深度网络生成式伪造人脸检测方法研究综述

杨睿[1,2,3], 胡心如[1], 黄卓超[1], 张玉书[4], 蓝如师[1,2,3]*, 邓珍荣[2], 罗笑南[1,3]

[1] (桂林电子科技大学广西图像图形与智能处理重点实验室 桂林 541004)
[2] (桂林电子科技大学计算机与信息安全学院 桂林 541004)
[3] (桂林电子科技大学南宁研究院 南宁 530033)
[4] (南京航空航天大学计算机科学与技术学院 南京 210016)
(rslan@guet.edu.cn)

摘 要: 随着深度网络生成式伪造人脸技术的迅速传播, 不法分子通过伪造人脸图像和视频实施电信诈骗等犯罪活动, 如何从海量数据中高效、准确地检测出伪造人脸成为研究焦点. 文中从深度网络生成式伪造人脸图像和生成式伪造人脸视频 2 个角度出发, 系统归纳、分析、比较了当前伪造人脸检测方法. 针对伪造人脸图像, 从基于数字图像处理基础、深层次特征提取、空间域特征分析、多特征融合分析和指纹检测 5 个类别详细介绍了检测方法; 并从生理信号、身份信息、多模态和时空不一致 4 个类别对伪造人脸视频的检测方法进行了探讨. 分析表明, 目前深度网络生成式伪造人脸检测方法的泛化能力有待提高, 在未来的研究中, 应当着重提升模型的跨数据集泛化能力、准确性和实用性, 从而更好地防范虚假信息传播, 以保护个人隐私和维护网络安全环境.

关键词: 伪造人脸检测; 生成式伪造人脸; 人脸图像; 人脸视频; 深度网络
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.2023-00615

Review of Deep Network Generative Fake Face Detection Methods

Yang Rui[1,2,3], Hu Xinru[1], Huang Zhuochao[1], Zhang Yushu[4], Lan Rushi[1,2,3]*, Deng Zhenrong[2], and Luo Xiaonan[1,3]

[1] (*Guangxi Key Laboratory of Images and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin* 541004)
[2] (*School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin* 541004)
[3] (*Nanning Research Institute, Guilin University of Electronic Technology, Nanning* 530033)
[4] (*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016)

Abstract: With the rapid spread of deep network generated fake face technology, criminals perpetrate telecom fraud, manipulate public opinion, and disseminate obscenity by forging face images and videos. How to efficiently and accurately detect fake faces from massive data has become a research focus. In this review, we systematically summarize, analyze and compare the current deep network generative forgery face detection methods from two fields: generative forgery face image and generative forgery face video. For the forged face images, the detection methods are introduced in detail from five categories: digital image processing foundation, deep feature extraction, spatial domain feature analysis, multi-feature fusion analysis and fingerprint detection.

收稿日期: 2023-09-17; 修回日期: 2024-03-08. **基金项目**: 广西自然科学基金(2019GXNSFFA245014, AD20159034); 广西科技计划(AB20238013, AB22035052); 国家自然科学基金(62172120, 62002082, 6202780103); 广西图像图形与智能处理重点实验室项目(GIIP2209, GIIP2211, GIIP2003); 桂林电子科技大学研究生教育创新计划(2023YCXB09). **杨睿**(1996—), 女, 博士研究生, CCF 学生会员, 主要研究方向为伪造人脸检测、目标检测、图像描述; **胡心如**(1999—), 女, 硕士研究生, 主要研究方向为伪造人脸视频检测; **黄卓超**(1999—), 男, 硕士研究生, 主要研究方向为伪造人脸图像检测; **张玉书**(1987—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为区块链、多媒体信息安全、人工智能; **蓝如师**(1986—), 男, 博士, 研究员, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为图像处理、超分辨率重建; **邓珍荣**(1977—), 女, 硕士, 研究员, 硕士生导师, 主要研究方向为深度学习、目标检测; **罗笑南**(1963—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机图形学、数字图像处理.

The detection methods of fake face videos are also discussed from four categories: physiological signals, identity information, multi-modal and spatio-temporal inconsistency. The analysis shows that the generalization ability of the current deep network generative fake face detection method needs to be improved. In future research, we should focus on improving the cross-dataset generalization ability, accuracy and practicality of the model, so as to better prevent the spread of false information, protect personal privacy and maintain network security environment.

Key words: fake face detection; generative fake face; face image; face video; deep network

随着神经网络和深度学习技术的蓬勃发展，深度网络模型已成为训练大规模图像和视频数据集以生成高度逼真虚拟人脸的主要工具. 深度伪造技术借助深度学习技术进行人脸合成，最初的动力源于为用户提供娱乐、辅助视觉特效、扩充数据集等目的. 然而近年来，深度伪造技术被不法分子滥用，进而产生电信诈骗、色情传播、肖像权侵犯以及舆论操控等诸多问题. 不仅给个人、组织和国家带来经济损失，还破坏了社会信任、社会稳定和法律秩序.

在电信诈骗方面，犯罪分子利用深度学习技术，通过恶意剪辑、换脸变声等手段伪造视频内容诱骗受害人上当. 在色情传播方面，不法分子利用AI(artificial intelligence)生成真实和动画风格的裸体人物图像，制作虚假的换脸淫秽视频并在网络社交平台上传播以谋取利益. 在肖像权侵犯方面，商家借助 AI 换脸技术生成虚拟的明星代言人，吸引消费者购买和打赏. 此外，通过对目标人物进行换脸，还可以生成虚假的发言视频，以影响社会舆论乃至选举结果. 因此，研究深度网络生成式伪造人脸检测模型具有重要的意义.

(1) 保护信息的真实性. 信息真实性是现代社会的基石，随着社交媒体和新闻传媒的普及，信息传播速度更快、范围更广. 然而，虚假信息可能导致误导、恐慌或对社会产生负面影响. 伪造人脸技术可以轻松地制作虚假人脸图像和视频，将其植入新闻报道、社交媒体帖子或在线视频中. 因此，研究伪造人脸检测模型有助于确保信息的真实性，这些模型能够探测潜在的伪造内容，降低虚假信息传播的风险，维护信息的可信度，保护公众免受虚假信息误导.

(2) 保护个人隐私. 个人隐私是数字时代的一项重要权利，伪造人脸技术可能被用于模拟个人面部特征，制造虚假身份，从而引发身份盗窃、网络欺诈等隐私侵犯行为. 通过研究伪造人脸检测模型，可以更有效地保护个人隐私，及时检测和阻止虚假身份的制造和滥用，尤其在金融领域可用于辨识试图冒充他人进行交易的个体，提高金融交易的安全性，降低电信诈骗风险.

(3) 增强网络安全. 随着伪造人脸技术的不断进步，攻击者可以更轻松地伪造虚假身份，从而混入网络并进行各种恶意活动，这给网络安全带来了严重的威胁. 因此，研究伪造人脸检测模型对于增强网络安全至关重要，这些模型能够帮助检测和识别潜在的网络威胁，如身份欺诈和网络犯罪. 在企业和政府机构中，还可以有效地防止未经授权的访问，提升网络的整体安全性.

针对上述问题，本文聚焦深度网络人脸生成技术及其在图像处理领域中的应用，主要阐述了 2 种伪造人脸生成模型生成对抗网络(generative adversarial networks, GAN)和扩散模型(denoising diffusion probabilistic model, DDPM)的发展现状以及 4 种常见伪造人脸检测模型，并通过对深度网络生成的虚假人脸的防范和深度伪造技术的检测技术分析，详细介绍了深度网络生成式伪造人脸图像和伪造人脸视频的检测方法.

1 深度网络生成式伪造人脸的检测方法

图 1 所示的基于深度网络的生成式伪造人脸检测方法可分为伪造人脸图像检测和伪造人脸视频检测两大类. 在伪造人脸图像检测方面，文中详细介绍了传统的深度伪造图像检测方法，并探讨了基于深层次特征的方法. 还介绍了基于空间域特征的方法，探讨了基于多特征融合的方法，从面部细节、角膜高光等方面提取特征进行伪造检测. 另外，本文还介绍了基于对抗网络生成图像指纹的方法，通过学习生成伪影或指纹实现精确的伪造人脸检测. 这些方法在不同场景和数据条件下具有各自的优势和局限性，对伪造人脸图像检测的研究提供参考.

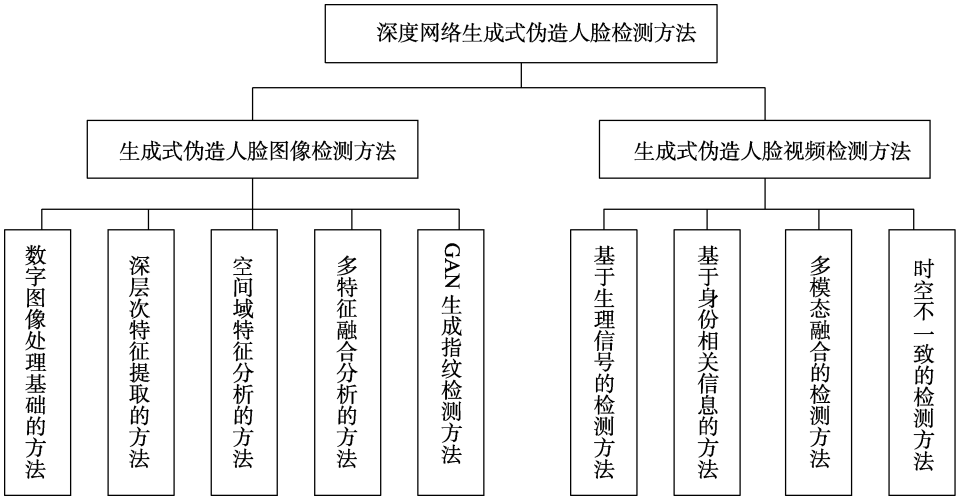

图 1　深度网络生式伪造人脸检测方法框架

下面对生成式伪造人脸视频检测方法进行了全面介绍. 这些方法涵盖了基于生理信号的方法, 利用光电容积描记法(photo plethysmo graphic, PPG)检测生物信号特征, 以及基于身份相关信息、多模态检测和基于视频帧之间的不一致性等方面的方法. 其中, 通过不一致性线索构建网络检测伪造人脸的方法特别值得关注. 随着技术的发展, 伪造人脸视频检测方法趋向于设计更小、更快的伪造人脸检测模型, 使得实现实时检测成为可能, 并减少伪造人脸带来的负面影响. 这些研究对于提高网络安全、防范伪造人脸技术带来的挑战具有重要意义.

2　人脸深度伪造技术

人脸深度伪造技术已在图像处理领域引起广泛关注. 随着深度学习技术的迅猛发展, 人脸伪造技术逐渐成为引人注目的研究方向, 这些技术涉及多种方法, 包括 GAN[1], FSGAN[2], Face2Face[3], Audio-Driven Face Animation[4], Lip Syncing[5]及 GAN 变体等. 本节将简单介绍这些深度人脸伪造方法的原理、应用以及相关研究的发展.

2.1　GAN 及其变体

GAN 作为人脸深度伪造使用最多的模型, 由 Goodfellow 等[1]在 2020 年提出通过 GAN 的生成器和判别器合作完成伪造人脸生成任务; 其中, 生成器负责生成图像, 判别器用于区分生成的图像与真实图像, 这种对抗过程的结果是生成器逐渐学会生成难以与真实图像区分的逼真图像. GAN 具备生成高质量数据、自动学习数据分布、创造性多样性等优势[6-8], 在人脸伪造中发挥着至关重要的作用. 在 GAN 的发展过程中产生了多种改进模型, 如 DCGAN(deep convolutional GAN)[9], StyleGAN[10], CycleGAN[11], StarGAN[12], PG-GAN(progressive growing of GAN)[13]等. DCGAN 采用卷积神经网络(convolutional neural networks, CNN)替代原始 GAN 的全连接层, 能更好地捕捉图像的空间结构, 并引入批量归一化、LeakyReLU 激活函数和权重初始化等策略, 以提升训练稳定性和生成效果; 其生成速度快, 适用于实时应用. StyleGAN 在 DCGAN 的基础上进一步演进, 引入逐步生长的生成器架构, 从低分辨率逐步生成高分辨率图像, 同时引入风格转移概念, 让用户控制生成图像的多种属性, 在人脸伪造中极具价值. CycleGAN[11]可将一种风格的人脸图像转换为另一种风格, 借助循环一致性损失函数进行约束, 避免不自然或矛盾的结果, 生成逼真多样的人脸. StarGAN[12]作为多领域图像转换模型, 其拥有生成器和判别器共享结构, 能够将输入图像转换为多个领域的图像; 并引入条件控制器实现不同特征邻域之间的转换, 生成的人脸图像具有多样化属性, 为特定应用场景提供重要价值. PG-GAN[13]以渐进式生成模型为特点, 生成的图像呈现更细腻纹理和丰富细节, 为人脸合成、表情操控、年龄改变、性别转换等方面增添新的数据集.

