中图法分类号： 文献标识码： 文章编号：1006-8961(XXXX)XX-0001-20

论文引用格式：. XXXX. A Review of Development and Application of Deepfake Face Detection Technology. Journal of Image and Graphics, XX(XX):0001-0020(. XXXX. 深度伪造人脸检测技术发展与应用综述. 中国图象图形学报, XX(XX):0001-0020)[DOI:10.11834/jig.250556]

深度伪造人脸检测技术发展与应用综述

李卫斌,冯雨婷,侯彪,焦李成西安电子科技大学人工智能学院,西安 710126

摘　要：由深度伪造技术(deepfake)生成的伪造图像和视频在网络上广泛传播,尤其针对名人的视频往往被用来损害他人名誉、引导舆论,极大威胁社会稳定,带来了诸多社会隐患。作为应对手段,深度伪造检测技术已成为学术界和业界的研究热点。本论文主要聚焦于深度伪造人脸检测任务,介绍了常用的伪造方法类型,按照模型基础架构将深度伪造人脸检测技术分为基于卷积神经网络(convolutional neural networks, CNN)的检测方法、基于Transformer的检测方法和新型范式三大类。基于卷积神经网络的检测方法是当前的主流方法,模型结构多样且成熟;基于Transformer的检测方法因其在长距离建模的优势,近年来快速发展;新型范式包括自监督/无监督学习方法和大模型检测方法,自监督/无监督学习方法能够有效避免特定数据、特定伪造方法所造成的偏差,大模型检测方法因加入了文本特征,能够提升检测模型的泛化性能和可解释性。此外,总结了深度伪造人脸检测领域的经典数据集和新一代多模态数据集,以及检测模型在分类性能、泛化性能和应用方面的评估指标。在实际应用方面,梳理了深度伪造人脸检测技术的四大应用场景,回顾了国内在深度合成相关的法律法规。最后,总结了深度伪造人脸检测领域的主要矛盾,并结合发展现状,提出与大模型融合、可解释性和泛化性、模型轻量化及行业法规细化等重要发展趋势和研究方向。相关内容总结开源地址:https://github.com/ytttkskr/2025-deepfake-detection。

关键词：深度伪造人脸;深度伪造人脸检测;卷积神经网络;Transformer;自监督/无监督学习;大模型

A Review of Development and Application of Deepfake Face Detection Technology

Abstract: With the rapid maturation and democratization of deep learning technologies, deepfake generation has evolved from a niche technical curiosity into a pervasive global phenomenon. As these tools become increasingly accessible, forged images and videos are proliferating across the Internet at an unprecedented scale. Among these, deepfake images and videos targeting human faces—particularly those involving identity manipulation of public figures and celebrities—are frequently exploited for malicious intent. These applications range from defamation and non-consensual pornography to the manipulation of public opinion and the dissemination of disinformation, thereby posing severe threats to digital trust, individual reputation, and social stability. Consequently, the development of robust deepfake face detection (DFD) technology has emerged as a critical research frontier and a strategic priority for both the academic community and the industrial sector. This review provides a systematic and comprehensive survey of the current landscape of deepfake face detection, with a specific focus on the prevalent manipulation technique of **face swapping**. The paper first scrutinizes the evolution of underlying generation mechanisms, detailing typical forgery methods including variational autoencoders (VAEs), generative adversarial networks (GANs), and the emerging diffusion models (DMs). According to changed attributes, face forgery methods can be summarized to four common forms: facial transfer, facial swapping, facial reenactment and facial editing. Subsequently, the review proposes a structured taxonomy of detection methodologies, categorizing them into three primary

收稿日期:2025-11-05;修回日期:2026-01-19

streams based on model architecture: **CNN-based methods, Transformer-based methods, and the new paradigms. CNN-based methods** currently represent the mainstream approach. This paper analyzes their diverse structural characteristics, further subdividing them into six distinct sub-categories based on their feature extraction strategies, such as basic CNN architecture and frequency domain analysis. And the advantages, disadvantages and applicable scenarios of each structure were analyzed. **Transformer-based methods** are highlighted for their recent rapid advancement. This review discusses the Transformer-based detection models in two branches. One type consists solely of the transformer structure. By adding modules and enhancing the structure, the detection accuracy can be improved and the computing efficiency can also be enhanced. While the other is CNN-Transformer hybrid architecture, make full use of the local receptive field of the CNN architecture and the global modeling capability of the Transformer. **The new paradigms** section explores cutting-edge innovations, specifically Self-supervised/Unsupervised Learning and combined with large language models (LLMs). The paper elucidates how self-supervised/unsupervised learning mitigates the dependency on labeled data, effectively avoiding the overfitting bias caused by specific datasets or forgery types. Furthermore, it examines how integrating large language models introduces semantic understanding and text features, significantly enhancing both the generalization capabilities and the interpretability of detection decisions. In the domain of data infrastructure, the paper traces the evolution of deepfake datasets. It contrasts classic datasets, which rely solely on visual data (images or videos), with the new generation of multimodal datasets that have emerged over the past two years. These modern datasets, enriched with human annotations and descriptive text prompts, are pivotal for training more context-aware detection models, thereby fundamentally redefining the objective from simplistic perceptual categorization to a more demanding exercise in cognitive reasoning and contextual justification. By using the multimodal dataset to enhance, the interpretability of the detection method will be improved. Evaluation is also a critical component of this survey. The paper provides a comprehensive summary of evaluation metrics, distinguishing between classification performance, generalization performance and practical deployment metrics. While acknowledging that DFD is a binary classification task, common classification evaluation metrics can be used. But considering the particularity of the deepfake face detection task, in general cases, there is an imbalance between the true and false categories of the data samples, some metrics such as accuracy have certain limitations. Therefore, it is advocated to adopt more reliable metrics. Regarding generalization, the paper details protocols for cross-forgery and cross-dataset evaluation, which are essential for measuring robustness against unseen attacks. Generalization is an indispensable indicator during model evaluation, especially in the current era where forgery techniques are diverse and rapidly evolving. Additionally, considering application requirements such as real-time monitoring on edge devices, the review summarizes efficiency metrics, including inference speed, memory footprint, and parameter count, which are important in the model during deployment. Moreover, from an application perspective, deepfake face detection technology can be used to prevent the spread of false information and to prevent criminals from using deepfake face technology to steal people's information, property and other targets. Thus, this review synthesizes four major deployment scenarios: the verification of media content authenticity, protection of digital identity security, judicial evidence collection/forensics, and the automated supervision of social network content In parallel, it reviews the developing legal landscape, specifically domestic regulations governing deep synthesis services, highlighting the intersection of technology and law. Finally, based on a dialectical analysis of the primary contradictions in the field (such as the failure of detection features due to the evolution of generation technology, insufficient data-driven and generalization capabilities, etc.) the paper identifies critical future trends. These include the deeper integration of detection systems with large language models, the pursuit of explainable and highly generalizable detection models to enhance trust, model lightweighting for edge devices deployment, and the continuous refinement of industry regulations to govern the ethical use of synthetic media. In summary, the deepfake face detection technology, as a core technology for safeguarding the truth and trust in the digital age, is currently in a stage of rapid evolution and multi-disciplinary integration. It is worthy of researchers' investment of time and effort for research and advancement. The github of the summary of relevant content: https://github.com/yttttkskr/2025-deepfake-detection.

Key words: deepfake face; deepfake face detection; convolutional neural network; Transformer; self-supervised/unsupervised learning; large language models

Li Weibin, Feng Yuting, Hou Biao, Jiao Lichen
School of Artificial Intelligence, Xidian University, Xi'an 710126, China

0 引 言

自2017年Reddit用户"deepfakes"发布第一个深度伪造算法起,深度伪造技术发展迅猛,网络上出现了越来越多的利用深度伪造技术生成的伪造图像、视频等,给社会带来了不安定因素。英国色情受害者救援组织"色情报复求助热线"数据显示:自2017年以来,使用"深度伪造"技术制作或传播色情影像的行为增幅超过400%(XinHuaNet, 2025);2020年,诈骗者利用深度假视频冒充美国海军上将在Skype上聊天,骗取了近30万美元(Mail, 2020)..... Check Point《AI安全报告2025》(Point, 2025)中指出,在2024年,全球67.4%的钓鱼事件已涉及AI技术,金融行业是受攻击最频繁的领域之一。尽管deepfake技术本身没有善恶属性,但却广泛被用于负面目的,其对社会的危害已显而易见。

目前已出现大量的深度伪造检测算法与平台(Intel, 2024;Ju等, 2024;Microsoft, 2020;DeepLive-Cam),但深度伪造与深度伪造检测技术是对抗性工作,尤其是在AIGC(artificial intelligence generated content)盛行的当下,深度伪造检测技术的发展仍将任重而道远。

本文相对于之前已有的综述文章(Ding等, 2024;Tan等, 2021;Wang等, 2022;Xie等, 2023;Zhou等, 2021;Li等, 2021;Yao等, 2025;Li等, 2023;Lai等, 2023),在模型介绍方面主要介绍针对"深度伪造人脸"的检测,尤其是面部替换,总结了最新的研究进展,并将深度伪造检测技术分为三大类:基于卷积神经网络的检测方法、基于Transformer的检测方法和新型范式进行介绍。此外,还包括深度伪造检测技术的应用场景、生成式人工智能伦理法律方面的内容。

1 深度伪造人脸技术

1.1 深度伪造技术

深度伪造技术指的是利用深度学习方法生成伪造图像、视频、音频等媒体形式的方法。近年来,常用的深度伪造方法是变分自编码器(variational auto encoder, VAE)(Kingma等, 2022)、生成对抗网络(generative adversarial network, GAN)(Ian等, 2020)和扩散模型(diffusion model, DM)(Ho等, 2020)。

VAE是在自编码器(autoencoder)的基础上,结合变分推断(variational inference)和贝叶斯理论提出的一种深度生成模型,其目标是学习一个能够生成与训练数据相似样本的模型。它假设隐变量服从某种先验分布(如标准正态分布),并通过编码器(encoder)将输入数据映射到隐变量的后验分布,再通过解码器(decoder)将隐变量还原成生成样本。虽然目前在视觉质量上可能不如基于GAN的模型,但它们通常具有训练稳定、生成速度快、易于控制和插值等优点。VAE很少被单独使用来生成高保真的人脸,它通常作为整个系统的一个核心组件,与其他技术(如GAN)结合,或者通过特定的结构设计来提升性能。

