深度伪造内容检测的挑战及可行技术路径

任 奎 林 峰 巴钟杰 刘振广 程 鹏
浙江大学

背景

生成式人工智能(generative artificial intelligence, GAI)近年来取得了突破性进展。其核心技术经历了从变分自编码器(variational autoencoder, VAE)[1]到生成对抗网络(generative adversarial network, GAN)[2],再到扩散模型(diffusion model, Diffusion)[3]和大模型生成技术(如 Stable Diffusion、DALL-E、Sora)[4]的持续演进。这些技术的不断发展与融合,推动了 GAI 在内容生成领域的广泛应用。随着算法效率的提升和算力成本的降低,GAI 的应用场景正在不断扩展,逐渐渗透到人们生活的方方面面。

深度伪造概述

生成式人工智能的飞速进步催生了一种名为深度伪造(deepfake)的人脸内容生成与操控方法。2017 年,Reddit 社区一位名为“deepfake”的用户首次公开了其利用深度学习算法生成的换脸视频,并开始引起大众的关注。该技术是一种基于深度学习的多媒体内容编辑方法,能够通过深度学习模型轻松生成逼真的视频和图像。随后出现的 ProGAN、StyleGAN 等网络进一步提升生成内容的质量和分辨率。随着 Diffusion 及相关生成式大模型出现,人们可以通过语义控制实现多模态模型的伪造生成。深度伪造技术凭借其高度的可操作性和拟真性,能够实现人脸替换、视频编辑等复杂操作,在影视、虚拟现实、教育等领域展现巨大潜力。

DOI:10.11991/cccf.202506003
基金项目:国家重点研发计划项目(2023YFB2904000);国家自然科学基金区域创新发展联合基金重点项目(U23A20306)
通信作者:任奎,E-mail:kuiren@zju.edu.cn

深度伪造技术的危害

尽管深度伪造技术具有广泛的应用前景,但其滥用也带来诸多社会问题。虚假新闻的传播可能误导公众舆论,身份盗窃和隐私泄露可能造成个人名誉和财产损失。例如,2022 年 3 月一段伪造的乌克兰总统泽连斯基呼吁乌克兰士兵投降的虚假视频在推特上广泛传播;2024 年 6 月韩国发生多起利用深度伪造技术制作色情内容的犯罪案件。此外,深度伪造还可能被用于政治操弄、商业欺诈等恶意行为。这些负面影响不仅威胁到个人权益,还可能破坏社会秩序和国家安全。

深度伪造检测能力概述

面对深度伪造技术带来的威胁,如何进行有效识别和防范成为亟待解决的问题。目前主要通过法律规制和技术手段相结合的方式应对这一挑战。在法律层面,中国于 2025 年 3 月正式发布强制性国家标准《网络安全技术人工智能生成合成内容标识方法》,其核心要求是通过显式和隐式标识对人工智能生成合成内容进行清晰标注。2025 年 4 月,《法治日报》提出需加快《人工智能法》立法,构建“伦理+法律+算法”综合治理体系,重点防范身份伪造、虚假信息扩散等风险。英国于 2025 年 1 月拟立法将制作与传播深度伪造色情内容列为刑事犯罪,即使内容为虚假生成。美国微软等科技公司推动国会立法,要求对人工智能(artificial intelligence, AI)生成内容强制标注“合成”标签,并建立

合成内容数据库以提高透明度。

在技术层面，深度伪造检测领域已形成多元化技术体系。英特尔公司开发的 Fake Catcher 工具通过实时检测面部血液流动的生理特征实现伪造识别；谷歌公司的 SynthID 开辟了主动防御新路径，通过深度学习模型为 AI 生成内容嵌入隐形水印，构建起从生成到识别的安全保障；瑞莱智慧的 DeepReal 系统采用多维度特征融合算法，在学术测试基准中表现卓越；中科睿鉴发布的睿鉴图灵采取了混合专家模型架构，整合多个垂直领域检测小模型，对深度伪造、大模型生成伪造、软件编辑篡改等进行多维度鉴别；美亚柏科推出了慧眼视频图像鉴真工作站，对利用深度伪造技术进行换脸、美颜、生成人脸等篡改的影像具有理想的鉴定效果；浙江大学 DFScan 检测平台通过构建千万级高质量且动态更新的伪造数据底座，集成自研的区域定位验证、数据增量学习、信息论分析、人脸基础模型等技术，在复杂现实场景下实现了精准泛化的检测。

