E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:D918.91;TP18　　**文献标识码**:A　　**文章编号**:1006-8961(2026)10-3502-21

论文引用格式:Guo T L, Li J S and Tang Y Q. 2026. Advances in forensic science of deepfake facial images evidence. Journal of Image and Graphics, 31(10):3502-3522(郭甜利, 李纪松, 唐云祁. 2026. 深度伪造人脸图像证据的司法鉴定研究进展. 中国图象图形学报, 31(10):3502-3522)[DOI:10.11834/jig.250610]

深度伪造人脸图像证据的司法鉴定研究进展

郭甜利,李纪松,唐云祁*

中国人民公安大学侦查学院,北京　100038

摘　要:随着生成对抗网络(generative adversarial network,GAN)、自编码器(autoencoder,AE)和扩散模型(diffusion model,DM)等深度伪造技术的快速发展,深度伪造人脸图像广泛用于制作非自愿色情内容、实施电信诈骗及传播政治虚假信息等恶意行为。作为新型数字证据,深度伪造不仅频繁出现在当前犯罪活动中,还可能被用于虚假指控或操控监控记录,严重削弱法庭证据科学中人脸图像分析的可靠性,进而威胁刑事司法的公正性。本文系统梳理了关键数据集的发展脉络与深度伪造人脸图像的技术演进路径,深入剖析其生成机制的核心架构,涵盖主要技术路线、理论基础与功能逻辑;基于Web of Science数据库筛选出的737篇文献,采用VOSviewer进行文献计量分析,识别该领域的研究趋势,并揭示现有检测技术研究与司法实践中证据鉴伪实际需求之间的显著脱节。本研究提出将深度伪造人脸图像检测方法划分为技术原理导向与应用阶段适配两类互补框架,以应对不同阶段的证据鉴定需求。针对深度伪造人脸图像作为新型犯罪载体所带来的证据可信性挑战,本文主张法庭证据鉴定人应推动检测向鉴定的转化,实现检测结果与贝叶斯推理框架的融合,提升其在司法证明中的可采性与说服力。本研究衔接计算机视觉技术发展与法庭证据实践需求,明确了从技术检测到司法鉴定的逻辑路径,为法庭科学研究人员与司法实务工作者提供了具有技术价值的理论参考。

关键词:司法;证据;人脸图像;深度伪造;检测;鉴定

Advances in forensic science of deepfake facial images evidence

Guo Tianli, Li Jisong, Tang Yunqi*

School of Criminal Investigation, People's Public Security University of China, Beijing　100038, China

Abstract: With the rapid advancement of deepfake technologies, including generative adversarial networks (GANs), autoencoders (AEs), and diffusion models (DMs), deepfake facial images have been increasingly exploited for malicious purposes, such as the creation of nonconsensual pornographic content, telecommunication fraud, and the propagation of political disinformation. In recent years, the continuous iteration of these core deepfake technologies has considerably lowered the technical threshold for generating high-fidelity fake facial images: GANs, with their adversarial training mechanism between generators and discriminators, enable the rapid synthesis of realistic facial features; AEs, relying on unsupervised learning and feature reconstruction capabilities, excel in simulating subtle facial expressions and skin textures; diffusion models, through step-by-step noise reduction and feature enhancement, have further broken through the limitations of traditional deepfake technologies in terms of image clarity and detail authenticity, making deepfake facial images increasingly

收稿日期:2025-12-15;**修回日期**:2026-02-13;**预印本日期**:2026-02-20

*通信作者:唐云祁　tangyunqi@ppsuc.edu.cn

基金项目:国家社会科学基金重大项目(25&ZD238);中国人民公安大学研究生科研项目(2024yjsky032)

Supported by: Major Project of the National Social Science Fund of China (25&ZD238); Graduate Research Project of People's Public Security University of China (2024yjsky032)

difficult to distinguish with the naked eye and even challenging preliminary manual identification in forensic investigations. As an emerging form of digital evidence, deepfakes not only pervade contemporary criminal activities but also risk being weaponized for false accusations or the tampering of surveillance records, thereby severely undermining the reliability of facial image analysis in forensic science and threatening the integrity of the criminal justice system. To address the aforementioned challenges, this study systematically traces the developmental trajectory of key datasets and the technological evolution of deepfake facial images while conducting an in-depth dissection of the core architecture underlying their generation mechanisms, encompassing primary technical routes, theoretical underpinnings, and functional logics. In terms of dataset development, the evolution process from early small-scale, single-scene datasets (such as CelebA and LFW) to large-scale, multiscenario, high-diversity datasets (such as DFDC, FF++, and Celeb-DF) is sorted out, with the advantages and limitations of each type of dataset analyzed in terms of sample size, fake generation technology coverage, and scene diversity Moreover, the lack of standardized, forensic-oriented datasets has become one of the key bottlenecks restricting the practical application of deepfake detection technologies. In terms of technological evolution, the development of deepfake facial image technology is divided into three stages: the initial exploration stage dominated by GANs, the improvement stage supplemented by AEs, and the mature stage driven by diffusion models. The theoretical breakthroughs and technical improvements of each stage are elaborated, and the internal logic of the continuous improvement of fake image fidelity and the continuous reduction of generation costs is revealed. In the in-depth analysis of generation mechanisms, the core principles of the three mainstream technologies are focused on, the differences in their technical routes and functional characteristics are clarified, and the common key links and potential vulnerabilities in the generation process are summarized, laying a theoretical foundation for the subsequent design of targeted detection and identification methods. On the basis of 737 relevant articles retrieved and screened from the Web of Science database, a bibliometric analysis is performed via VOSviewer to identify research trends in the field and illuminate the substantial disconnect between existing detection technology research and the practical requirements for evidence authentication in judicial practice. The retrieval and screening process strictly followed the PRISMA guidelines: "deepfake", "facial image", "forensic science", "identification", and "detection" are adopted as key search terms; the retrieval time range is defined from the emergence of deepfake technology to the present; and irrelevant literature, review papers with low innovation, and duplicate publications are excluded through preliminary screening and full-text reading. The bibliometric analysis focuses on three core dimensions: research hotspots, institutional cooperation, and research frontiers. Results show that current research in the field is mainly concentrated on the optimization of detection algorithms (such as deep learning-based feature extraction and classification models), while research on the translation of detection results into forensic authentication, the formulation of forensic identification standards, and the adaptation of detection technologies to judicial practice scenarios is relatively scarce. This disconnect is mainly reflected in two aspects: on the one hand, most existing detection technologies focus on technical accuracy, ignoring the admissibility requirements of judicial evidence (such as objectivity, authenticity, and relevance); on the other hand, forensic examiners lack effective methods to integrate technical detection results into judicial authentication, leading to the situation where advanced detection technologies fail to be effectively applied in judicial practice. In response to the identified research-practice disconnect, this study proposes a classification of deepfake facial image detection methods into two complementary frameworks: technology principle oriented and application scenario adaptive, designed to address evidence authentication needs across distinct stages. The technology-principle-oriented framework takes the technical principles of deepfake generation as the core, classifies detection methods in accordance with the technical characteristics of different generation technologies (such as GAN-based detection, AE-based detection, and diffusion model-based detection), and focuses on improving the accuracy and generalization ability of detection, thus being mainly applicable to the preliminary technical screening stage of evidence in forensic investigations. The application-scenario-adaptive framework takes the actual needs of different judicial stages (such as investigation, prosecution, and trial) as guidance, classifies detection methods in light of the scenario characteristics and evidence requirements of each stage, and focuses on enhancing the adaptability and operability of detection methods, thereby being mainly applicable to the formal evidence authentication stage in judicial practice. The two frameworks are complementary and mutually supportive: the technology-principle-oriented framework provides technical support for the application-scenario-adaptive framework, and

the application-scenario-adaptive framework guides the direction of technical optimization, thus forming a complete detection system covering the entire process of forensic evidence authentication. In response to the evidentiary credibility challenges posed by deepfake facial images as a novel criminal vector, this study argues that forensic examiners should facilitate the translation of detection outcomes into forensic authentication, integrating detection results with the Bayesian inference framework to enhance their admissibility and probative force in judicial proceedings. Specifically, a Bayesian inference-based forensic authentication model is proposed. Technical detection results are taken as the prior probability, forensic examiners' professional experience and case context are integrated to determine the likelihood ratio, and the posterior probability of the authenticity of facial images is ultimately calculated via the Bayesian formula to convert objective technical indicators into credible judicial authentication conclusions. Furthermore, forensic examiners should continuously improve their understanding of deepfake technologies, master the working principles and limitations of detection technologies, and avoid over-reliance on technical detection results to ensure the scientificity and rigor of forensic authentication. In conclusion, by bridging the advancements in computer vision technology with the practical demands of forensic evidence practice, this research articulates the logical pathway from technical detection to judicial authentication, providing a theoretically insightful and practically relevant reference for forensic science researchers and judicial practitioners alike. For forensic science researchers, this study clarifies the research direction of integrating technical research with judicial practice, identifies the key issues to be addressed in the future (such as the construction of forensic-oriented datasets, the optimization of detection algorithms adapted to judicial scenarios, and the formulation of forensic identification standards), and offers a theoretical framework for subsequent research. For judicial practitioners, this study provides a practical operational path for the forensic authentication of deepfake facial images, facilitates the better grasp of the application methods and precautions of deepfake detection technologies, and elevates the efficiency and accuracy of digital evidence authentication. Additionally, this study offers a reference for the formulation of relevant laws and regulations, which helps improve the legal system of digital evidence authentication and promote the sound development of the criminal justice system in the era of deepfake technology.

Key words: forensic science; evidence; facial image; deepfake; detection; identification

0 引　言

“深度伪造”作为“深度学习”与“伪造技术”的融合产物，特指通过深度学习模型生成或篡改图像、视频、音频等数字内容的技术过程，其中人脸替换是最典型的表现形式，此外还包括人脸合成以及视频音频伪造等衍生形态。2017年，美国社交平台Reddit上首次出现名人面部篡改内容的大规模传播，标志着深度伪造技术正式进入公众视野（Dayal和van Nieberk，2021）。经过多年迭代，该技术的影响力已渗透至社会治理、刑事司法等关键领域。2024年《自然》（Nature）杂志将深度伪造检测列为年度值得关注的七大技术之一（Eisenstein，2024），凸显了其全球范围内的技术重要性与治理紧迫性。

深度伪造技术的滥用已引发多维度、跨领域的严重危害。

在色情与隐私侵犯领域，韩国产生的相关内容占全球的53%，99%的受害者为女性（Ji，2025），相关黑色产业链已形成规模化；西班牙女学生也遭遇了此类网络欺凌。

在金融领域，诈骗者通过模仿高管语音、伪造投资广告等手段实施犯罪。相关案件造成了严重损失：2019年虚假信息引发英国大都会银行挤兑潮，导致股价下跌9%（Van Der Sloot和Wagensveld，2022），且单起案件授权转账金额最高达2 560万美元（Brown等，2025）；2023—2025年间，英国储户相关诈骗案涉案金额达3 500万美元，企业及金融部门年度损失均超数十万美元，欧洲、加拿大储户及西班牙相关诈骗案的涉案金额达2 000万美元。

在政治与国际关系层面，伪造领导人言论、外交虚假图像、战争宣传视频等内容已成为地缘政治博弈工具，相关案例持续加剧国际或地区紧张局势：特朗普、奥巴马相关虚假言论（Pesetski，2020；Ramluckan，2024），意大利前总理伦齐、加蓬总统的伪造视频争议（Schiff等，2025；Venema和Geradts，2020），外交虚假图像（Alhajjar和Lee，2022），2024年模仿拜登声音以压制选民投票的人工语音电话，战

争宣传视频(Boháček 和 Farid,2022),以及 2025 年印巴冲突中的伪造“道歉”视频均在其列。司法实践中,伪造音频还曾干扰监护权判决。

在军事与国家安全领域,篡改卫星图像导致军事误判(Çiftçi 和 Demir,2023);技术安全层面,78%的深度伪造内容成功欺骗微软 Azure 实验性服务 API(application programming interface),暴露了商业面部识别系统的脆弱性。

