视听觉深度伪造检测技术研究综述

梁瑞刚[1,2], 吕培卓[1,2], 赵 月[1,2], 陈 鹏[1,2], 邢 豪[3], 张颖君[4], 韩冀中[1,2], 赫 然[5], 赵险峰[1,2], 李 明[3], 陈 恺[1,2*]

[1] 中国科学信息工程研究所信息安全国家重点实验室 北京 中国 100093
[2] 中国科学院大学网络空间安全学院 北京 中国 100049
[3] 太原理工大学大数据学院 太原 中国 030024
[4] 中国科学院软件研究所 北京 中国 100190
[5] 中国科学院自动化研究所 北京 中国 100190

摘要 深度学习被广泛应用于自然语言处理、计算机视觉和无人驾驶等领域, 引领了新一轮的人工智能浪潮。然而, 深度学习也被用于构建对国家安全、社会稳定和个人隐私等造成潜在威胁的技术, 如近期在世界范围内引起广泛关注的深度伪造技术能够生成逼真的虚假图像及音视频内容。本文介绍了深度伪造的背景及深度伪造内容生成原理, 概述和分析了针对不同类型伪造内容(图像、视频、音频等)的检测方法和数据集, 最后展望了深度伪造检测和防御未来的研究方向和面临的挑战。

关键词 深度伪造; 深度学习; 生成对抗网络
中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.02.01

A Survey of Audiovisual Deepfake Detection Techniques

LIANG Ruigang[1,2], LV Peizhuo[1,2], ZHAO Yue[1,2], CHEN Peng[1,2], XING Hao[3], ZHANG Yingjun[4], HAN Jizhong[1,2], He Ran[5], ZHAO Xianfeng[1,2], LI Ming[3], CHEN Kai[1,2*]

[1] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Science, Beijing 100049, China
[3]College of Data Science, Taiyuan University of Technology, Taiyuan 030024, China
[4]Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[5]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract Deep learning has been widely used in fields such as natural language processing, computer vision, and driverless vehicles, leading a new wave of artificial intelligence. Deep learning advances however have also been used to create technologies that pose potential threats to national security, social stability, and personal privacy. For example, deepfakes that have recently attracted widespread attention worldwide, which could generate seemingly realistic fake images, audio and video content. This article introduces the background of deepfakes and the principles of deepfakes creation, and then outlines and analyzes the detection methods and datasets for different types of deepfakes, including images, videos, audios, etc. Finally, the article discusses potential research directions and challenges of deepfakes detection and prevention.

Key words deepfakes; deep learning; generative adversarial network

1 引言

近年来, 随着科技的快速发展和计算能力的飞速提升, 人工智能理论技术不断成熟, 现已在金融、医疗、城市服务、工业制造和生活服务等领域得到了广泛应用, 人工智能(Artificial intelligence, AI)技术正在引领新一轮的全方位产业变革, 推动人类世界进入智能化时代。机器学习(Machine learning, ML)是人工智能的核心研究领域, 据德勤 2019 年发布的《全球人工智能发展白皮书》统计[1], 89%的人工智能专利申请和 40%人工智能范围内的相关专利均属于机器学习。深度学习(Deep learning, DL)作为机器

通讯作者: 陈恺, 博士, 研究员, Email: chenkai@iie.ac.cn。
本课题得到国家重点研发项目(No.2016QY04W0805), 国家自然科学基金项目(No.61728209), 中国科学院青年创新促进会, 北京市科技新星计划, 北京市自然科学基金项目(No.JQ18011), 国家前沿科技创新项目(No. YJKYYQ20170070)资助。
收稿日期: 2019-12-31; 修改日期: 2020-03-04; 定稿日期: 2020-03-09

学习领域中的热门研究方向, 为计算机视觉、无人驾驶、自然语言处理和语音识别等领域的创新提供了有力的技术支撑。然而, 深度学习技术在引领新一轮人工智能浪潮的同时, 也对个人隐私数据, 社会稳定和国家安全等造成了潜在威胁。自 2017 年以来, 以“深度伪造”技术为支撑的语音诈骗[2]事件在世界范围内引起广泛关注, 恶搞政治人物[3]及公众人物[4]的“换脸”视频事件[5]也层出不穷, 造成了非常恶劣的负面影响[6], 甚至间接导致了加蓬共和国的军事政变[7]。

深度伪造一词译自英文“Deepfake”(“deep learning” 和 “fake” 的组合[8]), 最初源于一个名为“deepfakes”的 Reddit 社交网站用户, 该用户于 2017 年 12 月在 Reddit 社交网站上发布了将斯嘉丽·约翰逊等女演员的面孔映射至色情表演者身上的伪造视频[8]。深度伪造目前在国际上并没有公认的统一性定义, 美国在其发布的《2018 年恶意伪造禁令法案》[9]中将“deep fake”定义为“以某种方式使合理的观察者错误地将其视为个人真实言语或行为的真实记录的方式创建或更改的视听记录”, 其中“视听记录”即指图像、视频和语音等数字内容。由于深度伪造技术自身并不存在善恶, 所以本文更加倾向于赋予“深度伪造”一个中立化定义: 基于深度学习等智能化方法创建或合成视听觉内容(如图像、音视频、文本等)。

深度伪造技术可以推动娱乐与文化交流产业的新兴发展, 如可应用于在电影制作中创建虚拟角色、视频渲染、声音模拟; “复活”历史人物或已逝的亲朋好友, 实现“面对面”沟通, 创造了一种新型的交流方式。深度伪造技术也可用于误导舆论、扰乱社会秩序, 甚至可能会威胁人脸识别系统[10]、干预政府选举[11]和颠覆国家政权[12]等, 已成为当前最先进的新型网络攻击形式。基于深度伪造技术构建的图像/视频换脸、语音诈骗等事件数见不鲜, 相继出现了 FakeApp[13]、Faceswap[14]等(表 1)多个“一键式”内容合成(图像、视频、语音)应用程序[19], 2019 年 6 月甚至出现了“一键式”智能脱衣软件 Deepnude[20], 虽然该软件在发布之后即迫于舆论压力被开发者下架, 但仍在全球范围内引起了巨大恐慌。据 Deeptrace 公司近期的调研发现[21], 目前在互联网上使用深度伪造技术生成的伪造视频至少有 14678 个, 其中涉及色情信息的占比高达 96%。深度伪造内容的危害和影响已经蔓延至世界各地, 针对深度伪造内容的检测和防御现已成为世界各国政府、企业乃至个人所关注的热点问题之一。

表 1 代表性“深度伪造”应用软件

Table 1 Representative Deepfake Software

名称	出现时间	功能
FakeApp[13]	2018.01	图片和视频合成
Faceswap[14]	2019.02	图片和视频合成
DeepFaceLab[15]	2019.02	图片和视频合成
RealTalk[16]	2019.05	语音合成
Melnet[17]	2019.06	语音合成
ZAO[18]	2019.08	图片和视频合成

目前, 各国政府针对深度伪造的监管问题均已出台了相关的政策法规。欧盟于 2019 年 4 月 8 日发布了《人工智能道德准则》[22], 将隐私和数据管理作为可信赖人工智能需要满足的七个要素之一。美国针对深度伪造被恶意滥用的现状, 在联邦及州层面均提出了一系列法案, 如众议员 Yvette Clarke 于 2019 年 6 月 12 日提出的《深度伪造责任法案》[23], 目的是防止深度伪造内容干扰大选; 美国加州州长 Gavin Newsom 于 2019 年 10 月 3 日签署法令《AB 730》和《AB 602》, 禁止使用深度伪造技术干预选举和制作色情内容[24]。我国国家互联网信息办公室也于 2019 年 12 月 20 日发布了《网络信息内容生态治理规定》[25], 明确网络信息内容服务使用者和网络信息内容生产者、网络信息内容服务平台不得开展网络暴力、人肉搜索、深度伪造、流量造假和操纵账号等违法活动, 该规定自 2020 年 3 月 1 日起开始实施。

本文针对深度伪造内容检测技术的研究和发展状况进行梳理和归纳总结, 讨论其技术特点、存在问题, 进一步分析深度伪造检测和防御所面临的挑战, 旨在为深度伪造内容检测技术的研究和发展提供方向性指导。

本文组织结构如下: 第 2 章扼要介绍了深度伪造生成技术的基本原理及发展现状; 第 3 章对已存在的深度伪造内容检测技术进行了归纳和分析; 第 4 章对已有的深度伪造数据集进行了介绍; 第 5 章则对影响深度伪造检测的相关技术进行了介绍和分析; 第 6 章展望了深度伪造检测技术的未来研究方向和所面临的挑战。

2 深度伪造内容生成技术

深度伪造内容生成工具多依赖于深度学习技术开发, 任何人可通过现有的深度伪造软件(模型)轻松创建图像、视频和语音等伪造内容。深度学习技术具有准确表示复杂、高维、大规模数据, 进而直接提取特征的能力, 极大地推动了语音识别、视觉目标识

别、基因组学等新兴领域的快速发展。当前, 在深度伪造中广泛使用的深度学习技术主要有生成对抗网络(Generative adversarial networks, GAN)[26]、卷积神经网络(Convolutional neural network, CNN)[27]、循环神经网络(Recurrent neural network, RNN)[28]和变分自编码器(Variational auto-encoder, VAE)[29]。

GAN 源于博弈论中的“零和博弈”思想[31], 是一种通过生成模型(generative model)和判别模型(discriminative model)互相博弈的方法来学习数据分布的生成式网络[32]。其中生成模型通过给定某种隐含信息, 随机生成观测数据样本; 判别模型则需要预测数据样本是否属于真实训练样本。两者通过对抗式训练提升其能力, 最理想的状态是生成模型能够生成足以“以假乱真”的数据样本, 而判别模型却对其真伪性难以判别, 即判断正确的概率只有 50%。相比于其他生成式模型, GAN 具有以下特点和优势: 不依赖先验知识; 生成模型的参数更新来自判别模型的反向传播, 而非直接来自于数据样本, 故训练不需要复杂的马尔科夫链。GAN 已在图像编辑、数据生成、恶意攻击检测、肿瘤识别和注意力预测等领域得到了广泛应用。

CNN 是仿造生物视知觉机制构建的一类包含卷积计算且具有深度结构的前馈神经网络[32], 由一个输入层、输出层和多个隐含层构成。隐含层通常包含卷积层、池化层和全连接层三部分, 其中卷积层主要负责局部特征的提取, 池化层专注于参数降维, 全连接层则用于输出结果。基于 CNN 的相关技术现已在图像、音视频的智能化处理方面取得了重要的突破。

RNN 是一类以序列数据为输入, 在序列的演进方向进行递归, 且所有节点按链式连接的递归神经网络(Recursive neural network)[32]。RNN 由输入层、隐藏层和输出层组成, 其特点是每次都会将上一次的输出结果输入下一次的隐藏层中一起训练, 故较为擅长处理序列化数据。双向循环神经网络(Bidirectional recurrent neural networks, BRNN)[33]和长短期记忆网络(Long short-term memory networks, LSTM)[34]是目前被广泛使用的 RNN 模型。RNN 对诸如文本、语音及语言之类的顺序数据的智能化处理带来了新的探索方向。