基于GAN的生成算法是深度伪造技术的最常用的方法,尤其是在追求高保真度、高真实感的人脸合成方面。与VAE相比,GAN生成的图像细节更丰富,纹理更逼真。GAN由生成器(generator)和判别器(discriminator)组成。生成器的任务是生成尽可能接近真实数据的假数据,而判别器的任务是区分输入数据是真实数据还是生成器生成的假数据。二者通过相互竞争与对抗共同进化,最终生成器能够生成非常接近真实数据的样本。基于GAN的深度伪造技术已经发展得非常成熟。目前,以DeepFaceLab(DeepFaceLab)为代表的工业化流程工具和基于StyleGAN(Style-based Generative Adversarial Network)(Karras等, 2018;Karras等, 2020)潜空间编辑、SimSwap(Simple Swap)(Chen等, 2020;Chen等, 2024)等先进算法是生成高质量深度伪造内容的主流力量。这些方法在身份保真度、姿态表情迁移的自然度以及处理复杂场景(如遮挡)的能力上不断突破,使得检测任务变得越来越困难。

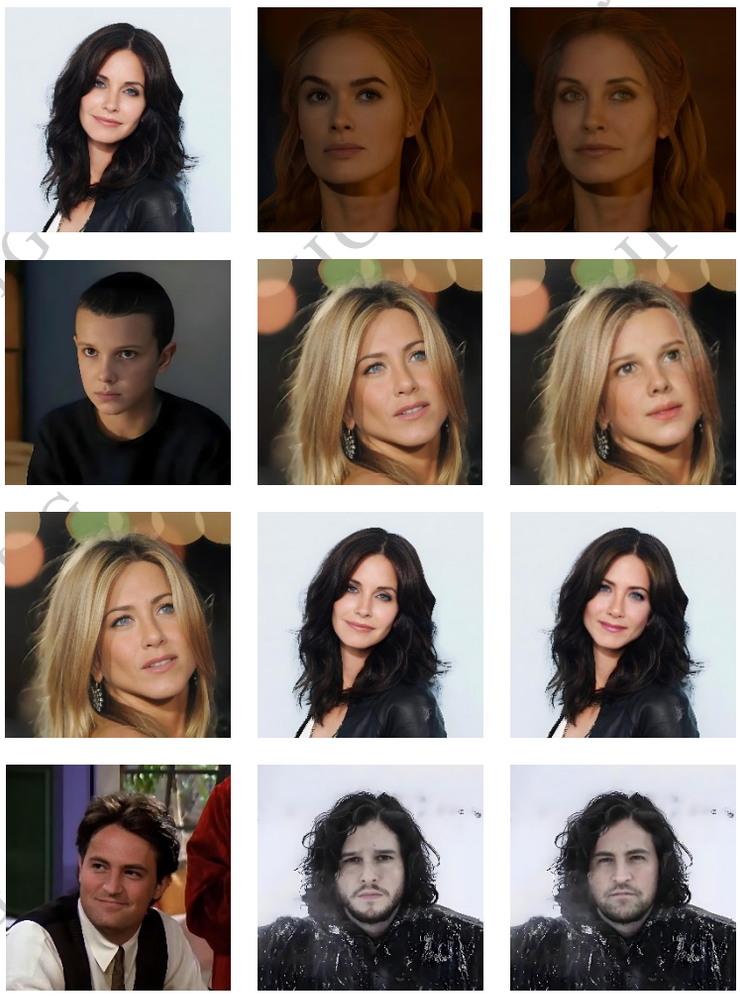
基于扩散模型的深度伪造生成算法凭借其更高的生成质量、更强的稳定性和在图像编辑方面的灵活性,是当前该领域发展最迅速、潜力最大的方向。DM分为两个过程:正向过程(或扩散过程)和反向过程(或逆扩散过程)。扩散过程通过设置不同的均匀超参数分别作为扩散过程中的权值,然后通过重参数化的方法逐步加入噪声,使最终的图像呈标准

高斯白噪声,可以看作是模拟了数据分布逐渐模糊的过程。逆扩散过程从噪声分布出发,通过不断采样逐步移除噪声,恢复到目标数据分布。DreamBooth(Ruiz 等,2022)+ Stable Diffusion 的组合是目前在图像领域获得最逼真效果的主流技术路径之一。而像 FaceDiffusion(Shiohara 等,2024)这样的专用模型,则展示了在特定任务上更优的性能和便利性。不过最大的挑战依然是推理速度,随着蒸馏技术(model distillation)的发展,该问题或将被解决。

1.2 深度伪造人脸技术

深度伪造领域内最成熟、最引人注目的一个应用就是深度伪造人脸技术,即伪造对象是包含人脸的图像或视频的算法。由于人脸是人的本质属性、是一个个体区分于另一个个体的重要"生物社会身份标识符",深度伪造人脸技术的错误应用通常会带来错误信息的传播、他人名誉的损毁等社会问题。

图 1 为深度伪造人脸的示例,使用 Deep Live Cam 算法,将源图像的人脸替换到目标图像,保留目标图像的表情、姿态等信息。



(a) 源图像 (b) 目标图像 (c) 合成图像

((a)original faces;(b) target faces;(c) synthesis faces)

图 1 深度伪造人脸示例

Fig. 1 images deepfake face

根据伪造属性,现有的深度伪造人脸技术大体可分为四类(Li 等,2023):

1)面部转换(facial transfer)是人脸伪造技术中最为经典的一类算法,将人脸身份信息和其他信息(如表情和面部动作信息)从源图像转移到目标图像。

2)面部替换(facial swapping)将目标图像的人物身份替换为源图像的人物身份,但保留其他信息。

3)面部重现(facial reenactment)保持人物身份,只对人物的表情特征进行修改,例如面部表情或者嘴型等。

4)面部属性编辑(facial editing)是较为传统的伪造类型,修改人脸的面部属性,例如头发与皮肤的颜色、皱纹等。

2 深度伪造人脸检测技术

深度伪造人脸检测技术(deepfake face detection, DFD),即识别、定位和验证人脸图像或视频的真实性,主要目标是区分真实内容与伪造内容。早期的检测方法主要基于机器学习,旨在检测生成过程中的伪影和缺陷,例如图像中的不自然纹理、边缘伪影等。这些方法通常依赖于手工设计的特征提取器,如局部二值模式(local binary patterns, LBP)、尺度不变特征变换(scale-invariant feature transform, SIFT)等,并使用支持向量机(support vector machine, SVM)、随机森林(random forest)等传统分类器进行分类。图2对比了深度伪造人脸技术和深度伪造人脸检测技术的流程对比。

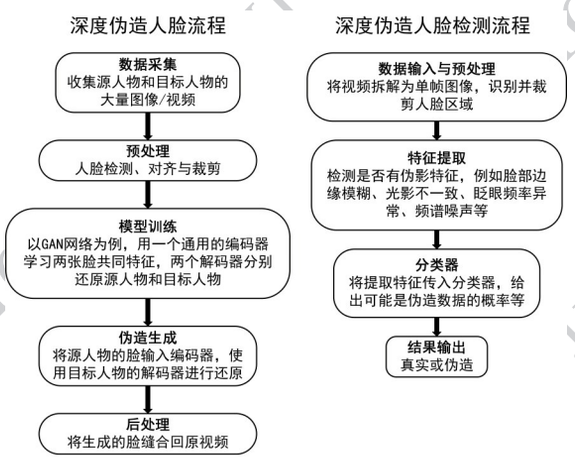


图2 深度伪造人脸技术与深度伪造人脸检测技术大体流程

Fig. 2 flowcharts of deepfake face and deepfake face detection technology

随着深度学习的发展,伪造内容变得更加真实、伪造过程生成的伪影越来越不明显,检测方法也转向基于深度学习的方法。深度学习主流的五大模型架构为:卷积神经网络、Transformer、自编码器(auto encoder, AE)、循环神经网络(recurrent neural network, RNN)、和生成对抗网络(generative adversarial network, GAN)。其中,RNN因擅长处理序列数据而主要应用于自然语言处理(natural language processing, NLP)领域,AE常用作数据处理、特征提取和生成模型,GAN则是生成任务的主力,因而在检测任务中应用较少。因此本文将深度伪造人脸检测技术分为基于卷积神经网络的检测方法、基于Transformer的检测方法、新型范式进行介绍。

2.1 基于卷积神经网络的深度伪造人脸检测方法

由于CNN模型提出较早,且基于CNN的检测算法展现出了较高的正确率,因此成为一种常用且持续发展中的检测手段。基于CNN的模型根据其结构特点大体可分为六类:基础CNN架构、胶囊网络、时序建模方法、结构增强、频域特征增强方法和多任务/多分支方法。

2.1.1 基础CNN架构

深度伪造人脸检测本质上是一个分类问题,因此前期研究人员使用较为经典的CNN结构来进行检测。XceptionNet (extreme inception network)(Chollet,2017)基于Xception架构,利用能够减少参数量、提高计算效率的深度可分离卷积,在Deepfake Detection Challenge(DFDC)数据集中表现优异,通常作为深度伪造检测的基准模型。MesoNet (meso network)(Afchar 等,2018)是一个轻量级卷积神经网络(CNN),使用端到端训练方式,从原始图像中直接学习特征。尽管其网络较浅、结构简单,但通过专注于特定特征的学习,能够有效地区分伪造图像和真实图像,避免了过多计算资源的浪费。EfficientNet-b5(Tan 等,2019)的复合缩放策略使其能有效捕捉深度伪造图像中不同尺度的异常特征,例如细粒度纹理(伪造区域)和全局一致性(光照或颜色差异)。Inception Res. V1 (inception-ResNet version 1)(Szegedy 等,2017)结合了Inception模块和ResNet (residual network)的残差连接,其多分支结构能够并行提取不同感受野的特征,非常适合捕捉多尺度信息。基础CNN架构渐渐难以对抗日益发展的伪造方法、无法满足应用场景,但不可否认的是,这些模型为该领域后续发展奠定了基础。

2.1.2 胶囊网络

胶囊网络(capsule network)是一种新兴的神经网络架构,其设计初衷是为了解决传统CNN的局部

敏感性和空间层次解析能力的不足。胶囊网络引入“胶囊(capsule)”的概念,每个胶囊都是一个小型的神经网络,能够识别特定类型的特征,并对其存在的概率和姿态参数进行编码。一个胶囊可能含有多个神经元,神经元之间传递的都是向量,并通过“动态路由(dynamic routing)”在不同胶囊之间传递信息。图3给出了传统神经元和胶囊神经元的对比。

