中国科学: 信息科学 2021 年 第 51 卷 第 9 期: 1451–1474
SCIENTIA SINICA Informationis
论文

可视身份深度伪造与检测

彭春蕾[1], 高新波[2*], 王楠楠[1*], 李洁[1]

1. 西安电子科技大学综合业务网理论及关键技术国家重点实验室, 西安 710071
2. 重庆邮电大学重庆市图像认知重点实验室, 重庆 400065
* 通信作者. E-mail: gaoxb@cqupt.edu.cn, nnwang@xidian.edu.cn

收稿日期: 2020–03–19; 修回日期: 2020–07–20; 接受日期: 2020–08–03; 网络出版日期: 2021–09–10

2030 新一代人工智能重大专项 (批准号: 2018AAA0103202)、国家自然科学基金 (批准号: 61806152) 和陕西省重点研发项目 (批准号: 2020ZDLGY08-08) 资助项目

摘要 随着深度学习技术在视频和图像生成领域的广泛应用, 视频和图像中的可视身份伪造, 特别是人脸伪造结果的逼真程度越来越高, 对于身份伪造数据的检测在国家安全和社会稳定等方面均具有重要的研究和应用价值, 近年来已成为研究的热点问题. 本文从有目标身份伪造和无目标身份伪造两个方面归纳和介绍了可视身份深度伪造的研究方法, 并从基于空域线索、时域线索的面向已知伪造类型检测方法、面向未知伪造类型的泛化能力研究, 以及面向对抗样本攻击的可信伪造检测研究等多个方面阐述了伪造检测的关键技术, 并在总结现有数据集和代表性算法的性能分析基础上, 进一步讨论了可视身份深度伪造与检测的关键问题和面临的挑战.

关键词 深度伪造, 人脸替换, 人脸编辑, 表情重演, 人脸生成, 伪造检测

1 引言

随着人工智能技术的发展, 特别是互联网和数字化产品的广泛应用, 视频和图像等可视媒体资料的获取、编辑和传输更加便捷, 然而随之引发的可视媒体身份伪造等安全问题也日益严重. 特别是利用深度学习等人工智能技术所生成的伪造人脸图像、伪造视频等虚假可视身份信息的仿真程度越来越高, 难以依靠人工或传统技术进行有效的检测甄别, 对维护国家安全和保障社会稳定带来严峻的挑战和威胁. 例如, 可视身份深度伪造技术最早引起公众和媒体的注意可以追溯到 2017 年 12 月, 一位名为 Deepfakes 的用户将一段伪造的不良视频上传到社交新闻网站 Reddit 上, 该视频利用深度学习技术将视频中女主角的人脸替换为明星人脸, 并且效果十分逼真, 引起公众和媒体的轰动. 随后在互联网上出现越来越多人脸替换后的不良视频, 以及一款名为 FakeAPP 的视频伪造软件. 然而这些伪造视频对被替换人脸图像的当事人肖像权、名誉权等均构成严重的伤害, 因而互联网上包括 Google 和 Twitter 在内的多个主流网站均对此类伪造不良视频进行了封禁和抵制.

引用格式: 彭春蕾, 高新波, 王楠楠, 等. 可视身份深度伪造与检测. 中国科学: 信息科学, 2021, 51: 1451–1474, doi: 10.1360/SSI-2020-0064
Peng C L, Gao X B, Wang N N, et al. Deep visual identity forgery and detection (in Chinese). Sci Sin Inform, 2021, 51: 1451–1474, doi: 10.1360/SSI-2020-0064

可视身份深度伪造与检测技术对于维护国家形象和安全具有重要的意义. 例如, 身份伪造技术可能被不法分子出于政治动机而恶意利用, 对政治人物进行恶意抹黑和损害国家形象. 政治人物往往经常出现在公众场合, 这也给不法分子获取政治人物的大量图像、视频等可视身份信息带来诸多便利 [1]. 在 2018 年 4 月, 伪造美国前总统奥巴马发言侮辱特朗普的视频在互联网上流传, 该伪造视频对美国政府形象造成了极大的不良影响. 再比如, 可视身份伪造技术容易被恶意用于伪造国家领袖或政治人物的演讲视频, 从而宣扬一些虚假讲话内容或外交政策, 使得伪造技术成为挑拨国家矛盾和影响国家安全的工具. 在 2019 年 1 月, 非洲加蓬总统的一段新年致辞讲话视频被反对势力声称属于伪造视频且认为该国总统已丧失行动能力或已去世从而引发兵变. 此外, 可视身份深度伪造技术还可能被间谍或恐怖分子恶意利用. 例如, 2019 年 6 月美联社新闻网站报道了利用人工智能技术生成的虚假人脸形象凯蒂琼斯 (Katie Jones) 在职业社交网站领英上进行身份注册, 并利用这一虚假身份与政客联系以从事间谍活动[1)]. 另一方面, 恐怖分子通常通过开展恐怖主义活动以引发社会恐慌和达到自身目的. 随着视频、图像伪造技术的日趋成熟, 很容易被恐怖分子所利用, 制作虚假恐怖行为视频并散播到互联网上, 以激起更多人的恐慌或仇恨, 从而可以在不采取实际恐怖行动的同时达到预期的恐怖效果, 给国家安全和稳定带来巨大隐患.

可视身份深度伪造与检测技术的研究还有助于保障社会稳定和司法公正. 由于伪造技术存在成本低和技术容易获取等特点, 不法分子可能制作并发布虚假信息, 例如伪造警方暴力执法视频, 来煽动普通民众, 破坏社会稳定和扰乱社会秩序. 尽管目前互联网上类似涉暴伪造视频或图像并不多, 但其潜在危害巨大. 如何尽早发现并识别伪造视频图像内容, 将类似隐患遏制在萌芽状态, 对保障社会稳定具有重要意义. 另一方面, 利用深度伪造技术制作虚假的不良视频图像, 对女性肖像权和名誉权产生严重危害. 在 2019 年 6 月, 一款基于深度伪造技术开发的 DeepNude 伪造软件能自动消除视频中人物的衣物穿着, 对女性权益和隐私具有严重危害; 2019 年 8 月, 换脸软件 ZAO 可以通过在手机上简单操作实现换脸短视频, 随后该软件被指责侵犯用户肖像隐私并下架. 2019 年 9 月, 美国 Boston University 法律系教授 Danielle K. Citron 在 TED 演讲中[2)] 以虚假伪造视频对印度记者 Rana Ayyub 的工作与生活的影响为例, 探讨了可视身份伪造技术对个人隐私的侵犯和社会信任体系的破坏. 类似伪造技术的恶意应用还可能在商业竞争中被用于制作和散播竞争对手的虚假新闻, 损害其商业声誉, 以及实施敲诈勒索等. 诈骗团伙可以利用该技术伪造被诈骗对象亲属的视频片段, 以实施网络诈骗行为. 深度伪造技术对司法公正也具有日趋严峻的威胁 [2]. 违法乱纪人员可以利用伪造技术生成虚假视频片段, 一方面恶意用于嫁祸他人以造成冤假错案, 另一方面可能用于法庭举证环节以逃脱法律制裁.

除了上述恶意应用外, 可视身份深度伪造技术在艺术创作和教育行业等领域也具有一定的积极帮助作用. 例如, 可以利用虚假人脸生成技术用于电影制作中的人物形象创作. 在教育行业中, 可以利用人脸面部表情重演等技术还原历史人物的动态样貌视频, 用于爱国历史教育过程中, 提高学生对历史典故的学习兴趣和代入感. 因此, 开展伪造与检测对抗技术的研究以协助检测视频和图像的真伪性, 在近几年来成为人工智能领域的研究热点.

早期的图像伪造和篡改技术主要依赖图像编辑软件, 而伪造结果往往存在明显的失真, 易于通过人眼或传统的多媒体取证技术进行篡改检测 [3]. 自 2015 年 5 月 *Nature* 杂志发表深度学习论文 [4] 以来, 深度学习技术被应用到包括视频图像生成在内的各个领域中. 2019 年 3 月, *Science* 杂志发表论文 [5] 探讨了图像和文本伪造行为在医疗行业的影响, 随后在 2019 年 10 月 *Nature* 杂志发表论文 [6] 讨论人工智能技术是否可能失控, 警示语音和人脸深度伪造技术可能会被应用到安防监控、控制和

1) http://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

2) http://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy.

行为篡改编辑等场景. 2019 年 11 月, 国家互联网信息办公室发布了《网络音视频信息服务管理规定》[3], 以应对伪造视频图像信息的潜在威胁. 该规定中指出自 2020 年 1 月起互联网上发布基于深度学习等新兴技术制作的非真实音视频信息应当以显著方式予以标识, 并且不得制作、发布和传播虚假新闻信息, 同时网络音视频服务信息提供者应当部署非真实音视频鉴别技术以及时发现伪造虚假音视频. 人工可以依据伪造视频图像中的细节不一致进行伪造判断. 例如, 伪造人脸图像中容易存在对称性异常的现象, 如图像中眼镜框、耳坠等面部装饰物的不对称性、头发不连贯, 以及背景变形等异常. 随着深度学习技术的不断发展, 伪造图像的逼真程度也日益提高, 使得依靠人工进行伪造检测变得更加困难. 为了应对这一挑战, 越来越多的研究团队开展了身份伪造与检测研究. 例如, 2019 年 9 月 Facebook 公司在著名的机器学习竞赛网站 Kaggle 上发起了深度伪造检测挑战竞赛 (Deepfake detection challenge)[4]. 在 2020 年 2 月, Google 子公司 Jigsaw 发布了一款名为 Assembler 的伪造图像检测平台[5] 以帮助新闻从业者快速检测一幅图像是否属于伪造图像.

目前针对可视身份深度伪造问题已有大量的研究工作, 然而面向伪造视频和图像的检测研究目前仍处于起步阶段. 本文主要针对基于视频和图像的可视身份深度伪造与检测研究, 将现有可视身份深度伪造方法按照是否具有伪造目标划分为有目标身份伪造和无目标身份伪造. 按照伪造检测的研究思路不同, 本文将身份伪造检测方法分为基于空域线索、时域线索地面向已知伪造类型检测方法、面向未知伪造类型的泛化能力研究方法, 以及面向对抗样本攻击的可信伪造检测研究. 在此基础上, 本文从实验角度在不同类方法及同类伪造检测方法之间进行性能对比分析, 重点分析了基于空域线索和基于时域线索的检测方法之间的异同. 最后, 本文讨论了可视身份深度伪造与检测领域当前存在的局限性和未来的发展趋势. 下面将依次展开介绍.

2 可视身份深度伪造方法

划分可视身份深度伪造方法的主要依据是伪造过程中是否有具体的伪造目标, 因而可分为有目标身份伪造 (target-specific face forgery) 和无目标身份伪造 (target-generic face forgery), 如图 1 所示. 有目标身份伪造方法通常在视频或图像伪造过程中, 将伪造目标的身份或属性信息输入到模型中, 实现特定目标身份的视频或图像伪造. 该伪造形式可能被用于进行特定身份的伪装与假冒, 例如可以伪造特定身份人员的动态视频用于对该目标人员进行恶意抹黑或虚假信息的传播. 无目标身份伪造方法通常以随机变量作为输入信息, 生成现实世界中不存在的虚假人脸图像, 生成过程中没有特定的伪造目标身份. 该伪造形式可以结合虚假音频或虚假文本等其他类型的伪造数据, 设定虚假的网络人物进行政治或经济欺骗, 或被恶意用于虚假新闻的杜撰导致潜在破坏性的负面社会影响. 因此, 可视身份伪造技术的恶意应用对维护国家和社会的稳定具有严峻的威胁.

