人脸生成(entire face synthesis).旨在利用人脸图片样本,通过生成对抗等技术,合成完全不存在的人脸或对目标人脸缺失部分进行合成.

图 3 展示了现有研究中常见的一些不同类型的人脸深度篡改样本,这些样本直接对人脸的呈现内容进行篡改或伪造生成,容易造成人类感知系统的判断出错.

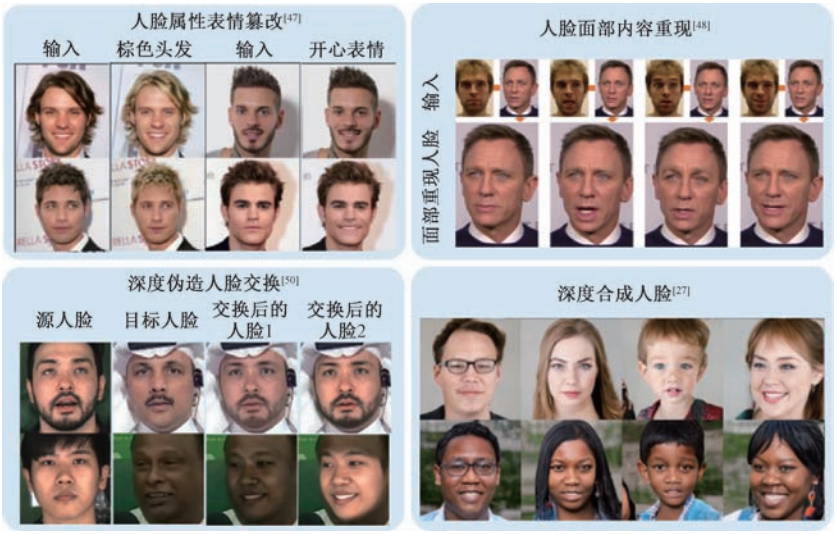

图 3　不同类型的人脸深度篡改样本

在人脸深度篡改中,从原理上讲,四种人脸深度篡改都可采用基于生成对抗网络(GAN)的技术实现,其中深度伪造人脸交换,多采用自编码器(autoencoder)的技术实现.人脸面部内容重现及深度伪造人脸交换是以目标人脸为参考,对人脸进行篡改;而人脸属性内容编辑及深度合成人脸生成则是直接对源人脸进行篡改或生成.从技术成熟度上讲,人脸属性内容编辑及人脸面部内容重现的研究起步相对较早,采用传统机器学习及图形图像学的算法也可实现以上两种人脸篡改.而深度伪造人脸交换及深度合成人脸生成基本都采用深度学习技术实现.由于人脸属性内容编辑、人脸面部内容重现及深度伪造人脸交换技术带来的潜在威胁更大,近几年内受到了更为广泛的关注及研究,其技术发展也相对较快.四种人脸深度篡改技术多基于 GAN、autoencoder 等生成方法进行优化改进,这些方法普遍存在生成效果不稳定等问题,造成现有方法整体上仍有较大的提升空间.但总体上,四种人脸深度篡改的成功攻击样本给学术界和工业界以及生活中都带来了严重的威胁.

本文分别对以上七种类型的虚假数字人脸内容的生成与检测相关研究进行归纳总结及展开介绍,并对现有的相关研究从算法、数据集、评测指标、存在的问题及挑战等多个维度进行了对比、分析及讨论.

3 人脸对抗样本生成与检测

早在 2014 年，Goodfellow 等人[29,31]通过对图像按一定规则添加微小的扰动从而生成对抗样本，造成人工智能图像分类、识别系统出错. 随后，该技术被优化拓展并应用到对智能人脸识别系统的攻击中，给目前广泛部署及使用的人脸检测、人脸识别等智能系统造成了巨大的安全隐患. 针对此问题，有研究人员[35]提出不同的检测方法来预防人脸对抗样本的攻击. 根据人脸对抗样本攻击任务的不同，本章将人脸对抗样本分为人脸识别对抗样本、人脸检测对抗样本、人脸属性对抗样本，并从对抗生成与对抗检测两个方面，对人脸对抗样本的研究进展进行介绍归纳与分析讨论.

3.1 人脸对抗样本生成技术

3.1.1 人脸识别对抗攻击

采用或借鉴文献[32,33]中通用图像对抗样本生成的思想，主流的人脸识别对抗样本的生成技术可分为物理实现的对抗人脸样本及全局数字对抗人脸样本两类[44,46,51-57].

物理实现攻击. Sharif 等人[44,46]提出，通过引入一种经过精心构造且不明显的、物理可实现的扰动(文中为眼镜配饰)，可实现对先进人脸识别系统①[23]的攻击，造成目标用户通过佩戴精心设计的眼镜，被误识别为其他用户. 针对白盒系统，作者采用文献[30][31]中的 LBFGS 方法实现攻击；针对黑盒系统，作者采用粒子群优化算法[58]实现攻击.

通过构造对抗性贴纸，Komkov 等人[51]及 Pautov 等人[52]实现了物理的扰动攻击. Komkov 构造一种具有对抗性的长方形贴纸，通过将贴纸放在额头上可成功迷惑公开的最先进的人脸识别模型[58]. 作者采用 FGSM[32]方法，并提出 TV loss 实现对抗样本的生成. Pautov 提出一种可贴至人脸多个部位的对抗性贴纸，实现对先进人脸识别白盒模型的攻击. 作者同样采用 FGSM[32]方法，并引入透视变换对形变的贴纸进行对抗性训练及生成，保证形变后的贴纸依然有效.

为了解决物理实现攻击隐蔽性差的问题，Zhu 等人[53]提出一种隐蔽性较强的人脸识别对抗攻击. 作者通过人脸伪造化妆的方式生成人脸识别对抗样本. 作者首先采用 GAN 实现化妆迁移网络，并设计对抗攻击网络将对抗扰动隐藏在化妆迁移样本中，实现人脸对抗样本生成及攻击.

虽然现有的物理实现的攻击方法实现了对白盒/黑盒系统的有效攻击，但此类方法生成的对抗样本隐蔽性不强，且容易受到现实环境变化的影响，如背景变化、光照变化、图像角度变化等，而导致对抗攻击失效. 并且在实际场景中，对抗性物理配饰本身的位置、形态、角度的微小变化都可能造成攻击失效. 因此如何生成更鲁棒、逼真且不易受环境影响的对抗样本，是物理实现人脸识别对抗样本的研究重点之一.

全局数字攻击. 与物理实现的攻击方式不同，全局数字的对抗样本通过引入全局的且不易被人眼感知的扰动，实现人脸识别对抗样本攻击. 在较新的研究中，Dong 等人[54]提出一种进化式的攻击算法，并通过降低搜索空间范围以最小的代价引入全局隐蔽的对抗样本扰动，实现对现有人脸识别模型以及多个商用的人脸识别系统的黑盒攻击. 除此之外，作者对多个主流人脸识别模型的鲁棒性进行评估，并分析其是否容易受对抗样本的攻击.

为了提升定向人脸识别攻击的能力，Song 等人[55]介绍一种注意力对抗攻击生成网络 A3GN，通过添加条件变分自编码器及注意力机制，学习并获取特定目标的语义信息，从而生成对抗性的人脸样本，实现人脸识别系统的定向白盒及黑盒攻击.

现有的全局数字攻击方法大多关注攻击的成功率及效率，鲜有研究致力于对抗样本的隐蔽性体提升. 虽然有研究评估了对抗样本的隐蔽程度，但是对抗扰动仍容易被人眼感知，尤其在高分辨人脸图像中，全局的扰动噪声明显可见. 因此，如何在保障攻击成功率及效率的前提下，不断提升对抗样本的隐蔽性，是未来此领域重要的研究方向之一.

表 1 对比了主流的针对人脸识别模型或系统的对抗性样本攻击. 其中白盒攻击是指攻击者能获取算法及模型参数；而在黑盒攻击中，攻击者则无法获取这些信息. 定向攻击是指对于一个分类网络，把输入分类错误判断到指定的类别上，非定向攻击指只生成对抗样本，而不指定误判的类别. Acc 代表攻击成功率(有些论文以 SR，即 success rate 表示).

虽然现有的人脸识别对抗样本攻击方法都取得了不错的效果，但仍存在诸多不足. 现有方法的实验设定、数据集选取、测试指标都不尽相同，如 Zhu[53]等使用模型出错率评价其攻击算法，Dong[54]等人通过计算对抗人脸的平均形变程度(MSE)评价其生

① Face++. http://www.faceplusplus.com

成样本的质量,而 Qing[55] 等人通过结构相似度(SSIM)评价生成人脸质量.这都导致难以对现有的方法做公平的效果对比.除此之外,诸多方法还缺乏对生成对抗样本的噪声程度进行量化评估,生成的数据中可能存在隐蔽性差、人眼明显可见的对抗攻击样本.因此,建立公平、全面、可量化的评测体系是人脸识别对抗样本技术发展的关键.

3.1.2　人脸检测对抗攻击

人脸检测通常是人脸识别系统的首要环节,人脸检测模块的失效将导致整个人脸识别系统无法运作.因此,针对人脸检测进行的对抗攻击也逐渐受到人们关注.现有的人脸检测对抗样本生成方法多基于 FGSM 方法[32] 进行改进优化,主流的人脸检测对抗攻击同样也可根据攻击方式不同大致分为物理域攻击和数字域攻击[34,62-64].

针对人脸检测模型的物理域攻击.Sharif 等人[46] 采用 FGSM 方法,将人脸的眼部及其周围区域作为扰动区域,打印出具有扰动图像的眼镜框架,攻击者通过佩戴具有扰动图案的眼镜即可对传统机器学习的人脸检测器 VJ detector[65] 中的一个级联分类器进行扰动攻击,从而造成整个人脸检测器的失效,使全部测试样本中的人脸无法被正常检测.该方法将已有的技术进行使用,仅对白盒系统有效,其实用性较差.

Kaziakhmedov[64] 等人设计多个黑白像素的对抗补丁,采用优化的 FGSM 算法对多个补丁联合训练,然后使用普通的黑白打印机打印出对抗补丁并将该补丁贴在口罩上进行佩戴或直接贴在人脸局部区域即可有效地攻击主流的人脸检测器 MTCNN[66].但该方法同样只适用于白盒的模型,导致其实用性较差.

表 1　主流的人脸识别对抗攻击算法对比　　(单位:%)

文献	算法	攻击类型	白盒/黑盒	定向/不定向	模型/系统	数据集	定向成功率/不定向成功率	结构相似度/平均形变度
Sharif[44,46]	处理边界约束 L-BFGS	物理攻击	白盒	定向/不定向	VGG-face①	PubFig[59]	91.67/100	—
	粒子群优化	物理攻击	黑盒	定向/不定向	Face++[23]	PubFig	60/100	—
Zhu[53]	化妆迁移网络	物理攻击	白盒	定向/不定向	ResNet50	自采集数据集[53]	90/97.5	—
Dong[54]	进化算法	数字攻击	黑盒	定向/不定向	ArcFace[60]	LFW[61]	100/100	—/0.0016
A^3GN[55]	注意力机制+生成网络(A^3GN)	数字攻击	白盒	定向	ArcFace	LFW	99.94/—	5.025/—

针对人脸检测模型的数字域图像对抗攻击.Lu 等人[62] 通过构造微小的扰动图案并添加到原始图像中,实现对通用目标检测模 Faster-RCNN[67] 和 YOLO9000[68] 的对抗攻击,并将此方法拓展到人脸数字图像上,实现人脸检测的白盒对抗攻击.Bose 等人[34] 采用约束优化的算法构造生成网络,生成对基于 Faster-RCNN 的人脸检测器攻击有效的对抗样本,并验证了该方法可有效攻击数字图像的 jpg 压缩防御[69] 方法.Yang 等人[63] 对目标检测的各个环节进行详细的分析,并基于优化算法设计了一种通用性的对抗补丁,通过将此对抗补丁贴在数字图像中人脸前额上即可有效地实现了对基于通用目标检测模型 FPN[70] 的人脸检测器的对抗攻击.

图 4 展示了人脸图像引入对抗样本扰动后使检测器无法检测到人脸的示例.目前针对人脸检测对抗攻击的研究较少,多数研究是基于通用目标检测的迁移或与人脸识别对抗攻击相结合的研究开展的,很少有研究针对人脸检测模型或系统地设计专门的对抗攻击方法,攻击的成功率也有待进一步提高.此外,现有的方法多针对白盒模型,实用性较差,而采用对抗性补丁方式的攻击不够隐蔽,且鲁棒性不强易受环境因素的影响.因此,设计黑盒有效且鲁棒隐蔽的人脸检测对抗样本是未来重要的研究方向之一.除此之外,现有的研究缺乏统一的基准数据集及评测方法,不同文献中攻击的检测模型、系统也各不相同,如何对现有的方法进行公平、全面的效果对比,建立量化评估体系是也该领域的重要研究方向之一.

图 4　人脸检测对抗样本示意图[62]

3.1.3　人脸属性对抗攻击

针对人脸属性识别对抗攻击的研究较少[45,71,72].人脸属性识别属于图像识别的一种,因此针对此的对抗攻击多数采用通用图像对抗样本生成的方法来优化和改进.

① VGG Face Descriptors. http://www.robots.ox.ac.uk/~vgg/software/vgg_face/

Mirjalili 等人[71]提出一种针对性别判断的对抗攻击,作者通过一种弯曲技术同时对人脸的一组像素点进行扰动修改,直到使性别识别器判断出错,同时作者引入了限制条件,保障生成的对抗样本几乎不影响人脸识别器的功能. Rozsa 等人[45]首先对基于深度神经网络的人脸属性识别模型的鲁棒性进行了评估,并基于 FGSM 方法提出一种快速翻转人脸属性的技术,使人脸属性的多个二分类器判断出错,如将涂口红判断为未涂口红,将男性判断为女性等.

Joshi 等人[72]利用语义信息设计了一种语义的对抗样本,采用 Attgan[73]的人脸属性语义修改算法,并在其中引入对抗扰动,使得在修改人脸语义信息的同时,如添加眼镜等,造成人脸属性识别器判断出错.