VAE 是基于自编码器(Autoencoder, AE)和高斯混合模型(Gaussian Mixture Model, GMM)构建的一种可以直接通过随机梯度下降进行训练的深度生成模型。VAE 主要由编码器和解码器两部分组成: 编码器负责将数据分布的高级特征映射到数据的低级表征(隐变量), 并加入了高斯噪声; 解码器则通过吸收数据的低级表征进而输出同样数据的高级表征, 相比 AE 具有更高的鲁棒性。VAE 能够直接比较生成数据和原始数据之间的差异性, 现已在数据可视化、自然语言处理、图像和音频合成等领域等到了广泛应用。

2.1 视觉深度伪造生成技术

2017 年, Korshunova 等人[35]提出了一种基于 GAN 的自动化实时换脸技术。同年, Suwajanakorn 等人[36]使用 LSTM 设计了一种智能化学习口腔形状和声音之间关联性的方法, 该方法仅通过音频即可合成对应的口部特征, 并基于美国前总统奥巴马在互联网上已有的音频和视频片段, 生成了非常逼真的假视频。这些技术一经问世, 便引起了广泛的关注, 基于其原理实现的相关开源项目也陆续问世, 进而为视觉深度伪造生成技术的孕育和发展提供了契机。

视觉深度伪造生成技术的实现总体可以分为数据收集、模型训练和伪造内容生成三个核心步骤, 本文参考文献[37-38], 以伪造人脸图像生成为例对深度视觉伪造生成技术的共性原理进行简单介绍, 假设我们的目标是将 Alice 的脸换至 Bob 的身体上, 伪造图像具体生成流程如图 1 所示。

1) 数据收集

数据收集顾名思义是通过各种渠道对 Alice 和 Bob 的已有图像进行大量收集, 以便为模型训练提供数据支撑。

2) 模型训练

目前, 深度伪造模型的构造主要基于 GAN 实现, 由编码器(encoder)和解码器(decoder)构成: 编码器用于提取人脸图像的潜在特征, 解码器则用于重构人脸图像, 基于该原理的典型工具如 DeepFake_tf[39]和 Dfaker[40]。为了实现换脸操作, 模型需要两个编码器/解码器对(编码器 A/解码器 A, 编码器 B/解码器 B), 分别基于已收集的 Alice 和 Bob 的图像集进行训练, 其中编码器 A 和编码器 B 具有相同的编码网络(即参数共享), 编码器的统一性能够保证模型学习到两组图像面部结构之间的相似性(如五官特征)。伪造模型具体的训练过程如图 1(a)所示。

3) 伪造图像生成

待模型训练完成之后, 通过将模型训练中 Alice 和 Bob 的解码器互换, 进而构建新的编码器/解码器对(编码器 A/解码器 B, 编码器 B/解码器 A), 然后选取 Alice 的一张图像作为目标图像, 在编码器 A 编码完成之后, 基于解码器 B 进行解码, 从而生成载有

Alice 面部、Bob 身体的深度伪造(换脸)图像, 如图 1(b)所示。

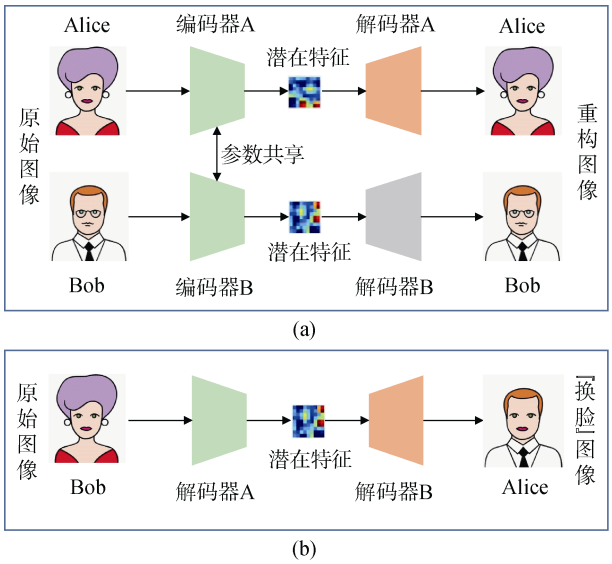

图 1 视觉深度伪造内容生成流程示例

Figure 1 Example of visual deepfake generation process

中国科学院自动化研究所的谭铁牛、赫然团队从全光人脸分析、视觉拓扑优先、生成对抗架构和身份保持结构四个人脸图像重组和编辑所面临的核心基础问题入手, 提出了轻量级神经网络 Light CNN[44]、超分辨率 Wavelet-SRNet[45]、视觉旋转[46-47]、表情编辑[48]、年龄变换[49]、像素补充[50]和跨光谱合成[51]等一系列理论及方法, 在内容和表观上实现了对图像的重组和编辑, 能够生成逼真的人脸伪造图像。

Faceswap-GAN[52]是当前较为热门的视觉深度伪造内容生成模型, 可用于生成分辨率为 64*64(默认分辨率)、128*128 和 256*256 的视觉伪造内容[38]。Faceswap-GAN 以 GAN 的改进版本 CycleGAN[53]为支撑, 融合了 VGGFace[54]和 FaceNet[55]模型的核心技术: 引入了多任务卷积神经网络(Multi-task convolutional neural networks, MTCNN)[56], 使得针对人脸的检测更加稳定, 增加了人脸对齐的可靠性; 将对抗性损失和感知性损失添加到 GAN 的编码器-解码器体系结构之中; 添加了知觉损失, 使眼睛部位的动作更加逼真并与输入图像的脸部保持一致。

近期, Fried 等人[57]证明能够在保持无缝视听流的前提下以类似修改文本的方式编辑视频内容, 如单词的增加、删除和修改, 句子拼接, 语言翻译及背景替换等操作。

2.2 听觉深度伪造生成技术

音频生成技术最初的研究主要专注于文本到语音的转换(Text-to-speech, TTS), 主要可分为两种方法: 拼接式语音合成方法和基于参数估计的语音合成方法[64]。在拼接式语音合成方法中, 音频的生成主要通过对语音索引词典中预先录制的小部分语音进行排序。基于参数估计的语音合成方法则通过将文本映射到语音的显著参数, 进而基于声码器来合成语音。其中典型语音参数估计方法为隐马尔可夫模型(Hidden markov model, HMM)[58-59]。随着人工智能技术的兴起, 研究人员借鉴图像、视频的新型智能化合成技术, 开始探索智能化辅助的语音合成方法, 陆续提出了基于声码器[57-62]、GAN[63-64]、自编码器(Denoising autoencoder, DAE)[65]、自回归模型(Autoregressive model, AR)[66-68]等一系列新兴的语音合成技术, 推动了语音合成产业的快速发展。

中国科学院自动化研究所的陶建华团队[41-43]从言语生成和感知深层机理理解, 语言、口语和情感处理的深层次分析等多角度出发, 提出了高效、鲁棒的自然口语语音交互技术。

百度 Ping 等人[60]提出了一种全卷积的特征到频谱的体系架构 Deep Voice 3, 该架构能够将字符、音素和重音等文本特征转换为各种声学特征, 进而将其作为声音波形合成模型的输入。相较于其历史版本 Deep Voice[61]和 Deep Voice 2[62], Deep Voice 3 提升了模型的训练速度, 更加适用于大规模的录音数据集。

Pascual 等人[63]基于 GAN 提出了一种增强原始音频语音的方法, 通过生成器完成对噪音的过滤, 有效提升了输出语音的清晰度。Donahue 等人[64]进一步基于 GAN 提出了一种无监督音频生成模型 WaveGAN, 该模型可以从人类语音的少量词汇中生成可理解的单词。基于 WaveGAN 合成的语音具有很好的音质。

Ren 等人[65]基于 DAE 和对偶转换技术设计了一种(近乎)无监督的语音合成方法, 该模型仅需 200 个配对的语音和文本数据(近 20 分钟的素材)及额外的非配对语音和文本数据即可实现对特定目标人物的语音伪造。

Facebook AI 研究团队的 Vasquez 等人[66]基于频谱图(spectrogram)设计了一种新型的端到端的语音生成模型 Melnet, 该模型结合了细粒度的自回归模型和多尺度模型[69-72], 能够同时捕获局部和全局结构。Melnet 生成的语音内容不仅可以重现人类的语调, 而且可以像真实的人一样说话, 性能优于

SampleRNN[67]和 WaveNet[68]等基于波形(waveform)的语音生成模型。MelNet 能够在几秒钟内重现 TED 演讲者的声音[17]。

Google AI 团队 Jia 等人基于序列到序列模型(Sequence to sequence, Seq2Seq)[73-74]开发了一款语音翻译模型 Translatotron[75], 该模型可以实现将语音从一种语言直接转换为另一种语言, 且在翻译后的语音中保留原始说话者的声音特征, 从而使得翻译后的语音听起来更加自然[76]。

3 深度伪造内容检测技术

由于深度伪造内容具有辨别难度大、制造成本低、传播速度快和破坏能力强等特点, 对个人隐私数据, 社会稳定甚至国家安全等造成严重的潜在威胁, 所以亟需提出切实有效的深度伪造内容检测方法来应对深度伪造内容带来的严峻挑战。

现有的深度伪造内容检测方法多依赖于深度学习模型, 基于深度伪造内容数据集的训练, 实现特征提取并构建分类器。特征提取可分为自动提取和手动提取两种类型: 自动提取指在数据集上直接训练模型, 即让模型自主学习和提取能够区分真伪内容的特征; 手动提取特征则需要对数据集进行预处理, 人工提取出部分特征, 进而基于已提取特征完成分类器的训练。

本章主要概述和分析了深度伪造相关的数字内容检测方法。根据检测伪造内容的不同, 现有的方法可以分为视觉深度伪造检测技术和听觉深度伪造检测技术。

3.1 视觉深度伪造检测技术

深度伪造内容颠覆了人们对“眼见为实”观念的认知, 近两年出现的视觉深度伪造主要有换脸、表情迁移和动作迁移等方式, 造成了全球范围内的“信任”危机。现有(已调研)的视觉深度伪造内容检测方法可分为深度伪造图像检测技术和深度伪造视频检测技术两大类。

3.1.1 深度伪造图像检测技术

Agarwal 等人[77]将基于 GAN 生成的深度伪造图像的检测归纳为一种假设性检验问题。针对这种假设性检验问题, Maurer 等人[78]将身份认证的信息论研究引入到统计框架之中, 并且定义了预言误差, 即真实图像与 GAN 生成的伪造图像之间的最小距离: 预言误差越大, 则 GAN 的精度越差, 而深度伪造图像的检测则相对越容易。