在儿童权益保护领域,2023—2024 年人工智能生成的儿童性虐待材料数量翻两番,勒索案件同步增加(Krishna 等,2024)。

从法庭证据科学视角来看,洛卡德物质交换原理作为刑事侦查的核心理论,其内涵已从物理环境延伸至数字空间——任何数字篡改行为都会留下可追溯的技术痕迹(Amundsen 和 Ovens,2017; Iqbal 等,2020; Kaur 和 Saini,2024)。但与指纹(Levanon 和 Tully,2025)、脚印、DNA(deoxyribonucleic acid)、声纹、面部特征、虹膜图案以及步态动态(Deng 等,2022)等传统生物识别证据不同,深度伪造人脸图像证据具有显著的数字特性:1)数字无形性,痕迹仅存在于数字系统中,无物理载体;2)技术异质性,生成模型与操纵算法的多样性导致伪造结果呈现高度差异化;3)传播突发性,数字内容可在短时间内实现大规模扩散,放大社会危害;4)鉴别复杂性,对抗性生成机制使人工鉴别可靠性极低,必须依赖自动化检测工具;5)刑事侦查模糊性,现有刑事侦查与司法框架中缺乏针对该类数字证据的可采性标准与证据权重评估规范。这些特性对法庭证据科学提出了全新的理论与实践挑战,亟需建立适配数字环境的证据分析与评估体系。

围绕深度伪造的生成机制与检测技术,现有研究已开展了大量探索。Seow 等人(2022)系统梳理了深度伪造的生成技术、检测方法、数据集及研究空白,为领域发展提供了全面参考;Ramanaharan 等人(2025)深入探讨检测模型的泛化能力,指出当前研究存在过度拟合数据集等局限性,并提出开发更逼真数据集的未来方向;Garg 和 Gill(2025)指出当前研究缺乏多模态方法、对混合模型探索不足的问题,进而提出多模态混合学习检测框架;Abbas 和 Taeihagh(2024)探讨了基于人工智能的深度伪造检测与生成技术,阐明了人工智能在该领域的应用潜力,并提出相关政策建议。然而,现有研究仍存在显著短板:1)研究视角集中于计算机视觉领域,缺乏与法庭证据科学的跨学科整合,未能回应司法实践对证据评估的核心需求;2)检测技术分类体系零散,未形成适配法庭证据应用阶段的结构化框架;3)对深度伪造证据的定量评估方法研究不足,难以支撑刑事诉讼程序中的证据采信。

法庭科学作为整合生物学、物理学、化学和计算机科学知识的应用学科,核心目标是对 DNA(生物类证据)、指纹(形态类证据,含足迹、枪弹痕迹等)以及电子证据(包含图像、音频、视频等)进行定量研究与概率评估,为刑事侦查和刑侦诉讼程序提供支撑。针对现有研究缺口与刑事司法实践需求,本研究以法庭证据科学核心逻辑为切入点,厘清深度伪造技术的核心原理与发展趋势,建立适配法庭证据场景的检测方法分类体系,提出科学可行的深度伪造人脸图像证据的鉴定可行性方案。

1　深度伪造人脸

深度伪造技术的演进路径,也是其作为犯罪载体的发展过程。

1.1　深度伪造人脸数据集

从时间维度看,深度伪造技术的发展呈现出多样化、精细化和向真实场景演进的趋势。详见图 1。

深度伪造人脸数据集的演进紧密跟随生成技术的迭代,历经了从奠基期、优化期、场景适配期,到体系化及爆发期的完整发展历程。这一过程反映了数据资源从早期单一合成向复杂多模态生成的转变,2018—2024 年间各阶段的代表性深度伪造数据集如表 1 所示。

2018 年为技术奠基阶段,核心生成框架初步建立,研究聚焦于低分辨率人脸合成与基础换脸技术,主要采用生成对抗网络(generative adversarial network, GAN)和自编码器(autoencoder, AE)。例如,DeepFakeTIMIT 数据集利用 GAN 结合卷积神经网络(convolutional neural network, CNN)分割与多任务卷积神经网络(mulit-task convolutional neural network, MTCNN)检测的面部特征点,实现初步换脸。UADFV 则通过自编码器配合仿射变换和边界平滑减少拼接痕迹。但由于依赖单一模型,伪造结果存在边缘模糊、表情僵硬等明显视觉缺陷。

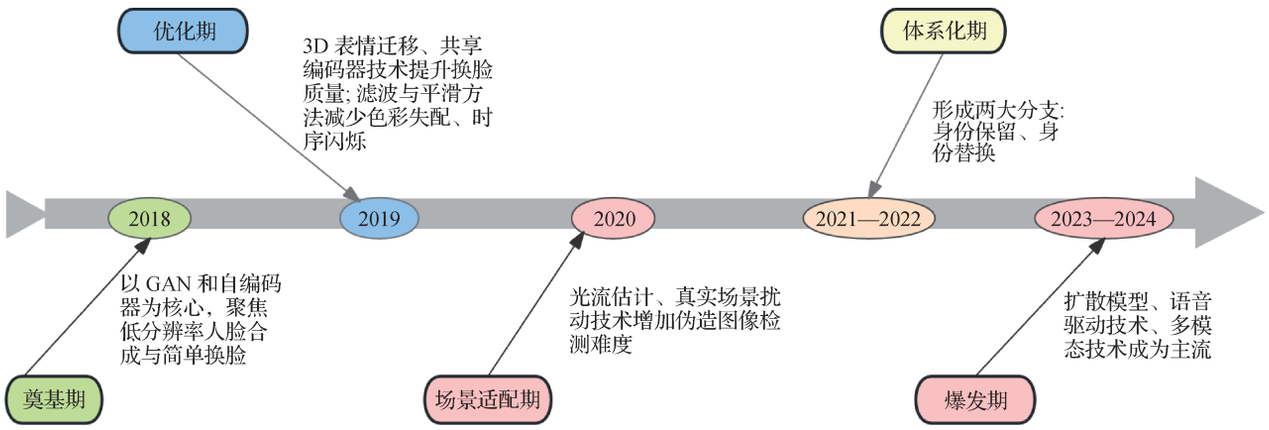


图 1　数据集时间发展的鱼骨图

Fig. 1　Fishbone diagram of the temporal development of the dataset

表 1　2018—2024 年深度伪造人脸的代表性数据集

Table 1　Representative datasets of deepfake faces from 2018 to 2024

数据集名称	基本信息
DeepFakeTIMIT （Korshunov 和 Marcel，2018）	时间：2018；机构：艾迪亚普研究所；大小：总计 640 段（分高清、标清版本）；类型：视频；方法：GAN 驱动；链接 https://www.idiap.ch/dataset/DeepFaketimit
UADFV（Yang 等，2019）	时间：2018；机构：奥尔巴尼大学等；大小：真实 49 段，伪造 49 段；类型：视频；方法：检测框定、仿射变换与边界平滑；链接 https://docs.google.com/forms/d/e/1FAIpQLScKPoOv15TIZ9Mn0nGScIVgKRM9tFWOmjh9eHKx57Yp-XcnxA/viewform
Fake Face in the Wild FFW database（Khodabakhsh 等，2018）	时间：2018；机构：奥勒松大学等；大小：伪造图像 5.3 万幅（150 段视频），真实图像 7.85 万幅（150 段视频）；类型：视频/图像；方法：编码器、手动篡改、CGI；链接 http://ali.khodabakhsh.org/ffw/
100K-Faces（Almars, 2021）	时间：2018；机构：英伟达；大小：伪造图像 10 万幅；类型：图像；方法：StyleGAN；链接 https://generated.photos/
PGGAN（Karras 等，2018）	时间：2018；机构：英伟达；大小：伪造图像 8 万幅；类型：图像；方法：PGGAN；链接 https://github.com/tkarras/progressive_growing_of_gans
FaceForensics++（Rössler 等，2019）	时间：2019；机构：慕尼黑工业大学等；大小：图像超 180 万幅（1 000 段原始视频，4 000 段伪造视频）；类型：图像/视频；方法：3D 建模、编码；链接 https://github.com/ondyari/FaceForensics
TPDNE（Karras 等，2019）	时间：2019；机构：英伟达；大小：图像 15 万幅；类型：图像；方法：StyleGAN + 潜空间生成；链接 https://thispersondoesnotexist.com
Celeb-DF-v2（Li 等，2020）	时间：2019；机构：奥尔巴尼大学等；大小：真实 590 段，伪造 5 639 段；类型：视频；方法：编解码；链接 http://www.cs.albany.edu/~lsw/celeb-DeepFakeforensics.html
Celeb-DF-v1（Li 等，2019）	时间：2019；机构：奥尔巴尼大学等；大小：真实视频 408 段；深度伪造视频 795 段；类型：视频；方法：传统伪造；链接 https://github.com/danmohaha/celeb-DeepFakeforensics
DFDC（Dolhansky 等，2020）	时间：2019；机构：脸书等；大小：总计 12.815 万段；类型：视频；方法：8 类伪造方法锐化，泊松融合；链接 https://ai.facebook.com/datasets/dfdc
WildDeepFake（Zi 等，2020）	时间：2020；机构：复旦大学等；大小：伪造视频 707 段，人脸图像 118 万幅；类型：视频/图像；方法：真实场景采集；链接 https://github.com/OpenTAI/wild-DeepFake
DF-1.0　DeeperForensics-1.0（Jiang 等，2020）	时间：2020；机构：南洋理工大学等；大小：真实 5 万段，伪造 1 万段；类型：视频；方法：光流时间一致性、真实场景扰动；链接 https://liming-jiang.com/projects/DrF1/DrF1.html
iFakeFaceDB（Neves 等，2020）	时间：2020；机构：葡萄牙贝拉内大学；大小：伪造图像 8.7 万幅；类型：图像；方法：卷积自编码器去除指纹；链接 https://github.com/socialabubi/iFakeFaceDB

续表1 2018—2024年深度伪造人脸的代表性数据集

Table 1 Representative datasets of deepfake faces from 2018 to 2024(continued)

数据集名称	基本信息
OpenForensics(Le等,2021)	时间:2021;机构:日本国立信息学研究所;大小:真实16.067 6万幅,伪造17.366 0万幅;类型:图像;方法:潜向量修改;链接https://sites.google.com/view/ltnghia/research/openforensics
KoDF(Kwon等,2021)	时间:2021;机构:钱脑公司;大小:403名亚洲志愿者;真实6.22万幅,伪造17.58万幅;类型:视频;方法:6类核心合成模型;链接https://moneybrain-research.github.io/kodf
ForgeryNet(He等,2021)	时间:2021;机构:中国商汤科技研究院等;大小:图像290万幅,视频22.124 7万段;类型:视频/图像;方法:FOMM, StarGAN2;链接https://yinanhe.github.io/projects/forgerynet.html
FakeAVCeleb (Khalid 等, 2021)	时间:2021;机构:韩国成均馆大学等;大小:真实500段,伪造1.95万段;类型:视频;方法:人脸相似度;链接https://sites.google.com/view/fakeavcelebdash-lab/
DFGC-21 testing dataset(Peng等,2021)	时间:2021;机构:深圳大学等;大小:真实1 000幅,伪造1 000幅;类型:图像;方法:仿射;链接https://github.com/yuezunli/celeb-DeepFakeforensics
Localized Audio Visual DeepFake LAV-DF(Cai等,2022)	时间:2022;机构:澳大利亚莫纳什大学等;大小:真实3.643 1万段,伪造片段9.987 3万段;153名个体;类型:视频;方法:唇形同步;链接https://github.com/ControlNet/LAV-DF
DFDM(Jia等,2022)	时间:2022;机构:美国纽约州立大学布法罗分校等;大小:伪造6 450段;类型:视频;方法:自编码器;链接https://github.com/shanface33/DeepFake_Model_Attribution
DeePhy DeepFake Phylogeny dataset(Narayan等,2022)	时间:2022;机构:印度焦特布尔印度理工学院等;大小:伪造5 040段,真实100段;类型:视频;方法:1～3次人脸换脸;链接http://iabrubric.org/deephy-database
GOTCHA(Mittal等,2024)	时间:2022;机构:纽约大学等;大小:伪造409段,真实5.583 8万段;类型:视频;方法:FSGAN v2;链接https://github.com/mittalgovind/GOTCHA-DeepFakes
DeepFakeFace (Song 等, 2023)	时间:2023;机构:清华大学等.;大小:真实3幅,伪造9万幅;类型:图像;方法:扩散模型;链接https://github.com/OpenRL-Lab/DeepFakeFace/ https://huggingface.co/datasets/OpenRL/DeepFakeFace
StyleGAN2-FFHQ (Bray 等, 2023)	时间:2023;机构:伦敦大学等;大小:图像100幅(深度伪造人脸50幅;真实人脸50幅);类型:图像;方法:StyleGAN2;链接https://osf.io/tfn7v/
Div-DF(Dagar和Vishwakarma,2023)	时间:2023;机构:德里技术大学等;大小:真实150幅,伪造250幅(换脸100幅,重演100幅;唇形同步50幅);类型:视频;方法:FSGAN;链接https://forms.gle/WhWDcVoM3wHToBYi9
DFFMD(Alnaim等,2023)	时间:2023;机构:伊玛目阿卜杜勒拉赫曼·本·费萨尔大学;大小:真实1 000段, 伪造1 000段;类型:视频;方法:FOMM;链接https://creativecommons.org/licenses/by-nc-nd/4.0/
DF40(Yan等,2024)	时间:2024;机构:北京大学等;大小:视频片段超10万段,图像超100万幅;类型:视频/图像;方法:10类换脸,12类人脸重演,12类全脸合成,5类人脸编辑;链接https://github.com/YZY-stack/DF40