随着 GAN 技术的不断发展, 其对人脸伪造研究影响日益深远, 并将朝着更高质量图像生成方向迈进; 同时也对伪造人脸检测技术提出更高要求, 包括更高的准确性与鲁棒性、更快的响应速度和效率、更强的安全性和保密性.

2.2　DDPM

近年来, DDPM 在图像研究领域崭露头角, 其基于 UNet 架构[14], 借助残差层、下采样卷积层以及带有上采样卷积的残差层堆叠, 通过跳跃层连

接具有相同空间大小的层. DDPM 在 2015 年由 Sohl-Dickstein 等[15]提出, 其扩展过程通过构建生成式马尔可夫链将简单的已知分布转换为目标分布. 然而, DDPM 的保真度不及 GAN 模型. 因此, Ho 等[16]通过在加权变分下训练 DDPM, 建立 DDPM 与朗之万动力学(Langevin dynamics)的联系, 但在性能指标方面仍不如 GAN 模型; Nichol 等[17]通过正态分布的预测 DDPM 的均值和方差, 进一步改善了生成结果; Dhariwal 等[18]引入分类引导模型采样和生成, 同时加速逆向采样速度, 提升生成图像的保真度. 尽管 Diffusion GANs[19]将 DDPM 与 GAN 结合以提高推理速度, 但其生成效率相对较慢. Phung 等[20]引入基于小波扩散方案弥补速度差距, 提高处理速度并保持生成质量; Huang 等[21]提出协作 DDPM, 通过动态扩散器实现多模态人脸的生成和编辑. 虽然 DDPM 生成的图像和视频已逐渐逼真, 但其训练和推理速度仍然有巨大提升潜力.

2.3 FSGAN

FSGAN 是一种通过深度学习技术实现人脸替换的方法. 通过从一张照片或视频中提取一个人的面部特征, 并将这些特征嵌入到另一张照片或视频中实现面部替换. 该方法的便捷性使其在电影制作和娱乐领域大受欢迎. 对抗性训练和深度神经网络的进步使得 FSGAN 方法生成的图像在逼真性和自然度方面得到了极大的提升.

2.4 Face2Face

Face2Face 是一种使用神经网络实现实时面部表情转移的技术. 该模型通过捕捉视频中的面部运动和表情特征, 将这些特征应用于目标人物的面部, 实现面部表情的同步. Face2Face 技术的应用不仅在娱乐领域, 还在医学和用户界面设计等方面展现出潜力, 推动了多个领域创新的引擎, 为深度网络生成技术的应用带来广阔前景.

2.5 Audio-Driven Face Animation

Audio-Driven Face Animation 是一种通过深度学习模型实现人脸表情生成和同步的技术, 其将音频信息与面部表情关联, 使模型能够生成与音频输入相匹配的逼真面部表情. Audio-Driven Face Animation 技术使得伪造的视频在表达情感和语音同步方面更加逼真, 推动了人工智能和深度学习技术在人机交互领域的演进.

2.6 Lip Syncing

Lip Syncing 是一种利用深度学习技术实现从音频到视频的同步的方法, 主要应用于视频制作、动画和虚拟人物的表情合成等多个方向. 其目标是确保嘴唇的运动与音频的发音相匹配, 创造出更逼真的视频效果. Lip Syncing 方法对于提高视频质量、减少后期制作工作量以及增强用户体验都具有重要意义.

上述深度伪造人脸生成方法展现出了高度的创造性和逼真度, 但同时也引发了对伦理、隐私和信息安全的广泛关注. 随着对抗性方法和生成技术的不断演进, 既为创新和艺术表达提供了更多可能性, 也在不同领域中引发了对其潜在影响的审慎思考. 伦理、隐私和信息安全等问题将在技术发展的过程中持续受到重视, 促使社会与科技同步进步.

3 生成式伪造人脸图像检测方法

本节主要通过 5 个小节对生成式伪造人脸图像检测方法进行阐述. 第 3.1 节是基于数字图像处理基础, 采用传统的深度伪造图像检测方法, 如图像增强和对比损失来强化颜色信息的差异和共同辨别特征. 第 3.2 节探讨基于深层次特征的方法, 使用 CNN 如 ResNet 和 DenseNet 提取深层特征, 并通过迁移学习、微调神经网络等方法提高检测性能. 第 3.3 节基于空间域介绍基于光响应非均匀性(photo response non-uniformity, PRNU)的检测方法, 利用 PRNU 提取特定的图像特征进行伪造人脸检测. 第 3.4 节讨论基于多特征融合的方法, 从面部的细节、角膜高光等方面提取特征进行伪造检测. 第 3.5 节着重介绍基于 GAN 生成图像指纹的方法, 通过学习特定的生成伪影或指纹进行准确的深度伪造图像检测. 上述方法在不同场景和数据条件下具有各自的优势和局限性, 有助于对生成式伪造人脸图像检测方法进行初步的了解和认识.

3.1 基于数字图像处理基础的检测方法

基于数字图像处理基础的检测方法是传统的图像检测方法, 对于深度伪造图像检测着重于图像像素的处理上. 由于通过深度网络生成的伪造人脸图像的分辨率较小, 为了提高检测模型在压缩图像上的伪造人脸检测准确率, Marra 等[22]通过加入图像增强和分辨率重建模块[23]保证了模型的健壮性; Hsu 等[24]采用对比损失寻找不同 GAN 生成伪造人脸图像的共同辨别特征, 使用级联分类器充分学习真假图像差异; McCloskey 等[25]对上述取证方法进行补充, 与真实图像相比, GAN 生成的图像在像素空间上具有更高的相关性, 使用双变量直方图[26]提取伪造线索, 通过分析 GAN 生成器结构观察其对图像统计信息的影响, 提出一种检

测伪造图像的方法，利用真实图像和 GAN 生成图像之间的颜色信息差异，并通过训练的线性支持向量机进行分类.

GAN 生成的图像在像素统计信息上与真实图像是不同的，基于这一偏差，通过计算图像残差或通过不同滤波器进行处理可以有效地检测 GAN 生成的伪造人脸图像. Li 等[27]首先计算高通滤波图像的残差，然后在这些残差上提取共现矩阵，再连接起来形成可以区分真实图像和 GAN 生成式图像的特征向量；与此不同, Nataraj 等[28]提出一种使用像素共现矩阵和深度学习的组合来检测 GAN 生成的伪造图像的方法，不需要计算图像的残差，先在图像像素域的 3 个颜色通道上计算共现矩阵，再使用深度 CNN 进行训练模型. 该模型分别在不同数据集上训练和测试，取得较好的测试精度，但是在噪声图像上检测效果下降. Li 等[29]发现图像颜色空间中的相邻像素之间有很大的关联性，真实图像和生成图像的残差域差异明显. 基于该线索，将在残差域色度分量中提取的特征连接到特征向量里，采用分类器来预测图像的真伪性，该方法在数据集内测试和跨数据集测试准确率效果都很好. Nowroozi 等[30]发现，GAN 生成图像和真实图像在光谱带上存在差异，利用颜色带之间的相关性，使用跨带共现矩阵和空间共现矩阵保存面部图像的数字，再输入到 CNN 模型中进行训练，该模型的检测准确率超过 92%；但其在 JPEG[31]压缩过的图像上进行测试时准确率下降较明显，在对低质量图像检测上，基于数图像处理技术的检测方法效果有待提升.

基于数字图像处理技术的优势在于其高效性、灵活性和自动化，能够快速地处理伪造人脸图像并实现定制化需求，提高处理效率和可移植性. 然而，数字图像处理技术也存在一些挑战，如容易受到对抗性攻击、对低质量图像的处理能力有限、对计算资源需求高以及对数据集质量和数量的依赖性. 尤其面对日益更迭的深度伪造人脸图像生成模型，需要不断地改进算法以提高鲁棒性；同时需要加强对低质量图像的处理能力，以应对不同场景下的挑战，进一步提升数字图像处理技术对伪造人脸图像检测的实际应用效果和可靠性. 表 1 所示为基于数字图像处理基础的检测方法的优缺点总结.

表 1　基于数字图像处理基础的检测方法总结

模型	主要方法	优点	缺点
Marra 等[22]	图像增强，分辨率重建	保证模型健壮性，寻找共同辨别特征	JPEG 压缩下效果不佳
DeepFD[24]	对比损失	学习真假图像差异	精度不高
McCloskey 等[25]	双变量直方图	在像素空间有更高的相关性	时间代价大
Li 等[27]	高通滤波残差，共现矩阵	有效检测 GAN 生成的伪造人脸图像	依赖滤波方法，对噪声敏感
Nataraj 等[28]	像素共现矩阵和深度学习	无需计算残差，较好的测试精度	噪声图像上检测效果下降
Li 等[29]	色度残差特征提取，分类器训练	利用色度分量差异，测试效果良好	低质量图像未显示优势
Nowroozi 等[30]	跨带共现矩阵和空间共现矩阵	利用颜色带相关性，准确率超过 92%	JPEG 压缩图像准确率下降

3.2　基于深层次特征的检测方法

随着 GAN 生成式模型的不断改进，传统的伪造人脸检测方法不再适用于高分辨率的伪造人脸图像，需要深度神经网络提取图像的更深层次特征[32]. CNN 是常用的深度学习架构，现有的网络架构大多数是 CNN 的变种，如 ResNet[33]，VGGNet[34]，XceptionNet[23]等，它们能够提取到图像更加深层的特征. CNN 的基本结构有卷积层、池化层和全连接层，其中，卷积层的作用是对图像进行特征提取，池化层是对卷积层提取的特征图进行降维，全连接层对特征进行重新整理，减少特征信息的丢失.

虽然伪造人脸检测方法不断进步，但是伪造人脸技术也在不断发展，两者形成了对抗性. 随着 GAN 生成图像的潜在伪影和特定图案的显著减少，对于这些图像的检测就变得越来越有挑战性. Jeon 等[35]提出可转移 GAN 图像检测框架 T-GD，首次把迁移学习引用到 GAN 生成图像检测，该框架由一个教师和多个学生模型组成，通过相互教学和评估进行自训练，克服了灾难性的遗忘，并且仅需要少量数据训练就能够有效地检测出伪造人脸图像；Jeon 等[36]还提出一个微调神经网络架构 FDFtNet，其在 MBblockV3 模型的基础上进行微调，同时引入一个自注意力模块来关注上层语义信息和下采样层，该模型可移植性高，容易与现有的 CNN 架构集成；盛文俊等[37]提出有监督注意力机制，从而提高有用信息的权重；杨挺等[38]通过使用伪影图生成器生成伪影图加深伪造人脸与真实人

脸之间的特征差异, 从而提高伪造人脸图像检测准确率; Hsu 等[39]使用成对学习检测深度伪造图像, 将 DenseNet 转换为双流网络结构, 使用输入的成对信息训练共同伪特征网络, 再连接到分类层检测输入图像的真假. 不同 GAN 生成的图像可能存在未知的图像失真, 如模糊、噪声和 JPEG 压缩等. Liu 等[40]发现与传统方法相比神经网络更加专注于纹理信息, 基于此线索该团队提出 Gram-Net, 在图像训练网络 ResNet 中添加 Gram 结构, 在不同的语义级别网络层合并图像纹理信息, 使模型在失真图像的检测上具有鲁棒性. 当有新的伪造人脸生成方法出现时, 许多伪造人脸检测模型的检测性能会下降. 为此, Khalid 等[41]提出 Ocfakedect 方法, 利用由编码器和解码器组成的变分自动编码器(variational auto-encoders, VAE), 将输入编码转为潜在空间中的分布, 提高伪造人脸图像检测准确率. Barni 等[42]利用光谱带之间的不一致性和跨带共现矩阵训练 CNN 模型, 自动学习人脸图像的颜色不一致性, 提高检测效率. 为了解决神经网络对逆图像问题的局限性, Sabour 等[43]提出一个由多个胶囊构成的稳健架构, 将胶囊网络应用到伪造人脸检测, 通过胶囊网络提高模型泛化能力; Nguyen 等[44]提出胶囊人脸取证的方法, 使用人脸检测算法剪裁面部区域, 引入 VGG-19[34]进行图像预处理, 再将图像特征输入到胶囊网络中. 与传统的 CNN 相比, 该方法所用到的参数较少, 降低了计算成本; Xue 等[45]根据胶囊网络提出生成对抗胶囊网络 Caps-GAN, 当数据集标记样本有限时, 可有效地提高胶囊网络的泛化能力.

