面向人脸识别可信应用的隐私保护计算研究综述

袁 霖[①] 武雁尚[①] 张力元[①] 张玉书[②] 王楠楠[③] 高新波*[①③]

[①](重庆邮电大学图像认知重庆市重点实验室 重庆 400065)

[②](江西财经大学计算机与人工智能学院 南昌 330032)

[③](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)

摘 要：该文聚焦人脸识别生态，系统梳理了面向人脸识别可信应用的隐私保护计算研究进展。首先，概述了人脸识别系统的基本架构与流程，剖析非授权采集、信息泄露、梯度泄露、成员推理、人脸重建及非授权识别等关键隐私风险。随后，围绕数据变换、分布式、图像合成和对抗扰动4类主流隐私保护范式，解析加密计算、联邦学习、频域学习、特征模板保护、合成图像训练、身份保持匿名化、虚拟身份识别、差分隐私、重建攻击防御与对抗性隐私保护等10类代表性技术。最后，展望未来研究方向，包括隐私保护计算的效率提升、生成式大模型带来的新机遇与挑战、新型识别范式的构建以及标准化评估体系的建立。该文旨在为可信人脸识别研究提供系统性参考，推动其在信息物理系统中的安全与可信应用，进一步强化个人信息保护。

关键词：隐私保护计算；信息物理系统；人脸识别；身份信息

中图分类号：TN919.8 **文献标识码**：A **文章编号**：1009-5896(2026)04-1549-20

DOI: 10.11999/JEIT251063 **CSTR**: 32379.14.JEIT251063

1 引言

当前，身份认证已成为生活与工作的重要环节。相较于账号密码和证件核验，人脸识别因直观便捷、无感操作等优势，被广泛应用于手机解锁、支付验证和政务核验等场景，并在智慧城市、智能交通与公共安全等信息物理系统中发挥关键支撑作用。然而，人脸识别在提升效率与便利性的同时，也带来了严重的隐私风险。系统的训练与运行依赖大量人脸图像，一旦数据被恶意获取，可能被用于身份冒用、深度伪造，甚至黑市交易，造成严重后果。在全球范围内，欧盟、美国和中国已分别通过《通用数据保护条例》《人脸识别技术授权法案》《人脸识别技术应用安全管理办法》等，将人脸识别纳入高风险监管体系，明确要求遵循目的限定、最小必要和严格安全防护原则，凸显了人脸信息保护的紧迫性与重要性。鉴于人脸识别生态系统蕴含的巨大隐私风险，本文聚焦其在数据采集、模型训练，以及识别推理等阶段面临的隐私威胁(第2节)，系统梳理了应对各类威胁的隐私保护计算方法，包括其应用场景、工作原理、优缺点(第3节)，以及实际表现效果(第4节)，并对该研究方向的未来发展趋势进行了展望(第5节)。

2 人脸识别系统及其隐私风险概述

图1概述了人脸识别系统的整体流程、潜在隐私风险及相应的隐私保护策略。完整的人脸识别系统通常包括数据准备、模型训练以及部署与推理3个阶段。在数据准备阶段，需要构建覆盖多种人脸变化的人脸数据集，并通过清洗、归一化和数据增强等操作提升数据质量。在模型训练阶段，传统方法如主成分分析(Principal Component Analysis, PCA)用于特征降维与身份建模[1]，而基于卷积神经网络(Convolutional Neural Network, CNN)的深度学习方法则能够自动学习判别性身份特征[2–4]。在部署与推理阶段，系统通过模板注册与特征匹配完成身份识别。贯穿上述流程，各阶段均可能引入不同形式的隐私风险，整体可归纳为以下几类：

(1)非授权人脸采集：人脸识别模型训练依赖大规模、带有身份标签的人脸图像数据集。为获取足够数据，一些组织在未经授权情况下通过网络爬虫从社交平台、论坛等公开网站大规模抓取用户人脸图像，暴露严重的隐私风险。例如，美国Clearview AI公司曾从网络平台收集超300亿张人脸图像用于执法服务，引发广泛争议。为应对数据集隐私问题，研究者开始尝试利用生成模型合成虚拟人脸数据替代真实图像用于训练[5–8]。但由于合成图像在特征分布与多样性等方面仍与真实人脸存在差距，

收稿日期：2025-10-09；改回日期：2026-02-04；网络出版：2026-02-15

*通信作者： 高新波 gaoxb@cqupt.edu.cn

基金项目：国家自然科学基金(62201107, U22A2096)，重庆市教育委员会科学技术研究项目(KJQN202300606, KJQN202300619)

Foundation Items: The National Natural Science Foundation of China (62201107, U22A2096), The Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202300606, KJQN202300619)

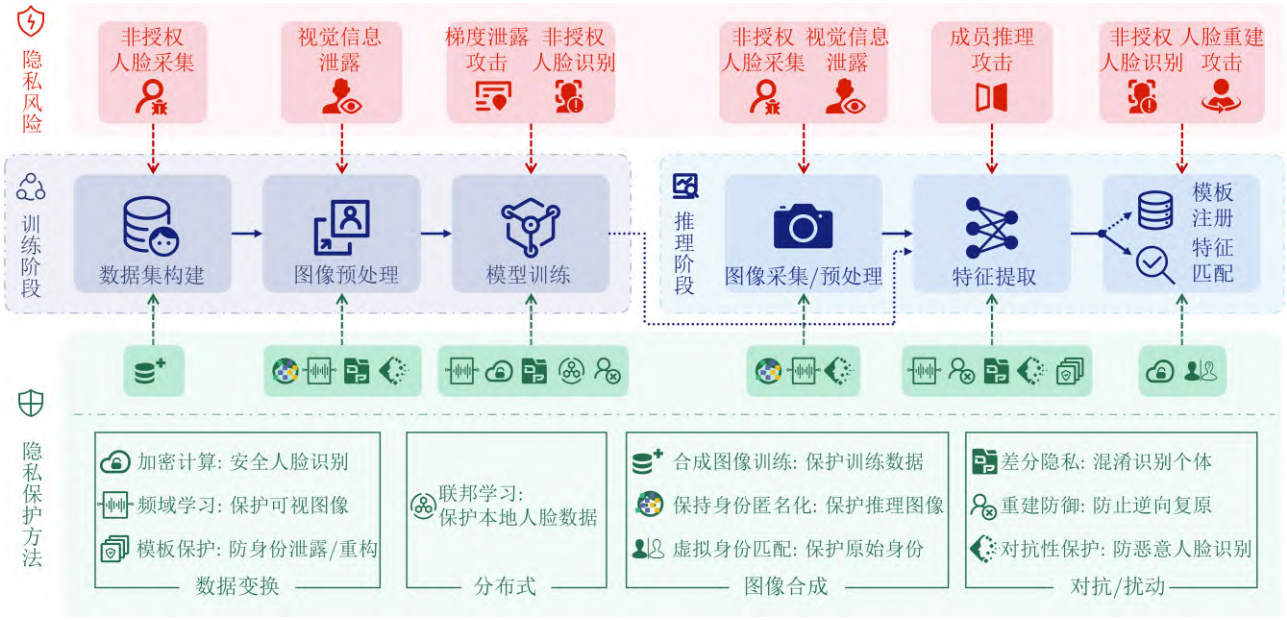

图 1 人脸身份识别系统工作流程、各阶段涉及的隐私风险，以及相应的隐私保护策略示意图

基于此类数据训练的模型用于现实场景人脸识别时的准确率仍有待提升。

(2)视觉信息泄露：传统人脸识别依赖可见光采集清晰人脸并以明文参与计算，图像在采集、处理和传输过程中均存在视觉泄露风险。加密可保护传输安全，但识别仍需解密原图；同态加密虽支持密文计算，但计算开销过高，难以实用。研究[9,10]表明，只利用图像的部分高频分量即可实现身份识别，基于此，研究者尝试采用频域学习与分离的方法，在不暴露原始清晰图像的前提下完成识别任务，然而，该类方法在识别精度与通信开销上仍面临挑战。

(3)成员推理攻击：成员推理攻击[11]是指攻击者试图判断某个样本是否曾用于目标模型的训练，通过分析模型输出的置信度、概率分布等差异，攻击者可推测样本是否属于训练集。在人脸识别中，模型对训练样本与非训练样本的响应差异可被用于推测训练数据。攻击者通常构建与识别模型结构相似的“影子模型”，获取其对不同样本的输出特征，用于训练二分类器，再利用目标模型的输出判断某个样本是否为其训练成员。为应对该类攻击，研究者尝试使用合成图像代替部分真实人脸进行训练。然而，合成数据同样面临成员推理风险。例如，Shahreza等人[12]发现，多个合成数据集中存在大量与真实人脸特征高度相似的样本，使攻击者能够推测某个个体是否出现在原始训练集中，这种现象被称为“身份泄露”。

(4)梯度泄露攻击：梯度泄露攻击是针对深度学习训练过程的隐私威胁，攻击者通过访问梯度信息可反推出原始输入甚至标签。由于梯度由样本与标签共同决定，攻击者可构造伪输入并迭代优化，使其梯度与目标一致，从而重建出与原始数据高度相似的样本。此类攻击在联邦学习场景尤为突出，即便避免了原始数据共享，若上传梯度未加保护，仍可能泄露敏感信息。经典方法如文献[13]已通过实验证明可利用模型的输入输出，以及中间梯度来进行训练数据的还原，暴露了联邦学习的隐私风险。常见防御手段包括差分隐私(向梯度加噪，简单易用但影响精度)、梯度裁剪(限制梯度幅度，成本低但效果有限)、以及加密计算(如同态加密和多方安全计算，保护最强但开销较大)。

(5)人脸重建攻击：人脸重建攻击是指攻击者利用人脸识别系统的输出信息(如预测置信度或特征向量)重建原始输入图像的过程。该类攻击主要包括两类方法：一是基于置信度反馈[14,15]，通过梯度下降迭代优化噪声图像，以最大化目标类别概率，从而重建目标人脸；二是基于特征向量映射[16,17]，将人脸特征模板输入生成模型还原图像，该类方法通常需要训练特征编码器，将人脸特征模板映射至生成模型的隐空间进而完成图像重建。为防止重建攻击，常用防护手段包括特征加密或模版保护[18](如同态加密、哈希编码)、可撤销模板生成[19–21](替代原始特征)与差分隐私[22](在特征或梯度中加入噪声)。其中，特征加密能有效防止直接还原，但计算开销较大；可撤销模板便于重置，增强安全性，但设计复杂、可能影响识别精度；差分隐私虽然实现相对简单，但引入的噪声可能降低系统的识别性能。

(6)非授权人脸识别：非授权人脸识别是指在未获用户同意的情况下采集和使用人脸图像进行身份识别的行为，广泛存在于社交平台、公共监控及商业场景中。2021年央视315晚会曝光多家品牌门店在顾客不知情下采集人脸信息用于识别分析，引发公众强烈担忧。对此，常见的保护手段包括对抗样本扰动[23,24](在图像中添加对抗性信号导致识别失效)与物理遮挡[25](佩戴隐私眼镜、面具等干扰识别系统)。前者适用于数字图像保护但在现实场景中的泛化性存在问题，后者更加实用但会直接影响用户的使用体验。

3　面向可信人脸识别的隐私保护计算方法

为应对上述隐私风险，研究者从识别系统的不同环节提出了多种保护策略，主要可归纳为数据变换、分布式计算、图像合成和对抗扰动4类。数据变换方法通过加密计算、频域学习或模板保护等手段，将原始图像或特征转换为不可见或不可逆形式以完成识别；分布式计算以联邦学习为代表，实现多方在不共享原始数据的情况下协同训练；图像合成方法借助生成模型，在训练或推理阶段以合成图像、匿名化或虚拟身份匹配替代真实人脸；对抗扰动方法则通过差分隐私、噪声注入或对抗样本抑制人脸特征的重建与识别。

3.1　基于加密计算的安全人脸识别方法

加密计算作为隐私保护人脸识别的重要技术路径，使得识别过程能够在数据加密的状态下完成，从而避免原始人脸图像和中间特征在计算过程中被暴露。在高敏感度的身份认证场景中，它提供了可证明的安全保障。现有技术主要包括同态加密、安全多方计算和矩阵变换类方法。图2展示了这些加密方法在识别系统中的应用流程，该类方法通常采用同态加密、矩阵变换等手段对人脸特征进行加密，并在密文空间完成身份匹配，防止原始特征的暴露。