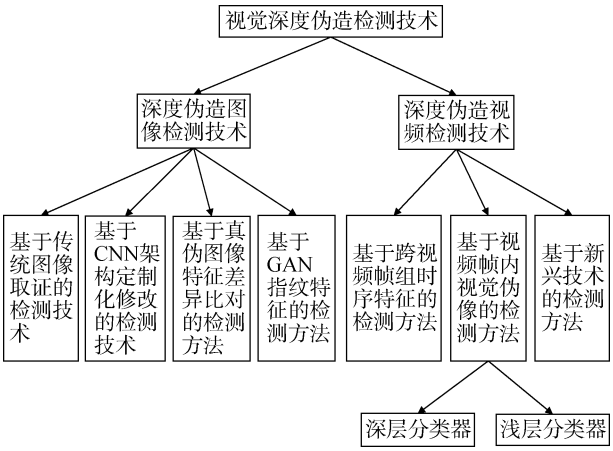

图 2 现有视觉深度伪造检测技术分类

Figure 2 Classification of visual deepfake detection methods

目前, 深度伪造图像的检测方法基于其判别原理可分为四种: 第一种是借鉴传统图像取证方法, 在像素级别构建模型检测深度伪造图像; 第二种是通过修改 CNN 架构和损失函数等方式深度伪造图像检测方法; 第三种则通过分析和提取真伪图像自身的差异化特征, 进而训练分类器实现深度伪造图像的检测; 第四种是通过真伪图像频谱中的差异化分析, 找出特定 GAN 的指纹特征, 最终实现对伪造图像的识别。

1) 基于传统图像取证的检测技术

Nataraj 等人[79]通过提取像素域中 RGB 通道上的共现矩阵(co-occurrence matrices), 基于 CNN 构建了一种像素级的图像检测模型来实现对 GAN 生成伪造图像的检测。Wang 等人[80]提出一种在商业软件 Adobe Photoshop 上编写脚本来检测合成图像内容的方法。然而, 这类借鉴传统图像取证技术的深度伪造图像检测模型[81-82]可通过在伪造图像中加噪声的方式绕过。

2) 基于 CNN 架构定制化修改的检测技术

Mo 等人[83]通过修改 CNN 架构(如输入图像的高通滤波器、层组数和激活函数), 进而以监督学习的方式实现了对深度伪造图像的检测。但是这种通过定制化修改 CNN 架构和损失函数等方式构建的深度伪造图像检测模型[84-85, 107]容易受到对抗样本的攻击[86]。

3) 基于真伪图像特征差异比对的检测方法

Zhang 等人[87]使用稳定特征加速算法(Sped up robust features, SURF)[88]和词袋模型(Bag of words,

BoW)[89]来提取图像特征, 并将其分别在支持向量机(Support vector machine, SVM)[90], 随机森林(Random forest, RF)[91]和多层感知器(Multilayer perceptron, 简称 MLP)[92]等分类器上进行了测试, 准确率均可达到92%以上。但该模型所使用的数据集相对较小, 仅包含 10000 张图像(伪造图像占 50%), 且其数据集中的伪造图像质量也未与其他深度伪造数据集进行比较。

Huh 等人[93]通过图像的 EXIF 元数据特征实现了对深度伪造图像的检测, 但 EXIF 元数据可以被修改和删除。Yang 等人[94]通过提取真实图像和伪造图像面部标志点位置之间的差异性特征来进行分类器训练, 进而实现对深度伪造图像的识别。但是随着 GAN 技术的改进和深度伪造内容生成模型性能的逐步提升, 真伪图像之间的差异性将会逐渐缩小, 甚至可能消失。

Hsu 等人[95]提出了一种新型的通用伪特征网络模型(Common fake feature network, CFFN)。该模型可分为两个阶段: 第一阶段基于多种 GAN[96-102]生成大量的<真实图像、伪造图像>对, 共计获得了 385198 张训练图像和 10000 张<真实图像、伪造图像>对。CFFN 通过已收集的<真实图像、伪造图像>对来学习真、伪图像的鉴别特征。其中 CFFN 基于 Siames 网络架构[103-104]设计, 包含 3~5 个密集单元, 具体数目取决于输入数据是人脸图像还是普通图像, 为了准确提取伪造图像的代表性特征, 每个密集单元包含了不同数量的密集块[78]。第二阶段是一个具有跨级别伪特征提供能力的 CNN, 通过将该 CNN 连接到 CFFN 的最后一个卷积层, 进而基于第一阶段提取的鉴别特征完成对深度伪造图像的识别。该方法在 CelebA 数据集[113]上进行了评估, 相比于[83, 105-107]等检测方法具有更高的性能。

4) 基于 GAN 指纹特征的检测方法

Zhang 等人[108]通过探索 GAN 指纹特征[109-110], 提出了一种基于频谱输入的分类器模型 AutoGAN[111], 该模型能够实现对基于 CycleGAN[53]等流行 GAN 模型所生成的伪造图像的准确检测。Zheng 等人[112]则提出了一种基于频域分析的深度伪造图像分类模型 SegNet, 该模型在高像素图像集 Faces-HQ[112]、中像素图像集 CelebA[113]和低像素图像集 FaceForensics++[114]上均具有较高的准确率。Wang 等人[86]提出了一种基于神经元覆盖的深度伪造图像检测方法, 其性能优于基于传统图像取证和 CNN 架构定制化修改的深度伪造图像检测模型。然而, 深度伪造图像生成模型可通过选用无指纹特征的 GAN 来绕过这类检测模型, 且 GAN 技术进展迅速, 所以上述检测方法所提取的 GAN 指纹特征并不具有持久性和通用性。

3.1.2 深度伪造视频检测技术

由于视频在被压缩后, 帧数据会产生严重的退化现象[115], 且视频帧组之间的时序特征存在一定的变化, 故多数基于静态特征的深度伪造图像检测方法无法直接用于深度伪造视频的检测。当前, 深度伪造视频检测方法可分为三大类: 第一类是基于跨视频帧组时序特征的检测方法, 第二类是基于视频帧内视觉伪像的检测方法, 第三类则是基于新兴技术的检测方法。

1) 基于跨视频帧组时序特征的检测方法

由于深度伪造内容检测模型经常使用在线收集的(静态)面部图像集进行训练, 无法实现对眨眼、呼吸和心跳等生理信息的准确伪造, 故可以基于生理信息的合理性来构建深度伪造视频检测方法。Li 等人[120]提出一种基于眨眼来鉴别深度伪造视频的方法: 首先在视频帧层面提取出面部区域和眼睛区域, 然后经过人脸对齐、提取和缩放眼睛区域标点的边界框等操作创建新的帧序列, 进而将其分配至长期循环卷积网络(Long-term recurrent convolutional networks, LRCN)[121]中实现对睁眼和闭眼状态的动态预测, 该方法在 EBV[119]等数据集上具有良好的性能。但是这种方法仅仅将是否眨眼作为伪造视频评判指标, 并未进一步考虑眨眼频率的合理性, 故容易通过后期处理或训练具备眨眼能力的更高级模型等方式绕过。

由于深度伪造内容生成模型对目标视频的重构多在逐帧操作的基础上实现, 且在合成阶段不能有效地增强视频帧组之间的时间连贯性, Güera 等人[119]证明深度伪造视频帧内和帧之间时序具有不一致的特性, 进而基于 CNN 和 LSTM 提出了一种时间感知管道方法来检测深度伪造视频。其中 CNN 用于提取视频帧级特征, 接着将其馈入 LSTM 中以创建时间序列描述符。然而, 该方法鲁棒性不足, 易受到对抗样本的攻击。Sabir 等人[116]基于递归卷积网络(Recursive cortical network, RCN)提出了一种基于视频流时空特征的检测方法。由于 RCN 集成了 DenseNet[117]和门控循环单元[118], 因而该模型能够利用帧组之间的时序差异实现对深度伪造视频的检测, 该方法在数据集 FaceForensics++[114]上具有较高的准确率。

2) 基于视频帧内视觉伪像的检测方法

基于视频帧内视觉伪像的检测技术主要通过探索视频帧内视觉伪像提取判别特征, 并将这些特征

分配至深层或浅层分类器中进行训练, 其中深层分类器基于神经网络模型实现, 而浅层分类器则采用简单的机器学习模型实现, 最终完成对深度伪造视频的准确检测。

1) 深层分类器

Afchar 等人[115]提出了两种关注于肉眼可见特征的检测方法 MesoNet, 包括 Meso-4 和 MesoInception-4。由于 MesoNet[14, 126]是轻量级的神经网络, 因此其在保障了高性能的同时, 参数数量也少于 ResNet-50[123], XceptionNet[85]等深度神经网络。Afchar 等人同时也证明了眼睛和嘴巴部位的特征在深度伪造视频检测中具有至关重要的作用。

深度伪造视频通常需要基于人脸仿射变形技术(如缩放、旋转和剪切)将目标人物的面部准确匹配到原始视频, 因而可能致使合成视频的面部区域与周围环境之间的分辨率存在不一致的情况, 因而可以基于 CNN[122-123]模型构建深度伪造视频检测方法。Li 等人[124]提出一种基于面部变形后帧内视觉伪像特征的检测模型。该模型通过提取原始图像的面部区域并从多个尺度实现对齐处理, 再对随机选取的缩放图像应用高斯模糊并将其扭曲回原始图像, 从而动态的生成深度伪造视频。由于该方法在模型训练之前无需生成深度伪造视频作为反例样本(negative examples), 相较于 Mesonet[115]减少了大量时间和计算资源。但是该模型未在大量压缩视频上进行性能评估。

Nguyen 等人[127]提出了一种基于胶囊网络的视觉伪造检测方法。Sabour 等人[128]证明了胶囊网络能够准确描述对象部件之间的层次关系。胶囊网络通过动态路由算法, 以胶囊作为基本的训练单元, 在多次迭代后将三个胶囊的输出路由到对应的输出胶囊, 进而分离伪造图像和真实图像。该方法在 replayattack 数据集[129]、Deepfake 人脸交换数据集[115]、使用 Face2Face 方法[130]生成的面部重现数据集 FaceForensics[131] 和由 Rahmouni 等人设计的 CG(computer graphics)和 PG(photographic images)图像数据集[132]上均具有较高的准确率, 然而该方法对其抵御对抗机器攻击的能力进行评估。该方法证明了胶囊网络在构建视觉深度伪造通用检测模型方面具有较大的潜力。

2) 浅层分类器

基于深度伪造视频部分区域像素关系存在突变性, Koopman 等人[135]提出了一种基于光响应非均匀性(Photo response non-uniformity, PRNU)的检测方法。PRNU 是一种噪声模式, 噪声源于数码相机的感光传感器的出厂缺陷。每个数字相机的 PRNU 都不相同, 通常被视为数字图像的指纹[136-137]。由于被交换面部会改变视频帧中面部区域的局部 PRNU, 所以被广泛应用于面部操作检测。该方法首先将视频转换为帧, 并裁剪有问题的面部区域。然后将裁剪的帧按照顺序分为八个组, 在其中为每个组计算平均 PRNU。为了比较这些组之间的 PRNU, 计算归一化的互相关分数, Koopman 等人创建了一个包含 10 个真实视频和 16 个通过 DeepFaceLab[15]制造的伪造视频的测试数据集, 分析结果表明深度伪造视频和真实视频的平均标准化互相关系数存在显著差异。但该方法未在较大的数据集上进行测试, 不能够准确区分伪造视频和真实视频之间的互相关性, 也无法确定准确的似然比。

