计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009839] [CSTR: 32024.14.csa.009839]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的伪造人脸检测技术综述①

赵　娅[1,2], 郜明超[1], 姚文达[1], 徐　锋[3]

[1](东北石油大学 计算机与信息技术学院, 大庆 163318)
[2](黑龙江省石油大数据与智能分析重点实验室, 大庆 163318)
[3](山东省滕州市公安局 网络警察大队, 滕州 277500)
通信作者: 郜明超, E-mail: 943029361@qq.com

摘　要: 近年来, 随着伪造人脸技术的快速发展, 通过伪造人脸技术合成的人脸已经非常逼真, 人眼很难鉴别, 部分不法分子对伪造人脸技术的非法应用已经对社会稳定、个人隐私造成了恶劣影响, 因此伪造人脸检测技术的重要性日益凸显. 本文系统地探讨了伪造人脸检测技术的现状, 主要从伪造人脸图像和伪造人脸视频的检测两个方面进行分析. 在伪造人脸图像检测方面, 重点讨论了基于图像空间域和频率域的方法、身份一致性检测以及人脸区域定位技术的应用. 在伪造人脸视频检测方面, 研究聚焦于时空特征融合、生理特征利用及视听信息的结合. 此外, 本文介绍了常用的评估指标, 系统分析了多种重要数据集, 包括其特点和适用场景. 同时还指出当前文献中的局限性, 例如对抗样本的鲁棒性不足、检测方法对新型伪造技术的适应性差等问题. 基于这些分析, 我们提出了未来可能的研究方向, 包括跨域检测技术的优化、新算法的探索及模型的可解释性研究. 本文不仅为研究者提供了对伪造人脸检测技术的全面了解, 也为后续研究指明了发展方向, 具有重要的理论价值和实际应用意义.

关键词: 伪造人脸检测; 深度学习; 频率域; 时空融合; 身份一致性

引用格式: 赵娅,郜明超,姚文达,徐锋.基于深度学习的伪造人脸检测技术综述.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9839.html

Review of Forged Face Detection Techniques Based on Deep Learning

ZHAO Ya[1,2], GAO Ming-Chao[1], YAO Wen-Da[1], XU Feng[3]

[1](School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)
[2](Heilongjiang Key Laboratory of Petroleum Big Data and Intelligent Analysis, Daqing 163318, China)
[3](Cyber Police Brigade, Tengzhou Public Security Bureau, Tengzhou 277500, China)

Abstract: In recent years, as the forged face technology rapidly develops, the face synthesized has been extremely hard for the human eyes to identify, and the application of this technology by some criminals has badly threatened social stability and personal privacy, so the importance of forged face detection technology has become increasingly prominent. This review systematically discusses the current status of forged face detection technology, mainly from two aspects of forged face image detection and forged face video detection. In the aspect of forged face image detection, the methods based on the image spatial domain and frequency domain, identity consistency detection, and the application of face region localization technology are discussed. In the field of forged face video detection, the research focuses on the integration of spatio-temporal features, the utilization of physiological features, and the combination of audiovisual information. In addition, the study introduces the commonly used evaluation indicators and systematically analyzes a variety of important data sets, including their characteristics and application scenarios. At the same time, it also points out the limitations in the current literature, such as the lack of robustness of adversarial samples and the poor adaptability of detection methods to new forgery techniques. Based on these analyses, this study puts forward the possible research

① 基金项目: 国家自然科学基金 (62471124); 黑龙江省自然科学基金 (LH2022F006); 黑龙江省教育科学规划重点项目 (GJB1421114)
收稿时间: 2024-10-11; 修改时间: 2024-11-12; 采用时间: 2024-12-04; csa 在线出版时间: 2025-02-26

directions in the future, including the optimization of cross-domain detection technology, the exploration of new algorithms, and the study of the model interpretability. This review not only provides researchers with a comprehensive understanding of fake face detection technology but also points out the development direction for subsequent research, possessing high theoretical value and practical application significance.

Key words: forged face detection; deep learning; frequency domain; spatio-temporal fusion; identity consistency

随着数字图像处理和深度学习技术的迅速发展, 人脸生成和伪造技术正日益成熟和普及. 利用生成对抗网络 (GAN)[1]等先进技术, 对真实的人脸图像进行伪造的能力得到了明显的增强, 从而使得伪造出的人脸在质量和逼真度上都达到了很高的水平. 然而, 随着伪造技术的不断进步, 伪造人脸的出现给社会带来了诸多安全隐患, 例如用于欺骗识别系统、网络诈骗等.

近些年, 深度伪造关键词频频出现在新闻中, 引发了公众和政策制定者的讨论. 在国内, 2024 年 2 月, 香港曝出 AI 换脸诈骗案, 犯罪团伙利用跨国公司高管公开视频伪造身份, 骗取香港分公司 2 亿港元; 2023 年 5 月, 内蒙古警方也通报了类似 AI 电信诈骗案件[2]. 而在国外, 伪造欺骗更是频繁出现, 2024 年 1 月, 美国出现“假拜登语音”干扰选举事件, 暴露监管滞后问题; 而 2022 年 3 月乌克兰总统泽连斯基的“投降视频”更引发国际舆论震荡[3]. 对此, 多个国际组织和政府机构开始制定相应的法律法规, 旨在打击伪造人脸带来的违法行为.

在学术界, 伪造人脸检测技术的研究已成为计算机视觉和深度学习领域的重要课题. 许多顶尖会议和期刊相继发表了相关研究, 展现了学者们对此问题的关注和探索. 例如, CVPR、ICCV 等国际会议中, 涉及伪造人脸检测的论文数量逐年增加, 显示出该领域的研究热度不断上升. 图 1、图 2 展示了近年来深度伪造相关文献的关注程度. 图 3 展示了伪造人脸检测技术发展时间线.

伪造人脸检测技术的研究涉及多个领域, 包括计算机视觉、深度学习、图像处理和模式识别等. 这一技术的研究不仅对信息真实性和个人隐私保护具有重要意义, 同时也对司法取证、视频内容审核、媒体报道验证等领域有着深远影响.

本综述根据训练样本的类型, 将现有伪造人脸检测技术分为伪造人脸图像检测和伪造人脸视频检测. 通过回顾近年来发布的伪造人脸检测技术, 评估了不同技术的优劣. 此外, 本文还介绍了常用的评估指标, 并对多个重要数据集进行了系统分析, 包括其特点和适用场景. 同时指出当前文献中的局限性, 例如对抗样本的鲁棒性不足、检测方法对新型伪造技术的适应性差等问题. 基于这些分析, 提出了未来可能的研究方向, 包括跨域检测技术的优化、新算法的探索及模型的可解释性研究.

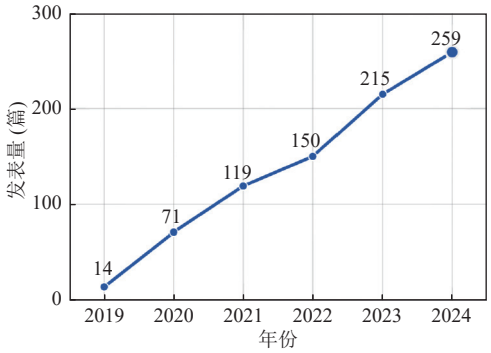

图 1 在中国知网以“深度伪造”为关键字检索的 2019–2024 年的文献数量

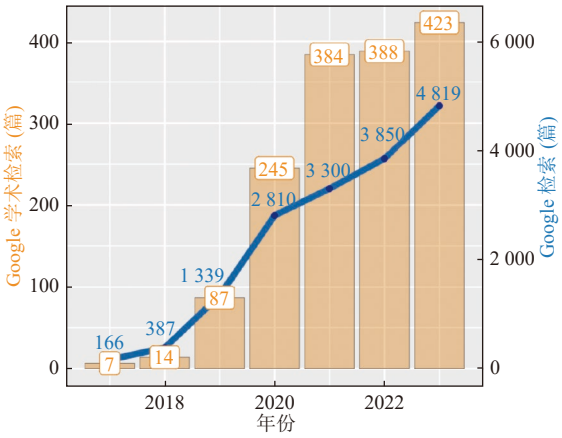

图 2 在 Google 上搜索关键词以及 Google 学术检索关于 DeepFake 的博客和文献数量

1 伪造人脸图像检测技术

伪造人脸图像检测是指利用计算机视觉和机器学

习等技术来识别和检测伪造或合成的人脸图像. 这项技术的目标是识别那些经过合成、修改或伪造的人脸图像, 以区分真实的人脸图像和伪造的人脸图像. 以下是一些常见的用于伪造人脸图像检测的技术和方法.

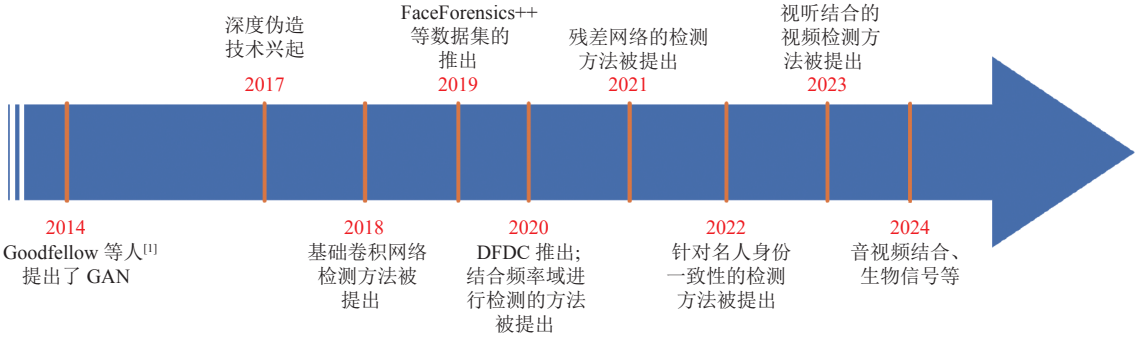

图 3 伪造人脸检测技术发展时间线图

1.1 基于传统机器学习的方法

基于传统机器学习的方法通常涉及使用经典的机器学习算法和特征工程技术来检测数字图像的伪造. 由于深度学习技术的快速发展, 近年来, 单独只使用传统机器学习的方法相对较少, 更多的是通过使用深度学习和传统机器学习相结合的检测方法.