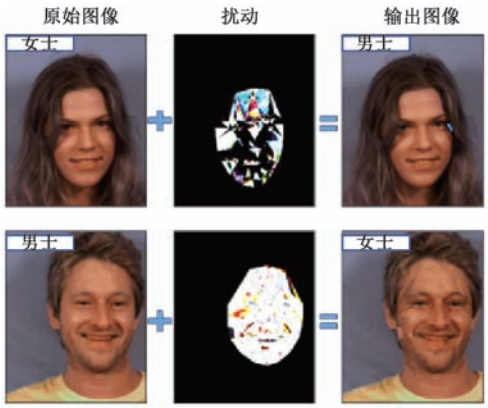

图 5 人脸属性对抗样本示意图[71]

图 5 展示了人脸图像引入对抗样本扰动后人脸属性识别模型出错的示例. 通过添加对抗噪声,原来识别为女性或男性的样本被错误的识别.

目前针对人脸属性识别对抗攻击的研究还处于初级阶段,现有的研究多直接将通用图像的攻击方法使用在人脸属性图像样本中以实现人脸属性对抗攻击,许多对抗样本不够隐蔽,容易被人眼区分. 除此之外,现有攻击方法的实用性较差,鲜有研究在黑盒的人脸属性识别模型或系统上实现对抗样本攻击. 由于此方面研究还处于初级阶段,因此目前还没有可以公平对比的统一数据集或针对性的评测指标. 设计针对性的对抗样本生成算法、统一的评价数据集及针对性评测指标是人脸属性识别对抗攻击技术发展的关键之一.

3.2 人脸对抗样本检测技术

3.2.1 人脸识别对抗检测防御

从 2018 年开始,针对人脸识别对抗检测防御的研究[35,36,74-77]开始引起人们的关注,根据防御方法的类型,现有研究可分为对抗攻击检测及对抗攻击防御,对抗攻击检测一般通过额外的模型对潜在的对抗样本进行检测,对抗攻击防御则是通过提升模型本身的鲁棒性及安全性,使其面对对抗攻击样本时仍有正确的输出.

对抗攻击检测. Agarwal 等人[35]采用传统的机器学习方法,提出了基于像素值及主成分分析的特征提取法,并利用支持向量机[78](SVM)作为分类器,实现了对全局人脸识别对抗攻击的检测.

利用深度学习可解释性分析,Tao 等人[77]提出了一种基于深度神经网络可解释性的人脸识别对抗样本检测技术. 该方法可有效地对可解释部分的神经元的激活值进行放大并同时抑制不可解释的神经元,从而通过内部神经元实现对对抗样本攻击的高效检测.

虽然现有的对抗攻击检测方法可有效地实现对对抗样本的检测预警,但是该方法普遍需要采用单独训练的分类模型,引入了额外的计算开销,不利于人脸识别系统的部署应用.

对抗攻击防御. Su 等人[75]研究人脸比对任务的对抗防御策略. 作者设计了一种新型的深度残差生成网络的训练方法,通过该方法可有效地清除针对人脸识别的对抗扰动,从而实现对人脸比对对抗攻击的有效防御.

除此之外,有研究对现有的人脸识别对抗样本检测及防御方法进行归纳总结. Goswami 等人[36,79]对现有的主流人脸识别模型/系统进行分析,认为现有的基于深度学习的人脸识别算法容易受到对抗样本扰动的攻击. 同时作者提出了可有效检测出对抗样本攻击的针对性的方法,并构建了一些缓解对抗样本攻击同时增强人脸识别模型鲁棒性的策略. Goel 等人[74]开发了一款名为 SmartBox 的用于检测人脸识别对抗样本及缓解对抗攻击的工具包,该工具包对现有的主流检测防御方法进行整理和实现,对多种通用的对抗攻击方法有一定效果.

然而,现有研究中多采用通用的检测防御方法,不是针对人脸识别对抗样本进行设计的,这导致现有的检测防御方法效果较差、效率较低,尤其是面对不断涌现的人脸识别对抗样本攻击算法,现有的检测防御方法已无法满足实际场景的需要,效果仍有待提升. 除此之外,不同的研究中多采用不同的数据集及不同的评价标准对所提出的方法进行评估,这导致难以对现有的人脸识别对抗检测及防御方法做公平的效果对比.

3.2.2 人脸检测对抗检测防御

针对人脸检测对抗攻击的检测防御策略还未得

到深入的研究,现有大多数研究都采用通用的攻击检测防御方案对此类攻击进行防御[69,80,81].如文献[62]中,作者采用了一种基于图像处理的方式[80]实现对人脸检测对抗攻击的检测防御.因此,研究并开发针对性的人脸检测对抗样本检测防御方法是该领域的重要研究方向.

3.2.3　人脸属性对抗检测防御

现有工作鲜有对人脸属性对抗攻击进行检测防御的研究.由于人脸属性攻击属于通用图像对抗攻击的一种,从理论上,通用的攻击检测及防御方案[69,82]对人脸属性对抗攻击应有一定效果,但目前还没有研究对此进行实验验证,也没有研究设计针对人脸属性攻击的检测防御方法.因此,实现人脸属性对抗样本检测防御并研究针对性的算法是保障人脸属性识别技术鲁棒、安全的关键.

3.3　人脸识别系统和通用人脸数据集

3.3.1　基于深度学习的人脸识别系统

基于深度学习的人脸识别系统一般包含人脸检测(Face Detection)功能、人脸对齐(Face Alignment)功能和人脸表征(Face Representation)功能,这通常也是基于深度学习的人脸识别的三个流程.人脸检测即检测出图像中人脸的具体位置,并通常用矩形框定位人脸;人脸对齐是使用人脸检测中的关键点对人脸做对齐校准,以缓解由于人脸姿势不同带来的识别误差;人脸表征指对预处理后的人脸图像进行特征提取.人脸识别对抗攻击多发生于人脸检测与人脸表征阶段,会导致整个识别系统的错误和崩溃.

现存的一些人脸识别系统虽然具有较高的识别准确率,但比较容易受到潜在的对抗算法的攻击.例如 FaceNet[22]是经典的人脸识别系统,其利用卷积神经网络(CNN)将人脸图像映射到欧式空间上,然后直接学习图像到欧式空间上点的映射,通过衡量两张图像所对应的特征的欧式空间上的点的距离来判断两个人脸是否相似,进而实现人脸识别功能.DeepID[83]可以判断两张人脸图像是否属于同一个人,该方法通过提取人脸特征将人脸划分为不同的区域,并训练多个单独的 CNN 来提取不同区域的人脸身份信息,再通过将提取到的多个区域的特征信息串联起来构建出整个人脸的特征向量,送入到分类器中完成人脸验证,DeepID 在 LFW 数据集取得了 97.45%的准确率.诸如此类的人脸识别系统都曾被攻击者作为目标进行攻击[84].

3.3.2　通用人脸数据集

LFW[61](Labeled Faces in the Wild)人脸数据集是目前人脸识别的最常用数据集之一,共包含 5749 个人的 13233 张人脸照片,其多数人脸图片来自于生活中的自然场景,均采集于互联网中.LFW 广泛运用于人脸识别的测试中,用于评价算法性能.

PubFig[59]与 LFW 类似,是公众人物脸部数据集,包含 200 个人的 58797 张人脸图像,相比于 LFW 数据量更大,每个人物拥有更多的图像样本.

CelebaA[85](CelebFaces Attribute)包含 10177 个公众人物身份的 202599 张人脸图像,且每张都包含了 40 个属性注释,常用于人脸属性的识别.

FFHQ[27](Flickr-Faces-High-Quality)是一个高质量的人脸数据集,包含 70000 张 1024x1024 的高清人脸图像,人物的年龄、种族、图像背景丰富多样,包含着多种人脸属性,例如眼睛、帽子、饰品等,可用于人脸识别及人脸属性分割.

VGGFace[23]是牛津大学在提出 VGG 网络时同时发布的人脸图片数据集,包含 2622 个人物,每个人包含 1000 张图片,数据集中的图像均来自互联网,在年龄、种族以及姿态方面有着较大的差异.

4　人脸深度篡改生成与检测

随着自编码器(autoencoder)、生成对抗网络(GAN)等神经网络图像生成技术[28,29]的发展,研究人员将此技术应用于人脸图像实现人脸的深度篡改生成.采用此类技术可任意修改、编辑、模拟人脸属性[42],实现人脸重现及面部表情、行为生成[43],进行人脸交换[41],甚至可以合成真实世界不存在的人脸[27].该技术最初被用做娱乐性质,但随后被用于政治、色情、恐怖主义、经济犯罪等,造成严重的个人、社会乃至国家安全危害.因此,针对人脸深度篡改攻击的检测防御技术也受到重视及深入研究,研究人员提出不同的方法[86-89]实现深度篡改人脸的检测防御.根据人脸篡改方式的不同,本章将人脸深度篡改分为人脸属性内容编辑、人脸面部内容重现、深度伪造人脸交换及深度合成人脸生成,并从篡改生成与篡改检测两个方面,对人脸深度篡改样本的研究进展进行介绍归纳与分析讨论.

4.1　人脸深度篡改生成技术

4.1.1　人脸属性内容编辑

随着生成对抗网络(GAN)技术的提出和发展,基于此技术进行人脸属性编辑篡改的方法也不断涌现[47,73,90-102].主流的人脸属性内容编辑主要是通过基于 GAN 的训练范式,也有少量研究是单纯基于

autoencoder 模型的方法，还有一些方法结合了对抗学习范式和 autoencoder 模型.

基于 GAN 训练范式的属性编辑. 最早的人脸属性内容编辑方法由 Perarnau 等人[103]提出，作者基于条件 GAN（cGAN）[104]方法，并通过引入编码器，构造可逆条件 GAN（IcGAN），成功实现对性别、头发颜色、化妆等多个人脸属性的编辑生成. 但是此生成方法不够稳定，容易对人脸的非编辑区域产生影响甚至产生完全无意义的人脸样本.

为了使生成的人脸更加稳定，并保证生成属性的非编辑区域不受影响，一系列的方法被人提出. Zhang 等人[105]将空间注意力机制引入 GAN 的框架中，并称之为 SaGAN. 该方法仅对特定属性的编辑区域进行修改，而保持其他部分不变从而成功实现更稳定的人脸属性的编辑. 该方法获得了比 ResGAN[106]、IcGAN 更好的人脸属性编辑效果.

为了实现跨域（domain）和跨数据集的人脸属性编辑，Choi 等人[47]提出 StarGAN 的模型，仅基于单个生成对抗模型并利用掩码向量的方法，即可实现跨域的人脸属性编辑. 与之前的方法不同，该方法中的生成器不是学习固定的转换，而是接收图像和域信息作为输入，以在相应的域中生成图像. StarGAN 还可以从包含不同类型标签的多个数据集中学习. 相比于之前的方法 StarGAN 更具有实用性.

在文献[107]中，作者提出一种全新的框架 InterFaceGAN，通过对隐层语义信息的解释实现语义人脸属性的编辑. 该方法首次对基于 GAN 的人脸属性编辑框架的语义信息进行可解释性分析，并对在 StarGAN 等方法中常产生的伪影有明显的修复效果.

除此之外，Jo 等人[108]提出一种异于之前工作的人脸属性编辑 SC-FEGAN. 该方法以用户对需要编辑区域的手绘草图为目标，实现对该区域的属性编辑.

基于 autoencoder 模型结构的属性编辑. 仅利用 autoencoder 技术实现人脸属性篡改的研究较少. Lample 等人[109]提出一种基于 autoencoder 结构的人脸属性编辑方式. 作者通过在编码过程中引入对抗实现属性编码和非属性分离，并通过解码过程对图像进行重构生成. Chen 等人[110]也提出一种 autoencoder 结构的基于语义框架的人脸属性操控方法. 作者将人脸属性分解为多个语义组件，并针对需要编辑的语义组件进行修改和重构，从而获得更精细的人脸属性编辑. 但是，相比于基于 GAN 或 GAN 与 autoencoder 相结合的生成篡改方式，此类方法在生成效果、稳定性等方面都处于劣势.

基于 GAN 训练范式和 autoencoder 模型结构结合的属性编辑. 在 autoencoder 和 GAN 结合的工作中，autoencoder 通常提供生成器的功能，可以与鉴别器一起构成 GAN 训练范式. He 等人[73]提出一种 AttGAN 的人脸属性编辑方法. 该方法不对属性的隐式表达施加约束，而是对编辑生成图像的属性分类进行约束，从而保证生成属性类别的正确性. 作者还提出重建损失方法来对非编辑区域的细节信息进行保留，基于 GAN 训练范式与 autoencoder 结构结合实现了人脸属性的编辑生成. 虽然方法鲁棒性有所提升，但是在面对大面积编辑区域时，仍可能产生错误的人脸样本.

为了解决人脸属性编辑精细化程度不够的问题，同时保持多种的属性编辑有效，Liu 等人[111]基于 AttGAN 提出从选择性迁移的视角解决人脸属性编辑问题的方法. 作者提出的模型 STGAN 以目标和源属性向量之间的差异作为输入，并可以自适应地选择和修改编码器特征用于增强属性编辑的效果. 这也使得 STGAN 相比于 AttGAN、StarGAN 在更精细的属性编辑中表现得更好.

为了提高人脸属性编辑过程中的交互性，Lee 等人[106]提出了一个支持交互式和多样化面部操作的新的框架，称为 MaskGAN. 框架主要由两部分组成：密集映射网络（DMN）和编辑行为模拟训练（EBST），其中 DMN 学习用户修改的 Mask 和目标图像之间的样式映射，从而支持不同的生成结果. EBST 可以对源 Mask 上的用户编辑行为建模，使整个框架对各种操作输入更加健壮.

表 2 对比了主流的人脸属性编辑生成方法. 主流的算法都采用同样的基准数据集及评价指标进行实验效果对比，评价指标主要包括峰值信噪比（PSNR）、结构相似度（SSIM）和 FID[112]（Fréchet Inception Distance）. PSNR 是使用最为广泛的一种衡量图像失真或是噪声水平的客观评价指标，通过对比当前图像与参考图像的均方误差评价图像质量，PSNR 数值越大表示图像质量越好. SSIM 分别从亮度、对比度、结构三方面度量图像相似性，取值越大表示图像质量越好. FID 可以表示生成图像的多样性和质量，FID 越小，则图像多样性越好，质量也越好. 除此之外，多数的研究采用生成样本图像可视化的方式与其他方法进行对比，从人眼主观角度对生成人脸的真实性进行判断及评价.