基于深层次特征提取的伪造人脸图像检测方法比基于数字图像处理基础的检测方法具有更高的准确率和鲁棒性, 但其在不同数据集上面的泛化能力仍需要提升. 表 2 所示为基于深层次特征的检测方法的优缺点总结.

表 2 基于深层次特征的检测方法总结

模型	主要方法	优点	缺点
T-GD[35]	T-GD 框架, 迁移学习	自训练, 克服灾难性遗忘, 少量数据训练	对特定图案检测有挑战
FDFtNet[36]	自注意力模块	整合自注意力模块, 易与 CNN 架构集成	依赖微调
盛文俊等[37]	有监督注意力机制	提高有用信息权重, 检测效果良好	未充分解决对抗性问题
杨挺等[38]	伪影图生成器	加深伪造与真实特征差异, 优于 CNN	需要额外生成伪影图
CFFN[39]	双流网络结构	使用成对信息训练共同伪特征网络, 检测准确	仍存在对抗性问题
Gram-Net[40]	GramNet 架构	利用全局图像纹理信息, 鲁棒性, 泛化能力好	新生成方法泛化能力下降
Ocfakedect[41]	VAE	利用 VAE, 仅使用真人脸训练	新生成方法性能下降
Barni 等[42]	CNN	学习颜色信息不一致性, 对抗性强	依赖像素共现矩阵
Sabour 等[43]	胶囊网络	提高泛化能力, 参数少, 降低计算成本	对逆图像问题有局限性
Capsule-Forensics[44]	胶囊网络	使用胶囊网络进行取证, 降低计算成本	依赖有限标记样本
Xue 等[45]	Caps-GAN	结合 GAN, 提高泛化能力	依赖标记样本

3.3 基于图像空间域的检测方法

XceptionNet[23], EfficientNet[46]等深度神经网络虽然能够提取到图像的深层次特征, 但是它们是通用目标检测网络, 当用于伪造人脸图像检测时, 对于未知的伪造人脸数据集, 这些通用的目标检测网络性能会降低[47]. 基于此, 不少专家学者针对伪造讨论图像的空间结构, 设计了基于图像空间域的检测方法[48-49].

Li 等[50]针对之前的伪造人脸图像检测方法重点检测已知的伪造人脸图像、对未知伪造人脸图像的检测泛化能力较低的问题, 提出一种伪造人脸检测方法 Face X-ray. 首先通过混合 2 个不同的真实人脸来建立重要的面部 X-ray, 设计特征学习模型学习混合图像, 并输出混合图像的特征权重, 然后基于混合图像的权重来预测图像是真实的或混合的概率. X-ray 方法不依赖于特定伪造图像的伪影, 且在不使用任何面部操作方法生成的图像的情况下检测效果良好; 但是, 该方法对于对抗性图像和全合成图像是失效的. Shiohara 等[51]提出自混合图像(self-blended images, SBI)方法, 其使用更通用的伪造图像鼓励分类器学习特征表示. 与 Face X-ray 不同, SBI 对输入的原始人脸图像先进行不同的篡改操作将篡改后的伪源图像与目标图像混合, 再通过混合源和目标图像的边界不一致来判断伪造人脸图像; 其不依赖特定的伪造人脸图像生成方法, 对未知生成方法的伪造人脸图像具有较好的泛化性能. 和 X-ray 方法相反, SBI 方法通过自混合伪造图像, 对全合成图像进行检测的

效果不佳. Guo 等[52]针对深度网络伪造方法生成的图像和图像编辑篡改的图像伪造痕迹差异较大的问题，提出通过分层次对图像进行多特征融合伪造人脸图像检测方法，该方法使用多分支特征提取器使每个分支学习一个级别伪造属性，并对不同的伪造属性进行层次分类；同时使用定位模块对像素伪造区域进行分割，联合多分支特征提取器检测图像伪造区域，其对全合成图像具有泛化能力.

基于图像空间域的检测方法比基于数字图像处理基础的检测方法和基于深层次特征的检测方法具有更高的泛化能力，但是其对于未知伪造人脸图像数据集的检测效果不佳. 表 3 所示为基于图像空间域的检测方法的优缺点总结.

表 3　基于图像空间域的检测方法总结

模型	主要方法	优点	缺点
左菊仙等[47]	深度神经网络	提取深层次特征，适用伪造人脸检测	对未知伪造方法效果差
Face X-ray[50]	面部 X 射线	不依赖伪影混合真实人脸，泛化性能良好	对抗性图像和全合成图像失效
SBI[51]	伪造图像鼓励分类器	学习通用表示，不依赖具体操作方法	对全合成图像效果不佳
Guo 等[52]	多特征融合伪造检测	分层次分类，泛化性能强	依赖多层次特征提取

3.4　基于多特征融合的检测方法

随着神经网络的层数增加，不同的网络层输出的特征预测与最终真假人脸判断结果息息相关. 朱新同等[53]提出一种特征融合的伪造人脸检测方法，将频域特征与空间特征融合对真假人脸进行二分类. 虽然 GAN 模型生成的人脸图像在人脸面部细节上处理得很好，但对于不同面部组件的配置缺乏约束. Yang 等[54]通过比较真实人脸和 GAN 合成人脸上检测的面部坐标位置来量化不一致性，对人脸标准进行仿射变换. 通过最小化对准误差，把检测到的地标扭曲为标准配置. 其使用面部中心区域的面部标志估计扭曲变换，当 GAN 生成的人脸地标边缘分布出现差异时，把这些地标位置矢量化成向量作为特征向量，构建 GAN 生成人脸和真实人脸的分类系统. 所提出的方法在低维输入、轻加权模型和对尺度变化的鲁棒性等方面是有效的，但在现实场景的使用中，图像有可能不符合预期假设. 为此，Matern 等[55]针对 GAN 生成的图像会留下一些视觉伪影的问题，通过构建简单的手工特征来表征操作痕迹；但这些特征描述了特定的伪影，仅使用小型的神经网络分类模型，对训练数据和时间要求都很低. 虽然 Hu 等[56]注意到之前的 GAN 合成的人脸图像存在虹膜颜色不一致、镜面反射可能缺失或者简化为白点，但是随着 GAN 生成模型不断发展，这些问题很大程度上地得到了改善. Matern 等[55]基于在相同条件下双眼的角膜镜面高光形状大致相同为线索，提出一种自动比较两只眼睛的角膜镜面高光的相似度的方法，使用人脸检测器和地标提取器定位和提取人脸轮廓；虽然该方法取得了不错的效果，但是当光源非常靠近被摄体或周边光源在双眼中都不可见时，其不适用于不存在镜面图案的图像. PRNU[57]也叫传感器噪声，源于数字图像传感器中光电转换期间单个像素之间的微小变化，数字图像传感器将这种弱噪声样信号投射到所采集的图像中，在相同的条件下，不同的相机 PRNU 不一样. 因为图像移动或者拼接会改变 PRNU 特征，所以可以将其作为深度伪造的水印. Scherhag 等[58]针对面部变化对 PRNU 的空间和光谱特性的影响，提出一种基于 PRNU 的方法，从面部图像中提取面部区域并将其归一化；除提取光谱特征外，还研究了空间特征，构造多特征融合的检测方法. Rathgeb 等[59]也是基于 PRNU 对空间和光谱特征得的标准化分数进行融合，该模型训练效率较高.

多特征融合可以综合利用不同特征的信息，提高检测模型对于伪造人脸的识别能力. 不同特征之间的互补和协同作用可以有效地弥补各自的不足，使得检测结果更加准确、可靠. 即使在面对未知的伪造方法时，模型也能够通过多特征的综合分析进行有效的识别. 基于多特征融合检测模型虽然能够获得较高的检测准确率，但是仅局限于某类特定的数据集，其在复杂背景下的鲁棒性较差. 表 4 所示为基于多特征融合的检测方法的优缺点总结.

3.5　基于 GAN 生成图像指纹检测方法

随着 GAN 不断发展，生成的图像让人越来越难以分辨，但这些图像总是会留下一些特定的生成伪影. 早期的方法是识别每个 GAN 模型生成的伪影，但是不同 GAN 模型生成的图像伪影是不同的，当 GAN 架构改变时检测准确率会变低. 为此，

表 4 基于多特征融合的检测方法总结

模型	主要方法	优点	缺点
朱新同等[53]	特征融合伪造人脸检测	频域与空间特征融合	对现实场景的使用有限
Yang 等[54]	面部坐标位置比较	标准配置，适用于低维输入	对被遮挡面部不适用
Matern 等[55]	视觉伪影特征	少量数据和时间要求短	对存在光源问题的图像不适用
Hu 等[56]	角膜高光相似度	自动评估，相对稳健	不适用于不存在镜面图案的图像
Scherhag 等[58]	PRNU	适用于相机 PRNU 检测	泛化能力差，仅局限于特定数据集
Rathgeb 等[59]	PRNU	时间训练较少，检测错误率低	泛化能力差，仅局限于特定数据集

Marra 等[60]发现, GAN 模型生成的图像中都会留下一些特定的痕迹，跟现实中相机拍摄的照片通过 PRNU 模式的痕迹标记采集的图像是类似的，这个独特的标记可以被称为指纹，可以根据这些指纹进行可靠的伪造人脸检测. 基于这一发现，Yu 等[61]首次提出对于 GAN 生成的图像指纹的研究，不同于 GAN 生成的图像与真实图像之间的差异，GAN 生成的指纹是肉眼难以分辨的，并且不受 GAN 伪影的影响. 与基于 PRNU 的方法不同，该研究通过训练一个归因分类器学习统一的嵌入空间，同时通过修改归因网络架构和图像源学习指纹图像预测图像来源. 虽然检测准确率较高，但是在低质量图像上的检测效果欠佳. 为此，Guarnera 等[62]认为，伪造人脸图像中像素局部相关性取决于 GAN 中所有层执行的操作，尤其在转置卷积层中，提出通过期望最大化算法[63]捕获图像像素相关性的数学模型，并提取一组专门针对基础卷积生成过程建模的局部特征向量；通过类似的思想，该团队通过期望最大化算法提取生成图像的卷积痕迹[64]，设计的方法不仅可以识别生成图像，还可以识别到生成该图像的 GAN 框架，提高了模型对未知伪造人脸图像的检测能力，具有更好的鲁棒性. Wang 等[65]根据伪造人脸的检测器与人眼的视角不同，提出一个基于生成网络(称为伪造人脸摧毁者)的有效流水线，通过训练扰动生成器，同时保护源图像不受伪造人脸操纵算法的影响. Jeong 等[66]提出一个由指纹生成器和 GAN 生成的图像检测器组成的新框架，其中，检测器通过混合采样的真实图像和生成图像减少对真实图像的数据依赖，提高了模型对 GAN 生成高质量图像的鲁棒性. 另外对于观察到的手工指纹，可以采用仅使用真实图像进行训练的自监督模型[67]，该模型可以避免训练时对数据的依赖，从而增强模型泛化能力.

基于 GAN 生成图像指纹的检测方法能够在更高层次上分析图像特征，增强了对抗攻击的能力，提高了伪造人脸图像检测方法的鲁棒性，可保护用户免受欺骗和虚假信息的影响. 表 5 所示为基于 GAN 生成图像指纹检测方法的优缺点总结. 图 2 所示为代表性的生成式伪造人脸图像检测方法时间轴.