2.1 有目标可视身份伪造

有目标可视身份伪造方法, 根据伪造程度的不同, 可以分为人脸替换 (face swap) 和人脸编辑 (face manipulation) 两个类别. 人脸替换技术通常对视频或图像中的人脸面部区域整体进行替换, 而人脸编辑技术往往在保持人脸面部大部分特征不变的前提下编辑修改特定的属性, 例如头发颜色、性别、年龄、面部表情动作等. 其中, 人脸编辑技术可进一步分为人脸属性编辑 (attribute manipulation)、人

3) http://www.cac.gov.cn/2019-11/29/c_1576561820967678.htm.
4) https://ai.facebook.com/datasets/dfdc/.
5) https://projectassembler.org/.

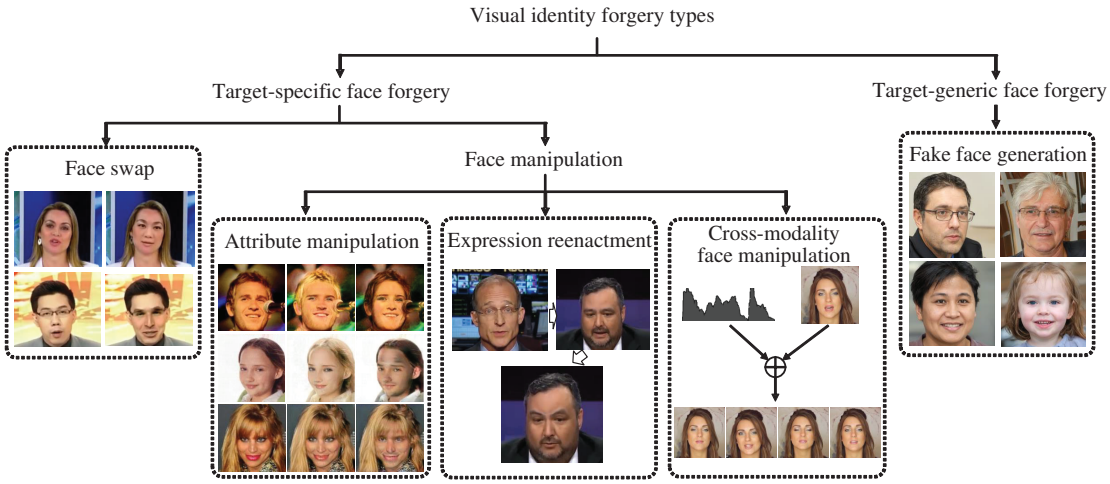


图 1 (网络版彩图) 可视身份深度伪造类型示例

Figure 1 (Color online) Example types of deep face forgery

脸表情重演 (expression reenactment) 和基于语音或文本描述的跨模态人脸编辑 (cross-modality face manipulation).

2.1.1 人脸替换

人脸替换通常是将一张人脸图像中的面部区域替换到另一张人脸图像中, 而图像的背景和头发等区域保持不变. 传统的人脸替换技术主要采用计算机图形学的方法实现[6]. 近年来随着大规模人脸数据集的发布和不断改进的深度学习模型, 基于深度学习特别是生成对抗网络 (generative adversarial networks, GAN) 的人脸替换方法取得了以假乱真的替换效果. 例如, 开源代码 Faceswap[7]和 DeepFaceLab[8] 被广泛用来制作人脸替换视频和构建人脸替换数据集.

比利时 Ghent University 和 Twitter 研究团队 [7] 提出一种基于卷积神经网络的快速人脸替换方法. 该方法将人脸替换问题看作图像风格转换任务, 通过将一幅图像渲染成另一幅图像风格的方式实现. 该方法存在和开源代码 Faceswap 与 DeepFaceLab 类似的不足之处, 即需要对每组人脸替换图像对单独训练深度模型. 以色列 Nirkin 等 [8] 提出利用全卷积神经网络实现包括人脸分割、人脸替换和人脸身份识别在内的多个任务. 在人脸替换过程中, 该方法首先从图像中进行人脸分割, 随后拟合到三维人脸模型上进行姿态、表情等风格迁移后, 嵌入到目标背景图像中实现人脸替换. Natsume 等 [9] 提出一种基于人脸面部区域划分的生成对抗网络模型用于人脸替换. 在该方法基础上, Natsume 等 [10] 进而提出一种只对人脸面部待替换区域提取隐变量空间特征, 在保持头发背景区域不变的情况下进行人脸替换的方法, 以应对不同光照和姿态角度下的人脸替换问题.

上述人脸替换方法存在只能面向训练集中已有的人脸图像对的不足, 文献 [11] 提出一种面向开放条件的人脸替换方法. 该方法只需要一张人脸图像提供身份信息, 另外一张图像提供人脸姿态、表情、光照和背景等信息, 即可实现对任意身份的人脸替换功能. Nirkin 等 [12] 提出一种不需要针对特定对象进行模型训练的人脸替换方法, 利用循环神经网络对人脸姿态、表情和身份进行建模, 借助三角剖

6) http://github.com/MarekKowalski/FaceSwap.
7) http://github.com/deepfakes/faceswap.
8) http://github.com/iperov/DeepFaceLab.

分技术对人脸面部区域进行连续插值和面部表情重演, 进而利用图像修复模型处理人脸图像中的遮挡现象. 微软亚洲研究院 [13] 最近提出一种高质量和对遮挡鲁棒的人脸替换算法 FaceShifter, 通过属性编码器提取多级人脸属性特征以挖掘目标人脸图像中的属性信息, 并提出一种启发式的深度神经网络模型自适应修复图像中的遮挡区域.

2.1.2 人脸编辑

人脸编辑, 也称为人脸篡改, 指的是修改图像或视频中的人脸特征, 如头发颜色、性别、年龄、面部表情等. 该类方法主要受图像翻译技术所启发, Isola 等 [14] 率先提出图像翻译方法 Pix2Pix. 该方法利用成对训练数据进行训练, 可用于图像风格迁移、图像彩色化等图像翻译任务中. 随后 Zhu 等 [15] 进一步提出基于循环一致性对抗网络的图像翻译方法 CycleGAN, 解决了非成对训练数据的场景下图像翻译问题, 具有更强的泛化能力. 本文根据人脸编辑过程中目标身份信息的来源不同, 分为人脸属性编辑、人脸表情重演和跨模态人脸编辑展开介绍.

(1) 人脸属性编辑. 人脸属性编辑方法通常致力于修改人脸图像的特定属性特征. 北京富士通研发中心的研究人员 [16] 将人脸图像属性编辑前后的差异定义为残差图像, 并利用生成对抗网络进行建模. 该模型中包含两个图像生成网络用于实现人脸属性编辑过程及其逆过程, 另外包含一个判别网络用于判断属性编辑后的生成图像与真实图像之间的相似性, 并利用对偶学习理论进行网络优化, 实现佩戴眼镜、修改表情的人脸属性编辑效果. Facebook 研究团队 [17] 提出利用编码 – 解码器的网络结构, 通过在隐变量空间中分解图像的显著性信息和属性信息实现人脸属性编辑.

考虑到 Pix2Pix [14] 和 CycleGAN [15] 等图像翻译模型只能应用于两种图像类型之间的转换, 由于人脸属性类型较多, 基于成对类型的图像翻译模型训练效率低, 韩国高丽大学研究团队 [18] 提出一种多领域图像翻译的统一框架 StarGAN. 该方法可以仅训练一个统一的生成对抗网络模型, 用于多种类型的图像翻译任务. 针对该模型在每个属性下的生成结果多样性不足的问题, 韩国 Naver 公司研究人员 [19] 近期提出改进算法 StarGANv2, 通过将原 StarGAN 模型中的各领域类标替换为领域内的风格编码向量, 以实现多样化的人脸属性编辑.

中国科学院山世光教授团队 [20] 同样提出一种仅训练单一网络模型实现多种人脸属性编辑任务的算法 AttGAN. 同 StarGAN [18] 模型相比, 该方法利用编码 – 解码器的结构对人脸属性和隐层表达之间的联系进行建模, 并通过属性分类约束来确保人脸属性编辑结果的准确性. 哈尔滨工业大学左旺孟教授团队 [21] 提出改进版本的 STGAN 模型, 只考虑属性向量之间的差异而不将整个属性向量作为类标, 并在编码 – 解码器的结构中增加选择性输入单元, 提高了人脸属性编辑的精度和伪造图像的质量. 北京交通大学刘一副研究员团队 [22] 提出直接在编码 – 解码器的隐层表达空间中的属性相关区域内添加人脸属性类别约束, 然后通过属性分类器来保障合成结果中包含所需要修改的人脸属性.

腾讯优图团队 [23] 提出一种通用且灵活的高质量人脸属性编辑网络 Facelet-Bank, 进而与香港中文大学、Adobe 研究院、字节跳动人工智能实验室 [24] 合作提出一种基于语义部件分解的人脸属性编辑方法, 以解决现有方法存在的属性编辑结果难以控制调整的不足. 该方法将人脸属性分解为多个语义部件, 其中每个语义部件对应人脸图像的特定区域, 这样不仅有助于用户精确地控制调整属性编辑的结果, 还可以方便地移除合成结果中多余的属性编辑效果.

针对高分辨率人脸图像的属性编辑问题, 上述方法由于缺乏充足的训练数据, 往往难以取得令人满意的效果. 为此, 商汤科技研究团队 [25] 提出一种可加性焦距变分自编码器, 通过对图像重建和相对熵损失进行弱监督训练, 实现高分辨率人脸图像的属性修改. 商汤科技与香港中文大学 [26] 提出一种具有交互能力的人脸图像编辑方法 MaskGAN, 解决了已有人脸属性编辑方法缺乏用户交互能力的

不足. 该团队同时发布了一个大规模人脸语义分割数据集 CelebAMask-HQ 用于人脸语义分割与编辑研究.

上述现有方法在人脸属性编辑过程中均需要指定具体的目标身份属性信息, 然而对一些特殊的人脸属性 (例如光照程度、细节纹理或形状) 难以用数字化的属性标签进行定量刻画. 因此, Google 人工智能实验室与 Boston University [27] 联合提出一种基于示例图像指导的图像属性编辑方法 PuppleGAN. 该方法不需要指定明确的属性标签, 只需要提供一些示例图像来演示期望的属性编辑效果, 即可实现人脸属性编辑任务.

(2) 人脸表情重演. 人脸表情重演任务主要指的是将目标视频中的人脸面部表情与输入源视频中的人脸表情保持一致. 德国 Friedrich-Alexander-Universität Erlangen-Nürnberg 的 Thies 等 [28] 首次提出利用 RGB-D 深度传感器的人脸面部表情迁移和重演方法. 然而该方法需要专门的传感器采集人脸深度信息, 为此 Thies 等 [29] 提出只需要视频数据进行实时人脸面部表情重演算法 Face2Face. 该方法克服了对人脸深度信息的依赖, 是当前人脸表情重演的基准算法之一. 在 Face2Face 算法基础上, Thies 等 [30] 进一步提出利用三维建模的方式将人脸面部区域的表情重演拓展到包含上半身区域内的头部姿态、眼睛注视角度的实时重演. 类似的, 德国 Max-Planck-Gesellschaft 的 Kim 等 [31] 同样提出包含头部姿态、眼睛注视方向和眨眼等动作在内的人脸表情重演方法.