表 2　主流的人脸属性编辑算法对比

文献	算法	数据集	峰值信噪比/结构相似度	Fréchet Inception 距离
Perarnau[103]	IcGAN	CelebA[85]	15.28/0.430	—
He[72]	AttGAN	CelebA	24.07/0.841	—
Choi[49]	StarGAN	CelebA	22.80/0.819	30.17
Liu[111]	STGAN	CelebA	31.67/0.948	—
Lample[109]	FaderNet	CelebA	30.62/0.908	—
Lee[106]	MaskGAN	CelebA	—	46.67

关于人脸属性编辑的研究一直是领域内的热点问题,现有的研究已经形成了初步的体系.然而,现有的研究仍存在诸多不足:1)目前的人脸属性编辑方法普遍存在训练中模型参数震荡、稳定性不足的问题,多数方法需要大量的数据及较长的训练时间才能获得相对稳定的生成样本;2)多数的人脸属性编辑样本真实性不足,尤其是在分辨率较高的情况下,经过编辑的图像样本容易被人眼感知;3)现有的人脸属性编辑多采用正面、相对固定分辨率的人脸作为训练及测试,其在真实复杂场景中的有效性及鲁棒性还未得到验证.这些问题及挑战也是人脸属性内容编辑未来研究的重点.

4.1.2　人脸面部内容重现

人脸面部内容重现可分为图像驱动人脸重现和音频驱动人脸重现.图像驱动人脸重现的任务目标为从源人脸图像中提取表情、姿态等信息并在目标人脸图像中重现出来,表现为将源人脸的表情、姿态迁移至目标人脸上[43,48,113,114];音频驱动人脸重现的任务目标为根据语音的指导,生成目标人脸说话行为的视频图像[49,90].早期的人脸面部内容重现采用传统 3D 人脸建模、机器学习及图像处理的方法,随着深度学习技术的普及,借助于 DNN、GAN 等深度神经网络实现人脸面部内容重现的方法也不断涌现.

图像驱动人脸重现.在人脸 3D 建模的基础上,Thies 等人[43,48,113]先后提出多种方法实现人脸面部内容的重现.在文献[43]中,作者采用 RGB-D 传感器实时采集源人脸和目标人脸图像,并将每一帧需要迁移重构的信息进行学习和生成,实现源人脸的面部细节信息及表情、姿态的迁移重现.为了解决视频中面部特征恢复受限的问题,作者[43]提出非刚性模型捆绑的方法,采用密集的光度一致性度量跟踪源人脸和目标人脸的面部信息,从而实现更精准的面部重现.在文献[113]中,作者提出延迟神经渲染的技术,通过将可学习的组件融入到传统的图形学处理流程中,并利用三维的神经纹理,实现人脸面部内容重现.该方法可以获得更清晰的重构细节信息.

为了实现更真实的人脸重现,Kim 等人[114]首次提出全三维的人头位置、旋转、面部表情、眼部信息等的人体、面部内容重现.作者提出基于 GAN 及空间时序编码和图像解码方法,实现了高质量的模仿源人脸面部表情的面部重现.

基于人脸 3D 建模的方式,由于其预定义的人脸 3D 参数模型难以覆盖到所有的人脸运动,且需要非常复杂的参数设置,很少有广泛的应用.随着生成对抗网络的发展,Pix2Pix[115]和 CycleGAN[116]等方法的提出,Wu 等人[117]提出了一个使用 GAN 的重现方法,并且引入面部边界潜空间,将图像映射到边界空间,并调整其适应目标人脸,缓解了 CycleGAN 会在极端条件下生成非自然图像的问题,并且可以在没有成对数据的情况下实现图像驱动的面部重现.

由于人脸重现时,驱动人脸和目标人脸的身份信息的不同会导致重现质量的下降,在少样本重现的情况下尤其严重.驱动人脸的身份信息会泄露至重现图像,或在处理未见的大姿态时,模型会丢失目标的身份信息.Ha 等人[118]提出了一个使用人脸标识检测器,并引入图像注意力机制的模型 MarioNETte.通过人脸标识检测器解耦人脸的身份和姿态信息以及通过扭曲结构使驱动人脸标识适应目标身份,减轻了身份信息泄露的问题,实现了更高质量的人脸重现.

Zakharov 等人[119]借鉴风格迁移的思路,使用 AdaIN[120]将嵌入网络输出的外观信息注入到生成模型中,实现了少样本的人脸重现.由于使用人脸关键点对图片进行表情、姿态信息提取,会缺少较多的细节,并且许多数据集不具备标注的结构基础,Burkov[121]在此基础上提出了使用姿态编码器编码姿态向量,并将其利用 AdaIN 注入到生成模型中的方法.仅利用姿态提取网络,便能够实现姿态和身份的解耦.除此之外,空间自适应归一化模块(SPADE)[122]的提出在带来了条件图像生成模型革新的同时,也推动了图像驱动人脸重现模型的发展迭代.传统方法中使用卷积、归一化、非线性激活函数的处理会因归一化层而清洗掉特征中的语义信息,利用条件归一化层通过空间自适应去调节归一化层的激活情况,可以有效传播语义信息.Hao 等人[123]在 Zakharov 工作的基础上,选择使用嵌入网络学习人脸标识信息,将驱动人脸的结构信息作为条件使用 SPADE 模块注入生成器网络,实现了较

高质量的单样本人脸重现.

然而上述方法使用人脸标识进行重现通常需要大量的人脸标识的标注信息,或需要关于对象的特定先验信息. Aliakasndr 等人[124]提出了使用自监督的方式解耦外观和运动信息,并通过自监督学习的关键点和局部仿射变换估计光流并扭曲的一阶动作模型. 模型采用了端到端的训练方式,避免了大量的人脸标识标注花费;同时使用了遮挡感知对不能用扭曲重现的部分进行修复,实现了质量高、泛化性好的人脸重现,并且在动作迁移等任务上也拥有较好的表现.

虽然现有的方法以较高的质量实现了图像驱动的人脸重现,但是这些方法普遍存在训练数据成本较高,训练模型迁移泛化性不足,在非训练语料库范围内的数据中表现较差等问题. 除此之外,目前针对人脸面部重现技术的评估缺乏公平、统一的量化评测指标,现有方法多采用可视化的方式,直接对比人脸重现部分的细节信息进行不同方式优劣的评判,不利于该技术的体系化发展.

音频驱动人脸重现. 最早期的音频驱动人脸生成采用结构信息的拼接技术[125,126],如 Bregler 等人[125]直接修改嘴巴的轮廓. 随着深度学习模型的发展,人脸面部内容重现的思想和技术的迭代,研究人员[49,109-115]利用语音信息,实现人脸面部说话行为的生成. Suwajanakorn 等人[49]用互联网的图像及奥巴马的声音信息,成功地合成了准确的嘴唇同步说话的高质量视频,作者利用循环神经网络学习从原始音频特征到嘴部形状的映射,并综合嘴部纹理信息及采用三维姿态进行匹配拟合. 虽然此方法可逼真地生成人脸说话的行为,但却需要对特定的目标进行大规模数据的训练和建模.

为了解决以上问题,研究人员寻求建立一种通用模型,可以通过一个模型处理所有身份,其关键思想为确定音频和视频之间的同步关系. Zhou 等人[90]提出仅用一张照片和一段任何人说话的语音,通过听觉和视觉信息的对抗分解方式,实现高分辨率逼真人脸说话视频的生成. 作者采用了将说话视频映射到人脸身份特征空间及说话内容特征空间的方法,实现了信息互补和逼真生成. 由于从人脸的动作中分离解耦头部的姿势较为困难,因此以上根据语音和视频同步的重建方式容易忽视头部的运动.

与上述学习音频到视频的同步映射的方法不同,人脸的结构信息如人脸标识等在基于生成对抗网络的模型中经常被用于作为中间表示. 如 Chen 等人[91]设计了一种级联 GAN 的模型将音频信息迁移到高维的面部特征点结构,同时提出了具有注意力机制的动态可调像素损失,加强了网络对视听相关区域的关注,实现了对人脸形状、角度、噪声音频等都具有较好的鲁棒性的视频生成. Das 等人[127]将该问题划分为运动学习和纹理学习两部分,同样利用级联 GAN 模型分别学习嘴唇运动和纹理生成,具有较高的泛化能力. 此外,还有一些使用 3D 标识的方法[128,129]. 然而在单样本或极端条件下,此类方法精度显著下降.

以上方法均能实现唇形与音频同步的人脸重现,但是在驱动头部运动和重现头部姿态方面仍有所欠缺. 在极端环境下,使用人脸结构信息的方式容易因为信息的不确定性而引起重现质量退化. Zhou 等人[130]提出了一个隐式模块化视听表达的模型,作者将说话的人脸表征模块化为语音空间、姿态空间和身份空间,从姿态源视频学习姿态动作,并通过一个低维姿态特征指导人脸姿态生成. 实现了姿态可控、唇形同步的说话人脸重现.

虽然音频驱动人脸重现的研究起步较晚,但是这方面的工作已经形成了初步研究体系,主流的研究多基于同样的数据集开展实验,并采用统一的评价指标进行效果对比. 表 3 对比了主流的音频驱动人脸重现的方法. 多数对比方法使用峰值信噪比(PSNR)、结构相似度(SSIM)和人脸标识距离(LMD)的定量指标进行评价,除此之外,连续帧生成样本的可视化主观评估也是主流的评价手段之一,如 Yi[92]采用用户评分的方式对重现质量进行评判,20 名参与者共比较了 30 组重现视频,并从图像质量、嘴唇同步性和自然程度进行打分评判.

然而现有的音频驱动人脸重现研究也存在一定的不足,包括连续帧生成不自然,面部表情固化等,如何重现出更连续且自然逼真的说话行为是未来重要的研究方向之一.

4.1.3 深度伪造人脸交换

广义的深度伪造包括多种针对人脸、语音等内容的修改、生成的技术[41],本节关注于深度伪造人脸交换的讨论. 现有的深度伪造人脸交换技术大多基于 autoencoder 及 GAN 等深度学习生成模型进行实现及改进优化[96-99].

2017 年底,在美国 Reddit 论坛上,出现最早的深度伪造人脸交换视频/图像,一位匿名的用户通过深度伪造技术[46],将一位明星的人脸与色情人员的

表 3　主流的音频驱动人脸重现算法对比　　(单位:%)

文献	算法	数据集	PSNR/SSIM/LMD	图像质量/嘴唇同步/自然程度
Chen[91]	级联 GAN	LRW[95]	30.91/0.81/1.37	5.67/32.33/9.50
Chung [93]	编解码 CNN	LRW	29.91/0.77/1.63	3.50/20.50/4.17
Zhou[90]	视听信息分解	LRW	29.90/0.73/1.73	2.17/2.33/2.67
Yi[92]	记忆增强 GAN	LRW	30.94/0.75/1.58	88.67/44.83/83.67
Wiles[94]	x2Face 网络	LRW	29.82/0.75/1.60	—

人脸进行了交换①.虽然此应用的性质恶劣且非法,但它也引起工业界和学术界的兴趣与重视,推动该技术的发展.目前许多国家也已经通过立法手段保障深度伪造人脸交换技术的积极正确发展及应用②③.

在商业应用领域.许多公司开发了用于娱乐的伪造换脸程序④⑤.在著名的代码服务平台 Github 上,有研究人员分享深度伪造人脸交换相关的项目⑥⑦,提供目前流行且实用的深度伪造人脸交换方法,他们的方法都是基于 autoencoder 生成技术实现.在文献⑦中,作者通过 encoder 提取到源人脸的隐层信息并通过 decoder 实现了与目标人脸的生成、交换.在此基础上,作者利用 GAN 对基于 autoencoder 的深度伪造人脸技术进行优化,通过引入人脸分割掩码(segmentation mask)及对抗损失和感知损失,提出了细节更加逼真的深度伪造换脸技术.

在学术领域.深度伪造人脸交换侧重于提升算法的性能及建立标准数据集.Korshunova 等人[100]基于 CNN 提出一种前馈神经网络实现自动化实时换脸的方式,使用较少的目标人脸图片完成人脸的转换,将人脸转换提高到近乎达到实时换脸的速度.在损失函数上通过考虑环境光照条件的影响,对损失函数进行改进使得图像有较高的真实感.

除了针对完整的人脸进行人脸交换,Nirkin 等人[131]针对原图像中有部分面容被遮挡的人脸,进行人脸交换的研究.作者使用运动线索和 3D 数据增强来扩充数据集,通过输入图像建立面部形状的 3D 姿势和面部表情的界标,然后使用全卷积网络从面部的遮挡中分割出可见部分完成分割.该方法在准确度及速度上有很好的表现.

FSGAN 是 Nirkin 等人[132]提出的一种新的非特定目标换脸以及重现方法,其实现主要分为三个步骤,面部重现及面部分割、面部修补和面部融合.面部重现及面部分割完成对源人脸面容重现以及对目标人脸部分进行语义分割,只保留人脸部分,获得人脸掩码(mask).面部修补根据前一步得到的重现后的源人脸利用分割出来的目标面部区域范围对缺少的部分进行填充修补.面部融合根据修补后的源人脸、目标人脸图像以及其语义分割图像进行混合,完成换脸,在融合阶段提出了使用生成图片与原始图片的融合来对细节上进行优化,减少了由于使用人脸掩码而在掩码周围可能产生的伪影.FSGAN 支持图像和视频的面部交换,在实际运用中有着良好的表现.

Faceshifter[133]是一种基于 GAN 训练方法的高保真、能够感知遮挡的深度伪造人脸交换算法,包含两个独立的训练阶段.第一阶段中的 AEI-Net 用于完成换脸,加强了对目标人脸的属性特征提取,通过设计一个类似 U-Net 网络结构的网络来逐层输出目标人脸的属性特征.源人脸的身份信息(identity)则是通过使用 Arcface 产生身份特征向量,同时使用一个全新的自适应注意力生成器(AAD)来完成源人脸身份信息与目标人脸属性特征进行融合.第二阶段设计利用 HEAR-Net 网络用于解决面部遮挡问题,包括刘海,眼镜以及口罩等遮挡.