从技术演进维度观察，传统检测模型主要基于各种架构的卷积神经网络，针对特定的伪造模式设计分类模型，从而实现伪造检测。近年来，自注意力机制的广泛应用使一些基于 Transformer[5] 的检测网络能够更好地捕捉图像的局部和全局伪造痕迹，实现架构统一并促进了模型的性能扩展性。未来，随着多模态大模型和通用人工智能（artificial general intelligence, AGI）的发展，检测技术可与大模型结合，在提升检测效果的同时进一步增强支持多模态伪造内容端到端统一识别、检测可解释交互、思维链（chain of thought, CoT）推断伪造意图等新兴能力。

深度伪造检测面临的挑战

在生成式人工智能呈现指数级发展的背景下，深度伪造正经历着从单模态到多模态、从局部伪造到全局操纵的范式转变。现有检测方法在面对日益复杂和多样化的深度伪造技术时，正面临越来越大的系统性挑战，如图 1 所示。

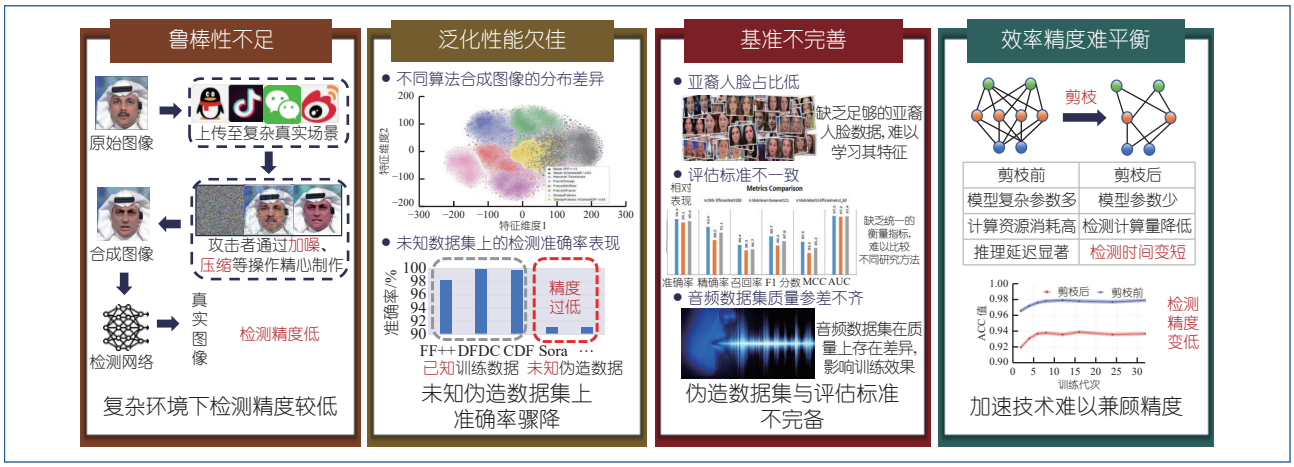


图1　深度伪造检测面临的挑战

首先，现有检测方法在复杂场景下的鲁棒性不足，难以应对现实环境中低分辨率、噪声干扰等的挑战；其次，检测系统的泛化能力有限，难以有效识别未知类型的伪造内容；此外，数据底座和评估基准的构建仍不完善，缺乏持续更新的动态数据集和可靠的评估体系，难以快速迭代检测模型且合理评估实际检测表现；最后，如何在保持高精度的同时提升检测效率，也是当前技术面临的重要问题。这些挑战的存在，凸显了深度伪造检测领域仍需进一步研究和突破的必要性。

鲁棒性问题

深度神经网络一方面在复杂现实场景中，尤其是面临社交平台压缩、媒体噪声干扰，甚至恶意对抗攻击等非理想条件时，其性能往往会出现显著下降；另一方面，考虑到现实应用的需求，伪造检测系统需要确保在各种复杂场景下对伪造内容的高可靠检出率。然而，现有基于深度神经网络的深度伪造检测器的鲁棒性严重不足，即在复杂场景中的表现并不理想[6]。例如，低