表 5 基于 GAN 生成图像指纹检测方法总结

模型	主要方法	优点	缺点
Marra 等[60]	深度网络	不受 GAN 架构改变的影响	对低质量图像检测效果不佳
Yu 等[61]	CNN	对未知攻击具有更好的鲁棒性	对低质量图像检测效果不佳
Guarnera 等[62]	CNN	提高模型鲁棒性	对低质量图像检测效果不佳
Wang 等[65]	深度网络	对伪造人脸操纵算法具有防御性	泛化能力弱
Jeong 等[66]	GAN 检测器	避免模型对数据依赖	时间消耗大

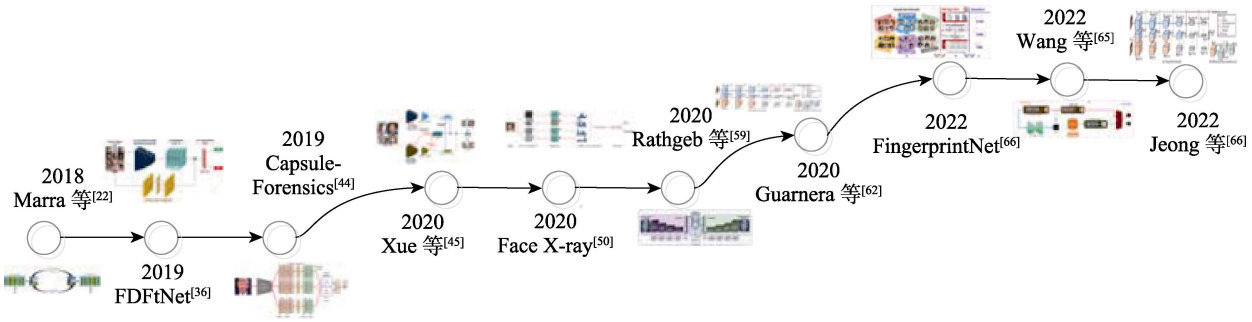

图 2 深度网络生成式伪造人脸图像代表性检测方法时间轴

4　生成式伪造人脸视频检测方法

本节主要通过 4 个小节对生成式伪造人脸视频检测方法进行阐述. 第 4.1 节介绍了基于生理信号的方法，包括利用 PPG 检测视频中的生物信号特征，以及如何通过信号变换和特征集构建分类器来检测伪造内容；还探讨了使用自回归(auto regressive, AR)特征和 DenseNet 结构提升检测效果. 第 4.2 节探讨基于身份相关信息的检测方法，介绍了通过对人物面部表情、头部运动等特征建模，以及将身份作为先验来检测伪造人脸视频的方法；还讨论了如何在特定人物身份检测上应用这些方法. 第 4.3 节讨论多模态检测方法，包括结合视觉、音频、嘴型等信息的方法，以及如何通过不同模态之间的匹配和对比来检测伪造人脸视频. 第 4.4 节针对时空不一致的检测方法介绍了基于视频帧之间时空不一致性，包括利用不一致性线索构建网络、利用时间变换网络和局部面部区域位移轨迹检测伪造人脸视频；并讨论了如何通过细微运动引起的不一致性区分不同人脸伪造类型.

4.1　基于生理信号的检测方法

Ciftci 等[68]认为人像视频中存在着生物信号的时空特征，利用 PPG 真实录制的视频中人类心跳与脉搏的信号具有时空一致性，这些特征在伪造视频中是不存在的，基于这种思路提出基于活体生理信号的检测方法，取得了较高的准确率. 通过收集具有面部部分的固定长度视频片段，定义每个面部内的感兴趣区域以及从这些时间片段中的该区域提取若干生物信号；再对成对的分离进行几次信号变换，检查提取的生物信号，通过分析建议的信号变换和相应的特征集，为虚假内容制定通用的分类器；最后生成新的信号图，并使用 CNN 改进用于检测合成内容的传统分类器. Jin 等[69]指出在传统的 PPG 方法检测伪造人脸视频中，为了消除环境变化造成的噪声、提取清晰的生理信号，通常会对视频进行去噪与滤波处理，但这会破坏伪造视频中的异常信号、造成低效性问题. 因此提出将同一块中不同帧的信号值按行排列，再通过深度网络进行分类，达到基于不同心率提取算法的伪造人脸检测视频效果.

Mao 等[70]提出另一种生理信号检测的方案，即以 PPG 特征和 AR 特征作为时间域和空间域的取证基础，利用 PPG 原理追踪人脸皮肤血细胞对光的吸收来估计人像视频的时域心率，不规则的心率波动可以被视为篡改的痕迹；同时，AR 系数不仅能够反映像素间的相关性，也能够反映在生成伪造人脸的过程中由上采样引起的平滑痕迹. 该方案还结合了基于 ACBlock 结构和改进的 DenseNet 进行真实性求证，采用非对称卷积结构强化了图像倒置与左右翻转的鲁棒性；在多个伪造人脸数据集中具有更高的准确率和更好的泛化能力.

伪造人脸视频检测在大规模数据和强大的深度模型上表现良好，但现有方法的泛化性仍需要提升. Yu 等[71]认为，单峰外观和 rPPG 特征容易受到高保真人脸 3D 掩模和视频回放攻击，其利用视觉外观和生理 rPPG 线索建立了人脸欺骗和伪造检测基准；使用面部时空 rPPG 信号图及其连续小波变换对应物作为输入，增强了 rPPG 周期性识别；对外观和 rPPG 特征进行加权以及层归一化处理，以减轻模态偏差并提高融合效率. Stefanov 等[72]基于音频、视频和生理信号，提出一种多模态融合检测方案，通过用图形卷积网络将学习视频与生理图 2 种模态融合来检测真实视频和伪视频. 表 6 所示为基于生理信号的检测方法的优缺点总结.

表 6　基于生理信号的检测方法总结

模型	主要方法	优点	缺点
Ciftci 等[68]	PPG	高准确率，生物信号时空一致性	存在低效性问题
Jin 等[69]	PPG_MAP	不受噪声影响，基于更多维度的信号计算	对低质量图像检测效果不佳
Mao 等[70]	PPG 和 AR 特征	结合时空域证据，非对称卷积结构提高鲁棒性	泛化问题仍待解决
Yu 等[71]	双分支生理网络	高精度	泛化性能较差
Stefanov 等[72]	图形卷积网络架构	高准确率	需要更多数据训练

4.2　基于身份相关信息的检测方法

目前，基于身份信息的伪造人脸旨在让视频中的人物做出特定的表情与头部运动，这种伪造方式会对个人的习惯表述模式造成破坏. 为此，Agarwal 等[73]提出一种痕迹检验技术，通过对具体人物的面部表情与头部运动进行建模，再根据特定的模型与特定的人物进行一致性判断，由于伪造人脸视频往往会破坏人物表情与运动的一致性，因此可以通过一致性进行认证. 该技术利用开源的 OpenFace2.0[74]提取输入视频的面部和头部运动单元，通过采集 20

种运动单元, 产生了 190 维特征, 再构建检测模型(单类支持向量机[75]), 由该模型结果判断视频内容真伪; 但由于这种方法的泛化能力弱, 因此更多用于对重要人物的检测[76]. Dong 等[77]根据视频中人员的身份信息对伪造人脸进行检测, 提出将身份信息作为一个强大的语义级别先验器, 把对伪造人脸检测以伪影驱动为主导转变为以身份驱动作为主导; 为了促进以身份驱动的伪造人脸检测研究发展, 建立一个大规模伪造人脸数据集 Vox-DeepFake, 其中明确指明每个视频的参考身份信息. 在伪造人脸视频的检测中, 一些研究者将注意力放在面部表情、嘴部和音频匹配等动态特征是否同步上, 忽视了静态特征. 为了解决该问题, Agarwal 等[78]提出将人耳等静态特征与动态特征结合的辩证方法. 为了使伪造人脸检测有更好的通用性, Cozzolino 等[79]提出 ID-reveal 方法, 它可以避免对特定操作后的视频进行训练, 只需要对未经过篡改的视频进行训练, 扩大了可使用数据集的范围. ID-reveal 方法基于时序 ID 网络和 3DMM 生成网络 2 个神经网络, 它们以对抗的方式相互作用, 在处理不同身份的视频时, 利用 3DMM 生成网络将 *A* 身份的脸部替换到 *B* 身份的脸部上, 模拟伪造人脸视频; 将生成的视频传入时序 ID 网络进行对抗性训练, 使训练后的模型更加专注于身份上的差异. 通过这个过程, 模型能够有效地判别视频真伪的身份特征, 该方法可以准确地区分真实和伪造视频, 保障模型在应对伪造人脸威胁时的高效性和可靠性. Cozzolino 等[80]还提出另一种思路的改进方法, 给定一组真实身份, 确定锚点视频、相同身份视频以及不同身份视频, 在训练迭代中, 分析多个视频段, 并从音频和视频信号中提取嵌入向量, 分别计算仅视频、仅音频和音频视频的 3 个相似性度量矩阵, 再对矩阵进行评估和对比. 这个过程有助于优化伪造人脸视频检测模型, 使其在嵌入空间中更好地捕捉相同身份的特征, 并将不同身份的特征更加明显地分隔开来. 该方法利用数据集 VoxCeleb2[81]进行训练, 但并不直接寻找伪造人脸视频中的操作痕迹, 而是依赖高级语义寻找出间接的线索来证明视频中的人为伪造, 高级语义特征的使用在未来研究方向上有很好的前景. Dong 等[82]延续 Face X-ray[50]的思路, 提出一种通过设计身份一致性变化器[83]来检测伪造面部视频的方法. 所设计的变化器使用 Transformer 同时学习内部身份和外部身份, 其灵感来源于视觉 Transformer 在面部分类学习中身份信息的成功应用. Transformer 对内部身份和外部身份的输出使用 Arcface[84]中提出的基于余弦的 Softmax 损失作为分类损失; 为了进一步鼓励这种期望的特性, 在交换的面部特征上引入一致性损失, 确保同一个人在不同视频帧的外部特征和内部特征上的距离尽可能小. 该方法在泛化能力和对低质量伪造人脸视频检测的鲁棒性方面达到了当前的先进水平.

大部分基于身份伪造检测模型仅适用于闭集场景, 需要参考身份集来提供候选身份. 如何让检测器识别不明身份的人脸伪造是值得研究的问题. 为了解决这个问题, Liu 等[85]让模型专注于视频帧之间的身份不一致信息, 提出时间身份不一致网络 TI^{2}Net 框架. 该框架对数据预处理后通过身份编码器转化三元组向量, 利用差分运算强调视频帧之间的特征, 特别是帧之间的身份不一致, 并利用三元组损失作为损失函数最小化锚点正距离以及最大化锚点负距离, 其在 Faceforensics++数据集上的接收者操作特征曲线下的面积(area under the curve, AUC)达到 99.95%. 表 7 所示为对基于身份相关信息的检测方法的优缺点总结.

4.3 基于多模态的检测方法

随着伪造人脸越来越逼真, 普通的单模态伪造人脸检测模型难以准确地检测出伪造人脸, Lewis 等[86]提出一种多模态检测方法 NOLANet. 该方法将输入视频进行预处理, 采用子网络 VSNet 处理视觉特征和频谱, FourierNet 处理音频信号和频谱, Lip-Speech 处理嘴型特征和语音序列, 最后将 3 种自网络输出传入到长短记忆(long short term memory, LSTM)网络; 在 2020 年脸书伪造人脸检测挑战数据集 DFDC 上实现了 61.95%的检测准确率.