MegaFS[134]是基于 GAN 模块的高清换脸算法.在编码端基于 ResNet 和 FPN 来预测人脸图像的隐空间(latent space)将人脸图像编码为潜向量(latent code)并划分为中高低三部分,只对部分最高部分的特征信息进行交换,在解码端利用表现优秀的 StyleGAN2 预训练模型用于解码,生成出 1024x1024 的高清人脸图像.MegaFS 在训练时可以使用模块化训练,即编码模块和特征信息交换模块的训练可分开进行,以减少在训练时间和硬件资源上的花费.同时 MegaFS 是一种探究 GAN 的逆映射(GAN Inversion)的换脸算法,在对使用不同隐空间下的图像潜向量编码进行对比时,使用了感

① Porn Producers Offer to Help Hollywood Take Down Deepfake Videos. https://www.yahoo.com/entertainment/porn-producers-offer-help-hollywood-173047833.html

② Malicious Deep Fake Prohibition Act of 2018. https://www.congress.gov/bill/115th-congress/senate-bill/3805

③ Deepfake Report Act of 2019. https://www.congress.gov/bill/116th-congress/senate-bill/2065

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

⑤ ZAO. https://apkproz.com/app/zao

⑥ Deepfakes github. https://github.com/deepfakes/faceswap

⑦ DeepFaceLab. https://github.com/iperov/DeepFaceLab

知图像块相似度(LPIPS)[134]作为指标,证实了作者所提出的 W++隐空间相比于 W+隐空间在 LPIPS、MSE 以及换脸成功率上都有着更好的表现.其中 LPIPS 用于度量两张图像之间的差别,相比传统度量指标,如 L2/PSNR, SSIM,更符合人类感知系统衡量标准,LPIPS 的值越低表示两张图像越相似,反之,则差异越大.

目前公开的学术深度伪造人脸交换数据集质量并不高,与网络上呈现出的伪造视频质量相差甚远,因此,有部分研究工作利用采集的精细数据集生成更逼真的深度伪造换脸.Li 等人[135]提出一个新的伪造换脸视频数据集,Celeb-DF.它由 590 个 YouTube 真实视频和 5639 个伪造换脸视频组成.作者通过提高人脸图像的分辨率,减少伪造视频中的颜色失真,改进数据集中掩码的生成步骤并使用卡尔曼平滑算法对人脸的时间序列进行滤波,以减少每帧中真实值的不精确变化.Celeb-DF 在视频质量上远高于现有的伪造换脸公开数据集.最近的研究中,为了促使深度伪造人脸视频帧之间有更好的连贯性,文献[47]在基于 autoencoder 技术的深度伪造人脸方法上引入光流法,使连续多帧之间的人脸差异变化更加平衡,从而获得更逼真的伪造人脸.另外,作者还制造并发布目前最大的深度伪造人脸数据集 DeeperForensics-1.0,该数据集包含 60,000 个视频及对应的视频抽帧.

除了生成更加逼真的伪造换脸图像视频以外,Gandhi 等人[136]采取另外一种思路,他们基于对抗样本攻击(FGSM)提出使用对抗性的扰动来增强伪造换脸图像的真实性、鲁棒性及抗检测性,并可以欺骗普通的深度伪造检测器.

深度伪造人脸交换的研究已形成了初步的研究体系,并在实际场景中已有较多的应用,同时也有文献对现有的方法进行了多方面的对比.然而,现有的研究仍存在一些问题及挑战,主要归纳为:1)伪造图像样本在高分辨率下的噪声较明显,真实度不够高,容易被人眼识别;2)伪造视频样本在视频帧之间的连贯性较差且背景信息易受影响,容易被人眼或算法感知,许多检测算法利用这一弱点实现了精确的伪造视频检测;3)现有的研究中存在源数据集多样、不统一,评测指标单一、不全面等问题,这使得难以对现有的伪造生成方法进行全面公平的对比.这些问题与不足也是未来研究工作的重点.表 4 对比了主流的深度伪造人脸交换的方法,针对现有交换方法的评估,目前主要有三种指标[50,132,133,135].其中掩码结构相似度是对比真实人脸和伪造人脸的结构相似度指标,表中所列的指标值为源数据集人脸和由其生成的伪造人脸的掩码结构相似度;真实度是通过人工参与者直接判断人脸交换检测基准数据集中的虚假数据是否为真实人脸的"真实度"指标;姿态损失(pose error)是通过人脸识别模型计算出人脸关键点(landmark),获取到的人脸姿态信息常用三个欧拉角(pitch,yaw,roll)并以角度为单位计算出交换后的人脸与源人脸之间的误差.但是这三种指标并不能全面地评估伪造人脸的效果,需要提出新的评价标准对生成效率、多帧一致性等指标进行评测.

表 4 主流的人脸交换算法对比

算法	源数据集	人脸交换检测基准数据集	掩码结构相似度	真实度/%	姿态损失
FakeApp①	CEW Dataset[137]	UADFV[138]	0.82	14.1	—
Faceswap-GAN②	VidTIMIT③	DF-TIMIT[139]	0.80	12.3	—
FaceSwap④	FF++-YouTube	FF++-FS[89]	0.81	8.4	—
DeepFake⑤	FF++-YouTube	FF++-DF[89]	0.50	—	4.14
DF-VAE[50]	Collected Source[47]	DF-1.0[47]	—	64.1	—
FSGAN	FF++-YouTube	DFDC[140], ForgeryNet[141]	0.51	—	—
Nirkin[131]	FF++-YouTube	—	0.49	—	3.29
FaceShifter	FF++-YouTube	FF++-FaceShifter[89], ForgeryNet	—	—	2.96
MegaFS	CelebA-HQ	MegaFS Dataset[142]	—	—	3.81

4.1.4 深度合成人脸生成

深度合成人脸生成技术可分为完全合成与缺失合成,即合成完全不存在的人脸或对目标人脸缺失部分进行合成.现有的深度合成人脸生成技术多基于 GAN 的方法进行优化改进[143-148],本节主要对基于深度学习的合成人脸方法进行总结并展开讨论.

完全人脸合成.Karras 等人[143]提出一种渐进式生长的 GAN,即 PGGAN.在训练过程中,通过对低分率的输入图像添加增长式的修正细节信息,以

① FakeApp. https://www.malavida.com/en/soft/fakeapp
② Faceswap-GAN. https://github.com/shaoanlu/faceswap-GAN
③ VidTIMIT Dataset. http://conradsanderson.id.au/vidtimit/
④ Faceswap. https://github.com/MarekKowalski/FaceSwap/
⑤ Deepfakes github. https://github.com/deepfakes/faceswap

获取更稳定、质量更高的合成人脸.基于图像风格迁移学习,作者还提出一种风格生成框架 StyleGAN[27],该方法可自动学习人脸的高级属性(如表情、身份等)并进行无监督的分割,并且生成图像还具备雀斑、头发等随机变化的特点.采用此方法合成的虚假人脸在分布质量评价指标上达到了当时的最优结果.

基于 GAN 的人脸合成研究多采用两者对抗的方式,Shen 等人[146]提出了一种三者对抗的 GAN,即 Faceid-GAN.作者通过引入分类器,对传统判别器和生成器之间的对抗进行扩展,确保生成的图像同时具有高质量并可以保留原始的身份信息.除此之外,该方法还可以生成多视角的合成人脸及表情.

为了使深度合成图像检测器失效,Neves 等人[145]提出一种消除深度合成人脸 GAN 印记的策略.作者利用 autoencoder 的方式对隐含的、具有辨识力的 GAN 印记特征进行擦除,并同时保持合成图像的原有整体和细节信息,从而导致检测器对该合成人脸失效.

缺失人脸合成.除了直接合成完全不存在的虚假人脸内容,对目标人脸缺失部分进行合成是深度合成人脸另的一个研究方向[144,149-151].Lu 等人[144]介绍了一种人脸属性引导的人脸生成.该方法以低像素的人脸为输入,并获取该人脸缺少部分的其他高清图像内容,基于此高清人脸内容,对原低分辨率人脸进行高清合成.作者提出属性引导及身份引导的条件式 CycleGAN,将高清人脸属性映射到低清的人脸上,实现人脸低分辨率部分的高清合成.

基于深度生成模型,Li 等人[147]提出一种人脸补全的算法,该算法不同于之前的采用搜索补丁进行合成的方法,它基于 autoencoder 网络直接对缺失的区域进行像素的生成,从而可以解决大面积人脸缺失情况下的补全合成.同样是基于 autoencoder 结构,Song 等人[148]提出一种几何感知的人脸补全网络 FCENet.作者首先通过面部几何预估器预测面部特征点的热力图并对其生成剖析映射图,之后基于生成网络对缺失部分进行合成.

针对深度合成人脸的效果评估,现有的研究并没统一的评价指标,采用的源数据集也不尽相同.多数文献中[144,146,147]采用非量化的可视化效果进行生成效果的对比评价.部分研究采用针对 GAN 的评价指标如 inception score 对合成人脸进行评价[143]或采用 PSNR/SSIM 对补全合成人脸进行评价[148].采用相同的源数据集及统一的针对性设计的量化评测指标,是深度合成人脸评估的重要研究发展方向.

除此之外,深度合成人脸生成技术同样存在稳定性不足的弱点,轻微的样本扰动或参数设定变化就可能导致合成的样本效果变差,同时该技术还存在易产生异常斑点、背景出错、人脸属性不对称等弱点,因此如何提升深度合成人脸技术的稳定性,并不断提升合成细节的真实性是未来研究的重点方向之一.

4.2　人脸深度篡改检测防御技术

4.2.1　人脸属性内容编辑检测

针对深度学习人脸属性内容编辑进行检测防御,现有的研究主要可分为结合传统机器学习的方法[137-142]如 SVM 等,以及基于深度学习的方法[86]如 RBM、CNN 等的检测.

结合传统机器学习方法.为了判断人脸属性内容是否经过 GAN 的编辑处理,Jain 等人[150]提出采用基于 CNN 的方法对人脸图像进行特征提取,并结合 SVM 等传统机器学习方式对特征进行分类,实现对基于 StarGAN 编辑的人脸属性内容的检测.面向多种不同 GAN 方法,如 Style-GAN、PGGAN 等实现的人脸属性编辑,Wang 等人[149]发现深度人脸识别系统中的每层神经元激活纹理可以被用来区分真实和虚假(经过编辑的)的人脸图像.基于此作者提出了 FakeSpotter,通过监控神经元覆盖行为进行深度特征提取,并采用 SVM 设计特征二分类器,实现虚假人脸属性的检测.

基于深度学习方法.文献[86]中,Stehouwer 等人提出一种利用深度学习注意力机制的处理方式,学习到的注意力图可有效地高亮出可能经过编辑的区域,从而有利于 CNN 分类器区分真实和虚假的人脸,并可对图像中经过编辑的区域可视化.Nataraj 等人[152]提出一种结合共生矩阵与深度学习的方法,实现 StarGAN 生成的人脸属性编辑检测.作者从图像像素的三通道中提取共生矩阵,并基于此训练深度神经网络判断人脸真伪.

除了直接对图像样本进行检测,Zhang 等人[153]提出一种基于图像频谱的分类器,对 GAN 生成图片时的图像上采样环节进行了分析研究,并使用一种信号处理分析的方式,实现对多种 GAN 如 CycleGAN 生成的人脸属性编辑图像的检测.

针对人脸属性编辑内容,虽然现有研究中的检测方法使用的数据集多采用统一的方法生成(如 StarGAN),但是其源数据多为作者自己整理,造成

生成的数据集各异，这导致难以对此类方法做公平的效果对比. 表 5 对比了主流的人脸属性编辑的检测防御方法. 多数的方法采用检测准确率(Acc)或 ROC 曲线下面积(AUC)作为评价指标.

现有的人脸属性内容编辑检测已经取得了一定的效果，然而这些方法没有对时间复杂度进行评测，在实际场景中的检测效率未知. 除此之外，多数的检测方法仅在一至两个数据集上进行评测，其算法的迁移性及泛化性也有待进一步评估.

表 5 主流的人脸属性编辑的检测算法对比 (单位：%)

文献	算法	数据集	准确率/ROC 曲线下面积
Jain[150]	CNN+SVM	StarGAN 数据集[47]，ND-IIITD[154]	99.7/—
Wang[149]	CNN+SVM	StyleGAN 数据集[27]，StarGAN 数据集	84.7/—
Stehouwer[86]	CNN+注意力机制	DFFD[86]	—/99.9
Zhang[153]	GAN+频谱	CycleGAN 数据集[136]	100/—
Nataraj[152]	CNN	StarGAN 数据集，CycleGAN 数据集	99.4/—
Marra[155]	CNN +增量学习	StarGAN 数据集	99.3/—

4.2.2 人脸面部内容重现检测

针对人脸面部内容重现的检测方法[86-89,156-161]未形成完整的技术体系，主流的方法多采用 CNN 设计二分类器，实现人脸面部内容重现的检测.

传统机器学习方法. Matern 等人[162]提出了一种简单有效的检测方法，通过对图像全局一致性及几何估计的建模，采用传统机器学习中的逻辑回归及多层感知机算法，实现面部内容重现的检测.

卷积神经网络方法. Afchar 等人[163]提出 MesoNet 方法构建深度卷积神经网络，并关注和学习图像的细微属性差异，以此实现 Face2Face 人脸面部内容重现的检测. Rossler 等人先后采集并公开两个针对人脸面部内容重构的数据集[89,157]，并提出基于 Xception[160]网络的真实人脸、重构人脸二分类器，实现面部内容重现的检测.

Sabir 等人[158]采用图像预处理的方式并结合循环卷积神经网络构造面部重现的检测模型. 较新的研究[87]中，作者提出一种多流的神经网络，对面部重现样本的不同区域进行分别学习，并将学习特征进行融合从而实现更精确的检测.

自编码器方法. 先前的研究多仅判断人脸是否经过修改，不同于此，Nguyen 等人[164]提出可同时检测重现人脸并定位面部经过修改的区域的方法. 作者采用 encoder-decoder 的方式实现虚假内容检测和虚假区域的定位分割. 文献[156]中，作者将伪造人脸检测视为单分类异常检测问题，提出了基于自编码器架构的 OC-FakeDect 网络，该网络的训练数据仅包含真实图像，由于真实图像的重建图像与输入图像的差异较小，所以通过设置重建分数阈值可以区分虚假的异常图像，基于此思想，作者提出了两种基于 OC-VAE 的方法来检测真实和虚假图像.