Erkin等人[26]提出了一种结合Eigenface特征[1]与安全多方计算的人脸识别方案，利用Paillier半同态加密[27]和分布式密钥生成技术[28]对特征进行加密，并在密文中完成人脸匹配。Ma等人[29]提出了基于边缘计算的轻量级方案，通过加法秘密共享将人脸特征分发到多个边缘服务器，协同完成AdaBoost识别训练[30]，实现结果的秘密共享。Osadchy等人[31]采用加性同态加密与不经意传输协议，由客户端加密特征，服务器完成匹配，最后安全返回识别结果。此外，Troncoso-Pastoriza等人[32]提出基于全同态加密与Gabor特征的人脸验证方案，Jin等人[33]则结合全同态加密和不经意传输，用第三方数据库构建稀疏字典，计算人脸向量间的欧氏距离，并安全检索结果。鉴于同态加密计算开销较大，部分研究提出了优化策略以提高效率。部分研究面向云服务场景，设计了全同态加密的人脸识别协议，将计算任务外包至云端，以减轻本地负担。Boddeti等人[18]提出一种高效的同态加密人脸识别方法，利用批处理技术将多个值编码进1个多项式，从而在1次操作中完成多个同态乘法。Ibarrondo等人[34]则引入分组测试思想，使用Cheon-Kim-Kim-Song (CKKS)同态加密[35]对人脸特征分组后计算组间最大相似度，显著减少了所需的加密计算次数。除了安全多方计算和同态加密，还有一些研究采用矩阵变换类加密方法保护人脸识别过程中的数据隐私。Guo等人[36]通过密钥种子生成仿射变换(如置换、扩散、移位)对特征脸进行加密，服务器可在加密域中计算与查询图像的余弦相似度以实现匹配。Kou等人[37]构建分布式智能家居识别系统，利用随机矩阵加密特征向量，并在加密域中完成身份相似度计算。Gao等人[38]提出基于豪斯霍尔德变换(Householder Transformation)的方案，使用密钥控制变换矩阵生成，并在变换域内进行特征匹配。

加密计算在理论上可以同时保障数据隐私和识别准确性，为人脸识别系统提供了一种兼顾安全与可用性的方案。然而，由于其加密操作复杂，往往带来较高的计算开销和通信负担。当前的软硬件条件仍难以支持加密计算在多场景下的高效运行。虽然图形处理器(Graphics Processing Unit, GPU)、

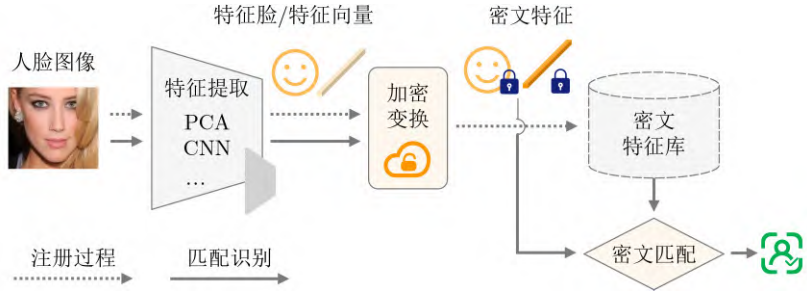

图 2 基于加密计算的安全人脸识别方法示意图

现场可编程门阵列(Field Programmable Gate Array, FPGA)或可信执行环境(Trusted Execution Environment, TEE)等硬件加速可部分缓解性能问题，但在通用性和成本方面仍有限制。因此，加密计算在实际人脸识别系统中的应用仍需在算法、系统和硬件协同上进一步突破。

3.2 基于频域学习的视觉隐私保护人脸识别方法

频域学习的基本思想是将人眼敏感的视觉信息进行频域转换或弱化，使图像在视觉上更难识别，却仍在频域保留足够的结构信息供神经网络进行特征提取。在人脸识别中，研究者利用图像中不易被人察觉、却具有区分性的高频特征，实现“人眼不可见，机器能识别”的效果，从而在不依赖高算力的条件下实现高效且更具隐私性的身份验证。图3展示了频域学习在人脸隐私保护中的应用，该类方法通过频域转换保留识别关键特征，抑制视觉敏感信息，实现人眼难辨、机器可识别的隐私保护。

Wang等人[9]提出一种基于频域掩蔽的人脸识别方法，该方法首先利用离散余弦变换(Discrete Cosine Transform, DCT)将人脸图像转为频域，再对对人眼敏感但识别作用小的频率分量施加随机扰动，从而掩蔽敏感信息，降低泄露风险，同时基本不影响识别效果。Ji等人[39]也基于DCT，在差分隐私框架下向高频特征添加拉普拉斯噪声，以降低图像可视性，保护隐私。Mi等人[10]提出DuetFace框架，在端侧设备将图像转换为频域，只上传不易被人眼识别的高频特征，同时生成补偿特征图以弥补低频信息缺失，再在云端融合这些特征，保证识别精度。为增强安全性，Mi等人[40]又提出改进版PartialFace，在高频特征上传前执行随机采样和通道打乱，进一步防止重建攻击。Henry等人[41]利用无透镜的FlatCam系统[42]采集图像，基于DCT特征进行识别，并因缺少原始图像信息而难以被重建。Mi等人[43]还提出MinusFace方法，模拟图像压缩过程，用频域残差图进行识别训练，在不暴露清晰图像的情况下实现接近原图的识别性能。

基于频域学习的人脸隐私保护方法虽具独特优势，但实际应用仍面临多重挑战。其一，通过抑制频率信息带来的隐私增强常伴随识别性能下降，且由于主流识别模型多基于空间域，频域特征的引入易导致结构不兼容。其二，现有机制难以抵御扩散模型等强重建攻击，缺乏可验证的不可逆性保障。此外，频域处理所带来的额外计算与通信开销在实际应用中也不可忽视。

3.3 人脸识别特征模版保护方法

人脸识别系统通常在注册阶段将用户的人脸特征模板存入数据库，用于后续身份匹配。这些模板如同用户的“生物密码”，一旦泄露，攻击者可能借此非法访问用户银行账户、门禁或支付系统，甚至通过重建攻击实施深度伪造重建。随着数据合规要求和隐私保护法规的不断加强，如何在保证识别性能的同时对原始特征模板进行保护，以防其被滥用，已成为人脸识别系统安全落地和规模化应用中的关键问题。图4展示了该类研究方法的基本思路，该类方法将人脸特征映射为不可逆、可撤销的二进制模板，防止特征泄露与重建攻击。

人脸特征模板保护的基本思路是将原始特征映射为不可逆、可撤销且能保持识别精度的二进制表示，称为“人脸哈希”，以实现安全高效的身份匹配。Pandey等人[44]提出为每位用户分配最大熵二进制(Maximum Entropy Binary, MEB)码，通过卷积神经网络将不同拍摄条件下的人脸图像稳定映射为相同MEB码，并使用安全哈希算法(Secure Hash Algorithm, SHA-512)生成不可逆模板存储。Talreja等人[45]结合深度哈希与神经网络解码器，利用预训练的VGG-19模型提取图像特征并通过纠错机制增强哈希模板在多种干扰下的鲁棒性。赵铖辉等人[46]设计了一个包含特征转换与生物加密的双阶段方法：前者利用轻量级网络与正交映射生成紧凑哈希模板，后者通过模糊承诺方案绑定纠错码与哈希值，以增强安全性。Zhou等人[47]基于纠错码构建模板，将人脸特征映射至用户专属编码后再计算

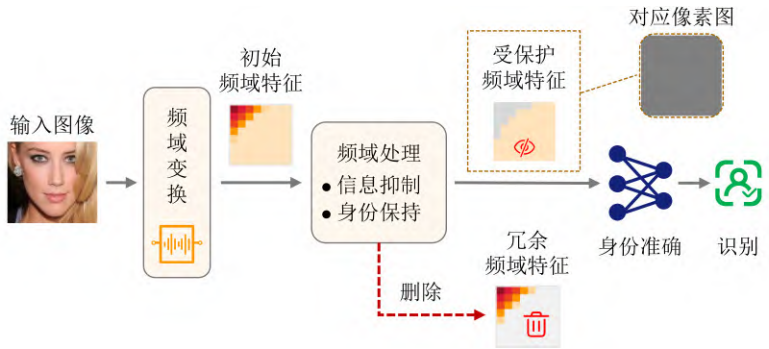

图 3 基于频域学习的视觉隐私保护人脸识别示意图

为哈希进行存储，确保身份唯一性与抗攻击能力。Shahreza等人[21]提出MLP-Hash方法，利用受用户密钥控制的随机多层感知器，将原始特征转换为二进制模板。Mai等人[48]提出的SecureFace方法通过引入随机密钥并生成安全草图(Secure Sketch)避免存储明文密钥，提升模板安全性。Gao等人[49]采用多重部分沃尔什变换[50]与SimHash[51]对特征进行投影与离散处理，生成最终的受保护模板。Zhong等人[19]提出SlerpFace方法，通过将模板扰动为类噪声分布抵御生成模型反推，并结合注意力机制对特征分组加权，兼顾识别性能与不可逆性。

人脸模板保护通过将特征映射为特定的二进制编码，实现不可逆性(难以还原原始图像)与可撤销性(泄露后可重新生成)的隐私目标。然而，二值化过程中常伴有特征精度损失，影响识别性能。例如，MLP-Hash[21]方法在抵御重建攻击方面效果显著，但会引入1%～3%的识别率下降。因此，如何减少模版保护带来的识别性能损失，是该领域的关键研究方向。

3.4　基于联邦学习的分布式人脸识别方法

联邦学习是一种分布式机器学习范式，各参与方在本地训练模型，仅上传模型参数供服务器聚合，从而在不共享原始数据的情况下完成全局模型更新。该机制减少了数据集中传输带来的隐私泄露风险。部分研究[52,53]进一步将联邦学习与前述同态加密相结合，增强隐私保护能力。在人脸识别应用中，联邦学习可用于多个机构或用户设备间联合训练模型(见图5)，该类方法允许多个客户端在本地使用各自人脸数据训练模型，仅将模型参数上传至中心服务器进行聚合，无需共享原始数据），从而在保证模型识别性能的同时保护本地人脸隐私，特别适合于医疗、金融等对本地人脸数据保护要求较高的应用场景。

Aggarwal等人[54]提出了FedFace架构，通过联邦学习实现隐私保护人脸识别。该方法在服务器端初始化全局模型，将参数下发至客户端，客户端本地训练后上传更新，服务器端再聚合优化全局模型，并通过正则化平衡特征分布。该过程不断迭代直至模型收敛。为解决联邦学习中的单点故障[55]与参与度不足[56]问题，Zheng等人[57]引入区块链，提出BFLFace框架，借助链上存储与去中心化聚合增强系统鲁棒性，并设置激励机制提升用户参与度。Niu等人[58]提出FedGC框架，引入梯度校正和Softmax正则化，提升异构数据下的识别准确性与训练稳定性。Liu等人[59]的FedFR框架通过解耦模块实现用户特征的本地定制化，增强识别个性化。Woubie等人[60]还引入安全聚合器以增强模型更新过程中的隐私性，并在边缘设备上利用生成对抗网络生成伪造人脸样本，提升模型泛化能力。Kim等人[61]提出在每轮聚合后用部分客户端评估新旧模型，若新模型无效则回滚，提升系统稳定性和适应动态数据的能力。

联邦学习通过在不共享原始人脸数据的前提下训练模型，有助于降低训练阶段的隐私风险。但其仍面临如下3方面挑战：频繁同步导致通信成本高；上传梯度可能泄露隐私；客户端数据分布差异易造成“客户端漂移”，影响泛化与个性化效果。因此，提升联邦学习的效率与安全性仍是未来重要的研究方向。

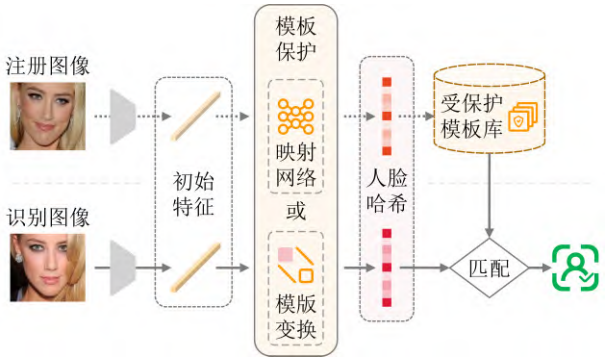

图 4　人脸识别特征模版保护示意图

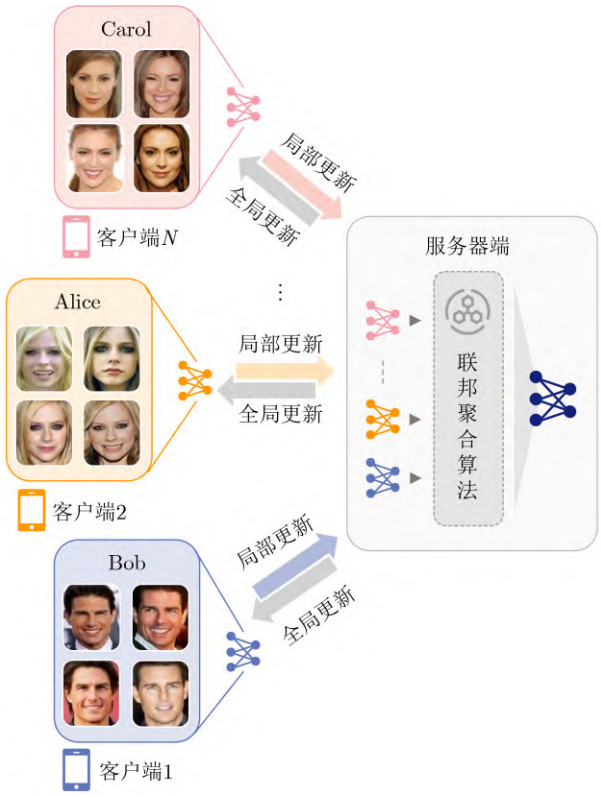

图 5　基于联邦学习的分布式人脸识别方法示意图

3.5 基于合成图像训练的人脸识别方法

主流人脸识别模型的构建依赖大规模真实人脸数据训练，数据采集、存储与使用过程中普遍存在隐私与安全风险，甚至可能遭受推理攻击获取敏感信息。鉴于人脸识别基于特征相似性匹配，无需测试身份出现在训练集中，合成人脸识别(Synthetic Face Recognition, SFR)通过生成模型合成虚拟人脸替代真实数据用于训练，在降低隐私风险的同时支持模型预训练与能力扩展，尤其适用于数据获取受限或合规成本较高的场景。图6展示了该类方法的基本原理，该类方法利用生成模型合成虚拟人脸数据替代真实图像训练，从源头规避身份泄露。