分辨率图像或视频中的细节丢失、噪声干扰(如压缩伪影、模糊或光照变化)都会显著降低检测器的准确性。此外,深度伪造生成器还可能利用对抗性机器学习(adversarial machine learning, AML)技术,通过在输入中添加微小扰动或噪声来欺骗检测器。研究表明,基于卷积神经网络(convolutional neural network, CNN)的检测器在面对梯度攻击时,准确率可能降至接近 0%[7],这进一步凸显了鲁棒性问题的严重性。

泛化性问题

深度伪造检测技术在应对新兴伪造手段时面临显著的泛化性问题。大多数现有的检测算法依赖于特定类型的数据集进行训练,导致其在实际应用中难以有效识别未知或新型的伪造内容。这种局限性主要源于深度伪造技术的快速演进,使得检测模型难以适应伪造手段的变化。

当前大多数深度伪造检测算法是基于特定数据集训练的,其检测能力局限于已知的伪造模式。随着深度伪造技术的不断演进,检测模型不得不频繁进行重新训练来应对新的伪造模式,从而大幅削弱了检测系统的泛化能力。一些研究工作尝试通过引入频域信息来提升深度伪造检测模型的泛化能力[8]。具体而言,这些方法通过分析图像或视频的频域特征,试图捕捉伪造内容在频域中留下的痕迹。然而,现有方法通常局限于特定的频带(如高频或低频),未能充分利用频域信息的全频谱特性,这主要是其依赖于预定义的频率滤波器或对伪造线索所在频段的先验假设,导致其检测能力受到限制。这种频谱解析的碎片化特征提取模式导致模型难以捕捉跨频段耦合的篡改痕迹,使其在面对具备频谱自适应特性的新型攻击手段时表现出泛化性衰减。此外,最近的研究热点聚焦于通过模拟伪造或操纵的痕迹,如模仿换脸的边界伪影,进行数据增强。虽然这些尝试在一定程度上提升了泛化能力,但仍受限于特定先验假设,未能聚焦困境本质,即算法架构与动态对抗环境的结构性矛盾,导致检测模型的泛化性能受限。

数据底座与评估基准构建问题

在深度伪造检测领域,高质量的数据底座与标准化评估基准的缺失已成为制约技术发展的关键瓶颈。深度神经网络的性能高度依赖数据的质量与多样性,但现有数据集的局限性使得检测模型的训练、验证与跨场景应用面临严峻挑战。数据底座方面,公开可用的深度伪造数据集数量有限且质量参差不齐。首先,数据收集过程复杂且成本高昂,需协调人脸采集、多模态伪造技术模拟及伦理审查等环节,导致数据规模难以满足需求[9]。其次,现有数据集存在显著异质性:视频分辨率差异大、时长普遍较短,且缺乏场景多样性。这种偏差导致模型在训练中过度拟合特定特征,难以泛化至真实场景中的复杂伪造内容。评估基准构建方面,当前缺乏统一的评估标准和测试框架。现有评估多局限于特定数据集上的性能指标,难以全面反映模型在实际应用场景中的表现。评估指标单一,主要关注检测准确率,而忽视了模型对新型伪造技术的适应能力、跨场景泛化性能等重要维度。此外,评估过程缺乏动态性,无法及时反映快速演进的深度伪造技术带来的新挑战。

效率与性能平衡问题

深度伪造检测技术在实际部署中面临效率与检测精度难以兼顾的核心矛盾。一方面,基于深度学习的检测模型需处理海量视频帧数据,导致计算资源消耗高、推理延迟显著,难以满足实时检测需求[10]。另一方面,深度伪造技术持续迭代,检测模型需不断升级复杂度以捕捉细微伪造痕迹,进一步加剧了效率与精度的冲突。例如,量化技术可能导致模型对高频伪造伪影的响应能力下降,而轻量模型在对抗性样本攻击下的鲁棒性显著弱于全量模型。现有优化方案普遍面临“精度–效率”的权衡困境。

深度伪造检测可行技术路线

针对这些挑战,研究者们相应提出了多种可行的技术路径,如图 2 所示,包括通过数据增强、防御对抗攻击等手段提高鲁棒性,针对图像域、音频域和视频域提出不同策略来增强泛化能力,采用标准化的数据集和评估平台来衡量模型的效果,以及利用蒸馏和量化等加速技术提高检测效率。借助这些创新技术手段,深度伪造检测在精度和适应性方面取得了显著突破,同时也为实时应用和大规模部署提供了坚实的基础。