相较于其他虚假人脸篡改检测任务研究，人脸面部重现的检测研究有比较统一的评测数据集及评测指标，可较有效、公平地对主流的方法进行评估对比. 然而，现有的检测方法鲜有在检测性能方面进行评测，使得这些方法的实用性无法得到保证. 表 6 对比了主流的人脸面部内容重现的检测防御方法，多数的方法采用检测准确率(Acc)作为评价指标，表中所列的指标值为各文献所提出的方法在人脸面部重现数据上的最优结果，其中 FF++数据集中包含三种不同质量的视频，分别为 Raw(无损视频)、HQ(量化参数为 23 的 H.264 压缩高质量视频)和 LQ(量化参数为 40 的 H.264 压缩低质量视频).

表 6 主流的人脸面部重现的检测算法对比

文献	算法	数据集	准确率/%
Afchar[163]	CNN	FF++-F2F(Raw/HQ/LQ)[89]	96.8/93.4/81.3
Matern[162]	LR，MLP	FF++-F2F	AUC=86.6
Rossler[89]	CNN	FF++-F2F(Raw/HQ/LQ)	99.61/98.36/91.56
		FF++-NT(Raw/HQ/LQ)[89]	99.36/94.5/82.11
Nguyen[164]	autoencoder	FF-F2F[157]	93.63
Sabir[158]	CNN+RNN	FF++-F2F(LQ)	94.35
Amerini[159]	CNN+光流法	FF++-F2F	81.6
Kumar[87]	CNN+融合法	FF++-F2F(Raw/HQ/LQ)	99.96/99.10/91.20

4.2.3 深度伪造人脸交换检测

针对深度伪造人脸交换的检测，现有的研究多采用基于深度卷积神经网络的分类器，对交换前后的人脸视频或图像本身进行分析，通过人脸部分的特征差异或背景环境的变化，判断该人脸是否是深度伪造人脸[164,165]. 深度伪造人脸交换检测技术从数据模态上可分为两种：静态图像检测及动态视频检测.

静态图像检测. 在早期的伪造人脸检测研究中，Zhang 等人[166]提出一种基于传统机器学习方法的伪造人脸检测方法，作者采用 surf 特征[167]及 bag of words 算法[168]对人脸的静态图像(视频抽帧获得)进行特征提取分析，并采用 SVM 等作为分类器，实现伪造人脸检测. Yang 等人[169]利用人头的

姿态信息构建神经网络,对人脸头部的姿态方向进行建模和分析,并通过判断头部的姿态向量对真实和伪造的人脸进行区分.

由于深度伪造人脸交换的面部区域和周围背景之间的分辨率不一致,造成了小范围内的扭曲,而这种扭曲留下了明显的伪像.基于这些伪像,Li 等人[165]设计专用的卷积神经网络(CNN)模型,将生成的面部区域及其周围区域进行比较来检测此类伪影,从而实现深度伪造人脸的检测.采用此检测方法的优势是不需要生成大量的伪造换脸样本进行模型训练,降低了训练的时间复杂度.

基于 Siamese 网络架构,Hsu 等人[161]使用对比损失来检测深度伪造人脸,称为通用伪特征网络(CFFN).作者以真实图像、伪造图像对,作为网络输入提取伪造人脸的特征.然后通过将 CNN 连接到 CFFN 的最后一个卷积层,基于提取的深度特征完成对深度伪造图像的检测.

除了对人脸交换过程中的生成、篡改部分进行检测,研究[170]中提出一种针对人脸交换人脸融合过程的检测方法.作者提出深度伪造的人脸在融合过程中都会存在一个特殊的掩码(mask),其边缘信息即为人脸 x 射线(face x-ray),利用此特征可区分真实与伪造的人脸,其实验结果也证明该方法优于之前最好的检测方法.

结合注意力机制,Zhao 等人[171]将深度伪造人脸交换检测任务视为细粒度分类问题,提出了一种结合了多注意力机制的检测网络,该网络利用多头注意力机制使网络关注不同的局部区域,利用纹理特征增强模块突出浅层特征中的细微局部差异特征,利用双线性注意力池化操作聚合由注意力特征图引导的低级纹理特征和高级语义特征,此外,为了解决网络学习困难的问题,此方法提出了一种新的区域独立性损失和注意力引导的数据增强策略迫使网络学习到不同的注意力特征,经实验验证,该方法在多个基准数据集上均有优异的检测能力.

一些研究通过挖掘频域内的信息来进行检测[172-174],Qian 等人[174]利用了可分离的频率分量 FAD(Frequency-aware Image Decomposition)和局部频率统计信息 LFS(Local Frequency Statistics)去深入挖掘深度篡改人脸图像的异常模式,基于上述两种特征,作者设计了 Frequency in Face Forgery Network(F^3-Net),该网络为双分支网络,分别学习 FSD 和 LFS,然后再通过融合模块 MixBlock 来融合双路网络中的特征进而对图像进行分类.

动态视频检测.利用深度伪造人脸视频不同帧之间的差异,基于循环神经网络(recurrent neural network),文献[158,175]分别提出不同的深度伪造人脸交换检测的方法.作者[158]借鉴视频理解、行为识别任务中对多帧信息的处理方法,采用单向或双向的循环神经网络实现深度伪造人脸检测.文献[175]中,作者使用长短期记忆(LSTM)对连续 24 帧的图片进行分析及特征提取,从而判断输入为真实或伪造人脸.

结合多分支网络的思想,Wu 等人[176]提出了 SSTNet,该网络采用多分支网络的架构,在检测深度伪造人脸时融合了空间、隐写和时序特征,以连续多帧人脸图像作为网络输入,利用 XceptionNet[160]提取图像的空间特征,提出了一种带有限制的卷积过滤器,并结合截断的 XceptionNet 提取隐写特征,最后将空间特征和隐写特征融合,利用 LSTM 检测融合特征中存在的时序差异.同样地,Masi 等人[177]提出了一个双分支网络,一个分支利用传统的卷积网络提取 RGB 域图像特征,另一个分支利用多尺度高斯拉普拉斯算子核提取频域维度特征,其中高斯拉普拉斯算子核用来降低低层特征图中的高维度语义信息,同时,在训练阶段作者利用基于特征空间的损失函数来聚合真实类内样本,疏远相对于真实样本的虚假类间样本.

许多基于动态视频的伪造换脸检测研究是利用人脸上存在的生物特征信息而实现的[138,162,178].Li 等人[138]提出利用多帧间眨眼行为的连续性和一致来区分真实的人脸和深度伪造的人脸.作者构造由多个 LSTM 组成的深度神经网络 LRCN,实现深度伪造人脸的检测.Mittal 等人[178]提出一种基于 Siamese 网络架构的方法来检测深度伪造视频,通过观察情感与行为之间的联系,例如眼睛的扩张、眉毛的抬高、音量、步调和声音的音调进行检测.训练时通过输入真实视频和深度伪造视频获取人物的面部和语言的情态情感并用此计算三重损失函数,通过语音与面部模态进行检测,以及语音所表达出的情感与面部神态表达出的情感相似性进行检测.

表 7 和表 8 分别对近期和较早期[135]提出的深度伪造人脸交换检测方法在常用的深度伪造检测基准数据集上进行了评价对比,由于早期的数据集(如 UADFV、DF-TIMIT、DFD)中利用的操纵方法单一且生成的虚假数据质量较差,所以近期提出的检测方法鲜在这些数据集上进行验证,两个表中的多数方法采用 ROC 曲线下面积 AUC(%)作为评价指

标.由表 7 和表 8 可知,(多帧)动态视频的检测方法没有明显优于(单帧)静态图像的检测方法.

表 7 主流的深度伪造人脸交换检测算法对比

数据集 / 文献	FF++-All[157] Raw	FF++-All[157] HQ	FF++-All[157] LQ	Celeb-DF[135]	DFDC[140]	DeeperForensics-1.0-rand[50]	ForgeryNet[141]
Xception[89]	99.26	96.3	89.3	48.2	49.9	94.75	90.12
Two-branch[177]	—	96.43	86.34	73.41	—	—	—
Face X-ray[170]	98.52	87.35	61.60	74.76	80.92	—	—
F^3-Net[174]	99.8	98.1	93.3	65.17	—	—	90.15
FFD(Xception+Reg.)[86]	—	—	—	71.2	—	—	—
Patch[179]	99.77	97.2	78.3	84.88	65.6	81.8	—
Multi-attention[171]	99.80	99.29	90.40	67.44	recall=0.1, precision=0.92	—	—

表 8 较早期的主流深度伪造人脸交换检测算法对比

数据集 / 文献	UADFV[138]	DF-TIMIT[139] LQ	DF-TIMIT[139] HQ	FF++-FS	DFD①	DFDC	Celeb-DF
Two-stream[180]	85.1	83.5	73.5	70.1	52.8	61.4	53.8
HeadPose[169]	89.0	55.1	53.2	47.3	56.1	55.9	54.6
FWA-Res50[165]	97.4	99.9	93.2	80.1	74.3	72.7	56.9
VA-MLP[162]	70.2	61.4	62.1	66.4	69.1	61.9	55.0
Xception-Raw[89]	80.4	56.7	54.0	99.7	53.9	49.9	48.2
Multi-task[164]	65.8	62.2	55.3	76.3	54.1	53.6	54.3
Capsule[181]	61.3	78.4	74.4	96.6	64.0	53.3	57.5
DSP-FWA[135]	97.7	97.7	99.7	93.0	81.1	75.5	64.6

虽然关于深度伪造人脸交换检测的研究已有一定的成果并形成初步的体系,但现有的研究中还存在诸多问题:(1)现有的研究缺乏对检测性能的评测,难以保证各类算法在实际应用中的效果;(2)现有的检测方法普遍存在泛化性较差的问题,这导致相关研究中发表的检测精度指标值在数据质量较差的现实场景中是无法达到的,如表 7 和表 8 所示,各检测方法在跨数据集(跨操纵方法)的条件下进行评估时检测性能均有下降,并且在多个困难数据集上,如 DFDC、Celeb-DF 等,现有的检测算法效果普遍较差.另外,如表 7 所示,在带有压缩、扰动噪声的困难数据上,现有的检测算法精度较低且泛化性较差.因此针对不同的数据集需重新训练生成新的模型才能获得较好的检测效果;(3)现有的评测方法也不够公平,不同的检测模型多使用不同的数据进行训练,缺乏统一的训练集;(4)评测指标较单一,缺乏检测性能、模型复杂度、泛化能力等评价指标,无法全面、公平地评价不同深度伪造人脸交换的检测方法.这些问题与挑战也是深度伪造换脸检测技术未来发展的关键.

4.2.4 深度合成人脸检测

为了实现深度合成人脸的检测,早期的研究曾利用传统机器学方法进行检测,目前多数研究采用深度卷积神经网络直接对图像进行分类,还有一部分研究是通过检测 GAN 合成图像留下的指纹而实现的[86,88,149,182-184].

传统机器学方法.在早期的研究中,Li 等人[185]利用相机成像和深度网络生成之间的差异,分析了生成图像与真实图像在不同的色度通道之间的区别,他们发现生成图像和真实图像在 HSV 和 YCbCr 颜色空间中的色度分量统计特征是不一致的,而且这种不一致性在残差图像中更显著.基于此图像统计特征,作者提出了一个包含不同颜色分量条件下残差图像共生矩阵的特征集用来识别生成图像,这个低维特征集在仅有少量训练集数据的条件下也可以有较好的检测性能.

卷积神经网络.针对采用 PGGAN[143]生成的深度合成人脸,McCloskey 等人[182]提出使用卷积网络对图像的颜色线索进行特征提取,即生成的图像颜色和照相机拍的图像颜色有明显差异,并结合 SVM 实现深度合成的人脸的检测.但是该方法的整体检测精度还有待提高.同样是对 PGGAN 生成的深度合成人脸进行检测,Tariq 等人[183]提出一种基于神经网络的分类器,配合数据预处理及增强的方法,取得了效果更好的深度合成人脸检测.

在最新的研究中,Liu 等人[184]提出一种全局纹理增强的方法实现对利用 StyleGAN 和 PGGAN 生成的深度合成人脸的检测.作者利用 CNN 对全局纹理特征进行提取并统计分析,发现此信息对于不同方法及不同数据集合成的人脸都更加鲁棒,有较好的检测效果.Chai 等人[179]提出了具有有限感受野的基于 patch 的检测器,该方法将 Resnet 和 Xception 模型从中间层截断,利用此截断模型作为检测分类器,从而基于不同尺度的中间层特征图对输入图像进行分类,其中 Patch Resnet Layer1 和 Patch Xception Block2 这两个分类器有较好的表

① Google AI Blog. Contributing Data to Deepfake Detection Research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

现.除此之外,文献[86,149]中提出的用于人脸属性内容编辑的 CNN 方法也同样适用于深度合成人脸图像的检测.

GAN 图像指纹.另一类辨别深度合成人脸的方法是通过检测基于 GAN 合成图像的图像指纹来实现的[88,186].作者[186]利用一个去噪滤波器,对原始图片及深度合成图片进行过滤,并对过滤后的图像进行对比,找出 GAN 合成图像留下的图像指纹,从而实现深度合成图像的检测.基于此思想,Yu 等人[88]提出一种基于 autoencoder 的 GAN 指纹检测机制,通过此机制证明不同的 GAN 携带有独特的模型指纹,并且该指纹同样保留在其生成的图片中.基于此特征,实现了更精确的深度合成人脸检测.