该类方法的核心挑战在于构建兼具身份多样性与类内自然变化的高质量合成数据，使虚拟身份之间区分明确，身份内部在姿态、表情和光照等方面具有真实分布特征。Qiu等人[7]利用DiscoFace-GAN[62]提出身份混合与域混合策略，通过潜在空间插值生成中间身份，并结合少量真实图像增强模型的泛化能力。Li等人[63]提出LightFace，基于Bayesian GAN[64]生成不含真实身份的人脸图像，同时在标签中加入差分隐私噪声，防止隐私泄露。Boutros等人[6]提出SFace，首先在真实数据上预训练人脸识别模型，再利用由StyleGAN2-ADA[65]生成的人脸数据对识别模型进行微调，提升合成数据的域适应性。Bae等人[66]构建了DigiFace-1M数据集，采用3D可变形模型[67]和渲染技术[68]合成覆盖姿态、表情和光照变化的二维人脸图，并结合图像增强以扩展多样性。Kolf等人[69]提出IDnet，通过引入身份网络ID-3模块，强化身份可分性，提升识别性能。Kim等人[8]提出DCFace，使用双重条件控制身份和风格调节范围，优化类间与类内聚类结构。Boutros等人[70]提出IDiff-Face，基于条件潜在扩散模型生成多样人脸，进一步提升了识别精度。Melzi等人[71]提出GAN-DiffFace，先用GAN合成身份图，再通过扩散模型添加丰富的属性，如配饰、姿态和场景等。Xu等人[72]提出ID3模型，结合身份和属性双条件指导图像生成，并通过扰动引入类内变化，保持身份一致性的同时增强多样性。Sun等人[5]发现“半困难”样本(与类中心身份相似度介于0.70～0.76)对模型训练最有效，进而提出Cemi-Face方法生成这类样本用于提升精度。Shahreza等人[73]提出HyperFace，将数据生成视为超球面嵌入空间布局优化问题，通过扰动与正则化调控类间距离与类内多样性，提升数据质量。Mi等人[74]提出了一种名为MorphFace的扩散式人脸生成方法，通过结合3D人脸模型提取细粒度风格和预训练识别模型提取身份信息，实现了兼顾身份一致性与风格多样性的虚拟人脸生成。

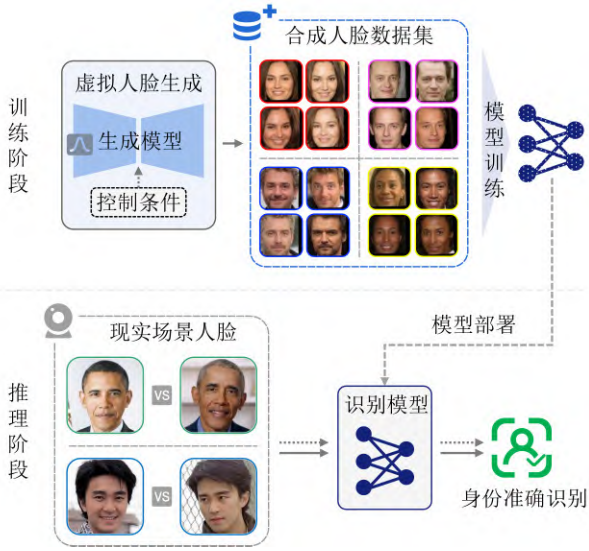

图 6 基于合成图像训练的隐私保护人脸识别示意图

合成人脸图像不包含真实人脸，可在训练阶段有效降低隐私泄露风险。然而，基于合成数据的人脸识别仍面临挑战。其一，合成数据在类间和类内多样性上的不足可能影响识别精度；其二，生成模型基于真实人脸训练，生成图像仍可能泄露身份信息，即“身份泄露”问题[12]。此外，现有研究多聚焦于合成数据的构建方法，而在合成数据条件下如何优化模型训练过程仍然缺乏深入探讨。

3.6 保持原始身份特征的人脸匿名化方法

人脸匿名化[75]是一类用于隐藏图像中视觉身份信息的技术，通常通过模糊、像素化、遮挡或人脸替换等方式，在一定程度上保留人脸结构的同时，降低其被人眼或算法识别的可能性。近年来，研究者提出了“人眼不可识别、机器可识别”的新范式，以满足智能视频监控、刷脸支付等场景中“认人不看图”的应用需求。该类方法在生成匿名图像时，并非简单地抹除身份信息，而是以隐式方式保留对机器识别至关重要的判别特征，使匿名人脸难以被人眼辨认，却仍可被后端识别系统稳定识别。由此，该范式在保障身份验证功能正常运行的同时，有效隐藏了前端可见的视觉隐私。图7展示了

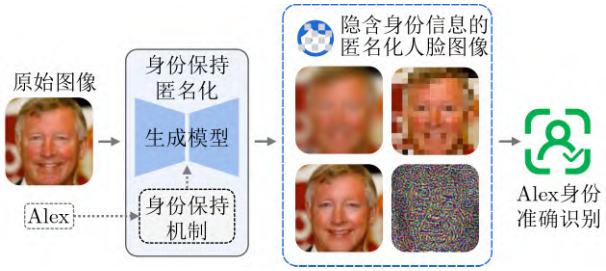

图 7 保持原始身份特征的人脸匿名化方法示意图

其基本原理，该类方法对图像进行视觉匿名处理(模糊、替换或扰动)，隐去人眼可辨身份，同时保留机器可识别的判别特征。

Li等人[76]最早提出身份保持的人脸匿名化方法，借鉴深度换脸思想，将人脸的外观与身份特征解耦，并与他人身份特征融合，在身份分类器的引导下生成视觉匿名、但机器可识别的换脸图像。然而，该方法在匿名化强度与识别精度间仍存在权衡问题。为进一步提升效果，Li等人[77,78]引入身份感知区域检测模块，生成面部重要性解析图，引导匿名化处理更精准地作用于关键区域，从而在增强匿名效果的同时保持较高的识别率。Yuan等人[79]提出一种通用的身份保持匿名化框架PRO-Face，借鉴图像隐写思想，将原始图像嵌入已匿名处理(如模糊、像素化和换脸)后的图像中。通过孪生结构的隐写网络，确保视觉效果接近匿名图像，同时保留机器可提取的真实身份信息。为进一步提升匿名化人脸的身份识别性能，同一研究团队[80]在此基础上提出了PRO-Face C架构。该方法在原有人脸识别模型中引入特征补偿机制，对匿名化过程中被削弱或丢失的身份判别特征进行显式补充，从而显著提高了基于匿名图像恢复和识别原始身份的准确率。Su等人[81]受对抗样本的启发，向图像添加强干扰噪声，几乎完全抹除视觉信息，但仍保留可被识别模型提取的身份特征，识别精度接近原始图像。此外，Wang等人[82]则利用身份与外观特征解耦引导预训练的StyleGAN生成高质量、身份可识别的匿名化人脸，进一步提升识别性能与隐私保护效果。

人脸识别与匿名化在目标上天然冲突，使“可识别的匿名化”看似成为悖论，现有方法往往难以同时兼顾识别性能与匿名化强度。然而，人眼与机器对视觉信息的感知机制存在本质差异，使二者可从同一图像中提取不同特征，这一现象也构成了对抗样本与信息隐藏的理论基础。因此，在强化视觉匿名化的同时保留机器可识别特征仍具有一定理论可行性，值得进一步研究。此外，传统模糊化等匿名化手段易受到超分辨率重建[83]等还原攻击，其安全性仍有待审视。

3.7　基于虚拟身份的隐私保护人脸识别方法

基于人脸合成的思路，部分研究提出在采集图像后对原始人脸进行“虚拟化”处理，将其转换为在外观和身份特征上显著不同的虚拟图像，并确保同一身份的图像在特征空间中仍然紧密聚类，不同身份则保持区分，这一特性被称为“虚拟身份一致性”。借此，系统可在不暴露原始人脸视觉信息与身份的情况下，实现虚拟人脸的模板注册与识别。这一思路尤其适用于用户需要在公开或半公开平台(如元宇宙、线上会议)使用人脸进行认证或交互，但不愿暴露真实容貌的场景，其目标是为用户生成一个可长期使用且一致的“数字面具”。图8展示了其基本原理与流程，该类方法为每个真实身份生成一个外观不同的虚拟人脸用于系统注册与识别。虚拟身份在视觉和特征层面均与原始身份分离，从而实现图像层与特征层的双重隐私保护。

Yuan等人[84]首次提出了基于虚拟身份一致性的虚拟人脸匹配方法(Identifiable Virtual Face Generator, IVFG)，该方法利用特定密钥将原始人脸的身份向量转换为虚拟特征向量，并将其映射为StyleGAN[85]的潜在输入，生成高度逼真、具备身份一致性的虚拟人脸，实现隐私保护下的人脸识别。然而，IVFG未充分考虑生成图像在姿态与背景上的一致性，限制了其实际应用效果。为提升实用性，Wang等人[86]构建了基于虚拟身份变换的识别系统，为每位用户分配虚拟身份特征向量，并通过身份与属性解耦和重建机制，生成姿态与表情保持一致的虚拟人脸用于匹配，从而在保障隐私的同时实现准确识别。Wang等人[82]设计了解耦人脸表示的虚拟身份转换方法，在实现视觉隐私保护的同时保持较高识别性。该团队后续提出CanFG方法[87]，通过去除物理身份(Physical IDentity, PID)并嵌入虚拟身份(Virtual IDentity, VID)，结合距离保持机制与辅助模块，生成可撤销、可更新的匿名人脸图像，增强了身份管理灵活性。近期，Wang等人[88]提出密钥驱动的人脸匿名与身份认证识别框架(Key-driven Face Anonymization and Authentication Recognition, KFAAR)，该方法结合密钥控制的虚拟人脸生成模块与身份认证模块，在保留面部姿态与表情的前提下生成可在虚拟特征域匹配的虚拟人脸，并创新性地引入一个基于密钥的身份复原模块，以实现虚拟人脸的原始身份认证，为隐私保护下的人脸识别提供了新思路。

第3.6节所介绍的方法期望匿名化人脸所提取

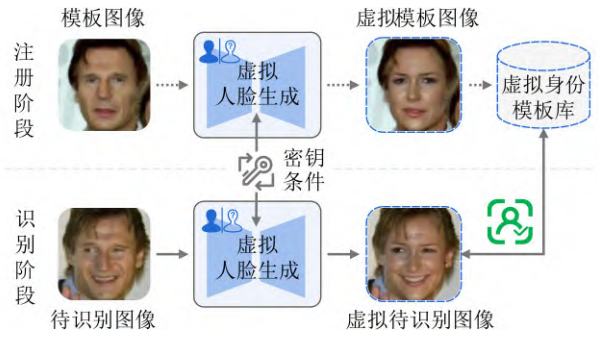

图 8 基于虚拟身份的隐私保护人脸识别示意图

的身份特征仍与原始人脸高度相似。相比之下，本节所述方法更关注构建一个与原始身份不同的“虚拟身份”，并保证识别系统在数据(图像与特征)传输和存储过程中始终以该虚拟身份作为表征对象，从而同时降低图像和特征层的隐私泄露风险。然而，该类研究仍面临两大挑战：一是生成足够逼真的虚拟人脸以保障视觉迷惑性；二是构建能有效区分样本身份的虚拟特征空间，即在保持同一身份虚拟人脸聚集的同时，拉开不同身份之间的距离。

