视觉身份隐私保护：人脸匿名化研究方法

彭春蕾[1)] 苗紫民[1)] 刘德成[1)] 王楠楠[1)] 高新波[2)]

[1)](西安电子科技大学空天地一体化综合业务网全国重点实验室 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)

摘 要 随着深度学习的广泛应用，身份伪造技术的发展越来越迅猛. 各种伪造的图像和视频在社交媒体平台上的传播直接影响了公共隐私安全，人脸身份隐私保护已成为当前研究热点. 本文从基于图像和视频两个方面的匿名化方法阐述和归纳了人脸隐私保护研究现状，并将人脸图像匿名化方法从图像语义修改、图像语义保持、视觉可恢复以及深度学习过程中的人脸隐私保护四个方面进行分类，将人脸视频匿名化方法从聚焦面部区域隐私的视频匿名化方法和面向生物特征隐私的视频匿名化方法两个方面进行分类. 在此基础上，本文进一步介绍目前广泛使用的数据集及匿名算法评价标准，分析现有人脸匿名技术生成人脸图像的可靠性和实用性，并对此领域的未来研究进行了展望.

关键词 深度学习；身份伪造；隐私保护；人脸图像匿名；人脸视频匿名；公共安全

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2023.02431

Visual Identity Privacy Protection: Research Methods of Face Anonymization

PENG Chun-Lei[1)] MIAO Zi-Min[1)] LIU De-Cheng[1)] WANG Nan-Nan[1)] GAO Xin-Bo[2)]

[1)](*State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an* 710071)

[2)](*Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

Abstract With the wide application of deep learning, the development of identity forgery technology is more and more rapid. At present, face forgery has reached the level of fake to real, which can generate fake videos with real facial expressions and body movements. In addition, various fake images and videos spread rapidly on social media platforms, which directly affects public privacy security. What is more, these fake images or videos have a significant impact on national security and personal privacy. Therefore, face privacy protection should be paid attention to and corresponding measures should be taken. Recently, more and more people begin to devote themselves to studying how to prevent the abuse of identity forgery technology, and face identity privacy protection has become a current research hotspot. Therefore, this paper describes and summarizes the research status of face privacy protection from two aspects of image and video anonymization methods. And this paper classifies the face image anonymization methods from four aspects: the anonymization method of image semantic modification, the anonymization method of image semantic preservation, the anonymization method of visual recovery, and the face privacy protection in the process of deep learning. The anonymization method of image semantic modification refers to the differ-

收稿日期：2022-05-29；在线发布日期：2023-01-19. 本课题得到国家自然科学基金（No.62276198，No.U22A2035，No.U22A2096，No.61806152）、陕西省重点研发计划（No.2023-YBGY-231）、广西自然科学基金（No.2021GXNSFDA075011）资助. **彭春蕾**，博士，副教授，CCF会员，主要研究领域为计算机视觉、模式识别和机器学习. E-mail: clpeng@xidian.edu.cn. **苗紫民**，硕士，主要研究领域为计算机视觉、模式识别和机器学习. **刘德成**，博士，讲师，主要研究领域为计算机视觉、模式识别和机器学习. **王楠楠**（通信作者），博士，教授，主要研究领域为计算机视觉、模式识别和机器学习. E-mail: nnwang@xidian.edu.cn. **高新波**，博士，教授，主要研究领域为计算机视觉、模式识别和机器学习.

ence in visual effect between the original image and the image after anonymity. Semantics-preserving image anonymization means that the visual identity information of the original image and the image after anonymization are the same. Visually recoverable anonymization means that the anonymized image can be restored to the original image. Face privacy protection methods in the process of deep learning generally study the impact of detail factors in the process of privacy protection on the anonymity effect. And this paper classifies the face video anonymization methods from two aspects: the video anonymization methods focusing on facial region privacy and the video anonymization methods oriented to biometric privacy. The video anonymization method focusing on facial region privacy aims to generate anonymous face videos that are different from the original face, while making the anonymous face still retain other identity- independent biometric characteristics. Video anonymization for biometric privacy refers to the removal of physiological signals to achieve video anonymization. On this basis, this paper further introduces the widely used datasets and evaluation criteria of anonymous algorithms. In addition, this paper collects the experimental datasets and main evaluation indicators of representative anonymization methods to facilitate the subsequent research of researchers. Furthermore, this paper compares the experimental results of representative anonymization methods, and analyzes the reliability and practicability of the existing face anonymization techniques to generate face images. With people's attention to identity privacy protection, face anonymization technology is more and more advanced. However, this paper finds that although some methods generate good visual effects of face images, many methods have higher requirements for datasets, and the generalization performance of the model needs to be further improved. As people pay more attention to personal privacy data, this paper puts forward the corresponding research directions and prospects the future research in this field. With the continuous development of technology, privacy protection research will become more and more popular to prevent the malicious use of deepfake technology.

Keywords deep learning; identity forgery; privacy protection; face image anonymization; face video anonymization; public safety

1 引 言

近年来，人工智能技术的飞速发展影响了人类社会生活的各个方面，特别是人脸识别技术的广泛应用，给国家安防、社会医疗、互联网多媒体等领域带来了极大的便利. 然而，人脸识别技术的滥用所带来的个人隐私安全问题也越来越引起人们的担忧. 例如，售楼处等公共场所擅自使用人脸识别技术收集消费者人脸信息进行记录和营销分析；小区物业强制业主录入人脸信息并将人脸识别作为进出小区的唯一途径；移动应用迫使用户上传非必要的人脸信息；犯罪分子盗取人脸和身份信息用于贷款、诈骗和公开网上销售等.

在 AI 时代，大数据和算法的发展推动了全球化进程的加快，也导致了个人隐私数据的泄露. 某些私营企业收录并利用用户的个人照片及信息，创建人脸信息数据库，各种身份数据交织在一起形成了一张信息网，处于这张网下的用户处于“裸露”状态，因此保护公共隐私安全成为当务之急. 2018 年，剑桥分析被曝出该公司已经导致了 8000 多万用户数据泄露. 在 2016 年美国总统大选期间，剑桥分析向特朗普竞选团队提供了详细的美国选民数据，并利用 Facebook 带有偏重的算法向用户投放广告，在很大程度上影响了当年的美国总统大选. 2019 年 2 月，国内一家专注于安防领域的人工智能公司 SenseNet 发生了大规模的数据非法获取及泄露事件. 同年 9 月，某商家在网上公开售卖近 17 万的人脸数据信息，部分当事人表示自己毫不知情. 2020 年 11 月，“国内人脸识别第一案”宣判，法院判决动物园删除郭兵办理年卡时提交的面部特征信息并赔偿当事人损失①. 现如今，接连不断的隐私泄露案件使得人们对现有规章制度、运行机制和法律体系等进行了逐步地修改和完善. 2019 年 4 月，全国人

① 国内人脸识别第一案，http://www.xinhuanet.com/legal/2020-11/25/c_1126786381.htm.

大常委会审议的《民法典人格权编（草案）》里，正式加入了一条规定：任何组织和个人不得利用信息技术手段伪造的方式侵害他人的肖像权①. 为促进网络音视频信息服务健康有序发展，保护公民、法人和其他组织的合法权益，维护国家安全和公共利益，国家互联网信息办公室、文化和旅游部、国家广播电视总局制定了《网络音视频信息服务管理规定》，该规定于 2020 年 1 月 1 日正式实行②. 美国近年来陆续出台了《恶意伪造禁令法案》、《2019 年商业人脸识别隐私法案》、《停止秘密监视条例》③和《深度伪造责任法案》等法律法规. 与此同时，新加坡通过了《防止网络假信息和网络操纵法案》，旨在使政府有权要求个人或网络平台更正或撤下对公共利益造成负面影响的假新闻. 2018 年 5 月 25 日，欧盟出台的《通用数据保护条例》规定所有在欧盟范围内的公司或地点在欧盟范围外、但与欧盟个人相关、提供服务等行为的公司都属于该条例适用范围内，这是一部面向世界的约束法规，为全球化的发展提供了有利的安全保障.

综上所述，视觉身份隐私数据泄露对国家安全和社会稳定都可能产生不可估量的影响. 除了出台相关立法，人们需要采取更多措施保护人脸隐私安全. 近年来，解决该问题的通用性方法便是人脸匿名化处理技术. 人脸匿名化是指将人脸信息经过处理使其无法识别出特定身份的过程，如图 1 所示. 根据匿名化的目标不同，人脸匿名化技术可以分为基于视觉内容修改的方法、基于视觉内容保持的方法、视觉可恢复匿名方法和深度学习过程中的隐私保护问题等. 其中，视觉内容修改指在人眼观看上，图像匿名化前后视觉效果不一样；视觉内容保持指匿名前后人脸视觉样貌不变，然而人脸识别系统却认为匿名前后的图像身份不一致. 视觉可恢复匿名方法是指匿名化后的图像能够恢复成匿名前的图像，整个过程是“可逆”的. 深度学习中的隐私保护则探讨了图像匿名过程中的细节因素对匿名效果产生的影响，例如数据集的处理方式对训练效果的影响以及模型训练过程中是否泄露隐私数据等问题. 传统的人脸匿名化手段包括模糊、马赛克和遮挡等，即通过掩盖人脸面部区域实现匿名化. 随着深度生成网络技术的发展，人脸匿名化可以在保持面部表情、姿态等身份无关属性不变的前提下，生成虚假人脸且不影响匿名后人脸的视觉质量. 2021 年，Pataranutaporn 等人[1]指出，使用人工智能算法生成的虚拟人脸可以代替真实人脸，能够应用在视频会议研讨、视频医疗问诊等场合进行隐私保护. 与此同时，人脸匿名化生成的虚假图像逼真程度对于其应用推广十分重要，需要利用人脸伪造检测技术进行监督评估和对抗改善. 匿名人脸很大程度上保护了个人隐私安全，阻碍了不法机构窃取身份信息数据. 人脸匿名化技术的发展在解决隐私泄露问题的同时，也促进了人工智能领域技术的全面发展，在国家政治安全、经济安全、社会安全及网络安全等

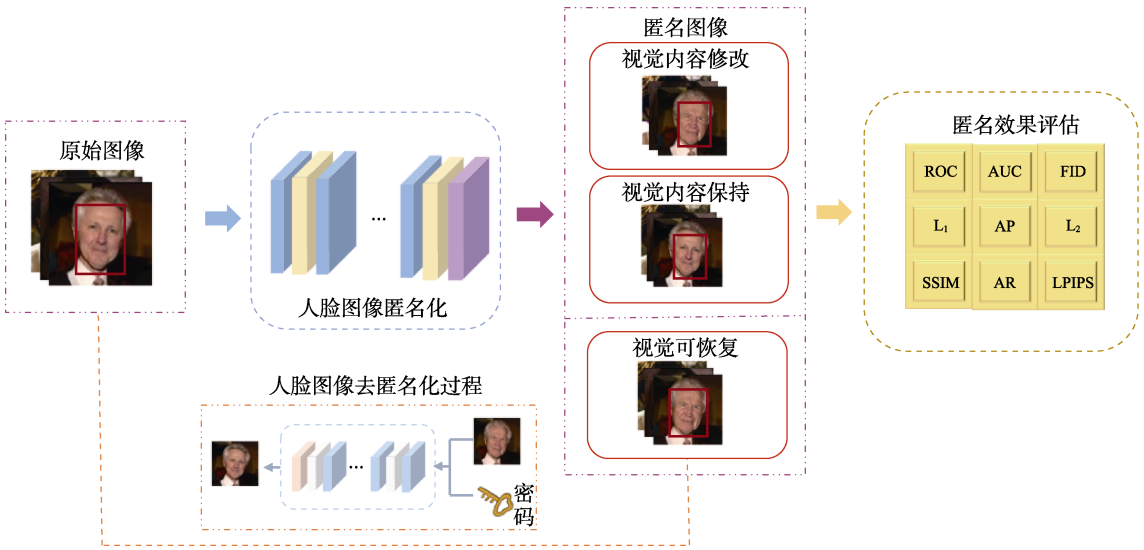


图 1　人脸匿名化定义和目标

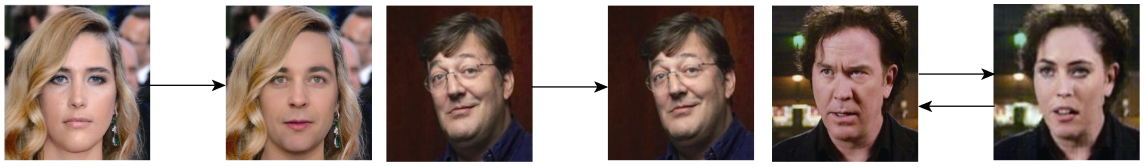
①《民法典》，http://www.npc.gov.cn/npc/c36502/201904/263ca093f6d0- 4741b1cbb703a0015a29.shtml.
②《网络音视频信息服务管理规定》，http://www.cac.gov.cn/2019-11/ 29/c_1576561820967678.htm.
③《停止秘密监视条例》，https://www.eff.org/document/stop-secret- surveillance-ordinance-05062019.

领域扮演着重要的角色.