考虑到直接在像素空间进行人脸表情重演容易产生失真现象, 由人脸关键点组成的面部轮廓图像被广泛用来辅助人脸表情重演过程 [32]. 例如, 商汤科技研究团队 [33] 提出将人脸图像转化到关键点组成的人脸轮廓隐空间上, 进而利用人脸轮廓图像进行端到端的表情重演. Otberdout 等 [34] 提出利用希尔伯特 (Hilbert) 超球面对人脸表情变化过程中的人脸关键点信息进行建模, 进而生成表情运动模型后增加纹理信息实现人脸表情重演. Huang 等 [35] 提出一种两级生成对抗网络算法, 首先合成线条画形式的人脸表情初始轮廓, 在此基础上合成高清人脸图像.

为了提高人脸表情重演过程中的表情控制能力, University of Maryland, College Park 研究人员 [36] 提出一种可控制的人脸表情生成方法. 该方法设计了人脸面部表情控制单元, 以控制人脸表情合成过程中的面部表情动作幅度. Pumarola 等 [37] 提出基于解剖学中面部肌肉运动单元的生成对抗网络模型 GANimation, 通过将人脸面部表情动作分解为多个运动单元, 控制每个运动单元的激活程度实现自动的人脸表情重演. 最近, Tripathy 等 [38] 提出一种可解释和可控制的人脸表情重演算法 ICface, 该方法同样利用了面部运动单元和头部姿态角度来表示人脸面部表情属性, 在实现人脸表情重演任务的同时, 还可以人工控制表情和头部姿态的变化程度.

为了解决基于少量人脸数据甚至从单张图像进行人脸表情重演的问题, 三星人工智能研究中心 [39] 提出基于小样本学习和元学习的人脸表情重演方法. 韩国 Hyperconnect 公司 [40] 同样利用了小样本学习的思想进行人脸表情重演, 以解决人脸脸型差异较大、身份未知或者头部姿态变化较大时人脸表情重演效果较差的问题. 该方法通过图像注意力机制、特征对齐和关键点转换等模块, 在保留目标人脸身份特征的同时完成人脸表情重演.

(3) 跨模态人脸编辑. 人脸编辑的目标身份信息来源还可能是音频片段或文本段落等跨模态信息, 例如, University of Washington 研究团队 [41] 提出一种根据音频片段合成动态人脸视频的算法. 为了简化问题, 该方法只关注嘴巴周边区域来伪造奥巴马演讲特定对话的虚假视频. Massachusetts Institute of Technology 人工智能实验室 [42] 提出基于音频合成人脸图像的方法 Speech2Face, 通过采集上百万音视频数据进行深度神经网络的训练, 以学习音频和人脸视频之间的映射关系. 类似地, Carnegie Mellon University 研究团队 [43] 也提出从音频片段中提取身份信息进行人脸合成的方法.

除了以单独的音频数据作为身份信息来源外, Imperial College London 和三星人工智能研究中

心 [44] 提出综合单张人脸图像和一段演讲音频作为输入, 生成与音频同步的人脸伪造视频. University of Oxford 的 Jamaludin 等 [45] 同样以静态人脸图像和一段音频作为模型输入, 分别通过身份编码器和音频编码器进行编码, 然后输入到解码器中, 实现基于音频数据的视频伪造. 最近, 中国科学院赫然研究团队与商汤科技 [46] 合作提出一种新的基于音频片段的人脸视频编辑方法. 考虑到在缺乏大规模训练数据的情况下难以直接学习映射关系, 且不同音频和人脸图像存在巨大的多样性, 该方法提出将人脸视频分解为由表情、几何形状和姿态在内的多个子空间, 然后将输入音频数据转化到与语音相关的表情子空间中, 而几何形状与姿态子空间保持不变, 即可完成基于音频的人脸视频编辑任务.

文本段落也可以作为目标身份来源指导人脸编辑. Stanford University 的 Fried 等 [47] 提出一种基于演讲文本内容修改的人脸视频编辑方法. 给定一段演讲视频和对应的文本段落, 该方法首先将人脸视频中的语音信息进行分解, 对于给定的文本内容修改操作, 从输入视频中寻找出与修改文本发音嘴型相似的视频片段生成伪造人脸图像, 并嵌入到原视频中实现视频中的人脸编辑伪造. 四川大学研究团队 [48] 提出一种基于生成对抗网络的文本到人脸图像生成方法, 借助长短时记忆网络所组成的文本编码器从文本段落中提取语义特征, 然后利用卷积神经网络所构成的图像解码器将语义特征转化为合成图像, 从而伪造出与文本描述对应的人脸图像.

2.2 无目标可视身份伪造

上述身份伪造方法通常依据人脸属性标签、人脸表情视频或音频文本等作为伪造过程中的目标身份信息, 而无目标身份伪造方法通常在没有身份指信息指导下进行人脸图像伪造, 从而生成现实世界中完全不存在的虚假人脸. 该类方法最早开始于 Ian Goodfellow 所提出的生成对抗网络 GAN 方法 [49]. 随后, Radford 等 [50] 利用卷积神经网络替换原始生成对抗网络模型中的多层感知机, 提出了深度卷积生成对抗网络模型 DCGAN 以提高生成图像的质量. Miyato 等 [51] 将谱理论应用到生成对抗网络训练中, 提出谱归一化的模型训练方法 SNGAN, 以解决梯度消失导致训练不稳定的问题. Ian Goodfellow 团队 [52] 在 SNGAN 模型基础上将人眼视觉注意机制引入进来, 提出基于视觉注意驱动的模型 SAGAN, 取得了更好的合成图像质量.

英伟达研究团队 [53] 于 2018 年提出一种基于渐进式训练的生成对抗网络模型 ProGAN, 该方法从 4×4 像素尺度开始训练, 在训练过程中逐步增大图像分辨率, 最终可以生成 1024×1024 分辨率的高清虚假人脸图像. 随后, 英伟达研究团队于 2019 年借鉴图像风格迁移的思路, 提出一种基于图像风格的生成对抗网络框架 StyleGAN [54]. 该方法是在 ProGAN 模型的基础上, 以无监督学习的形式生成高质量人脸图像的同时对图像细节进行控制, 以提高伪造人脸图像的逼真程度. 在该方法基础上, 英伟达研究团队 [55] 近期进一步分析讨论了 StyleGAN 模型中存在的不足, 并从模型结构和训练方式上进行改进提出了 StyleGAN2 算法. Ivan Braun 等利用 GAN 模型 [49] 和 StyleGAN 模型 [54] 生成大量伪造人脸图像并按照性别、年龄、种族、头发颜色、表情等属性分类在网站上[9]进行展示. 为了让更多的人意识到虚假人脸伪造技术达到的逼真程度, 以及警示人们所看到的人脸图像有可能是伪造的, University of Washington 的 Jevin West 等创建了一个在线伪造人脸辨别测试网站[10], 所展示的真实人脸来源于高清人脸数据集 Flickr-Faces-HQ (FFHQ) [54], 而展示的虚假人脸图像来自基于 StyleGAN2 算法生成并展示伪造人脸图像的网站[11].

最近, DeepMind 团队 [56] 通过对网络结构进行调整, 在更大的训练参数模型和训练数据集上提出

9) http://generated.photos/faces/.
10) http://www.whichfaceisreal.com/.
11) http://www.thispersondoesnotexist.com/.

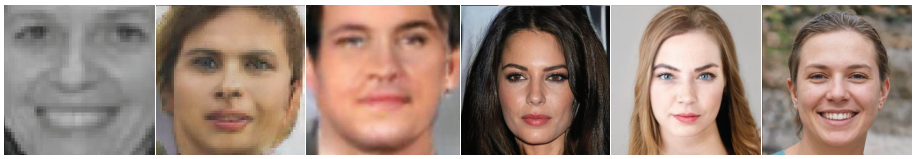


图 2 (网络版彩图) 虚假人脸合成结果示例

Figure 2 (Color online) Examples of target-generic face forgery

一种高分辨率和高质量的图像伪造算法 BigGAN. 该算法可以在 128×128, 256×256, 512×512 分辨率图像数据上直接进行训练, 并能够提高合成结果的保真度和多样性. 图 2 展示了上述介绍的虚假人脸伪造结果示例. 从图中可以看出, 随着网络结构的改进和训练数据规模的增大, 虚假人脸伪造质量逐步提高, 目前已达到人工难以辨别的逼真程度. 关于生成对抗网络的更多方法介绍可以参考综述论文 [57~59] 进行进一步了解.

3 身份伪造检测方法

目前, 面向可视身份伪造数据的检测研究仍处于起步阶段. 早期的伪造检测方法主要针对图像编辑软件的篡改检测, 其伪造痕迹相对比较明显, 本文主要讨论基于深度学习的可视身份伪造检测技术. 华南理工大学胡永建教授研究团队 [60] 对 5 种早期的人脸视频伪造检测方法进行了性能比较. 按照检测对象的类型不同, 澳大利亚 Nguyen 等 [61] 将伪造检测方法分为伪造图像检测和伪造视频检测, 其中后者又进一步分为单帧伪造检测和基于多帧时域特征的伪造检测. 该分类方式过于复杂, 且基于单帧信息的伪造视频检测方法与伪造图像检测方法往往存在一定的相似性. 按照检测对象的伪造程度不同, Tolosana 等 [62] 将现有伪造方法分为整幅人脸合成、人脸替换、人脸属性编辑和人脸表情编辑 4 个类别. 然而在该分类方式中, 人脸表情编辑与人脸属性编辑均属于人脸编辑范畴, 例如人脸编辑算法 StyleGAN [54] 可以同时实现人脸发色、年龄等属性编辑和人脸表情编辑的功能. Verdoliva [63] 将深度伪造问题与图像风格转换、图像修复, 以及反取证技术等问题包含到可视媒体安全研究当中. 考虑到本文主要面向可视身份伪造的检测研究, 如图 3 所示, 本文根据伪造检测研究的侧重点不同, 将现有方法划分为面向已知伪造检测的模型性能研究、面向未知伪造类型的泛化能力研究和面向对抗样本攻击的可信检测研究 3 个类别进行介绍. 其中, 在面向已知伪造类型的检测方法中根据伪造类型的特点所设计的检测线索有所差异, 可以进一步划分为基于空域线索的检测方法和基于时域线索的检测方法. 下面将进行详细介绍.