3.8 基于差分隐私的识别信息保护方法

差分隐私(Differential Privacy, DP)是一种用于保护个体数据隐私的数学框架，旨在数据发布或共享过程中防止泄露个人信息。其核心思想是在数据处理时引入随机性(如加噪)，使得加入或移除任意个体数据对算法输出的影响微乎其微。换言之，即使观察者看到算法结果，也难以判断某个具体数据是否被使用，从而有效保护个体隐私。差分隐私不仅关注如何加噪声，还提供了一种可理论证明的概率框架，用于判断算法是否满足隐私保护标准。其中，Dwork等人[89]提出的ϵ-差分隐私(ϵ-DP)模型是最具代表性的差分隐私框架。该模型通过隐私预算ϵ约束随机算法在相邻数据集上的输出差异，从而限制单条记录对结果的影响；ϵ越小，算法输出越难区分，隐私保护强度越高。在人脸识别领域，差分隐私理论框架尤其适用于需要公开发布人脸数据集，或在研究与商业应用中共享训练模型的场景。

差分隐私通常通过在特征、模型参数或输出结果中引入随机噪声来限制单个样本对模型输出的影响(如图9所示，该类方法在特征提取、识别模型或预测标签中添加可控噪声，通过满足差分隐私定义，提供可证明的隐私保障)，从而降低身份信息被推断或重构的风险。常见加噪机制包括拉普拉斯、高斯和指数机制，依据加噪位置可分为本地差分隐私与全局差分隐私，其噪声强度由数据敏感度决定。Chamikara等人[22]提出特征脸差分扰动(Privacy using EigEnface Perturbation, PEEP)协议，通过PCA提取特征脸(Eigenface)，并对其应用局部差分隐私机制，在特征上加入拉普拉斯噪声。实验表明，在较强隐私保护条件下($\epsilon \leq 9$)，该方法能显著混淆视觉信息，有效防止原始图像被恢复，同时保持较高识别准确率[90]。Mao等人[91]提出基于边缘计算的差分隐私人脸识别框架，将识别模型的第一层部署在用户设备上，并在该层输出的特征图上加入满足ϵ-差分隐私的高斯噪声，再传输至边缘服务器继续识别任务。由于噪声的引入，服务器难以还原原始图像信息，同时保护了模型参数的隐私。此外，频域学习方法[39]也引入差分隐私机制，该研究设计了动态隐私预算分配策略，根据频率通道的敏感度调整噪声强度，并通过损失函数优化隐私预算，以平衡隐私保护与识别精度。在联邦学习框架下，PrivacyFace方法[92]对模型参数添加高斯噪声，以实现差分隐私保护。而在基于合成数据的识别方法中，Li等人[63]在训练阶段向标签引入拉普拉斯噪声，确保即使攻击者访问模型输出，也无法还原原始数据。

差分隐私在数据发布场景提供了严格的隐私保护验证，但在人脸识别应用中仍存在局限：由于人眼对图像高频区域不敏感，差分隐私在此类区域添加的噪声往往不易察觉，难以有效保护视觉隐私[93,94]；其次，为增强隐私保护，通常需使用较小的隐私预算，这会显著影响人脸识别的准确性。因此，差分隐私用于面部视觉信息保护的实用性值得进一步探讨。

3.9 人脸识别系统图像重建攻击防御方法

研究表明，主流人脸识别模型普遍容易遭受人脸重建攻击[14,16,17]，即借助人脸识别模型的输出(身份特征、相似度、预测概率等)重建出用户人脸图像。即使具备一定隐私保护能力的人脸识别方法在应对重建攻击时也往往防御效果有限[11,12,17,58]。为降低重建攻击风险，研究者在训练阶段引入对抗学习或特征混淆机制，借助模拟攻击网络，引导识别模型学习具备抗重建能力的识别特征，同时保持识别性能。图10展示了该类方法的基本思路，该类方法在训练或推理阶段引入对抗性模块或特征扰动机制，使提取的人脸特征难以被攻击者用于重建原始图像。

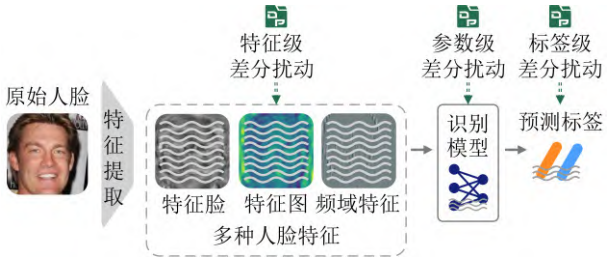

图 9 基于差分隐私的识别信息保护示意图

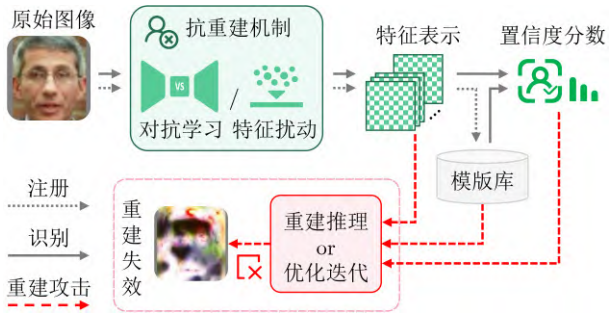

图 10 人脸识别系统图像重建攻击防御方法

Xiao等人[95]提出了一种对抗性学习框架，用于抵御基于身份特征的人脸重建攻击。该框架由编码器、分类器和解码器组成：编码器提取特征，分类器执行识别任务，解码器模拟攻击者试图从特征重建人脸图像。在训练过程中，编码器与解码器之间形成对抗关系，使编码器能够输出难以被准确重建的特征表示，从而提升隐私保护能力。Li等人[96]提出了DeepObfuscator框架，通过引入混淆器以抑制特征中可被用于图像重构或敏感属性推断的信息。Wang等人[20]提出的AdvFace方法，在服务器端训练影子模型模拟重建攻击行为，通过其梯度信息向人脸特征中注入对抗性噪声，从而干扰重建过程。实验表明，该方法在轻微影响识别准确率(下降约1%)的前提下，显著增强了对重建攻击的防御效果。Jin等人[97]则提出了FaceObfuscator方法，基于高频特征混淆策略来提升抗重建能力。该方法首先在DCT频域中去除低频信息，仅保留高频特征用于识别；随后对高频图施加随机方向和尺度扰动，阻断重建网络的反向传播路径，干扰其对图像的还原。Wei等人[98]提出了一种无需重新训练现有模型即可直接部署的特征变换机制。该方法通过将人脸特征随机分片，并经异构卷积网络进行非线性映射与融合，生成受保护的特征。通过联合优化识别与隐私损失，该方法能有效抵御黑盒与白盒重建攻击。此外，前述的部分频域学习方法也结合了特征扰乱机制以提升安全性，例如PartialFace[40]通过对高频特征图进行随机采样与通道打乱，提升特征的唯一性；MinusFace[43]则采用通道随机洗牌策略，有效增强了模型对图像重建攻击的抵抗力。

在训练人脸识别模型时引入对抗学习或特征扰动机制，有助于减缓重建攻击带来的隐私泄露风险。但现有方法多针对特定攻击模型设计，缺乏对更复杂生成模型(如扩散模型)的有效防御，泛化能力仍待提升。同时，多数攻击场景设定理想化，难以贴合真实应用。Zhang等人[15]的研究也指出，在更接近实际的环境中，现有防护手段表现有限。此外，这些防御机制往往会削弱识别准确性。因此，提升抗重建方法的泛化性、鲁棒性及实用性，是未来研究的关键方向。

3.10　防人脸识别的对抗性隐私保护方法

随着社交媒体中人脸图像的广泛公开，其被滥用于未授权模型训练或跨平台身份关联分析的风险不断上升。为此，“防人脸识别”的对抗性隐私保护机制受到关注。该方法通过在面部添加人眼难以察觉的微小扰动，在不影响视觉观感的前提下，有效干扰人脸识别模型的训练或推断过程。与此前以提升识别准确性为核心的思路不同，对抗性保护以“反识别”为目的，二者构成了互补的隐私安全闭环：传统方法主要缓解系统正常运行中的“内生”隐私风险；而对抗性方法则旨在防范技术被恶意滥用所产生的“衍生”风险。两者结合，方能构建更完整的人脸识别应用隐私保护体系。图11展示了这类方法的基本流程，该类方法通过在面部图像中添加人眼难以察觉的微小扰动，干扰未经授权的识别系统，防止人脸被恶意识别。

Shan等人[99]提出了Fawkes法，在人脸图像中添加像素级扰动(称为“cloaks”噪声)，使识别模型在训练后无法正确识别原始图像。该方法在多种训练条件下可达95%以上的保护效果，即使部分清晰图像泄露，防护率仍高于80%。Yang等人[100]则针对特定目标身份优化生成扰动掩码，以误导模型识别。Zhong等人[101]提出“一人一掩码”(One Person One Mask, OPOM)的方法，为每个人定制通用掩码，通过远离原始身份特征空间的方向优化扰动，并设计多种子空间建模方式，如仿射包、类中心、凸包等，在黑盒识别模型中验证有效性。Liu等人[102]提出了AdvCloak框架，通过两阶段训练为每个个体生成专属对抗掩码：先提升掩码对面部变化的适应性，再增强其在特征空间中对同一身份图像的泛化能力。相比OPOM[101]等需多次迭代优化的方法，AdvCloak 仅需1次前向传播即可生成掩码，将平均处理时间由数秒级降至约 0.05 s，在保护效果、视觉质量与效率之间实现了更好的平衡。Dong等人[103]发现多数模型提取的人脸特征趋于平均，因此利用海量数据的均值特征指导扰动生成，并通过离散余弦变换限制低频扰动，提升视觉质量。Tang等人[23]提出通用扰动策略，通过优化扰动使其远离多个哈希模型的特征中心，实现跨模型迁移防护。由于对抗扰动往往带来明显视觉伪影，Adv-Makeup[104], AMT-GAN[24], CLIP2-Protect[105], DiffAM[106]等方法将对抗性扰动融入人

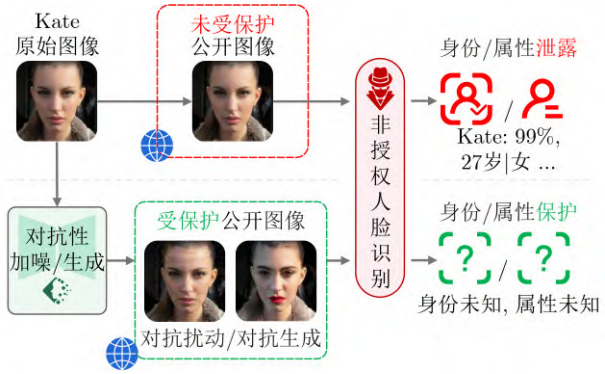

图 11 防人脸识别的对抗性隐私保护示意图

脸妆容的改变，实现在“美颜”的同时保护身份信息。由于对抗性扰动通常被视为高频噪声，易在扩散模型的去噪过程中逐步被消除，这一现象被称为“扩散净化效应”。为减弱该效应，Salar等人[107]提出对潜在编码进行对抗性修改，并引入无条件嵌入以引导图像生成。此外，相关研究还包括可迁移对抗性保护方法[108]，可逆对抗性保护方法[109]，以及软生物属性(如年龄、性别、种族)保护方法[110, 111]。

对抗性人脸隐私保护方法旨在尽可能保持图像视觉质量的同时，阻止未经授权的识别模型提取人脸身份或属性信息。然而，现有方法仍面临多重挑战：尽管许多方法在客观图像质量指标上表现良好，但生成图像仍可能出现肉眼可见的噪声或伪影，且对主观视觉质量的系统性评估仍较欠缺。此外，大多数方法在应对未知识别模型时防护效果显著下降，暴露出其在隐私保护泛化能力方面的局限。

4 代表性方法性能对比与分析

本节对前述各类方法中的若干代表性研究工作进行了深入剖析，列举了主流方法所采用的测试数据集，并从人脸识别性能、隐私保护效果和实际可用性等维度进行横向比较，系统梳理了不同类型方法的优势与局限。各代表性方法的具体对比如表1所示。

数据集：人脸识别及其隐私保护研究中广泛使用多个公开数据集，涵盖从静态图像到视频、多姿态、多年龄段等多种复杂条件。常用的人脸数据集包括LFW[112], CelebA[113], VGGFace2[114], CASIA-WebFace[115], CFP-FP[116], AgeDB[117], CPLFW[118], CALFW[119], FEI[120], IJB-B[121], IJB-C[122], Yale[123], ORL[124]和MOBIO[125]等。其中，LFW, CALFW和CPLFW系列聚焦于非受限环境下的人脸验证问