目前越来越多的科研工作者投身于身份匿名化和隐私保护的研究中，提出了各种各样的隐私保护方法. 本文将针对图像和视频两个方面介绍现有方法，如表 1 所示. 由于人们对基于视频的隐私保护研究正处于起步阶段，本文将重点放在基于图像的研究中，并将人脸图像匿名化方法分为四类. 第一类为图像语义修改的匿名化方法，此类方法匿名前后的图片视觉效果不同，如早期的 K-Same 系列方法[2,3,16]；第二类为图像语义保持的匿名化方法，该类方法匿名前后图片的视觉身份信息相同，即不改变原图的外貌特征，如基于身份无关属性修改的方法[8,9,17-21]；第三类为视觉可恢复的匿名化方法，如基于密码的匿名化与去匿名化方法[10,11,22]，上述人脸匿名化方法效果示意图如图 2 所示；第四类为深度学习过程中的人脸隐私保护方法，该类别一般研究隐私保护过程中的细节因素对匿名效果的影响，如针对抠图任务的隐私保护方法[24]和图像匿名化方法对人脸检测模型的影响实验[25]. 在此基础上，本文介绍现有方法中常用的数据集及客观匿名化评估指标，并总结具有代表性方法的实验设置及结果，评估模型性能优劣. 最后，本文在分析现有匿名化方法的不足之处基础上，对未来隐私保护的研究方向和趋势进行了展望.

表 1 人脸匿名化方法

人脸匿名化方法			主要方法
人脸图像匿名化方法	图像语义修改的匿名化方法	传统人脸匿名化方法	K-Same[2]、K-Same-Net[3]
		基于深度学习的人脸匿名化方法	DeepBlur[4]、IdentityDP[5]
	图像语义保持的匿名化方法	基于对抗攻击的身份匿名化方法	FoggySight[6]、Fawke[7]
		身份保持的属性匿名化方法	SAN[8]、PrivacyNet[9]
	视觉可恢复的匿名化方法	–	FIT[10]、MfM[11]
	深度学习过程中的人脸隐私保护	–	PriMIA[12]
人脸视频匿名化方法	聚焦面部区域隐私的视频匿名化方法	–	CIAGAN[13]、JaGAN[14]
	面向生物特征隐私的视频匿名化方法	–	Pulseedit[15]



(a) 图像语义修改[23] (b) 图像语义保持[10] (c) 视觉可恢复[10]

图 2 人脸匿名化方法效果示意图

2 人脸图像匿名化方法

社交媒体的发展促进了图像的广泛传播和便捷获取，一些虚假信息经传播后，容易对国家政治安全、经济发展、网络安全等各方面造成巨大的影响. 例如，不法分子利用深度伪造技术故意抹黑政治领导人物影响国家发展、制造恶意视频破坏公众人物形象导致公司经济状况受到影响、在网络平台伪造虚假视频欺诈钱财对公民自身安全造成危害等. 因此，数据的肆意传播使得隐私保护成为有待解决的问题. 本章节将对基于图像数据的人脸匿名隐私保护方法展开介绍，并将人脸图像匿名化分为图像语义修改的匿名化方法、图像语义保持的匿名化方法、视觉可恢复的匿名化方法以及深度学习过程中的人脸隐私保护四部分，进而对每一部分内容进行详细分类和讨论.

2.1 图像语义修改的匿名化方法

本节将介绍图像语义修改的匿名化方法，并将其细分为传统人脸匿名化方法和基于深度学习的人脸匿名化方法，上述方法主要指视觉内容修改的匿名化. 在日常生活中，监控设备无所不在，人们早已习惯生活在摄像头的监视之下，商场、学校、公司、家庭等各种场所基本都需要安装监控设备来提供保障和便利. 然而在享受便利的同时，社会对监控设备也产生了恐惧. 人们不希望自己的个人隐私过于暴露，一般情况下会采取遮挡或模糊头部区域的方法对图像中的人脸隐私进行保护，但此类方法生成的匿名图像视觉效果极差，同时遮挡人脸依然

可能被人脸识别技术检测到[26]. 因此，越来越多的科研工作者开始研究视觉内容修改的匿名化方法，其目标是在保证匿名图像真实自然的前提下，改变原始人脸的视觉特征，达到人脸隐私保护的目的.

2.1.1　传统人脸匿名化方法

早期的人脸匿名化方法包括模糊、马赛克及掩蔽等. 模糊和马赛克是对图像进行下采样或平滑操作，将图像中的像素值进行平滑处理或在特定区域进行色块打乱，形成模糊或像素化的效果，从而达到隐匿人脸样貌隐私的目的. 掩蔽是指将选定的图像区域进行遮挡来移除原有的人脸外貌隐私. 上述方法均通过直接覆盖人脸区域实现人脸匿名效果，然而匿名后的图像往往视觉效果差且基本丢失了人脸数据，如图 3 所示.

图 3　不同匿名方法的生成图像（其中生成人脸图像参考文献[10]）

卡内基梅隆大学 Newton 等人[2]提出一种名为 K-Same 的人脸去识别方法. 该方法将数据集中所有人脸图像经旋转、剪裁等操作，大致对齐在相同位置（比如将每张图片中的人眼位置对齐）; 然后计算每张人脸图像之间的距离，匿名图像用 K 个最相似的人脸图像平均值表示. 然而，该算法的人脸识别效果有限，并且去识别后的人脸图像容易出现不清晰或伪影等现象.

由于计算平均脸的方法会引入严重的噪声，美国卡耐基梅隆大学 Gross 等人[16]提出一种基于人脸表征模型的人脸去识别方法,即 K-Same-M 算法. 该算法在 K-Same 算法的基础上进行改进，结合主动外观模型 AAMs（Active Appearance Models）计算人脸的外观参数,再根据 L2 范数得出最接近外观参数的 K 个向量，最后计算 K 个向量的均值作为人脸去识别图像. 在识别精度和数据利用率方面，该算法均优于 K-Same 算法.

卢布尔雅那大学计算机与信息科学学院 Meden 等人[3]将 GNN 与 K-Same 算法结合，提出了一种名为 K-Same-Net 的人脸去识别算法. 该团队首先利用 K-Same 算法生成 K 张图像的聚类，在原始图像集和替代的人脸图像集上建立一一对应的关系；然后将替代图像集输入到 GNN 中进行训练,并结合外观参数生成视觉特征可控的匿名人脸图像.

上述隐私保护方法试图隐藏敏感信息，在人脸图像上放置黑色掩膜或将与面部对应的图像区域替换成没有提供信息的代理图像[2,3,16]. 虽然这些方法避免了隐私数据泄露，但大多方法生成的匿名图像伪影现象较严重，影响了图像的后续使用.

2.1.2　基于深度学习的人脸匿名化方法

随着深度学习技术的发展，为了进一步提升人脸匿名图像的质量，人脸匿名化任务可以利用基于图像修复的方法实现. 为了解决社交媒体照片中身份模糊的问题，德国萨尔大学的 Sun 等人[27]提出了两阶段头部修复方法. 该方法将人脸关键点与被遮挡人脸的图像共同输入到头部生成模型中以获取头部姿态等信息. 其中第一阶段的网络结构由编码器和解码器组成，编码器将输入图像编码为隐向量，解码器根据隐向量获取人脸关键点坐标. 第二阶段网络结构由头部生成器和头部判别器组成，头部生成器基于 U-Net 体系结构，能够根据周围的环境和关键点等信息生成自然的头部图像. 头部判别器借鉴了 DCGAN[28]结构，以提高网络的稳定性. 该方法生成的人脸图像视觉效果较逼真，但图像的分辨率较低，人脸边界还会出现模糊现象.

普渡大学 Li 等人[4]提出了一种简单而有效的图像修复方法——DeepBlur. DeepBlur 能够利用无条件训练的生成对抗网络改变原始图像深层特征控制高质量图像合成. 该方法使用 VGG16[29]等模型提取输入图像的特征，并通过生成模型（如 StyleGAN[30]）将潜在特征合成输出. 与传统方法进行比较，DeepBlur 可以消除人脸上的伪影和遮挡现象，还能很好地平衡计算效率和图像质量之间的关系.

通过基于人脸替换的方法，也可以实现视觉内容的修改完成人脸匿名化. 首先 Sun 等人[27]仅通过简单的数据驱动方法大幅度修改人脸，并没有考虑图像生成过程缺乏可控性导致生成伪影的问题. 于是 Sun 等人[31]进一步提出了一种基于数据驱动和参数化人脸模型的人脸替换方法. 该团队使用参数脸部模型[32]控制人的身份，还能保留原始图像中诸如面部姿势和表情之类的属性. 团队具体通过两阶段头部替换方法来模糊图像中的身份，在第一阶段使用不同身份的渲染人脸替换原始人脸，从而改变与原始人脸相关的身份向量；在第二阶段使用生成对抗网络模型合成完整的头部图像，并基于细粒度特征学习得到逼真的输出图像. 其中判别器和损失函数的设置与先前工作类似，然而在数据集方面，该方法对图像进行了更加细致的处理，如剪除较强轮廓特

征的图像以及图像大小标准化等.

由于某些网络对特征的放大和抑制，一些方法[27,31]产生的匿名图像与原始图像具有显著不同的视觉外观，并且生成的匿名图像大多数不自然. 虽然基于对抗攻击的方法产生的对抗性扰动具有较小的强度变化，但生成的伪影仍然会影响人类感知的视觉质量. 因此，Yang 等人[33]提出了一种有针对性的身份保护迭代方法①, 可以生成覆盖在人脸图像上的对抗性身份面具，从而在不牺牲视觉质量的情况下隐藏原始身份. 为了模拟现实世界的场景，该团队收集了包含目标身份的人脸图像作为原始图像的替代图像，并采用白盒模型来生成受保护的图像，以提高图像对未知人脸识别系统的可转移性.

人脸匿名化任务还可以借助基于特征解构的方法将身份相关特征与身份无关特征拆分来实现. 给定一张人脸图像，如何才能在隐藏真实身份、仍允许人脸检测器工作的情况下, 创建另一张外观相似、背景相同的图像？模糊和马赛克等方法严重影响了视觉质量；K-Same 系列方法[2,3,16]无法充分利用现有的数据，而且大多数匿名图像的伪影现象严重；基于生成对抗网络的方法[34-36]很难生成视觉相似的匿名人脸，隐私保护效果和数据利用率之间也存在权衡问题. 为了应对这些挑战, Wen 等人[5]提出基于特征解构的人脸匿名化框架, 并命名为 IdentityDP. 它能够生成与原始人脸相似的匿名人脸图像，且不破坏人脸检测器的可用性. 该框架将基于数据驱动的深度神经网络与差分隐私机制相结合，主要包括三个阶段：人脸表征分解、在身份表征上添加扰动混淆身份和图像重建. 具体来说，该算法首先提取需要解构的身份特征和属性特征，然后在设计的 IdentityDP 机制下生成扰动身份表征，最后从扰动身份表征和原始属性表征中构造匿名人脸. 该方法能够有效模糊人脸的身份相关信息，保留重要的视觉相似性，并且生成的高质量图像能够用于身份不可知的计算机视觉任务，如检测和跟踪等.

Nousi 等人[37]提出同时包含有监督和无监督训练能力的匿名方法，可以利用多样化的度量矩阵来衡量匿名化人脸的真实性和自然性，以提高匿名化后人脸图像的视觉质量. 该方法利用自编码器提取有用的低维特征，并微调自编码器的编码部分，改变人脸身份信息并保留其他属性特征. 对于有监督的属性保留匿名方法来说，网络迫使编码后的人脸表征更远离冲突属性的表征, 更接近目标属性的表征，使得匿名人脸图像的视觉效果更自然. 为了确保属性保留，该团队利用吸引矩阵和排斥矩阵对人脸属性进行操作，合并这两个矩阵使得解码器生成人眼和人脸识别系统都辨别不出的匿名人脸. 对于无监督的去识别方法来说，该团队进一步定义了没有明确属性信息的吸引矩阵，迫使编码好的人脸表征被解码成不同的身份信息，并利用整张人脸或不同人脸部分的平均特征约束吸引矩阵的样本，使生成的人脸身份信息不同于原始人脸，人脸特征分布与原始人脸接近.

近些年来比较流行的生成对抗网络是一种基于卷积神经网络的深度学习方法，生成对抗网络的出现基于“博弈对抗”的思想，生成器生成的图像尽量真实可靠，以便可以欺骗判别器；而判别器的任务是要鉴别生成器生成图像的真假. 生成器和判别器互相“博弈”的过程，优化了模型参数，使网络结构更加健壮. 利用生成对抗网络进行虚假/伪造人脸的生成/编辑，也可实现人脸匿名化和隐私保护的目的. 因而，基于虚假人脸生成的方法可以应用到人脸匿名化任务当中来.