表 7 基于身份相关信息的检测方法总结

模型	主要方法	优点	缺点
Agarwal 等[73]	痕迹检验技术	用于认证, 适用于重要人物	对于非重要人物检测有限
Dong 等[77]	身份作为语义先验器	以身份驱动, 有助于重要人物检测	可能忽略其他伪造痕迹
Agarwal 等[78]	结合静态特征和动态特征进行检测	补充动态特征不足, 提高准确性	对于全面伪造可能不够鲁棒
Cozzolino 等[80]	时序 ID 网络和 3DMM 生成网络结合	高效, 专注身份差异, 通用性更好	数据集范围有限
Dong 等[82]	音频和视频信号嵌入向量分析	明显分隔身份特征	依赖高级语义
TI^{2}Net[85]	身份不一致网络 TI^{2}Net 框架	用于识别不明身份	需要数据预处理

Cai 等[87]对于多模态检测的内容做出更细的划分，观察到虽然假内容可能只是构成整段视频内容的一小部分，但是可以改变整段内容的真实含义和情感，因此创建一种由内容驱动操作的数据集 LAV-DF；同时提出一种基于视觉和音频信息的精确预测伪造区域边界的多模态方法 BA-TFD. 该方法根据输入的视频提取图像和音频特征，之后利用对比损失函数学习 2 个模态的特征，对于正对，对比损失使模态之间的差异最小化，而对于负对，对比损失使该差值大于一个常数；再用交叉熵损失训练分类器；最后用多模态融合预测边界. Shahzad 等[88]对另一种细化方向进行研究，提出基于唇部特征的多模态检测，以高级语义特征为目标，利用从视频中提取的嘴唇序列与 Wav2lip 模型从音频中生成的合成嘴唇序列之间的不匹配来检测伪造视频. 实验结果表明，在低维嵌入空间中，对于真实视频，嘴唇运动的嵌入与相应的合成嘴唇序列的嵌入接近，而对于假视频它们相对分离，该方法在多模态 FakeAVCeleb 数据集[89]上的性能优于其他方法. 多模态不仅可以从结合音频和文本信息等相关因素的角度出发，还可以从视频内部的角度进行分析. 如伪造模式的大小差异较大，伪造颜色不匹配，伪造痕迹通常出现在局部区域(如嘴角)等，为了进行伪造检测，将视频中的不同区域视为不同的模态. 基于这一假设，Wang 等[90]提出一种多尺度架构 ST-M2TR，旨在捕捉可能具有不同大小的伪造区域. 值得注意的是，该模型在特定的伪造人脸数据集下训练可能导致模型过度拟合该数据集的某些特征，从而在一定程度上影响检测方法的可推广性. 因此 Cheng 等[91]提出一种快速适应的方法，先进行预训练，再进行微调. 该方法根据同个人声音和人脸的同质性，提出一种实用的语音人脸匹配检测，即通过匹配相似度而不是直接寻找伪造人脸的伪影来检测；在微调部分，使用余弦相似度分类. 实验结果表明，该方法在 DFDC，FakeAVCeleb 数据集上分别取得 85.13%和 86.11%的准确率.

通过预训练学习更丰富的人脸特征，以及在特定方向的微调增强模型的泛化能力，是一种有效的伪造人脸视频检测方法；但这也伴随着一个问题，即预训练阶段需要对域外的未标记数据进行大量的预处理. 为此，Knafo 等[92]提出一种由自监督阶段和有监督微调阶段组成的多模态方法. 在自监督阶段，利用域外多模态视频为每种模态创建鲁棒的表示；而在有监督微调阶段，通过任务特定的检测器实现在伪视频检测方面的微调. 该方法能够降低预训练阶段对大量未标记数据的需求，提高模型在伪视频检测任务上的性能. 相比单一模态的检测方法，多模态方法能够更全面地分析和识别伪造人脸的特征，提高检测的可靠性. 表 8 所示为基于多模态的检测方法的优缺点总结.

表 8　基于多模态的检测方法总结

模型	主要方法	优点	缺点
Lewis 等[86]	视觉、音频、嘴形特征融合，LSTM 网络	多模态融合提高检测准确率	计算复杂度较高
Cai 等[87]	基于内容驱动操作，视觉和音频的融合	适用于真实内容分析	对于复杂伪造情况表现一般
Shahzad 等[88]	嵌入空间分析	在多模态数据集上性能优越	对于某些伪造模式不敏感
ST-M2TR[90]	多尺度架构，区域视为不同模态	有效捕捉伪造区域	在特定数据集上过度拟合
Cheng 等[91]	语音人脸匹配检测 VFD	高检测准确率	对于极高质量伪造性能下降
FakeOut[92]	自监督阶段与有监督阶段微调	降低对未标记数据的需求	复杂性高

4.4　基于时空不一致的检测方法

基于 GAN 生成的伪造人脸视频在每帧中都具有高度的真实感，为解决此问题，Zhang 等[93]利用视频帧之间存在的不一致性线索，构建时域分离三维卷积神经网络(time-domain separated 3D convolutional neural networks, TD-3DCNN). 其中，3DCNN(3D convolutional neural networks)用于特征提取以及分类操作，时域(time-domain, TD)模块处理视频帧；使用预训练的人脸检测网络提取每帧人脸特征；最后经过一系列卷积操作，得到压缩后的特征向量，通过 Softmax 得到预测结果. 与同类模型相比，该方法具有很好的泛化能力与准确性. Zheng 等[94]提出一个端到端的框架，由全时间卷积网络(full time convolutional networks, FTCN)和时间变换网络组成，重点关注如何使网络学习时间上的非相关性. 该框架通过限制网络处理空间相关性的能力，当空间维度受到限制时网络的演化将更加依赖于时间；同时，利用 Transformer 捕捉沿时间维度的长期差异. 该框架可以在没有任何预训练或人工操作的数据集训练中取得较好的结果.

为了充分利用深度假视频中不同帧的局部面

部区域中分布的时空不一致性, 2022 年, Zhang 等[95]提出一种简洁而高效的方法，即补丁级别的时空丢失变换器. 该方法将输入视频提取为面部帧序列, 并将其重新组合成一袋补丁包; 然后传入视觉变换器中, 学习对不同帧的局部面部区域中的动态时空线索进行区分表示. 该方法通过补丁级别的时空丢失操作, 充分探索了时空不一致性; 逐步减少补丁包中的部分面部补丁, 通过丢弃操作大幅减少需要处理的数据, 从而增强了该方法的鲁棒性. Gu 等[96]通过对伪造视频序列的局部运动进行探索, 提出一个采样单元 snippet 用于学习局部不一致性, 并构建了一个由片段内不一致性模块和片段间交互模块组成的动态不一致性框架. 其中, 片段内不一致性模块运用双向时间差运算和可学习卷积函数, 挖掘每个片段内的运动特征; 片段间交互模块促进跨片段的信息交互, 形成全局表示. 在 Faceforensics++, Celeb-DF, DFDC 和 Wild-Deepfake 数据集上, 与当前同类模型相比, 该方法的表现最佳. 大多数伪造人脸检测方法主要针对整个视频或在随机位置进行的时空修改, 涉及身份、面部属性和对抗性扰动. 为了准确地检测伪造视频, 可以将注意力集中在基于内容的视频篡改方向上. 基于这一思路, Gu 等[97]建立了一个内容驱动的视听伪造人脸数据集 LAV-DF, 为了有效地判别这类伪造, 并设计了边界感知时间感知检测方法 HCIL. 在融合局部与全局角度对比学习的方向上, 该方法取得很好的效果. 除了通过空间与时间序列抽取独立的多帧检测来确定视频的真伪外, 强制转换人脸必然使得视频逐帧轨迹序列存在误差. 为此, Sun 等[98]利用面部区域位移轨迹这个具有模式差异的鲁棒特征, 基于虚拟锚点的面部区域位移轨迹提取方法将多个局部区域作为跟踪目标, 在每个区域逐帧筛选具有良好跟踪特性的特征点, 有效地捕获被操纵视频轨迹中的时空异常.

基于时空不一致的检测方法通过分析视频序列中人脸的动态特征，识别出伪造视频中的人工痕迹，从而高效地检测伪造人脸. 表 9 所示为基于时空不一致的检测方法的优缺点总结. 图 3 所示为代表性的生成式伪造人脸视频检测方法的时间轴.

5 实 验

5.1 数据集

伪造人脸的数据集用于训练、测试和评估模型的性能, 虽然现有的数据集是由多种不同的生成模型生成的, 但数据集数量和种类较少. 表 10 展示了当前使用较多的数据集[13,99-107], 如 Faceforensics++[99], Celeb-DF[100], DFDC[101], PGGAN[13]等, 随着生成技术的发展, 伪造人脸数据集也在不断增加.

(1) UADFV[102]和 DF-TIMIT[103]

UADFV 数据集含 49 个真实视频, 用于创建 49 个伪造人脸视频, 这些视频的平均时长约为 11.14 s.

表 9 基于时空不一致的检测方法总结

模型	主要方法	优点	缺点
Zhang 等[93]	3DCNN 特征提取	对抗性学习的高泛化与准确性	计算过程相对较复杂
FTCN[94]	FTCN 和时间变换网络	不依赖预训练	对于复杂情况表现一般
SDT[95]	局部面部区域的时空不一致性	通过丢弃操作增强鲁棒性	对高质量伪造的适应性下降
Gu 等[96]	Snippet 采样单元, 局部不一致性模块	在多个数据集上表现优越	对于极端情况可能不敏感
HCIL[97]	边界感知, 时间感知检测方法	对特定伪造领域有效果	需要专门设计数据集
Sun 等[98]	基于虚拟锚点, 捕获时空异常	利用鲁棒特征有效地捕获异常	对于复杂数据不具备鲁棒性

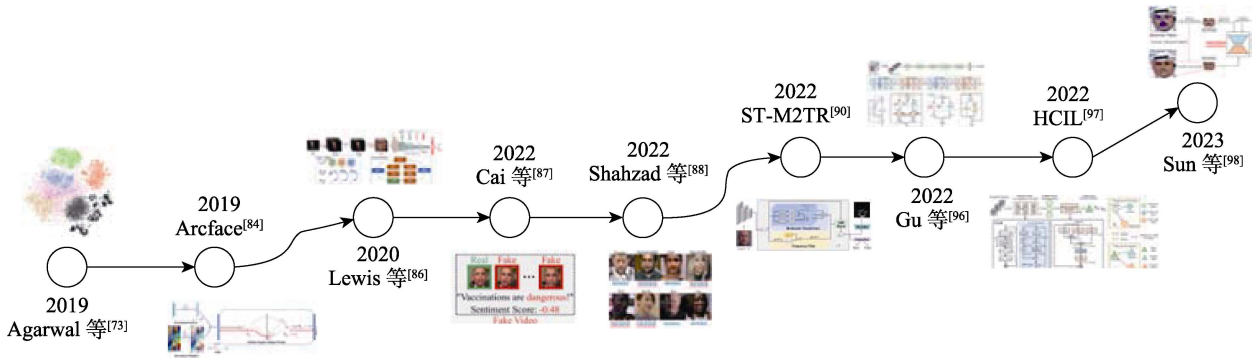

图 3 深度网络生成式伪造人脸视频代表性检测方法时间轴

表 10　伪造人脸人脸检测常用的数据集

数据集名称	数据集规模		数据类型	URL
	真实	伪造		
UADFV[102]	49	49	视频	https://arxiv.org/abs/1812.08685
DF-TIMIT[103]	0	49	视频	https://conradsanderson.id.au/vidtimit
Faceforensics++[99]	1 000	5 000	视频	https://github.com/ondyari/FaceForensics
DFDC[101]	23 564	104 500	视频	https://github.com/bomb2peng/DFGC_starterkit
Celeb-DF[100]	590	5 639	视频	https://github.com/yuezunli/celebdeepfakeforensics
Deeperforensic-1.0[104]	50 000	10 000	视频	https://github.com/EndlessSora/DeeperForensics-10
FFIW-10K[105]	10 000	10 000	视频	https://github.com/tfzhou/FFIW
Wilddeepfake[106]	3 805	3 509	视频	https://github.com/deepfakeinthewild/deepfake-in-the-wild
iFakeFaceDB[107]	494 414	330 000	图像	https://github.com/socialabubi/iFakeFaceDB
PGGAN[13]	0	80 000	图像	https://github.com/tkarras/progressive_growing_of_gans

DF-TIMIT 数据集是在 VidTimit 数据集 320 个真实视频的基础上创建的，包括 320 个低质量视频和 320 个高质量视频. UADFV 和 DF-TIMIT 数据集使用基于 GAN 的开源软件创建伪造人脸视频，它们的伪造种类较单一，视频的数量和质量都较低，多用于基础模型的训练.