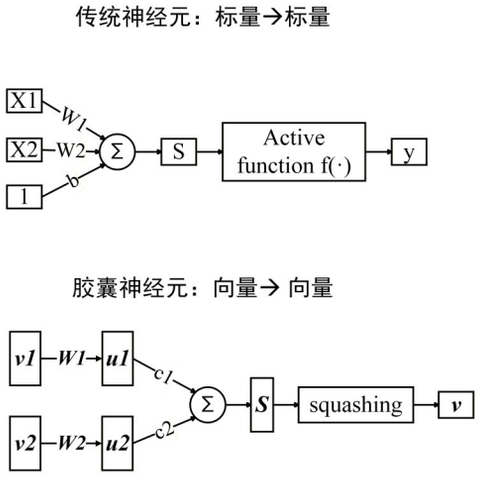


图3 传统神经元与胶囊神经元对比

Fig. 3 Traditional Neuron vs. Capsule Neuron

Capsule-Forensics(Nguyen 等,2019)是最早将胶囊网络引入深度伪造检测的工作之一。该模型首先利用预训练的 CNN 提取图像的底层特征,再通过统计池化(statistical pooling)获取特征的均值与方差,从而减少参数量并增强模型对伪造图像统计异常的敏感性。随后,这些特征被输入至主胶囊层(primary capsules),并通过动态路由机制聚合为更高层次的“real/fake”表示。实验结果表明,该方法在多种伪造场景下均能取得优异表现,并且相较于同规模 CNN 模型,其参数量减少了约 5 - 11 倍,展现出更强的泛化能力和轻量化特性。golden ratio capsule networks(Dincer 等,2024)将黄金比例引入胶囊网络的设计中,通过调整胶囊的形状和连接方式,进一步提升了模型对伪造图像的敏感性和鲁棒性。

2.1.3 时序建模方法

对于视频数据而言,除了在空间上进行检测,利用时间维度的特征能够更好地帮助模型进行检测。时序建模可以通过引入循环神经网络(recurrent neural network, RNN)或长短时记忆网络(long short-term memory networks, LSTM)等结构来实现。

例如,论文(Satpute 等,2024)采用 CNN-LSTM 两阶段串联框架,使用 CNN 捕捉单帧上的空间特征,LSTM 建模帧间序列特征,在 FF++ (FaceForensics++)和 Celeb-DF 数据集上取得了良好的分类效果。论文(Taviti 等,2023)利用 ResNet 提取更丰富的空间特征,再通过 LSTM 进行时序建模。模型在 FF++、Celeb-DF 和 DFDC (Deepfake Detection Challenge)数据集上,使用不同序列长度(20,40,60,80,100)进行单个数据集和混合数据集测试,关注序列长度对模型性能的影响。论文(Petmezas 等,2025)在此基础上引入 Transformer,构建 CNN-LSTM-Transformer 结构,增强帧间依赖建模能力。此外,利用人物的身份特征,将目标从真假分类转为身份验证,更接近实际应用场景。训练时只使用真实视频,让模型学习每个身份的“正常”面部特征。测试时将待测视频与同一人的参考视频进行比较,差异过大则被判定为伪造。总体来看,通过结合空间与时间特征,有效利用视频序列信息,在深度伪造人脸检测中能达到较好的检测效果,并为多模态或长序列建模提供了可行方案。

2.1.4 结构增强

论文(Zhao 等,2021)将深度伪造人脸检测视为细粒度的分类问题,在主干网络之外使用多注意力机制聚焦不同人脸区域,增强对局部伪造痕迹的感知能力。并通过双线性注意力池化(bilinear attention pooling)聚合局部纹理特征与高层语义特征。适用于高质量、局部伪造痕迹明显的伪造检测。论文(Shirley 等,2024)进一步将视觉与音频信息进行融合,分别提取视频的视觉特征和音频特征,然后进行特征融合,在融合后的特征上引入注意力机制,使模型能动态关注最相关的区域和特征。Guan 等人(Guan 等,2024)发现浅层特征的均值与方差(纹理统计)影响模型性能,因此通过梯度正则化降低检测模型对伪造纹理模式的敏感性,从而提高模型在未知伪造类型上的泛化能力。该方法具有通用性,适用于各种主干网络,也可与数据增强等其他方法结合,进一步提升性能。Yan 等人(Yan 等,2025)发现了一种在现有伪造视频中普遍存在但未被充分探索的时序伪造痕迹,称为 FFD(facial feature drift),并据此设计了一个轻量级、即插即用的时空适配器(spa-

tiotemporal adapter, StA),可插入预训练的图像模型中,增强其时空建模能力。

为避免模型关注“不必要”信息(非伪影信息),ID-unaware(Dong 等,2023)检测模型在模型中设计了伪影检测模块,以关注图像上的局部伪影区域,少关注全局身份特征。通过阻止模型学习图像的全局ID表示,可以减少隐式身份泄漏的影响。此外,为了便于伪影检测模块的训练,还提出了多尺度面部交换方法,利用伪影区域位置的地面真实值(groundtruth)生成伪图像,丰富了训练阶段的伪影特征。此外,Kim等人提出的FRIDAY(facial recognition identity attenuation)模型(Kim 等,2024)也针对于该现象,他们将人脸识别器作为一种辅助手段,减少模型对面部身份的识别从而更注重于deepfake伪影,从而提高模型泛化性能。

2.1.5 频域特征增强方法

频域中的高频成分往往包含了图像的细节信息,而这些信息在空域中可能被模糊或掩盖。因此,利用频域信息作为特征补充能够捕捉伪造图像中的细微差别。通过对图像进行傅里叶变换,可以提取出频域特征,并与空域特征进行融合,从而提高检测的准确性。

AutoGAN(Zhang 等,2019)分析了GAN网络在生成伪造图像时,上采样操作(如转置卷积)会在频域中引入频谱复制现象(spectra replications),并基于此构建一个GAN模拟器,使用真实图像生成“模拟伪影图像”。使用真实图像和模拟伪影图像训练检测器,对每个RGB通道进行2D离散傅里叶变换,使用频域特征而非像素值,直接学习频域中的伪影模式。F3-Net(frequency in face forgery network)(Qian 等,2020)包含两个互补的频域感知分支,频率感知分解(frequency-aware decomposition, FAD)和局部频率统计(local frequency statistics, LFS)。FAD从全局上关注出现异常的频率成分,LFS通过滑动窗口进行局部DCT(sliding window discrete cosine transform, SWDCT),关注局部区域的频率分布是否异常。并通过交叉注意力对两分支协同融合,实现了对伪造图像的精细识别。SPSL(spatial-phase shallow learning)(Liu 等,2021)进一步指出,伪造过程的上采样步骤会在相位谱引入更多高频成分,而这些成分在振幅谱中可能趋近于0,因此提出了一种空间-相位浅层学习框架,将RGB图像每个通道的相位谱转回空间域并与原始图像进行拼接后,输入进裁剪后的Xception网络,因此模型体量较小。这里使用裁剪后的浅层网络,是因为浅层网络感受野小,更关注于局部纹理。模型在FF++上训练,在DFDC数据集上达到了66.16%的AUC,Celeb-DF上76.88%的AUC(当时的SOTA)。MSIDSnet(multi-scale dual-stream network)(Cheng 等,2024)提出一种基于多尺度交互双流网络的检测方法。在空间域上,利用空洞卷积和通道注意力机制提取粗粒度空间特征;在频域上使用改进的BayarConv(bayar and stamm constraint convolution)提取细粒度高频噪声特征;使用交互双流模块(interactive dual-stream module, IDS)实现空间域和频域特征的融合。FreqDebias(frequency debiasing framework)(Kashiani 等,2025)致力于解决现有检测器容易过拟合到训练数据的特定频带,即频谱偏差(spectral bias)问题,提出一种频域去偏(frequency debiasing)框架。通过伪造混合增强(forgery mixup (Fo-Mixup) augmentation)动态地使训练样本的频率特性多样化;利用双重一致性正则化(dual consistency regularization)让模型学习对频域变化不敏感但保持判别性的特征表示。综合来看,利用频域特征增强方法能够帮助模型捕捉到高频、细节的信息,不仅能够提高检测的准确性,也有助于改善模型的泛化能力。

2.1.6 多任务/多分支方法

多任务是指在分类的同时,定位伪造区域或检测伪造方法等;多分支方法则是通过全局/细节分支、或不同模态分支等提升模型的检测能力。

Chen等人提出了一种基于双粒度伪造痕迹(Bi-granularity artifacts, BiG-Arts)的模型(Chen 等,2023),利用浅层卷积神经网络捕捉伪造的像素级痕迹,利用更深的CNN层和更大的感受野提取全局伪造特征,最终通过多层融合将细粒度和粗粒度特征结合起来,提高伪造检测的精度和鲁棒性。RECCE(reconstruction-classification consistency estimator)模型(Cao 等,2022)采用端到端的重建-分类学习结构,聚焦于通过重建网络和分类网络的组合进行伪造检测。重建网络的目标是恢复图像的原始面部特征,通过学习真实图像的特征来生成伪造图像的“重建”版本,重建的误差被用来揭示图像中的伪造痕迹。重建后的图像将被输入到一个分类网络中,将图像分类为真实或伪造。分类网络依赖于卷积神经

网络CNN提取图像中的深层次特征，并根据这些特征判断图像的真实性。模型UCF（uncovering common features）（Yan 等，2023）将图像特征分解为三部分：无关特征、特定伪造特征（与具体伪造方法有关）和通用伪造特征（跨不同伪造方法的共性特征）。使用两个分类头：一个通过学习特定特征预测伪造方法，另一个学习通用特征预测真假。既避免在某种伪造方法上出现过拟合，又能提升跨数据集的泛化能力。

表1列出了六种分类的优缺点对比。

表1　CNN六种分类优缺点对比

Table 1　Comparation Across Six CNN Classification Models

类别	优点	缺点	适用场景
基础CNN架构	架构成熟、训练稳定、计算高效；有丰富公开模型	局部纹理建模为主，泛化能力较弱；对抗攻击敏感	静态图片检测、基础对比实验
胶囊网络	建模空间层次结构，保留部分-整体关系	参数量大、训练困难；工程化较难	保留结构信息、对抗攻击研究
时序建模	捕捉视频帧间伪差一致性，适用于动态检测	训练成本高，依赖较长序列；抗抖动能力有限	视频伪造检测、帧序建模
结构增强	能动态关注重要区域；提升表征能力和解释性	可能增加过拟合风险；模型复杂度提高	多模态融合、注意力机制研究
频域特征增强	能捕捉伪造图像的频率异常，抗压缩强	对真实图像的频域变化较敏感；需要频域融合技巧	图像压缩环境下的伪造检测
多任务/多分支	能同时完成分类、定位等多任务，提升检测精度	架构复杂，训练策略需仔细设计	高精度应用场景、伪造区域定位