2019年进入技术优化期,转向多方法融合与细节优化。FaceForensics++集成Face2Face(3D表情迁移)和DeepFakes(身份替换)等多种技术,并通过不同比特率的压缩模拟真实传播环境。Celeb-DF-v2引入色彩校正与卡尔曼滤波,有效缓解色偏和画面闪烁,显著提升视频的时间连贯性与真实感。

2020年迈向实际应用,强调合成内容与真实环境的融合。DF-1.0采用光流估计保障时间一致性,并加入JPEG(joint photographic experts group)压缩、噪声、运动模糊等7类现实失真,更贴近数字媒体的实际退化情况。WildDeepFake收集来自网络平台的真实深度伪造视频,涵盖图像拼接与生成合成等多种手段,增强生态效度,贴近司法实践中面临的复杂证据形态。2021—2022年,深度伪造方法体系趋于系统化,逐渐形成“身份保留”和“身份替换”两大伪造模式。KoDF(Korean deepfake detection dataset)整合FaceSwap(身份替换)与FOMM(first-order motion model)(面部重演,身份保留),并引入FGSM(fast gradient sign method)等对抗扰动以增强抗检测能力。ForgeryNet明确分类操纵技术,区分StarGAN2类属性编辑与FaceShifter类高保真身份迁移,构建了结构化评估体系。2023—2024年,前沿伪造

方法与基准集中涌现，扩散模型与多模态融合逐步成为主流。DeepFakeFace基于Stable Diffusion进行局部精细编辑，结合ControlNet实现像素级操控；DF40基准收录多种前沿伪造技术，覆盖StyleGAN3全脸生成与SadTalker音频驱动唇动同步。这些进展标志着深度伪造进入高保真、强适应、低残留的新阶段，对电子数据检验与数字取证工作构成前所未有的挑战。

1.2 深度伪造人脸图像中的信息

人脸图像包含丰富信息。如图2所示，原始人脸图像可提取多种特征，包括语义掩码、RGB颜色、深度、频域和空域特征等。这些内在属性被生成模型用于生成深度伪造人脸。

尽管不同类型信息的具体表示与特征提取方法较为复杂，但图2中的各子图仍能直观反映不同模态的基本分布与结构。其中，语义掩码通过像素级标注实现结构化面部分割，明确眼、鼻和口等关键面部部位的边缘轮廓与边界，完成空间定位与语义区分；颜色与深度信息互补，颜色通道反映表面外观（如色彩分布与纹理），深度数据则通过空间距离刻画3D面部几何特征（如鼻梁凸起、下颌凹陷），二者共同提供多模态的面部外观描述；在频域分析中，高频成分体现皮肤纹理、睫毛等细节，中频成分连接局部结构与整体形态，低频成分主导全局轮廓与宏观布局（如脸型、五官排列），反映粗粒度结构特征；空域特征则随维度变化而呈现不同表征能力。在二维空间中，像素梯度与密度可以刻画局部复杂性，例如眼部区域因纹理更为丰富而表现出更剧烈的梯度变化；在三维空间中，不同坐标轴用于捕捉特定几何属性，z轴聚焦于深度方向的形变，x/y轴描述平面内的轮廓与纹理，从而形成多视角的空间表达。

另外，深度神经网络学习到的潜在特征通常是高维、抽象且不可直接可视化的。尽管难以用直观方式加以解读，这些潜在特征却编码了人脸生成所需的全局统计规律与结构先验，是深度伪造人脸合

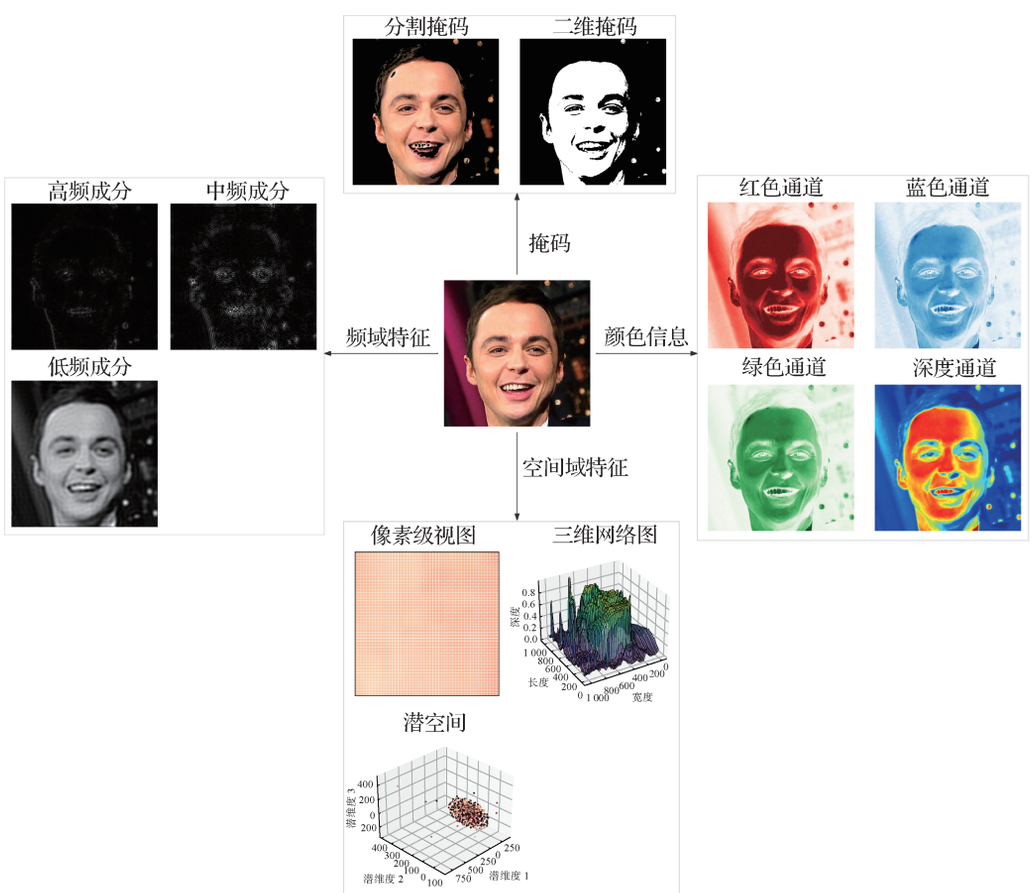


图2 从人脸图像中提取的关键信息

Fig. 2 Key information extracted from facial images

成与编辑的核心表征基础。

1.3　深度伪造人脸的技术本质

1.3.1　技术路线

深度伪造人脸的生成流程包含4个核心阶段，如图3所示。

1)准备阶段。利用深度神经网络从待伪造人脸中提取高维抽象特征。这些特征不同于传统法庭科学关注的几何比例、皮肤纹理等可观察形态，而是对颜色分布、空间结构和频域特性等多维信息的隐式编码，兼具高维度与强抽象性。同时提供待迁移特征的原始图像和伪造结果，并通过面部关键点检测与姿态对齐建立源与目标间的特征映射，为后续建模提供关键约束。

2)伪造建模阶段。使用生成模型，学习源与目标特征的关联，实现特征的定向迁移或全新生成。例如，将源数据的身份特征融入目标的姿态框架，或在目标约束下生成符合特定属性的新面部特征。

3)重新渲染阶段。将生成的抽象特征转换为可视图像，通过调整光照一致性、像素级重建和补充皮肤纹理等细节，赋予图像符合视觉感知的颜色、纹理与空间结构。

4)后处理阶段。通常进行扰动增强，如微调高频成分、添加噪声，提升图像真实感或增强逃避反深度伪造人脸检测的能力，最终输出符合预期的伪造结果。

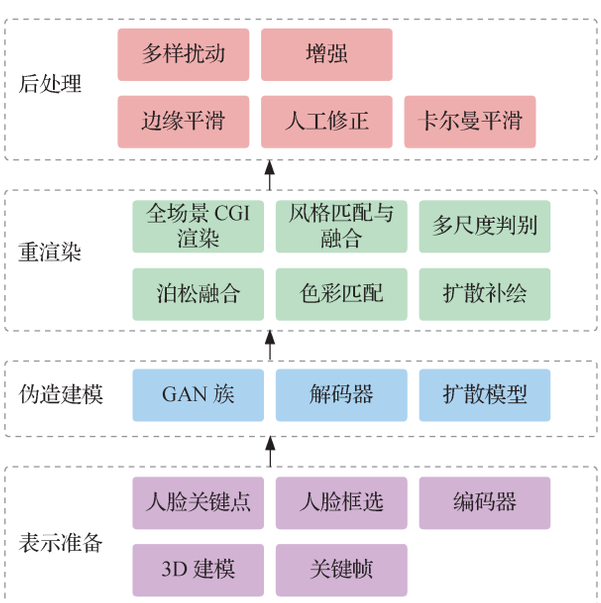


图3　生成深度伪造人脸图像的流程

Fig. 3　Deepfake image generation process

1.3.2　技术基础

1)生成对抗网络(GAN)及其变体。生成对抗网络由生成器和判别器组成，两者在对抗训练中不断提升各自能力，最终使伪造图像在视觉效果上接近真实图像，以至于判别器接近随机猜测。这一机制如图4所示。

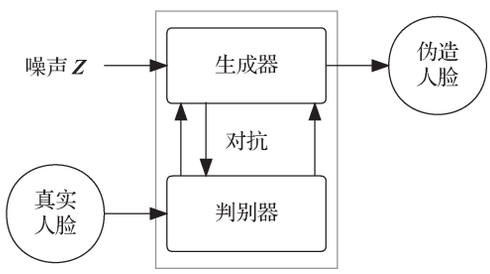


图4　基于生成对抗网络的伪造人脸图像生成机制示意图

Fig. 4　Schematic of GAN-based forged facial image generation

2)自编码器。自编码器通过编码器从人脸图像中提取潜在特征，并利用解码器将其重构为连贯的人脸图像。在面部替换中，采用两个共享编码器权重的对称编码器—解码器结构。共享编码器确保身份特征的一致性编码，各解码器则在其对应数据集上独立训练。解码路径相互解耦，便于独立优化以提升生成真实性。面部替换效果依赖于姿势、表情和头部形态的一致性。编码器通过分离身份与非身份属性实现特征的精确解缠，双解码器结构则在保持源图像外观上下文的同时重建目标身份。该过程如图5所示。

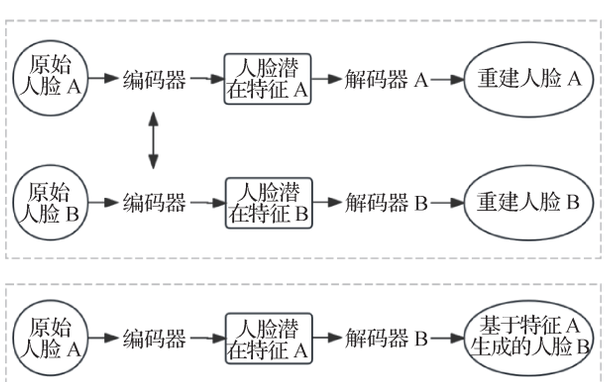


图5　基于自编码器的伪造人脸图像生成机制示意图

Fig. 5　Schematic of autoencoder-based forged facial image generation

3)扩散模型。扩散模型包含两个阶段：正向扩散和反向去噪。在正向过程中，噪声逐步添加到原始人脸图像，经过多步迭代后图像完全变为噪声。

该过程是固定的，由预定义的方差调度控制每一步的噪声强度。反向过程中，神经网络从纯噪声出发，逐步重建原始图像。该网络经训练可估计并去除每步噪声，从而实现对正向过程的有效逆转，其架构与原理如图6所示。

1.3.3　功能逻辑

从功能分类来看，深度伪造人脸图像主要包括面部重演、面部编辑、面部替换与面部合成4类，其核心差异体现在对身份信息的处理逻辑上，具体如表2所示。

面部重演聚焦面部动态迁移，将源个体的头部姿势、表情和唇部动作等特征传递至目标个体，同时完整保留目标主体的身份属性；面部编辑则针对特定面部属性进行局部修改，如调整年龄、发型或表情等，不改变个体的核心身份信息；面部替换通过特征替换实现身份重构，以源个体的面部特征替代目标个体特征，进而改变输出结果中感知到的身份归属；面部合成则基于生成模型从零构建逼真人脸图像，生成的内容不对应任何真实存在的个体。