在保留原始数据分布的情况下，生成匿名人脸是一个非常严峻的挑战. 挪威科技大学的 Hukkelås 等人[34]提出了基于条件生成对抗网络的模型——DeepPrivacy，生成的匿名人脸图像能够保留原始图像的姿态和背景信息. 其中生成器的架构使用的是 U-Net 模型, 由编码器和解码器组成. 编码器和解码器在每个卷积时都有相同数量的滤波，但解码器在每个跳跃连接后都增加了一个瓶颈卷积，以此改变特征图的维度大小；进一步地，该算法将背景信息作为条件信息输入到判别器中，去掉小批量标准差层，增加其可变性. 在实验部分，该团队使用 WIDER-FACE 数据集评估匿名图像对人脸检测的影响，结果表明在很大程度上匿名人脸能被人脸检测算法检测到，即能够生成真实有效的人脸. 一般情况下, DeepPrivacy 能够保证去除大部分隐私敏感信息，且生成高质量的人脸图像，但在一些非传统的姿态条件下（如面部遮挡）, 生成的人脸图像质量不佳.

除此之外，天普大学 Wu 等人[35]提出一种名为 PP-GAN（Privacy-Protective-GAN）的虚假人脸生成模型以实现隐私保护. DeepPrivacy[34]直接去除整个人脸区域, 确保 100%去除原始人脸中的隐私敏感信息. 不同的是，PP-GAN 在 GAN 模型基础上添加了身份识别模块来移除部分生物特征信息，以及图像质量评价模块来保持生成图像与输入图像的结构相

① 上述方法代码地址，https://github.com/shawnxyang/tip-im.

似度一致. PP-GAN 生成器使用与 DeepPrivacy 类似的 U-Net 结构，较浅的卷积网络作为鉴别器，并在对抗性环境中使用多个学习目标来确保所生成图像的质量.

传统的匿名人脸技术要么生成的图像不够真实，要么无法在隐私性和可用性之间取得平衡. 因此，普渡大学计算机科学系 Li 等人[36]提出一种衡量匿名化程度的人脸去识别方法，能够提升匿名化程度可用性、增强隐私性. 该方法分为四个阶段：人脸属性估计、以隐私度量为导向的人脸混淆、定向自然图像和对抗扰动. 该团队采用 GoogLeNet[38]网络提取人脸特征，然后使用隐私保留属性算法 PPAS（Privacy-Preserving Attribute Selection）选择并更新人脸属性，使得人脸属性分布与真实分布相近. 为了生成真实自然的人脸图像，网络使用 StarGAN[39]模型并添加属性分类损失和图像重构损失来约束匿名图像的质量. 该方法在实现匿名人脸的基础上，可以达到衡量匿名化程度的目标.

很多基于虚假人脸生成的研究能够在保留原始面部姿势信息[13,31,34]的前提下，生成不同身份的人脸图像. 然而由于现实世界中缺乏具有相同面部姿势的不同身份的人脸图像，模型很难生成逼真自然的匿名人脸. 为了解决上述问题，Wang 等人[40]提出了一种新的基于虚假人脸生成的模型来实现人脸匿名化①. 该结构能够将原身份对应的视觉信息替换为任意一幅图像提供的条件身份信息. 为了保留原始人脸的几何属性（面部姿势和表情），并促进更自然的脸部生成，该团队利用二分图和深度学习模型模拟原始身份的人脸关键点和条件身份信息之间的关系，并进一步使用关键点注意力模型对人脸关键点进行手动选择，允许网络对关键点进行加权，以达到最佳的视觉效果.

虽然 Sun[31]所提出的人脸图像匿名方法能够有效地保留原始人脸表情和头部姿势信息，但对于身份匿名程度仍然不可控. 为此，Jeong 等人[41]提出了一个人脸身份可控的模型，能够增强数据的有效性. 该算法混合不同人脸图像的身份信息作为目标身份，并使用 SphereFace 编码器来区分身份和非身份特征以获得详细的可控性，从而平衡去身份程度和人脸属性保留程度.

Khojaste 等人[42]提出了一种基于遗传算法处理人脸图像的模型，可对输入的人脸图像进行不可察觉的编辑，以保护图像中人脸隐私信息. 该模型主要由三个部分组成：人脸遮罩模块通过提取人脸关键点得到图像中的人脸部分，优化模块利用生成器生成去识别图像，合并模块将原始图像背景信息添加到由优化模块生成的去识别图像中. 其中该算法在优化步骤中使用了不同的损失函数，尽可能使模型生成与输入图像相似的高质量图像. 实验结果表明，在保证不降低匿名图像质量的前提下，该模型能够匿名原始图像的身份信息，且匿名图像不会被人脸识别系统识别，在社交网络上仍然可以共享. 然而该模型并不能保证对所有人脸识别方法有效，仍有改进的空间.

生成性对抗网络能够生成接近照片真实感的人脸，然而将这些网络扩展到整个人体仍然是一项具有挑战性的任务. 因此，Hukkelås 等人[43]提出了一种可以生成完整人体图像的匿名化方法②. 该方法通过学习像素-表面映射来设计对抗网络，引入 V-SAM(Variational Surface Adaptive Modulation)将生成器的输入特征投影到表面自适应隐空间，使得生成器直接将变化的潜在因素映射到相关的表面位置，从而提取特征空间的潜在特征. 该方法可以在复杂多变的场景中合成具有不同外观的人类图像，然而在人脸细节部分，生成的图像大多还是存在些许扭曲和伪影现象.

上述图像语义修改的匿名化方法能够生成视觉效果逼真的人脸图像，可用于匿名人脸图像和视频数据. 尤其基于图像修复、人脸替换、特征解构和虚假人脸生成等方面的研究，为隐私保护相关领域提供了丰富的借鉴和算法改进思路. 此外，由于深度学习模型能够提取多种不同的特征，匿名效果和特征选择之间产生了竞争性权衡. 尤其是虚假人脸生成的研究，使无监督模型能够生成更加清晰真实的图像，同时也为相关领域提供了对抗学习、对抗样本防御、模型鲁棒性等方面的启示.

2.2　图像语义保持的匿名化方法

本节将介绍图像语义保持的匿名化方法，并将其分为基于对抗攻击的身份匿名化方法和身份保持的属性匿名化方法，这些方法主要指视觉内容保持的匿名化. 通常用户在社交媒体平台上传照片时，希望人类观察者可以清楚辨识照片中人物，同时隐藏自己的某些个人信息. 因此，人们提出使用基于对抗攻击和属性匿名等方法来达到视觉内容保持的效果. 一般情况下，基于对抗攻击的身份匿名化方

① 上述方法代码地址，https://github.com/fodark/anonygan.

②上述方法代码地址，https://github.com/hukkelas/full_body_anonymization.

法即在原始样本上添加一些人眼无法察觉的扰动，致使机器做出错误判断. 身份保持的属性匿名化方法便是通过修改身份无关属性，欺骗属性判别器并保持身份一致.

2.2.1 基于对抗攻击的身份匿名化方法

塞萨洛尼基亚里士多德大学团队[44]提出使用对抗样本生成的方法进行人脸去识别，可以保证人脸图像最小失真的情况下，欺骗人脸识别系统. 这种新的对抗攻击方法称为惩罚快速梯度值法 P-FGVM（Penalized Fast Gradient Value Method），既能保护身份隐私，又能保持匿名前后视觉内容的一致. P-FGVM 结合对抗损失和真实性损失（realism loss）更新梯度下降方程，进而生成目标对抗样本. 该方法能够生成错分率较高的匿名人脸图像，从而更加有效地保护人脸隐私.

一些私营企业擅自将社交媒体平台上的用户身份信息与照片关联，导致了大量的隐私泄露. 于是 Evtimov 等人[6]提出了一种名为 FoggySight 的方案，用于抑制人脸图像库的面部搜索. 当使用人脸识别技术对用户图像库进行照片匹配和查询用户身份时，神经网络便会提取这张照片的特征向量；网络将特征向量与人脸识别公司从社交媒体抓取的照片集（查找集）进行比较，返回与查询照片最近的 K 个照片集合（召回集）. 具体来说，该方案在自愿提供照片的用户图片上添加扰动创建诱饵图片，并将这些照片上传到社交媒体上，人脸识别公司对受保护的用户身份进行查询时，计算查询照片与查找集中图片的距离，最终只会返回最接近查询照片的诱饵照片，从而保护了用户照片隐私.

美国休斯顿大学 Hadi 等人[45]提出了一种人脸检测评分过滤 FDSF（Face Detection Score Filtering）对策来保护训练数据的隐私免受重建攻击. 该方法主要思想是将高置信度分数返回给人脸检测分数 FDS（Face Detection Score）较低的人脸图像，具有高置信度的低质量人脸图像可以欺骗攻击者，致使他们在寻找最优人脸时形成错误的搜索路径，以此达到保护数据隐私的目的. 为测试 FDSF 检测虚假人脸的能力，该团队将合成人脸图像与真实人脸图像进行比较，证明了 FDSF 可以成功地通过虚假人脸的高置信度分数来欺骗攻击者.

Shan 等人[7]提出了一种能够帮助用户抵御人脸识别模型系统的匿名方法——Fawkes①. 在用户发布照片之前，该方法在照片上添加难以察觉的像素级信息来实现视觉内容共享的效果. 因此，当人脸识别模型检测用户图片时，模型会错误识别该用户身份. 一般情况下，该系统可以很好地欺骗人脸识别模型，但对于大量未伪装的用户图像（如名人图片），Fawkes 的有效性会极大地降低.

基于对抗攻击的身份匿名化方法通过添加不同的噪声或对图像的某些区域进行特殊处理进而生成对抗样本，以此样本对网络模型进行攻击，且生成图像对于其他任务（例如性别或年龄识别）仍然可行. 因此，将对抗攻击技术应用于人脸匿名化任务中，控制神经网络模型的输出并保持与非对抗性图像的视觉相似性是一种很好的匿名手段.

2.2.2 身份保持的属性匿名化方法

机器学习的进步使得网络从生物特征数据中提取软生物属性信息成为可能，如年龄、性别、种族属性等[46-48]，因此在身份保持的前提下对属性进行匿名和隐私保护同样重要. Suo 等人[49]提出了一种基于人脸组件分解的方法来实现性别转换. 该方法将原始人脸图像分解成几个人脸组件，用来自异性组别的模板替换这些人脸组件，并使用无缝图像编辑技术合成匿名人脸. 其中，该方法通过选择与原始组件相似的替换模板优化图像编辑步骤，进而保留原始图像的身份. 基于此，密西根州立大学 Othman 等人[17]提出了针对人脸软生物特征的隐私保护概念，即保护数据库中人脸图像的性别信息并保留身份信息. 与 Suo[49]最大不同的是，该方法可以生成与原始人脸视觉特征相似且性别信息被抑制的多幅图像. 该方法具体采用了人脸融合的思路，将一对人脸（如一男一女）的人脸关键点生成三角形人脸网格，然后组合两幅图像实现变形过程，从而通过变形步骤组合性别信息. 由于每一张脸的分布很容易控制，模型能够在性别混淆和身份保留之间实现不同的权衡. 然而从合成效果上看，该方法虽然可以混淆性别，但生成图像也会产生非常明显的伪影.

受 Othman[17]工作的启发，密西根州立大学 Mirjalili 等人[18]提出了一种利用生成对抗扰动对性别信息保护的方法，能够在改变性别信息的条件下，不影响生物特征匹配器和人类观察者的辨别情况. 该方法在人脸关键点上应用了和 Othman[17]同样的 Delaunay 三角剖分，再以预先训练的性别分类器优化了三角形人脸网格内的像素强度(即颜色信息). 由于颜色变化很小，匿名图像可以混淆性别分类器，同时保留生物特征匹配模型的效用. 然而该算法没有明确地设计保留身份信息的模块，身份识别性能

① Fawkes 代码地址，https://github.com/Shawn-Shan/fawkes.

较差.

在后续工作中，Mirjalili 等人[8]提出一种半对抗网络 SAN（Semi-Adversarial Networks）模型用于人脸图像的隐私保护①，即利用卷积自编码器将输入图像转化成对性别属性匿名化处理的人脸图像. SAN 能够混淆性别分类器，隐私增强的数据还可用于身份验证，且不会对验证性能造成影响. 该方法将性别属性分类模型与人脸识别模型组成半对抗模块添加到自动编码器当中，并在目标函数中约束匿名化后人脸图像的真实程度、性别属性的匿名程度以及用于人脸识别的身份特征一致性. 其中，身份保持的属性匿名化方法结构如图 4 所示. 然而，SAN 不适用于未经过训练的性别分类器，即没有纳入训练阶段的分类器仍然可以准确地对性别进行分类. 于是该团队[20]又提出一个集成的 SAN 模型来解决泛化问题，将多个独立训练的 SAN 网络集成，并为 SAN 模型的辅助性别分类器组件设计了三种不同的训练方案，增强了模型之间的多样性. 不同的数据增强技术使得模型为给定的输入图像生成一组不同的扰动输出图像，并确保至少有一种扰动的输出人脸图像可以混淆任意的性别分类器. 进一步针对以往 SAN 模型仅能隐藏性别信息的问题，Mirjalili 等人[9]又提出了一种新的面向多属性人脸隐私的半对抗网络模型——PrivacyNet，用于对多种属性信息进行处理. 该方法将 SAN 模块与生成对抗网络相结合来扰乱输入图像，并使用属性分类器将人脸图像转化到三个相互独立的属性方向（性别、年龄和种族），从而保持身份识别能力，并实现多种属性的隐私保护.