2.2　基于Transformer的深度伪造人脸检测方法

Transformer架构起初针对自然语言处理领域，但Vision Transformer（ViT）的提出，使Transformer架构与自注意力机制应用到计算机视觉领域，深度伪造人脸检测领域也开始引入Transformer架构进行检测。Transformer相较于CNN，拥有长距离建模能力和全局感受野，理论上能够更好地捕捉伪造过程中产生的与现实不符的特征，并且具有更高的泛化性能。现有研究（Ghita 等，2024）已经证明：与其他deepfake机器学习和深度学习检测方法相比，ViT模型的性能与已有的研究方法相近，值得进一步研究以评估其全部潜力。与基于CNN的模型相比，基于Transformer的检测方法具有以下优势：

1）更大的捕捉范围：Transformer模型使用的自注意力机制，能更好地进行全局上下文建模，能够有更高的伪影捕捉性能。

2）更好的泛化能力：Transformer模型已经在大规模的文本数据上进行了预训练，然后可以迁移到视觉任务上。这种迁移学习使得Transformer能够受益于大规模的多模态数据，从而提高泛化能力。

3）更强的鲁棒性：而Transformer模型训练需要更多的数据和算力，在训练时能包含更多的数据多样性和复杂性，从而常通常具有更强的鲁棒性。

然而，训练一个性能良好、泛化能力强的Transformer模型相对于CNN需要更强的算力要求（Transformer的自注意力机制计算复杂度为$O(n^2)$），所以通常情况下研究者会将Transformer作为检测模型中的一个模块，与其他架构（通常是CNN）进行杂交。因此该部分将分为纯Transformer模型和CNN-Transformer杂交模型两部分介绍。

2.2.1　纯Transformer模型

Bogdan Ghita等人在论文（Ghita 等，2024）中使用最简单的ViT结合MLP（multilayer perceptron）实现了检测任务，分析了参数如批大小（batchsize）、学习率（learning rate）等对分类过程和结果的影响，并说明基于CNN的检测方法在一定程度上更适用于检测相对应伪造模型生成的伪造图像，证明了ViT模型值得在深度伪造人脸检测方面作进一步研究。Dong等人提出的ICT（identity consistency Transformer）模型（Dong 等，2022）利用ViT提取人脸空间上下文信息，通过自注意力机制捕获脸部区域和脸部轮廓不一致信息，并引入了身份一致性监督（iden-

tity consistency supervision)，利用真实图像和伪造图像之间的身份差异作为监督信号，更敏感地捕捉伪造图像中身份特征的不一致性。

但由于Transformer计算成本高、局部特征提取能力不足、对数据依赖强、时间维度建模挑战等问题，研究人员通常会加入其他模块作为辅助，例如Xiong等人提出的AGIL-SwinT (attention-guided inconsistency learning SwinTransformer)模型(Xiong等，2024)利用Swin Transformer作为主干网络，引入注意力引导的不一致性学习模块(attention-guided inconsistency learning, AGIL)指导Swin Transformer专注于伪造区域，同时进行图像的分类和伪造区域的定位。Miao等人提出的多任务音视频提示学习方法(Miao等，2025)利用视觉提示和音频提示引导预训练的基础模型CLIP (contrastive language-image pre-training)和Whisper (web-scale supervised pre-training for speech recognition)关注伪造特征，同时训练视觉分类、音频分类和视频分类三个任务。参数量虽较大，但大部分参数冻结，从而实现了高效的多模态深度伪造检测任务。

2.2.2 CNN-Transformer杂交模型

CNN与Transformer两种架构均有其独特的特点与优势，因此一些研究者将两种结合，利用CNN的局部感受野和ViT的全局上下文依赖关系建模以达到更高的效果。常用的结合架构是先利用CNN网络提取图像局部特征，再将特征图传入Transformer Encoder捕获全局特征，再接MLP Head得到最终结果，如图4。

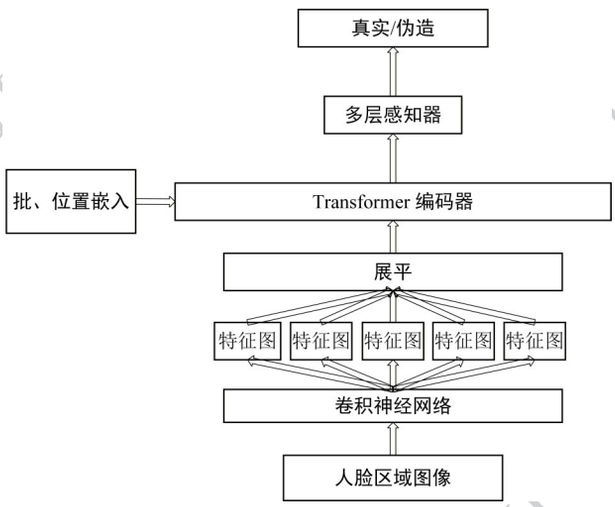


图4 CNN-ViT基础杂交架构

Fig. 4 basic hybrid structure of CNN-ViT

CViT(convolutional vison transformer)(Wodajo等，2021)和HCiT(Kaddar等，2024)等均采用这种结构。CViT的CNN网络使用类似VGG的架构，而HCiT则采用XceptionNet。此外，EfficientViT(Coccomini等，2022)的架构也与此相近，使用EfficientNet B0作为CNN网络，而在将特征图传入transformer编码器之前，未将批和位置嵌入替换为CLS token。其原因是传统的批和位置嵌入会导致输入长度随着分块数量增加，增加Transformer的计算开销。使用CLS token的方法只需在特征图上直接添加一个全局表征的特殊token，可以避免大规模分块操作，同时减少数据流动的复杂性，提升计算效率。Coccomini等人提出了MINTIME(Multi-Identity-size-iNvariant-TIMEsformer)模型(Coccomini等，2024)，采用预训练的Inception ResNet v1捕捉帧级空间信息，结合Transformer捕捉帧间的动态特征。而在将图像传入CNN结构之前，作者对图像中出现的每个人物都做了标记，进行身份聚类，然后再创建输入序列，确保模型能够独立处理每个身份的特征。在Transformer模块，使用TimeSformer with identity-based attention，从而使模型适用于多人视频检测。

还有一些其他杂交结构，例如ViXNet(Vision Transformer with Xception network)(Ganguly等，2022)，与上述结构的不同点在于，它采用双分支结构，使用带掩码的特征和Transformer编码器来提取局部特征，先将图片分成一个个批，在掩码步骤对每个批进行卷积操作获得权重，相乘得到带掩码的特征，再传入Transformer编码器来捕获各个批之间的关联。

但这样的杂交结构随着Transformer结构越来越深，会出现注意力崩塌(attention collapse)的问题，使模型的表示能力下降，因此Zhang等人提出了DTN(distilled Transformer network)模型(Zhang等，2024)，使用多注意力缩放(multi-attention scaling)来避免注意力崩塌，并指出该结构可以应用到任何Transformer的架构中去。此外，相比于简单的二分类的硬标签，该方法还通过自蒸馏结构(self distillation)充分利用了软标签信息，提升了分类精度，并结合混合专家系统(mixture of experts, MoE)来捕捉多样的伪造特征。

2.3 新型范式

在改进基于CNN、Transformer模型架构之外，研

究者也不断探索新的检测方法,主要包含以下两种:无监督/自监督学习方法和大模型检测方法。

2.3.1 自监督/无监督学习方法

模型在深度伪造样本数据中学习时,会不可避免地学习到一些伪影以外的特征。Dong等人指出(Dong等,2023)指出,深度伪造检测模型在训练过程中会在无意中学习了面部身份,更倾向于将伪造样本过多的人脸判定为伪造。而对于特定伪造方法生成的数据,论文(Marra等, 2018; Yu等,2019)表示,有监督的检测方法倾向于在特定伪造方法的特异性指纹上过度汇聚,从而影响模型的泛化性能。因此,无监督/自监督学习方法能够有效避免特定数据、特定伪造方法所造成的偏差。

audio-visual-forensics(Feng等,2023)提出利用音视频一致性进行自监督训练的方法。该方法通过音频-视频同步模型(audio-visual synchronization model)提取同步特征,并训练一个自回归Transformer模型,在真实视频上学习同步特征序列的条件概率分布。在检测阶段,模型在时间维度上对齐预测的同步特征与实际的同步特征,低概率则被识别为伪造视频。

AVH-Align(audio-video alignment)(Smeu等,2025)中分析了FakeAVCeleb和AVDeepfake1M数据集中伪造数据的"缺陷",从而指出:现有数据集可能存在"捷径(shortcuts)"特征,如非伪造内容的统计规律,使监督学习模型容易过拟合而缺乏泛化性。并针对该问题提出了一种无监督的跨模态特征对齐检测方法。使用预训练的AV-HuBERT(audio-visual hidden unit BERT)模型提取音视频特征,并训练一个对齐网络,在真实数据上学习音视频特征的跨模态对齐程度。该方法有效降低了模型对数据集特定偏差的依赖,提高了在不同数据集或真实场景下的检测稳健性。

此外,针对于信息传播领域,基于"传播虚假信息者一定会附加上一些真实信息"的发现(例如:若伪造视频者上传某人的伪造视频,则大概率会在文案里加上类似"这是某某某"的文字说明),论文(Reiss等,2023)提出了FACTOR(fact-checking for deepfake detection)模型,将需验证图像和真实图像提取特征图后,计算两者特征图相似度来判断是否为伪造图像。但该模型的局限性在于需要伪造对象身份已知且有真实图像。