表 1 代表性研究工作的实验方法与性能对比

类别	应对风险	代表性研究	测试数据集	人脸识别性能(%)	隐私保护性能	可用性评估
加密计算	(r4) (r5)	Secure Face-FH[18]	LFW, IJB-A, IJB-B, CASIA-WebFace	TAR ↑ (67.89～99.11)	理论性隐私保证	匹配耗时(2.45 ms) 存储开销(16 kB)
		Efficient-PPFR[37]	LFW	TAR ↑ (97.72)		注册耗时(2.614 s) 匹配耗时(2.718 s)
		Gao等人[38]	Yale, ORL	ACC ↑ (89.29～96.15)		识别总耗时(2.467 s)
频域学习	(r2)	PPFR-FD[9]	LFW, CFP-FP, AgeDB, CPLFW, CALFW, VGGFace2	ACC ↑ (90.78～99.68)	SSIM ↓ (0.713), PSNR ↓ (15.66)	N/A
		DuetFace[10]	LFW, CFP-FP, AgeDB, CPLFW, CALFW, IJB-B, IJB-C	ACC ↑ (92.10～99.82)	SSIM ↓ (0.866), PSNR ↓ (19.88)	
		PartialFace[40]		ACC ↑ (92.03～99.80)	SSIM ↓ (0.591), PSNR ↓ (13.70)	
模板保护	(r5)	MLP-Hash[21]	LFW, MOBIO	TAR ↑ (90.90～100)	不可链接性 ↓ (0.01) 不可逆性 ↑ (9.05)	保护执行时间 ↓ (62 μs)
		Simhash[49]	FEI, CASIA-WebFace, LFW	TAR ↑ (94.06～100)	字典攻击复杂度 ↑ (2^{411}～2^{430}次) 不可连接性 ↓ (0.039～0.043) 不可逆性 ↓ (0.04～0.08)	注册耗时 ↓ (10.1～10.8 ms) 匹配耗时 ↓ (69～72 μs)
		SlerpFace[19]	LFW, CFP-FP, AgeDB, CALFW, CPLFW, IJB-B, IJB-C	ACC ↑ (88.90～99.42)	不可链接性 ↓ (0.05) 暴力破解次数 ↑ (3.6^{49}次)	注册耗时 ↓ (0.35 s) 匹配耗时 ↓ (0.17 s)
联邦学习	(r3)	FedFace[54]	LFW, IJB-A, IJB-C	ACC ↑ (83.79～99.28)	经验性隐私保证	N/A
		FedFR[59]	IJB-C	ACC ↑ (85.21)		
合成图像训练	(r1) (r3)	HyperFace[73]	LFW, CFP-FP, AgeDB, CPLFW, CALFW	ACC ↑ (87.07～98.67)	经验性隐私保证	N/A
		CemiFace[5]		ACC ↑ (88.86～99.22)		
		MorphFace[74]		ACC ↑ (90.07～99.35)		
保持身份匿名化	(r2)	PRO-Face[79]	LFW, CelebA, VGGFace2	TAR ↑ (88.4～94.7)	SSIM ↓ (0.527～0.875) LPIPS ↑ (0.111～0.638)	N/A
		Li等人[78]	CelebA, VGGFace2	TAR ↑ (61～89)	属性识别准确率 ↓，如年龄23～49%，性别13～41%，种族10～35%	FID ↓ (41.64～54.04)
		Wang等人[82]	CelebA, VGGFace2	AUC ↑ (88.5～96.9)	SSIM ↓ (0.306～0.315) LPIPS ↑ (0.559～0.588) MAE ↑ (0.251～0.256)	HPS ↓ (2.315～3.831) EDR ↑ (69.2～73%)

续表 1

类别	应对风险	代表性研究	测试数据集	人脸识别性能(%)	隐私保护性能	可用性评估
虚拟身份匹配	(r2) (r5)	IVFG[84]	LFW, CelebA	AUC ↑ (99.4～99.9) EER ↓ (1.8～3.5)	PSR ↑ (98.8%)	FDR ↑ (100%), FID ↓ (6.17)
		CanFG[87]	CelebA, VGGFace2	AUC ↑ (95.1～98.8), EER ↓ (4.5～10.1)	PSR ↑ (98.2～99.2%)	FID ↓ (9.43), SSIM ↑ (0.823)
		KFAAR[88]	LFW, CelebA	AUC ↑ (97.3～99.2), EER ↓ (8.9～9.2)	PSR ↑ (92.2～96.2%)	FDR ↑ (100%), EDR ↑ (80.5～83.3%), FID ↓ (6.82～7.29)
差分隐私	(r2) (r3) (r4) (r5)	Mao等人[91]	LFW	ACC ↑ (82)	理论性隐私保证	N/A
		PEEP[22]	LFW, CFP-FP, AgeDB, CPLFW, CALFW, IJB-B, IJB-C	ACC ↑ (5.82～98.41)		
		Ji等人[39]		ACC ↑ (89.33～99.48)	PSNR ↓ (14.28), COS ↓ (0.214)	
重建攻击防御	(r5)	AdvFace[20]	LFW, CFP-FP, AgeDB-30	ACC ↑ (86.35～97.78)	SSIM ↓ (0.28), PSNR ↓ (6.97), MSE ↑ (0.206), SRRA ↓ (4.03%)	N/A
		MinusFace[43]	LFW, CFP-FP, AgeDB, CPLFW, CALFW, IJB-B, IJB-C	ACC ↑ (91.90～99.78)	SSIM ↓ (0.50), PSNR ↓ (10.98)	
		FaceObfuscator[97]		ACC ↑ (88.48～99.68)	SSIM ↓ (0.471), PSNR ↓ (12.71), MSE ↑ (0.057), COS ↓ (0.004)	存储开销：98 kB/pic 推理耗时：1.18 ms/pic
对抗性隐私保护	(r2) (r6)	OPOM[101]	Celeb-1M, LFW	N/A	PSR (1:N) ↑ (78～86.6%@R-1-U, 69.4～79.3%@R-5-U)	N/A
		CLIP2Protect[105]	CelebA, LFW		PSR (1:1) ↑ (64.9%), PSR (1:*N*) ↑ (82.2%@R-1-U, 23.4%@R-1-T)	PSNR ↑ (19.31), SSIM ↑ (0.75), FID ↓ (26.5)
		Salar等人[107]	CelebA		PSR (1:1) ↑ (79.17%)	PSNR ↑ (27.72), SSIM ↑ (0.84), FID ↓ (26.5)

题，并进一步引入年龄和姿态变化以增强挑战性；CelebA包含丰富的人脸属性标签和较少噪声，常用于属性编辑与训练任务；VGGFace2与CASIA-WebFace等提供大规模、高多样性的人脸图像数据，适合用于模型训练与大规模识别评估；AgeDB专注于跨年龄段识别评估；CFP-FP则聚焦于正侧面视角识别的性能分析；FEI数据集以表情变化为主，支持面部表情分析研究；IJB-B和 IJB-C数据集则整合了图像与视频数据，覆盖极为复杂的现实环境，是当前评估人脸识别系统鲁棒性的关键基准之一。

人脸识别性能评估：在人脸识别性能的评估方面，常用的评价指标包括准确率(ACCuracy, ACC)、正确接受率(True Acceptance Rate, TAR)、错误接受率(False Acceptance Rate, FAR)、错误拒绝率(False Rejection Rate, FRR)、AUC值(Area Under the Curve，即ROC曲线下面积)，等错误率(Equal Error Rate, EER)。ACC用于衡量整体识别的正确比例，适合类别分布均衡的任务；FAR反映系统误将一对不匹配的人脸错误识别为匹配的概率，是评估安全性的重要指标；TAR表示系统在特定FAR阈值下正确接受匹配身份的能力；AUC体现模型在不同阈值下的综合识别能力；EER则是在FAR和FRR相等时的错误率，常用作评估系统整体平衡性的标准。从表1可以看出，在较为简单的数据集(如LFW)上，多数方法已能够实现接近于原始无保护图像的识别性能；但在如IJB-B/C等复杂场景的数据集上，仍存在较大的提升空间。值得注意的是，一些隐私保护方法(如身份保持匿名化、差分隐私)通常会出现人脸识别准确率的下降，这源于其在保护过程中对图像语义特征造成的不可避免干扰。

隐私保护性能评估：在隐私保护性能的评估方面，不同方法的保护对象不同，所采用的评估手段也存在差异。基于加密计算和差分隐私的方法通过严格的数学建模与概率分析，为隐私保护提供了较强的理论保障。其他多数方法通常使用峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)、结构相似度指数(Structural Similarity Index Measure, SSIM)[126]、感知相似性(Learned Perceptual Image Patch Similarity, LPIPS)[127]以及余弦相似度(COsine Similarity, COS)等指标度量保护前后图像或特征之间的差异，用来量化隐私性。部分方法采用保护成功率(Protection Success Rate, PSR)作为隐私衡量指标，但该指标在不同场景具有不同的定义：在IVFG[84], CanFG[87]和KFAAR[88]等虚拟身

份匹配方法中，PSR(1:1) 表示原始人脸与匿名化人脸在1:1验证任务中未成功匹配的概率，这类方法的识别率通常较高；在对抗性隐私保护方法中(OPOM[101], CLIP2Protect[105]和Salar[107])，PSR被扩展至1:N查询，并区分为有目标(R-N-T)和无目标(R-N-U)两类：R-N-T表示Top-N结果中包含目标身份的比例，用于衡量将受保护人脸误识别为指定目标的能力；R-N-U表示Top-N结果中不含原始身份的比例，用于评估摆脱原始身份关联的效果。如表1所示，大部分对抗性方法的保护成功率仍有较大提升空间。人脸模版保护研究(如MLP-Hash[21], SimHash[49]和SlerpFace[19])通常采用不可关联性[128]、不可逆性分析，以及暴力破解所需计算开销等来衡量隐私强度。此外，联邦学习与基于合成图像训练的方法主要通过经验性假设来保证隐私性，即假设用户数据不上传至中心服务器或训练过程不使用真实人脸图像即保护了隐私，其理论性与严谨性有待提升。

可用性评估：隐私保护计算的核心在于在保障隐私的同时兼顾数据可用性，现有研究多从计算开销和结果质量等方面进行评估。加密计算与模版保护方法多通过注册时间、匹配时间和存储开销来衡量，例如，Secure Face-FH[18]匹配耗时约2.45 ms，存储开销为16 KB，而Gao方法[38]的识别耗时约为2.467 s，体现了不同加密策略的效率差异；基于图像合成的方法常用弗雷歇初始距离(Fréchet Inception Distance, FID)[129]或SSIM[126]来衡量生成图像的视觉质量，并通过人脸检测率(Face Detection Rate, FDR)、表情识别率(Emotion Detection Rate, EDR)，以及姿态相似度(Head Pose Similarity, HPS)等指标验证隐私保护后的视觉可用性。例如，KFAAR方法[89]生成的受保护人脸图像的FID低至6.82，同时保持了较高的人脸检测率与表情识别率。对抗性方法通常侧重通过PSNR, SSIM[126]等指标评估保护前后图像的视觉一致性。例如，Salar等人[107]在实现较高保护成功率的同时，受保护图像的PSNR达到27.72 dB、SSIM约为0.84，距离视觉无差异的图像保护仍有一定距离。然而，现有可用性评估多局限于算法层面，难以反映资源受限、通信开销及用户体验等现实约束，有必要将系统性能与使用体验纳入统一评估框架，推动隐私保护技术走向实际应用。

5 问题与展望

人脸识别系统涵盖数据构建、采集、模型训练与推理等环节，隐私保护计算研究因应用阶段与数据对象不同而呈现多样化方法，各有优劣。但该领域仍面临诸多问题与挑战，主要体现在以下方面。

(1)隐私保护计算的效率提升：当前人脸识别隐私保护方案普遍面临计算开销大与资源受限的双重挑战。为增强数据安全性，引入的加密计算、特征脱敏等机制往往显著增加计算、存储和通信负担，影响系统实时性。受限于算力、内存、能耗及网络带宽，移动端和边缘设备难以支撑复杂保护策略，迫使现有方法在隐私强度、识别性能与资源效率之间权衡。因此，有必要探索更轻量高效的隐私保护框架，以提升实际可部署性。

(2)生成式大模型带来的机遇与挑战：随着生成式AI的迅猛发展，生成技术为人脸识别隐私保护带来了新的机遇与挑战。一方面，生成式模型可用于构建更先进的隐私保护机制，如生成更加逼真的虚拟人脸数据代替真实数据，用于保护模型训练或识别推理阶段的视觉隐私。另一方面，其强大的视觉推理与逆向重建能力也加剧了识别系统的隐私风险，对防御成员推理、人脸重建、深度伪造等攻击提出了更高要求。

(3)隐私友好型的人脸识别混合架构：传统人脸识别采用中心化存储与处理模式，易因攻击或配置失误导致大规模隐私泄露，且现有隐私保护多为事后补丁，难以兼顾识别性能与效率。未来应以“隐私优先”为核心，构建融合多种技术的混合架构。通过联邦学习等机制降低数据集中风险，结合多模态采集与端侧匿名化提升抗重建能力，并引入同态加密、安全多方计算和可信执行环境实现端到端保护，推动人脸识别由“识别为中心”向“隐私为前提”的可信识别范式转型。

(4)标准化的隐私评估体系：人脸识别隐私风险不仅来自图像或特征模板本身，还涉及重建攻击、模型逆向和成员推理等多种威胁，使隐私评估复杂且难以统一。现有评估多依赖PSNR, SSIM或特征相似度等经验性指标，分散且难以覆盖不同攻击场景，尤其缺乏对视觉隐私的人眼主观评估机制。尽管 国际标准ISO/IEC 24745[130]提出“不可逆性”和“不可关联性”原则，但缺乏量化标准，实践指导有限。因此，亟需建立多维、可验证的标准化评估体系，系统衡量视觉感知、身份泄露风险与抗攻击能力。

6 结束语

本文聚焦于可信人脸识别生态系统中的隐私保护计算方法，系统梳理并深入分析该领域的研究现状。首先，简要介绍人脸识别的基本原理与系统流

程，剖析其在实际应用中面临的主要隐私风险。随后，围绕不同风险类型综述当前主流的隐私保护技术，解析其核心机制，并从训练与推理两个阶段探讨相关方法在实际部署中的挑战与局限。最后，展望未来研究方向，包括隐私保护与系统效率的协同优化、生成式大模型带来的新机遇与挑战、新型识别范式的构建，以及标准化评估体系的建立，旨在为信息物理系统中安全可信的人脸身份识别提供系统性参考。

参 考 文 献

[1] TURK M A and PENTLAND A P. Face recognition using eigenfaces[C]. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, USA, 1991: 586–591. doi: 10.1109/CVPR.1991.139758.