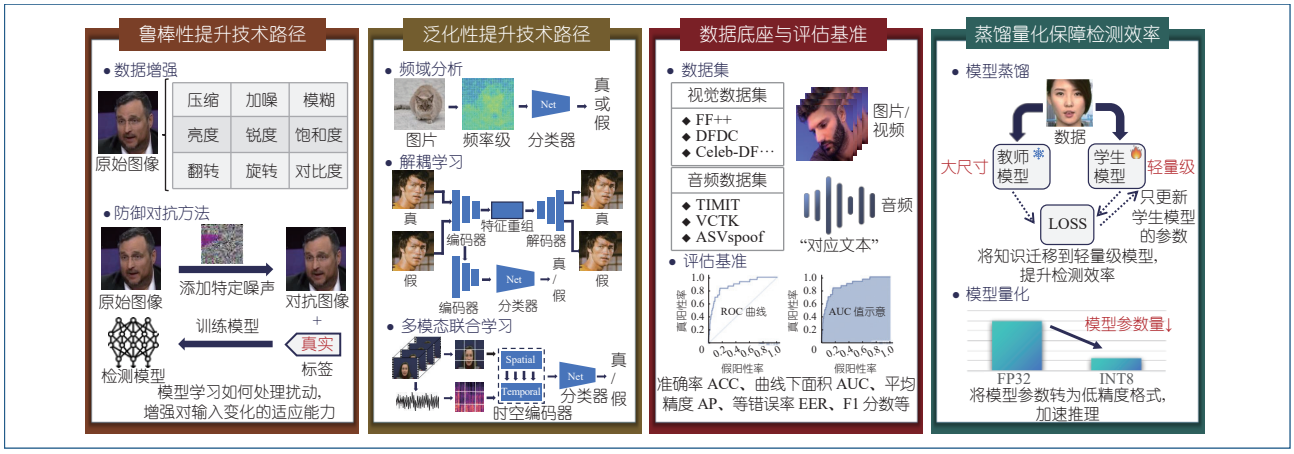


图 2　深度伪造检测可行技术路线

鲁棒性提升

在深度伪造内容检测领域，提高检测模型的鲁棒性是一个至关重要的挑战。该性能指标体现了系统在应对复杂多变的检测场景时，依然保持高准确性与稳定性的能力。为提升深度伪造检测模型的鲁棒性，通常采用数据增强和对抗攻击防御等策略。数据增强方法是通过对原始数据进行人工变换或扩展，生成多样化的训练样本，以提升模型的鲁棒性。在图像域中，传统的数据增强方法包含压缩、加噪、模糊、改变对比度、改变饱和度等操作。Wang 等[11]提出一种基于注意力的检测方法，通过自适应地遮掩最有可能被篡改的面部关键区域，从而实现数据增强。在音频域中，数据扰动方法通常包含压缩和扭曲 2 种。常见的压缩格式为 MP3、AAC、OGG、Opus 等，常见的扭曲方法包含在原始音频训练样本中添加混响、背景音和音乐等。此外，Tak 等[12]提出一种名为 Rawboost 的数据增强方法模拟电话场景中的音频变化，有效提高了深度伪造语音检测模型在面对电话场景中的各种编码和传输变化时的防欺骗(anti-spoofing)性能。在视频域中，数据增强方法不仅涵盖了前述图像域的空间增强与音频域的压缩扭曲，还能结合视频特有的多模态特性，设计更为复杂的复合扰动策略。例如，在时序维度上，可通过帧丢弃与重复(随机删除或复制关键帧)破坏伪造视频的动作连贯性，或通过帧重排(局部打乱相邻帧顺序)干扰模型对时序逻辑的依赖。防御对抗攻击方法主要指使用对抗训练(adversarial training)或特征正则化提升模型对干扰的免疫力。这些跨模态数据增强策略通过多样化训练样本分布并模拟真实场景干扰，结合对抗训练的防御机制，全面提升了检测模型对于压缩、噪声、时序篡改及对抗攻击的容错能力，从而提升深度伪造检测的鲁棒性。