2.3.2 大模型检测方法

随着大模型技术的发展,深度伪造人脸检测领域也引入大模型方法来增强模型性能,提高模型可解释性。RepDFD(reprogrammed deepFake detector)模型(Lin等,2024)通过在输入数据上引入可学习的视觉提示(visual prompts)和文本提示(text prompts),重新编程预训练的CLIP模型,而无需调整其内部参数,从而极大减少了需训练的参数。Wang等人(Wang等,2025)引入知识引导的提示学习和测试时提示调优,提出一种知识引导的即时学习方法,从大型语言模型中检索与伪造相关的提示作为专家知识,以指导可学习提示的优化,并设计了测试时提示调整,关注训练类别(例如自然和室内物体)和测试类别(例如细粒度的人脸图像)之间的领域转移,从而实现性能的改进。HEIE(a novel MLLM-based hierarchical explainable image implausibility evaluator)(Yang等,2025)是一种基于MLLM(multimodal large language models)的分层可解释的图像不合理评估器,同时生成全局热力图和局部热力图,分层捕捉大范围异常和细节缺陷,并使用链式思维(chain-of-thought, CoT)引导LLM:图像描述→问题识别→特殊map标记注入→问题分析→特殊分数标记注入,实现热力图、分数、文本解释三者之间的协同增强,提高可解释性。X^2-DFD(a framework for eXplainable and eXtendable Deepfake Detection)(Chen等,2025)则首先评估MLLM对不同伪造特征的敏感程度,筛选出强特征(模型检测时更敏感的特征)和弱特征(相对没那么敏感的特征)。对于强特征生成专门的提示词,构建针对性的训练样本,强化模型的检测能力和解释能力;对于弱特征则引入专门的特征检测器,作为额外的信息补充。基于大模型的检测方法引入了文本特征,对增强模型可解释性具有极大帮助。

3 常用数据集及评价方法

3.1 常用数据集

深度伪造人脸检测数据集的发展依赖于深度伪造技术的发展,表2介绍了一些深度伪造人脸检测领域的常用的经典视频数据集。近两年的数据集从单纯的"图像分类"任务转向"多模态理解与推理任务"转变,旨在利用大模型强大的理解与推理能力提

升深度伪造检测的性能和模型的可解释性，将模型从“黑盒”变为“白盒”。表3列出了几个深度伪造人脸的新一代多模态数据集。

3.2 评价方法

3.2.1 分类性能指标

深度伪造人脸检测本质是一种分类任务，将图片或视频分类到真/假类别当中，因此分类任务常用的指标都适用于该领域。常用的分类性能指标包括：

1）正确率（Accuracy）：仅适用于真假样本均衡的场景，否则会具有误导性。计算公式如公式（1）。值得注意的是，在深度伪造检测这种数据不均衡（大多数视频是真实的），且漏检和误判的后果不同的任务中，准确率是一个应该被谨慎对待甚至避免单独使用的指标。

2）精确率（Precision）：衡量模型的“判断能力”，在所有模型预测是“伪造”的数据中，多少真正是伪造的。

3）召回率（Recall）：衡量模型的“查全能力”，在所有真正是伪造的数据中，模型查出了多少。4）F1分数：精确率和召回率的调和平均，能准确反应模型在不均衡数据上的综合性能。

5）AUC（Area Under Curve）：即ROC曲线下面积，是当前评估模型能力的最常用指标。ROC曲线（receiver operating characteristic curve）是以假正例率（false positive rate，FPR）为横轴，真正例率（true positive rate，TPR）为纵轴绘制的曲线。AUC值越大，说明模型的分类性能越好，AUC值为1表示完美分类器，AUC值为0.5表示随机猜测。

表4列出了一些模型在FF++上训练，在FF++、DFDC、DFD、Celeb-DF v1 & v2和WildDeepfake上的测试结果（AUC %）。

3.2.2 泛化性能指标

1）跨伪造方法评估：在训练时使用由一种或几种伪造技术产生的数据，在测试时使用由另一种伪造方法产生的数据。也可进行逐步增量评估，例如，先在方法A上训练，在方法B上测试；然后在方法A、B上训练，在方法C上测试，观察随着训练数据中方法种类的增加，模型对未知方法的检测能力是否有提升。跨伪造方法评估更侧重于对未知攻击技术的泛化能力。

2）跨数据集评估：在一个或几个数据集（如：FF++等）上训练模型，然后在另一个不同的数据集上（如Celeb-DF等）上进行测试，得到测试的正确率或者AUC。跨数据集评估更侧重于对未知数据分布的鲁棒性。

3.2.3 应用指标

1）吞吐量：最直观的速度指标，单位时间内（通常为每秒）能处理的视频帧数或视频数量（FPS），直接决定系统能否满足实时检测的要求（例如，直播审核通常需要≥30FPS）。

2）延迟：处理单帧或单个视频从输入到输出所需的时间，对于交互式应用（如视频通话）至关重要，高延迟会破坏用户体验。

3）内存占用：模型加载到内存后，运行时所消耗的动态内存（RAM/VRAM），决定了模型能够在资源受限的边缘设备（如手机、摄像头等）上运行

4）模型参数量：模型所有可训练参数的总数，基本决定了模型的储存开销和部分内存占用。是模型轻量化的和核心衡量标准。

4 应用场景

深度伪造人脸技术给公众带来了一定的娱乐性，但是同时也极易传播虚假信息、实行诈骗手段，给社会带来危害。而深度伪造人脸检测技术的应用将在一定程度上帮助社会公众抵御这些危害，下面是一些深度伪造人脸检测技术的应用场景。

4.1 媒体内容真实性验证

随着社交媒体和短视频平台的广泛应用，各类内容传播的速度和范围大幅提升。然而，这也使得伪造人脸技术成为传播虚假信息的重要手段。而伪造人脸检测技术可以通过对视频和图像内容的真实性验证，帮助识别伪造信息，从而减少虚假内容的传播，并提高公众对媒体信息的信任。例如，微软推出的Microsoft Video Authenticator对由deefake技术产生的媒体内容进行验证，并与BBC等媒体公司进行合作，从而能继续优化模型。因此，社交媒体、信息传播平台将是该技术的一大应用场景。

4.2 数字身份安全

在人脸识别已被广泛应用于手机解锁、支付验证和门禁管理等领域的今天，伪造人脸技术的滥用带来了新的安全威胁。2019年，安全研究人员发现攻击者利用伪造视频成功骗过了一些银行和支付平

表2　深度伪造人脸检测经典数据集

Table 2　Deepfake Face Detection Traditional Datasets

数据集	年份	真实视频数量	伪造视频数量	伪造方法数量
UADFV(Yang 等, 2019) https://docs. google. com/forms/d/e/1FAIpQLScKPoOv15TIZ9Mn0nGScIVg-KRM9tFWOmjh9eHKx57Yp-XcnxA/viewform	*2018*	49	49	1
Deepfake-Timit(Korshunov 等, 2018) https://www.idiap.ch/en/scientific-research/data/deepfaketimit	*2018*	320	620	1
FaceForensics++(Rössler 等,2019) https://github.com/ondyari/FaceForensics/tree/master/dataset	2019	1000	4000	4
DFFD(Stehouwer 等, 2019) https://cvlab.cse.msu.edu/project-ffd.html	2019	1000	3000	多种方法
Celeb-DF v2(Li 等, 2020) https://github.com/yuezunli/celeb-deepfakeforensics	2020	590	5639	1
DFDC(Dolhansky 等, 2020) https://www.kaggle.com/c/deepfake-detection-challenge/data	2020	23,564	104,500	5
DeeperForensic-1.0(Jiang 等, 2020) https://github.com/EndlessSora/DeeperForensics-1.0	2020	50,000	10,000	1, 加 7 种扰动形式
FakeAVCeleb(Khalid 等, 2022) https://github.com/DASH-Lab/FakeAVCeleb	2020	500	19,500	5
WildDeepfake(Zi 等, 2020) https://github.com/OpenTAI/wild-deepfake	2020	0	707	未具体指出
ForgeryNet(He 等, 2021) https://yinanhe.github.io/projects/forgerynet.html	2021	99,630	121,617	15
KoDF(Kwon 等, 2021) https://deepbrainai-research.github.io/kodf/	2021	62,166	175,776	6
LAV-DF(Cai 等, 2022) https://github.com/ControlNet/LAV-DF	2022	540	6,480	9
AV-Deepfake1M(Cai 等, 2024) https://github.com/ControlNet/AV-Deepfake1M	2023	286,721	860,039	1
DGM4(Shao 等, 2023) https://github.com/rshaojimmy/MultiModal-DeepFake	2023	230,000 (图片)	185,267 (图片)	7
FFHQ-UV(Bai 等, 2023) https://github.com/csbhr/FFHQ-UV	2023	0	54,165	3
DeepFakeFace (DFF)(Song 等, 2023) https://github.com/OpenRL-Lab/DeepFakeFace	2023	30,000 (图片)	90,000 (图片)	3
AI-face(Lin 等, 2025) https://github.com/Purdue-M2/AI-Face-FairnessBench	2025	400,000 (图片)	1,200,000 (图片)	37

台的人脸识别系统。这些攻击通过从社交媒体获取目标用户的照片和视频生成逼真的伪造人脸，用于非法转账或其他身份冒用行为。伪造人脸检测技术作为身份认证系统的重要补充，能够识别伪造内容并降低风险，从而保障用户的数字身份安全和隐私。网易易盾与国内某城商行金融理财子公司合作，提出了基于人脸深度评估和注意力机制的活体检测算法，通过 SaaS 服务接口调用的方式快速对接，传入待审核的人脸视频等，实时返回人脸伪造检测结果，为深度伪造人脸检测的落地提供了应用案例。

4.3　司法取证

在法律诉讼和刑事调查中，影像证据的真实性

表3　深度伪造人脸检测新一代多模态数据集

Table 3　Deepfake Face Detection Multimodal Datasets

数据集	年份	数据类型及规模	特点
DD-VQA(Zhang Yue 等, 2024) https://github. com/Reality-Defender/Research-DD-VQA	2024	2,968 图像、14,782 问答对	**开创了 VQA(Visual Question Answering)范式**在深度伪造检测中的应用,将任务从分类提升到理解。
LOKI(Ye 等, 2025) https://github. com/opendatalab/LOKI? tab=readme-ov-file	2024	18K+视频、图像、3D、文本和音频	构建了**大规模基准**,主要应用于合成数据检测领域,用于**评测和开发多模态大模型**在该任务上的能力,推动通用 AI 解决该问题。
ExDDV(Hondru 等, 2025) https://github.com/vladhondru25/ExDDV	2025	5.4K 视频、文本描述和点击标记	提出了一个**系统化的可解释性框架**,通过层次化属性使模型的解释更规范、更细粒度、更易于验证。
DDL (Miao C 等, 2025) https://deepfake-workshop-ijcai2025github. io/main/index.html	2025	1.4M+视频、多层次标注	提升检测模型的**可解释性**和**精准定位**能力,已用于 IJCAI 2025 挑战赛

表4　模型在数据集上测试结果评估(AUC %)