McCloskey 等人[4]分析了当前一种流行的 GAN 实现的生成网络结构, 得出该网络对颜色的处理与真实相机拍摄存在明显不同, 通过构建颜色特征并使用支持向量机 (SVM) 进行分类. 汤桂花等人[5]采用了机器学习与深度学习相结合的方法. 首先使用深度对齐网络 (DAN) 来提取人脸的关键位置, 以获取更精确的面部信息. 接着, 通过使用主成分分析 (PCA) 技术将这些关键点映射到三维空间中, 这样可以减少多余的信息并降低特征的维度. 最后采用支持向量机 (SVM) 的五折交叉验证方法对特征进行了分类.

1.2 基于图像空间域的检测方法

在伪造人脸检测领域, 基于图像空间域的检测技术[6–23]被认为是一种经典的方法, 尽管如此, 它仍然具有很高的有效性. 随着计算机视觉技术的发展, 利用各种算法从不同角度对伪造人脸做出快速准确的识别已经成为一个重要课题. 以图像的像素域为核心的基于图像空间域的检测方法, 主要是通过卷积神经网络 (CNN) 的卷积、池化等手段来提取图像的特征. 这些特征的检测依赖于对图像的颜色、纹理和几何形态的深入分析, 并通过比较真实与伪造图像在这些特征上的区别来做出判断.

朱新同等人[6]首次尝试将 R 与 B 通道通过 Scharr 算子得到的一阶梯度边缘纹理图像与 G 通道通过 Laplacian 算子得到的二阶梯度边缘纹理图像相结合. 通过采用灰度共生矩阵 (GLCM) 进行融合, 他们成功地获取了图像的丰富特征. Zhou 等人[7]描述了一种双流篡改人脸检测技术, 分别用来检测图像补丁之间的低级别不一致和篡改人脸中特定的篡改证据, 通过此方法能够同时学习篡改伪影和隐藏噪声残差特征.

早期的研究主要集中在从图像的空间域中提取广泛特征, 但由于不同区域特征的重要性和贡献度存在差异, 这可能会使模型在有效挖掘具有区分性的局部分类特征时面临挑战. 因此, 一些学者开始从多尺度 (包括局部和全局特征) 和细粒度等方面进行深入研究[8–11]. Zhao 等人[8]从新的视角将深度伪造检测任务重新定义为细粒度分类问题, 提出了一种多注意力的检测框架. 在此框架中, 多个空间注意力模块使得网络能够关注不同局部区域, 同时, 纹理特征增强模块用于突出浅层特征中的微小伪影. Luo 等人[9]观察到, 在去除图像中的噪声和颜色纹理后, 真实区域与篡改区域之间的差异会更加突出. 基于这一发现, 他们提出了一种新方法, 通过分析高频噪声来检测人脸伪造. 这一方法引入了多尺度高频特征提取模块, 以捕获不同尺度的高频噪声, 并应用残差引导空间注意模块, 从而使低级 RGB 特征提取器能够更专注于识别伪造痕迹. Shao 等人[10]首次将适配器技术引入深度伪造检测领域, 它是一种由全局感知瓶颈适配器 (GBA) 和局部感知空间适配器 (LSA) 组成的双层适配器, 能够引导 ViT 的高级语义以双层方式与全局和局部低级特征进行交互, 该技术可以快速调整预训练的 ViT 以用于检测任务.

目前很多伪造人脸检测方法过度依赖于训练设置, 导致模型在未知的伪造类别数据中检测效果不理想, 然而实际场景中的测试图像总是来自未知来源. 为了解决模型泛化性问题, 一些学者开始设计更加通用普适的检测模型[12−23]. 戴昀书等人[12]将模型的学习目标从特定的伪造痕迹特征转变为更普遍的人脸图像局部相似度特征, 提出了一种基于局部相似度异常的深度伪造人脸检测算法. 该方法利用一组局部相似度预测器分别计算 RGB 图像中间层特征图的局部异常, 并引入可学习的空域富模型卷积金字塔 (SRMCP) 来提取多尺度高频噪声特征. 冯才博等人[13]引入了“图像块归属纯净性”和“残差图估计可靠性”的概念, 提出了一种结合纯净图像块比较与可靠残差图估计的多任务人脸伪造检测方法. 该方法通过分析纯净人脸图像块与背景图像块之间的特征差异来识别伪造图像, 并设计了一种距离场加权的残差损失 (DWRLoss), 旨在引导骨干网络更关注伪造边缘附近的真伪图像差异. Tan 等人[14]在对生成图像进行梯度提取时发现, 图像的内容会被过滤, 而与预训练模型的目标任务相关的判别像素得以保留. 基于这一发现, 他们提出了一个简单而新颖的检测框架, 称为梯度学习 (LGrad), 通过使用一种称为转换模型的预训练模型来将图像转换为梯度, 这些梯度作为广义伪像被输入到分类器中, 从而得到一个鲁棒检测器.

综合上述研究论文, 基于图像空间域的检测方法具有模型简单、特征自动提取等特点. 但同时也存在一些局限性, 主要问题包括对抗攻击下的检测性能不足, 以及面对日益复杂的伪造技术时的限制. 未来的改进方向应集中在增强对抗攻击的鲁棒性, 整合多模态信息以提高检测准确性, 并优化深度学习模型的泛化能力和结合传统方法的优势. 表 1 所示为基于图像空间域的检测方法的结果对比.

1.3 基于图像频率域的检测方法

基于图像频率域的检测方法[24−35]是一种用于处理图像、信号和数据的技术, 它通过将数据转换到频率域来揭示数据中的周期性和模式, 从而帮助人们理解和处理这些数据. 在伪造人脸图像检测中, 频率域可以帮助检测图像中的异常模式、纹理不连续性和频谱分布特征, 从而提高对伪造图像的识别能力.

Qian 等人[24]将频率概念应用于人脸伪造检测, 提出了一种新颖的频率面部伪造检测网络 (F3Net). 该网络由两个频率感知分支组成, 其中一个分支利用频率感知图像分解 (FAD) 来识别细微的伪造模式, 另一个分支则从局部频率统计 (LFS) 中提取高级语义, 以区分真实人脸和伪造人脸之间的频率统计差异. Li 等人[25]则提出了一种创新的频率感知判别特征学习框架 (FDFL), 结合了度量学习和自适应频率特征学习技术, 专注于人脸伪造检测. 在模型设计中, 他们引入了新的单中心损失 (SCL) 机制, 关注自然面孔的内部变化, 增强不同类别之间的差异, 从而降低优化难度并获取更多的判别特征. 此外, 他们开发了自适应频率特征生成模块, 通过完全数据驱动的方法挖掘频率线索, 使整个框架能够实现端到端的判别特征学习.

表 1 基于图像空间域的检测方法的结果对比

方法	数据集	AUC检测结果 (%)	跨数据集	跨库AUC检测结果 (%)
Two-stream network[7]	FaceSwap	92.7	—	—
Multi-attentional[8]	FaceForensics++	97.6	Celeb-DF	67.8
High-frequency[9]	FaceForensics++	98.4	Celeb-DF	73.8
DLA[10]	FaceForensics++	99.57	DFDC	72.66
IID[11]	FaceForensics++	99.32	Celeb-DF	82.04
LSP[12]	FaceForensics++	96.26	Celeb-DF	75.82
PuRe[13]	FaceForensics++	98.11	—	—
LGrad[14]	FaceForensics++	99.1	Celeb-DF	78.23
NPR[15]	FaceForensics++	99.8	Celeb-DF	80.1
CGS[16]	FaceForensics++	99.24	DFDC	81.65
AUNet[18]	FaceForensics++	99.89	DFDC	73.82
LAA-Net[20]	FaceForensics++	99.96	—	—
TUG[22]	FaceForensics++	99.50	Celeb-DF	76.75

上述方法主要侧重于图像的频域特征进行伪造检测, 未考虑其与空域特征的关系. 因此, 通过结合频域和空域特征, 可以弥补各自的不足, 进而增强模型的泛化能力[18−21]. 马欣等人[29]利用了大部分人脸伪造技术都可能导致图像频域发生显著变化的特性, 提出了一种基于多尺度 Transformer 融合多域信息的伪造人脸检测方法. 该方法通过整合图像的频域特征和 RGB 域特征来检测伪造的人脸, 从而补充图像 RGB 域中无法感知的伪影信息, 提高了检测的准确性. Liu 等人[30]提出了一种创新的空间-相位浅学习 (SPSL) 技术, 利用空间和频率数据. 这项技术通过结合空间图像与相位谱, 旨在捕捉人脸伪造中的上采样伪像, 从而提升人脸伪造检测的可迁移性. 同时, 他们还从理论上探讨了相位

谱的实用性.