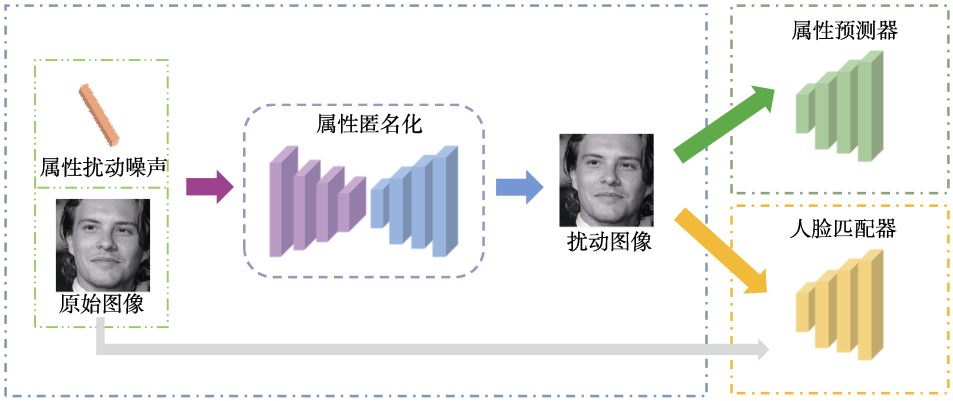


图 4　身份保持的属性匿名化方法示意图（方法流程主要参考文献[8]）

为了欺骗软生物特征分类器，印度新德里电子和信息技术部 Chhabra 等人[19]进一步利用对抗扰动的方法实现多个人脸属性的同步匿名化. 该方法能够在保留身份信息和视觉内容的前提下，改变一个或多个软生物特征属性. 在多属性匿名设计阶段，该方法基于 Carlini-Wagner L2 攻击[50]在图像中嵌入不可感知的噪声，使得分类器难以自动推断隐私增强的目标属性，从而根据用户的选择保留信息. 与上面讨论的 SAN 模型类似，这种方法不能很好地推广到任意的分类器.

Li 等人[21]提出一个能够混淆视觉外观，并保持身份可识别的人脸匿名化框架. 该框架由身份感知区域发现模块和身份感知人脸混淆模块组成，以便自适应地定位人脸与身份无关的属性，并利用原始人脸和定位的人脸属性生成匿名人脸图像. 相比已有方法[8,9,17-20]，该方法增加了匿名人脸属性的多样性. 具体来说，该方法首先计算出原始人脸图像的身份感知类激活热图，然后通过对各部分的激活分数进行排序，得到与身份无关的属性特征，再将原始人脸和属性特征送入条件人脸生成器得到隐私保护后的人脸图像. 其中鉴别器损失、属性分类器损失和人脸识别损失等损失函数共同约束人脸生成器，使模型可以根据实际需求修改不同数量的人脸属性，最终实现识别效用和外观匿名之间的平衡.

上述图像语义保持的方法能够生成身份匿名但视觉信息相似的匿名人脸，便于人们在社交平台共享个人照片等场景应用. 本节通过对基于对抗攻击的身份匿名化方法和身份保持的属性匿名化方法的介绍，为社交媒体等公开人脸图像的隐私泄露问题提供了应对方案. 一方面软生物特征信息和个人的其他公开数据本身可导致身份盗窃，另一方面生物特征数据和非生物特征数据可能存在链接攻击，从而泄露更多信息. 因此，图像语义保持的方法给隐私保护研究提供了不错的改进思路. 该方法结合对

① SAN 代码地址，https://github.com/iPRoBe-lab/semi-adversarial-networks.

抗攻击和属性匿名等技术，可达到修改一个或多个软生物特征信息的效果，进而匿名身份或属性信息以防止隐私泄露.

2.3 视觉可恢复的匿名化方法

现有的匿名化方法大多都存在一些局限性，例如隐私保护主要集中在数据保护过程中，而且一般是不可逆的. 因此本节将介绍视觉可恢复的匿名化方法，此类方法大多通过密码控制人脸图像的身份变化，通常满足两个要求：一是匿名化，二是去匿名化.

现实生活中，一方面可视数据的拥有者希望看到原始数据，如在司法领域，警察希望在视频中看清罪犯的真实面孔. 另一方面人们需要一个可以匿名敏感区域的系统，阻止黑客获得个人隐私信息. Ren 等人[51]提出的人脸匿名器在保留活动相关信息的同时，能够修改人脸的身份. 然而该技术没有考虑到视频/图像的授权使用者（如朋友、家人、执法部门等）可能希望看到原始身份的需求. 于是，在数据的可访问性和隐私性之间出现了权衡问题. Gu 等人[10]提出一种新的人脸身份转换模块，能够自动地对数据库中的人脸进行基于密码的匿名化和去匿名化操作. 通过设计人脸分类对抗损失、重建损失、背景一致损失和照片真实损失等损失函数，该方法可以在匿名化后移除人脸身份信息，并且当给定正确密码时能够恢复原始人脸，而在给定一个错误密码时返回一张错误身份且自然的人脸. 在 CASIA 和 LFW 数据集上，团队将该方法与传统匿名化方法和 DIIM 方法[11]进行对比实验，依据特征的距离估量和低阶感知度量等评价指标，证明该方法在不牺牲人脸隐私的情况下，能够进行基于密码的匿名化和去匿名化操作.

受 Gu 等人[10]利用多个神经网络训练身份转换器的启发，台湾大学团队[22]提出了一种基于条件编码器和解码器的框架实现视觉可恢复的人脸隐私保护，能够根据密码和多因素属性方案实现匿名化人脸的多样性和可控性. 继承 Gu[10]的密码方案并组合每张图像的潜在向量设计密码，该方法能够在正确密码下生成接近原始图像的重建图像，否则它便生成真实且多样化的匿名图像. 匿名化与去匿名化架构主要由两个并行编码器（风格编码器和内容编码器）和一个解码器组成，引入人脸属性和风格特征，从而提高匿名人脸的多样性和差异性. 与 Gu[10]方法相比，该方法可以根据身份属性来控制匿名图像的多样性，并在一定条件下成功地实现人脸图像的高保真匿名化，同时在不改变人脸数据分布的情况下实现去匿名化.

上海交通大学团队[11]提出一种基于深度生成模型的可恢复隐私保护方法. 受到 IdentityDP[5]模型特征解构的启发，该方法通过身份特征和属性特征的解耦操作，能够保留匿名图像的表情、姿势、光照等属性细节特征，并保持去匿名化图像与原始图像身份的一致性. 其网络架构主要由编码器、生成器和身份修改模块组成，编码器提取人脸的身份特征和属性特征，身份修改模块计算受保护的身份特征，最后生成器生成基于身份特征和属性特征的去识别结果. Zhu 等人[23]使用 ArcFace[52]将身份特征转换到超球面空间，并使用基于角度的余弦相似度进行身份修改操作①. 上述研究促使该团队通过改变身份嵌入的阶段来实现去身份处理，实现更有效的身份信息改变. 与先前提到的可恢复隐私保护方法[10,12]不同的是，该方法在设置密码方案的基础上添加了能够控制隐私级别的参数，增加了匿名人脸图像的多样性. 特别地，Li 等人[53]提出一种利用集成目标属性生成高保真图像的人脸交换算法. 为确保高保真的人脸交换结果，该网络通过多层属性编码提取不同空间分辨率的目标属性，并使用 AAD（Adaptive Attentional Denormalization）残差块自适应地学习集成属性或身份嵌入的位置信息.

大多数可恢复方法[10,11,22]的设计都需要在匿名过程中添加额外信息，导致应用的灵活性和隐私保护的安全性会受到影响. 于是 You 等人[54]提出了一种基于马赛克的可恢复人脸隐私保护方案，能够平衡匿名图像隐私性和可用性. 该方案首先使用预训练的人脸识别网络 YOLO[55]定位原始人脸的面部区域，然后用马赛克覆盖局部人脸区域得到马赛克人脸图像，再通过编码器将原始人脸特征隐藏到马赛克人脸中得到受保护的人脸图像. 并且，团队使用受保护的图像训练了一个用于面部表情识别的分类器，以供低特权用户执行计算机视觉任务. 这样一来，人脸图像上传到云端之前已经转化为受保护的匿名图像，于是在云服务上的受保护图像对于恶意攻击者来说能够窃取到的身份信息便很少. 对于低权限用户，他们可以使用提供的分类器对受保护的图像执行人脸识别等任务. 对于授权用户仍然可以重建出原始未匿名人脸图像以供使用. 然而该方案只验证应用于人脸表情识别任务，还未适用于其他计算机视觉任务.

已有方法[10,11,22,55]都是对于整张人脸进行恢复

① 上述方法代码地址，https://github.com/zyainfal/One-Shot-Face-Swapping-on-Megapixels.

操作，而 Ying 等人[56]提出了一种针对人脸局部区域的图像自恢复技术. 该技术能够直接定位到图像的被篡改位置，从而恢复原始图像信息. 该网络架构基于 U-Net 模型[57]，其中编码器生成带有恶意攻击的被攻击图像，验证网络预测被攻击图像的掩码并生成校正图像，解码器在给定校正图像的情况下重建图像. 在篡改分类损失、图像重建损失和判别损失的约束下，模型生成的被攻击图像与原始图像视觉外观相同，并且恢复后的图像具有很高的视觉质量. 然而对于纹理丰富的图像，重建图像很难完全恢复一些细节信息.

上述视觉可恢复的匿名化方法过程可分为三个阶段：（1）提取人脸图像的身份和属性特征，并确保属性特征在去身份过程中保持不变；（2）基于密码或其他控制参数，使用身份修改模块改变匿名图像的身份；（3）实现图像重建. 这类匿名化方法实现了身份信息的可逆化，使得拥有正确密码的人员可以查看用户数据，而非法偷取数据者无法得到正确的人脸内容信息，只能得到匿名人脸信息，从而防止了人脸隐私数据的泄露.

2.4　深度学习过程中的人脸隐私保护

目前很多方法在图像的匿名化和去匿名化方面表现出了不错的效果，然而现有方法很少对深度学习过程中的隐私保护进行探讨. 本节将介绍深度学习过程中的人脸隐私保护，其内容主要研究了深度学习数据的处理、模型训练过程的细节因素以及结合数据加密等其他领域对视觉身份隐私保护的影响.

北京大学王奕森团队[58]考虑了深度学习训练数据中的人脸隐私保护问题，并对人脸数据进行适当修改，使深度学习模型无法利用该样本进行训练. 该团队提出使用错误最小化噪声来防止数据被深度学习模型利用，并通过人脸识别的案例研究验证了该方法在创建不可学习样本方面具有有效性. 此外，错误最小化噪声可以很容易地从现有的公共数据集转移，使神经网络无法从私有数据集中获取信息，但在大规模应用方面仍然存在许多实际障碍. 这项工作开辟了防止数据被自由获取利用的新方向，为隐私保护的研究奠定了基础.

由于人脸图像可以从人脸模板中重建[59-61]，存储在系统中的生物测定模板必须受到保护. 于是，Mai 等人[62]提出一种端到端的方法来生成受保护的生物特征模板. 该框架由随机 CNN 和草图构造组件构成，并制定了两个损失函数，即随机三联体损失和正交三联体损失，共同优化受保护生物特征模板的验证性能和安全性能.

最近，通过联邦学习和数据加密实现隐私保护的研究取得了不错的成绩[63,64]，尤其在医学方面发挥了巨大的价值. MIME[65]提出一种局部应用全局动量的联邦学习框架，比仅应用客户端动量[66]方法更利于训练效率的提升. 因此，Bai 等人[67]提出利用联邦学习实现人脸数据的隐私保护. 该方法使用多方数据训练人脸识别模型，以避免隐私风险. 为了提高联邦人脸识别的性能，作者使用 PFM（Partially Federated Momentum）算法局部应用全局动量来有效地逼近集中动量，进而使用 FV（Federated Validation）算法在一些私有验证数据集上测试聚合模型来反复搜索更好的聚合权重，提高模型的泛化能力.

Kaissis 等人[12]设计了一个免费的医学图像联邦学习开源框架——PriMIA（Privacy-preserving Medical Image Analysis），该框架将联邦模型训练与模型更新的加密聚合以及加密的远程推理结合起来，用于对医学图像进行端到端的隐私保护. 该框架与多种医疗成像数据格式兼容，易于用户配置，并引入了联邦学习训练的功能改进（加权梯度下降、联邦平均、多样化数据增强、局部提前停止、联邦范围超参数优化、差分隐私数据集统计交换等），提高了灵活性、可用性、安全性等性能.

Froelicher 等人[68]提出了一种基于多方同态加密的联邦分析系统，能够在不泄露任何中间数据的情况下产生高精度的结果，实现对分布式数据集的隐私保护分析. 查询者首先将查询信息以明文形式发送给所有数据提供方，然后本地计算明文数据，并用集体公钥加密结果. 最后数据提供方将最终结果从集体公钥切换到查询者的公钥，一段时间后，查询者可解密最终结果.

Drozdowski 等人[69]提出了一种基于信息融合、数据隐私保护与同态加密相结合的面部生物特征识别检索方法. 该方法基于人脸模板的相似性进行智能配对和特征级融合，所创建的搜索结构便于多级生物识别检索，从而在级联的每个步骤中依次缩短检索的候选列表. 该系统在生物识别性能、计算效率和隐私保护方面取得了很好的平衡.