Table 4　Evaluation of the model´s test results on the datasets(AUC %)

算法模型	发布年份	FF++	DFDC	DFD	Celeb-DF v1	Celeb-DF v2	Wild Deepfake
FACTOR(Reiss 等, 2023)https://github. com/talreiss/FACTOR	2022	-	**99.70**	96.30	-	97.00	-
RECCE(Cao 等, 2022)https://github. com/VISION-SJTU/RECCE	2022	-	69.06	-	-	68.71	64.31
BiG-Arts(Chen 等, 2023)	2023	99.39	80.48	89.92	77.04	-	-
ID-unaware (Dong 等, 2023) https://github.com/megvii-research/CADDM	2023	**99.78**	73.74	-	-	93.08	-
UCF(Yan 等, 2023)	2023	88.30	80.50	94.50	-	82.40	-
AGIL-SwinT(Xiong 等, 2024)	2024	99.61	77.14	95.91	-	84.07	-
DTN(Zhang 等, 2024)	2024	99.70	80.01	**97.60**	-	75.32	-
FRIDAY(Kim 等, 2024)	2024	99.18	-	83.95	85.27	83.88	-
Guan et al. (Guan 等, 2024)	2024	99.17	68.04	-	-	87.34	72.28
RepDFD(Lin 等, 2024)https://github. com/KQL11/RepDFD	2024	-	77.34	-	-	80.00	**88.05**
FreqDebias(Kashiani 等, 2025) https://github. com/chuangchuangtan/FreqNet-DeepfakeDetection	2025	97.5	74.1	86.6	**87.5**	83.6	-
CLIP+StA(Yan 等, 2025)	2025	-	84.3	96.5	-	**94.7**	84.8
X^2-DFD(7B) (Chen 等, 2025)https://github. com/SCLBD/X2DFD	2025	-	83.7	92.3	-	90.4	-
X^2-DFD(13B) (Chen 等, 2025) https://github. com/SCLBD/X2DFD	2025	-	83.4	92.5	-	91.3	-

注:加粗字体为每列最优值。

至关重要。然而,伪造人脸技术可能被不法分子用于掩盖真实身份或制造伪造证据,从而干扰司法公正。伪造人脸检测技术能够对视频和图像的真实性进行鉴别,为司法机关提供可靠的技术支持。例如,通过检测犯罪嫌疑人伪造的视频供述或虚假现场影像,帮助查明案件真相并维护司法的公信力。

4.4 社交网络监管

随着社交网络和短视频平台的繁荣,伪造人脸技术被大量用于制作恶搞视频或实施隐私侵害。例如,将普通用户的面部图像替换到特定场景中可能导致名誉受损或隐私泄露。2021年,Facebook与密歇根州立大学(MSU)合作,展示了一种检测和归因(attribution) deepfake图像的研究方法,通过逆向工程来追溯伪造图像是使用哪种伪造方法生成的。伪造人脸检测技术能够帮助社交媒体平台实现自动化监管,及时识别和屏蔽不当内容,维护用户体验和平台环境的健康发展。

5 伦理与法律

由于人脸是一个多维度的、复杂的标识符,深度伪造人脸技术的滥用会引发巨大的伦理和法律冲击。因为它篡改或盗用的,不仅仅是几张图片,而是个体在社会中赖以存在的核心身份标识,动摇了信任、真实和责任的社会根基。而AI技术发展虽然迅速,但还未产生自我意识,犯罪的渊薮依旧指向了操刀的人。由于国家单独针对深度伪造人脸的政策较少,因此表5简要介绍了国内颁布的与深度合成有关的法律、政策和标准,以期为相关从业者提供参考。

6 发展趋势与展望

深度伪造人脸检测是持续发展的技术,与深度伪造的斗争将会持续多年。当前,该领域的核心矛盾聚焦于:

1) 生成技术进化导致检测特征失效

这是促进深度伪造人脸检测领域的最根本原因,深度伪造人脸的真实性、分辨率与细节一致性不断提升,导致传统检测模型依赖的伪造痕迹迅速失效。伪造方法的多样性也对检测模型的泛化能力提出了挑战。

2) 数据驱动与泛化能力不足

当前检测模型的性能提升在一定程度上依赖于数据,并且在已知数据集能达到不错的检测效果,但在新伪造方法、新数据集上性能往往断崖式下降。

3) 模型复杂度的提升与实际部署的条件约束

随着模型检测性能不断提升,模型越来越复杂、参数量与计算量都不断提高,但在实际应用中,往往是在边缘设备,并且可能需要能达到实时检测的效果。这对模型的轻量化、系统级优化提出了需求。

4) 黑盒模型与解释性

现有模型大多只给出检测结果,未给出检测依据。但在实际应用中,尤其牵扯到法律领域,需要模型给出具体的检测依据,哪些区域具有决定性影响,模型的可解释性也亟待提高。

表5 深度合成有关法律法规

Table 5 Deepfake related laws and regulations

法律法规/标准	年份	发布机构	特点
《互联网信息服务深度合成管理规定》	2022	国家网信办、工信部、公安部	聚焦“深度合成”技术强制要求对生成的或编辑的信息内容进行显著标识。
《网络安全标准实践指南——生成式人工智能服务内容标识方法》	2023	全国信息安全标准化技术委员会	详细规定了如何对文本、图片、音频、视频等不同形式的生成内容进行元数据嵌入或外在标识。
《生成式人工智能服务管理暂行办法》	2023	国家网信办、国家发改委、教育部、科技部等七部门联合发布	目前针对生成式AI服务最全面、最核心的监管规定。
《人工智能生成合成内容标识办法》	2025	国家互联网信息办公室、工业和信息化部、公安部、国家广播电视总局	是一部强制要求对AI生成内容进行“显性+隐性”全程标识,并明确各方责任的源头治理法规。

因此,深度伪造人脸检测领域未来将在以下方面着重发展:

1)与大模型的融合

在研究范式层面,一场从“感知”到“认知”的范式转移正在发生。其核心驱动力是检测模型与数据集对大语言模型等基础模型的全面拥抱。具体而言:新一代多模态数据通过引入问答对,为模型提供了需要理解与推理的训练和评测基准;而在模型架构上,研究前沿也从设计专用网络,转向探索如何与大模型对齐以利用其通用知识。这标志着深度伪造检测正作为一个专业模块、被内化至更宏观的、具备批判性思维能力的多模态AI系统中。融合大模型的推理能力与检测领域的专业知识,不仅是技术发展的必然趋势,更是应对未来AIGC安全挑战、实现高效且可信赖的深度伪造防御的根本路径。

2)可解释性与泛化性

当前深度伪造人脸检测的研究已超越单纯追求高精度的阶段,转而致力于破解模型的“黑箱”决策。通过集成可视化技术与注意力机制等方法,旨在使模型的判断依据(如所关注的伪造伪影)对研究者而言透明可信,这是构建可靠安全系统的基石。

此外,随着深度伪造技术不断进化,检测算法需要具备更强的泛化能力,以应对未知或未见伪造手段的挑战。当前一些深度伪造检测模型在训练时依赖于特定的数据分布,容易过拟合于已有数据集,对新的伪造技术(如基于扩散模型的deepfake、跨模态伪造等)检测能力不足、泛化能力较弱。未来的研究将开发通用性更强的特征提取方法和跨领域迁移能力,确保算法在不同伪造类型、分布和质量条件下均能保持高效的检测性能。

3)轻量化和实时检测

为了适应移动设备和边缘计算场景的需求,未来的伪造人脸检测模型应当更加轻量化。在保持高精度的同时,优化计算效率和资源消耗,使检测技术能够在终端设备上实时运行,推动其在社交媒体、内容审核等高频场景中的广泛应用。例如使用知识蒸馏结构、利用大型教师模型指导轻量级学生模型进行学习,使其在资源受限设备上保持良好检测性能,或使用量化、剪枝工具等。Amin等人(Amin 等,2025)提出了一种基于迭代幅度剪枝的模型压缩方法。该方法通过多轮剪枝与再训练,逐步移除网络中幅度较小的权重,从而识别出对检测性能至关重要的稀疏子网络。实验结果表明,即使在高达80%的剪枝率下,部分模型仍能保持接近原始模型的检测性能。这一方法为实现轻量化和实时检测提供了有效途径。MobileNetV3(Howard 等,2019)已在图像分类中实现高效轻量化,但原生架构在细粒度、时序建模和全局特征捕捉方面存在不足,或许可结合其他模块来完成检测任务。

4)法律与伦理的进一步发展

包含深度伪造在内的生成式人工智能的出现给我们带来了许多的便利,但随着深度伪造的社会影响日益扩大,行业和政府应进一步细化相关法规,对伪造内容的制作、传播和使用加以明确约束。国内现在对生成式人工智能愈发重视,下一步应当在一些敏感领域如金融业、娱乐业、新闻传播业等建立更细化的行业标准,使得技术为我们所用,而非自掘坟墓。这将推动深度伪造检测技术的发展成为规范化、标准化的工具,广泛应用于司法鉴定、内容审查和隐私保护等场景。

7 总 结

本文对深度伪造人脸检测技术进行了全面回顾,系统梳理了当前检测方法的研究现状,并按照模型基础架构将其归为三大类:基于卷积神经网络的检测方法、基于Transformer的检测方法和新型范式,分析了各个结构的特点、核心原理和发展现状。总结了深度伪造人脸检测领域的常用数据集,并按照数据的特征类型分为经典数据集和新一代多模态数据集;按照分类性能、泛化性能和应用层面三部分总结了模型的评估指标。此外,列举了深度伪造人脸检测技术的应用场景,揭示了其在社会安全、隐私保护等方面的重要价值。并结合当前面对的核心矛盾和行业发展提出了四个重要发展方向:一是与大模型融合,结合文本信息从而提升模型性能与可解释性;二是提升检测算法的可解释性和泛化性能,使得模型的决策过程更加透明,能够应对多样化伪造手段;三是实现模型轻量化设计,满足移动设备和实时检测的需求;四是推动行业内法规和标准的进一步细化,以规范技术使用。

深度伪造检测技术的持续发展,不仅依赖于技术的迭代,还需法律法规的保驾护航。未来,我们应进一步探索创新性的技术路径,推动深度伪造人脸

检测技术朝着高效化、规范化和实用化方向迈进。

参考文献(References)

Afchar D, Nozick V, Yamagishi J and Echizen I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network//Proceedings of the 10th IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]

Amin L A, Hossain Md I, Nguyen T T, Jahan T, Islam M and Quader F. 2025. Uncovering Critical Features for Deepfake Detection through the Lottery Ticket Hypothesis[EB/OL]. [2025-07-21]. https://arxiv.org/abs/2507.15636.