综合上述研究论文, 基于图像频率域的检测方法具备频谱特性和傅立叶变换等优点, 这使得模型能够从图像的频域信息中提取关键特征, 从而有助于区分真实和伪造的人脸图像. 尽管如此, 仍然有一些明显的限制, 例如受限于频域特性的描述能力和对复杂攻击的稳健性. 在未来, 可以思考如何优化频域特征的抽取技术, 并与其他技术领域相结合, 以增强检测的效果. 表 2 展示了基于图像频率域的检测方法的结果比较.

表 2 基于图像频率域的检测方法的结果对比

方法	数据集	AUC检测结果 (%)	跨数据集	跨库AUC检测结果 (%)
F3-Net[24]	FaceForensics++	95.84	—	—
FDFL[25]	FaceForensics++	92.4	—	—
GFS[26]	FaceForensics++	97.23	Celeb-DF	71.21
双流频率域[27]	FaceForensics++	97.60	Celeb-DF	72.98
FrePGAN[28]	FaceForensics++	99.11	DFDC	74.71
MST[29]	FaceForensics++	99.2	Celeb-DF	80.28
SPSL[30]	FaceForensics++	96.91	Celeb-DF	76.88
SFDG[31]	FaceForensics++	98.19	DFDC	73.64
CD-Net[34]	FaceForensics++	98.5	Celeb-DF	80.67
M2TR[35]	FaceForensics++	99.51	Celeb-DF	68.2

1.4 人脸伪造主动防御方法

人脸伪造主动防御技术是应对深度伪造 (如通过生成对抗网络 (GAN) 制造的伪造人脸) 与对抗攻击的一种新兴技术. 传统的伪造人脸检测方法通常依赖于静态的图像分析或特征提取, 主要是通过识别伪造痕迹来进行判断. 然而, 随着深度伪造技术的不断发展, 攻击者生成的伪造样本越来越难以区分, 传统的防御方法面临较大挑战. 深度伪造主动防御技术通过实时、动态地响应伪造攻击, 增强模型的鲁棒性, 进而提升对伪造人脸的检测效果. 该技术通常结合了对抗训练、特征增强、跨域迁移学习等手段, 旨在通过持续学习和适应, 自动识别和防御不同类型的深度伪造攻击.

一些相关的研究已提出了不同的深度伪造主动防御方法. Dong 等人[36]探讨了人脸交换模型 DeepFake 生成对抗示例的问题, 从对抗性可转移性和潜在表征的角度提出了针对 DeepFake 自编码器的 3 种新型对抗性攻击, 并成功证明了针对 DeepFake 的对抗性示例在参考和非参考图像质量评估方面的有效性. Wang 等人[37]将对抗性训练引入到深度假检测模型的训练过程中, 提出了一种新的对抗性训练方法, 通过引入逐像素高斯模糊来模糊这些伪像, 使模型被迫学习更多的判别和可概括的特征. Chen 等人[38]提出了一种可以在未知场景下很好泛化的深度伪造检测方法, 利用合成器和对抗训练框架来动态生成伪造, 通过训练来识别这些生成的伪造品, 使网络能够学习到更鲁棒的特征表示, 从而生成一个更通用的深度伪造探测器. 王喻等人[39]为了解决人脸图像未经授权情况下被伪造或篡改的问题, 提出一种基于注意力掩码与特征提取的人脸伪造主动防御模型, 旨在采取攻击性措施, 向图像中加入可干扰伪造模型的对抗样本, 从源头上预防图像被伪造, 同时提高被保护图像的视觉质量.

尽管深度伪造主动防御技术在伪造人脸检测中展现了巨大的潜力, 但也面临一些挑战. 首先, 这类方法往往需要较高的计算资源, 特别是在实时检测和大规模应用时, 可能会导致延迟较高. 其次, 随着对抗攻击技术的不断进化, 生成的伪造样本也越来越难以识别, 防御技术必须不断更新以适应新的攻击手段. 最后, 跨域适应性问题仍然是一个难点, 现有防御方法可能在不同的场景和数据集之间表现不一致. 因此, 虽然深度伪造主动防御技术具有提高鲁棒性的优势, 但在效率、适应性和可扩展性方面仍需进一步优化.

1.5 针对特定应用场景的检测方法

随着伪造人脸技术日益成熟和普及, 有学者开始尝试研究一些针对特定应用场景的伪造人脸检测技术, 例如在需要确保身份一致性或者需要定位伪造区域的特定应用场景. 其中基于身份一致性的方法依赖于人脸识别技术, 通过比对人脸特征来验证图像或视频中的身份真实性. 而伪造人脸区域定位则聚焦于分析和检测图像中可能存在的伪造迹象, 如不一致的眼睛或嘴巴区域. 这些方法不仅有助于提高检测的准确率, 还能为特定场景下的伪造人脸检测提供更精确和可靠的解决方案.

基于身份一致性的检测方法[40–43]与经典的基于图像纹理特征的检测方法在本质上存在差异, 主要是因为它并不试图识别图像中的伪影, 相反, 它关注的是嫌疑人图像中的身份是否真实. Dong 等人[40]提出了一种基于身份的深度伪造检测技术, 该技术使用疑似图像和目标的身份信息作为输入数据, 最终得出一个判定疑似图像身份是否与目标身份一致的结论. 借助 Transformer 在图像领域的兴起, Dong 等人[43]在之前的研究

基础上, 提出了一种名为身份一致性转换器 (identity consistency Transformer) 的新方法. 这种方法主要关注高级语义, 尤其是身份信息, 并通过在面部的内部和外部检测身份的不一致性来识别可疑的人脸. 经过实验验证, 身份一致性转换器能够轻松地利用附加的身份数据 (在这些数据可用的情况下) 进行身份增强, 这使其成为检测涉及名人面部伪造的理想工具.

现有的大多数伪造人脸检测方法本质上是二分类问题, 输出真实伪造预测结果, 然而对于伪造区域没有明确的定位标注, 有学者尝试在伪造人脸检测方法中加入区域定位技术, 来实现伪造区域的定位[44–49]. Guo 等人[48]在伪造人脸检测模型的最后加入了定位模块, 该定位模块采用了自注意力机制, 借助基于深度度量学习的目标生成伪造掩码, 提高真实像素与伪造像素的分离程度, 将检测模块中的细粒度分类特征作为改进定位的先验条件, 得出准确的定位区域. Fabrizio 等人[49]将伪造定位任务视为一个监督二值分割问题, 通过将噪声伪影信息与 RGB 图像的高级特征相结合, 提出了 CMX 架构, 它是一种跨模态融合框架, 利用语义分割的方法, 在具有共享编码器结构的两个并行分支上提取输入图像和噪声打印的特征.

综合上述研究论文, 特定应用场景下的检测方法具有明显的特性, 它在特定应用场景中比一般的检测方法更为适用, 并且效果更为出色. 但是, 通过对比近些年的研究论文, 可以观察到, 这种方法在处理跨数据集时的检测效果并不总是令人满意, 这可能是由于该方法过分依赖数据集的某一特性, 导致模型的泛化性能下降. 表 3 所示为基于针对特定应用场景的检测方法的结果对比.

表 3 基于针对特定应用场景的检测方法的结果对比

方法	数据集	AUC检测结果 (%)	跨数据集	跨库AUC检测结果 (%)
TI2Net[41]	FaceForensics++	98.56	—	—
AVD[42]	FaceForensics++	99.95	—	—
ICT[43]	DFDC	90.66	—	—
PWLA[45]	FaceForensics++	92.44	—	—
UADFV[46]	DFFD (自建)	98.4	Celeb-DF	71.2
DLFMNET[47]	FaceForensics++	97.42	Celeb-DF	71.55
IFDL[48]	FaceForensics++	98.11	—	—

2 伪造人脸视频检测技术

伪造人脸视频检测是指通过分析视频中的一系列人脸图像, 借助计算机视觉和深度学习技术, 来检测视频中的人脸图像是否存在编辑、合成或其他形式的篡改. 该技术面临着大数据量、复杂度高、实时性要求等挑战. 以下是一些常见的用于伪造人脸视频检测的技术和方法.

2.1 基于时空特征融合的检测方法

基于时空融合特征的检测方法[50–63]在伪造人脸视频检测中是一种先进的技术, 通过结合视频中的时间序列信息和每一帧的空间域特征, 能够有效提高对伪造视频的检测准确性和鲁棒性. 这种方法利用深度学习模型或特定算法从视频中提取综合的时空特征, 并通过特征融合技术增强检测模型的能力, 使其能够更全面地分析和识别潜在的伪造人脸视频.

在自然语言处理领域, 循环神经网络 (RNN) 广泛用于提取上下文的语义关系, 而在视频处理方面, 它则被应用于分析相邻帧之间的联系. 近年来, 基于 RNN 的时空融合特征检测方法在检测人脸深度伪造视频方面逐渐发挥了重要作用[50–52]. Sun 等人[50]提出了一种高效且鲁棒性强的框架, 称为 LRNet. 该框架通过对精确几何特征进行时间建模来识别深度伪造视频, 并设计了一个全新的校正模块, 以提高几何特征的准确性, 从而增强其区分能力. 该模型基于双流循环神经网络 (RNN) 来提取时间特征. 朱春陶等人[51]提出了一种利用多域时序特征挖掘技术来检测伪造人脸的新方法. 该方法是基于视频在空域和频域中的时序特性, 利用人脸统计特征数据的一致性和时间动作趋势的一致性, 对被篡改的特征进行了区分和增强. 在这个模型中, 采用经过优化的长短记忆网络 (LSTM) 作为核心网络, 以深入探索帧之间的时间序列特性.