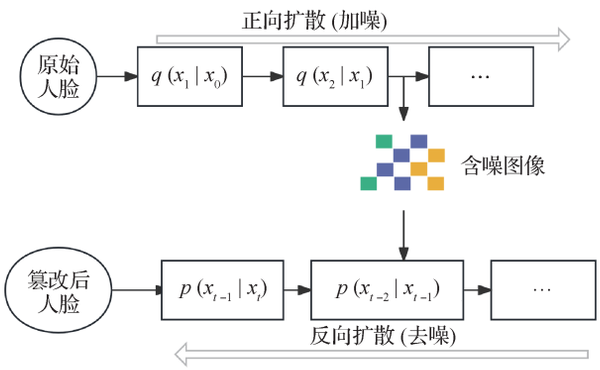


图6　基于扩散模型的伪造人脸图像生成机制示意图

Fig. 6　Schematic of diffusion-based forged facial image generation

表2　深度伪造人脸图像的分类

Table 2　Classification of deepfake facial images

类别	身份变化	技术实现	潜在风险
面部重演	保留目标身份	迁移源面部表情与头部姿态至目标，复刻动作	伪造公众人物言论或行为
面部编辑	保留目标身份	基于GAN或编解码器模型，修改年龄、发型、表情等属性，不改变身份	篡改数字证据、生成误导性生物特征数据
面部替换	替换为源身份	通过面部对齐、特征提取与图像融合，迁移源面部结构与纹理至目标	身份欺诈、深度伪造色情内容、冒充攻击
面部合成	生成全新身份	基于生成模型或3D建模，从噪声生成逼真人脸	传播虚假信息、伪造网络身份实施诈骗

1.3.4　后处理

基于68或81个关键点的面部遮罩技术已从早期方法的边缘模糊，发展为自适应高斯模糊，并结合泊松融合，使伪造区域与背景梯度对齐。伪造人像通过模拟自然阴影，使传统边缘检测失效，显著增加伪影检测难度。

深度伪造后处理已超越基础视觉融合，向对抗司法取证方向演进。为规避真伪判别，后处理算法在优化伪造痕迹的同时引入对抗策略。除追求视觉逼真外，更通过特征混淆与去标识化切断图像与特定生成工具的关联，掩盖生产方式与动因，增加溯源难度，从而削弱其法庭证据效力。

2　深度伪造人脸检测

深度伪造人脸生成与检测属于典型的猫鼠博弈关系。随着生成技术快速发展，会持续给人脸伪造检测带来挑战，削弱已有检测方法的效果，进而推动检测技术不断改进，也使深度伪造人脸检测成为备受关注的研究方向。

2.1　研究热点

本研究以“(“deepfake” OR “face forgery” OR “face manipulation” OR “face swap” OR “facial deepfake”) AND (“face” OR “facial”)”为主题，在Web of Science核心合集(core collection)检索近5年文献，共获取845篇文档；经筛选排除与深度伪造人脸不相关的内容及综述类文章后，剩余737篇有效文献。保留文献的题目与摘要信息并导入VOSviewer，以Binary Counting方式计算关键词频数，选取频数大于20的关键词共196个，展示前60%(共118个)；将标签最大长度设置为10、线长设置为1 000，最终生成网络可视化图谱。

图 7 展示了 2022—2026 年深度伪造人脸图像检测领域文献的发表趋势。2022 年与 2023 年发文量分别为 101 篇和 108 篇，呈现稳步增长态势，标志着该领域处于技术积累与认知拓展阶段。随着深度伪造技术引发的安全风险逐步显现，研究重点集中于检测方法的构建，致力于建立技术框架与理论体系。2024 年发文量跃升至 265 篇，2025 年维持在 260 篇的高位水平，进入显著增长期。这一爆发式增长源于双重驱动机制：1）学术界对该议题的关注持续升温，高精度检测模型与大规模标注数据集的相继推出为技术迭代提供了坚实支撑；2）现实中深度伪造引发的虚假信息传播、身份冒用等安全事件频发，对社会信任体系与个体权益构成严峻挑战，催生了迫切的实际防控需求。上述因素共同推动研究范式从技术可行性验证向实际应用阶段适配演进。截至 2026 年统计时点（2026 年 1 月 14 日）仅收录 3 篇文献，该年度研究工作正处于起步布局阶段。

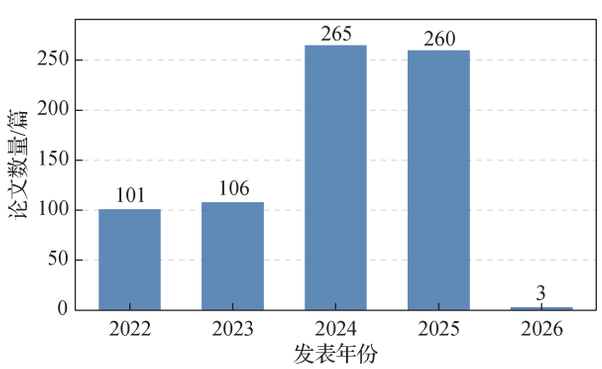


图 7　深度伪造人脸图像检测文献近 5 年发表数量

Fig. 7　The number of published papers on deepfake face image detection in the past five years

从图 8 的网络聚类可见，红色与橙色节点构成以“deepfake”为核心的技术聚类，涵盖伪造类型（如“faceswap”、“fake video”、“fake image”）和技术路径（如“deep neural network”、“feature extraction”、“convolution”），表明计算机视觉领域已围绕深度伪造图像构建起系统化的研究框架，并持续聚焦于特征提取、卷积模型优化等关键技术方向，展现出较强的技术积累与演进能力，具备支撑实际鉴伪应用的基础条件。同时，该聚类关联“precision”、“accuracy”“F1 score”等评估指标，进一步凸显研究对检测性能精细化评估的高度重视，反映出该领域正由方法探索向性能验证深化，是其研究成熟度提升的重要体现。绿色节点形成以“detection”为中心的实践聚类，延伸至“benchmark”、“generalization”、“artifact”等关键术语，不仅体现了计算机视觉领域当前的研究重心，也恰好契合司法实践中对证据鉴伪在可验证性、稳定性与可解释性方面的需求。

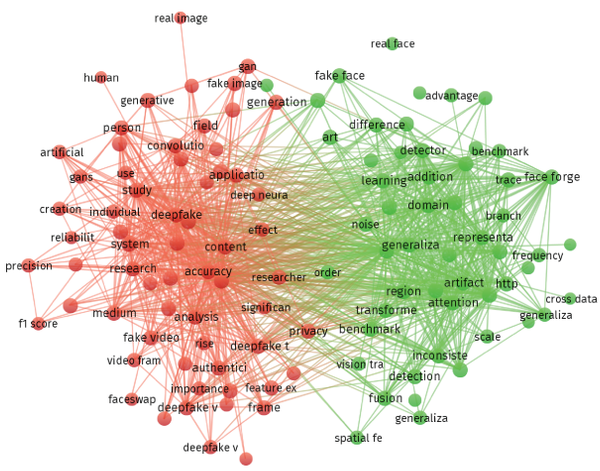


图 8　使用 VOSviewer 的关键词共现可视化

Fig. 8　Keyword co-occurrence visualization using VOSviewer

然而，值得注意的是，现有研究仍主要集中于技术性能的内部优化，尚未主动构建计算机视觉方法与法庭证据标准之间的系统性对接机制。尽管学界对“artifact”（伪造痕迹）、“frequency”（频率特征）等底层信号的分析已在事实上触及证据真实性的判断依据，而司法体系对证据可靠性与程序公正的要求亦亟需技术手段的支持。但目前技术供给与司法需求之间仍呈现并行发展、缺乏深度融合的局面，二者在术语体系、验证逻辑与应用阶段上的衔接尚待加强。

综上所述，从深度伪造人脸图像作为司法证据的视角出发，该网络图谱不仅反映出计算机视觉领域在深度伪造人脸检测研究方面的技术成熟度与发展深度，同时也揭示了当前技术研究与司法实践中证据鉴伪需求之间存在明显脱节的现状。未来的研究应主动构建技术与司法之间的跨学科衔接机制，推动计算机视觉成果在刑事侦查中更高效、精准地应用于证据真实性检验，实现技术能力与司法需求之间的双向对接。

2.2　检测方法分类

随着深度伪造人脸图像技术的演进，检测技术也随之不断发展（丁峰 等，2024）。深度伪造人脸检测通常被视为二分类任务，即通过深度模型区分真实与伪造的人脸图像。为了便于与刑事司法需求对

接，本文从两个维度对深度伪造人脸取证方法进行归类（姚文达 等，2025）。1）按技术原理划分，聚焦算法框架、特征提取和模型构建等底层逻辑，回答“如何判定图像为伪造”的核心问题；2）按应用阶段划分，根据技术在伪造流程中的介入时点，将方法分为伪造前的源头预防、伪造中的实时监测与拦截以及伪造后的溯源追责等3个关键环节。

2.2.1 按技术原理分类

按技术原理划分，深度伪造人脸检测方法大致可分为3类：特征分析、模型能力和对抗策略，如表3所示。

表3 按技术原理对深度伪造检测方法的分类

Table 3 Classification of deepfake detection methods by technical principle

技术原理	方法	解决问题	主流方法	技术优势	局限性
特征分析	时空法	视频中因人脸伪造引发的帧间逻辑矛盾	帧间运动一致性检测、光照与姿态的时序连贯性分析（Guan 等，2025; Shahzad 等，2025; Xu 等，2024a）	可捕捉人眼难以察觉的伪造痕迹	对视频压缩和帧率变化敏感，难以有效检测无明显帧间异常的伪造视频
	频率法	挖掘伪造人脸图像技术在频域留下的信息	傅里叶变换频谱分析、小波变换异常检测（Jin 等，2024）	对图像压缩和噪声扰动具有鲁棒性	对扩散模型等新型伪造技术响应较弱
	多模态法	整合多维度人脸图像证据信息，解决单一模态鉴伪不全面的问题	音视频同步性验证、基于CLIP等大模型的视觉—文本语义一致性分析（Joel 等，2025; Zhou 等，2025）	鲁棒性强，能识别复杂伪造人脸图像，降低误判	依赖高质量多模态数据，模型结构复杂且成本较高
	生物信号法	鉴定人脸证据的生物真实性	人脸生理信号（心跳、呼吸）检测与眨眼频率一致性分析（Jin 等，2021; Ni 等，2024）	泛化能力强，生物特征难伪造	对视频质量、光照和头部姿态要求高，实用性受限
	深度特征法	提取人脸图像证据中的深层伪造痕迹	CNN/Transformer 特征提取，基于注意力机制的细微伪影检测（Liang 等，2023; Wang 等，2024）	特征表达能力强，检测精度高，可适配多种伪造类型及高质量证据鉴伪	依赖训练数据，易发生过拟合，跨数据集泛化能力弱
模型能力	自监督学习	减少对标注数据的依赖，提升模型对多样伪造证据的适配性	基于对比学习与掩码重建的特征学习（Qiao 等，2024; Shao 等，2025）	无需大规模标注数据，泛化能力强，适配司法场景数据稀缺现状	模型训练难度高，部分场景检测精度低于监督学习，性能验证需更多司法案例支持
	多类别学习	区分深度伪造人脸图像证据的伪造类型，为法庭提供溯源依据，辅助案件事实认定	基于类别标签的伪造技术分类模型（如区分GAN类与扩散模型类伪造）（Bunluesakdikul 等，2025; Deng 等，2025）	分类结果明确，可为案件定性提供技术支撑，证据指向性强	对新型伪造类型的识别能力弱，需持续更新训练数据中的伪造类型标签
	元学习	快速适配新型伪造证据鉴伪需求，解决司法实践中伪造技术迭代快的问题	采用少样本/零样本检测模型与元知识迁移学习（Huang 和 Zhang，2024; Liu 等，2024）	样本需求小，可快速响应新型伪造手段，满足司法场景时效性要求	模型稳定性有待提升，小样本训练易致误判，需结合其他方法交叉验证
	异常检测	识别偏离正常模式的伪造证据，降低误用异常证据为有效证据的风险	基于正常人脸数据的分布建模与重构误差异常判定（Rosca 和 Stancu，2025; Wang 和 Liao，2024）	无需标注伪造样本，可适配未知伪造类型，应用阶段灵活	易受数据分布差异影响，异常阈值设定主观，可能导致合理证据误判
	鲁棒性优化	提升模型在复杂场景下的鉴伪可靠性	对抗训练、跨域自适应优化与噪声鲁棒性增强设计（Krasilnikov 等，2025）	模型抗干扰能力强，输出稳定，证据可信度高	优化过程复杂，可能牺牲检测精度，鲁棒性验证需覆盖多样化的司法证据场景
	可解释性增强	破解深度模型“黑箱”问题，满足法庭证据可采性要求	梯度可视化、注意力权重分析、特征归因方法（Peng 等，2022; Yang 等，2024）	呈现鉴伪依据，降低法庭解释难度	可解释性与检测精度存在权衡，部分解释方法结果主观性强
	伪造定位	定位证据中的伪造区域，为法庭明确证据中真实部分与伪造部分的边界	伪造区域分割模型、像素级异常检测（Waseem 等，2023; Zeng 等，2025）	能区分证据中“真”、“伪”成分，服务于鉴定人	定位精度受图像质量影响大，边缘区域定位准确性不足，需结合其他方法验证