Klomp 等人[25]首次分析了图像匿名化方法对人脸检测模型的影响并进行了一系列实验分析. 该团队将传统的人脸匿名方法与基于生成对抗网络的方法比较，并在预先训练人脸检测模型上评估匿名图像的效果. 在实验过程中，该团队重点研究了以下三方面内容. 一是匿名化方法在保持人脸检测模型性能方面的适用性，二是检测模型过度训练对匿名化图像伪影的影响，三是用于训练模型的数据集尺

寸以及使用不同网络对匿名化训练时间的影响. 在此基础上，团队继续讨论了常见的生成对抗网络评估指标与预训练人脸检测模型性能之间的相关性. 实验结果表明：尽管所有经过测试的匿名化方法都会降低预训练人脸检测模型的性能，但使用生成对抗网络进行人脸匿名化导致的性能下降程度远小于传统方法. 因此，生成对抗网络在匿名方面的性能通常优于传统方法，且普遍应用于隐私保护领域. 然而生成对抗网络是否实现了绝对的隐私保护？模型在训练过程中是否会导致信息泄露呢？针对以上问题,圣母大学 Tinsley 等人[70]讨论了生成对抗网络模型的隐私泄露问题，并证明了训练数据的人脸信息会泄露到生成的伪造人脸当中. 该团队使用 5 个不同的人脸匹配模型和 StyleGAN2[71]模型进行实验，发现在一些方法上匿名人脸与原始人脸分布接近，即生成对抗网络会产生身份泄露问题.

随着新冠病毒的爆发，越来越多的人带上了口罩，那么口罩能保护人们隐私不被泄露吗？于是，Seneviratne 等人[26]探讨了戴口罩是否会保护个人隐私的问题①. 该团队训练了一个基于 ResNet-50 体系结构的卷积神经网络，通过对戴口罩的人脸图像进行性别、种族和年龄预测，发现戴口罩对隐私侵犯没有显著差异，戴口罩人脸的生物特征仍然能被人脸识别系统准确预测. 该方法可以作为评估隐私入侵性的基准工具，为后续研究工作提供了一定的参考价值.

上述深度学习过程中人脸隐私保护方法的研究为人脸匿名化研究领域提供了新的启示. 一方面人们可以对人脸数据进行处理，使得深度学习模型无法从中训练[58];另一方面人们可以对模型进行处理，使其处理人脸数据过程中无法泄露隐私信息等. 后续工作应多加考虑人脸匿名化过程中的隐私问题，多方面注意整个深度学习过程中的隐私泄露因素.

3 人脸视频匿名化方法

除人脸图像匿名化研究外，近年来人脸视频匿名化方法逐渐引起关注. 本节将从聚焦面部区域隐私和面向生物特征隐私两方面来介绍人脸视频匿名化方法. 对于视频序列来说，每一帧中所有的生物数据都需要进行隐私保护处理. 如果存在单个人脸数据未被处理的现象，那么该视频仍然具有隐私泄露风险. 因此，人脸视频匿名化难度较大且在隐私保护中扮演着重要的角色.

① 上述方法代码地址，https://github.com/sachith500/MaskedFaceRepresentation.

3.1 聚焦面部区域隐私的视频匿名化方法

视频隐私保护是一项具有挑战性的任务，需要对每一帧图像进行修改，保证不造成闪烁或其他视觉伪影和失真的前提下改变身份信息. 针对面部区域隐私的视频匿名化，大多研究使用了对抗性学习的思想，一方面是提高视频中的数据利用率，另一个方面是通过修改敏感生物特征（如人脸）等信息来确保隐私保护. 因此，聚焦面部区域隐私的视频匿名化方法旨在生成与原始人脸不同的匿名人脸视频，同时使匿名人脸仍然保留其它身份无关的生物特征.

Ren 等人[51]在动作识别系统的背景下研究了视频人脸隐私增强问题，提出了一个基于生成对抗网络的人脸视频匿名方法. 该网络结构由视频匿名器、动作检测器和人脸分类器组成. 视频匿名器修改原始视频以移除隐私敏感信息，同时最大化增强动作检测器的性能；动作检测器检测每个视频帧中人的动作试图从匿名视频中提取隐私敏感信息；人脸分类器对不同身份的人脸图像进行分类. 该框架通过动作识别、身份分类和人脸修改相关的学习目标的组合来约束匿名人脸的生成效果并提高动作检测器的性能. 然而该方法生成的人脸质量不佳，在头发和下巴边界有明显的伪影. 当有遮挡物时，生成的人脸大多存在模糊不清的现象.

传统的人脸去识别方法基本覆盖了整张人脸，使得人们无法分析匿名图像的面部行为. 然而在某些医学诊断中，人脸关键点和身体关键点都是非常必要的. 为了保护医疗数据中的隐私信息，清华大学神经调控技术国家工程实验室团队[72]提出利用人脸替换算法进行人脸隐私保护. 该团队将深度伪造技术应用于帕金森病检查视频中，对被试目标进行去识别，如图 5 所示. 对于每一对要交换人脸的受试者，该算法都要专门训练一个模型. 这个模型由一个共享编码器和两个解码器组成，编码器将输入的人脸图像编码为特征向量，然后解码器根据人脸表征重新创建人脸. 该算法保证了匿名人脸的人脸关键点和身体关键点几乎不变，克服了医疗数据共享的隐私保护问题.

Facebook 人工智能实验室 Gafni 等人[73]首次提出一种面向视频的实时人脸匿名化方法. 人脸视频匿名化的目标不仅要去除可用于识别的身份信息，还要保证视频帧间人脸的姿态、光照表情等一致性. 该团队借鉴 Sun[31]对头部的处理，使用了较浅的 U-Net 网络作为头部生成器，并通过将自动编码器与人脸分类器表征层连接起来，嵌入了身份和表情

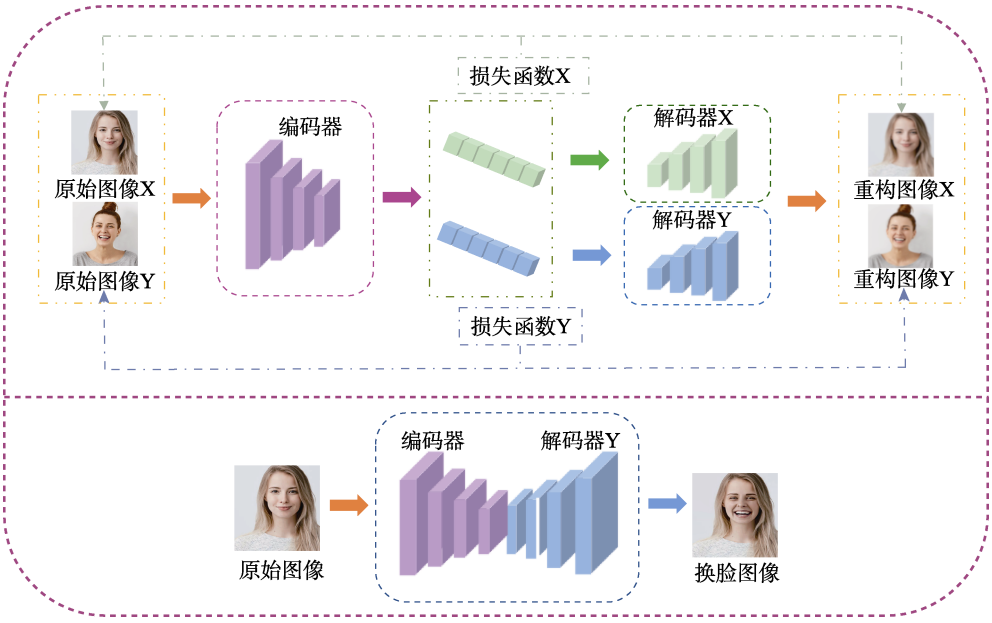


图 5　基于人脸替换的视觉内容修改与人脸匿名化示意图（方法流程主要参考文献[72]）

信息. 但由于常见的重建损失不能满足高度语义级别的任务，于是作者提出使用吸引/排斥感知损失来达到消除身份这一目的，即通过在几个中低级抽象层上采用未失真的原图特征和生成图特征之间的感知损失，同时添加目标图特征和生成图特征之间的高抽象层感知损失约束来实现.

Maximo 等人[13]提出了一种可以混合不同身份信息的人脸视频匿名化模型——CIAGAN 模型. 先前，Gafniet 等人[73]在姿势、光线和表情等保留的情况下，最大限度地提取整张人脸信息输入到网络中. 而 CIAGAN[13]则利用人脸关键点和独热编码去除人脸识别特征并保留必要的其他特征，以使人脸和身体检测器正常工作. 该模型提供了一个通用框架，为目标身份提供标签，还可混合不同风格的身份信息，从而直接控制去身份化过程. 该方法将掩蔽人脸和目标身份信息输入到编码器提取低维特征，并将其馈送到生成器的瓶颈部分. 然后，解码器将原图像和身份信息解码成匿名图像. 其中作者引入了身份鉴别网络，向生成器提供关于目标身份的引导信号. 该方法能够去除人脸和身体的识别特征，生成高质量视频的同时，还可用于其它计算机视觉任务，如检测或跟踪等. 在实验部分，该团队将模型用于 MOTS[74]数据集上，发现匿名人脸与原始人脸姿态基本一致，仅是衣服颜色和身体的其他部分发生了变化.

以上方法[13,51,73]大多基于生成对抗网络模型，在人脸去识别研究中占有相当大的比例. 然而有些模型受到各种条件信息的影响，需要大量的人力物力，更不用说解决表情变化、姿态多样等问题. 考虑到以上限制，中国科学院大学团队[75]提出了一种多样化的人脸匿名方法，可以灵活操纵视频中匿名人脸的身份，并保证其他属性基本不变. 借助特征解构的思想，该团队构造了一个身份解耦网络，引入条件多尺度重构损失 CMR（Conditional Multi-scale Reconstruction）和身份损失，将身份与其他属性解耦，以实现更好的身份匿名化. 其中，通过改变角度空间中的畸变角得到的匿名身份向量，可以直接控制匿名人脸的身份. 团队成员在 FaceForensics++[76]视频数据集上进行模型验证，结果表明匿名视频和原始视频中的人脸表情、姿势和其他属性基本一致，并且匿名视频具有良好的时间一致性.

对于复杂多变环境下的视频数据，大多匿名方法表现效果较差，容易出现质量不佳和时间不一致等问题. 因此 Balaji 等人[14]发布了一个在年龄、性别、种族和头部姿势等方面具有多样性的大规模人脸视频数据集. 基于此数据集，该团队提出一种名为 JaGAN 的人脸去识别方法，对没有标记面部特征点的人脸视频进行匿名化. 该方法首先识别并掩蔽单帧的人脸图像，然后用基于时间相关性的生成对抗网络修复缺失的内容，最终能够在视频序列中生成一致的人脸，并减少单帧之间的身份偏移. 此模型的实用性有助于促进人脸视频匿名化领域的进一步研究.

人脸识别技术的进步和滥用个人隐私数据的事件推动了图像和视频的匿名化研究发展，上述聚焦面部区域隐私的匿名化方法在视频和图像数据集上

都显示了良好的结果. 然而这些方法或多或少存在一些挑战，要么生成的人脸缺乏多样性，要么原始人脸的修改程度不可控，或是人脸属性没有得到很好的保存. 因此，对于人脸区域的隐私保护还需要进一步地研究和创新.

3.2 面向生物特征隐私的视频匿名化方法

除了修改人脸区域，面向生物特征隐私的修改也是一类有效的视频匿名化方法. 在日常生活中，监控设备可以同时记录一个人的外表和生理状态等信息，如心脏活动. 这种生理信息可能会被滥用，比如不法分子可通过秘密收集分析他人的生理特征，推断其生理状态，以在关键任务中获得优势. 因此，面向生物特征的隐私保护也是一个值得研究的课题.

视频隐私泄露日益严重，尤其针对基于云空间的视频监控系统，这导致了人们对基于云空间的安全视频应用程序产生了新需求. 目前很少有方法对复杂和动态场景具有稳健的性能，于是针对实际监控任务的场景，Tian 等人[77]提出了一种高效、稳健的隐私保护运动检测和多目标跟踪方案. 该方案能够直接对加密的视频流进行实时分析和处理，以便异常报警、异常目标检测和视觉跟踪等应用在智能监控系统中正常运行，最终达到了令人满意的隐私级别，有利于快速视频处理.

远程光电容积脉搏波描记法能够收集用户的生理状态，如心跳和呼吸等，这可能导致不必要的隐私泄露. 为了避免该技术的隐私滥用，Chen 等人[15]开发了一种新颖而高效的算法，它可以在不影响视觉外观的情况下编辑视频中的生理信号，从而保护用户的生理信号不被泄露，即通过去除视频中生理信号的痕迹，或者将视频转换为用户选择的目标生理信号来实现. 该算法能够对人脸视频中的生理信号进行编辑，在防御基于远程光电容积脉搏波描记法的活体检测和深度伪造检测中具有有效性.