基于真实视频和伪造视频之间的帧内伪像或固有特征的区别, Matern 等人[134]基于眼睛、牙齿和面部轮廓等视觉特征设计了一种深度伪造视频检测方法。该方法利用眼睛和牙齿区域中缺失的反射和细节、面部区域的纹理特征和面部标志等生成特征向量, 采用逻辑回归和小型全连接神经网络(3 层, 64 个神经元)两个分类器实现对真伪视频的判别。然而, 该方法能够准确检测的前提是伪造视频需满足某些先决条件(如眼睛为睁开状态或牙齿为可见状态)。

3) 基于新兴技术的检测方法

Hasan 等人[138]基于区块链和智能合约构建了一种深度伪造视频检测方法, 该方法的前提假设是视频只有来源可追溯才是真实的。每个视频都与一个智能合约相关联, 该智能合约链接到其父视频, 并且每个父视频在其层次结构中都有一个指向其子视频的链接。通过该链, 即使视频已被多次复制, 用户也可以可靠地追溯到其与原始视频关联的初始智能合约。智能合约的一个重要属性是星际文件系统(Inter planetary file system, IPFS)具有独特的哈希值, 该哈希值可用于以分散和内容可寻址的方式存储视频[139]。Hasan 等人[138]进而对智能合约的关键特性和功能进行了测试, 以应对中间人(Man in the middle, MITM)、重放和分布式拒绝服务(Distributed denial of service, DDoS)等常见安全攻击。实验证明, 这种方法可以扩展到图像、音频和文本等其他数字内容的伪造检测之中。

3.2 听觉深度伪造检测技术

随着听觉深度伪造的流行和技术能力不断的提升, 针对恶意使用(如语音诈骗)的听觉深度伪造的检测变得越来越重要。现有的听觉深度伪造检测技术主要通过语速、声纹和频谱分布等生物信息的差异

化特征实现。

Digger[144]是由 Fraunhofer IDMT、Athens 技术中心和 Deutsche Welle 共同组成, Google DNI 依据其“数字新闻倡议”计划[151]资助的一个语音伪造内容检测项目, 旨在以知识共享的方式构建深度语音伪造检测领域的一个全球性合作社区。该项目通过利用先进的音频取证技术开发可检测视频中音频伪造的工具包, 进而将其集成于视频验证等应用程序之中。

2019 年自动说话人识别欺骗攻击与防御对策挑战赛(Automatic speaker verification spoofing and countermeasures challenge, ASVspoof)[140]由英国爱丁堡大学、法国 EURECOM 和日本 NEC 等多个世界领先的高校和科研机构共同发起, 旨在针对虚假语音攻击对声纹识别系统所带来的严重安全威胁寻求检测和防御方案。ASVspoof 2019 包含伪造语音和录音重放两项子挑战。Wu 等人[44]提出了一种使用最大特征图(max-feature-map, MFM)激活函数的轻量级神经网络 Light CNN, 由于该框架具有提炼度高、空间占用小等特点[145], 被 ASVspoof 2019 挑战中的模型[146, 149]广泛使用。Gomez-Alanis 等人[148]通过融合 Light CNN[44]和基于门递归单元(Gated recurrent units, GRU)的 RNN, 提出了一种光卷积门控递归神经网络(Light convolutional gated recurrent neural network, LC-GRNN), 并将其作为深度特征提取器辅助完成分类器的训练。其中 LC-GRNN 既具有 Light CNN 在帧级别提取判别特征的能力, 又包含(基于 GRU 的)RNN 学习深层特征的能力。针对基于单一特征的虚假语音检测算法存在泛化性较差的问题, Li 等人[150]提出了一种基于多特征融合和多任务学习(Multiple features integration and multi-task learning, MFMT)的虚假语音检测框架。MFMT 所选取的特征主要有梅尔频率倒谱系数(Mel frequency cepstrum coefficient, MFCC)、常量 Q 倒谱系数(Constant Q cepstral coefficient, CQCC)和 FBank 等, 进一步基于蝶形单元(Butterfly unit, BU)完成多任务学习。上述方法在 ASVspoof 2019 所提供的数据集中均具有较高的检测准确率。

基于真伪音频频谱之间的差异性特征, Dessa 公司[141]构建了一种基于频谱图的深度伪造语音检测模型。频谱图是声音的视觉表示, 根据其条带的清晰程度能够判别目标音频是否为伪造。该模型采用了时间卷积网络(Temporal convolutional network, TCN)[142]架构: 首先将目标音频转换为频谱图, 作为模型的输入; 其次在频谱图的时间范围内执行卷积, 并使用屏蔽池以防止过度拟合; 最后, 将输出传递到一个密集层和一个 S 型的激活函数, 进而输出判别概率(介于 0(伪造)和 1(真实)之间)。该模型也在 ASVspoof 2019 中 Google 所提供的竞赛数据集[167]上进行了测试, 准确率可达到 90%。Dessa 公司现已将该模型开源[143], 期望以此来推动深度伪造语音检测技术的发展。

4 深度伪造内容数据集

目前, 深度伪造检测模型的训练和评估多依赖于大规模的深度伪造内容数据集, 数据集的质量直接影响着检测模型的准确率, 因此对高质量深度伪造视频数据集的需求不断增长。当前具有代表性的深度伪造内容数据集如表 2 所示。

1) 图像数据集

香港中文大学 Liu 等人[113]创建了 CelebA 数据集, 该数据集包含 10177 个名人的图像(人均约有 20 张), 共计 202599 张。此外, 该数据集提供了 5 个关于人脸的关键点坐标及 40 个属性特征。CelebA 常用于人脸检测模型的训练。

100K-Faces 数据集[152]包含基于 StyleGAN[153]生成的 100000 张深度伪造图像。

Neves 等人创建的 TPDNE 数据集[154]包含从网站[155]上收集的 150000 张的深度伪造图像, 这些图像基于在 FFHQ 数据库[156]上训练得到的 StyleGAN 模型生成。

Durall 等人创建的 Faces-HQ 数据集[112]包含 40000 高质量图像, 其中真实图像和伪造图像各占一半, 数据集共计 19GB。

SwapMe 和 FaceSwap 数据集[157]包含 1005 张人脸合成图像, 2300 张真实图像, 其中训练集包含 750 张合成图像, 1400 张真实图像; 测试集包括 300 张合成图像, 900 张真实图像。人脸合成图像由 SwapMe 和 Faceswap[14]应用程序生成(SwapMe 是一款 iOS 应用软件, 当前已在苹果官方应用市场下架)。

2) 视频数据集

UADFV 数据集[158]包含 49 个基于原生 YouTube 真实视频生成的深度伪造视频, 其中深度伪造视频基于 FakeApp[13]生成。

Khodabakhsh 等人[159]创建了假脸数据集 FFW(Fake face in the wild), 共包含 150 个分辨率大于 480p 的高质量视频, 其中有 50 个视频是由 FakeApp[13]生成的深度伪造视频。

DeepFake-TIMIT 数据集[160]包含 640 个基于 Vid-TIMIT 数据集[162]和 Faceswap-GAN[52]生成的深度伪造视频, DeepFake-TIMIT 被分为两个相等大小

表 2 深度伪造数据集

Table 2 Deepfake Datasets

类别	数据集	真实内容		伪造内容		发布日期
		个数	大小(单位: 帧)	个数	大小(单位: 帧)	
图像	*CelebA*[112]	*202599*	/	*0*	/	2015
图像	*100K-Faces*[151]	*0*	/	*100k*	/	2018
图像	*TPDNE*[153]	*0*	/	*150k*	/	2019
图像	*Faces-HQ*[111]	*20k*	/	*20k*	/	2019
图像	*SwapMe 和 FaceSwap*[156]	*2300*	/	*1005*	/	*2018*
视频	*UADFV*[157]	*49*	*17.3k*	*49*	*17.3k*	2018
视频	*FFW*[158]	*0*	/	*50*	*97.0k*	2018
视频	*DF-TIMIT-LQ*[159]	*320*	*34.0k*	*320*	*34.0k*	2018
	DF-TIMIT-HQ[159]			*320*	*34.0k*	
视频	*FaceForensics ++*[113]	*1000*	*509.9k*	*1000*	*509.9k*	2019
视频	*Google/Jigsaw 检测*[160]	*363*	*315.4k*	*3068*	*2242.7k*	2019
视频	*Facebook 检测*[162]	*1131*	*488.4k*	*4113*	*1783.3k*	2019
视频	*Celeb-DF*[163]	*590*	*225.4k*	*5639*	*2116.8k*	2019
视频	*DeeperForensics-1.0*[164]	50000	14666.7k	10000	2933.3k	2020
音频	*Baidu 克隆语音*[165]	*10*	/	*124*	/	2018
音频	*Google 伪造语音*[166]	*内容来自英文报纸, 且包含 68 种不同口音的伪造语音*				2019

的子集: 64×64 像素的 DF-TIMIT-LQ 数据集和 128×128 像素的 DF-TIMIT-HQ 数据集。

FaceForensics++数据集[114]包含 1000 个原生 YouTube 真实视频和基于 Faceswap[14]生成的相同数量的深度伪造视频。

Google/Jigsaw 数据集[161]包含 3068 个深度伪造视频, 这些伪造视频基于性别, 年龄和种族各异的 28 个人的 363 个原始视频生成。合成算法的具体细节尚未公开。

Facebook 的 Dolhansky 等人基于 66 个不同性别, 年龄和种族的 1131 个原始视频构建了深度伪造检测挑战数据集[163], 包含 4113 个深度伪造视频。该数据集由两种不同的合成算法创建, 算法具体细节目前也尚未公开。

Celeb-DF 数据集[164]基于 59 位性别, 年龄和种族各异的采访者的 590 个原始视频创建的 5639 个深度伪造视频, 算法细节尚未公开。

基于发布时间和算法的综合分析, Li 等人[164]提出将 UADFV 数据集[158] 、DeepFake-TIMIT 数据集[160]和 FaceForensics++数据集[114]归为第一代深度视频伪造数据集, 而 Google/Jigsaw 数据集[161], Facebook 深度伪造检测挑战数据集[163]和 Celeb-DF[164]为第二代数据集。相比于第一代数据集, 第二代数据集的数量和质量均有较大的提升。

DeeperForensics-1.0 数据集[165]包含 60000 个视频(其中 10000 个为伪造视频), 1760 万帧图片, 在规模上是现有同类型数据集的 10 倍。Jiang 等人[165]为了检测当前主流深度伪造视频数据集的质量, 召集了 100 位计算机视觉领域的专家对其进行了真实性评估。专家意见总共分为 5 个等级: 等级 1-强烈不认可, 等级 2-稍微不认可, 等级 3-中立, 等级 4-稍微认可, 等级 5-强烈认可。进一步, Jiang 等人将等级 4、等级 5 的得分总和定义为该数据集整体的"真实度", 其评估结果如表 3 所示, 相比于 UADFV、FaceForensics ++等主流数据集, Celeb-DF 和 DeeperForensics-1.0 数据集更加真实。

3) 音频数据集

百度 Arik 等人[166]构建了一个克隆音频数据集, 该数据集包含 10 个真实音频样本, 124 个伪造音频样本(其中 120 个克隆音频, 4 个变形音频)。