与利用 RNN 提取时间序列特征不同, 基于卷积的时空融合特征检测更关注卷积核的设计. 常用的方法是为视频帧构建具有时间维度的卷积核, 旨在更有效地捕捉帧与帧之间的连续性和相关性等关键特征[53–62]. Zhang 等人[53]利用视频帧之间可能存在的不一致线索, 提出时间 Dropout 三维卷积神经网络 (TD3DCNN) 方法来识别深度伪造视频. 在此方法中, 他们将固定长度的帧输入三维卷积神经网络 (3DCNN), 以提取不同尺度的特征并进行真伪鉴别. 在此基础上, Zhang 等人[54]后来发现了视频级检测器的关键是充分利用深度假视

频中不同帧间局部人脸区域分布的时空不一致性, 由此提出了基于时空 Dropout-Transformer 的深度假视频检测方法如图 4 所示, 通过使用时空 Dropout 操作, 充分挖掘斑块级时空线索, 进一步增强模型的鲁棒性和泛化能力. 程燕[57]以光流法为基础, 提出了一种基于关键帧与时空特征融合的人脸伪造检测方法. 首先, 采用加权光流能量阈值分析法筛选出视频中能量较大的关键帧, 将关键帧的光流和 LBP 纹理特征进行融合, 构成具有时间和空间特性的融合特征图, 经过增强处理后输入 CNN 模型进行学习.

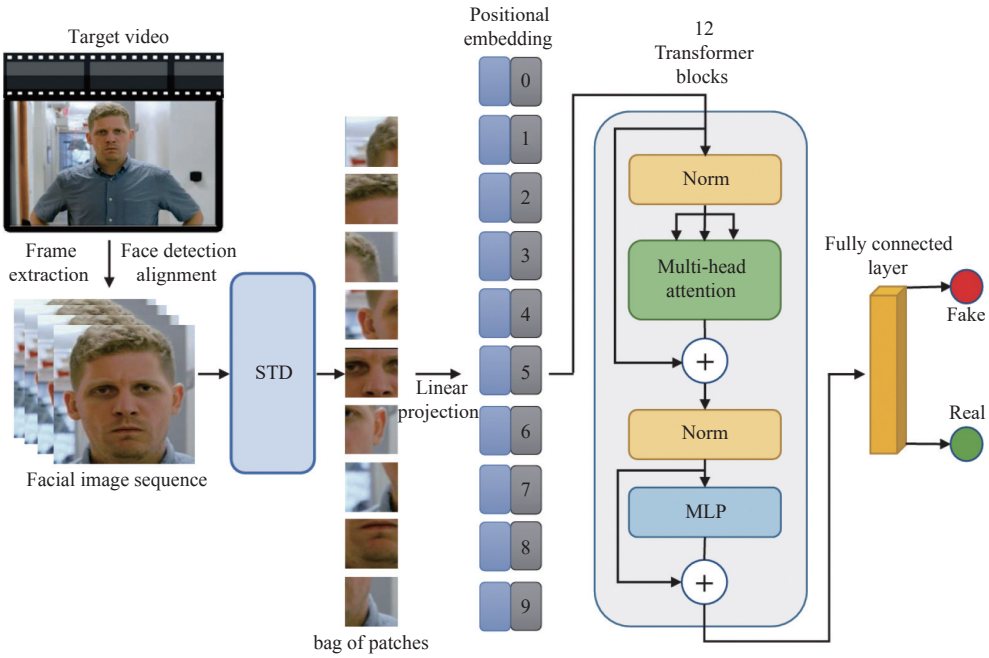

图 4 基于时空 Dropout-Transformer 的深度假视频检测方法

综合上述研究论文, 基于时空特征融合的检测方法具有能够有效捕捉视频中的动态和时序信息的特点, 能够帮助模型分析视频序列中帧间的信息, 从动态和时序角度识别真实和伪造人脸图像. 尽管时空特征融合的方法在伪造人脸检测中具有一定优势, 但仍存在一些局限性, 如对高计算成本、时空对齐、模型复杂度与泛化能力的平衡等问题, 未来的改进方向包括优化时空对齐方法、提升计算效率、引入多模态信息、增强模型的泛化能力以及集成深度学习与传统方法, 以提高检测系统的准确性和鲁棒性. 表 4 所示为基于时空特征融合的检测方法的结果对比.

2.2 基于生理特征的检测方法

目前, 基于生理特征的检测方法[64–73]依然处于研究阶段, 但已经出现了一些相关的研究方向和论文. 这些方法试图利用人类的生理特征, 如眼球运动、面部表情、呼吸等, 来区分真实的人脸视频和伪造的人脸视频.

Li 等人[64]基于视频中人物眨眼这一生理信号, 提出一种将 CNN 和递归神经网络相结合的深度神经网络模型, 称为长期递归 CNN (LRCN), 通过考虑之前的时间知识来区分睁眼和闭眼状态进行检测. 还有一些学者尝试从人嘴部的运动进行检测. Alexandros 等人[65]提出了 LipForensics 方法, 该方法旨在解决嘴巴运动中的高级语义不规则性问题. 首先, 他们通过预先训练一个时空网络来进行视觉语音识别 (唇读), 从而学习与自然口腔运动相关的丰富内部表征. 然后, 在真实和伪造数据的固定嘴巴嵌入上, 时间网络进行微调, 以便检测基于嘴巴运动的假视频. Agarwal 等人[66]发现在伪造视频中, 人物嘴部可能与音频很好地同步, 但耳朵运动的动态将与嘴部和下巴运动分离这一现象, 在原有的方法上加入耳朵运动区域模块, 使得模型同时关注人嘴部和耳朵的运动情况来进行检测.

近年来, 有学者通过结合生物等领域技术如远程光电容积脉搏波 (rPPG), 来获取更为细粒度的特征. Javier 等人考虑使用远程光电容积脉搏波 (rPPG) 与心率相关的信息, 提出了伪造人脸视频检测器 Deep-

FakesON-Phys, 该模型模仿传统手工 rPPG 技术行为的视频中提取时空信息, 而生理特征是通过用户面部颜色的变化来提取的, 该变化是由于血液中氧浓度的变化引起的, 从而达到检测目的[69].

表 4 基于时空特征融合的检测方法的结果对比

方法	数据集	AUC检测结果 (%)	跨数据集	跨库AUC检测结果 (%)
LRNet[50]	FaceForensics++	97.3	—	—
MDT[51]	FaceForensics++	99.26	Celeb-DF	76.7
TD-3DCNN[53]	Celeb-DF	88.83	FaceForensics++	68.88
STDT[54]	Celeb-DF	97.2	FaceForensics++	74.62
TALL[55]	FaceForensics++	98.65	DFDC	76.78
MKFFT[56]	Celeb-DF	98.64	—	—
STFF[57]	FaceForensics++	92.1	Celeb-DF	72.3
RATI[59]	FaceForensics++	99.64	Celeb-DF	76.50
FInfer[60]	Celeb-DF	93.30	WildDeepfake	69.46
DIL[61]	DFDC	92.7	Celeb-DF	77.65
ESLF[62]	FaceForensics++	99.1	—	—
FADE[63]	FaceForensics++	99.52	Celeb-DF	77.46

综合上述研究论文, 基于生理特征的检测方法具有一些技术特点, 如生物特征采集设备、特征提取算法、生理信号处理等, 通过分析不易伪造的生理特征, 提高人脸真伪判别的准确性和可靠性. 但同时也带来了一些问题, 如设备成本高、使用限制等, 对于技术的落地应用有一定的影响, 未来可以探索改进生理特征检测算法的方法, 降低成本、提高模型的实用性. 表 5 所示为基于生理特征的检测方法的结果对比.

表 5 基于生理特征的检测方法的结果对比

方法	数据集	AUC检测结果 (%)	跨数据集	跨库AUC检测结果 (%)
LipForensics[65]	FaceForensics++	97.1	Celeb-df	82.4
RealForensics[67]	FaceForensics++	99.2	DFDC	75.9
DeepFakesON-Phys[68]	FaceForensics++	98.2	—	—
Multi-phoneme[70]	DFDC	91.6	NLFD (自建)	64.05
L-rPPG[71]	FaceForensics++	99.28	—	—
VRPS[73]	DFDC	93.1	—	—

2.3 基于视听结合的检测方法

基于视听结合的伪检测方法[74–80]旨在利用音频和视频之间的关联信息, 以提高伪造人脸视频检测的准确性和鲁棒性. 这种方法通过分析音频和视频在时空一致性、情绪表达一致性等方面的关系, 来识别和区分真实和伪造的人脸视频. 深度学习模型常被用来融合和分析多模态数据, 以帮助区分伪造视频中可能存在的不一致性.

Chao 等人[74]通过实验发现经过处理的视频通常包含视觉和音频信号之间微妙的不一致, 提出了一种基于自监督异常检测的视频篡改检测方法如图 5 所示, 将检测被操纵视频的问题表述为异常检测问题. 具体通过设计视觉编码器和音频编码器获得视频片段和音频片段的特征信息, 再将两者进行融合输入到提出的概率函数中计算两者在时间上同时出现的可能性.