上述面向生物特征隐私的视频匿名化方法通过去除生理信号来实现对视频的匿名处理，然而现阶段面向生物特征隐私的视频匿名化方法研究较少，目前一些解决方案虽然达到了令人满意的效果，但大多方法还是无法完全兼顾每一个视频帧的信息. 因此，有效保护生理信号依旧是匿名化任务中一个重要的难题.

4 数据集和模型性能分析

目前，人脸隐私保护方法通常在现有的公开数据集或自建数据集上进行分析. 本节将介绍现有匿名方法中常用的人脸数据集及相关图像客观匿名化评估指标，讨论代表性匿名方法的实验设置情况及结果，总结并评价模型性能优劣.

4.1 主流数据集简介

在隐私保护的任务中，模型或算法一般在主流数据集上进行训练评估. 研究人员为了测试模型在特殊场景的效果，也会自己创建特殊数据集用在人脸隐私保护方法中. 数据集通常是由一系列人脸图像组成，这些图像包括不同身份的人脸. 特别地，图像场景和人脸姿态多样的数据集对模型构建有着很大的挑战. 本小节将介绍常用的数据集 LFW[78]、MUCT[79]、CelebA[80]、RaFD[81]、WIDER-FACE[82]及自建的数据集 FDF[34]、UNI[70]等.

LFW（Labled Faces in the Wild）人脸数据集由美国马萨诸塞州立大学阿默斯特分校计算机视觉实验室整理完成，是目前人脸识别的常用测试集之一. 该数据集中的人脸图像均来源于生活中的自然场景，人脸所受的光照情况和人脸姿态多种多样，有的人脸还存在部分遮挡的情况，因此识别难度较大. LFW 数据集共有 13 233 张人脸图像，每张图像均被标识出对应的人名，共有 5 749 人，且绝大部分人仅有一张图片. 其中多数图片为彩色图像，但也存在少许黑白人脸图片.

MUCT 数据集由开普敦大学于 2010 年发布，是一个主要用于身份鉴定的人脸数据集，其中每张图像包含 76 个关键点，在种族、年龄等方面表现出极大的多样性.

CelebA（CelebFaces Attribute）是香港中文大学开源的大规模人脸检测基准数据集. 该数据集包含了 10 177 个名人身份的 202 599 张人脸图片，其中每张图片都做好了特征标记，包含人脸边界框、人脸特征点坐标以及 40 个属性标记. 目前，CelebA 数据集被广泛用于人脸图像相关的计算机视觉训练任务中，在人脸属性标识训练、人脸检测训练以及人脸关键点标记等方面发挥着重要作用.

RaFD 人脸数据集于 2010 年由荷兰的拉德伯德大学发布，总共 8 040 张图，包含 67 个模特，分别为 20 名白人男性成年人、19 名白人女性成年人、4 名白人男孩、6 名白人女孩、18 名摩洛哥男性成年人. 并且该数据集包含 8 种表情，即愤怒、厌恶、恐惧、快乐、悲伤、惊奇、蔑视和中立，每一个表情包含 3 个不同的注视方向，且使用 5 个相机从不同的角度同时拍摄.

香港中文大学提出的 WIDER-FACE 数据集是

目前主流的人脸检测数据集. 该数据集总共有32 203 张图片和 393 703 张人脸，其中人脸尺寸、姿势、遮挡、表情、妆容和光照级别多种多样. 在标注边界框的同时，该数据集还提供了遮挡和姿态等信息，自发布后广泛应用于卷积神经网络的性能评估中.

FDF（Flickr Diversified Faces）是 Hukkelås 等人提出的一个多姿态人脸数据集，包含人脸关键点和人脸边界框注释. 该数据集从交通、体育赛事和户外活动相关的场景中提取，在年龄、种族、姿势、图像背景和面部遮挡方面具有巨大的多样性.

UNI 数据集是 Patrick 等人收集的 12 004 张高质量的人脸图像集，用于评估各种人脸匹配器的人脸识别准确率. 该数据集有 333 名受试者，每个受试者的正面照片在 4 到 78 张之间，在良好的光照下以 1200×1600 的分辨率拍摄.

4.2　客观匿名化评估指标简介

客观匿名化评估指标主要分为两类，一类是身份匿名效果指标，即用人脸识别等方法判断匿名前后身份是否一致；另一类为图像客观质量指标，即用 SSIM 等方法评估匿名人脸图像的图像质量.

（1）身份匿名效果指标简介

2018 年，伦敦帝国理工学院团队提出了 ArcFace[52]人脸识别模型. 在 SphereFace[83]基础上改进了对特征向量归一化和加性角度间隔，加强相同类别内部紧度的同时，提高了不同类别间差异. ArcFace 模型性能高、易于编程实现，且复杂性低、训练效率高. 该人脸识别模型可用于人脸身份匿名效果的评估任务，即通过计算图像特征向量的距离，判断匿名人脸与原始人脸是否属于同一身份.

IntraFace[84]是卡内基梅隆大学研究人员设计的一种新型面部识别软件. 这一软件反应迅速、操作简单，可以在智能手机上使用. 目前 IntraFace 是最常用的人脸标定方法之一，可准确预测 49 个关键点. 在匿名属性信息时，IntraFace 通常用于性别分类，通过给匿名前后人脸图像评分，评估模型性能优劣.

（2）图像客观质量指标简介

结构相似性指标 SSIM（Structural SIMilarity）[85]由德州大学奥斯丁分校的图像和视频工程实验室提出，在匿名化领域通常被用来衡量原始图像与匿名图像之间的相似度. SSIM 从亮度、对比度和结构三个方面进行比较，结构相似性的范围为-1 到 1. 当两张图像一模一样时，SSIM 的值等于 1.

学习感知图像补丁相似度 LPIPS（Learned Perceptual Image Patch Similarity）[86]度量使用深度特征测量图像相似度，是一种用于量化两幅图像间结构相似性的指标. 不同网络结构判断图片的感知相似度时，SSIM 指标可能给出与人类感知不同的结论，相比之下，基于学习的感知相似度度量要更符合人类的感知. LPIPS 的值越低表示两张图像越相似，更高的 LPIPS 意味着两张图像差异大.

FID 距离得分（Frechet Inception）[87]是测量两个数据集图像之间相似性的度量，即计算真实样本和生成样本在特征空间中的距离. FID 从原始图像的计算机视觉特征方面来衡量两组图像的相似度，这种视觉特征是使用 Inceptionv3 图像分类模型计算得到的. 较低的 FID 分数表示两组图像分布更接近，也就意味着生成图像的质量较高、多样性较好. 在最好情况下 FID 的得分为 0.0，表示两组图像相同. FID 分数也被广泛用于评估由生成对抗网络生成的图像质量，较低的分数与较高质量的图像有很高的相关性. 然而，这些度量指标是用来评价两张图片的相似度，而不是两个人脸的相似度. 同一个人在不同光照（姿态或表情）情况下的两张照片，FID 和 SSIM 的值可能相差很多.

ROC（Receiver Operating Characteristic）曲线又称接受者操作特征曲线，是反映敏感度和特异度连续变量的综合指标. 该曲线最早应用于雷达信号检测领域，用于区分信号与噪声，后来人们将其用于评价模型的预测能力. ROC 曲线距离左上角越近，证明分类器效果越好. AUC（Area Under Curve）为 ROC 曲线与坐标轴围成的面积，是一个二分类模型的评价指标. AUC 的本质是从样本集中随机选择一个正样本和负样本，模型预估正样本得分大于负样本得分的概率. 因此，AUC 的值越接近 1，其模型的性能越好. 一般情况下，ROC 曲线和 AUC 值通常用来对比匿名前后人脸图像的身份信息是否不同，并评估视觉内容可恢复阶段的人脸恢复效果[22,58,70].

AP（Average Precision）是指精度-召回率曲线下的面积，面积越大，模型性能越好. 性能较优的模型应是在召回率（R）增长的同时，精度（P）值也会保持在一个较高的水平，而性能较低的模型往往需要牺牲很多 P 值才能换来 R 值的提高. AR（Average Recall）即平均召回率，对于不同的交并比（IOU）取最大的召回率再求平均值.

$L1$（Manhattan Distance）用来对比两张图片的相似程度，即对图片逐个像素求差值，然后将所有差值加起来得到的数值为 $L1$ 距离. 如果两张图片一模一样，那么 $L1$ 距离为 0，反之 $L1$ 值将会非常大. $L2$（Euclidean Distance）同样是计算像素间的差值，

只是先求差值的平方,再将平方和进行开方操作. $L1$ 距离更依赖于坐标轴的选定，坐标轴选择不同 $L1$ 距离也会跟着变化. 相对来说，$L2$ 距离与坐标系的关联度较低，不跟随坐标轴变化. 当图像中有特殊类型的特征时可以选择 $L1$ 距离,当对图像中所有元素未知时，$L2$ 距离会更自然一些.

4.3 实验结果

本文从现有方法中选择一些代表性的方法进行对比：基于生成对抗网络的方法（记作 DP[34]）、基于特征解构的方法（记作 IdentityDP[5]）、基于属性修改的方法（记作 SBP[18]和 K-F[19]）、视觉内容可恢复的隐私保护方法（记作 FIT[10]、DIIM[11]和 MfM[22]）、针对医学研究隐私保护的人脸替换方法（记作 DF[72]）以及人脸视频匿名化方法（记作 CIAGAN[13]、AIG[75]和 JaGAN[14]）.

接下来本文对各类代表性方法的实验结果进行分析，其中各评估指标的结果为原文献中汇报的结果，细节部分及方法间的对比见表 2 和表 3. DP 方法①使用 FDF 数据集训练，以及 WIDER-FACE 数据集评估匿名化的影响，在匿名数据集上测量人脸检

表 2 代表性方法实验细节

算法类型	方法	发表时间	实验数据集	主要评估指标
基于虚假人脸生成	DP [34]	2019	WIDER-FACE[82]	AP_{Easy} ↑ (0.959)
			CelebA-HQ[88]	SSIM ↑ (0.7808)
特征解构	IdentityDP[5]	2021	CelebA-HQ[88]	PSNR ↑ (0.908) FDR ↑ (0.997) SSIM ↑ (0.8606)
属性修改	SBP[18]	2017	MUCT[79]	性别预测错误率 ↑ (0.901)
	K-F[19]	2018	CelebA[80]	属性分类准确率 ↓ (0.28)
视觉内容可恢复	FIT[10]	2019	CelebA-HQ[88]	FID ↓ (110) LPIPS ↓ (0.17)
			CASIA-WebFace[89]	SSIM ↑ (0.87) user study ↓ (12.2%, 2.7%)
	DIIM[11]	2021	CelebA-HQ[88]	LPIPS ↓ (0.062)
			CASIA-WebFace[89]	SSIM ↑ (0.902)
	MfM[22]	2021	CelebA-HQ[88]	LPIPS ↓ (0.35)
			CASIA-WebFace[89]	FID ↓ (28) SSIM ↑ (0.95)
人脸替换	DF [72]	2020	FaceForensics++[76]	$AP^{0.95}_{Swapped}$ ↑ (0.990)
			COCO[90]	$AP^{0.95}_{Masked}$ ↑ (0.443) $AP^{0.95}_{Blurred}$ ↑ (0.408)
人脸视频匿名化	CIAGAN[13]	2020	CelebA[80]	FID ↓ (2.663) LPIPS ↓ (0.221) SSIM ↑ (0.718)
	AIG[75]	2021	CelebA[80]	FID ↓ (2.193) LPIPS ↓ (0.177) SSIM ↑ (0.865)
	JaGAN[14]	2021	CC-BY[14]	FVD ↓ (59) IDI ↑ (0.48)
			FDF[34]	FID ↓ (1.97)

表 3 代表性方法对比

方法	创新思路	不足之处
DP, 2019[34]	使用 U-Net 模型保留背景信息	不适用于非传统的姿态场景
IdentityDP, 2021[5]	利用人脸表征解纠缠并在身份表征上添加扰动模糊身份信息	防止人脸识别表现方面较差
SBP, 2017[18]	基于 Delaunay 三角测量和卷积自动编码器的方法，在保留面部身份的同时翻转性别信息	身份识别性能较差
K-F, 2018[19]	使用 Carlini-Wagner L2 攻击对图像添加噪声	对任意分类器不适用
FIT, 2019[10]	使用多个神经网络和密码指令控制匿名人脸	匿名人脸的眼睛和头发等细节部分虚化现象较严重
DIIM, 2021[11]	引入多因素属性向量控制匿名人脸	部分匿名人脸的头发细节伪影较严重
MfM, 2021[22]	引入隐私参数控制匿名程度	部分匿名人脸的性别、年龄等因素不匹配
DF, 2020[72]	使用人脸替换算法生成匿名人脸	部分匿名人脸会保留目标人脸的属性特征
CIAGAN, 2020[13]	通过有条件的生成对抗网络和身份鉴别模块生成随机身份的图像	多数匿名图像视觉效果较不自然
AIG, 2021[75]	将身份特征和其他特征解构实现身份匿名化	少数匿名图像眼神方向改变
JaGAN, 2021[14]	使用虚假人脸填充单帧缺失的图像块	修复显示相邻视频帧之间的时间相关外观的图像块较差

① DP 方法代码地址，www.github.com/hukkelas/DeepPrivacy.