Google 新闻团队及 AI 研究部门合作构建了一个语音语料库[167], 该语料库包含 Google 基于其 TTS 模型合成的数千个短语。这些短语取自于英语报纸上的文章, 涵盖了 68 种不同区域的口音。

表 3 主流视频伪造数据集真实度评测

Table 3 Reality Evaluation of Mainstream Video Falsification Datasets

数据集	等级					真实度/%
	1	2	3	4	5	
UADFV[157]	*29.2*	*36.0*	*20.7*	8.9	*5.2*	*14.1*
DF-TIMIT[159]	*31.4*	*31.4*	*24.8*	*9.6*	*2.7*	*12.3*
FaceForensics ++[113]	*46.8*	*31.4*	*13.4*	4.4	*4.0*	*8.4*
Facebook 检测[162]	*25.4*	*29.7*	*22.0*	*11.9*	*11.1*	*23.0*
Celeb-DF[163]	*5.6*	*14.8*	*18.6*	*24.2*	*36.9*	*61.0*
DeeperForensics-1.0[164]	*4.3*	*8.9*	*22.6*	*29.8*	*34.3*	*64.1*

5 相关技术

深度伪造检测技术多基于深度学习模型构建, 所以针对深度学习模型本身或围绕深度学习模型的相关研究进展(成果)在一定程度上会影响深度伪造检测和防御技术的研究方向, 本章将对影响深度伪造检测和防御的相关技术进行简单介绍。

5.1 人工智能对抗技术

人工智能对抗技术的主要研究目标是通过构建对抗样本实现对特定人工智能模型的攻击。对抗样本是一种通过指定算法处理的内容, 通过在原始样本加入部分扰动, 进而使目标模型出错[168]。针对分类模型, 对抗样本的目标是改变其对于原有样本的分类[169]; 针对检测模型, 对抗样本的目标则是使其无法发现特定目标或对特定目标识别错误, 如针对智能语音系统[170]或物理目标检测系统的对抗攻击[171]。

由于视觉深度伪造模型依赖于面部检测器对人脸进行识别与定位, 故可针对面部检测模型易于受到对抗样本的攻击的弱点, 通过在原始真实数据内容中添加对抗扰动, 干扰面部检测结果, 从而实现对视觉深度伪造的防御。文献[172]提出对原始数字内容增加对抗性扰动, 使伪造模型的面部检测器无法准确定位原始目标人脸区域, 进而影响深度伪造效果。文献[173]则提出了 Patch-IoU 方法, Patch-IoU 是一种对抗性补丁的优化方法, 优化后的对抗性补丁会被面部检测器错误的识别成人脸, 从而阻碍深度伪造模型的面部检测器对人脸的正常检测。以上方法均为深度伪造防御技术的研究提供了新的思路。

由于现有深度伪造内容检测模型鲁棒性不足, 易受到对抗样本的攻击, 因而可以将人工智能对抗技术引入深度伪造内容检测模型的训练阶段中来提高算法的鲁棒性, 使其能够更好的抵御对抗样本的攻击。文献[174]提出了一种在无标签数据上进行对抗训练的方法, 文献[175]则提出了一种将对抗训练扩展到大型模型和数据集的方法, 这些方法对深度伪造检测模型的训练提供了新的方法。

5.2 数字水印技术

数字水印(Digital watermark)是一种将特定标识信息嵌入图像、音视频等数字载体中, 但不影响其使用价值且不易被人的直觉系统直接察觉的方法[176]。数字水印可以用于验证信息的真实性和完整性, 目前主要应用于防伪溯源、欺诈和篡改检测、版权保护等领域。

由于现有的深度伪造内容检测算法多依赖于大规模真实和伪造数据集的训练来保障检测算法的准确度, 真实数据集素材多来自于网络, 而未来存在于网络上的大量数据可能是深度伪造的内容, 故无法保障所收集真实数据集的真实性。通过鼓励机构或个人基于数字水印工具在其所创作或发布的原始(真实)图像、音视频等数字内容中嵌入创建者或所有者的特定标示信息, 进而保障深度伪造检测算法所依赖数据集的可信度。文献[177]提出了一种基于多级矢量量化的多用途图像水印算法, 文献[178]提出了两种在 RGB 颜色空间中具备鲁棒和不可见特性的图像水印方法, 文献[179]提出了一种基于 QR 码的视频水印算法, 文献[180]提出了一种鲁棒、透明和大容量的盲音频水印技术, 这些数字水印技术均可为数字内容可信体系的构建提供技术保障。

5.3 模型可解释性技术

目前, 针对深度学习模型的可解释性研究工作仍然处于初级阶段。深度学习模型, 作为一个具有强大功能的“黑盒子”, 由于其参数规模庞大、神经元结构复杂及内部状态的不透明性, 使得对其内在机理的理解和研究工作面临巨大的挑战。目前, 业界对可解释性还没有公认的统一性定义, 研究者基于其各自的角度赋予了“可解释性”不同的定义, 如 Miller 等人[180]将可解释性定义为“人们可以理解决策原因的程度”, 而 Kim 等人[181]将其定义为“人类可以一致地预测模型结果的程度”。直观来讲, 可解释性是回答针对一个特定的输入, “黑盒子”内部是

如何“运作”得到相应的输出的问题。现有的深度学习模型推动了无人驾驶、语音交互等生活服务类产业的快速发展, 但由于其在可解释性研究工作上的困难, 使人们对基于深度学习模型的应用产生了信任危机。所以, 模型具有可解释性是人们对智能化系统产生信任的基石。

由于现有的深度伪造内容检测方法多基于深度学习模型构建, 如果模型内在的检测机理未知且无法解释, 则无法对检测方法的性能和机理进行评测和解释, 使得人们对现有伪造检测方法的信任度降低, 甚至影响伪造数字内容取证的可靠性。Google 大脑团队的 Olah 等人[183]对可解释性技术的原则进行了总结: 首先需要理解隐藏层的工作原理, 因为神经网络的强大之处在于其隐藏层, 每一层对输入都有一个新的表示形式; 其次是对激活向量认知, 理解在同一空间位置一起激活的互相连接的神经元组, 通过互相连接的神经元组分割网络可以从更简单的抽象层次来理解其功能; 最后是决策形成的原因, 即深度学习模型如何对单个部分进行组装而得出最后的决策(输出)。文献[184]进而提出了一种基于概念激活向量(Concept activation vectors, CAVs)的线性可解释性方法 TCAV(Testing with CAV)[185], 使用方向导数来量化模型预测对 CAVs 学习到的底层高级概念的敏感度。TCAV 使用显著性图完成了对出租车概念的理解。针对现有可解释性方法无统一的评测方法, 文献[186]提出了一种名为 PDR(predictive, descriptive, relevant)的可解释性评估框架, 即基于预测准确性、描述准确性和相关性、相对于人类判断的相关性来实现对可解释性方法的评估。基于模型可解释性技术, 对伪造检测方法中所使用的深度学习模型进行可解释性分析, 构建能够自我解释的智能化伪造检测方法, 提高检测可信度, 进而保障未来对于伪造数字内容取证的可靠性。

6 总结与展望

近年来, 深度伪造内容生成技术依托于深度学习, 正在以前所未有的速度发展, 不仅可以生成换脸图像、模仿真人说话和动作、表情等, 还可以创造出现实中不存在的人物, 真正意义上实现了“以假乱真”。恶意的深度伪造内容可借助全球普及化的互联网和移动互联网实现快速传播, 甚至被作为新一轮信息战的武器, 对网络安全、数据安全、信息安全、隐私安全和国防安全均带来了巨大威胁。但是深度伪造技术本身并无恶意性, 可用于电影特效制作、虚拟角色创建和语音模拟等领域, 所以通过研究切实有效的深度伪造内容检测和防御方法, 制定相关政策法规等方式防止深度伪造内容生成技术被恶意使用是深度伪造内容检测和防御的未来研究方向。

当前, 主流的深度伪造内容检测技术主要依赖两点: 基于伪造内容数据集完成对模型检测器的训练, 以及基于生物信息不一致性实现对伪造内容的判别。针对第一点, 当伪造图像、音视频等内容来源于新型伪造内容生成技术时, 或训练数据集不包含某一种内容伪造技术生成的样本时, 则检测器对该类伪造内容无法实现良好的检测效果; 针对第二点, 受限于当前的伪造内容生成技术水平, 伪造图像、音视频等内容存在生物信息以及习惯等不一致性, 如眨眼频率、手部动作等, 基于这些差异化特征可实现伪造内容的检测, 然而随着生成技术水平的不断提升, 深度伪造内容将趋近逼真, 基于生物不一致性的检测也将变得越来越困难。

针对以上深度伪造内容检测技术面临的挑战以及难点问题, 我们可以从多角度出发, 探索针对深度伪造内容的检测。

1) 构建数字内容可信体系

从保障数字内容来源出发, 可以尝试将数字水印技术、基于区块链的数字内容溯源技术等方法引入互联网真实数字内容的保护机制中, 构建针对互联网中数字内容的可信体系, 有力保障数字媒体内容的来源安全。

2) 研究高效、准确的深度伪造内容检测技术

一方面, 探索多方位特征融合提取算法, 如在听觉方面可以专注于语速、声纹、频率分布、音素的过渡与连接等特征的融合, 在视觉方面尝试综合面部表情与动作、五官位置、纹理特征、眨眼和心跳频率, 肤色或光照变化, 以及脸部位置或轮廓变化等特征, 进而获取多角度深度融合特征。引入元学习和小样本学习等技术, 摆脱深度伪造检测模型对大规模、高质量深度伪造数据集的过度依赖, 有效降低模型训练成本。将对抗样本攻击加入到检测模型的训练阶段中来提高算法的鲁棒性, 使其能够更好的抵御对抗样本的攻击。对深度伪造检测方法中使用的深度模型进行可解释性分析, 在评测模型脆弱性的同时提高检测方法的可信度。

另一方面, 虽然深度伪造内容生成技术层出不穷, 但是“魔高一尺、道高一丈”: 由于深度伪造内容生成技术依赖于 GAN 实现, 故可将博弈论思想引入深度伪造检测模型的构建中: 通过实时关注 GAN 技术的最新研究成果, 基于其主动构建深度伪造生成模型, 创建对应的伪造数据集; 进而基于该伪造数

据集重新对检测模型进行训练, 修正模型缺陷、提升检测能力, 使其可以成功检测基于新型 GAN 模型生成的伪造内容, 从而有力保障互联网中数据生态的可信度。

3) 制定深度伪造相关的法律法规

通过在国家及各级政府层面制定深度伪造相关的法律法规, 约束深度伪造技术的使用范围; 鼓励数字内容分享平台及社交媒体制定相应的发布规则, 对发布恶意伪造内容的用户进行永久封号, 情节严重则可直接移交政府相关机构进行处理。将深度伪造装进制度的“笼子”里, 使其更好的为人类的美好生活服务。

通过法规约束和技术保障并行推进的模式可有效防止深度伪造技术被恶意滥用, 使其成为新一轮人工智能浪潮得以飞速前进的助推剂。

致　谢　本课题得到国家重点研发项目(No.2016QY04W0805), 国家自然科学基金项目(No.61728209), 中国科学院青年创新促进会, 北京市科技新星计划, 北京市自然科学基金项目(No.JQ18011), 国家前沿科技创新项目(No. YJKYYQ20170070)资助。

参考文献

[1] Deloitte, Global artificial intelligence industry whitepaper, 2019.