Bai H, Kang D, Zhang H, Pan J and Bao L. 2023. FFHQ-UV: Normalized Facial UV-Texture Dataset for 3D Face Reconstruction//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [DOI: 10.1109/CVPR52729.2023.00043]

Cai Z, Ghosh S, Adatia A P, Hayat M, Dhall A, Gedeon T and Stefanov K. 2024. AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset//Proceedings of the 32nd ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery: 7414-7423 [DOI: 10.1145/3664647.3680795]

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 the 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA). Sydney, Australia: 1-10 [DOI:10.1109/DICTA56598.2022.10034605]

Cao J, Ma C, Yao T, Chen S, Ding S and Yang X. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 4103-4112 [DOI: 10.1109/CVPR52688.2022.00408]

CCTV News. 2024. The US election sees the emergence of AI-generated fake content, raising concerns over slow legislation. (央视新闻, 2024. 美国大选现人工智能造假立法慢引担忧) https://news.cctv.com/2024/01/31/ARTIF8yjxViHGFxsuWPleO20240131.shtml.

CheckPoint. 2025. Ai Security Report. https://engage.checkpoint.com/resources-chinese-simpl/items/report-ai-security-report-2025-ch

Chen H, Li Y, Lin D, Li B and Wu J. 2023. Watching the BiG artifacts: Exposing DeepFake videos via Bi-granularity artifacts. Pattern Recognition, 135: 109-179 [DOI: https://doi.org/10.1016/j.patcog.2022.109179]

Chen R, Chen X, Ni B and Ge Y. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping//Proceedings of the MM '20: The 28th ACM International Conference on Multimedia: 2003–2011 [DOI: 10.1145/3394171.3413630]

Chen X, Ni B, Liu Y, Liu N, Zeng Z and Wang H. 2024. SimSwap++: Towards Faster and High-Quality Identity Swapping. IEEE Trans. Pattern Anal. Mach. Intell., 46(1): 576-592 [DOI: 10.1109/TPAMI.2023.3307156]

Chen Y, Yan Z, Cheng G, Zhao K, Lyu S W B. 2025. X^2-DFD: A framework for eXplainable and eXtendable Deepfake Detection//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. [DOI: 10.48550/arXiv.2410.06126]

Cheng Z, Wang Y, Wan Y and Jiang C. 2024. DeepFake detection method based on multi-scale interactive dual-stream network. Journal of Visual Communication and Image Representation, 104: 104263 [DOI: 10.1016/j.jvcir.2024.104263]

Chollet F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1800-1807 [DOI: 10.1109/CVPR.2017.195]

Coccomini D A, Zilos G K, Amato G, Caldelli R, Falchi F, Papadopoulos S and Gennaro C. 2024. MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection. IEEE Transactions on Information Forensics and Security, 19: 6084-6096 [DOI: 10.1109/TIFS.2024.3409054]

Coccomini D, Messina N, Gennaro C and Falchi F. 2022. Combining EfficientNet and Vision Transformers for Video Deepfake Detection [EB/OL]. [2021-07-06]. http://arxiv.org/abs/2107.02612.

DailyMail. 2020. Scammer used deepfake video to impersonate U.S. Admiral on Skype chat and swindle nearly $300,000 out of a California widow. https://www.dailymail.co.uk/news/article-8875299/Scammer-uses-deepfake-video-swindle-nearly-300-000-California-widow.html

DeepLiveCam. https://github.com/hacksider/Deep-Live-Cam.

DeepFaceLab. https://www.deepfacelab.cn/download.

Dincer S, Ulutas G, Ustubioglu B, Tahaoglu G and Sklavos N. 2024. Golden ratio based deep fake video detection system with fusion of capsule networks. Computers and Electrical Engineering, 117: 109234 [DOI:10.1016/j.compeleceng.2024.109234]

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(02): 0295-0317 (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(02): 0295-0317) [DOI:10.11834/jig.230088]

Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M and Ferrer C C. 2020. The DeepFake Detection Challenge (DFDC) Dataset [EB/OL]. [2020-06-12]. https://arxiv.org/abs/2006.07397.

Dong S, Wand J, Ji R, Liang J, Fan H and Ge Z. 2023. Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 3994-4004

[DOI: 10.1109/cvpr52729.2023.00389]

Dong X Y, Bao J M, Chen D D, Zhang T, Zhang W M, Yu N H, Chen D, Wen F and Guo B N. 2022. Protecting Celebrities from DeepFake with Identity Consistency Transformer//Proceedings of the 2nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE: 9458-9468 [DOI: 10.1109/CVPR52688.2022.00925]

Feng C, Chen Z and Owens A. 2023. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 10491-10503 [DOI: 10.1109/CVPR52729.2023.01011]

Ganguly S, Ganguly A, Mohiuddin S, Malakar S and Sarkar R. 2022. ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection. Expert Systems with Applications, 210: 118423 [DOI:10.1016/j.eswa.2022.118423]

Ghita B, Kuzminykh I, Usama A, Bakhshi T and Marchang J. 2024. Deepfake Image Detection Using Vision Transformer Models//Proceedings of the 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). IEEE: 332-335 [DOI:10.1109/BlackSeaCom61746.2024.10646310]

Guan W N, Wang W, Dong J and Peng B. 2024. Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization. IEEE Transactions on Information Forensics and Security, 19: 5345-5356 [DOI:10.1109/TIFS.2024.3396064]

He Y N, Gan B, Chen S Y, Zhou Y C, Yin G J, Song L C, Sheng L, Shao J and Liu Z W. 2021. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 4358-4367 [DOI:10.1109/CVPR46437.2021.00434]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: 6840-6851 [DOI:10.5555/3495724.3496298]

Hondru V, Hogea E, Onchis D, Ionescu R T. 2025. ExDDV: A New Dataset for Explainable Deepfake Detection in Video [EB/OL]. [2025-03-18]. https://doi.org/10.48550/arXiv.2503.14421

Howard A, Sandler M, Chen B, Wang W, Chen L C, Tan M, Chu G, Vasudevan V, Zhu Y, Pang R, Adam H and Le Q. 2019. Searching for MobileNetV3//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 1314-1324 [DOI: 10.1109/ICCV.2019.00140]

Ian G, Jean P A, Mehdi M, Bing X, David W F, Sherjil O Aaron C and Yoshua B. 2020. Generative adversarial networks. Commun. ACM, 63(11): 139-144 [DOI:10.1145/3422622]

Intel. Trusted Media: Real-time FakeCatcher for Deepfake Detection. https://www. intel. com/content/www/us/en/research/trusted-media-deepfake-detection.html

Jiang L M, 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 the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2886-2895 [DOI: 10.1109/cvpr42600.2020.00296]

Ju Y, Sun C Z, Jia S, Hou S W, Si Z F, Datta S K, Ke L P, Zhou R, Nikolich A and Lyu S W. 2024. DeepFake-O-Meter v2.0: An Open Platform for DeepFake Detection[EB/OL]. [2024-04-19]. https://arxiv.org/abs/2404.13146

Kaddar B, Fezza S A, Akhtar Z, Hamidouche W, Hadid A and Serra-Sagrista J. 2024. Deepfake Detection Using Spatiotemporal Transformer. ACM Transactions on Multimedia Computing, Communications, and Applications, 20(11): 1-21 [DOI:10.1145/3643030]

Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J and Aila T. 2020. Training Generative Adversarial Networks with Limited Data. Advances in Neural Information Processing Systems. 33: 12104--12114 [DOI: 10.48550/arXiv.2006.06676]

Karras T, Laine S and Aila T. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks[EB/OL]. [2018-12-12] http://arxiv.org/abs/1812.04948.

Kashiani H, Talemi N A and Afghah F. 2025. FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 8775-8785 [DOI:10.1109/CVPR52734.2025.00820]

Khalid H, Tariq S, Kim M and Woo S S. 2022. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset [EB/OL]. [2022-03-01] https://arxiv.org/abs/2108.05080

Kim Y, Kwon M J, Lee W and Kim C. 2024. FRIDAY: Mitigating Unintentional Facial Identity in Deepfake Detectors Guided by Facial Recognizers[EB/OL]. [2024-12-19]. http://arxiv.org/abs/2412.14623

Kingma D P and Welling M. 2022. Auto-Encoding Variational Bayes [EB/OL]. [2022-12-10]. https://arxiv.org/abs/1312.6114

Korshunov P and Marcel S. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection[EB/OL]. [2018-12-20]. https://arxiv.org/abs/1812.08685

Kwon P, You J, Nam G, Parl S and Chae G. 2021. KoDF: A Large-Scale Korean DeepFake Detection Dataset//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 10744-10753 [DOI: 10.1109/iccv48922.2021.01057]

Lai Z M, Zhang Y and Li D. 2023. A Survey of Deepfake Detection Techniques Based on Transformer. Journal of Guangdong University of Technology, 40(06): 155-167 (赖志茂,章云,李东. 基于Transformer的人脸深度伪造检测技术综述.广东工业大学学报, 2023, 40(06): 155-167) [DOI: 10.12052/gdutxb.230130]

Li X R, Ji S L, Wu C M, Liu Z G, Deng S G, Cheng P, Yang M and

Kong X W. 2021. Survey on Deepfakes and Detection Techniques. Journal of Software, 32(2): 496-518 (李旭嵘,纪守领,吴春明,刘振广,邓水光,程鹏,杨珉,孔祥维.深度伪造与检测技术综述.软件学报,2021,32(2):496-518) [DOI: 10.13328/j.cnki.jos.006140]

Li Y Z, Sun P, Qi H G and Lyu S W. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Lin K Q, Lin Y Z, Li W X, Yao T P and Li B. 2024. Standing on the Shoulders of Giants: Reprogramming Visual-Language Model for General Deepfake Detection[EB/OL]. [2024-12-29]. http://arxiv.org/abs/2409.02664

Lin L, Santosh S, Wu M Y, Wang X and Hu S. 2025. AI-Face: A Million-Scale Demographically Annotated AI-Generated Face Dataset and Fairness Benchmark//Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). 3503-3515 [DOI: 10.1109/cvpr52734.2025.00332]

Li Z Y, Zhang X H, Pu Y W, Wu Y M and Ji S L. 2023. A Survey on Multimodal Deepfake and Detection Techniques. Journal of Computer Research and Development, 60(6): 1396-1416 (李泽宇,张旭鸿,蒲誉文,伍一鸣,纪守领.多模态深度伪造及检测技术综述.计算机研究与发展,2023,60(6):1396-1416). [DOI: 10.7544/issn1000-1239.202111119]