泛化性增强

在深度伪造检测领域，检测模型的泛化能力是提升其对多样化篡改手段适应性的关键因素。该性能指标直接决定了检测算法在面对不断进化的伪造技术时，能否突破训练数据局限，实现对未知伪造特征的跨域识别与判别。在图像域方面，提高深度伪造模型检测泛化性的主要技术方案有频域分析、解耦学习等。为了进一步提高检测模型泛化性，研究人员开发了频域分析算法与空频协同建模框架。Tan 等[13]提出的 FreqNet 通过强调高频信息并引入频率域学习模块来增强检测器的泛化性。解耦学习通过隐空间正交分解机制，解决了深度伪造检测模型中内容特征与伪造特征在隐空间耦合的问题，进一步提高了检测模型的泛化性。UCF[14]基于对比正则化技术，将图像信息分解为 3 个组成部分：与伪造无关的特征、方法特定的伪造特征和共同伪造特征，从而有助于模型在不同伪造类型之间实现更好的泛化性能。在音频域方面，传统的深度伪造检测方法往往将原始音频语句通过各种变换方法转换为时频谱特征，例如转换为谱系系数，如梅尔频率倒谱系数(mel frequency cepstral coefficients, MFCC)[15]和线性频率倒谱系数(linear frequency cepstrum coefficients, LFCC)[16]等，或者转换为基于谱图的表示方式，如短时傅里叶变换谱图(short-time fourier transform, STFT)[17]、恒 Q 变换谱图(constant-Q

transform, CQT)[18] 等。之后再将处理得到的时频谱特征用于分类模型训练，得到深度伪造音频检测模型。为了提高检测模型泛化性，研究者们提出了基于可训练的网络层来提取音频特征，例如使用正交约束型网络层（sine constraints network, SincNet）[19]、可学习的前端网络层（learnable frontend, LEAF）[20] 等提取用于分类训练的特征。在视频域方面，除却单一模态检测方案，研究者还提出音视联合检测策略，利用图像、视频、声音等多种信息源的互补特性，综合不同模态的线索，从而提升伪造检测的准确性和泛化能力。Yang 等[21] 提出了一种基于 Transformer 架构的方法，该方法首先利用时空编码器分别对音频和视觉输入进行特征提取，然后在带有双向交叉注意力模块的解码器中实现 2 种模态特征的交互融合。通过时序维度与空间维度的联合建模，显著提升了音画一致性异常特征的捕捉能力。

数据底座与评估基准构建

在深度伪造检测领域，研究者针对图像域、音频域和视频域等不同模态域的伪造数据，提出了单模态特征增强、跨模态协同表征及对抗鲁棒性强化等模型框架，旨在增强检测的泛化性与鲁棒性。然而，鲁棒性和泛化能力的提升不仅依赖于模型结构的优化，还需要在数据评估阶段进行科学、全面的检验。在这一背景下，学术界广泛使用了多个标准化数据集和统一的评估指标来评估模型的性能，确保研究成果的可比性与可靠性。

在公开发布的数据集方面，图像域的常见数据集包含 FaceForensics++（FF++）、Celeb-DF、DiffusionFace 等。FF++作为深度伪造领域的基准伪造数据集，涵盖了多种典型的面部篡改技术，集成了 DeepFakes、Face2Face、FaceSwap 及 NeuralTextures 等典型算法生成的伪造样本，为检测模型性能评估提供了多维度的实验基准。Celeb-DF 系列数据集基于变分自编码器的生成框架，实现了跨身份的面部动态特征同步建模。DiffusionFace 是首个专门针对基于扩散模型的人脸伪造的数据集，通过创新性地采用 11 种扩散模型架构，涵盖了从无条件生成到复杂的人脸交换技术等多种伪造手段，为基于扩散模型的深度伪造检测研究提供了多样性的标准化评估体系。音频域的常见数据集包含 ASVspoof 2021(DF Task)[22]、ASVspoof 2024[23] 等。ASVspoof 2021(DF Task)[22] 数据集通过集成逾百种语音转换（voice conversion, VC）与文本到语音（text to speech, TTS）合成系统，累计生成 58.9 万条涵盖多语种、多音色的伪造语音样本，为语音伪造检测模型的泛化能力验证提供了多维压力测试环境。ASVspoof 2024[23] 数据集首次系统整合对抗攻击系统作为核心伪造引擎，累计生成 81.5 万条伪造语音样本，构建了具备高隐蔽性的伪造语音库，标志着对抗性攻击技术在语音生物特征伪造检测研究中的范式迁移与评估体系升级。视频域常见的数据集为 DeepFake Detection Challenge、FakeAVCeleb[24] 和 Joint Audio-Video Deepfake[25]，这些数据集不仅涵盖高保真视觉伪造样本，还创新性地引入精准音唇同步的合成声纹或在既有视觉数据集中植入语音特征篡改层，从而实现了视听双通道的深度耦合伪造，为视频域数据提供了测试基准。