测模型的 AP 均达到 90%以上. 一般情况下，DP 方法能够生成真实自然的匿名人脸，但模型对数据集的要求较高，在遮挡、不同姿态等复杂场景下，模型生成的图像质量可能不佳，如图 6(a)所示.

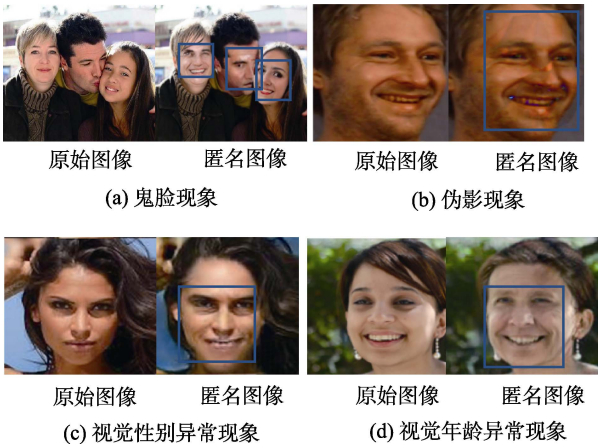


图 6　现有方法匿名化效果异常现象示例

与传统方法比较，IdentityDP 方法在视觉相似性和真实感方面取得了很大的优势. 研究人员使用 CASIA 和 VGGFace2 数据集进行模型测试，并得到各方法的匿名图像；进一步使用 FaceNet 识别模型识别匿名图像，计算匿名图像与原始图像的身份差异值. 实验结果表明身份差异值都达到了 1.1 以上，并且传统方法无法平衡隐私及效用权衡问题，在防止人脸识别方面表现较差. 值得一提的是，相较于 DP 方法，该方法在 CelebA-HQ 数据集上得到的匿名人脸的 SSIM 值有所提升.

SBP 方法使用 MUCT 数据集评估模型，在 IntraFace 性别分类模型上进行性别评分. 实验结果表明：其性别评分最初为负值（即女性）的图像，经过 31 个三角形更新步骤后变为正值（即男性），初始为正值（即男性）的图像也逐渐被成功地翻转为负值（即女性）. 对比原始图像、性别扰动后的图像在 IntraFace 和 GCOTS 上的得分，实验结果表明：SBP 并不适用于翻转整个数据集的人脸图像的性别属性，会出现图像失真、性别失败翻转等情况，如图 6(b)所示. 对于 K-F 方法来说，K-F 可以控制多个属性的匿名，一般情况下匿名人脸的视觉效果与原始人脸相同. 文献[19]对比该方法在抑制不同属性的匿名图像的分类准确率，发现超过 96%的人脸图像中的属性被成功抑制.

FIT 方法①计算 CASIA-WebFace 和 CelebA-HQ 数据集重构图像的 LPIPS、SSIM 及 FID 等指标，评估模型的图像重建能力. 实验结果显示 LPIPS 和 SSIM 值都达到了不错的效果，均优于传统方法. 与此同时，在给定不同密码的情况下评估模型创建不同面孔的能力方面，团队成员通过 AMT 感知研究对多模态进行定量评估（即用户调查问卷），要求 AMT 工作人员比较不同密码生成的匿名和错误恢复的人脸图像，认为是同一个人的比例分别为 12.2%和 2.7%（user study 指标越低越好），结果表明 FIT 方法可以在给定不同密码的情况下生成不同身份的人脸图像，且大部分匿名图像较真实. 然而某些图像在眼睛和头发部分出现断层虚化现象，且没有考虑年龄性别等因素，导致图像合成效果不佳. FIT 方法需要针对不同密码重新训练网络，而 DIIM 的加密过程相对独立于深度生成网络，可以更加灵活定义密码形式，并且操作复杂度和身份多样性得到了很大的改善. 该方法主要基于身份特征解构，即将图像表征分离为身份特征和属性特征，并利用属性特征模块和身份修改模块计算匿名身份特征，从而生成匿名图像. 实验设置和 FIT 方法相同，实验结果表明 SSIM 和 LPIPS 两个指标均优于 FIT 方法. MfM 是一种新的基于条件编码器和解码器框架的可恢复隐私保护方法，能够根据人脸属性因素和密码共同控制匿名化程度. 在匿名图像的质量和多样性方面，MfM 在 FID 和 SSIM 指标中的结果均优于 FIT，其中值得注意的便是 MfM 的 FID 值达到了 28，相比原始图像只增加了 12.9 的 FID 值，相比 FIT 降低了 82 的 FID 值[12]. 在多因素组合实现图像去匿名化的视觉效果上，大部分人脸看起来比较真实. 然而还是有些许人脸图像效果不佳，比如在密码和头发两种因素控制下，头发部分出现虚化现象，而且匿名人脸的性别、年龄可能不匹配，如图 6(c)(d)所示.

在将 Deepfakes 技术应用于医学视频的 DF 方法中，团队成员将一个病人 X 和一个开源数据集人物 Y 的视频来训练 Faceswap 模型，把 X 的视频输入到模型里，能够将 X 脸换成 Y 脸. 该团队对换脸前后的人脸图像和使用模糊、掩蔽方法匿名的人脸图像进行关键点检测，评估关键点位置的变化，实验表明换脸后的人脸关键点与原始人脸关键点相似性高，且人脸关键点的 AP 值均高于 99%，AUC 曲线也达到了非常不错的效果. 虽然匿名后的人脸图像没有出现明显的伪影，但还是可以观察到如果目标人脸是男性，那么被换脸后的图像可能会出现男性特征，如图 6(c)所示.

① FIT 方法代码地址，https://github.com/laoreja/face-identity-transformer.

CIAGAN 方法①利用条件生成对抗网络的优势，以人脸关键点作为监督信息，指导匿名化人脸的生成过程. 该方法能够确保匿名视频数据具有时间一致性，对于人物跟踪或动作识别等任务有较好的效果. 与传统方法比较，CIAGAN 在 CelebA 数据集上生成的匿名图像的检测率几乎达到 100%. 虽然 CIAGAN 可以生成多样化的匿名人脸，但在视觉效果上大多数匿名图像不太自然，人脸属性也有所改变. AIG 方法在 FaceForensics++视频数据集上测试匿名视频的视觉效果，实验结果展示匿名人脸的表情、姿势和其他属性基本不变，同时在视频数据中能够保持良好的时间一致性. 该团队进一步在 CelebA 数据集上计算 CIAGAN 和 AIG 方法生成人脸的 FID、LPIPS 及 SSIM 的值，相较 CIAGAN 来说，AIG 方法展现了不错的效果. JaGAN 方法能够在视频数据集上生成与原始人脸身份不同的匿名人脸，该团队使用自建的视频数据集评估视频质量(FVD 和 IdI)以及跨多帧的时间相关人脸图像视觉效果. 实验结果表明匿名人脸身份标识在帧之间保持不变，并且 FVD 得分为 59，IdI 得分为 0.48，整体效果较好. 其中 IdI 为实际帧的 L2 平方距离的中位数与生成帧的 L2 平方距离的中位数的比率.

5 总结与未来研究展望

互联网形成了一个开放、动态、持续演化的虚拟空间，极大地提升了人与人之间的通信速度，扩大了社会的通信范围，也为个人隐私安全的保护带来了新的挑战. 与此同时，身份伪造技术的发展推动了人脸隐私保护的进一步研究，社会对公共隐私安全保护问题越来越关注. 除了制定法规条文等措施，研究人员也在积极地提出更多可行的匿名方法来保护数据信息安全. 本文通过将现有方法对比和研究，发现虽然目前有些方法生成的人脸图像视觉效果较好，但很多方法对数据集图像的要求较高，模型的泛化性能需要进一步提高. 随着人们对个人隐私数据的重视，未来研究可以向以下几个方面进行探索.

（1）复杂场景下的人脸隐私保护

对于可控环境下的人脸数据集，现有匿名化方法通常能够生成逼真自然的匿名图像. 然而针对复杂不可控场景下的人脸图像，大多匿名方法表现效果不佳，尤其是人脸遮挡情况. 未来人脸匿名化研究应考虑各种环境下匿名图像的质量问题，提升匿名图像的视觉效果. 后续研究也可以通过丰富数据集、增加数据种类以及采取数据增强等各种手段来模拟现实生活中的复杂场景，从而提升模型对遮挡人脸的检测和生成效果，方便对其进行人脸匿名处理. 进一步地，开展人脸匿名化过程的对抗攻击和防御研究对提升人脸隐私保护过程的安全性也具有重要意义.

（2）匿名人脸视觉真实感的修正

在一些情况下，匿名化方法生成的匿名人脸视觉真实感不佳. 比如原始图像是一张女性图片，经过性别翻转后生成了带有女性长发特征的男性图片[22]，人眼观看的视觉效果不好. 因此在匿名过程中，除了考虑身份匿名，研究人员还应考虑匿名后人脸的真实感问题. 在实现目标任务的同时，匿名方法生成的图像应达到真实且清晰的效果. 后续研究可以考虑在生成过程中增加匿名前后与身份无关等生物属性特征的一致性约束来解决该问题. 在未来研究中，解决匿名人脸视觉真实感的修正问题值得进一步关注.

（3）敏感行为意图的隐私保护

在日常生活中，人脸敏感信息可能会暴露个人的行为意图. 现有的人脸匿名化方法一般会生成一张与原始人脸朝向、眼神注视方向[8,18,19]和微表情[35]相同的匿名人脸图像，然而这些人脸敏感信息对个人行为意图分析有一定的辅助作用，因此敏感行为意图的隐私保护也值得关注. 现有工作针对敏感行为意图的隐私保护研究较少，一方面可以利用属性修改的方法改变能够暴露行为意图的属性特征（比如注视方向和微表情等）. 另一方面，将人脸替换和属性迁移等技术融合改变某些属性信息，也具有很大的研究意义和探讨价值.

（4）人脸匿名化的可解释性

对于深度学习来说，模型的解释性很难通过精确的数学表达来形式化定义. 深度学习可解释性本身就是一个热门的研究方向，可解释性一般通过隐层分析法[91-93]、模拟模型[94-97]和注意力机制的引入等方法来进行，后续工作可以通过借鉴其他相关领域的研究提高人脸匿名化的可解释性. 更易于解释的模型有助于确保其完整性和隐私性，能够反映出人脸匿名前后之间的关系.

（5）人脸匿名化的评价指标

现有针对人脸匿名化的评价指标大多都是通过计算图像质量、图像相似度、人脸识别率或图像真实度来评价人脸匿名化模型的效果，其指标繁多且计算量较大. 未来研究可以在人脸匿名化的评价指

① CIAGAN 方法代码地址，https://github.com/dvl-tum/ciagan.

标上进行创新，设计统一的评价指标. 一个简单的解决思路是通过自适应的权重学习将已有指标关注的评价标准进行融合. 现有研究工作中针对人脸匿名化的评价指标研究相对较少，因此在解决隐私保护问题的同时，统一人脸匿名化的评价指标问题对于完善相关技术的理论研究具有重要意义.

致　谢　在此，我们向对本文的工作给予支持和宝贵建议的评审老师和同行表示衷心的感谢!

参　考　文　献

[1] Pataranutaporn P, Danry V, Leong J, Punpongsanon P, Novy D, Maes P, SraM. Ai-generated characters for supporting personalized learning and well-being. Nature Machine Intelligence, 2021, 3(12): 1013-1022

[2] Newton E M, Sweeney L, Malin B. Preserving privacy by de-identifying face images. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(2): 232-243

[3] Meden B, Emersic Z, Struc V. k-same-net: neural-network-based face deidentification//2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI). Funchal, Portugal, 2006: 161-161

[4] Li T, Choi M S. Deepblur: A simple and effective method for natural image obfuscation. arXiv preprint arXiv: 2104.02655, 1, 2021

[5] Wen Y, Song L, Liu B, Ding M, Xie R. Identitydp: Differential private identification protection for face images. Neurocomputing, 2022, 501: 197-211

[6] Evtimov I, Sturmfels P, Kohno T. Foggysight: A scheme for facial lookup privacy. arXiv preprint arXiv: 2012.08588, 2020

[7] Shan S, Wenger E, Zhang J, Li H, Zheng H, Zhao BY. Fawkes: Protecting privacy against unauthorized deep learning models// 29th USENIX Security Symposium (USENIX Security 20). Boston, USA, 2020: 1589-1604

[8] Mirjalili V, Raschka S, Namboodiri A, Ross A. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images//2018 International Conference on Biometrics (ICB). Gold Coast, Australia, 2018: 82-89

[9] Mirjalili V, Raschka S, Ross A. Privacynet: semi-adversarial networks for multi-attribute face privacy. IEEE Transactions on Image Processing, 2020, 29: 9400-9412

[10] Gu X, Luo W, Ryoo M S, Lee YJ. Password-conditioned anonymization and deanonymization with face identity transformers. Lecture Notes in Computer Science. Springer, Cham, 2020, 12368: 727-743

[11] Cao J, Liu B, Wen Y, Xie R，Song L. Personalized and invertible face de-identification by disentangled identity information manipulation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 3334-3342

[12] Kaissis G, Ziller A, Passerat-Palmbach J, Ryffel T, Usynin D, Trask A, Lima I, Mancuso J, Jungmann F, Steinborn MM, Saleh A. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence, 2021, 3(6): 473-484