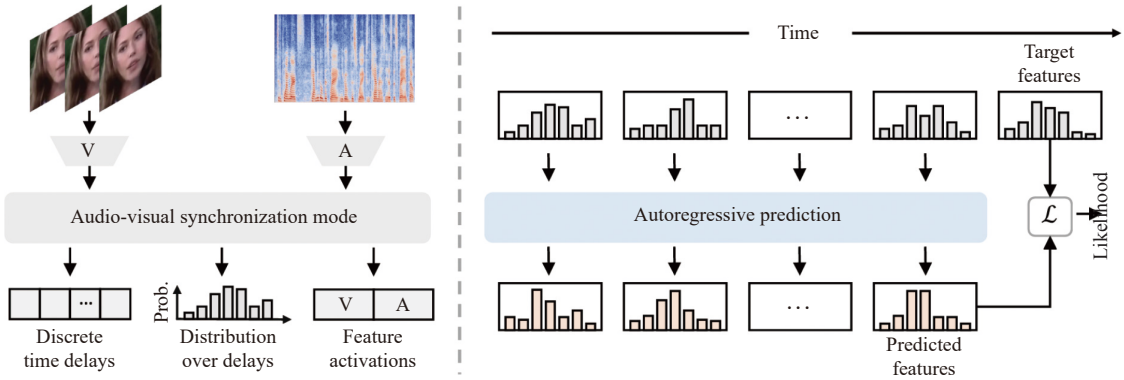

图 5 基于自监督异常检测的视频篡改检测方法

Trevine 等人[75]提出了视听特征融合方法 (AVFF), 这是一种两阶段的跨模态学习方法, 可以明确捕获音频和视觉模态之间的对应关系, 以改进深度伪造检测, 在第 1 阶段通过对真实视频的自我监督进行表征学习, 并引入一种新的视听互补掩蔽和特征融合策略, 以捕捉内在的视听对应关系. 在第 2 阶段, 将学习到的表示输入到卷积网络中进行调整, 通过对真实和虚假视频的监督学习来进行深度分类.

基于视听结合的检测方法利用视频和音频多模态信息融合, 提升伪造人脸视频检测的准确性和鲁棒性,

尤其在对抗新型伪造技术时具有优势. 然而, 目前面临数据获取困难、多模态信息融合复杂以及计算效率低等挑战. 未来的改进方向包括扩展真实数据集、优化深度学习模型、加强跨学科合作, 以及提高实时性和算法效率. 表 6 所示为基于视听结合的检测方法的结果对比. 表 7 中总结了伪造人脸检测的主要方法.

3 数据集

为训练和评估伪造人脸检测算法, 研究人员和开发者通常会使用专门设计的数据集. 这些数据集包含了各种类型的伪造人脸图像和视频, 以及真实的人脸图像和视频, 以便为算法提供充分的训练和测试数据. 如表 8 是常用的伪造人脸检测数据集, 具体介绍如下.

表 6 基于视听结合的检测方法的结果对比

方法	数据集	AUC检测结果 (%)	跨数据集	跨库AUC检测结果 (%)
AVAD[74]	FakeAVCeleb	91.1	—	—
AVFF[75]	FakeAVCeleb	98.5	KoDF	93.1
Multi-phoneme[76]	FakeAVCeleb	99.2	FaceForensics++	80.75
VFD[78]	DFDC	85.13	—	—
AVoiD-DF[79]	DFDC	96.6	—	—

表 7 伪造人脸检测方法总结

方法	输入	特点	局限性
图像空间域	图像	直接在像素级别进行分析, 实时处理性能较好	对图像质量依赖高, 易受光照、角度变化影响
图像频率域	图像	提取频率特征, 抗干扰能力强可捕捉细微伪造痕迹	计算复杂度高, 需要较高的频率域分析经验
身份一致性	图像	利用人物身份信息验证	针对特定人物场景使用
时空特征融合	视频关键帧	结合时间序列数据, 增加检测的可靠性	处理延迟较高, 需要大量时序数据支持
生理特征	视频关键帧	基于面部生理特征 (如眼动、微表情)进行分析	实际应用场景受限, 可能需要额外的传感器支持
音视频结合	视频关键帧	结合音频特征, 提升伪造检测的准确性	对环境噪声敏感, 需要多模态数据处理

表 8 伪造人脸检测常用的数据集

数据集	规模	特性	适用场景	局限性
FaceForensics++	fake: 5 000 real: 1 000	包含多种伪造方法	人脸伪造检测算法的训练与评估	数据集主要集中于特定的伪造技术, 对其他伪造技术的泛化能力有限
Celeb-DF v2	fake: 5 639 real: 590	包含高质量的DeepFake视频, 专注于名人图像	深度伪造检测研究, 尤其是名人视频分析	由于主要使用名人图像, 可能在泛化到普通人脸图像时表现不佳
DFDC	fake: 104 500 real: 23 564	数量庞大, 适用于广泛的研究	提高对不同伪造类型的检测能力	数据集的复杂性导致训练时间较长
FFIW-10K	fake: 10 000 real: 10 000	来源于社交媒体, 提供真实与伪造人脸的配对, 便于对比分析	社交媒体内容的真实性验证	对于特定背景或环境的泛化能力有限
DeepFake MNIST+	fake:100 000 real:10 000	在MNIST数据集的基础上, 加入了多种变形和伪造特征	训练模型以提高其对抗性, 增强其对伪造数据的抵抗力	尽管扩展了 MNIST 数据集, 但仍然局限于手写数字, 不能广泛应用于其他图像类型

(1) FaceForensics++: FaceForensics++数据集[81]是一个用于伪造人脸检测研究的关键资源, 由 Technische Universität München 的研究人员开发. 该数据集包含超过 1 000 个由 4 种不同类型的伪造技术生成的人脸视频, 这 4 种技术分别是 DeepFakes、Face2Face、FaceSwap 和 NeuralTextures. 这些视频涵盖了多个演员和不同姿势的面部表情, 使得数据集具有一定的多样性和挑战性.

FaceForensics++的独特之处在于其合成的视频看起来非常逼真, 这使得该数据集在模拟真实世界中可能出现的伪造情况方面具有很高的可信度. 这种真实性为研究人员提供了一个更接近实际情况的测试平台, 有助于他们开发和评估更有效的伪造人脸检测算法. 作为一个用于伪造人脸检测研究的基准数据集, FaceForensics++已经被广泛应用于训练和评估各种伪造人脸检测算法. 由于其规模和多样性, 该数据集已成为该领域的标准之一, 为研究人员提供了一个基准测试和对比实验的基础. 通过使用这个数据集, 研究人员可以更好地了解各种伪造技术的特点, 推动伪造人脸检测技术的发展.

(3) DFDC: DFDC 数据集[83]是 DeepFake detection challenge (DFDC) 的一部分, 由 Facebook 和多个合作伙伴共同创建和维护. 这个数据集是为了推动深度伪造视频 (DeepFake) 检测算法的研究而设计的. DFDC 数据集包含了大量的合成视频和原始视频, 其中合成

视频是使用深度学习技术生成的, 而原始视频则是真实录制的.

(2) Celeb-DF v2: Celeb-DF v2 是一个用于伪造人脸检测研究的重要数据集[82], 由南洋理工大学和新加坡科技设计大学的研究人员联合开发. 这个数据集包含大量的合成人脸视频, 覆盖 DeepFakes、Face2Face、FaceSwap 和 NeuralTextures 等多种伪造技术, 以及原始真实视频作为对照. Celeb-DF v2 的结构包括了源视频、伪造视频、人脸检测标签等内容, 为研究人员提供了一个重要的实验基准, 有助于他们开发和评估伪造人脸检测算法. 通过使用这个数据集, 研究人员可以更好地了解伪造视频的特点, 推动伪造人脸检测技术的发展.

DFDC 数据集的目的是建立一个用于测试和评估深度伪造视频检测算法的标准基准. 这个数据集覆盖了不同人物、不同场景和不同质量的视频, 从而更好地模拟了真实的深度伪造视频传播环境. 研究人员可以使用 DFDC 数据集来训练、验证和评估他们的深度伪造视频检测算法, 以便更好地识别和区分深度伪造视频和真实视频. DFDC 数据集的开发旨在促进对深度伪造视频检测技术的研究和发展, 以应对深度伪造视频可能带来的信息不实和隐私风险. 这个数据集的发布为相关研究提供了一个标准化的实验基准, 有助于推动对深度伪造视频检测算法的改进和创新.

(4) FFIW-10K: FFIW-10K 是一个专为人脸伪造检测设计的数据集[84], 包含 1 万个高保真的处理视频, 每个视频涉及多个个体, 平均长度为 12 s, 总共有 33 个小时的视频. 同时数据集提供视频和人脸级别的注释, 允许对伪造分类和定位任务进行基准测试. 此外, 为了使研究进入更自然的环境, FFIW10K 提供了两个基准设置. 在第 1 种设置中, 基准测试方法可以利用表面层面的监督. 然而, 在第 2 种情况下, 只允许在训练期间访问视频级别的标签. 这使得该任务在学术和实践上都更有价值.

(5) DeepFake MNIST+: DeepFake MNIST+是一个人脸动画视频数据集[85], 是作为对 MNIST 数据集广泛使用的响应而开发的, 但用于不同的 DeepFake 检测问题. 该数据集包含 10 种不同动作的 10 000 个面部动画视频, 加上 10 000 个真实面部视频, 以实现监督检测器训练. 这些虚假的面部动画具有更高的保真度, 能够欺骗市场上流行的活力检测器.

4 评估指标

在伪造人脸检测领域, 选择合适的评估指标至关重要. 评估指标能够客观评价不同算法或方法的性能, 帮助研究者全面了解各种技术的优劣之处. 目前研究者们普遍将伪造人脸检测问题看作成图像的一个二分类问题, 因此本文整理了二分类问题中所使用的评估指标.