续表 3　按技术原理对深度伪造人脸检测方法的分类(续)

Continued Table 3　Classification of deepfake detection methods by technical principle

技术原理	方法	解决问题	主流方法	技术优势	局限性
对抗策略	对抗防御	抵御伪造技术对鉴伪模型的攻击，确保鉴伪结果稳定可靠	对抗样本训练、防御性蒸馏、模型集成策略(Dong 等，2023; Lei 等，2025)	提升模型抗攻击能力，减少恶意伪造技术对鉴伪结果的干扰	增加模型复杂度与计算成本，部分防御策略可能影响正常的检测效率
	源模型溯源	追溯伪造证据的生成模型，为法庭提供伪造行为溯源线索，辅助追查相关责任人	生成模型指纹提取、模型输出特征匹配溯源(Sun 等，2025)	能为案件侦破提供技术线索，强化伪造行为与责任人的关联性	溯源精度受模型迭代、参数微调影响大，跨平台溯源难度高，技术成熟度不足
	联邦学习	在保护数据隐私的前提下开展多机构鉴伪模型协作	面向跨办案机构的联邦训练框架、模型参数加密传输(Zhang, 2025)	保护涉案数据隐私，实现司法资源共享，提升模型泛化	系统部署与协调成本高，模型训练通信开销大，需顶层协调
	知识蒸馏	构建轻量化鉴伪模型，满足司法一线快速鉴伪需求，降低设备部署与使用门槛	教师—学生模型蒸馏、关键特征迁移学习(Khan 等，2024; Wang 等，2025)	模型体积小，推理速度快，适配司法一线现场终端鉴伪场景，部署成本低	蒸馏过程可能损失部分检测精度，复杂伪造证据鉴伪能力弱于原始大模型

1)特征分析。特征分析是法庭场景下深度伪造人脸图像证据鉴伪的基础，通过时空分析、频率分析、多模态分析、生物信号分析以及深度特征分析(杨少聪 等，2022)等方法，识别图像的物理或生物特性，为刑事侦查提供依据。时空分析检测面部运动或表情不一致，排查伪造痕迹；频率分析揭示频率域中的压缩失真与边缘模糊，甄别虚假特征；多模态分析交叉验证图像、音频与文本，强化证据链判断；生物信号分析利用心率、瞳孔运动等难以伪造的生理特征，提供本质性鉴伪依据；深度特征分析借助深度学习发现伪造区域的细微纹理异常，辅助精准判定。

2)模型能力。模型能力指通过自监督学习、多类别学习、元学习、异常检测、鲁棒性优化、可解释性增强和伪造定位，提升证据取证的可靠性与刑事侦查适配性。其中，自监督学习利用图像修复、时间预测等无标签数据模式，降低对标注数据的依赖；多类别学习区分真实内容与特定伪造类型，实现细粒度分类；元学习基于少量样本快速适应新型伪造，应对新形态虚假证据的挑战；异常检测通过建模正常数据分布标记异常，辅助识别篡改内容；鲁棒性优化增强对噪声、压缩和篡改的抵抗能力，保障复杂场景下的稳定性；可解释性增强通过可视化可疑区域提高结果可信度，满足法庭对过程可追溯的要求；伪造定位精准识别被篡改的面部或合成背景，明确证据篡改范围。

3)对抗策略。对抗策略通过反制与溯源技术，既实现抗规避检测，又可追踪证据来源，为刑事追责提供支撑。具体包括对抗防御、源模型溯源、联邦学习与知识蒸馏。对抗防御抵御微小扰动等规避攻击，提升模型韧性，确保鉴伪结果可靠；源模型溯源通过模型特有噪声等独特痕迹识别生成模型，助力追查伪造源头；联邦学习支持隐私保护下的分布式训练，在保障案件数据安全的同时提升泛化性；知识蒸馏将复杂模型知识迁移至轻量模型，提升移动端与边缘设备部署效率，适配现场取证与庭审快速鉴伪需求。

文献统计如图 9 所示，特征分析是法庭证据鉴伪的基础，占 32.93%。其中，时空特征占比最高，为 37.23%，其次是频率特征占比 31.39%，另有多类模态特征占 13.14%、深度特征占 5.11%、生物信号占 13.14%，这些特征是鉴伪判断的核心依据。

模型能力研究占文献的 53.12%，是刑事侦查适配性优化的核心。其中，鲁棒性占 39.37%、定位追踪占 18.55%、异常检测占 17.65%、自监督学习占 6.79%、可解释性占 9.05%、多类别学习占 6.79%、元学习占 1.81%。以上覆盖模型全周期，契合刑事侦查鉴伪的严苛要求。

对抗策略研究占 13.94%，聚焦攻防对抗与证据溯源。其中，对抗攻击占 53.45%、模型溯源占 34.48%，知识蒸馏占 8.62%、联邦学习占 3.45%，这恰满足刑事侦查实践中多样化的鉴伪与追责需求。

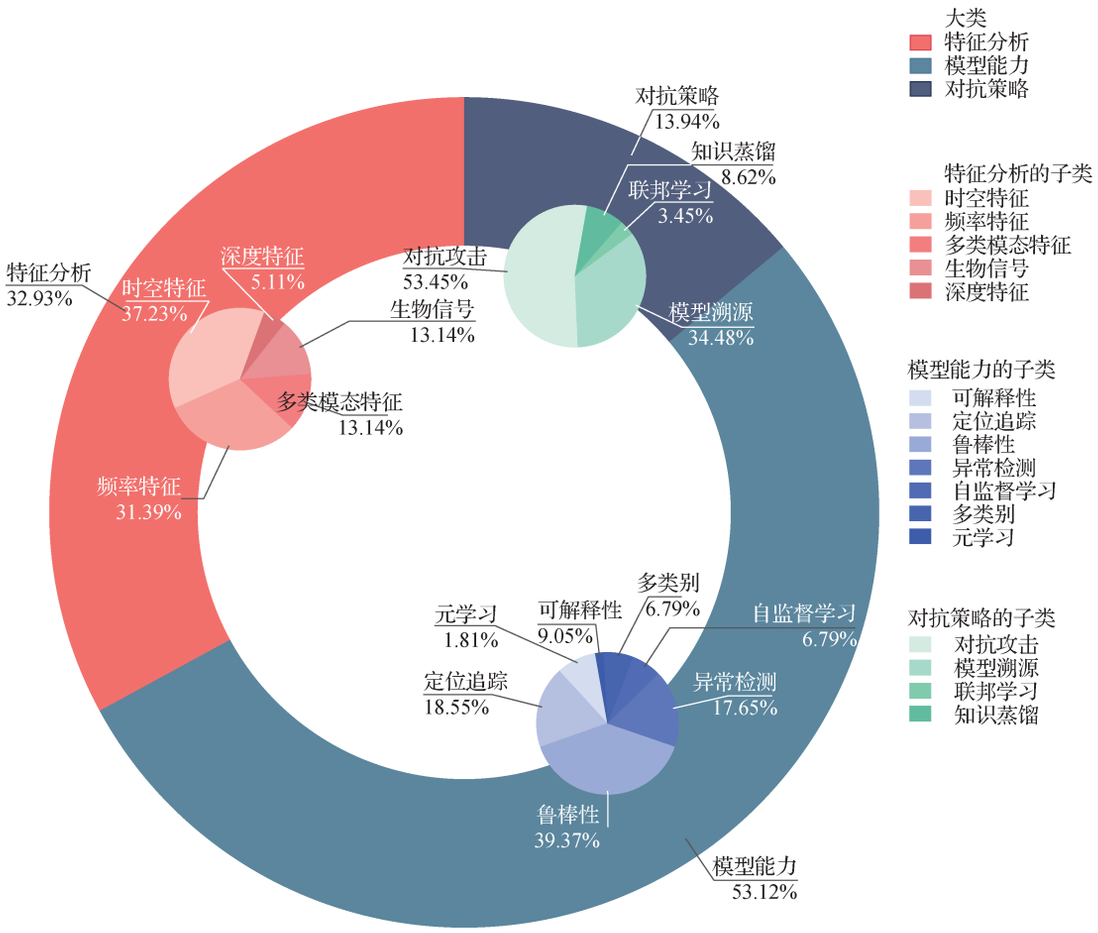


图9　按技术原理对深度伪造人脸检测方法的分类

Fig. 9　Classification of deepfake detection methods based on technical principles

2.2.2　按应用阶段分类

深度伪造人脸的检测技术按照应用阶段可分类为事前预防、事中监测和事后追溯，具体如表4所示。实现从源头阻断到案件响应的端到端刑事司法过程全覆盖，适配刑事侦查与法庭证据。

事前预防依托指纹水印与身份关联，可服务于侦查源头防控。指纹水印在图像创建时嵌入不可见标识符，通过验证其完整性可快速排查篡改。身份关联将伪造内容与伪造人绑定，减少伪造图像进入刑事诉讼程序的风险，降低取证干扰。

事中监测采用特征异常检测与对抗防御，支撑侦查取证。特征异常检测利用时空、频率和多模态方法分析传输的人脸图像，识别并拦截伪造内容，为固定真实证据提供支持；对抗防御应对小噪声干扰和对抗样本，确保模型在规避攻击下仍能稳定输出，满足证据有效性要求。

事后追溯包含4项核心任务，适配法庭证据审查与侦查需求。1）人为决策支持，借助可解释AI输出伪造区域标签和热图，辅助人工审查，确保证据符合质证标准；2）伪造定位，精确标记图像中的篡改区域，为认定篡改范围提供直观依据；3）模型迭代，基于已破案件通过元学习、联邦学习和知识蒸馏更新模型，提升对新型伪造的识别能力；4）公平性验证，评估结果以防止因年龄、性别等属性导致偏差。