3.1 面向已知伪造类型的模型性能研究

3.1.1 基于空域线索的检测方法

早期的伪造检测方法主要基于传统机器学习算法实现. 例如, Zhang 等 [64] 首次提出借鉴传统人脸识别的思路进行伪造检测, 通过在人脸图像上提取人脸局部特征并构建词袋模型后, 利用支持向量机、随机森林和多层感知机等分类器进行分类, 以判断输入人脸图像是否属于伪造图像. 深圳大学谭舜泉教授团队 [65] 提出分别基于侵入式和非侵入式的方案检测图像是否由 GAN 模型伪造生成, 其中侵入式方案借助 GAN 模型中的判别模块进行检测, 而非侵入式方案采取图像质量评价指标、GAN 评分指标或利用卷积神经网络提出特征后进行检测. Korshunov 等 [66] 率先公开了一个小规模深度伪造

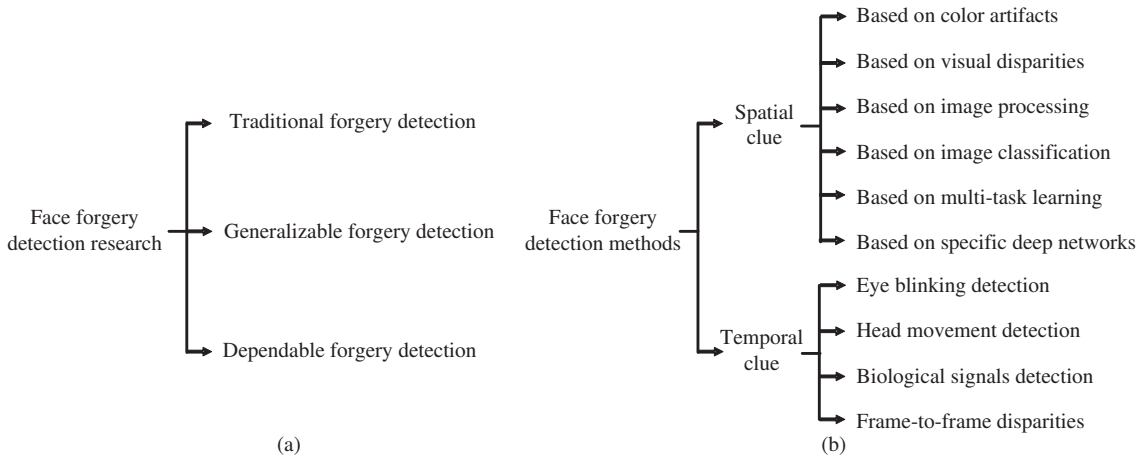


图 3 伪造检测方法分类

Figure 3 Classification of forgery detection methods. (a) Research on face forgery detection; (b) classification on face forgery detection

视频数据集 DeepfakeTIMIT, 其中伪造视频人脸替换代码[12)]生成, 并在该数据集上验证了包括主成分分析、线性判别分析、图像质量评价和支持向量机在内的一系列传统机器学习算法对于人脸替换伪造的检测能力. 表 1 [67~77] 对基于空域线索的代表性伪造检测方法进行了总结分析.

考虑到图像伪造方法大多在生成过程中未考虑颜色空间的约束, 因此可以检测图像中颜色失真进行伪造检测. 深圳大学黄继武教授团队 [67] 提出将人脸图像转换到不同颜色空间如 HSV 或 YCbCr 上, 利用在不同颜色通道之间的不一致性来判断该图像是否被伪造. McCloskey 等 [78] 利用图像中的颜色饱和程度和曝光度进行伪造人脸图像检测. Nataraj 等 [79] 提出在 3 个颜色通道上分别提取灰度共生矩阵等纹理特征作为检测线索. 考虑到上述基于颜色失真的方法易受到图像滤波等操作的影响, 四川大学王宏霞教授团队 [68] 提出在多个颜色空间提取特征后进一步进行特征融合的方法, 对每个颜色分量进行高通滤波后提取图像残差信息, 进而提取深度特征并进行拼接融合, 最后利用随机森林分类器进行伪造检测.

除颜色失真外, 可视身份伪造结果中还经常存在其他一些视觉细节的不一致现象, 因而可以通过检测视觉细节一致性特征作为伪造辨识的依据. 例如, 人脸替换结果中往往存在被替换人脸面部区域与背景之间的不协调现象, 而虚假人脸生成结果中往往包含与 GAN 模型相关的特定失真现象. 文献 [69] 提出通过检测视频中是否包含人脸图像进行缩放、旋转、剪切等仿射变换所产生的扭曲失真来判断是否属于伪造视频. Kumar 等 [80] 同样提出可以利用残差神经网络学习人脸替换算法伪造视频中的局部失真, 并在不同视频压缩程度下进行伪造检测. Martern 等 [70] 指出在伪造人脸图像的眼睛、鼻梁和牙齿等五官细节区域存在一些视觉细节一致性的不协调现象可以用于伪造人脸检测. 例如伪造人脸图像的眼睛区域会出现左右眼睛虹膜颜色不一致, 在鼻梁附近会出现异常的阴影, 以及头发和牙齿区域的细节缺失等. 文献 [81] 发现基于 GAN 模型生成的伪造图像往往存在一种特定的失真现象, 并可将图像转化到频域光谱图中进行真伪分辨. 除了关注伪造图像中的细节不一致外, Michigan State University 刘小明教授团队 [82] 引入了视觉注意模型来进一步关注图像中的伪造区域.

传统的图像处理技术也可以应用于伪造检测任务中提高伪造检测准确率. Durral 等 [71] 提出一种无需大规模训练数据的非深度学习方法, 对图像进行离散傅里叶 (Fourier) 变换处理后, 将频域特征输

12) http://github.com/shaoanlu/faceswap-GAN.

表 1　基于空域线索的代表性伪造检测方法总览

Table 1　Summary of face forgery detection methods based on spatial clues

Category	Motivation	Method	Detection clues	Face forgery types
Based on color artifacts	Existing face forgery methods usually ignore constrains on color spaces	[67]	Translating RGB images into HSV, YCbCr color spaces	Target-generic fake face generation
		[68]	Translating RGB images into HSV, YCbCr color spaces, together with high-pass filtering	Target-generic fake face generation
Based on visual disparities	Existing forgery results contain visual disparities in details	[69]	Using CNN to detect face swap contours	Face swap
		[70]	Detecting visual disparities in details, such as eye colors, varying lighting, and teeth details	Face swap, reenactment, fake face generation
Based on image processing	Using traditional image processing techniques	[71]	Conducting discrete Fourier transform, and apply classification models on frequency domain	Face swap, fake face generation
Based on image classification	Considering forgery detection as image classification task	[72]	Assuming fake images generated from the same GAN model as one class	Target-generic fake face generation
		[73]	Using AdaBoost and XGBoost to deal with unbalanced data between real and fake images	Face swap
Based on multi-task learning	Performing forgery detection and location together	[73]	Using multi-task learning to detect and segment forgery faces	Face swap, reenactment
		[74]	Introducing semantic segmentation into forgery detection task	Fake face generation
Based on specific deep networks	Replying on specific deep networks for forgery detection	[75]	Monitoring neuron behavior for face forgery detection	Fake face generation, reenactment,
		[76]	Using capsule networks for forgery feature extraction	Face swap, reenactment
		[77]	Introducing Gram-Net architecture for face forgery detection	Fake face generation

入到逻辑回归或支持向量机分类器中进行图像真伪判断. 中山大学骆伟祺教授团队 [83] 提出利用高通滤波器进行预处理以提高检测效果. 中国科学院董晶教授团队 [84] 指出可以借助高斯 (Gauss) 模糊和增加高斯噪声等图像预处理操作来减少伪造图像和真实图像中存在的不稳定高频失真, 从而有助于学习更加关键的伪造特征以提高伪造检测的性能.

此外, 伪造检测任务也可以按照图像分类的思路进行处理. University of Maryland, College Park 的 Larry Davis 教授团队 [85] 提出一种基于双通道的伪造人脸检测框架, 其中一个通道利用卷积神经网络构成分辨真实或伪造人脸的二分类器模型, 另一个通道提取局部残差噪声和相机噪声进行二分类操作, 最后将两个分类器的结果进行融合. 韩国世宗大学研究团队 [86] 提出可以利用梯度提升算法如 AdaBoost 和 XGBoost 等分类器解决伪造检测问题中数据规模不一致导致的数据不平衡问题. Marra 等 [87] 指出基于 GAN 模型生成的伪造人脸图像中往往包含特殊的指纹特征, 为此 University of Maryland, College Park 的 Larry Davis 教授团队 [72] 提出一种基于 GAN 指纹特征检测的方法. 该方

表 2 基于时域线索的代表性伪造检测方法总览

Table 2 Summary of face forgery detection methods based on temporal clues

Category	Motivation	Method	Detection clues	Forgery types
Eye blinking	Detecting the frequency of eye blinking	[91]	Using CNN and RNN to detect eye blinking	Face swap
Head movement	Detecting head poses	[92]	Locating disparities between head pose and facial pose	Face swap
Biological signals	Extracting biological signals for detection	[93]	Using PPG signals with SVM	Face swap
		[94]	Detecting heart rates for forgery classification	Face swap
Frame-to-frame disparities	Detecting disparities between frames	[95]	Using RNN to extract spatial features	Face swap
		[96]	Using optical flow for forgery detection	Face swap

法将同一个 GAN 模型所生成的伪造图像标记为相同的类标, 且真实图像也看作属于单独一类, 然后将所有数据输入到卷积神经网络中进行图像分类训练. 实验发现该方法不仅可以分辨出真实和伪造图像, 还可以区分出由不同 GAN 模型生成的伪造人脸图像.

在进行伪造检测的同时还可以将伪造区域的定位分割任务考虑进来, 利用二者之间的共性, 通过多任务学习的方法实现. Nguyen 等 [73] 提出一种基于卷积神经网络的多任务学习框架, 实现在检测伪造视频的同时定位伪造区域的范围. Zafeiriou 等 [88] 指出在全卷积神经网络中增加图像分割任务的损失函数, 即可同时实现伪造图像检测和伪造区域定位. 中国传媒大学任慧教授与京东人工智能研究院 [89] 合作提出一种在像素级别进行伪造人脸图像检测和伪造区域分割的方法. 最近, Huang 等 [74] 将图像语义分割的思想引入进来, 针对多种基于 GAN 模型的人脸图像伪造类型进行伪造人脸检测和伪造区域分割.

最后, 针对伪造检测问题还可以设计专门的深度网络结构实现. 文献 [75] 提出一种伪造人脸图像检测算法 Fakespotter, 通过提取神经网络模型中每一层神经元的激活状态作为伪造检测的特征. Jeon 等 [90] 在现有预训练模型基础上, 设计了一种参数微调网络, 可以很好地与现有图像分类网络模型进行结合. Nguyen 等 [76] 将胶囊网络结构引入到伪造检测任务中. University of Oxford 的 Philip Torr 教授和香港中文大学贾佳亚教授团队 [77] 最新提出一种对人脸图像纹理增强的深度模型 Gram-Net, 通过在网络结构中增加了图像风格转化任务里常用的格拉姆 (Gram) 矩阵层用于检测全局纹理特征, 可以在进行伪造检测的同时对图像下采样、图像压缩、图像模糊和噪声等具有一定的鲁棒性.

3.1.2 基于时域线索的检测方法

上述基于空域线索的检测方法大多针对伪造人脸图像或者伪造视频中的单帧图像进行检测, 而对于伪造视频来说, 还可以利用时域线索提高伪造检测算法的性能. 表 2 [91~96] 中对基于时域线索的代表性伪造检测方法进行了概述. Afchar 等 [97] 首次考虑到伪造视频检测的问题, 并设计了两种卷积神经网络模型来检测视频中的伪造线索.

人类的眨眼行为是一种常见的生理现象且在时域上具有一定的规律性, 而现有的伪造方法往往忽略了虚假人脸的眨眼或闭眼行为. 为此, Li 等 [91] 提出了一种基于眨眼检测的伪造视频检测方法. 该方法首先进行人脸检测和对齐, 以减少视频中头部摆动和角度变化的干扰. 随后提取人眼区域, 利用卷积神经网络提取特征后输入到循环神经网络中判断眼睛的状态, 最后考虑到真实视频中的眨眼频率

通常在每分钟 30 次左右而伪造视频中的眨眼频率每分钟只有 3, 4 次, 因而可以设置眨眼频次阈值来区分是否属于伪造视频.