(1) 准确率 (*Accuracy*) 是指分类器正确分类的样本所占的比例, 通常用以下公式来计算:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

其中, *TP* (真正例) 代表被分类器准确识别为正类的样本数量; *TN* (真负例) 指被正确分类为负类的样本数量; *FP* (假正例) 则是误判为正类的样本数量; 而 *FN* (假负例) 表示被错误地识别为负类的样本数量.

准确率是对分类器整体性能的一个评估指标. 然而, 在不平衡数据集中, 准确率可能并不是最好的评估指标, 因为在某些情况下, 即使准确率很高, 分类器也可能存在问题. 因此, 在一些情况下需要结合其他指标来评估分类器的性能, 比如精确率 (*Precision*)、召回率 (*Recall*)、*F*1 分数、ROC 曲线和 AUC 值等.

(2) 精确率和召回率: 精确率 (*Precision*) 用于量化分类器所识别出的正例样本中, 真实正例的数量. 精确率的计算公式如下:

$$Precision = TP/(TP + FP) \tag{2}$$

其中, *TP* (true positive) 代表被分类器正确识别为正例的样本数量, 而 *FP* (false positive) 则代表被误判为正例的样本数量.

召回率 (*Recall*) 是一个衡量标准, 用于确定在所有真实的正例案例中, 有多少被分类器准确地标识为正例. 召回率的计算公式如下:

$$Recall = TP/(TP + FN) \tag{3}$$

其中, *FN* (false negative) 表示被错误地识别为负例的样本数.

精确率和召回率通常存在竞争关系, 提升一个指标可能会导致另一个指标的下降. 因此, 在评估分类器的表现时, 需要综合考虑这两个指标, 通常可以利用 F1 分数来进行全面评估.

(3) ROC 曲线是一条以假阳性率 (FPR) 为横轴, 真阳性率 (TPR) 为纵轴的曲线, 展示了分类器在不同阈值下的特性. 这条曲线能够直观地反映分类器在各个

阈值下的表现, 通常, 越接近左上角的曲线表示分类器性能越优. AUC 值则是 ROC 曲线下方的面积, 值越接近 1, 说明分类器的整体性能越好.

AUC 值代表 ROC 曲线下的区域面积, 其数值范围是 0–1. AUC 的数值越高, 意味着分类器的表现越出色, 一般来说, 当 AUC 值超过 0.5 时, 意味着分类器具备了良好的预测功能.

通过 ROC 曲线和 AUC 值, 可以直观地比较不同分类器的性能, 选择最佳的阈值或者最佳的分类器. 特别是在处理不平衡数据集时, ROC 曲线和 AUC 值能够更好地反映分类器的整体性能, 因此在深度伪造视频检测中也是常用的评估指标之一.

5 应用和未来展望

5.1 应用领域

伪造人脸识别技术在现代社会中的重要性日益凸显, 它的应用范围广泛, 包括但不限于金融安全、社交媒体以及司法系统等多个方面. 本节将探讨伪造人脸检测技术在不同应用领域中的重要性和实际应用情况, 以帮助读者全面了解该技术的应用范围和影响.

(1) 社交媒体平台: 随着人工智能技术的发展, 越来越多的社交媒体平台开始使用伪造人脸检测技术来防止用户上传虚假的人脸照片、视频或者进行人脸交易等行为, 以确保平台上的内容和用户的真实性.

(2) 金融行业: 在金融领域, 伪造人脸检测被用于身份验证和安全监测. 银行、支付机构等金融机构可以利用伪造人脸检测技术来识别和预防欺诈行为, 保护客户资金安全.

(3) 安防监控: 在安防领域, 伪造人脸检测可用于监控摄像头系统, 以检测和防范恶意入侵、非法通行等行为. 这有助于提高公共安全和保护个人隐私.

(4) 身份识别和管理: 伪造人脸检测技术在身份识别和管理方面也有着广泛的应用, 例如用于边境检查、机场安检、企业员工考勤等场景, 以确保身份信息的准确性和安全性.

(5) 法律与司法领域: 在刑侦和司法领域, 伪造人脸检测技术可以用于刑事案件调查和取证, 帮助警方和法院识别和鉴定嫌疑人或证人提供的人脸信息的真实性.

5.2 现有挑战

尽管在深度伪造检测方面取得了重大进展, 但目前的文献中仍然存在一些局限性, 概括为以下几点.

(1) 泛化性不足

目前的伪造人脸检测模型通常在特定的数据集上表现良好, 但在面对不同伪造技术或新的数据集时, 性能往往急剧下降. 原因在于大多数现有模型在训练过程中仅依赖于单一类型的伪造数据 (例如, 基于 GAN 生成的图像或视频). 随着伪造技术的不断演进, 新的生成方法 (如自监督学习生成的伪造图像) 不断出现, 现有模型难以适应这些新型伪造技术, 导致检测效果受限. 缺乏多样化的数据集和通用的特征学习使得模型难以做到跨伪造类型的良好泛化.

(2) 对抗性攻击的脆弱性

深度伪造检测系统的另一个显著问题是其对对抗性攻击的脆弱性. 通过对输入图像进行微小的扰动, 攻击者可以轻松让伪造图像绕过检测器, 导致检测模型错误地将伪造图像判定为真实图像. 许多深度伪造检测方法依赖于深度学习模型, 如卷积神经网络 (CNN) 或其他复杂模型, 这些模型在面对对抗扰动时表现出较强的敏感性. 随着生成对抗网络 (GAN) 技术的不断进步, 伪造图像的质量和真实性不断提升, 传统的检测方法逐渐无法应对这种“博弈式”的攻防关系, 导致检测鲁棒性不足.

(3) 计算成本与实时性问题

随着深度伪造检测技术的不断发展, 模型的计算需求也急剧上升. 大多数当前的深度伪造检测方法基于复杂的卷积神经网络 (CNN) 或 Transformer 架构, 这些模型拥有庞大的参数量和计算复杂度, 需要大量的计算资源和存储空间. 这使得它们难以在计算能力有限的设备 (如智能手机、物联网设备等) 上部署, 并且在实时检测场景中, 延迟较高的推理时间影响了应用的实用性. 尤其是在视频流处理或大规模在线监控系统中, 实时性成为一个不可忽视的问题.

(4) 缺乏多模态数据融合

当前大多数伪造人脸检测技术主要依赖单一模态数据 (如图像或视频), 忽略了多模态信息的潜力. 在实际应用中, 单一模态往往无法全面捕捉伪造的细节. 例如, 视频中的音频信息和图像中的视觉线索可以提供互补的伪造识别信息, 帮助提升检测的准确性. 多模态数据融合的缺乏, 使得现有检测技术在面对复杂场景时, 无法充分利用来自不同模态的信息, 导致检测精度和鲁棒性受到限制.

(5) 适应性差, 对新型伪造技术无能为力

伪造人脸技术的快速发展导致了新型伪造方法层出不穷, 如基于自监督学习或其他创新算法的伪造生成技术. 这些新兴方法通常能生成质量更高、更加逼真的伪造图像, 而现有的检测模型多依赖于对特定伪造生成管道的特征学习, 缺乏对新型伪造技术的适应能力. 现有模型往往无法应对这些新的生成方法, 导致它们在实际应用中的检测效果大打折扣. 伪造技术的不断变化要求检测模型具有更强的适应性和灵活性, 而当前技术在这一方面的能力仍然有限.

5.3 未来研究方向

为了解决目前存在的问题并提高深度伪造检测的有效性, 提出了几个未来的研究方向.

(1) 提升检测模型的泛化能力

未来的研究需要专注于提升伪造人脸检测模型的泛化能力. 当前的检测模型在特定数据集上表现良好, 但在面对不同伪造技术或未知的数据集时, 往往表现不佳. 为了解决这一问题, 未来研究应致力于构建更加多样化的训练数据集, 涵盖不同类型的伪造方法、场景和生成管道. 此外, 迁移学习和元学习等方法也应被广泛应用, 通过这些技术, 模型可以在面对新型伪造图像时, 快速适应并保持较高的检测精度.

(2) 强化对抗性攻击防御机制

随着深度伪造技术与对抗性攻击手段的不断发展, 未来的研究需要进一步强化伪造人脸检测模型的对抗性防御能力. 目前的检测方法普遍存在对抗攻击脆弱的问题, 因此研究人员应探索更为有效的对抗训练方法, 提升模型对微小扰动的鲁棒性. 此外, 探索基于对抗样本的生成和检测技术, 结合动态防御策略, 也将是未来防止对抗性攻击的关键方向之一. 通过增强模型在对抗环境中的鲁棒性, 可以大大提升深度伪造检测技术的实用性.

(3) 轻量化与高效性研究

在实际应用中, 尤其是在边缘计算设备上, 伪造人脸检测模型的计算资源消耗仍然是一个亟待解决的问题. 未来的研究应着重于开发轻量化和高效的深度学习模型, 这些模型能够在计算资源有限的设备上运行, 同时保证高精度的检测效果. 例如, 使用模型压缩技术 (如剪枝、量化、蒸馏等) 来减少模型的参数量和计算量, 或设计专为低功耗设备优化的架构 (如 MobileNet、EfficientNet 等). 此外, 硬件加速 (如使用 GPU、TPU 等) 和分布式计算等技术, 也将在提升实时检测能力和降低计算延迟方面发挥重要作用.