图10展示了伪造检测系统的双层环结构阶段划分及文献方法分布，各阶段占比契合刑事司法资源配置逻辑。

事前阶段占19.73%，聚焦主动风险控制。其中，指纹水印占38.64%、身份关联占61.36%，通过水印验证或身份认证实现早期拦截，减少后续取证障碍。

事中阶段占42.15%，构成实时检测核心（李启运等，2019）。其中，异常检测占54.26%、对抗攻击占45.74%，共同保障侦查中图像检测的鲁棒性与适应性。

事后阶段占38.12%，支撑法庭回顾性分析。

表 4　按应用阶段划分的深度伪造人脸检测方法分类

Table 4　Classification of deepfake detection methods by application stage

检测阶段	方法	解决问题	主流方法	技术优势	局限性
事前预防	指纹水印	在原始人脸图像中植入不可见标识，从源头防范伪造人脸图像作为犯罪载体	嵌入鲁棒性数字水印（如DWT域水印）、区块链存证水印（Ge 等，2025; Yang等，2023）	源头把控人脸图像真实性，水印可作为直接溯源依据，说服力强；不影响图像视觉效果	依赖原始数据采集环节介入，无法覆盖已流通的无水印图像；水印易受图像压缩、裁剪等操作破坏，影响验证有效性
	身份关联法	建立人脸图像与真实身份的唯一关联，防范伪造他人身份	人脸生物特征绑定（如虹膜/指纹与人脸多模态关联）、身份信息加密嵌入、权威数据库身份校验（Zhang 等，2025b）	身份关联性强，能快速排查伪造身份的证据；验证逻辑直观，易被法庭理解采纳；适配身份类案件证据鉴伪需求	依赖权威身份数据库支撑，跨区域/跨机构数据库协同难度大；对无身份信息的匿名图像无效，应用阶段受限
事中监测	异常检测	在深度伪造人脸图像实时识别异常篡改/伪造行为，避免伪造证据进入司法程序，降低错案风险	实时帧间一致性监控、图像哈希值动态校验、异常特征实时预警模型（Stamnas 和 Sanchez，2025）	实时性强；无需标注伪造样本，适配未知伪造类型；操作便捷，适配司法一线流转场景	易受环境干扰导致误预警；对轻微伪造行为敏感度不足；对实时监控算法和算力要求高
	对抗防御	抵御伪造技术对证据鉴伪系统的攻击，确保鉴伪过程稳定可靠，保障鉴伪结果的证据效力	对抗样本训练增强鲁棒性、防御性蒸馏模型、多模型集成防御策略（Galdi等，2024）	抗攻击能力强，能应对恶意伪造技术的针对性干扰；鉴伪结果稳定性高	模型训练复杂度高，计算成本大；防御范围有限，难以覆盖所有新型攻击手段；可能牺牲鉴伪效率
事后追溯	人工判定	对疑似伪造的证据进行最终人工鉴定，为法庭提供权威鉴伪意见	法医图像专家人工审核、多专家交叉验证、结合司法案例经验判定（Bharati等，2026）	与其他如指纹、枪弹痕迹等证据鉴定流程一致；可结合案件背景综合判定，灵活性强	专家需结合检测结果与可视化输出，进行图结构分析和案件事实判断，依赖算法且主观性强
	伪造定位	精准定位证据中伪造区域的范围与边界，明确证据“真”、“伪”成分	基于深度网络的伪造区域分割、像素级异常特征定位、伪造边缘轮廓提取（Peng 等，2023; Zhang等，2025a）	鉴伪结果精细化；定位依据客观，可通过可视化呈现增强说服力；更符合法院、检察院、公安以及案件相关人的常识	定位精度受图像质量影响大，模糊图像边缘定位准确性与精确性不足；对轻微篡改的定位敏感度低；需结合其他鉴伪方法交叉验证
	公平性优化	解决不同人群（如不同肤色、性别）人脸伪造鉴伪的偏差问题，保障鉴伪结果的公平性，避免因数据集偏差影响司法公正	公平性约束损失函数优化、多人群均衡数据集训练、跨人群迁移学习（Ju 等，2024; Xu 等，2024b）	鉴伪结果公平性强，避免对特定人群的歧视性误判；符合司法公正的核心要求；提升模型在多样化人群案件中的适配性	公平性与检测精度存在权衡，可能牺牲部分单一人群的鉴伪性能；多人群均衡数据集构建难度大，数据获取成本高
	模型交互法	通过多模型协同交互验证鉴伪结果，解决单一模型鉴伪的局限性，提升复杂伪造证据鉴伪的可靠性与说服力	多模型投票决策、专家系统与深度学习模型交互、跨模态模型协同验证（Sun 和 Li，2025）	鉴伪结果可靠性高，误判率低；能应对复杂多样的伪造场景；可通过多模型结果互补增强法庭说服力	系统架构复杂，部署与维护成本高；多模型协同存在时延，适配实时鉴伪场景受限；鉴定人的水平参差增加不确定性

其中，人工决策占 23.53%、定位追踪占 20.00%、公平性占 48.24%、交互分析占 8.24%，可助力物证鉴定人鉴定伪造图像。

综上所述，将深度伪造人脸图像检测方法按技术原理和应用阶段分类，为新型深度伪造犯罪案件的法庭证据鉴定提供系统且精准的支持。技术原理

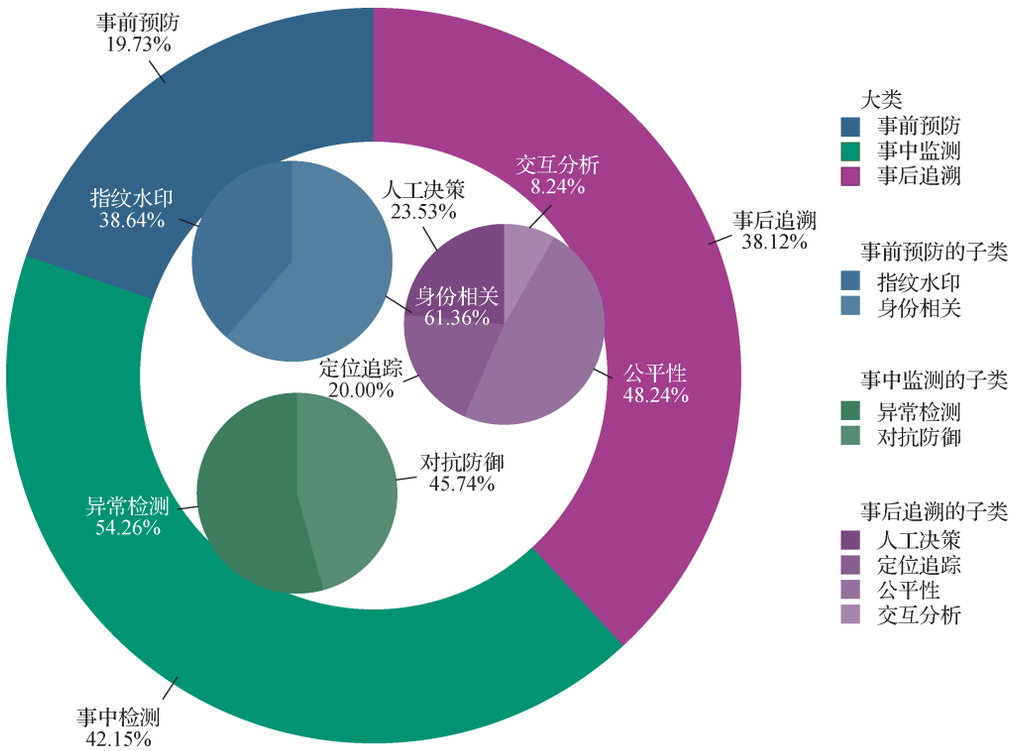


图 10　按应用阶段对深度伪造人脸检测方法的分类

Fig. 10　Classification of deepfake detection methods by application stage

上，特征分析、模型能力与对抗策略3类方法各有适用场景与优劣，明确了鉴伪依据，并为图像、视频和多模态等不同类型证据的技术选择提供指导，解决技术上检测伪造人脸图像的核心问题；应用阶段上，从事前防范、事中控制到事后追溯的全流程划分，契合社会治理与公共安全需求，既能减少伪造证据进入司法程序，保障取证真实，也可通过伪造定位、公平性验证等手段，帮助法庭界定证据真伪，提升质证说服力。以上分类体系有助于司法人员从技术上开展对该新型犯罪载体的研究。

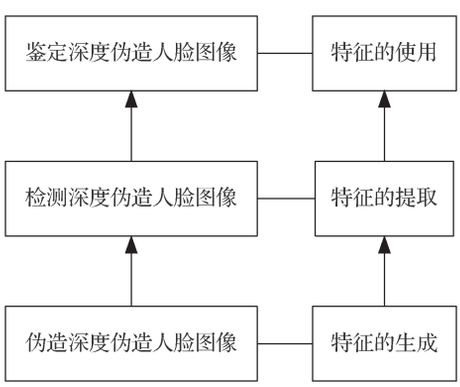


图 11　深度伪造人脸图像的鉴定前提

Fig. 11　Premises for the authentication of deepfake facial images

3　检测与鉴定

深度伪造人脸图像作为一种新型犯罪载体，其生成机制基于生成器与判别器之间的对抗博弈模型，导致伪造样本在特征空间中呈现出复杂的分布模式与高度隐蔽的伪影特征。此类技术的持续迭代，不仅表明传统判别器已难以精准识别特征空间中的伪造痕迹，更说明对该类图像的鉴定无法仅依赖专家对表层视觉特征的人工判读完成。如图 11 所示，唯有依托专业的检测模型，通过对高维特征空间的深度挖掘与系统性特征建模，才能实现科学、可靠的鉴定结论。

在“鉴定依赖检测”这一前提下，需进一步探讨当前检测技术为何尚难达到司法鉴定标准。若暂不考虑深度伪造图像的鉴定标准与技术规范问题，转而从可操作性角度分析，其根本原因在于：指纹、枪弹痕迹（Song 等，2020）、声纹（Morrison，2011）、足迹（Ma 等，2024）、DNA（Agudo 等，2024）以及人脸识别（Macarulla Rodriguez 等，2024）等物证均已纳入贝叶斯推理框架，且以似然比作为标准化的证据强度的表达形式（Van Lierop 等，2024），但是深度伪造人脸

图像尚未基于此进行深入研究。似然比框架基于控方假设与辩方假设进行推断，即

$$LR = \frac{f(score \mid Hp, I)}{f(score \mid Hd, I)} \tag{1}$$

式中，LR 代表似然比（likelihood ratio），$score$ 是现场提取的指纹检材与嫌疑人指纹样本之间的相似度，Hp 代表控方假设（prosecution hypothesis），Hd 代表辩方假设（defence hypothesis）。

例如，在犯罪现场提取一枚指纹并怀疑其属于犯罪嫌疑人 X 时，存在两种对立假设：控方假设为“该指纹来源于嫌疑人 X”，辩方假设为“该指纹来源于某一随机个体”。通过提取该指纹的特征，可计算其与嫌疑人指纹之间的相似度。依托已有数据库，可预先建立两类概率密度函数：1）两枚指纹同源（来自同一人）时相似度的分布函数；2）异源（来自不同人）时相似度的分布函数。将实际观测到的相似度分别代入这两个函数，所得同源概率密度与异源概率密度之比，即为该指纹证据的似然比，用以量化其证据强度。如 $LR = 10^3$，意味着该指纹相似度的观测结果，在“指纹属于嫌疑人 X”的前提下发生的可能性，是“指纹属于随机个体”前提下的 1 000 倍，属于强证据，能够有力支撑控方的主张。目前，在鉴定科学领域，所有证据普遍采用似然比作为统一的评估指标。

深度伪造人脸检测模型的典型输出为一组二分类概率值（p，1−p），其中，p 代表输入图像为深度伪造内容的预测概率，1−p 则对应图像为真实拍摄内容的预测概率。在工程化应用中，模型常以 0.5 为判定阈值，直接输出“伪造”或“真实”的离散结论。需明确的是，此类基于经验风险最小化的概率输出与阈值判定方式，并未直接纳入贝叶斯推理的完整分析框架——其本质是模型基于训练数据集的特征学习与概率拟合，未结合司法场景下的先验信息完成后验概率的推演。

类比指纹鉴定的法庭科学统计分析范式，深度伪造图像作为新型电子证据的鉴定工作，需构建一组相互对立的检验假设，以支撑证据证明力的量化评估。针对不同案件的待证事实，核心假设可分为两类。

1）基础属性假设。控方假设：送检的涉案人脸图像为深度伪造生成；辩方假设：送检的涉案人脸图像为真实拍摄所得。

2）关联属性假设。控方假设：送检的、与嫌疑人 X 面部特征高度相似的人脸图像为深度伪造生成；辩方假设：送检的、与嫌疑人 X 面部特征高度相似的人脸图像为 X 本人的真实人脸图像。

上述两类假设的量化分析，均需依托深度伪造人脸检测技术提取图像的鉴别特征或者分数，并基于特征或者分数构建概率密度函数，进而通过似然比等指标实现证据强度的客观量化。

上述计算过程可视为深度伪造图像证据的量化分析环节。鉴于司法实务人员对技术证据在理解与采信方面的需求，可通过检测流程可视化、特征分布可视化等方式，将抽象的量化结果转化为直观的视觉呈现，如图 12 所示。

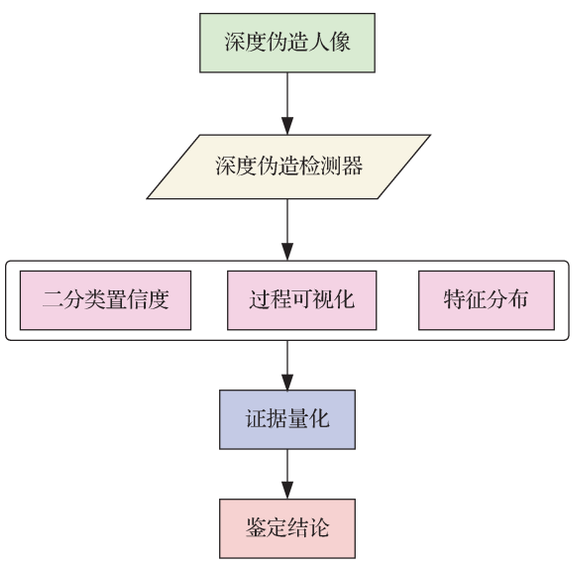


图 12　检测与鉴定的关系

Fig. 12　Relationship between detection and identification

尽管深度伪造人脸检测技术能够生成伪造区域热图、异常特征标注等可视化输出，其实质仍局限于技术特征的识别与量化，无法自主判断这些特征所承载的法律意义。例如，在诈骗案件中，检测到的人脸篡改痕迹是否指向被害人身份信息的伪造；在诬告案件中，图像的篡改区域是否影响关键时间戳的真实性。此类问题均需鉴定专家结合刑事侦查知识与具体案情进行专业解读。此外，涉案图像元数据的合法性、检测结果与待证事实之间的关联性，以及检测模型是否存在年龄、性别等属性上的系统性偏见，亦需由专家开展综合评估。唯有经过上述审查程序，方可将技术层面的量化结果转化为符合刑事侦查与司法规范要求的鉴定意见。