(2) Faceforensics++[99]

虽然目前最先进的面部视频操作处理方法在视觉上有很好的效果，但处理后的数据集质量较低，且伪造类型较少. 为此，Rössler 等[99]创建了伪造类型和数量更多的 Faceforensics++数据集，其在 YouTube 上下载了 1 000 个原始视频，使用 FSGAN, DeepFakes, Face2Face, NeuralTextures 和 FaceShifter 这 5 种先进的人脸伪造方法生成 5 000 多个伪造人脸视频，FSGAN 方法提取源视频的需要操作的面部区域并将其转移到目标视频. DeepFakes 有 2 个自动的共享编码器，分别用来训练重建面部和目标面部的图像. Face2Face 是一个面部再现方法，可以把源视频的人脸表情替换到目标视频的人脸上，并且保持目标人物的身份不变. Faceforensics++数据集数据规模较大，覆盖的生成技术类型较多，很多伪造人脸检测模型使用该数据集训练模型，但容易产生过拟合现象，所以通常使用其他数据集做跨数据集测试.

(3) DFDC[101]

DFDC 数据集的数据规模较大，是公开可用的人脸伪造视频数据集，共有来自 3 426 名付费演员的超过 100 000 个视频，使用几种 DeepFake、基于 GAN 的方法有 NTH, FSG, FSGAN 和 StyleGAN，包含多种伪造人脸类型，并且视频背景多样，可以推广到真实的野外场景，但视频的伪造质量不够好，人物运动时有伪影.

(4) Celeb-DF[100]

Celeb-DF 数据集由 590 个真实视频和 5 639 个使用真实视频生成的名人的高质量伪造视频组成；所有真实的视频都是在 YouTube 上公开的视频，包含的性别、年龄和种族分布各不相同，视频中的人脸面部大小、方向、光照以及背景都有较大变化；改进了之前数据集的视频存在视觉伪影的情况，最终的视频采用 MPEG4.0 格式. 虽然数据集的整体质量提高了，但由于其规模较小，真实视频和假视频样本分布不均，因此不使用该数据集训练伪造人脸检测模型，通常用于进行跨数据集测试，以评估模型的泛化能力.

(5) Deeperfrensic-1.0[104]

Deeperfrensics-1.0 是现有最大的人脸伪造检测数据集，共有 60 000 个视频，由 17 600 000 帧组成，其数量规模是现有数据集的几倍，并与 100 名付费演员签署正式协议，同意使用和操纵他们的面部视频；还使用各种扰动来更好地模拟真实场景中的视频，每个视频可能受到一种以上扰动的混合.该数据集具有高质量、大规模和高度多样性的优点，缺点是生成的过程较单一、数据集的伪造类型较少.

(6) FFIW-10K[105]

FFIW-10K 数据集有 10 000 个高质量的伪造视频和 10 000 个真实视频，平均每帧中有 3 幅人脸图像，其操作过程是全自动的，具有高可扩展性，且人工成本更低. 该数据集在野外收集原始视频，确保大量视频包含多个人脸. 先根据不同的关键词从 YouTube 上搜索一组视频，将每个视频分成 4 个统一的片段，并从每个片段中随机选择一个 12 s 的序列，总共产生了大约 12 000 个序列，将其用作面部操作的原始视频. 与之前的视频数据集相比，该数据集包含真视频和假视频，并且占比均匀，因为现有的数据集通过伪造技术的发展来更新数据

集，且其在稳定的条件下伪造视频有较强的一致性和较少的伪影，所以在数据更新方面是困难的.

(7) Wilddeepfake[106]

通常，现有的数据集包含的真实视频是由少数的志愿者或者演员在特定的场景拍摄，而假视频由特定的伪造人脸技术生成，所以只能覆盖少部分当下流行的伪造技术，并且假视频的制作也是在短期内完成，没有考虑到对灯光场景等仔细调整，导致数据集缺乏多样性和低质量.

Wilddeepfake 数据集由 707 个完全从互联网上收集的伪造人脸视频组成，包含了更多样化的场景、人脸和活动等，更加贴近真实的环境；因为其中的假视频由多种不同类型和版本伪造人脸软件制作，所以用于创建伪造人脸视频的软件是未知的，用该数据集训练的模型更具有鲁棒性. 但是，由于该数据集视频数量较少，场景复杂多变，训练出来的模型检测准确率较低.

(8) iFakeFaceDB[107]

由于之前的数据集中图像的伪造伪影过于明显，为了保持生成图像的视觉质量，Neves 等[107]提出基于 GAN 指纹去除自动编码器的方法 GANPrinter. 该方法使用卷积自动编码器(auto-encoder, AE)训练真实图像，AE 可以学习真实图像的核心结构；然后利用这些数据改进现有的伪造人脸的图像，对比 Faceforensics++等伪造人脸数据集，GANPrinter 操纵合成的数据集能够使现有的伪造人脸检测模型的检测性能显著降低. iFakeFaceDB 数据集由 GANPrinter 操纵合成，共有 87 000 幅伪造图像，没有真实图像.

(9) PGGAN[13]

PGGAN 数据集是使用 PGGAN 伪造人脸生成模型生成的 80 000 幅伪造人脸图像. 与之前 GAN 的训练方法不同，PGGAN 采用的是逐步增长的方式，从低分辨率开始训练网络，然后逐层增加分辨率，以更好地生成高分辨率的图像. 这种方法可以逐步提高图像细节的生成质量，而不是在所有尺度上同时学习；生成的图像质量比早期 GAN 提高了很多，且在大分辨率下训练是稳定的，但要做到真实图像具有的真实感，需要进一步深入研究.

5.2 实验结果

在评估检测模型性能时，通常采用准确率和 AUC 作为评估指标. 准确率是一种直观且简单的评估度量，能够反映各个检测器在所有样本中正确分类的比例；在很多情况下，特别是当正负样本数量相对平衡时，该指标是一个较精准的衡量指标. 然而，当进行跨数据集测试时，训练数据集与测试数据集之间可能存在较大的差异. 因为检测器可能会偏向于将其训练过的样本所占比例更高的类别进行分类，此时准确率可能会提供错误的结论，这种现象无法准确地反映模型的完整性. 为了更全面地评估模型性能，需要引入 AUC 作为另一项评估指标. AUC 是受试者工作特征曲线(receiver operating characteristic curve, ROC)下的面积，ROC 曲线呈现在不同阈值下真阳性率与假阳性率之间的关系. AUC 越接近 1，表明分类器的性能越优越. 与准确率不同，AUC 对于处理不平衡数据集也表现出较好的鲁棒性. 这是因为 AUC 关注整个 ROC 曲线而非单一点的表现，所以在评估检测器性能时，将准确率作为直观的度量标准，其特别适用于样本平衡的情景；在考虑数据集差异和样本不平衡性的情况下，使用 AUC 这一评估指标，以全面地评价模型在不同情况下的表现.

5.2.1 生成式伪造人脸图像检测

选择 Faceforensics++(简称 FF++)数据集进行模型的训练和测试，对比伪造人脸检测模型在图像数据集上的评估结果. 目前，伪造人脸图像检测模型的 AUC 指标已超过 99%，如表 11 所示，其中，SBI[51]模型的 AUC 高达 99.64%. 它在自伪造拼接图像上表现优异，然而对于完全合成的图像，其检测准确率下降显著. 表 12 所示为不同模型在 CelebA 数据集上的实验结果，CelebA 数据集涵盖了多种最先进的 GAN 架构，如 ProGAN，SNGAN，Cramer GAN 等. 可以看出，大多数模型在该数据集上的检测准确率均超过 90%. 值得一提的是，由于 Yu 等[71]的模型专注于 GAN 生成图像中的指纹，因此在 CelebA 数据集上的检测准确率达到 99.42%.

表 11 4 种模型在 FF++数据集上的实验结果 %

模型	准确率	AUC
Ocfakedect[41]	88.24	
Capsule-Forensics[44]	93.11	
Face X-ray[50]		98.52
SBI[51]		99.64

表 12 7 种模型在 CelebA 数据集上的实验结果 %

模型	准确率	AUC
DeepFD[24]	87.10	84.40
CFFN[39]	96.76	90.96
Gram-Net[40]	93.35	
Yang 等[54]		94.13
Yu 等[71]	99.42	
Guarnera 等[62]	90.22	
FingerprintNet[64]	92.60	

表 13 所示为伪造人脸检测模型在不同 GAN 生成图像上的实验结果. 可以看出，大多数伪造人脸检测模型都能在 GAN 生成数据集上取得良好的检测准确率，其中表现最为出色的是 Barni 等[42]提出的模型，其在 StyleGAN2 数据集上获得了 99.7%的准确率. 为了验证模型的泛化能力，将同一个模型在不同的数据集上进行实验，表 14 所示为 12 种模型在不同数据集上的性能比较. 可以看出，泛化能力较好的模型包括准确率为 97.84%的 Nataraj 等[28]模型，准确率为 96.73%的 Nowroozi 等[30]模型，以及 AUC 为 99.64%的 SBI[51]模型；泛化能力较差的有 DeepFD[24]模型，其在 DCGAN 数据集获得的召回率仅为 84.42%，在 WGAN-GP 数据集召回率仅为 83.51%，同样 Ocfakedect[41]模型在 Face2Face 数据集的召回率仅为 71.27%.

表 13　GAN 生成的数据集上的实验结果　%

模型	GAN 数据集	准确率	AUC
DeepFD[24]	LSGAN	93.20	90.30
FDFtNet[36]	PGGAN	90.29	95.98
Yang 等[54]	PGGAN		94.13
Nowroozi 等[30]	StyleGAN2	96.73	
Barni 等[42]	StyleGAN2	99.70	
Guarnera 等[62]	StyleGAN	99.31	
Nataraj 等[28]	StarGAN	93.42	
T-GD[35]	StarGAN		97.32

表 14　伪造人脸图像检测模型跨数据集泛化性比较

模型	数据集	性能/%			
		精确率 *P*	召回率 *R*	准确率	AUC
DeepFD[24]	LSGAN	94.74	92.22		
	DCGAN	87.12	84.42		
	WGAN	83.87	84.75		
	WGAN-GP	81.84	83.51		
	PGGAN	92.63	91.84		
Nataraj 等[28]	StarGAN			97.84	
Nowroozi 等[30]	StyleGAN2			96.73	
T-GD[35]	PGGAN				95.87
	StarGAN				97.32
	StyleGAN				97.83
	StyleGAN2				97.71
FDFtNet[36]	PGGAN			90.29	95.98
	Deepfake			97.02	99.37
	Face2Face			96.67	98.23
CFFN[39]	BigGAN	90.92	86.53		
	SA-GAN	93.06	93.61		
	SN-GAN	93.48	90.03		
Gram-Net[40]	StyleGAN			95.51	
	CelebA			89.26	
	PGGAN			92.28	
	FFHQ			90.00	
Ocfakedect[41]	Deepfake	98.66	98.49		
	NeuralTexture	97.33	88.63		
	FaceSwap	85.93	86.45		
	Face2Face	71.22	71.27		
Barni 等[42]	StyleGAN2			99.77	
Capsule-Forensics[44]	FF++			93.11	
	Face2Face			90.36	
	FaceSwap			92.79	
Face X-ray[50]	FF++				98.52
	DFD				95.40
	DFDC				80.92
	Celeb-DF				80.58
SBIs[51]	FF++				99.64
	CDF				93.18
	DFD				97.56
	DFDC				72.42
	DFDCP				86.15
	FFIW				84.83

综上所述，伪造人脸检测模型在图像数据集上取得了显著的成果，但在面对不同类型图像和数据集时的泛化性仍然是未来研究和改进的重要方向.