(4) 多模态数据融合

多模态数据融合是提升伪造人脸检测性能的重要研究方向. 未来的研究应探索如何有效地将图像、视频、音频、深度信息等多种模态数据进行融合, 利用多源信息协同提升检测准确性. 图像和视频中的伪造线索可能在不同模态之间有所差异, 而音频信息、面部表情、语音同步等特征也能够为伪造检测提供重要线索. 研究人员可以尝试设计新的多模态神经网络架构, 结合视觉、听觉等多个感知通道, 在跨模态学习和联合推理中, 提升对伪造人脸的识别能力.

(5) 应对新型伪造技术的适应性

伪造人脸技术不断演进, 新型生成方法 (如基于自监督学习或少量数据训练的伪造技术) 不断涌现, 这对现有检测方法提出了严峻挑战. 因此, 未来的研究应致力于开发适应性强的检测技术, 能够应对新型伪造方法. 探索通用的伪造图像特征, 如深层伪造图像的细节特征或生成过程中常见的伪造痕迹, 可能是解决这一问题的有效途径. 此外, 基于生成对抗网络 (GAN) 等技术的自适应检测系统, 能够根据不同伪造类型动态调整检测策略, 也将是未来的重要研究方向.

(6) 实时检测

随着深度造假变得越来越普遍, 对实时检测系统的需求也在增加. 未来的研究应着眼于优化检测模型的计算效率, 以实现实时处理. 可以采用模型修剪、量化和开发轻量级神经网络架构等技术来实现这一目标. 在边缘设备和分布式系统上实现检测算法也可以促进实时应用.

(7) 高效的数据增强

多样化和高质量数据集的可用性对于训练和评估深度伪造检测模型至关重要. 未来的研究应侧重于创建全面的基准数据集, 其中包括使用不同技术生成的各种深度伪造. 数据增强和综合技术可用于创建更健壮的训练数据集, 改进模型训练和评估.

(8) 持续学习和适应

开发能够持续学习和适应的深度假检测模型对于跟上不断发展的操纵技术至关重要. 持续学习框架可以增量地从新数据中学习, 而不会忘记以前学习的信息, 这可以增强模型的适应性. 可以研究基于峰值的稳健连续元学习和用于峰值驱动在线学习的神经形态架

构等技术.

(9) 可解释性和透明度

增强深度伪造检测模型的可解释性对于获得用户信任和促进其采用至关重要. 开发可解释的人工智能技术, 为检测决策提供明确的理由, 可以帮助实现这一目标. 分层相关传播 (LRP)、局部可解释模型不可知解释 (LIME) 和 SHAP 值等方法可用于创建可解释模型.

(10) 道德和法律框架

随着深度伪造技术的发展, 解决其道德和法律问题势在必行. 需要政策制定者、法律专家和科技公司共同努力, 为深度虚假内容的使用和传播制定指导方针和法规. 确保在开发和部署检测技术时采用合乎道德的人工智能实践, 对于保持公平、问责制和透明度至关重要.

6 结束语

本文系统地回顾了最近几年伪造人脸检测技术的发展历程和关键技术. 首先, 本文介绍了伪造人脸检测的背景和概念, 讲述伪造人脸技术的主要分类. 然后, 按照输入训练样本的格式分类讨论, 对当前的技术、方法、模型等进行整理和对比; 同时, 本文整理了伪造人脸检测常用的数据集和评价指标. 最后, 讲述了伪造人脸检测当前的应用领域以及未来面对的挑战. 希望本篇综述能够激发更多关于伪造人脸检测技术的研究, 并为构建更加安全、可信的数字世界贡献一份力量.

参考文献

1 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: ACM, 2014. 2672–2680.

2 邬林桦. 当“眼见不再为实”, 全新风险如何应对? 解放日报 . https://baijiahao.baidu.com/s?id=1791478272339787059&wfr=spider&for=pc. (2024-02-21).

3 张传鑫. 基于解耦学习的非特定 GAN 图像检测逃避方法 [硕士学位论文]. 广州: 华南理工大学, 2022. [doi: 10.27151/d.cnki.ghnlu.2022.004418]

4 McCloskey S, Albright M. Detecting GAN-generated imagery using color cues. arXiv:1812.08247, 2018.

5 汤桂花, 孙磊, 毛秀青, 等. 基于深度对齐网络的生成对抗网络伪造人脸检测. 计算机应用, 2021, 41(7): 1922–1927.

6 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. 信息网络安全, 2021, 21(8): 70–81. [doi: 10.3969/j.issn.1671-1122.2021.08.009]

7 Zhou P, Han XT, Morariu VI, *et al*. Two-stream neural networks for tampered face detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu: IEEE, 2017. 1831–1839. [doi: 10.1109/CVPRW.2017.229]

8 Zhao HQ, Wei TY, Zhou WB, *et al*. Multi-attentional DeepFake detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 2185–2194. [doi: 10.1109/CVPR46437.2021.00222]

9 Luo YC, Zhang Y, Yan JC, *et al*. Generalizing face forgery detection with high-frequency features. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 16312–16321. [doi: 10.1109/CVPR46437.2021.01605]

10 Shao R, Wu TX, Nie LQ, *et al*. DeepFake-adapter: Dual-level adapter for DeepFake detection. arXiv:2306.00863, 2023.

11 Huang BJ, Wang ZY, Yang JF, *et al*. Implicit identity driven DeepFake face swapping detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 4490–4499. [doi: 10.1109/CVPR52729.2023.00436]

12 戴昀书, 费建伟, 夏志华, 等. 局部相似度异常的强泛化性伪造人脸检测. 中国图象图形学报, 2023, 28(11): 3453–3470. [doi: 10.11834/jig.221006]

13 冯才博, 刘春晓, 王昱烨, 等. 结合图像块比较与残差图估计的人脸伪造检测. 中国图象图形学报, 2024, 29(2): 457–467. [doi: 10.11834/jig.230149]

14 Tan CC, Zhao Y, Wei SK, *et al*. Learning on gradients: Generalized artifacts representation for GAN-generated images detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 12105–12114. [doi: 10.1109/CVPR52729.2023.01165]

15 Tan CC, Liu H, Zhao Y, *et al*. Rethinking the up-sampling operations in CNN-based generative network for generalizable DeepFake detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 28130–28139.

16 Guo Y, Zhen C, Yan PF. Controllable guide-space for generalizable face forgery detection. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023. 20761–20770. [doi: 10.1109/ICCV51070.2023.01903]

17 Yan ZY, Zhang Y, Fan YB, *et al*. UCF: Uncovering common features for generalizable DeepFake detection. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023. 22355–22366.

18 Bai WM, Liu YF, Zhang ZP, *et al*. AUNet: Learning relations between action units for face forgery detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 24709–24719.

19 Hou Y, Guo Q, Huang YH, *et al*. Evading DeepFake detectors via adversarial statistical consistency. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 12271–12280.

20 Nguyen D, Mejri N, Singh IP, *et al*. LAA-Net: Localized artifact attention network for quality-agnostic and generalizable DeepFake detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 17395–17405. [doi: 10.1109/CVPR52733.2024.01647]

21 Yan ZY, Luo YH, Lyu S, *et al*. Transcending forgery specificity with latent space augmentation for generalizable DeepFake detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 8984–8994. [doi: 10.1109/CVPR52733.2024.00858]

22 Yao KL, Wang J, Diao BY, *et al*. Towards understanding the generalization of DeepFake detectors from a game-theoretical view. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023. 2031–2041. [doi: 10.1109/ICCV51070.2023.00194]

23 Kim DK, Kim K. Generalized facial manipulation detection with edge region feature extraction. Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa: IEEE, 2022. 2784–2794. [doi: 10.1109/WACV51458.2022.00284]

24 Qian YY, Yin GJ, Sheng L, *et al*. Thinking in frequency: Face forgery detection by mining frequency-aware clues. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 86–103. [doi: 10.1007/978-3-030-58610-2_6]

25 Li JM, Xie HT, Li JH, *et al*. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 6454–6563. [doi: 10.1109/CVPR46437.2021.00639]

26 Wang GJ, Li W, Jiang Q, *et al*. Using grayscale frequency statistic to detect manipulated faces in wavelet-domain. Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Melbourne: IEEE, 2021. 2986–2993.

27 李颖, 边山, 王春桃, 等. 基于双流网络结构的深度伪造人脸的检测方法. 计算机科学, 2022, 49(11A): 220100106. [doi: 10.11896/jsjkx.220100106]

28 Jeong Y, Kim D, Ro Y, *et al*. FrePGAN: Robust DeepFake detection using frequency-level perturbations. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 1060–1068. [doi: 10.1609/aaai.v36i1.19990]

29 马欣, 吉立新, 李邵梅. 基于多尺度 Transformer 融合多域信息的伪造人脸检测. 计算机科学, 2023, 50(10): 112–118. [doi: 10.11896/jsjkx.220900048]

30 Liu HG, Li XD, Zhou WB, *et al*. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 772–781.

31 Wang Y, Yu K, Chen C, *et al*. Dynamic graph learning with content-guided spatial-frequency relation reasoning for DeepFake detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 7278–7287. [doi: 10.1109/CVPR52729.2023.00703]

32 Li JM, Xie HT, Yu LY, *et al*. Wavelet-enhanced weakly supervised local feature learning for face forgery detection. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: IEEE, 2022. 1299–1308.

33 Wolter M, Blanke F, Heese R, *et al*. Wavelet-packets for DeepFake image analysis and detection. Machine Learning, 2022, 111(11): 4295–4327. [doi: 10.1007/s10994-022-06225-5]

34 Song LC, Fang Z, Li XD, *et al*. Adaptive face forgery detection in cross domain. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 467–484.