需要强调的是，人类对深度伪造内容的主观辨

识能力有限,且随着伪造技术的不断迭代而持续下降,这进一步凸显了技术检测与专家审查相结合的必要性。所以,未来研究需要致力于通过证据量化与专家审查,实现从技术特征检测向司法证据审查的转化。

4 结 语

本文所探讨的深度伪造人脸图像生成技术,已从生成对抗网络等早期技术演进至扩散模型与多模态驱动阶段,并初步构建起完整技术流程。该流程可实现篡改检测、工具追踪与来源溯源,充分满足刑侦证据的可追溯性需求,且配套检测技术兼具高精度与可解释性,能生成直观篡改标记并定位伪造范围。同时,文中提出的端到端框架覆盖事前水印管控、事中异常检测与事后溯源全环节,有效提升了鉴定工作的科学性。最后,本文基于鉴定领域对证据形式所要求的贝叶斯框架,提出基于检测技术的深度伪造图像鉴定可行性方法。综上,本文通过相关技术与框架的研究,推动了数字取证与法庭证据鉴定的深度融合,为应对深度伪造技术对刑事证据构成的挑战提供方法参考和实践启示。

参考文献(Reference)

Amundsen A E and Ovens K M. 2017. Forensics analysis of WI-FI communication traces in mobile devices//Proceedings of 2017 IEEE International Conference on Big Data (Big Data). Boston, USA: IEEE: 3632-3637 [DOI: 10.1109/BigData.2017.8258357]

Abbas F and Taeihagh A. 2024. Unmasking deepfakes: a systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications, 252: #124260 [DOI: 10.1016/j.eswa.2024.124260]

Agudo M M, Fantinato C, Roseth A, Aanes H, Gill P, Fonneløp A E, et al. 2024. A comparison of likelihood ratios calculated from surface DNA mixtures using MPS and CE technologies. Forensic Science International: Genetics, 73: #103111 [DOI: 10.1016/j.fsigen.2024.103111]

Alhajjar E and Lee K. 2022. The U.S. cyber threat landscape. European Conference on Cyber Warfare and Security, 21(1): 18-24 [DOI: 10.34190/eccws.21.1.197]

Almars A M. 2021. Deepfakes detection techniques using deep learning: a survey. Journal of Computer and Communications, 9(5): 20-35 [DOI: 10.4236/jcc.2021.95003]

Alnaim N M, Almutairi Z M, Alsuwat M S, Alalawi H H, Alshobaili A and Alenezi F S. 2023. DFFMD: a deepfake face mask dataset for infectious disease era with deepfake detection algorithms. IEEE Access, 11: 16711-16722 [DOI: 10.1109/ACCESS.2023.3246661]

Bharati N, Wong P, Mostéfaoui S K, Kbaier D and Collie J. 2026. Explainable deepfake detection: a multi-model framework with human-interpretable rationales for legal investigation purposes. Machine Learning with Applications, 23: #100819 [DOI: 10.1016/j.mlwa.2025.100819]

Boháček M and Farid H. 2022. Protecting President Zelenskyy against deep fakes [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2206.12043.pdf

Bray S D, Johnson S D and Kleinberg B. 2023. Testing human ability to detect "deepfake" images of human faces. Journal of Cyber Security, 9(1) [DOI: 10.1093/cybsec/tyad011]

Brown A, Johnson S D and Tuptuk N. 2025. A scoping study: crime and connected and autonomous vehicles. Crime Science, 14(1): #2 [DOI: 10.1186/s40163-025-00245-x]

Bunluesakdikul P, Mahanan W, Sungunnasil P and Sangamuang S. 2025. Deepfake video detection: a novel approach via NLP-based classification. International Journal of Computational Intelligence and Applications, 24(2): #2550001 [DOI: 10.1142/S1469026825500014]

Cai Z, Stefanov K, Dhall A and Hayat M. 2022. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization//Proceedings of 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA): 1-10 [DOI: 10.1109/DICTA56598.2022.10034605]

Çiftçi U A and Demir İ. 2023. Deepfake satellite imagery detection with multi-attention and super resolution//2023 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Pasadena, USA: IEEE: 4871-4874 [DOI: 10.1109/IGARSS52108.2023.10282927]

Dagar D and Vishwakarma D K. 2023. Div-Df: a diverse manipulation deepfake video dataset//Proceedings of 2023 Global Conference on Information Technologies and Communications (GCITC). Bangalore, India: IEEE: 1-6 [DOI: 10.1109/GCITC60406.2023.10426446]

Dayal S B and van Niekerk B. 2021. Deepfake video detection//Proceedings of the 20th European Conference on Cyber Warfare and Security. 2021 [DOI: 10.34190/ECCWS.21.110]

Deng L W, Wu B D and Wang J D. 2025. A multi-label classification method combined with texture enhancement for deepfake face detection. Multimedia Systems, 31(6): #410 [DOI: 10.1007/s00530-025-01996-y]

Deng W, Feng J, Huang D, Kan M, Sun Z, Zheng F, et al. 2022. Biometric recognition//Proceedings of the 16th Chinese Conference on Biometrics Recognition (CCBR). Beijing, China: [s.n.] [DOI: 10.1007/978-3-031-20233-9]

Ding F, Kuang R S, Zhou Y, Sun L, Zhu X G and Zhu G P. 2024. A

survey of Deepfake and related digital forensics. Journal of Image and Graphics, 29(2): 295-317 (丁峰, 匡仁盛, 周越, 孙珑, 朱小刚, 朱国普. 2024. 深度伪造及其取证技术综述. 中国图象图形学报, 29(2): 295-317) [DOI: 10.11834/jig.230088]

Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. 2020. The DeepFake detection challenge (DFDC) dataset [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2006.07397.pdf

Dong J H, Wang Y, Lai J H and Xie X H. 2023. Restricted black-box adversarial attack against DeepFake face swapping. IEEE Transactions on Information Forensics and Security, 18: 2596-2608 [DOI: 10.1109/TIFS.2023.3266702]

Eisenstein M. 2024. Seven technologies to watch in 2024. Nature, 625(7996): 844-848 [DOI: 10.1038/d41586-024-00173-x]

Galdi C, Panariello M, Todisco M and Evans N. 2024. 2D-malafide: adversarial attacks against face deepfake detection systems//Proceedings of 2024 International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany: IEEE: 1-7 [DOI: 10.1109/BIOSIG61931.2024.10786754]

Garg D and Gill R. 2025. Unmasking deepfakes: a review of current datasets, tools, and detection features//Proceedings of the 6th International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06), Held in Uttarakhand, India, 259: 1737-1748 [DOI: 10.1016/j.procs.2025.04.129]

Ge J W, Cao J X, Zhao Z X and Liu B. 2025. FSD-GAN: generative adversarial training for face swap detection via the latent noise fingerprint. Journal of Computer Science and Technology, 40(2): 397-412 [DOI: 10.1007/s11390-024-3337-8]

Guan W N, Wang W, Peng B, Dong J and Tan T N. 2025. ST-SBV: spatial-temporal self-blended videos for deepfake detection//Proceedings of the 7th Chinese Conference on Pattern Recognition and Computer Vision. Urumqi, China: Springer: 274-288 [DOI: 10.1007/978-981-97-8620-6_19]

He Y, Gan B, Chen S, Zhou Y, Yin G, Song L, et al. 2021. ForgeryNet: a versatile benchmark for comprehensive forgery analysis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE: 4358-4367 [DOI: 10.1109/CVPR46437.2021.00434]

Huang D Z and Zhang Y Q. 2024. Learning meta model for strong generalization deepfake detection//Proceedings of 2024 International Joint Conference on Neural Networks. Yokohama, Japan: IEEE: 1-8 [DOI: 10.1109/IJCNN60899.2024.10651482]

Iqbal A, Olegård J and Ghimire R. 2020. Digital forensic evidence—the missing link in threat modeling//Proceedings of 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI). Sakheer, Bahrain: IEEE: 1-5 [DOI: 10.1109/ICDABI51230.2020.9325650]

Joel E J, Kumar K P, Silar M M and Rohan J. 2025. Flexible multimodality deepfake identification with dynamic learning and cross-modality inconsistency//Proceedings of the 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA 2025). Coimbatore, India: IEEE: 1-6 [DOI: 10.1109/ICAECA63854.2025.11012276]

Ji S. 2025. #MeToo in an AI-generated deepfake sexual violence era in South Korea. Women's Studies International Forum, 112: #103146 [DOI: 10.1016/j.wsif.2025.103146]

Jia S, Li X and Lyu S. 2022. Model attribution of face-swap deepfake videos//Proceedings of the 29th IEEE International Conference on Image Processing (ICIP). Bordeaux, France: IEEE Computer Society: 2356-2360 [DOI: 10.1109/ICIP46576.2022.9897972]

Jiang L, Li R, Wu W, Qian C and Loy C C. 2020. DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 2886-2895 [DOI: 10.1109/CVPR42600.2020.00296]

Jin X, Wu N, Jiang Q, Kou Y R, Duan H X, Wang P M, et al. 2024. A dual descriptor combined with frequency domain reconstruction learning for face forgery detection in deepfake videos. Forensic Science International: Digital Investigation, 49: #301747 [DOI: 10.1016/j.fsidi.2024.301747]

Jin X L, Ye D P and Chen C X. 2021. Countering spoof: towards detecting deepfake with multidimensional biological signals. Security and Communication Networks, 2021(1): #6626974 [DOI: 10.1155/2021/6626974]

Ju Y, Hu S, Jia S, Chen G H and Lyu S W. 2024. Improving fairness in deepfake detection//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 4643-4653 [DOI: 10.1109/WACV57701.2024.00459]

Karras T, Aila T, Laine S and Lehtinen J. 2018. Progressive growing of GANs for improved quality, stability, and variation//Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada: ICLR

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]

Kaur M and Saini M. 2024. Role of artificial intelligence in the crime prediction and pattern analysis studies published over the last decade: a scientometric analysis. Artificial Intelligence Review, 57(8): #202 [DOI: 10.1007/s10462-024-10823-1]

Khalid H, Tariq S, Kim M and Woo S S. 2021. FakeAVCeleb: a novel audio-video multimodal deepfake dataset [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2108.05080.pdf

Khan S S, Hossain R, Bishal S H and Khan R. 2024. Smartphone-based deepfake detection through transfer learning and lightweight knowledge distillation technique//Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT). Cox's Bazar, Bangladesh: IEEE: 675-680 [DOI: 10.1109/

ICCIT64611.2024.11022522]

Khodabakhsh A, Ramachandra R, Raja K, Wasnik P and Busch C. 2018. Fake face detection methods: can they Be generalized?//Proceedings of 2018 International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany: IEEE: 1-6 [DOI: 10.23919/BIOSIG.2018.8553251]

Korshunov P and Marcel S. 2018. DeepFakes: a new threat to face recognition? Assessment and detection [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/1812.08685.pdf

Krasilnikov M, Nikitin M and Konushin A. 2025. VCF: a real-world video conference deepfake benchmark for face-swap detection and robustness evaluation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-2/W9-2025: 169-174 [DOI: 10.5194/isprs-archives-XLVIII-2-W9-2025-169-2025]

Krishna S, Dubrosa F and Milanaik R. 2024. Rising threats of AI-driven child sexual abuse material. Pediatrics, 153(2): #2023063954 [DOI: 10.1542/peds.2023-063954]

Kwon P, You J, Nam G, Park S and Chae G. 2021. KoDF: a large-scale Korean DeepFake detection dataset//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE: 10724-10733 [DOI: 10.1109/ICCV48922.2021.01057]

Le T N, Nguyen H H, Yamagishi J and Echizen I. 2021. OpenForensics: large-scale challenging dataset for multi-face forgery detection and segmentation In-the-wild [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2107.14480.pdf

Lei S Z, Song J F, Feng F Y, Yan Z Y and Wang A X. 2025. Deepfake face detection and adversarial attack defense method based on multi-feature decision fusion. Applied Sciences, 15(12): #6588 [DOI: 10.3390/app15126588]

Levanon L and Tully G. 2025. Fingerprint analysis and reporting in legal trials: a critical re-evaluation. Criminal Law Forum, 36(1): 1-32 [DOI: 10.1007/s10609-025-09497-3]

Li Q Y, Ji Q G and Hong S D. 2019. FastFace: a real-time robust algorithm for face detection. Journal of Image and Graphics, 24(10): 1761-1771 (李启运, 纪庆革, 洪赛丁. 2019. FastFace: 实时鲁棒的人脸检测算法. 中国图象图形学报, 24(10): 1761-1771) [DOI: 10.11834/jig.180662]