除了检测眨眼动作外, 还可以检测视频中的头部摆动规律来进行伪造视频的鉴别. 文献 [92] 考虑到在人脸替换过程中易存在嵌入人脸与背景人脸的姿态角度偏差, 提出通过检测人脸五官关键点与带下巴轮廓的整幅人脸关键点后, 分别根据关键点信息估计二者的头部姿态角度以判断是否一致. University of California, Berkeley 的 Agarwal 等 [98] 提出利用 OpenFace2 [99] 进行人脸检测, 并提取人脸面部运动单元特征用于完成伪造视频的检测.

从人脸视频中还可以提取生物信号特征作为视频伪造的线索. 文献 [93] 提出一种利用生物信号特征进行伪造视频检测的算法 FakeCatcher, 利用血压检测中常用的光电容积脉搏波描记法 (PPG, PhotoPlethysmoGraphy) 作为生物信号特征, 输入到支持向量机中进行伪造视频的鉴别. 此外, Fernandes 等 [94] 同样提出可以利用神经常微分方程模型进行伪造人脸视频中的心率估计.

伪造视频中往往还存在视频帧之间的不一致性, 因此可以通过提取帧间差异用于伪造视频的检测. Purdue University 研究人员 [95] 首先提出将伪造视频帧间不一致性作为检测线索. 该方法利用卷积神经网络提取每一帧图像的特征, 然后将其输入到循环神经网络中进行时序特征的提取, 最后利用全卷积神经网络进行分类以判断是否属于伪造视频. Sabir 等 [100] 同样提出利用卷积神经网络和循环神经网络进行级联的伪造视频检测的方案. Amerini 等 [96] 提出可以利用光流法检测视频帧间的差异性, 并将该差异性输入到卷积神经网络中完成伪造视频的检测任务.

3.2 面向未知伪造类型的泛化能力研究

上述可视身份伪造检测方法大多需要利用特定类型的伪造数据进行训练, 并在测试视频或图像的伪造类型与训练数据一致时可以取得较好的伪造检测效果. 然而, 当测试图像来自与训练样本不同的伪造类型时, 现有伪造检测方法的性能往往很差. 因此, 如何提高伪造检测模型的泛化能力, 对未知伪造类型的可视身份视频或图像数据进行有效的检测是当前研究的热点问题.

文献 [101] 借助迁移学习的思想, 提出基于弱监督领域自适应的方法提高伪造检测的泛化能力, 并可以应对当训练样本不足时模型的训练问题. Du 等 [102] 提出一种局部感知自编码器的方法, 利用主动学习的策略从未知伪造类型的数据中选择具有挑战性的样本进行标记, 以提高对未知伪造类型图像的检测能力. Marra 等 [103] 为了应对不断改进的 GAN 图像生成算法, 将增量学习的思想引入到检测模型的训练过程中, 并通过多任务学习来同时实现 GAN 图像类型分类任务和伪造检测任务. University of California, Berkeley 的 Alexix Efros 教授团队与 Adobe 研究院 [104] 提出借助数据增强的方案提高伪造检测算法的泛化能力, 在图像预处理和数据增强环节采取包括图像翻转、高斯模糊、图像压缩, 以及多种增强方案组合的形式对训练数据进行处理, 以得到具有较好泛化能力的伪造检测模型.

最近, 微软亚洲研究院和北京大学研究团队 [105] 提出一种人脸伪造检测方法 Face X-ray. 该方法主要针对人脸替换时将人脸图像面部区域嵌入到背景图像的步骤, 通过判断一张图像能否被分解为来自不同来源的两个部分, 进而显示出人脸替换的伪造边界作为检测的依据. 该方法还采取了自监督学习的策略, 在训练过程中不断利用真实人脸图像生成伪造数据, 并具有很好的泛化能力.

3.3 面向对抗样本攻击的可信检测研究

在伪造检测算法的设计过程中, 还存在一个不可忽视的问题就是伪造检测算法在遇到对抗样本攻击等欺骗攻击时能否继续提供可信的伪造检测结论. 算法的可信研究在信息安全相关领域是重要的研究课题之一 [106], 然而在可视身份伪造检测领域中所开展的相关研究还相对较少. 对抗样本攻击技

术 [107] 通过对输入图像添加肉眼难以察觉的微小扰动, 可以误导或欺骗模型做出错误的判断, 并成功应用在人脸识别 [108]、目标检测 [109] 等领域的研究当中. 考虑到对抗样本攻击对于深度神经网络的欺骗能力, 伪造检测技术的可信度问题也是重要的研究课题.

Nguyen 等 [110] 首次提出需要验证伪造检测算法应对对抗样本攻击的问题, 针对基于多任务学习的伪造检测算法 [73], 设计了基于梯度的白盒攻击方案, 并可迁移攻击 [97,101] 等其他的伪造检测算法. 考虑到伪造检测算法 [72,87] 主要通过检测基于 GAN 模型生成人脸中特定的指纹特征来进行伪造图像的鉴别, Neves 等 [111] 提出移除这些指纹特征来欺骗现有的人脸伪造检测算法 [79]. Boston University 的 Ruiz 等 [112] 借鉴了传统的对抗样本攻击算法干扰伪造人脸图像的生成过程, 从而在生成结果中添加对抗扰动以实现攻击深度伪造算法的效果. Huang 等 [113] 提出可以通过对生成图像进行局部细节重建来消除其中的伪造线索, 从而达到欺骗现有伪造检测方法的目的.

4 数据集和算法性能分析

目前的可视身份深度伪造检测算法通常通过两种方式进行评估: 一是基于公开的可视身份伪造数据集, 二是利用公开代码生成伪造数据后在自建数据集上进行算法验证. 本节主要介绍已有伪造检测算法所采用的数据集情况以及代表性算法在公开数据集上的实验结果与分析.

4.1 伪造检测数据集

在伪造检测任务中常用的真实人脸数据集有明星人脸数据集 CelebFaces Attributes Dataset (CelebA)[13)][114]、高清明星人脸数据集 CelebA-HQ[14)][53] 和高清人脸数据集 Flickr-Faces-HQ (FFHQ)[15)][54]. CelebA 数据集中包含约 20 万张明星人脸照片, 然而其图像分辨率较低. CelebA-HQ 数据集包含 3 万张取自 CelebA 的明星人脸图像, 并将其图像分辨率提高到 1024×1024. FFHQ 数据集收集了更大规模的约 7 万张高清人脸图像, 且分辨率为 1024×1024.

目前公开的真实和伪造检测数据集及相应数据类型与规模总结见表 3 [53~55,66,92,114~119]. 具体介绍如下:

(1) 100K-Faces-StyleGAN [54]. 共包含 10 万张伪造人脸图像, 利用无目标虚假人脸伪造算法 StyleGAN [54] 所生成, 并随同高清人脸数据集 FFHQ 一起公开发布[16)].

(2) 100K-Faces-StyleGAN2 [55]. 同样包含 10 万张伪造人脸图像, 利用伪造算法 StyleGAN2 [55] 所生成[17)]. 此外, 在网站 [11)] 上公开有基于 StyleGAN2 算法生成的伪造人脸图像, 算法 [71] 也曾从网站上采集该类型伪造人脸数据用于伪造检测研究.

(3) UADFV [92]. 共包含 49 个真实视频和 49 个伪造视频[18)]. 该数据集中真实和伪造视频的长度均为 11 s 左右, 视频图像分辨率为 294×500, 其中真实视频从 YouTube 网站中采集, 而伪造视频利用文献 [91] 中所介绍的人脸替换算法生成.

(4) Fake Face in the Wild dataset (FFW) [115]. 共包含 150 个伪造视频[19)]. 其中伪造视频采集自 Youtube 网站, 伪造方法包括人脸替换、计算机图形学处理等多种类型.

13) http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.
14) http://github.com/tkarras/progressive_growing_of_gans.
15) http://github.com/NVlabs/ffhq-dataset.
16) http://github.com/NVlabs/stylegan.
17) http://github.com/NVlabs/stylegan2.
18) http://github.com/danmohaha/WIFS2018_In_Ictu_Oculi.
19) http://ali.khodabakhsh.org/research/ffw/.

表 3 伪造检测数据集总览

Table 3 Summary of face forgery detection datasets

Dataset	Date	Types	Real data (number)	Fake data (number)
CelebA [114]	2015	Real face images	Image (202K)	–
CelebA-HQ [53]	2018	Real face images	Image (30K)	–
FFHQ [54]	2018	Real face images	Image (70K)	–
100K-Faces-StyleGAN [54]	2018	Target-generic fake faces	–	Image (100K)
100K-Faces-StyleGAN2 [55]	2019	Target-generic fake faces	–	Image (100K)
UADFV [92]	2018	Target-specific fake faces	Video (49)	Video (49)
FFW [115]	2018	Target-specific fake faces	–	Video (150)
DeepfakeTIMIT [66]	2018	Target-specific fake faces	Video (320)	Video (640)
FaceForensics++ [116]	2019	Target-specific fake faces	Video (1000)	Video (4000)
Celeb-DF [117]	2019	Target-specific fake faces	Video (408)	Video (795)
Celeb-DF(v2) [117]	2019	Target-specific fake faces	Video (590)	Video (5639)
DFDC [118]	2019	Target-specific fake faces	Video (1131)	Video (4113)
DeeperForensics-1.0 [119]	2020	Target-specific fake faces	Video (10000)	Video (10000)

(5) DeepfakeTIMIT [66]. 共包含 320 个真实视频和 640 个伪造视频[20]. 其中真实视频来源为澳大利亚 The University of Queensland 发布的 VidTIMIT Audio-Video Dataset[21], 伪造视频利用开源的人脸替换代码生成, 并且生成低分辨率 (64×64) 和高分辨率 (128×128) 伪造视频各 320 个.

(6) FaceForensics++ [116]. 共包含 1000 个真实视频和 4000 个伪造视频[22]. 其中, 真实视频采集自 Youtube 网站, 而伪造视频分别利用人脸表情重演算法 Face2Face [29]、基于计算机图形学的传统人脸替换算法人脸 FaceSwap[23], 以及基于深度学习的人脸替换算法 DeepFakes[24]、NeuralTextures [120] 生成. 该数据集中还提供了原始视频、高质量压缩视频和低质量压缩视频 3 种类型的伪造视频. 此外, 该数据库还包含了由谷歌及其子公司 Jigsaw 发起的深度伪造检测数据集 (deep fake detection dataset, DFD), 包含有 28 名演员在不同场景下的 3431 个伪造视频.

(7) Celeb-DF [117]. 该数据集共有两个版本. 初始版本中共包含 408 个真实视频和 795 个伪造视频[25], 其中真实视频采集自 Youtube 网站, 而伪造视频通过人脸替换算法生成. 随后, 在 2019 年 11 月更大规模的 Celeb-DF(v2) 发布, 共包含 590 个真实视频和 5639 个伪造视频[26].

(8) Deepfake detection challenge (DFDC) [118]. 该数据集为 Facebook 在竞赛网站 Kaggle 上发起深度伪造检测挑战竞赛数据[27], 共包含 1131 和 4113 个真实与伪造视频. 其中真实视频为 Facebook 通过发起视频征集活动后所选择 66 名用户自行拍摄的视频. 伪造视频通过人脸替换算法生成, 并且添加了降低视频每秒帧数、缩小视频图像分辨率和降低视频编码质量等方式处理后的伪造数据.