表 9　主流的深度合成人脸检测算法对比　(单位:%)

文献	算法	数据集	准确率 ROC 曲线下的面积
McCloskey[182]	色彩+SVM	NIST(PGGAN)	−/70
Marra[186]	图像处理	CelebA[85](PGGAN)	92.28/−
Yu[88]	CNN	CelebA (PGGAN)	99.5/−
Stehouwer[86]	CNN+注意力机制	DFFD(PGGAN)	−/100
Wang[149]	SVM	Celeb-DF(StyleGAN)	84.7/−
Liu[184]	纹理特征 CNN	CelebA (PGGAN)	98.78/−
Chai[179]	CNN	CelebA(PGGAN)/ CelebA(GMM[187])/ CelebA(StyleGAN)	AP=100.0 /80.69 /97.22

表 9 对比了主流的深度合成人脸的检测防御方法,多数的方法采用检测准确率(Acc)或 ROC 曲线下面积 AUC 作为评价指标,Chai[179]提出的方法利用平均查准率(Average Precision)作为评价指标,通过指标值可以看出利用 PGGAN 数据集训练出的模型在 GMM[187]操纵数据和 StyleGAN 操纵数据上进行测试评价时均存在由于模型泛化性较差导致的性能下降的问题.

4.3　人脸深度篡改检测防御技术对比

为应对不同类型的人脸深度篡改生成技术,研究者们提出了多种检测防御技术,其中不乏一些方法能够有效地检测多种类型的深度篡改伪造数字内容[86,165,170,188],这是由于不同类型的人脸深度篡改生成技术生成网络以及生成流程上具有共通之处.

通过探究人脸深度篡改生成网络的相同点,文献[188]中提出由 CNN 生成的图片具有共同的异常特征,基于此作者仅利用了由一种特定的 CNN 生成器(ProGAN)生成的图像训练出了一个泛化性极强的检测分类器,该分类器为 ResNet-50 网络,在训练过程中,作者对数据施加了数据预处理、数据后处理及图像增强的策略,经过实验证实,该分类器具有较强的检测泛化性,在包含由 11 种不同的 CNN 生成器生成的图像的数据集中表现良好,可以检测深度合成人脸数据和人脸交换数据,另外作者进一步探究了不同的图像后处理、数据增强策略、训练数据多样性对模型泛化性能的影响,发现了利用数据增强技术和增加训练数据多样性有助于提升检测模型泛化性,且模型在经过不同的后处理的数据上表现有较大差异. Durall 等人[189]提出由深度卷积生成对抗网络生成的图像无法重现真实样本中存在的光谱分布,这是由基于卷积的上采样过程造成的,这一现象广泛存在于包含此类反卷积操作的生成网络中,利用这一现象,作者提出了一个简单但准确率极高的检测器用来检测深度合成人脸数据、人脸交换数据和人脸面部内容重现数据.

基于人脸深度篡改生成流程的相同点,Li 等人[170]关注生成流程中面部融合步骤带来的异常特征,提出利用融合边界作为真假图像的差异特征进行检测.Li 等人[165]同样利用面部融合步骤产生的人脸前景和场景背景的分辨率不一致的特点作为异常特征对真假图像进行检测.

除了人脸深度篡改检测防御技术的相同点之外,在面对不同类型的深度篡改数据时,检测方法也有不同点.这主要表现在如[165,170]中的方法是基于生成流程中的面部融合步骤产生的异常特征设计的,因此生成流程中不包含面部融合步骤的深度篡改类型的数据,如深度合成人脸虚假数据,则无法利用这种方法进行检测.

4.4　对抗样本与人脸深度篡改防御技术

4.4.1　对抗样本攻击人脸深度篡改检测器

目前存在大量关于人脸深度篡改防御的研究,且这些研究在主流的基准测评数据集上都获得了很高的检测准确率,但随后的研究很快发现这些检测器很容易因为被对对抗样本攻击而失效[136,190,191].文献[191]中提出一种对人脸深度篡改视频添加扰动的方法,作者通过对虚假视频中的每一帧图像添加扰动从而得到一个被对抗样本更改过的虚假视频,使得检测器将其误分类为真实视频,该方法有较好的实用性和健壮性,可以在黑盒和白盒场景下应用,且对真实场景中经过图像或视频压缩的虚假数据也是有效的. Carlini 等人[190]分别利用白盒攻击和黑盒攻击评估了 3 种人脸深度篡改检测器的稳健性,结果显示一个在常规情景下 AUC 为 0.95 的检测器在受到 4 种白盒攻击方法的攻击后,AUC 最

低被降至了 0.085，并且在具有限制条件的黑盒攻击情景下，AUC 值也降至了 0.22 以下，这说明现存的深度篡改检测器仍有局限性，需要深入研究. Gandhi 等人[136]利用了通用的对抗扰动来欺骗检测器并提出了相应的提升深度篡改样本健壮性的方法，作者在黑盒和白盒条件下利用 FGSM(Fast Gradient Sign Method)方法和 Carlini and Wagner L2 Norm Attack(CW-L2)方法创建了对抗样本，使得原本检测准确率为 95%的检测器在检测带有扰动的深度篡改样本时准确率仅为 27%，同时作者提出了两种提升检测器性能的方法，分别为 Lipschitz 正则化和 Deep Image Prior(DIP)，Lipschitz 正则化让检测器在受到黑盒攻击的条件下增加了 10%的准确率，DIP 防御方法让检测器面对带有攻击的虚假样本时可以达到 95%的检测准确率. 针对带有对抗扰动的深度篡改人脸图像，Chen 等人[192]认为通过对抗样本使检测器失效的目的并不能达成，从而提出了 MagDR 网络，该网络可以检测输入检测器的图像是否为对抗样本，若是对抗样本，则去掉对抗样本的扰动并对图像进行重建得到未被扰动的深度篡改图像.

4.4.2 对抗样本攻击人脸深度篡改生成器

面对人脸深度篡改的威胁，除了研发防御检测器以外，研究者们还提出可以通过对未篡改前的真实源图像添加对抗扰动来扰乱深度篡改生成器，以致毁坏其输出来防止虚假数据产生带来的威胁[193,194].

为了防止恶意用户在未经源图像涉及人物同意的情况下生成一个人的修改图像，Ruiz 等人[193]通过利用对抗扰动干扰生成器，破坏生成的输出图像来进行防御，他们首次提出了可以转换操纵类别的对抗攻击，并提出了运用对抗训练作为训练生成器的第一步以增强生成器的健壮性，并且他们发现在灰盒场景中，模糊操作可以成功抵御对抗攻击，进而提出了一种可以躲避模糊防御的扩频对抗攻击. Yeh 等人[194]针对人脸深度篡改生成器提出了两种对抗攻击的方法，分别是利用 Nullifying 攻击生成与源图像相似的未更改图像和利用 Distorting 攻击生成被毁坏的图像. Li 等人[195]基于对抗样本提出一种主动保护的机制，通过对人脸图像加入对抗性扰动信息，从而降低作为深度伪造面部合成训练数据的质量，进而使得采用深度伪造技术生成的人脸出现明显的错误，实现主动防御.

4.5 数据集介绍

人脸属性内容编辑相关的研究中，许多文献中[86,149,152]的虚假人脸属性数据是基于 CelebA[85]生成的，但是总体而言，由于研究者可以很容易地利用大量公开的基于 GAN 的代码实现人脸属性内容的编辑，目前的研究还缺乏统一的基准数据集，这不利于不同方法的公平评测对比.

人脸面部内容重现相关的研究有两个较为通用的数据集[89,157]，其中[157]公开了一个包含 1004 个视频的数据集，该数据集由两部分组成：1 是源到目标的 Face2Face 面部内容重现数据集；2 是自面部内容重现(源和目标数据同源)数据集. 文献[89]公开了一个来自 1000 个视频超过 1.8 万张图片的真实和面部内容重现数据集，该数据集是目前已知的最大规模源到目标的人脸面部内容重现数据集. 采用统一的基准数据集进行训练及测试也有助于人脸面部内容重现的研究逐步形成体系.

深度伪造人脸交换相关的研究中，已经有多个公开的数据集常被用于深度伪造人脸交换检测的训练及测试. 如表 10 中提到的多个数据集，UADFV、DF-TIMIT、DFD、FF ＋ ＋、DFDC、Celeb-DF-v2、DeeperForensics-1.0、ForgeryNet 为常用的深度伪造人脸交换检测基准数据集，表中所列数据均为仅与深度伪造人脸交换操纵相关的数据. 其中 ForgeryNet[141]为目前最大的深度伪造人脸交换数据集，包含 2,900,000 张图像和 221,247 段视频，其中的虚假数据由 15 种深度伪造篡改生成方法生成，数据中共带有 36 种类型的扰动，提供图像级和视频级两种级别的标注，可以利用其开展 4 类深度篡改检测任务的学习，分别为伪造图像分类、伪造人脸区域定位、伪造视频分类和视频伪造片段定位，除此之外，作者利用该数据集对以上四种任务建立了一个比较全面的基准评估.

但目前的深度伪造人脸交换数据集多采用不同的数据源、不同的伪造生成方法、不同的视频/图像分辨率伪造生成，缺乏来自同源的整合数据集，造成在不同数据集上进行深度伪造检测模型的训练及测试的对比结果不置信，不利于深度伪造人脸交换及检测技术的长期发展.

深度合成人脸相关的研究中，有多个公开的通用数据集常被用于深度合成人脸的训练生成和检测[196]. 如 FFHQ[27]、CelebA[85]数据集被文献[149]用于合成人脸；CASIA-WebFace[151]和 VGGFace2 [197]被文献[145]用于合成人脸.

表 10　主流的深度伪造人脸交换数据集对比

数据集	人脸交换算法	真实视频数量	虚假视频数量	扰动类型	隐藏集
UADFV	FakeApp	49	49	—	—
DF-TIMIT	Faceswap-GAN	320	640	—	—
DFD	未知	363	3,068	—	—
Celeb-DF	未知	590	5,639	—	—
FF++	DeepFakes, FaceSwap, FaceShifter[133]	1,000	3,000	2	—
DF−1.0[50]	DF-VAE[50]	50,000	10,000	7	1
DFDC	DFAE[140], MM/NN[140], NTH[119], FSGAN[132], StyleGAN[27]	23,564	104,500	19	1
ForgeryNet	FSGAN[132], DeepFakes[198], FaceShifter[133]	99,630	121,617	36	1

除此之外,文献①中包含了 10 万张深度合成的人脸图像,这些图像由约 29000 张人脸及 69 个不同的模型基于 Style-GAN 的方法训练而成. 文献[86]中公开了一个大型深度合成人脸数据集,它包含了 20 万张由 PGGAN 和 10 万张由 StyleGAN 生成的深度合成人脸图片. 如何将现有的多个数据集进行整合,建立统一的基准数据集,有利于深度合成人脸检测技术的进一步发展.

5　研究难点与未来挑战

近年来,针对虚假数字人脸内容生成与检测的研究已取得了一定的成果,然而该研究整体上仍处于初级阶段,其生成、检测算法的精确度、性能、泛化能力等方面有许多关键问题尚待解决,尚未形成完整的技术体系,并且其数据集、评测指标多样且不统一,鲜有工作对现有的研究进行全面且公平的效果对比. 与此同时,以 GAN 和 autoencoder 为代表的深度学习的生成技术仍在不断的发展中,这使得无论是人脸对抗样本技术还是人脸深度篡改技术都还有广阔的发展空间,这也给虚假数字人脸内容生成与检测带来了持续的挑战.

5.1　人脸对抗样本技术

从算法模型层面上看,虽然现有的研究在人脸识别、人脸检测对抗攻击上取得了一定的成果,但现有的攻击方法多数对白盒攻击有一定效果,而对现实场景中更常见的黑盒模型、系统攻击效果较差,造成对抗攻击算法的实用性不强. 因此,如何设计更有效的黑盒攻击算法及模型是该领域未来的研究重点.

从结果及评价体系上看,现有的人脸识别、人脸检测对抗攻击方法各异,缺乏统一的评测数据集及通用的评价指标,这使得难以对现有的方法进行公平、合理的对比评价. 针对人脸识别对抗攻击的检测防御虽然取得了一定的研究成果,但是现有的方法都缺乏对检测性能的评测,这使得检测方法的实用性无法得到全面的评估. 同时,现有的检测策略多采用不同的数据集及不同的评价标准进行效果评估,导致难以对现有的方法做公平的效果对比. 因此,建立公平的评测数据和评测指标,是该领域未来研究的重点方向之一.

从技术完整性层面上看,关于人脸对抗样本技术的研究主要集中于人脸识别对抗攻击及对抗检测防御,而针对完整人脸识别、监控系统中的重要技术点和拓展技术点包括人脸检测、人脸活体检测判断、人脸属性识别等相关的对抗攻击与检测防御的研究目前还较少,这也使得现有的人脸识别系统在许多环节上都容易受到对抗样本的攻击,其安全性面临着较大的威胁和挑战. 因此,除了针对人脸识别对抗攻击与检测防御技术的探索研究,未来的工作应在人脸识别系统其他技术环节上的攻击与检测中投入更多的研究,以此保障人脸识别系统全生命周期的安全鲁棒.

目前对抗攻击技术在人脸检测、识别等任务中有着较好的表现,但是物理实现的对抗攻击易受到真实场景中光线、角度变化以及场景噪声等多种不稳定不可预知的因素的影响,这会显著降低对抗攻击的成功率. 此外,现有的物理实现的对抗攻击隐蔽性普遍较差,很容易被人类感知或机器学习算法检测识别. 如何实现更稳定、更隐蔽的物理域人脸检测、识别对抗攻击将是未来的研究重点之一. 同时,目前对抗训练被认为是最有效的防御方法之一,但其需要大量的对抗样本进行训练,时间复杂度较高,并且一定程度上会造成原始识别能力下降,除此之外,攻击者可轻易针对对抗训练后的模型设计新的对抗攻击算法,从而造成识别系统再次出错. 因此,如何在不影响原始识别能力的情况下设计通用高效的对抗防御方法已成为对抗样本防御的研究方向之一.

5.2　人脸深度篡改生成和检测技术

近几年关于人脸深度篡改虚假内容生成的研究

① 100,000 Faces Generated by AI. https://generated.photos/

工作不断涌现,无论是人脸属性内容编辑、面部重现还是人脸交换及深度合成人脸生成方面,现有的研究实现了逼真精细的人脸篡改.

从算法模型层面上看,现有的篡改、伪造方法多基于 GAN、autoencoder 等生成方法进行优化改进,这些方法普遍存在生成效果不稳定、需要大量训练样本、训练时间长等问题;并且现有的方法很难实现高分辨率的人脸深度篡改,在高分辨率情况下,人眼仍可以较容易地分辨真实和伪造的人脸样本.如何实现更稳定、更高清、更高效的人脸深度篡改是未来研究的重点之一.