[2] Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case, Catherine Stupp, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, Aug. 2019.

[3] Deepfake: The Good, The Bad and the Ugly, Nahua Kang, https://medium.com/twentybn/deepfake-the-good-the-bad-and-the-ugly-8b261ecf0f52, May. 2019.

[4] Deepfake video of Facebook CEO Mark Zuckerberg posted on Instagram, Queenie Wong, https://www.cnet.com/news/deepfake-video-of-facebook-ceo-mark-zuckerberg-posted-on-instagram/, Jun. 2019.

[5] 9 deepfake examples that terrified and amused the internet, Joseph Foley, https://www.creativebloq.com/features/deepfake-examples, Sept. 2019.

[6] Brundage M, Avin S, Clark J, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation[EB/OL]. 2018: arXiv:1802.07228[cs.AI]. https://arxiv.org/abs/1802.07228.

[7] The Bizarre and Terrifying Case of the “Deepfake” Video that Helped Bring an African Nation to the Brink, Ali Breland, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/, Mar. 2019.

[8] Porn Producers Offer to Help Hollywood Take Down Deepfake Videos, Janko Roettgers, https://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/, Aug. 2019.

[9] S.3805-Malicious Deep Fake Prohibition Act of 2018, Ben Sasse, https://www.congress.gov/bill/115th-congress/senate-bill/ 3805.

[10] Korshunov P, Marcel S. Vulnerability Assessment and Detection of Deepfake Videos[C]. *2019 International Conference on Biometrics (ICB)*, June 4-7, 2019. Crete, Greece. Piscataway, NJ: IEEE, 2019: 1-6.

[11] Fake videos could be the next big problem in the 2020 elections, Grace Shao, https://www.cnbc.com/2019/10/15/deepfakes-could-be-problem-for-the-2020-election.html, Oct. 2019.

[12] Deepfakes and the new disinformation war: The coming age of post-truth geopolitics, Robert Chesney and Danielle Citron, https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war, Jan/Feb. 2019.

[13] FakeApp, https://www.malavida.com/en/soft/fakeapp/.

[14] Faceswap, https://github.com/deepfakes/faceswap.

[15] DeepFaceLab, https://github.com/iperov/DeepFaceLab.

[16] RealTalk, https://medium.com/dessa-news/real-talk-speech-synthesis-5dd0897eef7f.

[17] MelNet, https://sjvasquez.github.io/blog/melnet/.

[18] ZAO, https://apkproz.com/app/zao.

[19] LIST OF DEEPFAKE TOOLS, https://vuild.com/deep-fake- tools, Jul. 2019.

[20] DeepNude, https://www.deepnude.com/

[21] The State of Deepfakes: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, Sept. 2019.

[22] Ethics guidelines for trustworthy AI, High-Level Expert Group on Artificial Intelligence, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[23] H.R.3600-Deepfakes Report Act of 2019, Yvette Clarke, https://www.congress.gov/116/bills/hr3600/BILLS-116hr3600ih.pdf.

[24] California makes ‘deepfake’ videos illegal, but law may be hard to enforce, Kari Paul, https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce.

[25] 网络信息内容生态治理规定, http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm.

[26] Regulations on Ecological Governance of Network Information Content, http://www.cac.gov.cn/2019-12/20/c_157837515950930 9.htm.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial nets[J]. *Advances in Neural Information Processing Systems (NIPS)*, 2014: 2672–2680.

[28] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based Learning Applied to Document Recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.

[29] Pearlmutter B A. Gradient Calculations for Dynamic Recurrent Neural Networks: A Survey[J]. *IEEE Transactions on Neural Networks*, 1995, 6(5): 1212-1228.

[30] Kingma D P, Welling M. Auto-Encoding Variational Bayes [EB/OL]. 2013: arXiv:1312.6114[stat.ML]. https://arxiv.org/abs/1312.6114.

[31] Bowles, Samuel., Microeconomics: behavior, institutions, and evolution[M]. Princeton University Press, 2009.

[32] LeCun, Yann, Yoshua Bengio et al. Deep learning[J], *nature*, 2015, 521(7553): 436-444.

[33] Schuster M, Paliwal K K. Bidirectional Recurrent Neural Networks[J]. *IEEE Transactions on Signal Processing*, 1997, 45(11): 2673-2681.

[34] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.

[35] Korshunova I, Shi W Z, Dambre J, et al. Fast Face-Swap Using Convolutional Neural Networks[C]. *2017 IEEE International Conference on Computer Vision (ICCV)*, October 22-29, 2017. Venice. Piscataway, NJ: IEEE, 2017: 3677-3685.

[36] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing Obama[J]. *ACM Transactions on Graphics*, 2017, 36(4): 1-13.

[37] Deep face swap with GAN, (CS 230) Chi Wang, and Jinil Jing, http://cs230.stanford.edu/projects_spring_2019/reports/18681213.pdf., 2019.

[38] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T. and Nahavandi, S., Deep Learning for Deepfakes Creation and Detection, arXiv preprint arXiv:1909.11573, 2019.

[39] DeepFake_tf, https://github.com/StromWine/DeepFake_tf.

[40] Dfaker, https://github.com/dfaker/df.

[41] Wen Z Q, Li K H, Huang Z, et al. Improving Deep Neural Network Based Speech Synthesis through Contextual Feature Parametrization and Multi-Task Learning[J]. *Journal of Signal Processing Systems*, 2018, 90(7): 1025-1037.

[42] Huang J, Li Y, Tao J H, et al. End-to-End Continuous Emotion Recognition from Video Using 3D Convlstm Networks[C]. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 15-20, 2018. Calgary, AB. Piscataway, NJ: IEEE, 2018: 6837-6841.

[43] Zheng Y B, Wang X, He L, et al. Forward-Backward Decoding for Regularizing End-to-End TTS[C]//Interspeech 2019, ISCA: ISCA, 2019: 234-243.

[44] Wu X, He R, Sun Z N, et al. A Light CNN for Deep Face Representation with Noisy Labels[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(11): 2884-2896.

[45] Huang H B, He R, Sun Z N, et al. Wavelet-SRNet: A Wavelet-Based CNN for Multi-scale Face Super Resolution[C]. *2017 IEEE International Conference on Computer Vision (ICCV)*, October 22-29, 2017. Venice. Piscataway, NJ: IEEE, 2017: 456-465.

[46] Rui Huang, Shu Zhang, Tianyu Li, and Ran He, "Be-yond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," IEEE International Conference on Computer Vision (ICCV), 2017.

[47] Huang R, Zhang S, Li T, et al. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis[C]. *2017 IEEE International Conference on Computer Vision (ICCV)*, October 22-29, 2017. Venice. Piscataway, NJ: IEEE, 2017: 567-576.

[48] Yibo Hu, Xiang Wu, Bing Yu, et al. Pose-Guided Photorealistic Face Rotation[C]. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018: 2345-2376.

[49] Lingxiao Song, Zhihe Lu, Ran He, et al. Geometry guided adversarial facial expression synthesis[C]. *ACM Multimedia Conference on Multimedia Conference (CORR)*, 2018: 346-365.

[50] Li P P, Hu Y B, Li Q, et al. Global and Local Consistent Age Generative Adversarial Networks[C]. *2018 24th International Conference on Pattern Recognition (ICPR)*, August 20-24, 2018. Beijing. Piscataway, NJ: IEEE, 2018: 345-365.

[51] He R, Zheng W S, Tan T N, et al. Half-Quadratic-Based Iterative Minimization for Robust Sparse Representation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(2): 261-275.

[52] Lingxiao Song, Man Zhang, Xiang Wu, et al. Adversarial discriminative heterogeneous face recognition[C]. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018: 387-395.

[53] faceswap-GAN, https://github.com/shaoanlu/faceswap-GAN.

[54] CycleGAN, https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

[55] VGGFace, https://github.com/rcmalli/keras-vggface.

[56] FaceNet, https://github.com/davidsandberg/facenet.

[57] Zhang K P, Zhang Z P, Li Z F, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. *IEEE Signal Processing Letters*, 2016, 23(10): 1499-1503.

[58] Ohad Fried, Ayush Tewari, Michael Zollhöfer, et al. Chris-tian Theobalt and Maneesh Agrawala[EB/OL]. Text-based Editing of Talking-head Video, 2019: arXiv preprint arXiv:1906.01524.

[59] Yoshimura, Takayoshi., Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems, PhD diss, Nagoya Institute of Technology, 2002.

[60] Tokuda K, Nankaku Y, Toda T, et al. Speech Synthesis Based on Hidden Markov Models[J]. *Proceedings of the IEEE*, 2013, 101(5): 1234-1252.

[61] Ping, Wei, Peng, Kainan, Gibiansky, et al. Deep Voice 3: 2000-speaker neural text-to-speech[C]. *ICLR*, 2018: 568-576.

[62] Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, et al. Deep voice: Real-time neural text-to-speech[C]. *34th International Conference on Machine Learning*, 2017: 195-204.

[63] Gibiansky, A., Arik, S., Diamos, G., et al. Deep voice 2: Multi-speaker neural text-to-speech[C]. *In Advances in neural information processing systems*, 2017: 2962-2970.

[64] Pascual S, Bonafonte A, Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network[C]. *Interspeech 2017*, ISCA: ISCA, 2017: 235-243.

[65] Donahue, C., McAuley, J., Puckette, M., Synthesizing audio with generative adversarial networks[EB/OL]. 2018: arXiv preprint arXiv:1802.04208.

[66] Ren Y, Tan X, Qin T, et al. Almost Unsupervised Text to Speech and Automatic Speech Recognition[EB/OL]. 2019: arXiv:1905.06791[eess.AS]. https://arxiv.org/abs/1905.06791.

[67] Vasquez, S., Lewis M.. MelNet: A Generative Model for Audio in the Frequency Domain[EB/OL]. 2019: arXiv preprint arXiv:1906.01083.

[68] Mehri, S., Kumar, K., Gulrajani, et al. Samplernn: An unconditional end-to-end neural audio generation model[EB/OL]. 2016: arXiv preprint arXiv:1612.07837.

[69] Oord A V D, Dieleman S, Zen H G, et al. WaveNet: A Generative Model for Raw Audio[EB/OL]. 2016: arXiv:1609.03499[cs.SD]. https://arxiv.org/abs/1609.03499.

[70] Oord, A. V. D., Kalchbrenner, N., Kavukcuoglu, K. Pixel recurrent neural networks[EB/OL]. 2016: arXiv preprint arXiv:1601.06759

[71] Dahl R, Norouzi M, Shlens J. Pixel Recursive Super Resolution[C]. *2017 IEEE International Conference on Computer Vision (ICCV)*, October 22-29, 2017. Venice. Piscataway, NJ: IEEE, 2017: 5439-548.

[72] Reed, S., van den Oord, A., Kalchbrenner, et al. Parallel multiscale autoregressive density estimation[J]. *34th International Conference on Machine Learning*, 2017,70: 2912-2921.