[2] SCHROFF F, KALENICHENKO D, and PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]. The 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 815–823. doi: 10.1109/CVPR.2015.7298682.

[3] DENG Jiankang, GUO Jia, YANG Jing, *et al*. ArcFace: Additive angular margin loss for deep face recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(10): 5962–5979. doi: 10.1109/TPAMI.2021.3087709.

[4] KIM M, JAIN A K, and LIU Xiaoming. AdaFace: Quality adaptive margin for face recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 18729–18738. doi: 10.1109/CVPR52688.2022.01819.

[5] SUN Zhonglin, SONG Siyang, PATRAS I, *et al*. CemiFace: Center-based semi-hard synthetic face generation for face recognition[C]. The 37th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 35612–35638. doi: 10.52202/079017-1123.

[6] BOUTROS F, HUBER M, SIEBKE P, *et al*. SFace: Privacy-friendly and accurate face recognition using synthetic data[C]. 2022 IEEE International Joint Conference on Biometrics, Abu Dhabi, United Arab Emirates, 2022: 1–11. doi: 10.1109/IJCB54206.2022.10007961.

[7] QIU Haibo, YU Baosheng, GONG Dihong, *et al*. SynFace: Face recognition with synthetic data[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 10860–10870. doi: 10.1109/ICCV48922.2021.01070.

[8] KIM M, LIU Feng, JAIN A, *et al*. DCFace: Synthetic face generation with dual condition diffusion model[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 12715–12725. doi: 10.1109/CVPR52729.2023.01223.

[9] WANG Yinggui, LIU Jian, LUO Man, *et al*. Privacy-preserving face recognition in the frequency domain[C]. The 36th AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2022: 2558–2566. doi: 10.1609/aaai.v36i3.20157.

[10] MI Yuxi, HUANG Yuge, JI Jiazhen, *et al*. DuetFace: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain[C]. The 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 6755–6764. doi: 10.1145/3503161.3548303.

[11] SHOKRI R, STRONATI M, SONG Congzheng, *et al*. Membership inference attacks against machine learning models[C]. 2017 IEEE Symposium on Security and Privacy, San Jose, USA, 2017: 3–18. doi: 10.1109/sp.2017.41.

[12] SHAHREZA H O and MARCEL S. Unveiling synthetic faces: How synthetic datasets can expose real identities[C]. The Third Workshop on New Frontiers in Adversarial Machine Learning, Vancouver, Canada, 2024.

[13] ZHU Ligeng, LIU Zhijian, and HAN Song. Deep leakage from gradients[C]. The 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019: 1323.

[14] LIU Yufan, ZHANG Wanqian, WU Dayan, *et al*. Prediction exposes your face: Black-box model inversion via prediction alignment[C]. The 18th European Conference on Computer Vision, Milan, Italy, 2025: 288–306. doi: 10.1007/978-3-031-72764-1_17.

[15] ZHANG Hui, DONG Xingbo, LAI Y, *et al*. Validating privacy-preserving face recognition under a minimum assumption[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 12205–12214. doi: 10.1109/CVPR52733.2024.01160.

[16] SHAHREZA H O and MARCEL S. Template inversion attack against face recognition systems using 3D face reconstruction[C]. 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 19605–19615. doi: 10.1109/ICCV51070.2023.01801.

[17] SHAHREZA H O and MARCEL S. Face reconstruction from facial templates by learning latent space of a generator network[C]. The 37th International Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 557.

[18] BODDETI V N. Secure face matching using fully homomorphic encryption[C]. 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems, Redondo Beach, USA, 2018: 1–10. doi:

10.1109/BTAS.2018.8698601.

[19] ZHONG Zhizhou, MI Yuxi, HUANG Yuge, *et al*. SlerpFace: Face template protection via spherical linear interpolation[C]. The 39th AAAI Conference on Artificial Intelligence, Philadelphia, USA, 2025: 10698–10706. doi: 10.1609/aaai.v39i10.33162.

[20] WANG Zhibo, WANG He, JIN Shuaifan, *et al*. Privacy-preserving adversarial facial features[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 8212–8221. doi: 10.1109/CVPR52729.2023.00794.

[21] SHAHREZA H O, HAHN V K, and MARCEL S. MLP-hash: Protecting face templates via hashing of randomized multi-layer perceptron[C]. 2023 31st European Signal Processing Conference, Helsinki, Finland, 2023: 605–609. doi: 10.23919/EUSIPCO58844.2023.10289780.

[22] CHAMIKARA M A P, BERTOK P, KHALIL I, *et al*. Privacy preserving face recognition utilizing differential privacy[J]. *Computers & Security*, 2020, 97: 101951. doi: 10.1016/j.cose.2020.101951.

[23] TANG Long, YE Dengpan, LV Yunna, *et al*. Once and for all: Universal transferable adversarial perturbation against deep hashing-based facial image retrieval[C]. The 38th AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024: 5136–5144. doi: 10.1609/aaai.v38i6.28319.

[24] HU Shengshan, LIU Xiaogeng, ZHANG Yechao, *et al*. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer[C]. The 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 14994–15003. doi: 10.1109/CVPR52688.2022.01459.

[25] YANG Xiao, LIU Chang, XU Longlong, *et al*. Towards effective adversarial textured 3D meshes on physical face recognition[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 4119–4128. doi: 10.1109/CVPR52729.2023.00401.

[26] ERKIN Z, FRANZ M, GUAJARDO J, *et al*. Privacy-preserving face recognition[C]. The 9th International Symposium on Privacy Enhancing Technologies, Seattle, USA, 2009: 235–253. doi: 10.1007/978-3-642-03168-7_14.

[27] PAILLIER P. Public-key cryptosystems based on composite degree residuosity classes[M]. STERN J. Advances in Cryptology–EUROCRYPT '99. Berlin: Springer, 1999: 223–238. doi: 10.1007/3-540-48910-X_16.

[28] DAMGÅRD I, GEISLER M, and KRØIGAARD M. Efficient and secure comparison for on-line auctions[C]. 12th Australasian Conference on Information Security and Privacy, Townsville, Australia, 2007: 416–430. doi: 10.1007/978-3-540-73458-1_30.

[29] MA Zhuo, LIU Yang, LIU Ximeng, *et al*. Lightweight privacy-preserving ensemble classification for face recognition[J]. *IEEE Internet of Things Journal*, 2019, 6(3): 5778–5790. doi: 10.1109/JIOT.2019.2905555.

[30] ZHANG Pengbo and YANG Zhixin. A novel AdaBoost framework with robust threshold and structural optimization[J]. *IEEE Transactions on Cybernetics*, 2018, 48(1): 64–76. doi: 10.1109/TCYB.2016.2623900.

[31] OSADCHY M, PINKAS B, JARROUS A, *et al*. SCiFI - A system for secure face identification[C]. 2010 IEEE Symposium on Security and Privacy, Oakland, USA, 2010: 239–254. doi: 10.1109/SP.2010.39.

[32] TRONCOSO-PASTORIZA J R, GONZALEZ-JIMENEZ D, and PEREZ-GONZALEZ F. Fully private noninteractive face verification[J]. *IEEE Transactions on Information Forensics and Security*, 2013, 8(7): 1101–1114. doi: 10.1109/TIFS.2013.2262273.

[33] JIN Xin, LIU Yan, LI Xiaodong, *et al*. Privacy preserving face identification in the cloud through sparse representation[C]. The 10th Chinese Conference on Biometric Recognition, Tianjin, China, 2015: 160–167. doi: 10.1007/978-3-319-25417-3_20.

[34] IBARRONDO A, CHABANNE H, DESPIEGEL V, *et al*. Grote: Group testing for privacy-preserving face identification[C]. The Thirteenth ACM Conference on Data and Application Security and Privacy, Charlotte, USA, 2023: 117–128. doi: 10.1145/3577923.3583656.

[35] CHEON J H, KIM A, KIM M, *et al*. Homomorphic encryption for arithmetic of approximate numbers[C]. 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, 2017: 409–437. doi: 10.1007/978-3-319-70694-8_15.

[36] GUO Shangwei, XIANG Tao, and LI Xiaoguo. Towards efficient privacy-preserving face recognition in the cloud[J]. *Signal Processing*, 2019, 164: 320–328. doi: 10.1016/j.sigpro.2019.06.024.

[37] KOU Xiaoyu, ZHANG Ziling, ZHANG Yuelei, *et al*. Efficient and privacy-preserving distributed face recognition scheme via FaceNet[C]. The ACM Turing Award Celebration Conference, Hefei, China, 2021: 110–115. doi: 10.1145/3472634.3472661.

[38] GAO Wenjing, YU Jia, HAO Rong, *et al*. Privacy-preserving face recognition with multi-edge assistance for intelligent security systems[J]. *IEEE Internet of Things Journal*, 2023, 10(12): 10948–10958. doi: 10.1109/JIOT.2023.3240166.

[39] JI Jiazhen, WANG Huan, HUANG Yuge, *et al*. Privacy-preserving face recognition with learnable privacy budgets in frequency domain[C]. 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 475–491. doi:

10.1007/978-3-031-19775-8_28.

[40] MI Yuxi, HUANG Yuge, JI Jiazhen, *et al*. Privacy-preserving face recognition using random frequency components[C]. The IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 19616–19627. doi: 10.1109/ICCV51070.2023.01802.

[41] HENRY C, ASIF M S, and LI Zhu. Privacy preserving face recognition with lensless camera[C]. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 2023: 1–5. doi: 10.1109/ICASSP49357.2023.10096627.

[42] ASIF M S, AYREMLOU A, SANKARANARAYANAN A, *et al*. Flatcam: Thin, lensless cameras using coded aperture and computation[J]. *IEEE Transactions on Computational Imaging*, 2017, 3(3): 384–397. doi: 10.1109/TCI.2016.2593662.

[43] MI Yuxi, ZHONG Zhizhou, HUANG Yuge, *et al*. Privacy-preserving face recognition using trainable feature subtraction[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 297–307. doi: 10.1109/CVPR52733.2024.00036.

[44] PANDEY R K, ZHOU Yingbo, KOTA B U, *et al*. Deep secure encoding for face template protection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, USA, 2016: 77–83. doi: 10.1109/CVPRW.2016.17.

[45] TALREJA V, VALENTI M C, and NASRABADI N M. Zero-shot deep hashing and neural network based error correction for face template protection[C]. 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems, Tampa, USA, 2019: 1–10. doi: 10.1109/BTAS46853.2019.9185979.

[46] 赵铖辉, 李勇, 张振江. BinaryFace: 基于深层卷积神经网络的人脸模板保护模型[J]. 信息安全学报, 2020, 5(5): 43–55. doi: 10.19363/J.cnki.cn10-1380/tn.2020.09.04.
ZHAO Chenghui, LI Yong, and ZHANG Zhenjiang. BinaryFace: The model of face template protection based on CNN[J]. *Journal of Cyber Security*, 2020, 5(5): 43–55. doi: 10.19363/J.cnki.cn10-1380/tn.2020.09.04.

[47] ZHOU Junwei, SHANG Delong, LANG Huile, *et al*. Face template protection through residual learning based error-correcting codes[C]. The 4th International Conference on Control and Computer Vision, Macau, China, 2021: 112–118. doi: 10.1145/3484274.3484292.

[48] MAI Guangcan, CAO Kai, LAN Xiangyuan, *et al*. SecureFace: Face template protection[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 262–277. doi: 10.1109/TIFS.2020.3009590.

[49] GAO Ce, ZHANG Kang, WANG Weiwei, *et al*. Protected face templates generation based on multiple partial Walsh transformations and Simhash[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 4100–4113. doi: 10.1109/TIFS.2024.3369322.