35 Wang JK, Wu ZX, Ouyang WH, *et al*. M2TR: Multi-modal multi-scale Transformers for DeepFake detection. Proceedings of the 2022 International Conference on Multimedia Retrieval. Newark: ACM, 2022. 615–623. [doi: 10.1145/3512527.3531415]

36 Dong JH, Xie XH. Visually maintained image disturbance

against DeepFake face swapping. Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/icme51207.2021.9428173]

37 Wang Z, Guo YW, Zuo WM. DeepFake forensics via an adversarial game. IEEE Transactions on Image Processing, 2022, 31: 3541–3552. [doi: 10.1109/TIP.2022.3172845]

38 Chen L, Zhang Y, Song YB, *et al*. Self-supervised learning of adversarial example: Towards good generalizations for DeepFake detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 18689–18698. [doi: 10.1109/CVPR52688.2022.01815]

39 王瑜, 方贤进, 杨高明, 等. 基于注意力掩码与特征提取的人脸伪造主动防御. 计算机应用. https://link.cnki.net/urlid/51.1307.TP.20240802.1428.006. (2024-08-05)[2024-11-15].

40 Dong XY, Bao JM, Chen DD, *et al*. Identity-driven DeepFake detection. arXiv:2012.03930, 2020.

41 Liu BP, Liu B, Ding M, *et al*. TI^{2}Net: Temporal identity inconsistency network for DeepFake detection. Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa: IEEE, 2023. 4680–4689. [doi: 10.1109/WACV56688.2023.00467]

42 Chugh K, Gupta P, Dhall A, *et al*. Not made for each other-audio-visual dissonance-based DeepFake detection and localization. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 439–447.

43 Dong XY, Bao JM, Chen DD, *et al*. Protecting celebrities from DeepFake with identity consistency Transformer. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 9458–9468.

44 Chu BL, You WK, Yang Z, *et al*. Protecting world leader using facial speaking pattern against DeepFakes. IEEE Signal Processing Letters, 2022, 29: 2078–2082. [doi: 10.1109/LSP.2022.3205562]

45 Boháček M, Farid H. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. Proceedings of the 2022 National Academy of Sciences of the United States of America, 2022, 119(48): e2216035119.

46 Stehouwer J, Dang H, Liu F, *et al*. On the detection of digital face manipulation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020. 5780–5789.

47 Chen P, Liu J, Liang T, *et al*. DLFMNet: End-to-end detection and localization of face manipulation using multi-domain features. Proceedings of the 2021 IEEE International Conference on Multimedia and Expo. Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/ICME51207.2021.9428450]

48 Guo X, Liu XH, Ren ZY, *et al*. Hierarchical fine-grained image forgery detection and localization. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 3155–3165.

49 Guillaro F, Cozzolino D, Sud A, *et al*. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 20606–20615.

50 Sun ZK, Han YJ, Hua ZY, *et al*. Improving the efficiency and robustness of DeepFakes detection through precise geometric features. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3608–3617. [doi: 10.1109/CVPR46437.2021.00361]

51 朱春陶, 尹承禧, 张博林, 等. 基于多域时序特征挖掘的伪造人脸检测方法. 网络与信息安全学报, 2023, 9(3): 123–134. [doi: 10.11959/j.issn.2096-109x.2023044]

52 Fei JW, Xia ZH, Yu PP, *et al*. Exposing AI-generated videos with motion magnification. Multimedia Tools and Applications, 2021, 80(20): 30789–30802. [doi: 10.1007/s11042-020-09147-3]

53 Zhang DC, Li CY, Lin FZ, *et al*. Detecting DeepFake videos with temporal dropout 3DCNN. Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal: IJCAI, 2021. 1288–1294. [doi: 10.24963/IJCAI.2021/178]

54 Zhang DC, Lin FZ, Hua YY, *et al*. Deepfake video detection with spatiotemporal dropout Transformer. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022. 5833–5841.

55 Xu YT, Liang J, Jia GY, *et al*. TALL: Thumbnail layout for DeepFake video detection. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023. 22601–22611.

56 祝恺蔓, 徐文博, 卢伟, 等. 多关键帧特征交互的人脸篡改视频检测. 中国图象图形学报, 2022, 27(1): 188–202. [doi: 10.11834/jig.210408]

57 程燕. 基于关键帧与时空特征融合的人脸伪造检测. 计算机科学, 2024, 51(11): 191–197. [doi: 10.11896/jsjkx.240100063]

58 Gu ZH, Yao TP, Chen Y, *et al*. Hierarchical contrastive inconsistency learning for DeepFake video detection.

Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 596–613. [doi: 10.1007/978-3-031-19775-8_35]

59 Gu ZH, Yao TP, Chen Y, *et al*. Region-aware temporal inconsistency learning for DeepFake video detection. Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna: IJCAI, 2022. 920–926.

60 Hu J, Liao X, Liang JW, *et al*. FInfer: Frame inference-based DeepFake detection for high-visual-quality videos. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 951–959.

61 Gu ZH, Chen Y, Yao TP, *et al*. Delving into the local: Dynamic inconsistency learning for DeepFake video detection. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 744–752.

62 Choi J, Kim T, Jeong Y, *et al*. Exploiting style latent flows for generalizing DeepFake video detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 1133–1143. [doi: 10.1109/CVPR52733.2024.00114]

63 Tan LF, Wang YH, Wang JF, *et al*. DeepFake video detection via facial action dependencies estimation. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: IEEE, 2023. 5276–5284.

64 Li YZ, Chang MC, Lyu SW. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong: IEEE, 2018. 1–7. [doi: 10.1109/WIFS.2018.8630787]

65 Haliassos A, Vougioukas K, Petridis S, *et al*. Lips don't lie: A generalisable and robust approach to face forgery detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2020. 5037–5047.

66 Agarwal S, Farid H. Detecting deep-fake videos from aural and oral dynamics. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 981–989.

67 Haliassos A, Mira R, Petridis S, *et al*. Leveraging real talking faces via self-supervision for robust forgery detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 14930–14942.

68 Hernandez-Ortega J, Tolosana R, Fiérrez J, *et al*. DeepFakesON-Phys: DeepFakes detection based on heart rate estimation. Proceedings of the 2021 Workshop on Artificial Intelligence Safety (SafeAI 2021) Co-located with the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021.

69 Chen ML, Liao X, Wu M. PulseEdit: Editing physiological signals in facial videos for privacy protection. IEEE Transactions on Information Forensics and Security, 2022, 17: 457–471. [doi: 10.1109/TIFS.2022.3142993]

70 Lin JY, Zhou WB, Liu HG, *et al*. Lip forgery video detection via multi-phoneme selection. Proceedings of the 2021 International Workshop on Safety and Security of Deep Learning. New York, 2021.

71 Wu JH, Zhu Y, Jiang XB, *et al*. Local attention and long-distance interaction of rPPG for DeepFake detection. The Visual Computer, 2023, 40(2): 1083–1094.

72 Yu ZT, Cai RZ, Li Z, *et al*. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. IEEE Transactions on Dependable and Secure Computing, 2024, 21(5): 4327–4342. [doi: 10.1109/TDSC.2024.3352049]

73 Kalin S, Bhawna P, Abhinav D. Visual representations of physiological signals for fake video detection. arXiv:2207.08380, 2022.

74 Feng C, Chen ZY, Owens A. Self-supervised video forensics by audio-visual anomaly detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 10491–10503.

75 Oorloff T, Koppisetti S, Bonettini N, *et al*. AVFF: Audio-visual feature fusion for video DeepFake detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 27092–27102. [doi: 10.1109/CVPR52733.2024.02559]

76 Mittal T, Bhattacharya U, Chandra R, *et al*. Emotions don't lie: An audio-visual DeepFake detection method using affective cues. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 2823–2832.

77 Bohacek M, Farid H. Lost in translation: Lip-sync DeepFake detection from audio-video mismatch. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 4315–4323.

78 Cheng H, GuoYY, Wang TY, *et al*. Voice-face homogeneity tells DeepFake. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(3): 76.

79 Yang WY, Zhou XY, Chen ZK, *et al*. AVoiD-DF: Audio-visual joint learning for detecting DeepFake. IEEE Transactions on Information Forensics and Security, 2023,

18: 2015–2029. [doi: 10.1109/TIFS.2023.3262148]

80 Cozzolino D, Pianese A, Nießner M, *et al*. Audio-visual person-of-interest DeepFake detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 943–952. [doi: 10.1109/CVPRW59228.2023.00101]

81 Rössler A, Cozzolino D, Verdoliva L, *et al*. FaceForensics++: Learning to detect manipulated facial images. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2020. 1–11. [doi: 10.1109/ICCV.2019.00009]

82 Li YZ, Yang X, Sun P, *et al*. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3204–3213. [doi: 10.1109/CVPR42600.2020.00327]

83 Dolhansky B, Howes R, Pflaum B, *et al*. The DeepFake Detection Challenge (DFDC) preview dataset. arXiv:1910.08854, 2019.

84 Zhou TF, Wang WG, Liang ZY, *et al*. Face forensics in the wild. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 5774–5784. [doi: 10.1109/CVPR46437.2021.00572]

85 Huang JJ, Wang XY, Du B, *et al*. DeepFake MNIST+: A DeepFake facial animation dataset. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021. 1973–1982. [doi: 10.1109/ICCVW54120.2021.00224]

(校对责编: 张重毅)