[73] Menick, J. , Kalchbrenner, N., Generating high fidelity images with subscale pixel networks and multidimensional upscaling[EB/OL]. 2018: arXiv preprint arXiv:1812.01608.

[74] Sutskever, I., Vinyals, O. , Le, Q. V., Sequence to sequence learning with neural networks[C]. *NIPS*, 2014:45-65.

[75] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, et al. Multi-task Sequence to Sequence Learning[C]. *ICLR*, 2016:478-486.

[76] Jia Y, Weiss R, Biadsy F D, et al. Direct Speech-to-speech Translation with a Sequence-to-sequence Model[EB/OL]. 2019: arXiv:1904.06037[cs.CL]. https://arxiv.org/abs/1904.06037.

[77] Audio samples from 'Direct speech-to-speech translation with a sequence-to-sequence model', Ye Jia, Ron J. Weiss, Fadi Biadsy, et al., https://google-research.github.io/lingvo-lab/translatotron/, 2019.

[78] Agarwal S, Varshney L. Limits of Deepfake Detection: A Robust Estimation Viewpoint[EB/OL]. 2019: arXiv:1905.03493[cs.LG]. https://arxiv.org/abs/1905.03493.

[79] Maurer U M. Authentication Theory and Hypothesis Testing[J]. IEEE Transactions on Information Theory, 2000, 46(4): 1350-1356.

[80] Nataraj L, Mohammed T, Chandrasekaran S, et al. Detecting GAN Generated Fake Images Using Co-occurrence Matrices[EB/OL]. 2019: arXiv:1903.06836[cs.CV]. https://arxiv.org/abs/1903.06836.

[81] Wang S Y, Wang O, Owens A, et al. Detecting Photoshopped Faces by Scripting Photoshop[EB/OL]. 2019: arXiv:1906.05856[cs.CV]. https://arxiv.org/abs/1906.05856.

[82] McCloskey, S. , Albright, M., Detecting GAN-generated Imagery using Color Cues[EB/OL]. 2018: arXiv preprint arXiv:1812.08247.

[83] Li, H., Li, B., Tan, S. , Huang, J, Detection of deep network generated images using disparities in color components[EB/OL], 2018: arXiv preprint arXiv:1808.07276.

[84] Mo H X, Chen B L, Luo W Q. Fake Faces Identification Via Convolutional Neural Network[C]. *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security* - IH&MMSec '18, June 14-22, 2018. Innsbruck, Austria. New York, USA: ACM Press, 2018: 346-365.

[85] Dang L, Hassan S, Im S, et al. Deep Learning Based Computer Generated Face Identification Using Convolutional Neural Network[J]. *Applied Sciences*, 2018, 8(12): 2610.

[86] François Chollet, Xception: Deep learning with depthwise separable convolutions[C]. *IEEE conference on computer vision and pattern recognition*, 2017: 246-253.

[87] Wang, R., Ma, L., Juefei-Xu, et al. Fakespotter: A simple baseline for spotting ai-synthesized fake faces[EB/OL]. 2019: arXiv preprint arXiv:1909.06122.

[88] Zhang Y, Zheng L L, Thing V L L. Automated Face Swapping and Its Detection[C]. *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, August 4-6, 2017. Singapore. Piscataway, NJ: IEEE, 2017: 15-19.

[89] Bay, Herbert, Tinne Tuytelaars et al. Surf: Speeded up robust features[C]. *In European conference on computer vision*, 2006: 404-417.

[90] Bag-of-words model, Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Bag-of-words_model&oldid=928524653, 2019.

[91] Wang X, Thome N, Cord M. Gaze Latent Support Vector Machine for Image Classification Improved by Weakly Supervised Region Selection[J]. *Pattern Recognition*, 2017, 72: 59-71.

[92] Bai S. Growing Random Forest on Deep Convolutional Neural Networks for Scene Categorization[J]. *Expert Systems With Applications*, 2017, 71: 279-287.

[93] Zheng L L, Duffner S, Idrissi K, et al. Siamese Multi-layer Perceptrons for Dimensionality Reduction and Face Identification[J]. *Multimedia Tools and Applications*, 2016, 75(9): 5055-5073.

[94] Huh M, Liu A, Owens A, et al. Fighting Fake News: Image Splice Detection Via Learned Self-Consistency[M]. Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 106-124.

[95] Yang X, Li Y Z, Lyu S W. Exposing Deep Fakes Using Inconsistent Head Poses[C]. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 12-17, 2019. Brighton, United Kingdom. Piscataway, NJ: IEEE, 2019: 8261-8265.

[96] Chih-Chung Hsu, Yi-Xiu Zhuang, Chia-Yen Lee, Deep Fake Image Detection based on Pairwise Learning[J]. *International Journal of Molecular Sciences* , 2020, 10(1): 370.

[97] Radford A, Metz L K, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[EB/OL]. 2015: arXiv:1511.06434[cs.LG]. https://arxiv.org/abs/1511.06434.

[98] Martin Arjovsky, Soumith Chintala , Léon Bottou, Wasserstein generative adversarial networks[C]. *International conference on machine learning*, 2017:378-382.

[99] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, et al. Improved training of wasserstein gans[C]. *Advances in neural information processing systems*, 2017:398-392.

[100] Xudong Mao, Qing Li, Haoran Xie, et al. Least squares generative adversarial networks[C]. *IEEE International Conference on Computer Vision*, 2017:29-34.

[101] Tero Karras, Timo Aila, Samuli Laine et al. Progressive growing of gans for improved quality, stability, and variation[EB/OL]. 2017: arXiv preprint arXiv:1710.10196.

[102] Han Zhang, Ian Goodfellow, Dimitris Metaxas et al. Self-attention generative adversarial networks[EB/OL]. 2018: arXiv preprint arXiv:1805.08318.

[103] Miyato T, Kataoka T, Koyama M, et al. Spectral Normalization for Generative Adversarial Networks[EB/OL]. 2018: arXiv:1802.05957[cs.LG]. https://arxiv.org/abs/1802.05957.

[104] Sumit Chopra, Raia Hadsell , Yann LeCun, Learning a similarity metric discriminatively, with application to face verification[C]. *CVPR*, 2005:38-42.

[105] Jane Bromley, Isabelle Guyon, Yann LeCun, et al. Signature verification using a 'siamese' time delay neural network[C]. *In Advances in neural information processing systems*, 1994: 737-744.

[106] Farid H. Image Forgery Detection[J]. *IEEE Signal Processing Magazine*, 2009, 26(2): 16-25.

[107] Marra F, Gragnaniello D, Cozzolino D, et al. Detection of GAN-Generated Fake Images over Social Networks[C]. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, April 10-12, 2018. Miami, FL. Piscataway, NJ: IEEE, 2018: 356-362.

[108] Hsu C C, Lee C Y, Zhuang Y X. Learning to Detect Fake Face Images in the Wild[C]. *2018 International Symposium on Computer, Consumer and Control (IS3C)*, December 6-8, 2018. Taichung, Taiwan, China. Piscataway, NJ: IEEE, 2018: 465-468.

[109] Zhang X, Karaman S, Chang S F. Detecting and Simulating Artifacts in GAN Fake Images[EB/OL]. 2019: arXiv:1907.06515[cs.CV]. https://arxiv.org/abs/1907.06515.

[110] Marra F, Gragnaniello D, Verdoliva L, et al. Do GANs Leave Artificial Fingerprints?[C]. *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, March 28-30, 2019. San Jose, CA, USA. Piscataway, NJ: IEEE, 2019: 245-251.

[111] Michael Albright , Scott McCloskey, Source generator attribution via inversion[C]. *CVPR Workshop on Media Forensics*, 2019: 88-92.

[112] AutoGAN, https://github.com/TAMU-VITA/AutoGAN.

[113] Durall R, Keuper M, Pfreundt F, et al. Unmasking DeepFakes with Simple Features[EB/OL]. 2019: arXiv:1911.00686[cs.LG]. https://arxiv.org/abs/1911.00686.

[114] Ziwei Liu,, Ping Luo, Xiaogang Wang et al. Deep learning face attributes in the wild[C]. *IEEE international confer-ence on computer vision*, 2015: 3730-3738.

[115] Rössler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images[EB/OL]. 2019: arXiv:1901.08971[cs.CV]. https://arxiv.org/abs/1901.08971.

[116] Afchar D, Nozick V, Yamagishi J, et al. MesoNet: A Compact Facial Video Forgery Detection Network[C]. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, December 11-13, 2018. Hong Kong, China. Piscataway, NJ: IEEE, 2018: 1-7.

[117] Sabir E, Cheng J X, Jaiswal A, et al. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos[EB/OL]. 2019: arXiv:1905.00582[cs.CV]. https://arxiv.org/abs/1905.00582.

[118] Huang G, Liu Z, van der Maaten L, et al. Densely Connected Convolutional Networks[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017. Honolulu, HI. Piscataway, NJ: IEEE, 2017: 4700-4707.

[119] Cho K, van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation[EB/OL]. 2014: arXiv:1406.1078[cs.CL]. https://arxiv.org/abs/1406.1078.

[120] Guera D, Delp E J. Deepfake Video Detection Using Recurrent Neural Networks[C]. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, November 27-30, 2018. Auckland, New Zealand. Piscataway, NJ: IEEE, 2018: 1-6.

[121] Li Y Z, Chang M C, Lyu S W. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, December 11-13, 2018. Hong Kong, China. Piscataway, NJ: IEEE, 2018: 1-7.

[122] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 7-12, 2015. Boston, MA, USA. Piscataway, NJ: IEEE, 2015: 2625-2634.

[123] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos[EB/OL]. 2014: arXiv:1406.2199[cs.CV]. https://arxiv.org/abs/1406.2199.

[124] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 770-778.

[125] Li Y Z, Lyu S W. Exposing DeepFake Videos by Detecting Face Warping Artifacts[EB/OL]. 2018: arXiv:1811.00656[cs.CV]. https://arxiv.org/abs/1811.00656.

[126] Zhou P, Han X T, Morariu V I, et al. Two-Stream Neural Networks for Tampered Face Detection[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 21-26, 2017. Honolulu, HI, USA. Piscataway, NJ: IEEE, 2017: 1831-1839.

[127] Zakharov E, Shysheya A, Burkov E, et al. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models[EB/OL]. 2019: arXiv:1905.08233[cs.CV]. https://arxiv.org/abs/1905.08233.

[128] Nguyen H H, Yamagishi J, Echizen I. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos[C]. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 12-17, 2019. Brighton, United Kingdom. Piscataway, NJ: IEEE, 2019: 2307-2311.

[129] Sara Sabour, Nicholas Frosst , Geoffrey E. Hinton, Dynamic routing between capsules[C]. *In Advances in neural information processing systems*, 2017: 3856-3866.

[130] Ivana Chingovska, André Anjos , Sébastien Marcel, On the effectiveness of local binary patterns in face anti-spoofing[C]. *BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, 2012: 1-7.

[131] Thies J, Zollhofer M, Stamminger M, et al. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 2387-2395.