评估基准方面，深度伪造检测通常被视为二分类任务，广泛采用的评估指标包括准确率（accuracy, ACC）、ROC 曲线下面积（area under curve, AUC）、F1 分数（F1-score）、平均精度（average precision, AP）等。这些指标共同构成了深度伪造检测算法判别能力的量化评价基准。

在现有评估指标体系为深度伪造检测算法提供量化基准的基础上，为应对深度伪造技术快速迭代的现实挑战，浙江大学区块链与数据安全全国重点实验室研究团队构建了深度伪造检测算法评估平台，通过开源收集和自主生成来构建多种测试模块，根据评估需求，基于测试模块动态构建测试数据，从有效性、鲁棒性和泛化性 3 个层面综合评估检测算法的现实场景应用能力。

效率与性能权衡

深度伪造检测模型的应用领域存在终端设备部署、内容实时检测等效率需求，以满足各种现实场景的高可用性。蒸馏与量化技术通过模型压缩与计算优化提升深度伪造检测系统的效率与实用性。知识蒸馏[26] 将复杂教师模型的知识迁移至轻量化学生模型，利用软标签与特征匹配保留伪造痕迹的细粒度表征能力。例如 Albadawy 等[27] 利用知识蒸馏来针对生成对抗网

络遗留的频域伪影进行针对性学习。量化技术[28]通过低位宽参数替代浮点运算，结合动态范围校准与量化感知训练，在保持卷积核对微妙伪造特征敏感度的同时，将模型尺寸压缩为原尺寸的 1/3 甚至 1/4，推理速度提升 2 倍以上。蒸馏与量化二者协同优化时采用渐进式量化蒸馏策略，先在原始精度空间完成特征对齐，再分阶段降低位宽并微调，有效缓解精度损失。Lim 等[29]在此基础上提出“DistilDIRE”的轻量级扩散生成伪造检测框架，通过对基于扩散模型的逆向重构过程及其预训练分类器的知识进行联合蒸馏，构建出小规模且高效的检测网络。该方法在保持较高检测准确率的同时，显著减少了对扩散模型逆向推理的依赖，推理速度可提升至原始框架的 3 倍以上，为实际应用中实现快速、低成本的深度伪造检测提供了新的思路。

未来展望

深度伪造检测技术的突破需围绕鲁棒性、泛化性、数据评估与效能平衡四大维度实现系统性协同演进。

1）现有检测模型在抵御对抗攻击时的鲁棒性不足，这归因于模型架构固有的梯度可溯性。并且随着扩散模型等具有隐式梯度特性的生成技术普及，传统对抗训练策略的防御效果出现系统性退化。这种攻防不对称性导致检测系统在实际部署中面临较高的误判风险，特别是在金融身份认证等关键场景中可能引发灾难性后果。

2）针对泛化性不足的问题，当前主流研究聚焦于减轻模型在源域分布上的过拟合问题，或引入伪造方法的先验假设知识，但这种“打补丁式”的研究范式也导致检测技术始终滞后于伪造技术的演进速度。

3）研究社区缺乏统一且动态更新的数据与评估体系，无法匹配深度伪造技术的快速迭代。这也限制了建立实时的攻击–防御闭环验证环境。

4）现有检测模型在应用即时性上的研究严重匮乏。现有方法多基于离线批处理模式，而面对实时视频会议或直播流媒体等场景时，主流框架往往因计算复杂度过高导致响应延迟超标，制约了在这些场景中的应用价值。