(9) DeeperForensics-1.0 [119]. 该数据集为近期发布的最大规模人脸伪造数据集, 共包含约 10000 个

20) http://www.idiap.ch/dataset/deepfaketimit.
21) http://conradsanderson.id.au/vidtimit/.
22) http://github.com/ondyari/FaceForensics.
23) http://github.com/MarekKowalski/FaceSwap/.
24) http://github.com/deepfakes/faceswap.
25) https://github.com/danmohaha/celeb-deepfakeforensics/blob/master/Celeb-DF-v1/README.md.
26) https://github.com/danmohaha/celeb-deepfakeforensics.
27) https://ai.facebook.com/datasets/dfdc/.

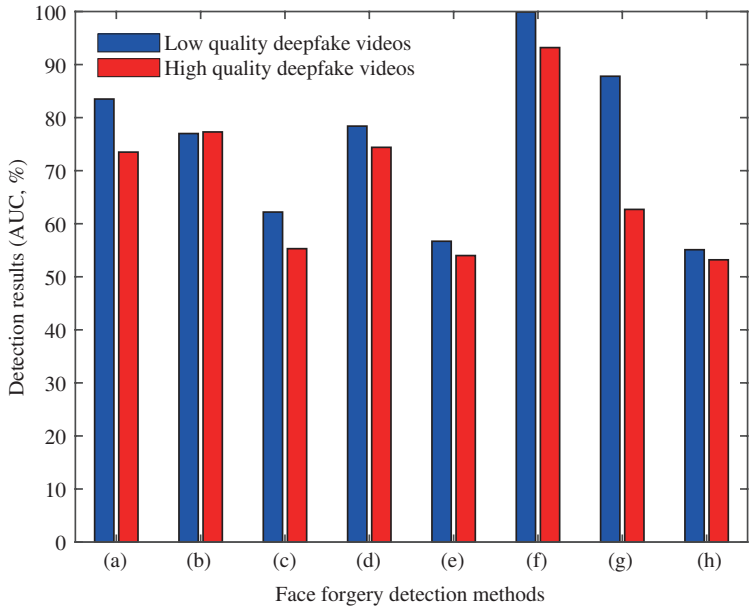


图 4 (网络版彩图) 在 DeepfakeTIMIT 数据集不同视频分辨率下的伪造检测结果对比图

Figure 4 (Color online) Experimental results on DeepfakeTIMIT dataset. (a) Two-streamNN [85], (b) EVA [70], (c) Multi-task [73], (d) Capsule [76], (e) Xception [116], (f) FWA [69], (g) Mesonet [97], (h) HeadPose [92]

真实视频和 50000 个伪造视频[28]. 其中真实视频采集自 26 个国家的 100 名演员拍摄数据, 并在光照环境、姿态、表情下利用不同的语言进行自然说话. 伪造数据利用文献 [119] 生成, 并采取添加高斯白噪声、图像压缩、视频压缩等 35 种不同类型的数据处理, 以模拟真实场景下视频传输对伪造检测的影响. 该数据集从数据规模和数据真实程度上均在目前已公开数据集中处于领先水平.

已有的公开数据集大多提供的是人脸替换伪造类型的视频数据. 除此之外, 还有一些方法会采用公开的可视身份伪造算法生成伪造人脸图像后, 在自建数据集上进行有目标人脸属性编辑伪造和无目标虚假人脸生成等类型的伪造检测研究. 例如, 深圳大学黄继武教授团队 [67] 采用了 DCGAN 和 ProGAN 等算法生成的虚假人脸图像数据验证基于颜色失真的伪造检测算法. 文献 [72] 在基于 ProGAN, SNGAN 在内的多种算法生成的虚假人脸图像上检测 GAN 特有的指纹特征用于伪造图像分类. 文献 [74,101∼104] 也采集了包括 CycleGAN, StarGAN, ProGAN, StyleGAN, BigGAN 在内的多种人脸生成算法制作伪造人脸数据以验证伪造检测算法的泛化能力.

4.2 算法性能分析

在现有的伪造检测任务中, 主要采用的是机器学习中常见的评价指标, 如性能评价指标有接受者操作特性 (receiver operating characteristic, ROC) 曲线下面积 (area under the curve, AUC)、平均错误率 (equal error rate, EER) 和检测准确率 (accuracy, Acc) 等. 目前在伪造检测问题中仍没有统一的评价指标标准, 且缺乏针对该问题的专用型评估方法. 我们下面对一些代表性伪造检测算法在部分公开数据集上的性能进行介绍. Celeb-DF (v2) 数据集、DFDC 数据集和 DeeperForensics-1.0 数据集由于发布时间较新, 暂时缺乏其他算法在这些数据集上进行实验仿真, 因此可以参考数据集对应论文 [117∼119] 了解基准算法在这些数据集上的表现.

28) http://github.com/EndlessSora/DeeperForensics-1.0.

表 4　不同算法在 FaceForensics++ 部分子集上的伪造检测结果 (%)

Table 4　Experimental results on FaceForensics++ dataset (%)

Category	Method	Result (FaceSwap)	Result (Face2Face)	Result (DFD)
Spatial clue	Two-streamNN [85]	AUC (70.1)	**Acc** (**NC: 99.9**, HQ: 96, LQ: 86.8)	AUC (52.8)
	EVA [70]	AUC (78)	AUC (86.6)	**AUC (77.2)**
	Multi-task [73]	AUC (76.3), EER (HQ: 15.1)	EER (HQ: 7.1)	AUC (54.1), **Acc (92.8)**, EER (8.2)
	Capsule [76]	AUC (96.6)	Acc (NC: 99.3, HQ: 96.5, LQ: 81)	AUC (64)
	Xception [116]	**AUC (99.7)**	–	AUC (53.9)
	FWA [69]	AUC (80.1)	–	AUC (74.3)
	Multi-Resnet [80]	–	**Acc** (**NC: 99.9**, HQ: 99.1, LQ: 91.2)	–
	Attention [82]	–	**AUC (99.4)**, EER (3.4)	–
	DFT [71]	–	–	Acc (91)
Temporal clue	Mosenet [97]	AUC (84.7)	Acc (NC: 96.8, HQ: 93.4, LQ: 83.2)	AUC (76)
	Headpose [92]	AUC (47.3)	–	AUC (56.1)
	RCNN [100]	AUC (LQ: 96.3)	Acc (LQ: 94.3)	–
	Opticalflow [96]	–	Acc (81.6)	–

图 4 展示了代表性算法在 DeepfakeTIMIT 数据集上的检测结果. 该数据集规模较小, 并且提供了低分辨率和高分辨率两种类型的视频数据, 以验证视频分辨率对伪造检测模型的影响. 从图中可以看出, 当视频分辨率提高后, 伪造视频的逼真程度也相应提高, 导致伪造检测难度增大, 检测模型在高分辨率数据上的表现更差. 然而, 该数据集中高分辨率数据的视频分辨率仅为 128×128, 随着视频分辨率的进一步提升, 伪造检测问题也将面临更大的挑战.

FaceForensics++ 数据集为首个大规模伪造视频数据集, 包含有人脸替换伪造 FaceSwap、面部重演伪造 Face2Face, 以及谷歌深度伪造检测数据 DFD 在内的多个子集. 由于缺乏统一的评价指标标准, 现有方法在该数据集上汇报了包括 AUC, EER, Acc 在内的多种性能结果, 如表 4 所示. 与此同时, 该数据集提供了原始数据、高质量数据和低质量数据 (表中分别记为 NC, HQ, LQ) 3 种类型的伪造视频, 以验证在实际场景中视频压缩对伪造检测模型的影响. 对于其中的人脸替换 FacSwap 子数据集来说, 现有方法大多采取了 AUC 作为测试评估指标. 基于空域线索的代表性方法 Capsule [76] 和基于时域线索的 RCNN [100] 算法均取得了 96% 的 AUC 值, 而提取空域特征的神经网络 Xception [116] 在人脸替换检测上达到了 99.7% AUC 值的最佳效果. 在评估面部重演伪造类型 Face2Face 和谷歌深度伪造检测数据集 DFD 时, 现有研究采取了多种评估指标. 例如, 基于视觉注意模型的 Attention [82] 算法对于面部重演伪造检测可以达到 99.4% 的 AUC 值. 在基于检测准确率 Acc 指标的评估过程中, 两种神经网络模型 Two-steamNN [85] 和 Multi-Resnet [80] 均达到了 99.9% 的准确率. 在谷歌深度伪造检测数据集 DFD 的实验中, EVA [70] 算法可以在评估指标 AUC 方面取得最好的 77.2%, 而基于多任务学习的 Multi-task [73] 算法在该数据集上达到最高的检测准确率 92.8%. 从实验结果可以看出, DFD 子集的伪造检测难度最高. 当检测准确率 Acc 难以有效区分不同算法的优劣时, ROC 曲线下面积 AUC 值可以更好地刻画不同算法之间的性能差异.

表 5 中展示了代表性算法在 UADFV 小规模数据集和 Celeb-DF 数据集上的实验结果. 其中, 由于 UADFV 数据规模较小, 现有方法大多可以达到 80% 以上的 AUC 指标值. 基于头部运动检测的

表 5 不同算法在 UADFV 和 Celeb-DF 数据集上的伪造检测结果 (%)

Table 5 Experimental results on UADFV and Celeb-DF datasets (%)

Category	Method	Result (UADFV)	Result (Celeb-DF)
Spatial clue	Two-streamNN [85]	AUC (85.1)	AUC (55.7)
	EVA [70]	AUC (70.2)	AUC (48.8)
	Multi-task [73]	AUC (65.8)	AUC (36.5)
	Capsule [76]	AUC (61.3)	–
	Xception [116]	AUC (80.4)	AUC (38.7)
	FWA [69]	**AUC (97.4)**	AUC (53.8)
Temporal clue	Mesonet [97]	AUC (84.3)	AUC (53.6)
	Headpose [92]	AUC (89)	AUC (54.8)
Generalizability	Face X-ray [105]	–	**AUC (80.6)**

Headpose [92] 算法可以达到 89% 的 AUC 值, 而基于视觉失真检测的算法 FWA [69] 在 UADFV 小规模数据集上取得了最高的 97.4%. Celeb-DF 数据集移除了伪造视频中存在的视觉细节不一致现象, 因此该数据集更具有挑战性. 例如, 利用神经网络提取空域线索的算法 Two-streamNN [85] 和 Xception [116] 只取得了 55.7% 和 38.7% 的 AUC 值, 而基于眨眼运动检测的算法 Mesonet [97] 和基于头部运动检测的算法 Headpose [92] 所达到的 AUC 值均低于 55%. 面向未知伪造类型的检测算法 Face X-ray [105] 通过提高检测模型的泛化能力, 可以在该数据集上取得最好的结果 (80.6%). 随着可视身份伪造算法的不断改进, 伪造数据中的颜色或视频失真、头部动作等细节将难以继续作为伪造检测的有效线索, 因而设计面向未知伪造类型的检测算法将是未来的研究热点之一.

5 问题与挑战

可视身份伪造与检测问题的研究目前还处在起步阶段, 尽管已有大量基于空域线索和时域线索的伪造检测方法, 也初步提出一些提高模型泛化能力以应对未知伪造类型的算法. 随着伪造技术的不断发展, 现有模型和研究思路逐渐显示出一定的局限性, 并且有以下待解决的问题与挑战:

(1) 复杂场景伪造检测. 在伪造检测研究之初, 伪造算法相对比较简单, 伪造视频或图像中往往存在明显的视觉失真等线索. 然而伴随着伪造算法的不断变革更新, 伪造检测的对象也将逐渐扩展到更加复杂的伪造类型和场景中. 例如, 现有伪造检测问题大多只考虑了正面人脸图像或视频, 而真实场景中还可能面临侧面视角人脸伪造、带遮挡人脸伪造和检测问题. 此外, 现有伪造检测方法通常只考虑一种伪造类型的数据, 然而在实际场景下可能存在多种伪造类型的叠加问题, 例如人脸属性修改与表情重演的复合伪造. 如何对复合伪造类型的数据进行检测, 并准确鉴别出所经过的伪造方式, 在当前研究中仍未曾涉及.