[13] Maximov M, Elezi I, Leal-Taix´e L. Ciagan: Conditional identity anonymization generative adversarial networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5447-5456

[14] SchT, Blies P, Göri G, Mitsch R, Wasserer M. Temporally coherent video anonymization through gan inpainting. arXiv preprint arXiv: 2106.02328, 2021

[15] Chen M, Liao X, Wu M. Pulseedit: Editing physiological signals in facial videos for privacy protection. IEEE Transactions on Information Forensics and Security, 2022, 17: 457-471

[16] Gross R, Sweeney L, De La Torre F. Model-based face de-identification//2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06). New York, USA, 2006: 161-161

[17] Othman A, Ross A. Privacy of facial soft biometrics: Suppressing gender but retaining identity//European Conference on Computer Vision. Zurich, Switzerland, 2014, 8926: 682-696

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

[19] Chhabra S, Singh R, Vatsa M, Gupta G. Anonymizing k-facial attributes via adversarial perturbations//International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 656-662

[20] Mirjalili V, Raschka S, Ross A. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers//2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). Redondo Beach, USA, 2018: 1-10

[21] Li J, Han L, Chen R, Zhang H, Han B, Wang L, Cao X. Identity-preserving face anonymization via adaptively facial attributes obfuscation//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China, 2021: 3891-3899

[22] Pan Y L, Chen J C, Wu J L. A multi-factor combinations enhanced reversible privacy protection system for facial images// 2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China, 2021: 1-6

[23] Zhu Y, Li Q, Wang J, Xu CZ, Sun Z. One shot face swapping on megapixels//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 4834-4844

[24] Li J, Ma S, Zhang J, Tao D. Privacy-preserving portrait matting//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China, 2021: 3501-3509

[25] Klomp SR, Van Rijn M, Wijnhoven RG, Snoek CG, De With PH. Safe fakes: Evaluating face anonymizers for face detectors//2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). Jodhpur, India, 2021: 1-8

[26] Seneviratne S, Kasthuriarachchi N, Rasnayaka S, Hettiachchi D, Shariffdeen R. Does a face mask protect my privacy? : Deep learning to predict protected attributes from masked face images//Australasian Joint Conference on Artificial Intelligence.

Perth, Western Australia, 2022: 91-102

[27] Sun Q, Ma L, Oh SJ, Van Gool L, Schiele B, Fritz M. Natural and effective obfuscation by head inpainting//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018: 5050-5059

[28] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks//Proceedings of the International Conference on Learning Representations (ICLR). Puerto Rico, USA, 2016, abs/1511.06434: 0

[29] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of International Conference on Learning Representations. Banff, Canada, 2014, abs/1409.1556: 0

[30] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,43(12): 4401-4410

[31] Sun Q, Tewari A, Xu W, Fritz M, Theobalt C, Schiele B. A hybrid model for identity obfuscation by face replacement//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 553-569

[32] Tewari A, Zollhofer M, Kim H, Garrido P, Bernard F, Perez P, Theobalt C. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction//Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, Italy, 2017: 1274-1283

[33] Yang X, Dong Y, Pang T, Su H, Zhu J, Chen Y, Xue H. Towards face encryption by generating adversarial identity masks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 3897-3907

[34] Hukkelås H, Mester R, Lindseth F. Deepprivacy: A generative adversarial network for face anonymization//International Symposium on Visual Computing. Springer, Cham, 2019, 11844: 565-578

[35] Wu Y, Yang F, Xu Y, Ling H. Privacy-protective-gan for privacy preserving face de-identification. Journal of Computer Science and Technology, 2019, 34(1): 47-60

[36] Li T, Lin L. Anonymousnet: Natural face de-identification with measurable privacy//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA, 2019: 56-65

[37] Nousi P, Papadopoulos S, Tefas A, Pitas I. Deep autoencoders for attribute preserving face de-identification. Signal Processing: Image Communication, 2020, 81: 115699

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

[39] Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018: 8789-8797

[40] Dall'Asen N, Wang Y, Tang H, Zanella L, Ricci E. Graph-based generative face anonymisation with pose preservation//International Conference on Image Analysis and Processing. Lecce, Italy, 2022, 13232: 503-515

[41] Jeong Y, Choi J, Kim S, Ro Y, Oh TH, Kim D, Ha H, Yoon S. Ficgan: Facial identity controllable gan for de-identification. arXiv preprint arXiv: 2110.00740, 2021

[42] Khojaste MH, Farid NM, Nickabadi A. Gmfim: A generative mask-guided facial image manipulation model for privacy preservation. arXiv preprint arXiv: 2201.03353, 2022

[43] Hukkelås H, Smebye M, Mester R, Lindseth F. Realistic full-body anonymization with surface-guided gans//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Hangzhou, China, 2023: 1430-1440

[44] Chatzikyriakidis E, Papaioannidis C, Pitas I. Adversarial face de-identification//2019 IEEE International Conference on Image Processing (ICIP). Taipei, China, 2019: 684-688

[45] Mansourifar H, Shi W. Vulnerability of face recognition systems against composite face reconstruction attack. arXiv preprint arXiv: 2009.02286, 2020

[46] Fu Y, Guo G, Huang T S. Age synthesis and estimation via faces: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(11): 1955-1976

[47] Lu X, Jain A K. Ethnicity identification from face images//Proceedings of SPIE, 2004, 5404: 114-123

[48] Makinen E, Raisamo R. Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(3): 541-547

[49] Suo J, Lin L, Shan S, Chen X, Gao W. High-resolution face fusion for gender conversion. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2010, 41(2): 226-237

[50] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//2017 IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[51] Ren Z, Lee YJ, Ryoo MS. Learning to anonymize faces for privacy preserving action detection//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 620-636

[52] Deng J, Guo J, Xue N, Zafeiriou Ss. Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 44(10): 4690-4699

[53] Li L, Bao J, Yang H, Chen D, Wen F. Advancing high fidelity identity swapping for forgery detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5074-5083

[54] You Z, Li S, Qian Z, Zhang X. Reversible privacy-preserving recognition//2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China, 2021: 1-6

[55] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779-788

[56] Ying Q, Qian Z, Zhou H, Xu H, Zhang X, Li S. From image to

imuge: Immunized image generation//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China, 2021: 3565-3573

[57] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//International Conference on Medical Image Computing and Computer-assisted Intervention. Berlin, Germany, 2015: 234-241

[58] Huang H, Ma X, Erfani SM, Bailey J, Wang Y. Unlearnable examples: Making personal data unexploitable. arXiv preprint arXiv: 2101.04898, 2021

[59] Feng Y C, Lim M H, Yuen P C. Masquerade attack on transform-based binary-template protection based on perceptron learning. Pattern Recognition, 2014, 47(9): 3019-3033

[60] Mai G, Cao K, Yuen P C, et al. On the reconstruction of face images from deep face templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(5): 1188-1202

[61] Cao K, Jain A K. Learning fingerprint reconstruction: From minutiae to image. IEEE Transactions on Information Forensics and Security, 2014, 10(1): 104-117

[62] Mai G, Cao K, Lan X, Yuen P C. Secureface: Face template protection. IEEE Transactions on Information Forensics and Security, 2020, 16: 262-277

[63] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Seattle, USA, 2017: 1273-1282

[64] Zheng W, Yan L, Gou C, et al. Federated meta-learning for fraudulent credit card detection//Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Yokohama, Japan, 2021: 4654-4660

[65] Karimireddy S P, Jaggi M, Kale S, Mohri M, Reddi S J, Stich S U, Suresh AT. Mime: Mimicking centralized stochastic algorithms in federated learning. arXiv preprint arXiv: 2008.03606, 2020

[66] Wang J, Tantia V, Ballas N, Rabbat M. Slowmo: Improving communication-efficient distributed sgd with slow momentum. Clinical Orthopaedics and Related Research, 2019, abs/1910. 00643: 0

[67] Bai F, Wu J, Shen P. Federated face recognition. arXiv preprint arXiv: 2105.02501, 2021

[68] Froelicher D, Troncoso-Pastoriza J R, Raisaro J L, Cuendet M A, Sousa J S, Cho H, Berger B, Fellay J, Hubaux J P. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nature Communications, 2021, 12(1): 1-10

[69] Drozdowski P, Stockhardt F, Rathgeb C, Osorio-Roig D, Busch C. Feature fusion methods for indexing and retrieval of biometric data: Application to face recognition with privacy protection. IEEE Access, 2021, 9: 139361-139378

[70] Tinsley P, Czajka A, Flynn P. This face does not exist... but it might be yours! Identity leakage in generative models//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021: 1320-1328

[71] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8110-8119

[72] Zhu B, Fang H, Sui Y, Li L. Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, USA, 2020: 414-420

[73] Gafni O, Wolf L, Taigman Y. Live face de-identification in video//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 9378-9387

[74] Voigtlaender P, Krause M, Osep A, Luiten J,Sekar BB, Geiger A, Leibe B. Mots: Multi-object tracking and segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7942-7951

[75] Ma T, Li D, Wang W, Dong J. Face anonymization by manipulating decoupled identity representation. arXiv preprint arXiv: 2105.11137, 2021

[76] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: Learning to detect manipulated facial images//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 1-11

[77] Tian X, Zheng P, Huang J. Robust privacy-preserving motion detection and object tracking in encrypted streaming video. IEEE Transactions on Information Forensics and Security, 2021, 16: 5381-5396

[78] Huang GB, Mattar M, Berg T, Learned-Miller E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments//Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition. Marseille, France, 2008

[79] Milborrow S, Morkel J, Nicolls F. The muct landmarked face database. Pattern Recognition Association of South Africa, 2010, 201: 32-34

[80] Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3730-3738

[81] Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg AD. Presentation and validation of the radboud faces database. Cognition and Emotion, 2010, 24(8): 1377-1388

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

[83] Liu W, Wen Y, Yu Z, Li M, Raj B, Song L. Sphereface: Deep hypersphere embedding for face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 212-220

[84] Padmanabhan M. Intraface: Negotiating gender-relations in agrobiodiversity. FZG-Freiburger Zeitschrift für Geschlechterstudien, 2016, 22(2): 11-12

[85] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[86] Zhang R, Isola P, Efros A A, Shechtman E, Wang O. The unrea-

sonable effectiveness of deep features as a perceptual metric// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018: 586-595

[87] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems, 2017, 30: 6629-6640

[88] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv: 1710.10196, 2017

[89] Yi D, Lei Z, Liao S, Li SZ. Learning face representation from scratch. Clinical Orthopaedics and Related Research, 2014, abs/1411.7923

[90] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Doll P, Zitnick CL. Microsoft coco: Common objects in context// European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[91] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4829-4837

[92] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5188-5196

[93] Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 6-11

[94] Ba J, Caruana R. Do deep nets really need to be deep? Advances in Neural Information Processing Systems, 2014, 3: 2654-2662

[95] Bastani O, Kim C, Bastani H. Interpreting blackbox models via model extraction. arXiv preprint arXiv: 1705.08504, 2017

[96] Che Z, Purushotham S, Khemani R. Distilling knowledge from deep networks with applications to healthcare domain. Clinical Orthopaedics and Related Research, 2015, abs/1512.03542

[97] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. Computer Science, 2015,14(7): 38

PENG Chun-Lei, Ph.D., associate professor. His current research interests include computer vision, pattern recognition, and machine learning.

MIAO Zi-Min, master student. Her current research interests include computer vision, pattern recognition, and machine learning.

LIU De-Cheng, Ph.D., lecture. His current research interests include computer vision, pattern recognition, and machine learning.

WANG Nan-Nan, Ph.D., professor. His current research interests include computer vision, pattern recognition, and machine learning.

GAO Xin-Bo, Ph.D., professor. His current research interests include computer vision, pattern recognition, and machine learning.

Background

The Internet forms an open and dynamic virtual space, which greatly improves the speed of communication between people, but brings new challenges to the protection of personal privacy security. With the development of deep learning, identity forgery technology becomes more and more advanced, which further promotes the in-depth research on the protection of face privacy, and the society pays more and more attention to the protection of public privacy security. The leak of private data and the malicious use of deep forgery could have incalculable implications for national security and even world order. In addition, people need to take more measures to protect face privacy security. In recent years, the universal method to solve this problem is the anonymity of visual identity and privacy protection. People have higher and higher requirements for anonymous face, so anonymous identity information and anonymous attribute information should be weighed. This paper explains the concept of face anonymization and illustrates the model structure of representative face anonymization methods with diagrams. By classifying existing anonymization methods, this paper analyzes and summarizes face anonymization methods, and then summarizes the current mainstream data sets and objective evaluation indicators. By comparing the results of the existing face anonymization methods, it is found that the face anonymization technology is still in the development stage. Although the visual effect of the synthesized face image generated by some methods is better, the requirement of the data set image is higher. Therefore, the generalization performance and robustness of the model need to be further improved. The generation of anonymous face fundamentally protects personal privacy and prevents illegal organizations from stealing identity information data. The development of facial anonymization technology not only solves the problem of privacy disclosure, but also promotes the overall development of artificial intelligence technology, playing an important role in national political security, economic security, social security and network security.