5.2.2　生成式伪造人脸视频检测

表 15 和表 16 所示为伪造人脸视频检测模型在检测任务中的实验结果，判断给定的人脸视频是真实或伪造. 表 15 所示为在 FF++数据集上 6 种主流模型的检测性能对比. 可以看出，TI^{2}Net[85]模型的 AUC 达到了 99.95%，显示出卓越的性能；然而，当应用于特定领域外数据集时，一些在特定领域内表现优异的模型的性能显著下降，以 FF++数据集为训练基础的 HCIL[97]模型，Celeb-DF 数据集上仅实现了 79.0%的 AUC，而在 FF++数据集上的 AUC 却高达 98.32%；基于多模态的检测模型，如 FTCN[94]模型也展现了出色的泛化能力，但由于这类模型通常需要大量的计算资源来支撑其性能，因此其计算效率相对较低. 从表 16 可以看出，模

型的泛化能力表现不一，根据模型在不同数据集上的性能表现，FakeOut[92]，SDT[95]和 HCIL[97]在多个数据集上展现了较高的 AUC，表现出较好的泛化能力；相反，ID-Reveal[79]和 TI^{2}Net[85]在某些数据集上的性能相对较差，呈现较差的泛化能力.

表 15　6 种模型在 FF++数据集上的实验结果　%

模型	准确率	AUC
ID-Reveal[79]	78.32	87.04
TI^{2}Net[85]		99.95
ST-M2TR[90]	96.71	98.23
FTCN[94]		99.75
SDT[95]	97.04	99.85
HCIL[97]	97.92	98.32

表 16　7 种模型跨数据集泛化性比较

模型	数据集	性能/%	
		准确率	AUC
ID-Reveal[79]	DFD	81.83	90.03
	DFDC	80.44	91.04
	Celeb-DF	64.45	80.06
TI^{2}Net[85]	DFD		72.03
	Deeper		76.08
	CDF1		66.65
	CDF2		68.22
ST-M2TR[90]	Celeb-DF		95.53
	SR-DF		86.74
FakeOut[92]	FaceShifter		99.52
	DFo		99.95
	Celeb-DF		72.85
	DFDC		75.17
FTCN[94]	Celeb-DF		86.99
	DFDC		74.03
	FaceShifter		98.82
	DFo		98.85
SDT[95]	Celeb-DF	91.73	97.26
	DFDC	97.44	99.14
	FaceShifter	98.64	99.88
HCIL[97]	Celeb-DF	98.31	79.06
	DFDC	98.34	69.29

综上所述，不同的伪造人脸视频检测模型不仅在相同数据集上存在准确率差异，而且在不同数据集上也存在泛化性的差异.

6　结　语

本文针对深度网络生成式伪造人脸的检测模型进行综述，归类总结如下:

(1) 目前，伪造人脸生成技术主要是由 GAN 及其衍生模型和 DDPM 生成伪造人脸图像和视频. GAN 在人脸伪造中起到关键作用，能生成高质量数据、学习数据分布、创造多样性. DDPM 可以生成更高质量的伪造人脸图像、视频，具有巨大潜力. 未来，GAN 将朝着高质量、多模态、交互方向发展; DDPM 虽然生成质量逐渐增强，但其生成速度仍需提高.

(2) 生成式伪造人脸图像检测模型包括 5 类方法: 基于数字图像处理基础的检测方法、基于深层次特征的检测方法、基于空间域的检测方法、基于多特征融合的检测方法和基于 GAN 生成图像指纹检测方法. 传统的基于数字图像处理的检测方法侧重于像素处理，然而在低质量图像检测方面并未表现出明显的优势. 随着 GAN 技术的进步，传统伪造人脸检测方法难以适应高分辨率伪造人脸，因此许多学者转向深度神经网络，以提取更深层次的特征. 基于空间域的方法探索了伪造与检测之间的对抗性关系，为创新性伪造人脸的检测提供了新思路. 基于图像多特征融合的方法，如 X-ray 和 PRNU 也显著地提升了检测模型的准确性. 基于 GAN 生成图像指纹的检测方法，利用其独特的生成痕迹也实现了可靠的伪造人脸检测.

(3) 生成式伪造人脸视频检测方法包括 4 类: 基于生理信号的检测方法、基于身份相关信息的检测方法、基于多模态的检测方法和基于时空不一致的检测方法. 基于生理信号的方法利用 PPG 捕获心跳和脉搏信号，在伪造人脸视频中检测出时空不一致性. 基于身份相关信息的方法通过建模个人面部表情与运动，以及引入身份作为先验，实现对伪造视频的检测. 基于多模态的方法结合视觉、音频和嘴型等多种信息，应对不同类型的伪造，取得了较好效果. 基于时空不一致的方法利用视频帧的不一致性线索构建不同的检测网络，捕捉时空异常. 这些方法的出发点各异，通过对生理信号、身份信息、多模态信息和时空不一致性的分析，为伪造人脸检测提供了多样化且有效的解决方案.

未来，伪造人脸检测方法将有许多潜在的研究和发展方向，以应对不断演进的伪造技术和挑战. 一些可能的工作方向如下:

(1) 对抗性攻击与对抗训练. 随着 GAN 和伪造技术的不断发展，检测模型也面临着对抗性攻击，需要研究如何提高模型对对抗性样本的鲁棒性，以及如何进行对抗性训练以增强检测模型的

稳定性.

(2) 跨数据集泛化. 目前，许多伪造人脸检测模型在特定数据集上表现良好，但泛化到其他数据集时性能可能下降. 如何实现更好的跨数据集泛化，使检测模型能够适应不同类型的伪造图像，是一个重要的方向.

(3) 小样本学习. 由于伪造图像和视频数据集有限，因此如何在小样本情况下进行伪造检测，包括元学习、迁移学习、半监督学习等方法，是一个具有挑战性的问题.

(4) 实时性与效率. 实际应用中，伪造人脸检测需要实时性和高效性，设计轻量级的检测模型以及提高检测速度和准确率，都是需要关注的问题.

(5) 不确定性建模. 对于新型人脸伪造技术，模型往往面临不确定性，因为它们可能不遵循已知的生成规律. 建模和处理这种不确定性以及进行可靠的检测，是一个前沿课题.

(6) 可解释性和解释性. 提高伪造检测模型的可解释性，使模型能够解释伪造人脸判定依据，有助于用户理解检测结果，也可以用于后续取证工作.

(7) 自适应方法. 随着人脸伪造技术的不断变化，设计自适应的检测方法使其能够及时适应新型伪造技术，是一个迫切需要解决的问题.

综上所述，伪造人脸检测还有许多挑战和机会等待着研究者的探索和创新，从而保护数字信息的安全性和可信性，减少伪造人脸带来的负面影响.

参考文献(References):

[1] Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144

[2] Nirkin Y, Keller Y, Hassner T. FSGAN: subject agnostic face swapping and reenactment[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7184-7193

[3] Thies J, Zollhöfer M, Stamminger M, *et al*. Face2Face: real-time face capture and reenactment of RGB videos[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2387-2395

[4] Karras T, Aila T, Laine S, *et al*. Audio-driven facial animation by joint end-to-end learning of pose and emotion[J]. ACM Transactions on Graphics, 2017, 36(4): Article No.94

[5] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing Obama: learning lip sync from audio[J]. ACM Transactions on Graphics, 2017, 36(4): Article No.95

[6] Zhu Min, Ming Zhangqiang, Yan Jianrong, *et al*. A survey on generative adversarial network based person re-identification method[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(2): 163-179(in Chinese)
(朱敏，明章强，闫建荣，等. 基于生成对抗网络的行人重识别方法研究综述[J]. 计算机辅助设计与图形学学报，2022, 34(2): 163-179)

[7] Zhou Rui, Jiang Cong, Xu Qingyang, *et al*. Multi-conditional generative adversarial network for text-to-video synthesis[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(10): 1567-1579(in Chinese)
(周瑞，姜聪，许庆阳，等. 多条件生成对抗网络的文本到视频合成方法[J]. 计算机辅助设计与图形学学报，2022, 34(10): 1567-1579)

[8] Yao Zhihao, Qiao Yuehan, Xu Qianyao. Research on virtual human design in smart home[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(2): 221-229(in Chinese)
(姚智皓，乔玥涵，徐千尧. 智能家居中虚拟人的设计与生成[J]. 计算机辅助设计与图形学学报，2023, 35(2): 221-229)

[9] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[OL]. [2023-09-17]. https://arxiv.org/abs/1511.06434

[10] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4401-4410

[11] Zhu J Y, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2223-2232

[12] Choi Y, Choi M, Kim M, *et al*. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8789-8797

[13] Karras T, Aila T, Laine S, *et al*. Progressive growing of GANs for improved quality, stability, and variation[OL]. [2023-09-17]. https://arxiv.org/abs/1710.10196

[14] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241

[15] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, *et al*. Deep unsupervised learning using nonequilibrium thermodynamics[C] //Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR.org, 2015: 2256-2265

[16] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2020: Article No.574

[17] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models[C] //Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 8162-8171

[18] Dhariwal P, Nichol A Q. Diffusion models beat GANs on image synthesis[OL]. [2023-09-17]. https://arxiv.org/abs/2105.05233

[19] Xiao Z S, Kreis K, Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs[OL]. [2023-09-17].

https://arxiv.org/abs/2112.07804
[20] Phung H, Dao Q, Tran A. Wavelet diffusion models are fast and scalable image generators[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 10199-10208
[21] Huang Z Q, Chan K C K, Jiang Y M, *et al*. Collaborative diffusion for multi-modal face generation and editing[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 6080-6090
[22] Marra F, Gragnaniello D, Cozzolino D, *et al*. Detection of GAN-generated fake images over social networks[C] //Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval. Los Alamitos: IEEE Computer Society Press, 2018: 384-389
[23] Chollet F. Xception: deep learning with depthwise separable convolutions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 1251-1258
[24] Hsu C C, Lee C Y, Zhuang Y X. Learning to detect fake face images in the wild[C] //Proceedings of the International Symposium on Computer, Consumer and Control. Los Alamitos: IEEE Computer Society Press, 2018: 388-391
[25] McCloskey S, Albright M. Detecting GAN-generated imagery using color cues[OL]. [2023-09-17]. https://arxiv.org/abs/1812.08247
[26] McCloskey S, Chen C, Yu J Y. Focus manipulation detection via photometric histogram analysis[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 1674-1682
[27] Li H D, Li B, Tan S Q, *et al*. Detection of deep network generated images using disparities in color components[OL]. [2023-09-17]. https://arxiv.org/abs/1808.07276v1
[28] Nataraj L, Mohammed T M, Manjunath B S, *et al*. Detecting GAN generated fake images using co-occurrence matrices[C] // Proceedings of the Media Watermarking, Security, and Forensics. Burlingame: Society for Imaging Science and Technology, 2019
[29] Li H D, Li B, Tan S Q, *et al*. Identification of deep network generated images using disparities in color components[J]. Signal Processing, 2020, 174: Article No.107616
[30] Nowroozi E, Mauro C. Detecting high-quality GAN-generated face images using neural networks[M] //Maleh Y, Alazab M, Tawalbeh L, *et al*. Big Data Analytics and Intelligent Systems for Cyber Threat Intelligence. Gistrup: River Publishers, 2022: 235-252
[31] Luo W Q, Huang J W, Qiu G P. JPEG error analysis and its applications to digital image forensics[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(3): 480-491
[32] Dong Lin, Huang Liqing, Ye Feng, *et al*. Survey on generalization methods of face forgery detection[J]. Computer Science, 2022, 49(2): 12-30(in Chinese)
(董琳，黄丽清，叶锋，等. 人脸伪造检测泛化性方法综述[J]. 计算机科学, 2022, 49(2): 12-30)
[33] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2023-09-17]. https://arxiv.org/abs/1409.1556
[35] Jeon H, Bang Y, Kim J, *et al*. T-GD: transferable GAN-generated images detection framework[C] //Proceedings of the 37th International Conference on Machine Learning. New York: PMLR, 2020: 4746-4761
[36] Jeon H, Bang Y, Woo S S. FDFtNet: facing off fake images using fake detection fine-tuning network[C] //Proceedings of the 35th IFIP International Conference on ICT Systems Security and Privacy Protection. Heidelberg: Springer, 2020: 416-430
[37] Sheng Wenjun, Cao Lin, Zhang Fan. Forged facial video detection based on supervised attention network[J]. Computer Engineering and Design, 2023, 44(2): 504-510(in Chinese)
(盛文俊，曹林，张帆. 基于有监督注意力网络的伪造人脸视频检测[J]. 计算机工程与设计, 2023, 44(2): 504-510)
[38] Yang Ting, Zhu Xi'an, Zhang Fan. Fake face video detection method based on improved triplet loss[J]. Application Research of Computers, 2021, 38(12): 3771-3775(in Chinese)
(杨挺，朱希安，张帆. 基于改进三元组损失的伪造人脸视频检测方法[J]. 计算机应用研究，2021，38(12)：3771-3775)
[39] Hsu C C, Zhuang Y X, Lee C Y. Deep fake image detection based on pairwise learning[J]. Applied Sciences, 2020, 10(1): Article No.370
[40] Liu Z Z, Qi X J, Torr P H S. Global texture enhancement for fake face detection in the wild[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8060-8069
[41] Khalid H, Woo S S. OC-FakeDect: classifying deepfakes using one-class variational autoencoder[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 656-657
[42] Barni M, Kallas K, Nowroozi E, *et al*. CNN detection of GAN-generated face images based on cross-band co-occurrences analysis[C] //Proceedings of the IEEE International Workshop on Information Forensics and Security. Los Alamitos: IEEE Computer Society Press, 2020: 1-6
[43] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 3859-3869
[44] Nguyen H, Yamagishi J, Echizen I. Use of a capsule network to detect fake images and videos[OL]. [2023-09-17]. https://arxiv.org/abs/1910.12467
[45] Xue Z X. A general generative adversarial capsule network for hyperspectral image spectral-spatial classification[J]. Remote Sensing Letters, 2020, 11(1): 19-28
[46] Tan M X, Le Q V. EfficientNet: rethinking model scaling for convolutional neural networks[C] //Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 6105-6114
[47] Zuo Juxian, Liu Benyong. Detection for typical tampering operations in a forged image[J]. Journal of Image and Graphics, 2012, 17(11): 1367-1375(in Chinese)
(左菊仙，刘本永. 伪造图像典型篡改操作的检测[J]. 中国图象图形学报, 2012, 17(11): 1367-1375)
[48] Cao Shenhao, Liu Xiaohui, Mao Xiuqing, *et al*. A review of human face forgery and forgery-detection technologies. Journal of Image and Graphics, 2022, 27(4): 1023-1038(in Chinese)
(曹申豪，刘晓辉，毛秀青，等. 人脸伪造及检测技术综述[J]. 中国图象图形学报, 2022, 27(4): 1023-1038)
[49] Zhou Wenbo, Zhang Weiming, Yu Nenghai, *et al*. An overview