[132] Rössler A, Cozzolino D, Verdoliva L, et al. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human

Faces[EB/OL]. 2018: arXiv:1803.09179[cs.CV]. https://arxiv.org/abs/1803.09179.

[133] Rahmouni N, Nozick V, Yamagishi J, et al. Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks[C]//2017 IEEE Workshop on Information Forensics and Security (WIFS), December 4-7, 2017. Rennes. Piscataway, NJ: IEEE, 2017: 1-6.

[134] Haiying Guan, Mark Kozak, Eric Robertson, et al. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation[C]. *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019: 63-72.

[135] Matern F, Riess C, Stamminger M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations[C]. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, January 7-11, 2019. Waikoloa Village, HI, USA. Piscataway, NJ: IEEE, 2019: 83-92.

[136] Marissa Koopman, Andrea Macarulla Rodriguez , Zeno Geradts, Detection of Deepfake Video Manipulation[C], *IMVIP*, 2018: 23-28.

[137] Luka J, Fridrich J, Goljan M. Digital Camera Identification from Sensor Pattern Noise[J]. *IEEE Transactions on Information Forensics and Security*, 2006, 1(2): 205-214.

[138] Kurt Rosenfeld , Husrev Taha Sencar, A study of the robustness of PRNU-based camera identification[J]. *Media Forensics and Security*, 2009, 7254:245-234..

[139] Hasan H R, Salah K. Combating Deepfake Videos Using Blockchain and Smart Contracts[J]. *IEEE Access*, 2019, 7: 41596-41606.

[140] IPFS powers the Distributed Web, https://ipfs.io/.

[141] ASVspoof 2019, https://www.asvspoof.org/.

[142] Detecting Audio Deepfakes With AI, https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35.

[143] Bai S J, Kolter J Z, Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling[EB/OL]. 2018: arXiv:1803.01271[cs.LG]. https://arxiv.org/abs/1803.01271.

[144] fake-voice-detection, https://github.com/dessa-public/fake-voice-detection.

[145] Digger-Deepfake Detection, https://twitter.com/deepfakedigger?lang=en.

[146] Himawan I, Madikeri S, Motlicek P, et al. Voice Presentation Attack Detection Using Convolutional Neural Networks[M]. Handbook of Biometric Anti-Spoofing. Cham: Springer International Publishing, 2019: 391-415.

[147] Lavrentyeva G, Novoselov S, Malykh E, et al. Audio Replay Attack Detection with Deep Learning Frameworks[C]. *Interspeech 2017*, ISCA: ISCA, 2017: 56-61.

[148] Monteiro J, Alam J, Falk T H. End-To-End Detection of Attacks to Automatic Speaker Recognizers with Time-Attentive Light Convolutional Neural Networks[C]. *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, October 13-16, 2019. Pittsburgh, PA, USA. Piscataway, NJ: IEEE, 2019: 678-681.

[149] Gomez-Alanis A, Peinado A M, Gonzalez J A, et al. A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection[C]. *Interspeech 2019*, ISCA: ISCA, 2019: 1068-1072.

[150] Zeinali H, Stafylakis T, Athanasopoulou G, et al. Detecting Spoofing Attacks Using VGG and SincNet: BUT-Omilia Submission to ASVspoof 2019 Challenge[C]. *Interspeech 2019*, ISCA: ISCA, 2019: 245-251.

[151] Li R J, Zhao M, Li Z, et al. Anti-Spoofing Speaker Verification System with Multi-Feature Integration and Multi-Task Learning[C]. *Interspeech 2019*, ISCA: ISCA, 2019: 1048-1052.

[152] Google News Initiative, https://newsinitiative.withgoogle.com/.

[153] Unique, worry-free model photos, https://generated.photos/.

[154] Karras T, Laine S, Aila T M. A Style-Based Generator Architecture for Generative Adversarial Networks[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019. Long Beach, CA, USA. Piscataway, NJ: IEEE, 2019: 34-41.

[155] João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, et al. Real or Fake? Spoofing State-Of-The- Art Face Synthesis Detection Systems[EB.OL]. 2019: arXiv preprint arXiv:1911. 05351.

[156] Thispersondoesnotexist, https://thispersondoesnotexist.com/.

[157] Flickr-Faces-HQ Dataset (FFHQ), https: //github.com/NVlabs/ffhq-dataset.

[158] Zhou P, Han X T, Morariu V I, et al. Two-Stream Neural Networks for Tampered Face Detection[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 21-26, 2017. Honolulu, HI, USA. Piscataway, NJ: IEEE, 2017: 1831-1839.

[159] Li Y Z, Chang M C, Lyu S W. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, December 11-13, 2018. Hong Kong, China. Piscataway, NJ: IEEE, 2018: 23-30.

[160] Khodabakhsh A, Ramachandra R, Raja K, et al. Fake Face Detection Methods: Can they be Generalized?[C]. *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, September 26-28, 2018. Darmstadt. Piscataway, NJ: IEEE, 2018: 1-6.

[161] Pavel Korshunov , Sébastien Marcel, Deepfakes: a new threat to face recognition? assessment and detection[EB.OL].2018: arXiv preprint arXiv:1812.08685.

[162] Nicholas Dufour, Andrew Gully, Per Karlsson, et al.Deepfakes detection dataset by google & jigsaw.

[163] Sanderson C, Lovell B C. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference[M]. Advances in Biometrics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 199-208.

[164] Dolhansky B, Howes R, Pflaum B, et al. The Deepfake Detection Challenge (DFDC) Preview Dataset[EB/OL]. 2019: arXiv:1910.08854[cs.CV]. https://arxiv.org/abs/1910.08854.

[165] Li Y Z, Yang X, Sun P, et al. Celeb-DF: A New Dataset for DeepFake Forensics[EB/OL]. 2019: arXiv:1909.12962[cs.CR]. https://arxiv.org/abs/1909.12962.

[166] L. Jiang, W. Wu, R. Li, et al. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection[EB/OL].2020: arXiv preprint arXiv:2001.03024v1.

[167] Sercan Arik, Jitong Chen, Kainan Peng, et al. Neural voice cloning with a few samples[C]. *Neural Information Processing Systems(NIPS)*, 2018:234-241.

[168] Advancing research on fake audio detection, Daisy Stanton, https://www.blog.google/outreach-initiatives/google-news-initiative/advancing-research-fake-audio-detection/.

[169] Goodfellow, Ian J., Jonathon Shlens, et al. Explaining and harnessing adversarial examples[EB/OL]. 2014: arXiv preprint arXiv:1412.6572.

[170] Miyato, Takeru, Andrew M. Dai et al. Adversarial training methods for semi-supervised text classification[EB/OL].2016: arXiv preprint arXiv:1605.07725.

[171] Xuejing Yuan, Yuxuan Chen, Yue Zhao et al. Commandersong: A systematic approach for practical adversarial voice recognition[C]. *27th {USENIX} Security Symposium*, 2018: 49-64.

[172] Zhao, Y., Zhu, H., Liang, R., et al. Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors[C]. *ACM SIGSAC Conference on Computer and Communications Security*, 2019: 1989-2004.

[173] Li, Y., Yang, X., Wu, B. et al. Hiding Faces in Plain Sight: Disrupting AI Face Synthesis with Adversarial Perturbations[EB/OL].2019: arXiv preprint arXiv:1906.09288.

[174] Yang, X., Wei, F., Zhang, H. et al. Design and Interpretation of Universal Adversarial Patches in Face Detection[EB/OL]. 2019: arXiv preprint arXiv:1912.05021.

[175] Miyato T, Maeda S I, Koyama M, et al. Distributional Smoothing by Virtual Adversarial Examples[EB/OL]. 2015: arXiv:1507.00677[stat.ML]. https://arxiv.org/abs/1507.00677.

[176] Kurakin, Alexey, Ian Goodfellow et al. Adversarial machine learning at scale[EB/OL]. 2016: arXiv preprint arXiv:1611.01236.

[177] Cox I J, Miller M L, Bloom J A, et al. Steganography[M]. Digital Watermarking and Steganography. Elsevier, 2008: 425-467.

[178] Lu Z M, Xu D G, Sun S H. Multipurpose Image Watermarking Algorithm Based on Multistage Vector Quantization[J]. *IEEE Transactions on Image Processing*, 2005, 14(6): 822-831.

[179] Vaishnavi D, Subashini T S. Robust and Invisible Image Watermarking in RGB Color Space Using SVD[J]. *Procedia Computer Science*, 2015, 46: 1770-1777.

[180] G. Prabakaran, R. Bhavani, and M. Ramesh, "A robust QR-Code video watermarking scheme based on SVD and DWT composite domain," Pattern Recognition, Informatics and Mobile En-gineering (PRIME), pp. 251–257, 2013.

[181] Prabakaran G, Bhavani R, Ramesh M. A Robust QR-Code Video Watermarking Scheme Based on SVD and DWT Composite Domain[C]. *2013 International Conference on Pattern Recognition*, Informatics and Mobile Engineering, February 21-22, 2013. Salem. Piscataway, NJ: IEEE, 2013: 251-257.

[182] Hu H T, Hsu L Y. Robust, Transparent and High-capacity Audio Watermarking in DCT Domain[J]. *Signal Processing*, 2015, 109: 226-235.

[183] Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences[J]. *Artificial Intelligence*, 2019, 267: 1-38.

[184] Kim, Been, Rajiv Khanna, Oluwasanmi O. Koyejo., Examples are not enough, learn to criticize! criticism for interpretability[C]. *Advances in Neural Information Processing Systems*, 2016: 23-32.

[185] Olah, C., Satyanarayan, A., Johnson, et al., The building blocks of interpretability[J]. *Distill*, 2018, 3(3): 10.

[186] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)[EB/OL]. 2017: arXiv:1711.11279[stat.ML]. https://arxiv.org/abs/1711.11279.

[187] tcav, https://github.com/tensorflow/tcav.

[188] Murdoch W, Singh C D, Kumbier K, et al. Interpretable Machine Learning: Definitions, Methods, and Applications[EB/OL]. 2019: arXiv:1901.04592[stat.ML]. https://arxiv.org/abs/1901.04592.

梁瑞刚 于 2017 年在中国科学院大学计算机技术专业获得硕士学位。现在中国科学院大学信息工程研究所攻读博士学位。研究领域为软件与系统安全。研究兴趣包括: 软件分析与测试、人工智能安全。Email: liangruigang@iie.ac.cn

吕培卓 现为中国科学院大学信息工程研究所客座学生(2020 年直博生), 现在西安电子科技大学计算机科学与技术专业攻读学士学位。研究兴趣包括: 人工智能对抗技术, 隐私安全。Email: lpz13201827326@163.com

赵月 于 2017 年在中国科学院大学信号与信息处理专业获得硕士学位。现在中国科学院大学信息工程研究所攻读博士学位。研究领域为人工智能安全与深度学习模型对抗研究。Email: zhaoyue@iie.ac.cn

陈鹏 于 2016 年在山东大学(威海)计算机科学与技术专业获得学士学位。现在中国科学院信息工程研究所攻读博士学位。研究领域为计算机视觉、人工智能安全。研究兴趣包括: 目标检测、对抗样本、深度伪造的生成和检测。Email: chenpeng@iie.ac.cn