(2) 算法性能评估. 现有伪造检测工作仍然缺乏可靠统一的性能评估标准. 在现有工作中大多采取了机器学习中常用的指标进行性能评估, 然而这些传统评估指标与人类对伪造数据的主观感受之间仍存在一定的差异. 因此, 有必要针对可视身份伪造与检测问题进行深入剖析, 设计与人类主管感受一致的伪造检测评估模型, 以帮助伪造与检测算法进行性能分析和模型优化. 与此同时, 现有公开数据集大多以人脸替换伪造类型数据为主, 伪造类型缺乏多样性, 构建包含人脸属性编辑、表情重演, 以及跨模态人脸编辑伪造在内的多样性伪造数据集也是目前亟待解决的问题.

(3) 与其他类型非真实身份图像的联系. 研究本文所介绍的可视身份深度伪造类型数据与一些传统的非真实身份图像之间的联系, 例如图像编辑软件 Photoshop 等篡改图像、人脸活体检测任务中常见的打印欺骗、重放欺骗、三维面具欺骗等, 也是未来待探索的方向之一. 由于非真实身份数据具有和真实数据分布差异性等共同特点, 一方面可以借鉴传统篡改图像检测任务、人脸活体检测任务中的研究思路, 启发深度伪造检测的算法设计. 另一方面, 如何借助多任务学习、联合学习等方案, 实现一种模型检测多种类型的非真实身份数据, 也具有巨大的理论研究与实际应用价值.

(4) 可信伪造检测研究. 在图像分类研究中, 分类器模型极易受到对抗样本的影响. 由于伪造检测问题在一定程度上也可以看作是分类问题, 因此也容易受到因为数据中的微小扰动干扰伪造检测模型的判断结果. 现有研究工作中针对伪造检测算法的可信研究相对较少, 因此在进行伪造检测算法设计的同时, 解决伪造检测结果是否可信的问题, 对于完善相关技术的理论发展和提升应用推广价值具有重要意义.

6 结束语

本文介绍了可视身份深度伪造与检测的研究方法, 首先针对可视身份深度伪造问题, 根据伪造过程中是否有具体的伪造目标从有目标身份伪造和无目标身份伪造两个方面进行阐述, 其中根据伪造程度的不同, 从人脸替换、人脸属性编辑、人脸表情重演和跨模态人脸编辑等方面对有目标身份伪造研究进行了归纳. 其次, 对于身份伪造检测方法从基于空域线索的方法、基于时域线索的方法、面向未知伪造类型的检测方法和可信伪造检测研究等多个方面进行了总结. 进而, 介绍了现有工作中常见的数据集以及代表性方法的算法性能分析后, 指出了可视身份深度伪造检测领域面临的问题和未来研究中的挑战.

参考文献

1 Long K, Ma Y, Zhu Q C. How will deepfake technology influence national security: emerging challenges and policy implications. China Inform Secur, 2019, 10: 21–34 [龙坤, 马钺, 朱启超. 深度伪造对国家安全的挑战及应对. 信息安全与通信保密, 2019, 10: 21–34]

2 Cao J F. Deepfake technology: the legal challenge and response. China Inform Secur, 2019, 10: 8 [曹建峰. 深度伪造技术的法律挑战及应对. 信息安全与通信保密, 2019, 10: 8]

3 Yang R, Luo W Q, Huang J W. Multimedia forensics. Sci Sin Inform, 2013, 43: 1654–1672 [杨锐, 骆伟祺, 黄继武. 多媒体取证. 中国科学: 信息科学, 2013, 43: 1654–1672]

4 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444

5 Finlayson S G, Bowers J D, Ito J, et al. Adversarial attacks on medical machine learning. Science, 2019, 363: 1287–1289

6 Leslie D. Raging robots, hapless humans: the AI dystopia. Nature, 2019, 574: 32–33

7 Korshunova I, Shi W, Dambre J, et al. Fast face-swap using convolutional neural networks. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 3677–3685

8 Nirkin Y, Masi I, Tuan A T, et al. On face segmentation, face swapping, and face perception. In: Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, 2018. 98–105

9 Natsume R, Yatagawa T, Morishima S. RSGAN: face swapping and editing using face and hair representation in latent spaces. 2018. ArXiv:1804.03447

10 Natsume R, Yatagawa T, Morishima S. Fsnet: an identity-aware generative model for image-based face swapping. In: Proceedings of Asian Conference on Computer Vision, Perth, 2018. 117–132

11 Bao J M, Chen D, Wen F, et al. Towards open-set identity preserving face synthesis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 6713–6722

12 Nirkin Y, Keller Y, Hassner T. Fsgan: subject agnostic face swapping and reenactment. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 7184–7193
13 Li L Z, Bao J M, Yang H, et al. FaceShifter: towards high fidelity and occlusion aware face swapping. 2019. ArXiv:1912.13457
14 Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017. 1125–1134
15 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 2223–2232
16 Shen W, Liu R J. Learning residual images for face attribute manipulation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017. 4030–4038
17 Lample G, Zeghidour N, Usunier N, et al. Fader networks: manipulating images by sliding attributes. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 5967–5976
18 Choi Y, Choi M, Kim M, et al. Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 8789–8797
19 Choi Y, Uh Y, Yoo J, et al. StarGAN v2: diverse image synthesis for multiple domains. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, 2020
20 He Z L, Zuo W M, Kan M N, et al. AttGAN: facial attribute editing by only changing what you want. IEEE Trans Image Process, 2019, 28: 5464–5478
21 Liu M, Ding Y K, Xia M, et al. STGAN: a unified selective transfer network for arbitrary image attribute editing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 3673–3682
22 Guo J T, Qian Z Z, Zhou Z W, et al. MulGAN: facial attribute editing by exemplar. 2019. ArXiv:1912.12396
23 Chen Y C, Lin H J, Shu M, et al. Facelet-bank for fast portrait manipulation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 3541–3549
24 Chen Y C, Shen X H, Lin Z, et al. Semantic component decomposition for face attribute manipulation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 9859–9867
25 Qian S J, Lin K Y, Wu W, et al. Make a face: towards arbitrary high fidelity face manipulation. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 10033–10042
26 Lee C H, Liu Z W, Wu L Y, et al. MaskGAN: towards diverse and interactive facial image manipulation. 2019. ArXiv:1907.11922
27 Usman B, Dufour N, Saenko K, et al. PuppetGAN: cross-domain image manipulation by demonstration. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 9450–9458
28 Thies J, Zollhöfer M, Nießner M, et al. Real-time expression transfer for facial reenactment. ACM Trans Graph, 2015, 34: 183
29 Thies J, Zollhofer M, Stamminger M, et al. Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 2387–2395
30 Thies J, Zollhöfer M, Theobalt C, et al. Headon: real-time reenactment of human portrait videos. ACM Trans Graph, 2018, 37: 1–13
31 Kim H, Garrido P, Tewari A, et al. Deep video portraits. ACM Trans Graph, 2018, 37: 1–14
32 Song L X, Lu Z H, He R, et al. Geometry guided adversarial facial expression synthesis. In: Proceedings of ACM International Conference on Multimedia, Seoul, 2018. 627–635
33 Wu W, Zhang Y X, Li C, et al. ReenactGAN: learning to reenact faces via boundary transfer. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 603–619
34 Otberdout N, Daoudi M, Kacem A, et al. Dynamic facial expression generation on Hilbert hypersphere with conditional Wasserstein generative adversarial nets. 2019. ArXiv:1907.10087
35 Huang Y, Khan S M. Generating photorealistic facial expressions in dyadic interactions. In: Proceedings of British Machine Vision Conference, Newcastle, 2018
36 Ding H, Sricharan K, Chellappa R. ExprGAN: facial expression editing with controllable expression intensity. In: Proceedings of AAAI Conference on Artificial Intelligence, New Orleans, 2018
37 Pumarola A, Agudo A, Martinez A M, et al. GANimation: anatomically-aware facial animation from a single image.

In: Proceedings of European Conference on Computer Vision, Munich, 2018. 818–833
38 Tripathy S, Kannala J, Rahtu E. ICface: interpretable and controllable face reenactment using gans. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, Colorado, 2020. 3385–3394
39 Zakharov E, Shysheya A, Burkov E, et al. Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 9459–9468
40 Ha S, Kersner M, Kim B, et al. MarioNETte: few-shot face reenactment preserving identity of unseen targets. 2019. ArXiv:1911.08139
41 Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing Obama: learning lip sync from audio. ACM Trans Graph, 2017, 36: 1–13
42 Oh T H, Dekel T, Kim C, et al. Speech2Face: learning the face behind a voice. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 7539–7548
43 Wen Y, Raj B, Singh R. Face reconstruction from voice using generative adversarial networks. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 5266–5275
44 Vougioukas K, Petridis S, Pantic M. Realistic speech-driven facial animation with GANs. Int J Comput Vis, 2020, 128: 1398–1413
45 Jamaludin A, Chung J S, Zisserman A. You said that?: synthesising talking faces from audio. Int J Comput Vis, 2019, 127: 1767–1779
46 Song L, Wu W, Qian C, et al. Everybody's talkin': let me talk as you want. 2020. ArXiv:2001.05201
47 Fried O, Tewari A, Zollhöfer M, et al. Text-based editing of talking-head video. ACM Trans Graph, 2019, 38: 1–14
48 Chen X, Qing L B, He X H, et al. FTGAN: a fully-trained generative adversarial networks for text to face generation. 2019. ArXiv:1904.05729
49 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, Montreal, 2014. 2672–2680
50 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015. ArXiv:1511.06434
51 Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks. 2018. ArXiv:1802.05957
52 Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks. 2018. ArXiv:1805.08318
53 Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation. 2017. ArXiv:1710.10196
54 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 4401–4410
55 Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of StyleGAN. 2019. ArXiv:1912.04958
56 Brock A, Donahue J, Simonyan K. Large scale gan training for high fidelity natural image synthesis. 2018. ArXiv:1809.11096
57 Gui J, Sun Z J, Wen Y G, et al. A review on generative adversarial networks: algorithms, theory, and applications. 2020. ArXiv:2001.06937
58 Cheng X Y, Xie L, Zhu J X, et al. Review of generative adversarial network. Comput Sci, 2019, 46: 74–81 [程显毅, 谢璐, 朱建新, 等. 生成对抗网络 GAN 综述. 计算机科学, 2019, 46: 74–81]
59 Liang J J, Wei J J, Jiang Z F. Generating against network GAN overview. J Front Comput Sci Technol, 2019, 14: 1–17 [梁俊杰, 韦舰晶, 蒋正锋. 生成对抗网络 GAN 综述. 计算机科学与探索, 2019, 14: 1–17]
60 Gao Y F, Hu Y J, Yu Z Q, et al. Evaluation and comparison of five popular fake face detection networks. J Appl Sci, 2019, 37: 590–608 [高逸飞, 胡永健, 余泽琼, 等. 5 种流行假脸视频检测网络性能分析和比较. 应用科学学报, 2019, 37: 590–608]
61 Nguyen T T, Nguyen C M, Nguyen D T, et al. Deep learning for deepfakes creation and detection. 2019. ArXiv:1909.11573
62 Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: a survey of face manipulation and fake detection. 2020. ArXiv:2001.00179
63 Verdoliva L. Media forensics and deepfakes: an overview. 2020. ArXiv:2001.06564
64 Zhang Y, Zheng L L, Thing V L L. Automated face swapping and its detection. In: Proceedings of IEEE International