of deepfake forgery and defense techniques[J]. Journal of Signal Processing, 2021, 37(12): 2338-2355(in Chinese)
(周文柏，张卫明，俞能海，等. 人脸视频深度伪造与防御技术综述[J]. 信号处理, 2021, 37(12): 2338-2355)

[50] Li L Z, Bao J M, Zhang T, *et al*. Face X-ray for more general face forgery detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 5001-5010

[51] Shiohara K, Yamasaki T. Detecting deepfakes with self-blended images[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 18699-18708

[52] Guo X, Liu X H, Ren Z Y, *et al*. Hierarchical fine-grained image forgery detection and localization[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 3155-3165

[53] Zhu Xintong, Tang Yunqi, Geng Pengzhi. Detection algorithm of tamper and deepfake image based on feature fusion[J]. Netinfo Security, 2021, 21(8): 70-81(in Chinese)
(朱新同，唐云祁，耿鹏志. 基于特征融合的篡改与深度伪造图像检测算法[J]. 信息网络安全, 2021, 21(8): 70-81)

[54] Yang X, Li Y Z, Qi H G, *et al*. Exposing GAN-synthesized faces using landmark locations[C] //Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York: ACM Press, 2019: 113-118

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

[56] Hu S, Li Y Z, Lyu S W. Exposing GAN-generated faces using inconsistent corneal specular highlights[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2021: 2500-2504

[57] Lukas J, Fridrich J, Goljan M. Digital camera identification from sensor pattern noise[J]. IEEE Transactions on Information Forensics and Security, 2006, 1(2): 205-214

[58] Scherhag U, Debiasi L, Rathgeb C, *et al*. Detection of face morphing attacks based on PRNU analysis[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2019, 1(4): 302-317

[59] Rathgeb C, Botaljov A, Stockhardt F, *et al*. PRNU‐based detection of facial retouching[J]. IET Biometrics, 2020, 9(4): 154-164

[60] Marra F, Gragnaniello D, Verdoliva L, *et al*. Do GANs leave artificial fingerprints?[C] //Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval. Los Alamitos: IEEE Computer Society Press, 2019: 506-511

[61] Yu N, Davis L, Fritz M. Attributing fake images to GANs: learning and analyzing GAN fingerprints[C] //Proceedings of the IEEE/CVF international conference on computer vision. Los Alamitos: IEEE Computer Society Press, 2019: 7556-7566

[62] Guarnera L, Giudice O, Battiato S. Deepfake detection by analyzing convolutional traces[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 666-667

[63] Moon T K. The expectation-maximization algorithm[J]. IEEE Signal Processing Magazine, 1996, 13(6): 47-60

[64] Guarnera L, Giudice O, Battiato S. Fighting deepfake by exposing the convolutional traces on images[J]. IEEE Access, 2020, 8: 165085-165098

[65] Wang X Y, Huang J J, Ma S Q, *et al*. DeepFake disrupter: the detector of deepfake is my friend[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 14920-14929

[66] Jeong Y, Kim D, Ro Y, *et al*. FingerprintNet: synthesized fingerprints for generated image detection[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 76-94

[67] Grill J B, Strub F, Altché F, *et al*. Bootstrap your own latent a new approach to self-supervised learning[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2020: Article No.1786

[68] Ciftci U A, Demir I, Yin L J. FakeCatcher: detection of synthetic portrait videos using biological signals[OL]. [2023-09-17]. https://ieeexplore.ieee.org/document/9141516

[69] Jin X L, Ye D P, Chen C X. Countering spoof: towards detecting deepfake with multidimensional biological signals[J]. Security and Communication Networks, 2021, 2021: Article No.6626974

[70] Mao M Y, Yang J. Exposing deepfake with pixel-wise AR and PPG correlation from faint signals[OL]. [2023-09-17]. https://arxiv.org/abs/2110.15561

[71] Yu Z T, Cai R Z, Li Z, *et al*. Benchmarking joint face spoofing and forgery detection with visual and physiological cues[J]. IEEE Transactions on Dependable and Secure Computing, 2024: 1-15

[72] Stefanov K, Paliwal B, Dhall A. Visual representations of physiological signals for fake video detection[OL]. [2023-09-17]. https://arxiv.org/abs/2207.08380

[73] Agarwal S, Farid H, Gu Y M, *et al*. Protecting world leaders against deep fakes[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern RecognitionWorkshops. Los Alamitos: IEEE Computer Society Press, 2019: 38-45

[74] Baltrusaitis T, Zadeh A, Lim Y C, *et al*. OpenFace 2.0: facial behavior analysis toolkit[C] //Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 59-66

[75] Schölkopf B, Platt J C, Shawe-Taylor J, *et al*. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471

[76] Wu Wei, Zhu Jianyu, Zhang Yan, *et al*. A deepfake face image detection model supporting privacy protection[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(10): 1510-1520(in Chinese)
(吴畏，朱剑宇，张延，等. 具有隐私保护特性的深度伪造人脸检测模型[J]. 计算机辅助设计与图形学学报, 2023, 35(10): 1510-1520)

[77] Dong X Y, Bao J M, Chen D D, *et al*. Identity-driven DeepFake detection[OL]. [2023-09-17]. https://arxiv.org/abs/2012.03930

[78] Agarwal S, Farid H. Detecting deep-fake videos from aural and oral dynamics[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 981-989

[79] Cozzolino D, Rössler A, Thies J, *et al*. Id-reveal: identity-aware DeepFake video detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 15108-15117

[80] Cozzolino D, Pianese A, Nießner M, *et al*. Audio-visual person-of-interest deepfake detection[C] //Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2023: 943-952

[81] Chung J S, Nagrani A, Zisserman A. VoxCeleb2: deep speaker recognition[C] //Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad: ISCA, 2018: 1086-1090

[82] Dong X Y, Bao J M, Chen D D, *et al*. Protecting celebrities from deepfake with identity consistency transformer[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 9468-9478

[83] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[OL]. [2023-09-17]. https://arxiv.org/abs/2010.11929

[84] Deng J K, Guo J, Xue N N, *et al*. ArcFace: additive angular margin loss for deep face recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4690-4699

[85] Liu B P, Liu B, Ding M, *et al*. TI^{2}Net: temporal identity inconsistency network for deepfake detection[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 4691-4700

[86] Lewis J K, Toubal I E, Chen H, *et al*. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning[C] //Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop. Los Alamitos: IEEE Computer Society Press, 2020: 1-9

[87] Cai Z X, Stefanov K, Dhall A, *et al*. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization[C] //Proceedings of the International Conference on Digital Image Computing: Techniques and Applications. Los Alamitos: IEEE Computer Society Press, 2022: 1-10

[88] Shahzad S A, Hashmi A, Khan S, *et al*. Lip sync matters: a novel multimodal forgery detector[C] //Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Los Alamitos: IEEE Computer Society Press, 2022: 1885-1892

[89] Khalid H, Tariq S, Kim M, *et al*. FakeAVCeleb: a novel audio-video multimodal deepfake dataset[OL]. [2023-09-17]. https://arxiv.org/abs/2108.05080

[90] Wang J K, Wu Z X, Ouyang W H, *et al*. M2TR: multi-modal multi-scale transformers for deepfake detection[C] //Proceedings of the International Conference on Multimedia Retrieval. New York: ACM Press, 2022: 615-623

[91] Cheng H, Guo Y Y, Wang T Y, *et al*. Voice-face homogeneity tells deepfake[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(3): Article No.76

[92] Knafo G, Fried O. FakeOut: leveraging out-of-domain self- supervision for multi-modal video deepfake detection[OL]. [2023-09-17]. https://arxiv.org/abs/2212.00773

[93] Zhang D C, Li C Y, Lin F Z, *et al*. Detecting deepfake videos with temporal dropout 3DCNN[C] //Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal: ijcai.org, 2021: 1288-1294

[94] Zheng Y L, Bao J M, Chen D, *et al*. Exploring temporal coherence for more general video face forgery detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 15044-15054

[95] Zhang D C, Lin F Z, Hua Y Y, *et al*. Deepfake video detection with spatiotemporal dropout transformer[C] //Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 5833-5841

[96] Gu Z H, Chen Y, Yao T P, *et al*. Delving into the local: dynamic inconsistency learning for deepfake video detection[C] // Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 744-752

[97] Gu Z H, Yao T P, Chen Y, *et al*. Hierarchical contrastive inconsistency learning for deepfake video detection[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 596-613

[98] Sun Y Y, Zhang Z Y, Echizen I, *et al*. Face forgery detection based on facial region displacement trajectory series[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 633-642

[99] Rössler A, Cozzolino D, Verdoliva L, *et al*. FaceForensics++: learning to detect manipulated facial images[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 1-11

[100] Li Y Z, Yang X, Sun P, *et al*. Celeb-DF: a large-scale challenging dataset for deepfake forensics[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 3207-3216

[101] Dolhansky B, Bitton J, Pflaum B, *et al*. The DeepFake detection challenge (DFDC) dataset[OL]. [2023-09-17]. https://arxiv.org/abs/2006.07397

[102] Yang X, Li Y Z, Lyu S W. Exposing deep fakes using inconsistent head poses[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2019: 8261-8265

[103] Korshunov P, Marcel S. DeepFakes: a new threat to face recognition? Assessment and detection[OL]. [2023-09-17]. https://arxiv.org/abs/1812.08685

[104] Jiang L M, Li R, Wu W, *et al*. DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 2889-2898

[105] Zhou T F, Wang W G, Liang Z Y, *et al*. Face forensics in the wild[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 5778-5788

[106] Zi B J, Chang M H, Chen J J, *et al*. WildDeepfake: a challenging real-world dataset for deepfake detection[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 2382-2390

[107] Neves J C, Tolosana R, Vera-Rodriguez R, *et al*. GANprintR: improved fakes and evaluation of the state of the art in face manipulation detection[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 1038-1048