从结果及评价体系上看,现有的检测方法普遍鲁棒性和泛化性较差,面对带有压缩、扰动噪声的困难样本或跨操纵方法的样本时,这些检测算法容易失效.虽然现有的研究有较为统一的数据集对算法进行评测,但是评测指标单一,无法全面地描述各个算法在不同维度上的优劣.因此,如何实现高效、泛化的人脸深度篡改检测防御,及制定更全面的评测体系是该领域未来的研究重点.

从技术完整性层面上看,针对深度篡改、伪造人脸进行的检测,现有的研究多集中在如何提高检测的精度,而鲜有研究考虑检测的性能,这使得真实世界中,系统面对大规模虚假数字人脸内容攻击时,许多低效的检测算法难以有效应用.因此,如何公平评估现有的虚假数字人脸内容的检测方法的实用性,并结合如模型所需计算力、模型参数量、模型推理时间等设计模型实用性的评价指标及完整评估体系,以及提出高效的虚假数字人脸内容检测算法是该领域未来的研究重点.

6 结束语

近年来,由于人工智能深度学习技术的发展与成功,基于此技术的智能系统尤其是围绕人脸研发的智能识别、智能监控等系统广泛地应用于人类的日常生活中.与此同时,针对人工智能系统的攻击及利用人工智能技术伪造虚假数字内容的方法也不断涌现,这给个人隐私安全、社会安全乃至国家安全都带来了巨大的威胁和挑战.

本文面向人工智能技术安全问题,针对国内外研究人员在虚假数字人脸内容生成与检测的研究工作展开调研和分析,并从人脸对抗样本、人脸深度篡改两个方面对现有的研究进行了归纳总结.本文整理总结了虚假数字人脸内容从生成攻击到检测防御的全流程相关研究及主流数据集,并讨论了现有研究中存在的问题、不足及面临的挑战.本文进一步讨论指出了该领域潜在的研究方向和研究重点,旨在为推动虚假数字人脸内容生成与检测技术的进一步发展,以及为保障人工智能人脸相关技术的安全应用提供指导和参考.

参 考 文 献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[2] Goodfellow I, Bengio Y, Courville. A. Deep learning. Cambridge, USA: MIT Press, 2016

[3] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2015, 61: 85-117

[4] Nilsson NJ. Principles of artificial intelligence. Berlin, Germany: Springer Science & Business Media, 1982

[5] Mitchell R, Michalski J, Carbonell T. An artificial intelligence approach. Berlin, Germany: Springer, 2013

[6] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012:1097-1105

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[8] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[9] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[10] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6):82-97

[11] Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks//Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649

[12] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks//Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 1764-1772

[13] Sak H, Senior A, Rao K, Beaufays F. Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947. 2015

[14] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12 (ARTICLE): 2493-2537

[15] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2013: 3111-3119

[16] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[17] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 3104-3112

[18] Ahmed E, Jones M, Marks TK. An improved deep learning architecture for person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3908-3916

[19] Grigorescu S, Trasnea B, Cocias T, Macesanu G. A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 2020, 37(3): 362-386

[20] Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ. Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics, 2016, 21(1): 4-21

[21] Taigman Y, Yang M, Ranzato MA, et al. Deepface: Closing the gap to human-level performance in face verification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1701-1708

[22] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 815-823

[23] Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition//Proceedings of the British Machine Vision Conference. Swansea, UK, 2015: 41.1-41.12

[24] Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 499-515

[25] Lu C, Tang X. Surpassing human-level face verification performance on LFW with GaussianFace//Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2307-2319

[26] Yang S, Luo P, Loy CC, Tang X. Wider face: A face detection benchmark//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5525-5533

[27] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4401-4410

[28] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[29] Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013

[30] Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396, 2016

[31] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013

[32] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014

[33] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016

[34] Bose AJ, Aarabi P. Adversarial attacks on face detectors using neural net based constrained optimization//Proceedings of the 2018 IEEE 20th International Workshop on Multimedia Signal Processing. Vancouver, Canada, 2018: 1-6

[35] Agarwal A, Singh R, Vatsa M, Ratha N. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? //Proceedings of the 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems. Los Angeles, USA, 2018: 1-7

[36] Goswami G, Ratha N, Agarwal A, Singh R, Vatsa M. Unravelling robustness of deep learning based face recognition against adversarial attacks//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 6829-6836

[37] Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. Journal of Software, 2020, 31(1): 67-81(in Chinese)
(潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述. 软件学报, 2020, 31(1): 67-81)

[38] Chen YF, Shen C, Wang Q, et al. Security and privacy risks in artificial intelligence systems. Journal of Computer Research and Development, 2019, 56(10): 2135-2150 (in Chinese)
(陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险. 计算机研究与发展, 2019, 56(10): 2135-2150)

[39] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 2018, 6: 14410-14430

[40] Li XR, Ji SL, Wu CM, et al. Survey on deepfakes and detection techniques. Journal of Software, 2020, 32(2): 496-518

(in Chinese)
(李旭嵘，纪守领，吴春明，等.深度伪造与检测技术综述.软件学报，2020，32(2)：496-518)

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

[42] Zheng X，Guo Y，Huang H，et al. A survey of deep facial attribute analysis. International Journal of Computer Vision，2020，128(8)：1-33

[43] Thies J，Zollhöfer M，Nießner M，et al. Real-time expression transfer for facial reenactment. ACM Transactions on Graphics，2015，34(6)：183:1-183:14

[44] Sharif M，Bhagavatula S，Bauer L，et al. Adversarial generative nets：Neural network attacks on state-of-the-art face recognition. arXiv preprint arXiv：1801.00349，2017：1556-6013

[45] Rozsa A，Günther M，Rudd E M，et al. Facial attributes：Accuracy and adversarial robustness. Pattern Recognition Letters，2019，124：100-108

[46] S Sharif M，Bhagavatula S，Bauer L，et al. Accessorize to a crime：Real and stealthy attacks on state-of-the-art face recognition//Proceedings of the 2016 ACMSigsac Conference on Computer and Communications Security. Vienna，Austria，2016：1528-1540

[47] Choi Y，Choi M，Kim M，et al. Stargan：Unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA，2018：8789-8797

[48] Thies J，Zollhofer M，Stamminger M，et al. Face2face：Real-time face capture and reenactment of rgb videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA，2016：2387-2395

[49] Suwajanakorn S，Seitz S M，Kemelmacher-Shlizerman I. Synthesizing obama：Learning lip sync from audio. ACM Transactions on Graphics，2017，36(4)：1-13

[50] Jiang L，Wu W，Li R，et al. DeeperForensics-1.0：A large-scale dataset for real-world face forgery detection. arXiv preprint arXiv:2001.03024，2020

[51] Komkov S，Petiushko A. AdvHat：Real-world adversarial attack on arcface face id system. arXiv preprint arXiv：1908.08705，2019

[52] Pautov M，Melnikov G，Kaziakhmedov E，et al. On adversarial patches：Real-world attack on arcface-100 face recognition system//Proceedings of the 2019 International Multi-Conference on Engineering，Computer and Information Sciences. Ural Hi-Tech Park，Russia，2019：0391-0396

[53] Zhu ZA，Lu YZ，Chiang CK. Generating adversarial examples by makeup attacks on face recognition//Proceedings of the 2019 IEEE International Conference on Image Processing. Taipei，China，2019：2516-2520

[54] Dong Y，Su H，Wu B，et al. Efficient decision-based black-box adversarial attacks on face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach，USA，2019：7714-7722

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

[56] Dabouei A，Soleymani S，Dawson J，et al. Fast geometrically-perturbed adversarial faces//Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision. Hawaii，USA，2019：1979-1988

[57] Deb D，Zhang J，Jain A K. Advfaces：Adversarial face synthesis. arXiv preprint arXiv:1908.05008，2019

[58] Eberhart RC ，Kennedy J. A new optimizer using particle swarm theory//Proceedings of the Sixth International Symposium on Micro Machine and Human Science. Nagoya，Japan，1995：39-43

[59] Kumar N，Berg AC，Belhumeur PN，et al. Attribute and simile classifiers for face verification//Proceedings of the Computer Vision，2009 IEEE 12th International Conference. Kyoto，Japan，2009：365-372

[60] Deng J ，Guo J ，Xue N ，et al. ArcFace：Additive angular margin loss for deep face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach，USA，2019：4690-4699

[61] Huang GB，Mattar M，Berg T，Learned-Miller E. Labeled faces in the wild：A database forstudying face recognition in unconstrained environments//Proceedings of the Workshop on Faces in'Real-Life'Images：Detection，Alignment，and Recognition. Marseille，France，2008：1-11

[62] Lu J，Sibai H，Fabry E. Adversarial examples that fool detectors. arXiv preprint arXiv:1712.02494，2014

[63] Yang X，Wei F，Zhang H，et al. Design and interpretation of universal adversarial patches in face detection. arXiv preprint arXiv:1912.05021，2019

[64] Kaziakhmedov E，Kireev K，Melnikov G，et al. Real-world attack on MTCNN face detection system//Proceedings of the 2019 International Multi-Conference on Engineering，Computer and Information Sciences. Ural Hi-Tech Park，Russia，2019：0422-0427

[65] Viola P，Jones M. Rapid object detection using a boosted cascade of simple features//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai，USA，2001(1)：I

[66] Zhang K，Zhang Z，Li Z，Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters，2016，10(23):1499-1503

[67] Ren S，He K，Girshick R，Sun J. Faster r-cnn：Towards real-time object detection with region proposal networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal，Canada，2015:91-99

[68] Redmon J，Farhadi A. Yolo9000：better，faster，stronger. arXiv preprint arXiv:1612.08242，2016

[69] Dziugaite GK, Ghahramani Z, Roy DM. A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853, 2016

[70] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017:2117-2125

[71] Mirjalili V, Ross A. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender//Proceedings of the 2017 IEEE International Joint Conference on Biometrics. Denver, USA, 2017:564-573

[72] Joshi A, Mukherjee A, Sarkar S, et al. Semantic adversarial attacks: Parametric transformations that fool deep classifiers//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 4773-4783

[73] He Z, Zuo W, Kan M, Shan S, Chen X. Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 2019, 28(11): 5464-5478

[74] Goel A, Singh A, Agarwal A, et al. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition//Proceedings of the 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems. Los Angeles, USA, 2018: 1-7

[75] Su Y, Sun G, Fan W, et al. Cleaning adversarial perturbations via residual generative network for face verification//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, 2019: 2597-2601

[76] Massoli FV, Carrara F, Amato G, et al. Detection of face recognition adversarial attacks. arXiv preprint arXiv: 1912.02918, 2019

[77] Tao G, Ma S, Liu Y, et al. Attacks meet interpretability: Attribute-steered detection of adversarial samples//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 7717-7728

[78] Cortes C, Vapnik V. Support vector machine. Machine Learning, 1995, 20(3):273-297

[79] Goswami G, Agarwal A, Ratha N, et al. Detecting and mitigating adversarial perturbations for robust face recognition. International Journal of Computer Vision, 2019, 127(6-7): 719-742

[80] Guo C, Rana M, Cisse M, Van Der Maaten L. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117, 2017

[81] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2016: 582-597

[82] Samangouei P, Kabkab M, Chellappa R. Defense-gan: Protecting classifiers against adversarial attacks using generative models//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-17

[83] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1891-1898

[84] Vakhshiteh F, Nickabadi A, Ramachandra R. Adversarial attacks against face recognition: A comprehensive study. IEEE Access, 2021, 9: 92735-92756

[85] Liu ZW, Luo P, Wang XG, Tang XO. Deep learning face attributes in the wild//Proceedings of the IEEE International Conference on Computer Vision. Las Condes, Chile, 2015: 3730-3738

[86] Stehouwer J, Dang H, Liu F, et al. On the detection of digital face manipulation. arXiv preprint arXiv: 1910.01717, 2019

[87] Kumar P, Vatsa M, Singh R. Detecting face2face facial reenactment in videos//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Colorado, USA, 2020: 2589-2597

[88] Yu N, Davis L S, Fritz M. Attributing fake images to gans: Learning and analyzing gan fingerprints//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 7556-7566

[89] Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: Learning to detect manipulated facial images//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 1-11

[90] Zhou H, Liu Y, Liu Z, et al. Talking face generation by adversarially disentangled audio-visual representation//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 9299-9306

[91] Chen L, Maddox R K, Duan Z, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7832-7841

[92] Yi R, Ye Z, Zhang J, et al. Audio-driven talking face video generation with natural head pose. arXiv preprint arXiv: 2002.10137, 2020

[93] Chung JS, Jamaludin A, Zisserman A. You said that? arXiv preprint arXiv:1705.02966, 2017

[94] Wiles O, Koepke A, Zisserman A. X2face: A network for controlling face generation using images, audio, and pose codes//Proceedings of the 15th European Conference. Munich, Germany, 2018:690-706

[95] Chung JS, Zisserman A. Zisserman. Lip reading in the wild//Proceedings of the 13t Asian Conference on Computer Vision. Taipei, China, 2016:87-103

[96] Badrinarayanan V, Kendall A, SegNet RC. A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561, 2015

[97] Guo Y, Jiao L, Wang S, et al. Fuzzy sparse autoencoder framework for single image per person face recognition. IEEE Transactions on Cybernetics, 2017, 48(8): 2402-2415

[98] Yang W, Hui C, Chen Z, et al. FV-GAN: Finger vein representation using generative adversarial networks. IEEE Transactions on Information Forensics and Security, 2019, 14(9): 2512-2524

[99] Liu F, Jiao L, Tang X. Task-oriented GAN for PolSAR image classification and clustering. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2707-2719

[100] Korshunova I, Shi W, Dambre J, et al. Fast face-swap using convolutional neural networks//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 3677-3685

[101] Ding H, Sricharan K, Chellappa R. Exprgan: Facial expression editing with controllable expression intensity//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 6781-6788

[102] Pumarola A, Agudo A, Martinez A M, et al. Ganimation: Anatomically-aware facial animation from a single image//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 818-833

[103] Perarnau G, Van De Weijer J, Raducanu B, et al. Invertible conditional gans for image editing. arXiv preprint arXiv: 1611.06355, 2016

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

[105] Zhang G, Kan M, Shan S, et al. Generative adversarial network with spatial attention for face attribute editing//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 417-432