Li Y, Yang X, Sun P, Qi H and Lyu S. 2020. Celeb-DF: a large-scale challenging dataset for DeepFake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Li Y Z, Yang X, Sun P, Qi H G and Lyu S W. 2019. Celeb-DF: a large-scale challenging dataset for DeepFake forensics [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/1909.12962.pdf

Liang B Y, Wang Z Y, Huang B J, Zou Q, Wang Q and Liang J J. 2023. Depth map guided triplet network for deepfake face detection. Neural Networks, 159: 34-42 [DOI: 10.1016/j.neunet.2022.11.031]

Liu X Y, Song P C, Lu P and Wang Y J. 2024. Meta-learning with relation embedding for few-shot deepfake detection. IEEE Access, 12: 180135-180145 [DOI: 10.1109/ACCESS.2024.3499353]

Ma X, Luo Y P, Yu Y, Liu Y W, Li S and Li H M. 2024. Similarity quantification of bare footprint based on linear measurement and shape context contour combined score-based likelihood ratio evaluation. Forensic Science International, 356: #111967 [DOI: 10.1016/j.forsciint.2024.111967]

Macarulla Rodriguez A, Geradts Z, Worring M and Unzueta L. 2024. Improved likelihood ratios for face recognition in surveillance video by multimodal feature pairing. Forensic Science International: Synergy, 8: #100458 [DOI: 10.1016/j.fsisyn.2024.100458]

Mittal G, Hegde C and Memon N. 2024. Gotcha: real-time video deepfake detection via challenge-response//Proceedings of 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P): 1-20 [DOI: 10.1109/EuroSP60621.2024.00009]

Morrison G S. 2011. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). Speech Communication, 53(2): 242-256 [DOI: 10.1016/j.specom.2010.09.005]

Narayan K, Agarwal H, Thakral K, Mittal S, Vatsa M and Singh R. 2022. DeePhy: on deepfake phylogeny//Proceedings of 2022 IEEE International Joint Conference on Biometrics (IJCB). Dhabi, United Arab Emirates: IEEE: 1-10 [DOI: 10.1109/IJCB54206.2022.10007968]

Neves J C, Tolosana R, Vera-Rodriguez R, Lopes V, Proença H and Fierrez J. 2020. GANprintR: improved fakes and evaluation of the state of the art in face manipulation detection. IEEE Journal of Selected Topics in Signal Processing, 14(5): 1038-1048 [DOI: 10.1109/JSTSP.2020.3007250]

Ni Y, Zeng W, Xia P, Yang G S and Tan R C. 2024. A deepfake detection algorithm based on fourier transform of biological signal. Computers, Materials and Continua, 79(3): 5295-5312 [DOI: 10.32604/cmc.2024.049911]

Peng B, Fan H, Wang W, Dong J, Li Y and Lyu S. 2021. DFGC 2021: a DeepFake game competition//Proceedings of 2021 IEEE International Joint Conference on Biometrics (IJCB). Shenzhen, China: IEEE: 1-8 [DOI: 10.1109/IJCB52358.2021.9484387]

Peng B, Lyu S W, Wang W and Dong J. 2022. Counterfactual image enhancement for explanation of face swap deepfakes//Proceedings of the 5th Chinese Conference on Pattern Recognition and Computer Vision. Shenzhen, China: Springer: 492-508. Lecture Notes in Computer Science, #13535 [DOI: 10.1007/978-3-031-18910-4_40]

Peng F, Zhang X L and Long M. 2023. F2DLNet: a face forgery detection and localization network based on SSIM error maps//Proceed-

ings of the 4th International Conference on Machine Learning for Cyber Security. Guangzhou, China: Springer: 355-369 [DOI: 10.1007/978-3-031-20099-1_30]

Pesetski A. 2020. Deepfakes: a new content category for a digital age. The William and Mary Bill of Rights Journal, 29(2): 503-532

Qiao T, Xie S C, Chen Y L, Retraint F and Luo X Y. 2024. Fully unsupervised deepfake video detection via enhanced contrastive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7): 4654-4668 [DOI: 10.1109/TPAMI.2024.3356814]

Ramanaharan R, Guruge D B and Agbinya J I. 2025. DeepFake video detection: insights into model generalisation — a systematic review. Data and Information Management: #100099 [DOI: 10.1016/j.dim.2025.100099]

Ramluckan T. 2024. Deepfakes: the legal implications. International Conference on Cyber Warfare and Security, 19(1): 282-288 [DOI: 10.34190/iccws.19.1.2099]

Rosca C M and Stancu A. 2025. AI anomaly-based deepfake detection using customized mahalanobis distance and head pose with facial landmarks. Applied Sciences, 15(17): #9574 [DOI: 10.3390/app15179574]

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Niessner M. 2019. FaceForensics++: learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]

Schiff K J, Schiff D S and Bueno N S. 2025. The liar's dividend: can politicians claim misinformation to evade accountability? American Political Science Review, 119(1): 71-90 [DOI: 10.1017/S0003055423001454]

Seow J W, Lim M K, Phan R C W and Liu J K. 2022. A comprehensive overview of deepfake: generation, detection, datasets, and opportunities. Neurocomputing, 513: 351-371 [DOI: 10.1016/j.neucom.2022.09.135]

Shahzad S A, Hashmi A, Peng Y T, Tsao Y and Wang H M. 2025. AV-lip-sync+: leveraging AV-HuBERT to exploit multimodal inconsistency for deepfake detection of frontal face videos. IEEE Transactions on Human-Machine Systems, 55(6): 973-982 [DOI: 10.1109/THMS.2025.3618409]

Shao C, Zhang F, Wang J X and Zhang B K. 2025. Sample based contrastive learning for DeepFake detection//Proceedings of the 2nd International Conference on Networks, Communications and Intelligent Computing (NCIC 2024). Singapore, Singapore: Springer: 993-1000 [DOI: 10.1007/978-981-96-5006-4_86]

Song H, Huang S, Dong Y and Tu W. 2023. Robustness and generalizability of deepfake detection: a study with diffusion models [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2309.02218.pdf

Song J, Chen Z, Vorburger T V and Soons J A. 2020. Evaluating likelihood ratio (LR) for firearm evidence identifications in forensic science based on the congruent matching cells (CMC) method. Forensic Science International, 317: #110502 [DOI: 10.1016/j.forsciint.2020.110502]

Stamnas S and Sanchez V. 2025. DiffFake: exposing deepfakes using differential anomaly detection//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Tucson, USA: IEEE: 647-657 [DOI: #10.1109/WACVW65960.2025.00079]

Sun C L and Li W J. 2025. A two-stage interaction approach for enhancing generalization of deepfake detection. Multimedia Systems, 31(5): #359 [DOI: 10.1007/s00530-025-01942-y]

Sun Z M, Chen S, Yao T P, Yi R, Ding S H and Ma L Z. 2025. Rethinking open-world DeepFake attribution with multi-perspective sensory learning. International Journal of Computer Vision, 133(2): 628-651 [DOI: 10.1007/s11263-024-02184-7]

Van Der Sloot B and Wagensveld Y. 2022. Deepfakes: regulatory challenges for the synthetic society. Computer Law & Security Review, 46: #105716 [DOI: 10.1016/j.clsr.2022.105716]

Van Lierop S, Ramos D, Sjerps M and Ypma R. 2024. An overview of log likelihood ratio cost in forensic science-where is it used and what values can we expect? Forensic Science International: Synergy, 8: #100466 [DOI: 10.1016/j.fsisyn.2024.100466]

Venema A E and Geradts Z J. 2020. Digital forensics, deepfakes, and the legal process. Scitech Lawyer, 16(4): 14-17, 23

Wang C P, Meng L S, Xia Z Q, Ren N and Ma B. 2025. Cross-domain deepfake detection based on latent domain knowledge distillation. IEEE Signal Processing Letters, 32: 896-900 [DOI: 10.1109/LSP.2025.3540941]

Wang H Y, Li S, He J, Qian Z X, Zhang X P and Fan S L. 2024. Exploring depth information for detecting manipulated face videos [EB/OL]. [2026-01-10]. http://arxiv.org/pdf/2411.18572.pdf

Wang Y F and Liao G J. 2024. Deepfake video detection based on image source anomaly//Proceedings of the 2nd IEEE International Conference on Image Processing and Computer Applications (ICIPCA). Shenyang, China: IEEE: 397-401 [DOI: 10.1109/ICIPCA61593.2024.10709022]

Waseem S, Abu-Bakar S A R S, Omar Z, Ahmed B A, Baloch S and Hafeezallah A. 2023. Multi-attention-based approach for deepfake face and expression swap detection and localization. EURASIP Journal on Image and Video Processing, 2023(1): #14 [DOI: 10.1186/s13640-023-00614-z]

Xu Y, Terhörst P, Pedersen M and Raja K. 2024a. Analyzing fairness in deepfake detection with massively annotated databases. IEEE Transactions on Technology and Society, 5(1): 93-106 [DOI: 10.1109/TTS.2024.3365421]

Xu Y T, Liang J, Sheng L J and Zhang X Y. 2024b. Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection. International Journal of Computer Vision, 132(12): 5663-5680 [DOI: 10.1007/s11263-024-02054-2]

Yan Z, Yao T, Chen S, Zhao Y, Fu X, Zhu J, et al. 2024. DF40:

toward next-generation deepfake detection [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2406.13495.pdf

Yang J, Sun Y R, Mao M Y, Bai L Z, Zhang S Y and Wang F. 2023. Model-agnostic method: exposing deepfake using pixel-wise spatial and temporal fingerprints. IEEE Transactions on Big Data, 9(6): 1496-1509 [DOI: 10.1109/TBDATA.2023.3284272]

Yang S C, Wang J, Sun Y L and Tang J H. 2022. Multi-level features global consistency for human facial deepfake detection. Journal of Image and Graphics, 27(9): 2708-2720 (杨少聪, 王健, 孙运莲, 唐金辉. 2022. 多级特征全局一致性的伪造人脸检测. 中国图象图形学报, 27(9): 2708-2720) [DOI: 10.11834/jig.211254]

Yang X, Li Y and Lyu S. 2019. Exposing deep fakes using inconsistent head poses//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 8261-8265 [DOI: 10.1109/ICASSP.2019.8683164]

Yang Y Q, Joukovsky B, Mogrovejo J O, Tuytelaars T and Deligiannis N. 2024. SNIPPET: a framework for subjective evaluation of visual explanations applied to DeepFake detection. ACM Transactions on Multimedia Computing, Communications and Applications, 20(8): #253 [DOI: 10.1145/3665248]

Yao W D, Li P C, Zhao Y and Wu H C. 2025. Review of research on face deepfake detection methods. Journal of Image and Graphics, 30(7): 2343-2363 (姚文达, 李盼池, 赵娅, 吴洪超. 2025. 人脸深度伪造检测方法研究综述. 中国图象图形学报, 30(7): 2343-2363) [DOI: 10.11834/jig.240586]

Zeng S D, Yi J Y, Tao J H, He J Y, Lian Z, Liang S, et al. 2025. Adversarial training and gradient optimization for partially deepfake audio localization//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10890470]

Zhang B L, Yin Q L, Lu W and Luo X Y. 2025a. Deepfake detection and localization using multi-view inconsistency measurement. IEEE Transactions on Dependable and Secure Computing, 22(2): 1796-1809 [DOI: 10.1109/TDSC.2024.3472064]

Zhang C H. 2025. Federated learning-based cross-domain face forgery detection for consumer IOT systems//Proceedings of the 4th International Conference on Electronic Information Engineering and Data Processing (Mohd Zain A B, Chen L, eds.) (EIEDP 2025. Kuala Lumpur, Malaysia: SPIE: #238 [DOI: 10.1117/12.3067515]

Zhang Z Y, Zhang J, Zhou W B, Zhou X H, Guo Q, Zhang W M, et al. 2025b. FaceTracer: unveiling source identities from swapped face images and videos for fraud prevention. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(12): 12021-12037 [DOI: 10.1109/TPAMI.2025.3601141]

Zhou C Z, Zhang D, Chen Z, Zhao J Q, Zhang G B and Wei C. 2025. A speaking face deepfake detection method, medium and device based on cross-modal learning: #CN120126223[P]. 2025

Zi B J, Chang M H, Chen J J, Ma X J and Jiang Y G. 2020. WildDeepfake: a challenging real-world dataset for deepfake detection//Proceedings of the 28th ACM International Conference on Multimedia: [s.l.]: [s.n.]

作者简介

郭甜利,女,博士研究生,主要研究方向为证据量化和深度伪造。E-mail: guotianli1@163.com

唐云祁,通信作者,男,教授,主要研究方向为电子数据检验。E-mail:tangyunqi@ppsuc.edu.cn

李纪松,男,硕士研究生,主要研究方向为计算机图形学。E-mail: 2023211410@stu.ppsuc.edu.cn