Liu H G, Li X D, Zhou W B, Chen Y F, He Y, Xue H, Zhang W M and Yu N H. 2021. 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). 772-781 [DOI: 10.1109/CVPR46437.2021.00083]

Marra F, Gragnaniello D, Verdoliva L and Poggi G. 2018. Do GANs Leave Artificial Fingerprints? //Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR): 506-511 [DOI: 10.1109/mipr.2019.00103]

Miao C T, Zhang Y, Gao W Z, Tan Z Y, Feng W W, Luo M, Li J S, Liu A, Diao Y F, Chu Q, Gong T, Li Z, Yao W B and Zhou J T. 2025. DDL: A Large-Scale Datasets for Deepfake Detection and Localization in Diversified Real-World Scenarios[EB/OL]. [2025-10-30]. https://arxiv.org/abs/2506.23292

Miao H, Guo Y F, Liu Z M and Wang Y H. 2025. Multi-modal Deepfake Detection via Multi-task Audio-Visual Prompt Learning//Proceedings of the AAAI Conference on Artificial Intelligence, 39(1): 612-621 [DOI:10.1609/aaai.v39i1.32042]

Microsoft. 2020. New Steps to Combat Disinformation. https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/

Nguyen H H, Yamagishi J and Echizen I. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos//Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 2307-2311 [DOI:10.1109/ICASSP.2019.8682602]

Petmezas G, Vanian V, Konstantoudakis K, Almaloglou Elena E. I. and Zarpalas D. 2025. Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification//Proceedings of the Multimed Tools Applications [DOI: 10.1007/s11042-024-20548-6]

Qian Y Y, Yin G J, Sheng L, Chen Z X and Shao J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues//Proceedings of the Computer Vision – ECCV 2020. Cham: Springer International Publishing: 86-103 [DOI: 10.1007/978-3-030-58610-2_6]

Reiss T, Cavia B and Hoshen Y. 2023. Detecting Deepfakes Without Seeing Any[EB/OL]. [2023-11-02]. http://arxiv.org/abs/2311.01458

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Niessner M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 1-11 [DOI:10.1109/ICCV.2019.00009]

Ruiz N, Li Y Z, Jampani V, Pritch Y, Rubinstein M and Aberman K. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation[EB/OL]. [2023-03-15]. https://doi.org/10.48550/arXiv.2208.12242

Satpute R and Onwe C P. 2024. CNN-LSTM Model for Deepfake Image Detection//Proceedings of the 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI): 1-6 [DOI: 10.1109/IDICAIEI61867.2024.10842840]

Shao R, Wu T and Liu Z. 2023. Detecting and Grounding Multi-Modal Media Manipulation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [DOI: 10.1109/cvpr52729.2023.00667]

Shiohara K and Yamasaki T. 2024. Face2Diffusion for Fast and Editable Face Personalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [DOI: 10.1109/cvpr52733.2024.00654]

Shirley C P, Berin Jeba Jingle I, Abisha M B, Venkatesan R, Yashvanth Ram R V and Elakkiya Elango. 2024. Deepfake Detection Using Multi-Modal Fusion Combined with Attention Mechanism//Proceedings of the 2024 4th International Conference on Sustainable Expert Systems (ICSES): 1194-1199 [DOI: 10.1109/ICSES63445.2024.10763221]

Smeu S, Boldisor D A, Oneata D and Oneata E. 2025. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning localization//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [DOI:

10.1109/cvpr52734.2025.01753]

Song H X, Huang S Y, Dong Y P and Tu W W. 2023. Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models[EB/OL]. [2023-09-05] https://arxiv.org/abs/2309.02218

Stehouwer J, Dang H, Liu F, Liu X M and Jain A K. 2019. On the Detection of Digital Face Manipulation//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 5780-5789. [DOI: 10.1109/cvpr42600.2020.00582]

Szegedy C, Ioffe S, Vanhoucke V and Alemi A A. 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI Press: 4278-4284 [DOI: 10.1609/aaai.v31i1.11231]

Tan M K, Xu S K, Zhang S H and Chen Q. 2021. A review on deep adversarial visual generation. Journal of Image and Graphics, 26(12):2751-2766 (谭明奎,许守恺,张书海,陈奇. 2021. 深度对抗视觉生成综述. 中国图象图形学报, 26(12):2751-2766) [DOI: 10.11834/jig.210252]

Tan M X and Le Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks//Proceedings of the 36th International Conference on Machine Learning: 97(6105-6114) [DOI: 10.1007/978-1-4842-6168-2_10]

Taviti R, Taviti S, Reddy P A, Sankar N R, Veneela T and Gou P B. 2023. Detecting Deepfakes With ResNext and LSTM: An Enhanced Feature Extraction and Classification Framework//Proceedings of the 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT): 1-5 [DOI: 10.1109/IConSCEPT57958.2023.10170580]

Wang H, Deng C and Zhao Z D, 2025. Knowledge-Guided Prompt Learning for Deepfake Facial Image Detection[EB/OL]. [2025-01-01]. http://arxiv.org/abs/2501.00700

Wang R Y, Chu B L, Yang Z and Zhou L N. 2022. An overview of visual DeepFake detection techniques. Journal of Image and Graphics, 27(01):0043-0062 (王任颖, 储贝林, 杨震,周琳娜. 2022. 视觉深度伪造检测技术综述. 中国图象图形学报, 27(01):0043-0062) [DOI:10.11834/jig.210410]

Wodajo D and Atnafu S. 2021. Deepfake Video Detection Using Convolutional Vision Transformer[EB/OL]. [2021-03-11]. https://arxiv.org/abs/2102.11126

Xie T, Yu L Y, Luo C W, Xie H T, and Zhang Y D, 2023. Survey of deep face manipulation and fake detection[J]. Journal of Tsinghua University (Science and Technology), 63(9): 1350-1365. (谢天, 于灵云,罗常伟,谢洪涛,张勇东. 深度人脸伪造与检测技术综述. 清华大学学报(自然科学版), 2023, 63(9):1350-1365) [DOI: 10.16511/j.cnki.qhdxxb.2023.21.002]

Xinhuanet. 2025. The UK plans to criminalize pornographic 'deepfake' content. 2025. (新华国际,2025.英国拟将涉色情"深度伪造"入刑).https://h.xinhuaxmt.com/vh512/share/12354202? d=134fdfb

Xiong W T, Chen H Y, Zhao G Y and Li X B. 2024. AGIL-SwinT: Attention-guided inconsistency learning for face forgery detection. Image and Vision Computing, 151: 105274 [DOI: 10.1016/j.imavis.2024.105274]

Yan Z Y, Zhang Y, Fan Y B and Wu B Y. 2023. UCF: Uncovering Common Features for Generalizable Deepfake Detection//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 22412-22423 [DOI: 10.1109/iccv51070.2023.02048]

Yan Z Y, Zhao Y Z, Chen S, Guo M Y, Fu X H, Yao T P, Ding S H, Wu Y S and Yuan L. 2025. Generalizing Deepfake Video Detection with Plug-and-Play: Video-Level Blending and Spatiotemporal Adapter Tuning//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 12615-12625 [DOI: 10.1109/CVPR52734.2025.01177]

Yang F, Zhen R, Wang J N, Zhang Y H, Chen H X, Lu H N, Zhao S C and Ding G G. 2025. HEIE: MLLM-Based Hierarchical Explainable AIGC Image Implausibility Evaluator//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 3856-3866 [DOI: 10.1109/cvpr52734.2025.00365]

Yang X, Li Y Z and Lyu S W. 2019. Exposing Deep Fakes Using Inconsistent Head Poses//Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 8261-8265 [DOI: 10.1109/ICASSP.2019.8683164]

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].

Ye J Y, Zhou B C, Huang Z L, Zhang J N, Bai T Y, Kang H R, He J, Lin H L, Wang Z H, Wu T, Wu Z Z, Chen Y P, Lin D H, He C H and Li W J. 2025. LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models//Proceedings of the 2025 ICLR. 70440—70522 [DOI: 10.48550/arXiv.2410.09732]

YU N, DAVIS L and FRITZ M. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 7555-7565 [DOI: 10.1109/ICCV.2019.00765]

Zhang X, Karaman S, Chang S F. 2019. Detecting and Simulating Artifacts in GAN Fake Images//Proceedings of the 2019 IEEE International Workshop on Information Forensics and Security (WIFS). 1-6 [DOI: 10.1109/WIFS47025.2019.9035107]

Zhang Y N, LI Q F, Yu Z T and Shen L L. 2024. Distilled Transformers with Locally Enhanced Global Representations for Face Forgery Detection[EB/OL]. [2024-12-28]. http://arxiv.org/abs/2412.20156

Zhang Y, Colman B, Guo X, Shanriyari A and Bharaj G. 2024. Common Sense Reasoning for Deepfake Detection[EB/OL]. [2024-07-18]. https://arxiv.org/abs/2402.00126

Zhao H Q, Wei T Y, Zhou W B, Zhang W M, Chen D D and Yu N. 2021. Multi-attentional Deepfake Detection//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 2185-2194 [DOI: 10.1109/CVPR46437.2021.00222]

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. 2382-2390 [DOI: 10.1145/3394171.3413769]

Zhou W B, Zhang W M, Yu N H, Zhao H Q, Liu H G. 2021. An Overview of Deepfake Forgery and Defense Technique. Journal of Signal Processing, 37(12): 2338-2355 (周文柏, 张卫明, 俞能海, 赵汉卿, 刘泓谷, 韦天一. 人脸视频深度伪造与防御技术综述. 信号处理, 2021,37(12):2338-2355) [DOI: 10.16798/j.issn.1003-0530.2021.12.007]

作者简介

李卫斌,通信作者,男,教授,主要研究方向为AIGC大语言模型、工业智能与工业互联网。 E-mail:weibinli@xidian.edu.cn

冯雨婷,女,硕士,主要研究方向为深度伪造人脸检测。E-mail:25171214003@stu.xidian.edu.cn

侯彪,男,教授,主要研究方向为遥感图像解译与目标识别、无人系统协同感知、人工智能芯片及系统。E-mail: avcodec@163.com

焦李成,男,教授,主要研究方向为图像理解与目标识别、智能感知与计算、深度学习与类脑计算。E-mail:lchjiao@mail.xidian.edu.cn