Conference on Signal and Image Processing, Singapore, 2017. 15–19
65 Li H D, Chen H, Li B, et al. Can forensic detectors identify gan generated images? In: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Honolulu, 2018. 722–727
66 Korshunov P, Marcel S. Deepfakes: a new threat to face recognition? Assessment and detection. 2018. ArXiv:1812.08685
67 Li H D, Li B, Tan S Q, et al. Detection of deep network generated images using disparities in color components. 2018. ArXiv:1808.07276
68 He P S, Li H L, Wang H X. Detection of fake images via the ensemble of deep representations from multi color spaces. In: Proceedings of IEEE International Conference on Image Processing, Taipei, 2019. 2299–2303
69 Li Y Z, Lyu S W. Exposing deepfake videos by detecting face warping artifacts. 2018. ArXiv:1811.00656
70 Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. In: Proceedings of IEEE Winter Applications of Computer Vision Workshops, Hawaii, 2019. 83–92
71 Durall R, Keuper M, Pfreundt F J, et al. Unmasking deepfakes with simple features. 2019. ArXiv:1911.00686
72 Yu N, Davis L S, Fritz M. Attributing fake images to GANs: learning and analyzing GAN fingerprints. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 7556–7566
73 Nguyen H H, Fang F M, Yamagishi J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos. 2019. ArXiv:1906.06876
74 Huang Y H, Juefei-Xu F, Wang R, et al. FakeLocator: robust localization of GAN-based face manipulations via semantic segmentation networks with bells and whistles. 2020. ArXiv:2001.09598
75 Wang R, Ma L, Juefei-Xu F, et al. Fakespotter: a simple baseline for spotting ai-synthesized fake faces. 2019. ArXiv:1909.06122
76 Nguyen H H, Yamagishi J, Echizen I. Capsule-forensics: using capsule networks to detect forged images and videos. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, 2019. 2307–2311
77 Liu Z Z, Qi X J, Jia J Y, et al. Global texture enhancement for fake face detection in the wild. 2020. ArXiv:2002.00133
78 McCloskey S, Albright M. Detecting GAN-generated imagery using saturation cues, In: Proceedings of IEEE International Conference on Image Processing, Taipei, 2019. 4584–4588
79 Nataraj L, Mohammed T M, Manjunath B S, et al. Detecting GAN generated fake images using co-occurrence matrices. 2019. ArXiv:1903.06836
80 Kumar P, Vatsa M, Singh R. Detecting Face2Face facial reenactment in videos. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, Colorado, 2020. 2589–2597
81 Zhang X, Karaman S, Chang S F. Detecting and simulating artifacts in gan fake images. 2019. ArXiv:1907.06515
82 Stehouwer J, Dang H, Liu F, et al. On the detection of digital face manipulation. 2019. ArXiv:1910.01717
83 Mo H X, Chen B L, Luo W Q. Fake faces identification via convolutional neural network. In: Proceedings of ACM Workshop on Information Hiding and Multimedia Security, Innsbruck, 2018. 43–47
84 Xuan X S, Peng B, Wang W, et al. On the generalization of GAN image forensics. In: Proceedings of Chinese Conference on Biometric Recognition, Zhuzhou, 2019. 134–141
85 Zhou P, Han X T, Morariu V I, et al. Two-stream neural networks for tampered face detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, 2017. 1831–1839
86 Dang L M, Hassan S I, Im S, et al. Face image manipulation detection based on a convolutional neural network. Expert Syst Appl, 2019, 129: 156–168
87 Marra F, Gragnaniello D, Verdoliva L, et al. Do GANs leave artificial fingerprints? In: Proceedings of IEEE Conference on Multimedia Information Processing and Retrieval, Shenzhen, 2019. 506–511
88 Tarasiou M, Zafeiriou S. Extracting deep local features to detect manipulated images of human faces. 2019. ArXiv:1911.13269
89 Li J, Shen T, Zhang W, et al. Zooming into face forensics: a pixel-level analysis. 2019. ArXiv:1912.05790
90 Jeon H, Bang Y, Woo S S. FDFtNet: facing off fake images using fake detection fine-tuning network. 2020. ArXiv:2001.01265
91 Li Y Z, Chang M C, Lyu S W. In ICTU oculi: exposing AI created fake videos by detecting eye blinking. In: Proceedings of IEEE International Workshop on Information Forensics and Security, Hong Kong, 2018

92 Yang X, Li Y Z, Lyu S W. Exposing deep fakes using inconsistent head poses. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, 2019. 8261–8265
93 Ciftci U A, Demir I. FakeCatcher: detection of synthetic portrait videos using biological signals. 2019. ArXiv:1901.02212
94 Fernandes S, Raj S, Ortiz E, et al. Predicting heart rate variations of deepfake videos using neural ODE. In: Proceedings of IEEE International Conference on Computer Vision Workshops, Seoul, 2019
95 Güera D, Delp E J. Deepfake video detection using recurrent neural networks. In: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance, Auckland, 2018
96 Amerini I, Galteri L, Caldelli R, et al. Deepfake video detection through optical flow based CNN. In: Proceedings of IEEE International Conference on Computer Vision Workshops, Seoul, 2019
97 Afchar D, Nozick V, Yamagishi J, et al. Mesonet: a compact facial video forgery detection network. In: Proceedings of IEEE International Workshop on Information Forensics and Security, Hong Kong, 2018
98 Agarwal S, Farid H, Gu Y, et al. Protecting world leaders against deep fakes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, 2019. 38–45
99 Baltrusaitis T, Zadeh A, Lim Y C, et al. Openface 2.0: facial behavior analysis toolkit. In: Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, 2018. 59–66
100 Sabir E, Cheng J X, Jaiswal A, et al. Recurrent convolutional strategies for face manipulation detection in videos. 2019. ArXiv:1905.00582
101 Cozzolino D, Thies J, Rössler A, et al. Forensictransfer: weakly-supervised domain adaptation for forgery detection. 2018. ArXiv:1812.02510
102 Du M N, Pentyala S, Li Y N, et al. Towards generalizable forgery detection with locality-aware autoencoder. 2019. ArXiv:1909.05999
103 Marra F, Saltori C, Boato G, et al. Incremental learning for the detection and classification of GAN-generated images. 2019. ArXiv:1910.01568
104 Wang S Y, Wang O, Zhang R, et al. CNN-generated images are surprisingly easy to spot... for now. 2019. ArXiv:1912.11035
105 Li L Z, Bao J M, Zhang T, et al. Face x-ray for more general face forgery detection. 2019. ArXiv:1912.13458
106 Sheng C X, Zhang H G, Wang H M, et al. Research and development of trusted computation. Sci Sin Inform, 2010, 40: 139–166 [沈昌祥, 张焕国, 王怀民, 等. 可信计算的研究与发展. 中国科学: 信息科学, 2010, 40: 139–166]
107 Zhang S S, Zuo X, Liu J W. The problem of the adversarial examples in deep learning. Chinese J Comput, 2019, 8: 15 [张思思, 左信, 刘建伟. 深度学习中的对抗样本问题, 计算机学报, 2019, 8: 15]
108 Dong Y P, Su H, Wu B Y, et al. Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7714–7722
109 Zhang H C, Wang J Y. Towards adversarially robust object detection. In: Proceedings of IEEE International Conference on Computer Vision, 2019. 421–430
110 Huang R, Fang F M, Nguyen H H, et al. Security of facial forensics models against adversarial attacks. 2019. ArXiv:1911.00660
111 Neves J C, Tolosana R, Vera-Rodriguez R, et al. Real or fake? Spoofing state-of-the-art face synthesis detection systems. 2019. ArXiv:1911.05351
112 Ruiz N, Bargal S, Sclaroff S. Disrupting deepfakes: adversarial attacks against conditional image translation networks and facial manipulation systems. 2020. ArXiv:2003.01279
113 Huang Y H, Juefei-Xu F, Wang R, et al. FakePolisher: making deepfakes more detection-evasive by shallow reconstruction. 2020. ArXiv:2006.07533
114 Liu Z W, Luo P, Wang X G, et al. Deep learning face attributes in the wild. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 3730–3738
115 Khodabakhsh A, Ramachandra R, Raja K, et al. Fake face detection methods: can they be generalized? In: Proceedings of International Conference of the Biometrics Special Interest Group, Darmstadt, 2018
116 Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: learning to detect manipulated facial images. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019
117 Li Y Z, Yang X, Sun P, et al. Celeb-DF: a new dataset for deepfake forensics. 2019. ArXiv:1909.12962

118 Dolhansky B, Howes R, Pflaum B, et al. The deepfake detection challenge (DFDC) preview dataset. 2019. ArXiv:1910.08854
119 Jiang L M, Wu W, Li R, et al. DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection. 2020. ArXiv:2001.03024
120 Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: image synthesis using neural textures. ACM Trans Graph, 2019, 38: 1–12

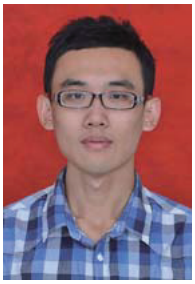
Deep visual identity forgery and detection

Chunlei PENG[1], Xinbo GAO[2*], Nannan WANG[1*] & Jie LI[1]

1. *State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an* 710071*, China*;
2. *Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing* 400065*, China*
* Corresponding author. E-mail: gaoxb@cqupt.edu.cn, nnwang@xidian.edu.cn

Abstract With the wide application of deep learning techniques in video and image generation, the quality of visual identity forgery, especially face forgery, is becoming increasingly high. The detection of visual identity forgery has been a hot issue because of its important influence on both national security and social stability. In this paper, we introduce recent researches on deep visual identity forgery from target-specific face forgery and target-generic face forgery. We further summarize the techniques of detecting visual identity forgery from multiple categories, including the spatial clue based methods, the temporal clue based method, the techniques for generalizable forgery detection and spoofing forgery detection models. We later present the public datasets and performance of representative approaches on these datasets. Finally, the issues and challenges in existing research are discussed.

Keywords deepfake, face swap, face manipulation, expression reenactment, face generation, forgery detection



Chunlei PENG received his B.Sc. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2012. He received his Ph.D. degree in information and telecommunications engineering from Xidian University in 2017. Now, he works with the School of Cyber Engineering at Xidian University. From September 2016 to September 2017, he has been a visiting Ph.D. student with Duke University, NC, USA. His current research interests include computer vision, pattern recognition, and machine learning.

Nannan WANG received his B.Sc. degree in information and computation science from Xian University of Posts and Telecommunications in 2009. He received his Ph.D. degree in information and telecommunications engineering from Xidian University in 2015. Now, he works with the State Key Laboratory of Integrated Services Networks at Xidian University. From September 2011 to September 2013, he has been a visiting Ph.D. student with the University of Technology, Sydney, NSW, Australia. His current research interests include computer vision, pattern recognition, and machine learning.

Xinbo GAO received his B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a postdoctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong professor of Ministry of Education, a professor of Pattern Recognition and Intelligent System, and the director of the State Key Laboratory of Integrated Services Networks, Xi'an, China. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications.

Jie LI received his B.Sc. degree in electronic engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems, from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively. She is currently a professor at the School of Electronic Engineering, Xidian University, China. Her research interests include image processing and machine learning.