[50] SWICK D. Walsh function generation (Corresp.)[J]. *IEEE Transactions on Information Theory*, 1969, 15(1): 167–167. doi: 10.1109/TIT.1969.1054251.

[51] CHARIKAR M S. Similarity estimation techniques from rounding algorithms[C]. The Thiry-Fourth Annual ACM Symposium on Theory of Computing, Montreal, Canada, 2002: 380–388. doi: 10.1145/509907.509965.

[52] 李亚红, 李一婧, 杨小东, 等. 基于同态加密和群签名的可验证联邦学习方案[J]. 电子与信息学报, 2025, 47(3): 758–768. doi: 10.11999/JEIT240796.
LI Yahong, LI Yijing, YANG Xiaodong, *et al*. A verifiable federated learning scheme based on homomorphic encryption and group signature[J]. *Journal of Electronics & Information Technology*, 2025, 47(3): 758–768. doi: 10.11999/JEIT240796.

[53] 郭显, 王典冬, 冯涛, 等. 基于同态加密的可验证隐私保护联邦学习方案[J]. 电子与信息学报, 2025, 47(4): 1113–1125. doi: 10.11999/JEIT240390.
GUO Xian, WANG Diandong, FENG Tao, *et al*. A verifiable privacy protection federated learning scheme based on homomorphic encryption[J]. *Journal of Electronics & Information Technology*, 2025, 47(4): 1113–1125. doi: 10.11999/JEIT240390.

[54] AGGARWAL D, ZHOU Jiayu, and JAIN A K. FedFace: Collaborative learning of face recognition model[C]. 2021 IEEE International Joint Conference on Biometrics, Shenzhen, China, 2021: 1–8. doi: 10.1109/IJCB52358.2021.9484386.

[55] MARTÍNEZ BELTRÁN E T, PERALES GÓMEZ Á L, FENG Chao, *et al*. Fedstellar: A platform for decentralized federated learning[J]. *Expert Systems with Applications*, 2024, 242: 122861. doi: 10.1016/j.eswa.2023.122861.

[56] GAO Liang, LI Li, CHEN Yingwen, *et al*. FIFL: A fair incentive mechanism for federated learning[C]. The 50th International Conference on Parallel Processing, Lemont, USA, 2021: 82. doi: 10.1145/3472456.3472469.

[57] ZHENG Haipeng, LI Bing, LIU Guozhu, *et al*. Blockchain-based federated learning framework applied in face recognition[C]. 2022 7th International Conference on Signal and Image Processing, Suzhou, China, 2022: 265–269. doi: 10.1109/ICSIP55141.2022.9886171.

[58] NIU Yifan and DENG Weihong. Federated learning for face recognition with gradient correction[C]. The 36th AAAI Conference on Artificial Intelligence, 2022: 1999–2007. doi: 10.1609/aaai.v36i2.20095.

[59] LIU C T, WANG C Y, CHIEN S Y, *et al*. FedFR: Joint optimization federated framework for generic and

personalized face recognition[C]. The 36th AAAI Conference on Artificial Intelligence, 2022: 1656–1664. doi: 10.1609/aaai.v36i2.20057.

[60] WOUBIE A, SOLOMON E, and ATTIEH J. Maintaining privacy in face recognition using federated learning method[J]. *IEEE Access*, 2024, 12: 39603–39613. doi: 10.1109/ACCESS.2024.3373691.

[61] KIM J, PARK T, KIM H, *et al*. Federated learning for face recognition[C]. 2021 IEEE International Conference on Consumer Electronics, Las Vegas, USA, 2021: 1–2. doi: 10.1109/ICCE50685.2021.9427748.

[62] DENG Yu, YANG Jiaolong, CHEN Dong, *et al*. Disentangled and controllable face image generation via 3D imitative-contrastive learning[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 5153–5162. doi: 10.1109/CVPR42600.2020.00520.

[63] LI Yuancheng, WANG Yimeng, and LI Daoxing. Privacy-preserving lightweight face recognition[J]. *Neurocomputing*, 2019, 363: 212–222. doi: 10.1016/j.neucom.2019.07.039.

[64] SAATCHI Y and WILSON A G. Bayesian GAN[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017. 2017: 3625–3634.

[65] KARRAS T, AITTALA M, HELLSTEN J, *et al*. Training generative adversarial networks with limited data[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 1015.

[66] BAE G, DE LA GORCE M, BALTRUŠAITIS T, *et al*. DigiFace-1M: 1 million digital face images for face recognition[C]. The IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2023: 3515–3524. doi: 10.1109/WACV56688.2023.00352.

[67] BLANZ V and VETTER T. A morphable model for the synthesis of 3D faces[C]. The 26th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, USA, 1999: 187–194. doi: 10.1145/311535.311556.

[68] WOOD E, BALTRUŠAITIS T, HEWITT C, *et al*. Fake it till you make it: Face analysis in the wild using synthetic data alone[C]. The IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 3661–3671. doi: 10.1109/ICCV48922.2021.00366.

[69] KOLF J N, RIEBER T, ELLIESEN J, *et al*. Identity-driven three-player generative adversarial network for synthetic-based face recognition[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Vancouver, Canada, 2023: 806–816. doi: 10.1109/CVPRW59228.2023.00088.

[70] BOUTROS F, GREBE J H, KUIJPER A, *et al*. IDiff-Face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model[C]. The IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 19593–19604. doi: 10.1109/ICCV51070.2023.01800.

[71] MELZI P, RATHGEB C, TOLOSANA R, *et al*. GANDiffFace: Controllable generation of synthetic datasets for face recognition with realistic variations[C]. The 2023 IEEE/CVF International Conference on Computer Vision Workshops, Paris, France, 2023: 3078–3087. doi: 10.1109/ICCVW60793.2023.00333.

[72] XU Jianqing, LI Shen, WU Jiaying, *et al*. ID^3: Identity-preserving-yet-diversified diffusion models for synthetic face recognition[C]. The 38th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 77777–77798.

[73] SHAHREZA H O and MARCEL S. HyperFace: Generating synthetic face recognition datasets by exploring face embedding hypersphere[C]. The Thirteenth International Conference on Learning Representations, Singapore, Singapore, 2025: 478–491.

[74] MI Yuxi, ZHONG Zhizhou, HUANG Yuge, *et al*. Data synthesis with diverse styles for face recognition via 3DMM-guided diffusion[C]. The 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2025: 21203–21214. doi: 10.1109/CVPR52734.2025.01975.

[75] 彭春蕾, 苗紫民, 刘德成, 等. 视觉身份隐私保护: 人脸匿名化研究方法[J]. 计算机学报, 2023, 46(11): 2431–2452. doi: 10.11897/SP.J.1016.2023.02431.
PENG Chunlei, MIAO Zimin, LIU Decheng, *et al*. Visual identity privacy protection: Research methods of face anonymization[J]. *Chinese Journal of Computers*, 2023, 46(11): 2431–2452. doi: 10.11897/SP.J.1016.2023.02431.

[76] LI Jingzhi, HAN Lutong, ZHANG Hua, *et al*. Learning disentangled representations for identity preserving surveillance face camouflage[C]. 2020 25th International Conference on Pattern Recognition, Milan, Italy, 2021: 9748–9755. doi: 10.1109/ICPR48806.2021.9412636.

[77] LI Jingzhi, HAN Lutong, CHEN Ruoyu, *et al*. Identity-preserving face anonymization via adaptively facial attributes obfuscation[C]. The 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 3891–3899. doi: 10.1145/3474085.3475367.

[78] LI Jingzhi, ZHANG Hua, LIANG Siyuan, *et al*. Privacy-enhancing face obfuscation guided by semantic-aware attribution maps[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 3632–3646. doi: 10.1109/TIFS.2023.3282384.

[79] YUAN Lin, LIU Linguo, PU Xiao, *et al*. PRO-face: A generic framework for privacy-preserving recognizable

obfuscation of face images[C]. The 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 1661–1669. doi: 10.1145/3503161.3548202.

[80] YUAN Lin, CHEN Wu, PU Xiao, *et al*. PRO-face C: Privacy-preserving recognition of obfuscated face via feature compensation[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 4930–4944. doi: 10.1109/TIFS.2024.3388976.

[81] SU Zhigang, ZHOU Dawei, WANG Nannan, *et al*. Hiding visual information via obfuscating adversarial perturbations[C]. The IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 4333–4343. doi: 10.1109/ICCV51070.2023.00402.

[82] WANG Tao, ZHANG Yushu, YANG Zixuan, *et al*. Seeing is not believing: An identity hider for human vision privacy protection[J]. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025, 7(2): 170–181. doi: 10.1109/TBIOM.2024.3449849.

[83] 任坤, 李峥瑱, 桂源泽, 等. 低分辨率随机遮挡人脸图像的超分辨率修复[J]. 电子与信息学报, 2024, 46(8): 3343–3352. doi: 10.11999/JEIT231262.
REN Kun, LI Zhengzhen, GUI Yuanze, *et al*. Super-resolution inpainting of low-resolution randomly occluded face images[J]. *Journal of Electronics & Information Technology*, 2024, 46(8): 3343–3352. doi: 10.11999/JEIT231262.

[84] YUAN Zhuowen, YOU Zhengxin, LI Sheng, *et al*. On generating identifiable virtual faces[C]. The 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 1465–1473. doi: 10.1145/3503161.3548110.

[85] KARRAS T, LAINE S, and AILA T. A style-based generator architecture for generative adversarial networks[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4396–4405. doi: 10.1109/CVPR.2019.00453.

[86] WANG Tao, ZHANG Yushu, ZHAO Ruoyu, *et al*. Identifiable face privacy protection via virtual identity transformation[J]. *IEEE Signal Processing Letters*, 2023, 30: 773–777. doi: 10.1109/LSP.2023.3289392.

[87] WANG Tao, ZHANG Yushu, XIAO Xiangli, *et al*. Make privacy renewable! Generating privacy-preserving faces supporting cancelable biometric recognition[C]. The 32nd ACM International Conference on Multimedia, Melbourne, Australia, 2024: 10268–10276. doi: 10.1145/3664647.3680704.

[88] WANG Miaomiao, HUA Guang, LI Sheng, *et al*. A key-driven framework for identity-preserving face anonymization[C]. 32nd Annual Network and Distributed System Security Symposium, San Diego, USA, 2025: 10168–10176. doi: 10.14722/ndss.2025.23729.

[89] DWORK C and LEI Jing. Differential privacy and robust statistics[C]. The Forty-First Annual ACM Symposium on Theory of Computing, Bethesda, USA, 2009: 371–380. doi: 10.1145/1536414.1536466.

[90] MAHAWAGA ARACHCHIGE P C, BERTOK P, KHALIL I, *et al*. Local differential privacy for deep learning[J]. *IEEE Internet of Things Journal*, 2020, 7(7): 5827–5842. doi: 10.1109/JIOT.2019.2952146.

[91] MAO Yunlong, YI Shanhe, LI Qun, *et al*. A privacy-preserving deep learning approach for face recognition with edge computing[C]. The USENIX Workshop on Hot Topics in Edge Computing, Boston, USA, 2018: 1–6. doi: 10.5555/3342665.3342676.

[92] MENG Qiang, ZHOU Feng, REN Hainan, *et al*. Improving federated learning face recognition via privacy-agnostic clusters[C]. The Tenth International Conference on Learning Representations, 2022: 237–245.

[93] WEN Yunqian, LIU Bo, DING Ming, *et al*. IdentityDP: Differential private identification protection for face images[J]. *Neurocomputing*, 2022, 501: 197–211. doi: 10.1016/j.neucom.2022.06.039.

[94] 张啸剑, 付聪聪, 孟小峰. 结合矩阵分解与差分隐私的人脸图像发布[J]. 中国图象图形学报, 2020, 25(4): 655–668. doi: 10.11834/jig.190308.
ZHANG Xiaojian, FU Congcong, and MENG Xiaofeng. Private facial image publication through matrix decomposition[J]. *Journal of Image and Graphics*, 2020, 25(4): 655–668. doi: 10.11834/jig.190308.

[95] XIAO Taihong, TSAI Y H, SOHN K, *et al*. Adversarial learning of privacy-preserving and task-oriented representations[C]. The 34th AAAI Conference on Artificial Intelligence, New York, USA, 2020: 12434–12441. doi: 10.1609/aaai.v34i07.6930.

[96] LI Ang, GUO Jiayi, YANG Huanrui, *et al*. DeepObfuscator: Obfuscating intermediate representations with privacy-preserving adversarial learning on smartphones[C]. The International Conference on Internet-of-Things Design and Implementation, Charlottesvle, USA, 2021: 28–39. doi: 10.1145/3450268.3453519.

[97] JIN Shuaifan, WANG He, WANG Zhibo, *et al*. FaceObfuscator: Defending deep learning-based privacy attacks with gradient descent-resistant features in face recognition[C]. The 33rd USENIX Conference on Security Symposium, Philadelphia, USA, 2024: 383.