[106] Lee CH, Liu Z, Wu L, Luo P, Maskgan: Towards diverse and interactive facial image manipulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5548-5557

[107] Shen Y, Gu J, Tang X, et al. Interpreting the latent space of gans for semantic face editing. arXiv preprint arXiv: 1907.10786, 2019

[108] Jo Y, Park J. SC-FEGAN: Face editing generative adversarial network with user's sketch and color//Proceedings of the IEEE International Conference on Computer Vision. Long Beach, USA, 2019: 1745-1753

[109] Lample G, Zeghidour N, Usunier N, et al. Fader networks: Manipulating images by sliding attributes//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA,2017: 5967-5976

[110] Chen YC, Shen X, Lin Z, et al. Semantic component decomposition for face attribute manipulation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9859-9867

[111] Liu M, Ding Y, Xia M, et al. STGAN: A unified selective transfer network for arbitrary image attribute editing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3673-3682

[112] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 6626-6637

[113] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics, 2019, 38(4): 1-12

[114] Kim H, Garrido P, Tewari A, et al. Deep video portraits. ACM Transactions on Graphics, 2018, 37(4): 1-14

[115] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1125-1134

[116] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2223-2232

[117] Wu W, Zhang Y, Li C, et al. Reenactgan: Learning to reenact faces via boundary transfer//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 603-619

[118] Ha S, Kersner M, Kim B, et al. Marionette: Few-shot face reenactment preserving identity of unseen targets//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA. 2020, 34(07): 10893-10900

[119] Zakharov E, Shysheya A, Burkov E, et al. Few-shot adversarial learning of realistic neural talking head models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 9459-9468

[120] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of the IEEE/CVF International Conference on Computer Vision. Venice, Italy, 2017:1501-1510

[121] Burkov E, Pasechnik I, Grigorev A, et al. Neural head reenactment with latent pose descriptors//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13786-13795

[122] Park T, Liu M Y, Wang T C, et al. Semantic image synthesis with spatially-adaptive normalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Bench, USA, 2019: 2337-2346

[123] Hao H, Baireddy S, Reibman A R, et al. FaR-gan for one-shot face reenactment. arXiv preprint arXiv:2005.06402, 2020

[124] Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation//Proceedings of the Advances in Neural Information Processing Systems 32. Vancouver, Canada, 2019: 7137-7147

[125] Bregler C, Covell M, Slaney M. Video rewrite: Driving visual speech with audio//Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques.

Los Angeles, USA, 1997: 353-360

[126] Brand M. Voice puppetry//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA. 1999: 21-28

[127] Das D, Biswas S, Sinha S, et al. Speech-driven facial animation using cascaded gans for learning of motion and texture//Proceedings of the European Conference on Computer Vision. 2020: 408-424

[128] Jiang Z H, Wu Q, Chen K, et al. Disentangled representation learning for 3D face shape//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11957-11966

[129] Deng Y, Yang J, Xu S, et al. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA, 2019

[130] Zhou H, Sun Y, Wu W, et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4176-4186

[131] Nirkin Y, Masi I, Tuan A T, et al. On face segmentation, face swapping, and face perception//Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. Xi'an, China, 2018: 98-105

[132] Nirkin Y, Keller Y, Hassner T. Fsgan: Subject agnostic face swapping and reenactment//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 7184-7193

[133] Li L, Bao J, Yang H, et al. Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457, 2019

[134] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 586-595

[135] Li Y, Yang X, Sun P, et al. Celeb-DF: A large-scale challenging dataset for deepfake forensics//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3207-3216

[136] Gandhi A, Jain S. Adversarial perturbations fool deepfake detectors// Proceedings of the 2020 International Joint Conference on Neural Networks. 2020: 1-8

[137] Song F, Tan X, Liu X, Chen S. Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. Pattern Recognition, 2014, 47(9):2825-2838

[138] Li Y, Chang MC, Lyu S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking//Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong, China, 2018:1-7

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

[140] Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC. The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854, 2019

[141] He Y, Gan B, Chen S, Zhou Y, Yin G, Song L, Sheng L, Shao J, Liu Z. ForgeryNet: A versatile benchmark for comprehensive forgery analysis// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4360-4369

[142] Zhu Y, Li Q, Wang J, et al. One Shot Face swapping on megapixels//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4834-4844

[143] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stability, and variation// Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-26

[144] Lu Y, Tai Y W, Tang C K. Attribute-guided face generation using conditional cyclegan//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 282-297

[145] Neves J C, Tolosana R, Vera-Rodriguez R, et al. Real or fake? spoofing state-of-the-art face synthesis detection systems. arXiv preprint arXiv:1911.05351, 2019

[146] Shen Y, Luo P, Yan J, et al. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 821-830

[147] Li Y, Liu S, Yang J, et al. Generative face completion// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3911-3919

[148] Song L, Cao J, Song L, et al. Geometry-aware face completion and editing//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 2506-2513

[149] Wang R, Ma L, Juefei-Xu F, Xie X, Wang J, Liu Y. FakeSpotter: A Simple baseline for spotting AI-synthesized fake faces. arXiv preprint arXiv:1909.06122, 2019

[150] Jain A, Singh R, Vatsa M. On detecting gans and retouching based synthetic alterations//Proceedings of the International Conference on Biometrics Theory. Applications and Systems. Los Angeles, USA, 2018:1-7

[151] D. Yi, Z. Lei, S. Liao, and S. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923, 2014

[152] Nataraj L, Mohammed TM, Manjunath BS, Chandrasekaran S, Flenner A, Bappy JH, Roy-Chowdhury AK. Detecting gangenerated fake images using co-occurrence matrices. arXiv preprint arXiv:1903.06836, 2019

[153] Zhang X, Karaman S, Chang SF. Detecting and simulating artifacts in ganfake images. arXiv preprint arXiv: 1907.06515, 2019

[154] Bharati A, Singh R, Vatsa M, Bowyer KW. Detecting fa-

cial retouching using supervised deep learning. IEEE Transactions on Information Forensics and Security,2016, 11(9): 1903-1913

[155] Marra F, Saltori C, Boato G, Verdoliva L. Incremental learning for the detection and classification of gan-generated images//Proceedings of the International Workshop on Information Forensics and Security. Delft, The Netherlands, 2019:1-6

[156] Khalid H, Woo S S. OC-FakeDect: Classifying deepfakes using one-class variational autoencoder//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 656-657

[157] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv: 1803.09179, 2018

[158] Sabir E, Cheng J, Jaiswal A, et al. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI), 2019, 3(1): 80-87

[159] Amerini I, Galteri L, Caldelli R, et al. Deepfake video detection through optical flow based CNN//Proceedings of the IEEE International Conference on Computer Vision Workshops. Seoul, Korea, 2019

[160] Chollet F. Xception: Deep Learning with depthwise separable convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1251-1258

[161] Hsu CC, Zhuang YX, Lee CY. Deep fake image detection based on pairwise learning. Applied Sciences, 2020, 10(1): 378

[162] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations//Proceedings of the IEEE Winter Applications of Computer Vision Workshops. Hawaii, USA, 2019: 83-92

[163] Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: Acompact facial video forgery detection network//Proceedings of the International Workshop on Information Forensics and Security. Hong Kong, China, 2018:1-7

[164] 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

[165] Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656, 201

[166] Zhang Y, Zheng L, Thing VL. Automated face swapping and its detection//Proceedings of the 2017 IEEE 2nd International Conference on Signal and Image Processing. Singapore, 2017:15-19

[167] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features//Proceedings of the European conference on computer vision. Graz, Austria, 2006: 404-417

[168] Qiu G. Indexing chromatic and achromatic patterns for content-based colour image retrieval. Pattern Recognition, 2002, 35(8): 1675-1686

[169] Yang, X., Li, Y., and Lyu, S. Exposing deep fakes using inconsistent head poses//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK 2019:8261-8265

[170] Li L, Bao J, Zhang T, et al. Face x-ray for more general face forgery detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5001-5010

[171] Zhao H, Zhou W, Chen D, et al. Multi-attentional deepfake detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2185-2194

[172] Li J, Xie H, Li J, Wang Z, Zhang Y. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6458-6467

[173] Chandrasegaran K, Tran NT, Cheung NM. A closer look at fourier spectrum discrepancies for cnn-generated images detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7200-7209

[174] Qian Y, Yin G, Sheng L, Chen Z, Shao J. Thinking in frequency: Face forgery detection by mining frequency-aware clues//Proceedings of the European Conference on Computer Vision. 2020: 86-103

[175] Güera D, Delp EJ. Deepfake video detection using recurrent neural networks//Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. Auckland, New Zealand, 2018: 1-6

[176] Wu X, Xie Z, Gao Y, Xiao Y. SSTNet: Detecting manipulated faces through spatial, steganalysis and temporal features//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain, 2020: 2952-2956

[177] Masi I, Killekar A, Mascarenhas R M, et al. Two-branch recurrent network for isolating deepfakes in videos//Proceedings of the European Conference on Computer Vision. 2020: 667-684

[178] Mittal T, Bhattacharya U, Chandra R, et al. Emotions don't lie: A deepfake detection method using audio-visual affective cues. arXiv preprint arXiv:2003.06711, 2020

[179] Chai L, Bau D, Lim SN, Isola P. What makes fake images detectable? understanding properties that generalize//Proceedings of the European Conference on Computer Vision. 2020: 103-120

[180] Zhou P, Han X, Morariu VI, Davis LS. Two-stream neural networks for tampered face detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 1831-1839

[181] Nguyen HH, Yamagishi J, Echizen I. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467, 2019

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

[183] Tariq S, Lee S, Kim H, Shin Y, Woo SS. Detecting both machine and human created fake face images in the wild//Proceedings of the International Workshop on Multimedia Privacy and Security. Toronto, Canada, 2018: 81-87

[184] Liu Z, Qi X, Torr P H S. Global texture enhancement for fake face detection in the wild//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle Washington, USA, 2020: 8060-8069

[185] Li H, Li B, Tan S, Huang J. Identification of deep network generated images using disparities in color components. Signal Processing, 2020,174:107616

[186] Marra F, Gragnaniello D, Verdoliva L, et al. Do gans leave artificial fingerprints? //Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval. San Jose, USA,2019: 506-511

[187] Richardson E, Weiss Y. On gans and gmms. arXiv preprint arXiv:1805.12462. 2018

[188] Wang SY, Wang O, Zhang R, et al. Cnn-generated images are surprisingly easy to spot. for now//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8695-8704

[189] Durall R, Keuper M, Keuper J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7890-7899

[190] Carlini N, Farid H. Evading deepfake-image detectors with white-and black-box attacks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 658-659

[191] Hussain S, Neekhara P, Jere M, et al. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3348-3357

[192] Chen Z, Xie L, Pang S, et al. MagDR: Mask-guided detection and reconstruction for defending deepfakes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9014-9023

[193] Ruiz N, Bargal S A, Sclaroff S. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems//Proceedings of the European Conference on Computer Vision. 2020: 236-251

[194] Yeh C Y, Chen H W, Tsai S L, et al. Disrupting image-translation-based deepfake algorithms with adversarial attacks//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Seattle, USA, 2020: 53-62

[195] Li Y, Yang X, Wu B, et al. Hiding faces in plain sight: Disrupting aiface synthesis with adversarial perturbations. arXiv preprint arXiv:1906.09288, 2019

[196] Dang-Nguyen DT, Pasquini C, Conotter V, Boato G. RAISE: A raw images dataset for digital image forensics//Proceedings of the 6th ACM Multimedia Systems Conference. New York, USA, 2015:219-1

[197] Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A. VGGFace2: A dataset for recognising faces across pose and age//Proceedings of the International Conference on Automatic Face & Gesture Recognition. Xi'an, China, 2018: 67-74

[198] Perov I, Gao D, Chervoniy N, Liu K. Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535, 2020

LIN Chen-Hao, Ph. D., research fellow. His research interests are artificial intelligence security, adversarial machine learning, deep fake and identity authentication.

SHEN Chao, Ph. D., professor. His research interests are trusted artificial intelligence, artificial intelligence security and cyber-physical power system security.

DENG Jing-Yi, Ph. D. candidate. Her research interest is digital image forgery detection.

HU Peng-Bin, M. S. candidate. His research interest is digital image forgery generation.

WANG Qian, Ph. D., professor. His research interests are artificial intelligence security, security and privacy of cloud computing, wireless systems security, big data securityand applied cryptography.

MA Shi-Qing, Ph. D., assistant professor. His research interests are system and software security, adversarial machine learning and software engineering.

LI Qi, Ph. D., associate professor. His research interests areinternet and cloud security, mobile security, machine learning security, big data security and blockchain security.

GUAN Xiao-Hong, Ph. D., professor. His research interests areoptimal scheduling of power and manufacturing system and cyber security.

Background

With the maturity and development of deep learning and artificial intelligence technologies, the research field of computer vision, especially the field of digital face, including the tasks of face detection, face recognition, and face generation, have made great progress. However, meanwhile the artificial intelligence technology has also been abused by some applications, such as the most popular adversarial face example attack technology and deep face manipulation technology. These applications generate digitally forged face content that seriously threaten the social security and invade people's privacy. To address this challenge, both academic community and industrial community have invested large efforts in researching the generation and detection methods. Thus, systematically summarizing the development status of digitally forged face content is an important thing to understand this field in-depth and maintaining the security of artificial intelligence. Although there are some literatures reviewing the artificialintelligence security, most of these works concentrates on the general security of artificial intelligence rather than the security of digital face content. With this goal, we introduce a systematic and comprehensive review on the development status of the digitally forged face content in this paper. We mainly focus on the adversarial face example attack technology and deep face manipulation technology. First summarizing the definition and subcategories of digitally forged face content. Then individually introducing the popular attack algorithms, detection algorithms, datasets and existing shortages of these two technologies. Finally indicating the challenge in this stage and support a potential research direction in the future. This review aims to promote the further development of forged face content and provide guidance and reference for ensuring the secure application of artificial intelligence face content-related technologies. This research is supported by the National Key Research and Development Program of China (2020AAA0107702), the National Natural Science Foundation of China (62006181, U20A20177, 61822309, 61703301, U21B2018) and the Shaanxi Province Key Industry Innovation Program (2021ZD LGY01-02).