[98] WEI Chenda, WANG Haoyue, QIAN Zhenxing, *et al*. Learning discrepant transformations for face privacy protection[C]. The 33rd ACM International Conference on Multimedia, Dublin, Ireland, 2025: 7672–7680. doi: 10.1145/3746027.3754881.

[99] SHAN S, WENGER Emily, ZHANG Jiayun, *et al*. Fawkes:

Protecting privacy against unauthorized deep learning models[C]. The 29th USENIX Conference on Security Symposium, Santa Clara, USA, 2020: 90.

[100] YANG Xiao, DONG Yinpeng, PANG Tianyu, *et al*. Towards face encryption by generating adversarial identity masks[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 3877–3887. doi: 10.1109/ICCV48922.2021.00387.

[101] ZHONG Yaoyao and DENG Weihong. OPOM: Customized invisible cloak towards face privacy protection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3590–3603. doi: 10.1109/TPAMI.2022.3175602.

[102] LIU Xuannan, ZHONG Yaoyao, CUI Xing, *et al*. AdvCloak: Customized adversarial cloak for privacy protection[J]. *Pattern Recognition*, 2025, 158: 111050. doi: 10.1016/j.patcog.2024.111050.

[103] DONG Xin, WANG Rui, LIANG Siyuan, *et al*. Face encryption via frequency-restricted identity-agnostic attacks[C]. The 31st ACM International Conference on Multimedia, Ottawa, Canada, 2023: 579–588. doi: 10.1145/3581783.3612233.

[104] YIN Bangjie, WANG Wenxuan, YAO Taiping, *et al*. Adv-makeup: A new imperceptible and transferable attack on face recognition[C]. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Montreal, Canada, 2021: 1252–1258. doi: 10.24963/ijcai.2021/173.

[105] SHAMSHAD F, NASEER M, and NANDAKUMAR K. CLIP2Protect: Protecting facial privacy using text-guided makeup via adversarial latent search[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 20595–20605. doi: 10.1109/CVPR52729.2023.01973.

[106] SUN Yuhao, YU Lingyun, XIE Hongtao, *et al*. DiffAM: Diffusion-based adversarial makeup transfer for facial privacy protection[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 24584–24594. doi: 10.1109/CVPR52733.2024.02321.

[107] SALAR A, LIU Qing, TIAN Yingli, *et al*. Enhancing facial privacy protection via weakening diffusion purification[C]. The 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2025: 8235–8244. doi: 10.1109/CVPR52734.2025.00771.

[108] 孙军梅, 潘振雄, 李秀梅, 等. 面向人脸验证的可迁移对抗样本生成方法[J]. 电子与信息学报, 2023, 45(5): 1842–1851. doi: 10.11999/JEIT220358.

SUN Junmei, PAN Zhenxiong, LI Xiumei, *et al*. Transferable adversarial example generation method for face verification[J]. *Journal of Electronics & Information Technology*, 2023, 45(5): 1842–1851. doi: 10.11999/JEIT220358.

[109] ZHANG Yushu, WANG Tao, ZHAO Ruoyu, *et al*. RAPP: Reversible privacy preservation for various face attributes[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 3074–3087. doi: 10.1109/TIFS.2023.3274359.

[110] MIRJALILI V, RASCHKA S, and ROSS A. PrivacyNet: Semi-adversarial networks for multi-attribute face privacy[J]. *IEEE Transactions on Image Processing*, 2020, 29: 9400–9412. doi: 10.1109/TIP.2020.3024026.

[111] MORALES A, FIERREZ J, VERA-RODRIGUEZ R, *et al*. SensitiveNets: Learning agnostic representations with application to face images[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(6): 2158–2164. doi: 10.1109/TPAMI.2020.3015420.

[112] HUANG G B, MATTAR M, BERG T, *et al*. Labeled faces in the wild: A database for studying face recognition in unconstrained environments[C]. The Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Marseille, France, 2008.

[113] LIU Ziwei, LUO Ping, WANG Xiaogang, *et al*. Deep learning face attributes in the wild[C]. The IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 3730–3738. doi: 10.1109/ICCV.2015.425.

[114] CAO Qiong, SHEN Li, XIE Weidi, *et al*. VGGFace2: A dataset for recognising faces across pose and age[C]. The 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, China, 2018: 67–74. doi: 10.1109/FG.2018.00020.

[115] YI Dong, LEI Zhen, LIAO Shengcai, *et al*. Learning face representation from scratch[EB/OL]. https://arxiv.org/abs/1411.7923, 2014.

[116] SENGUPTA S, CHEN Juncheng, CASTILLO C, *et al*. Frontal to profile face verification in the wild[C]. 2016 IEEE Winter Conference on Applications of Computer Vision, Lake Placid, USA, 2016: 1–9. doi: 10.1109/WACV.2016.7477558.

[117] MOSCHOGLOU S, PAPAIOANNOU A, SAGONAS C, *et al*. AgeDB: The first manually collected, in-the-wild age database[C]. The IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 1997–2005. doi: 10.1109/CVPRW.2017.250.

[118] ZHENG Tianyue and DENG Weihong. Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments[J]. *Beijing University of Posts and Telecommunications, Tech. Rep*, 2018, 5(7): 1-6.

[119] ZHENG Tianyue, DENG Weihong, and HU Jiani. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments[EB/OL].

https://arxiv.org/abs/1708.08197, 2017.

[120] THOMAZ C E and GIRALDI G A. A new ranking method for principal components analysis and its application to face image analysis[J]. *Image and Vision Computing*, 2010, 28(6): 902–913. doi: 10.1016/j.imavis.2009.11.005.

[121] WHITELAM C, TABORSKY E, BLANTON A, *et al*. IARPA JANUS benchmark-B face dataset[C]. The IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 592–600. doi: 10.1109/CVPRW.2017.87.

[122] MAZE B, ADAMS J, DUNCAN J A, *et al*. IARPA Janus Benchmark-C: Face dataset and protocol[C]. 2018 International Conference on Biometrics (ICB), Gold Coast, Australia, 2018: 158–165. doi: 10.1109/ICB2018.2018.00033.

[123] LEE K C, HO J, and KRIEGMAN D J. Acquiring linear subspaces for face recognition under variable lighting[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(5): 684–698. doi: 10.1109/TPAMI.2005.92.

[124] SAKTHIVEL S, LAKSHMIPATHI R, and ARUMUGAM M A M. Evaluation of feature extraction and dimensionality reduction algorithms for face recognition using ORL database[C]. The 2009 International Conference on Image Processing, Computer Vision, & Pattern Recognition, Las Vegas, USA, 2009: 367–373. doi: 10.1109/IPCV.2009.50.

[125] MCCOOL C, WALLACE R, MCLAREN M, *et al*. Session variability modelling for face authentication[J]. *IET Biometrics*, 2013, 2(3): 117–129. doi: 10.1049/iet-bmt.2012.0059.

[126] WANG Zhou, BOVIK A C, SHEIKH H R, *et al*. Image quality assessment: From error visibility to structural similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600–612. doi: 10.1109/TIP.2003.819861.

[127] ZHANG R, ISOLA P, EFROS A A, *et al*. The unreasonable effectiveness of deep features as a perceptual metric[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 586–595. doi: 10.1109/CVPR.2018.00068.

[128] GOMEZ-BARRERO M, GALBALLY J, RATHGEB C, *et al*. General framework to evaluate unlinkability in biometric template protection systems[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(6): 1406–1420. doi: 10.1109/TIFS.2017.2788000.

[129] HEUSEL M, RAMSAUER H, UNTERTHINER T, *et al*. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6629–6640.

[130] ISO. ISO/IEC 24745: 2022 Information security, cybersecurity and privacy protection — biometric information protection[S]. Geneva: ISO, 2022.

袁　霖：男，副教授，研究方向为人工智能安全与多媒体安全.
武雁尚：男，硕士生，研究方向为图像隐私保护.
张力元：男，硕士生，研究方向为图像隐私保护.
张玉书：男，教授，研究方向为多媒体隐私与安全、可信人工智能等.
王楠楠：男，教授，研究方向为计算机视觉、统计机器学习等.
高新波：男，教授，研究方向为机器学习、计算机视觉、图像分析等.

责任编辑：余　蓉

Privacy-preserving Computation in Trustworthy Face Recognition: A Comprehensive Survey

YUAN Lin[①]　WU Yanshang[①]　ZHANG Liyuan[①]　ZHANG Yushu[②]
WANG Nannan[③]　GAO Xinbo[①③]

[①](*Chongqing Key Laboratory of Image Cognition*, *Chongqing University of Posts and Telecommunications*, *Nan'an District*, *Chongqing* 400065, *China*)

[②](*School of Computing and Artificial Intelligence*, *Jiangxi University of Finance and Economics*, *Nanchang*, *Jiangxi* 330032, *China*)

[③](*State Key Laboratory of Integrated Services Networks*, *Xidian University*, *Xi'an*, *Shaanxi* 710071, *China*)

Abstract:

Significance　With the widespread deployment of face recognition in Cyber-Physical Systems (CPS), including smart cities, intelligent transportation, and public safety infrastructures, privacy leakage has become a central concern for both academia and industry. Unlike many biometric modalities, face recognition operates in highly visible and loosely controlled environments, such as public spaces, consumer devices, and online platforms,

where facial image acquisition is easy and pervasive. This exposure makes facial data especially vulnerable to unauthorized collection and misuse. Insufficient protection may lead to identity theft, unauthorized tracking, and deepfake generation, which threaten individual rights and reduce trust in digital systems. Therefore, facial data protection is not only a technical issue but also a significant societal and ethical challenge. This work integrates fragmented research across computer vision, cryptography, and privacy-preserving computation. It provides a unified perspective that guides the development of trustworthy face recognition ecosystems that balance usability, regulatory compliance, and public trust.

Contributions This paper systematically reviews recent advances in privacy-preserving computation for face recognition, covering both theoretical foundations and practical implementations. The architecture and application pipeline of face recognition systems are first examined, and privacy risks at each stage are identified. At the data collection stage, unauthorized or covert capture of facial images introduces immediate risks of misuse. During model training and deployment, gradient leakage, membership inference, and overfitting may expose sensitive information about individuals contained in training data. At the inference stage, adversaries may reconstruct facial images, perform unauthorized recognition, or associate identities across datasets, which compromises anonymity. To address these threats, existing approaches are classified into four major privacy-preserving paradigms: data transformation, distributed collaboration, image generation, and adversarial perturbation. Within these paradigms, ten representative techniques are analyzed. Cryptographic computation, including homomorphic encryption and secure multiparty computation, enables recognition without revealing raw data but often introduces substantial computational overhead. Frequency-domain learning converts images into spectral representations to suppress identifiable details while retaining discriminative features. Federated learning decentralizes model training and reduces centralized data exposure, although it remains vulnerable to gradient inversion attacks. Image generation techniques, such as face synthesis and virtual identity modeling, reduce reliance on real facial data during training and evaluation. Differential privacy introduces calibrated noise to provide statistical privacy guarantees, whereas face anonymization obscures identifiable visual traits. Template protection and anti-reconstruction mechanisms defend stored facial features against reverse engineering. Adversarial privacy protection introduces imperceptible perturbations that interfere with machine recognition yet preserve human visual perception. Several representative studies in each category are further examined. Commonly used evaluation datasets are summarized. A comparative analysis is conducted across multiple dimensions, including face recognition performance, privacy protection effectiveness, and practical usability. This analysis systematically identifies the strengths and limitations of different types of methods.

Prospects Several research directions are identified for future work. A primary challenge is to achieve a dynamic balance between privacy protection and system utility. Excessive protection may degrade recognition accuracy, whereas insufficient safeguards expose users to unacceptable risks. Adaptive mechanisms that adjust privacy levels according to context, task requirements, and user consent are therefore required. Another promising direction is the development of inherently privacy-aware recognition paradigms, such as feature representations that minimize identity leakage by design. The establishment of standardized evaluation frameworks for privacy risk and usability is also essential. Such frameworks would enable reproducible benchmarking and facilitate real-world deployment. The emergence of generative foundation models, including diffusion models and large multimodal models, further changes the research landscape. These models enable synthetic data generation and controllable identity representations. However, they also enable more advanced attacks, such as high-fidelity face reconstruction and identity impersonation. Addressing these dual effects requires interdisciplinary collaboration across computer vision, cryptography, law, and ethics, supported by appropriate regulation and continued methodological development.

Conclusions This paper provides a comprehensive reference for researchers and practitioners engaged in trustworthy face recognition. By integrating advances from multiple disciplines, it promotes the development of effective facial privacy protection technologies and supports the secure, reliable, and ethically responsible deployment of face recognition in practical scenarios. The long-term goal is to establish face recognition as a trustworthy component of CPS that balances functionality, privacy protection, and societal trust.

Key words: Privacy-preserving computation; Cyber-Physical Systems(CPS); Face recognition; Identity information