鉴于当前深度伪造检测技术发展遇到的种种问题，未来的发展方向应聚焦于提升检测系统的整体适应性和安全性。通过构建更为灵活的模型架构，使检测系统能够在面对未知伪造算法时，依然保持高效判别能力，突破仅针对特定分布数据训练的局限。主动防御机制的研究也十分重要，研究者需要探索动态防御策略，主动识别并防范对抗样本攻击，从而有效抑制因梯度可溯性带来的系统脆弱问题。实时检测与轻量化部署的实现也是未来研究的重要方向。针对实时视频会议和流媒体应用，需对算法进行深度优化，使之在移动端或边缘设备上实现低延迟、高效率地运行。此外，突破单一模态限制，将音频、图像、视频等多种数据形式整合，从而实现信息的互补与协同的多模态融合检测技术在提升检测精度，拓展应用场景等方面也极具发展潜力。

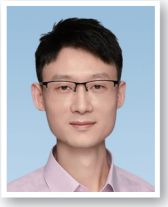
另一方面，大模型的能力涌现为扩展检测技术的边界提供了新的可能性。在检测算法层面，基于混合专家模型（mixture of experts, MOE）架构的万亿参数模型可通过隐空间特征解耦，实现对伪造痕迹的量子级感知；多模态大模型的跨模态对齐能力，为视频–音频–文本协同检测提供统一表征框架；防御机制创新方面，大模型的 CoT 推理能力使检测系统可构建伪造意图推演网络，通过分析内容传播路径、用户行为模式等元数据，实现从被动鉴别到主动防御的范式跃迁；未来，大模型技术的发展与深度伪造检测场景的结合有望推动现有的单一鉴别模式向智能分析与对抗的范式转变，迈入智能检测新阶段。

需要特别强调的是，技术突破必须与制度建设形成合力。在推进检测算法创新的同时，加强法律与道德规范建设势在必行。通过建立严格的监管机制和责任追溯体系，推动深度伪造技术在合法合规框架内应用。构建全球统一的技术标准和伦理准则，不仅有助于规范行业行为、预防技术滥用，更能为公众数字权益建立长效保护机制，最终形成技术治理与法律约束协同发展的良性生态。 ■

任　奎

CCF 会士、数据治理发展委员会副主任。AAAS/ACM/IEEE 会士。浙江大学求是讲席教授，计算机科学与技术学院院长、区块链与数据安全全国重点实验室常务副主任。主要研究方向为人工智能安全、数据安全与隐私保护。

kuiren@zju.edu.cn



林　峰

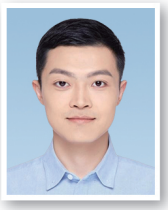
CCF 杰出会员。浙江大学网络空间安全学院/计算机学院百人计划研究员。主要研究方向为深度伪造检测、智能网联车安全、物联网安全。flin@zju.edu.cn

巴钟杰

CCF 专业会员。浙江大学网络空间安全学院/计算机学院百人计划研究员。主要研究方向为 AIGC 安全、深度伪造检测、物联网安全。zhongjieba@zju.edu.cn

刘振广

CCF 专业会员。浙江大学网络空间安全学院/计算机学院百人计划研究员。主要研究方向为人工智能、区块链。
liuzhenguang2008@gmail.com



程　鹏

CCF 专业会员。浙江大学区块链与数据安全全国重点实验室专职研究员。主要研究方向为人工智能生成内容安全、语音数据隐私与安全。
peng_cheng@zju.edu.cn

参考文献

[1] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. (2022−12−10)[2025−03−02]. https://arxiv.org/abs/1312.6114.

[2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. *Communications of the ACM*, 2020, 63(11): 139−144.

[3] NICHOL A Q, DHARIWAL P. Improved denoising diffusion probabilistic models[J]. *Proceedings of the 38th International Conference on Machine Learning*. 2021(139): 8162−8171.

[4] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10674−10685.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc. 2017: 6000−6010.

[6] HULZEBOSCH N, IBRAHIMI S, WORRING M. Detecting cnn-generated facial images in real-world scenarios[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020: 2729−2738.

[7] NEEKHARA P, DOLHANSKY B, BITTON J, et al. Adversarial threats to deepfake detection: a practical perspective[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE, 2021: 923−932.

[8] LI Hanzhe, LI Yuezun, ZHOU Jiaran, et al. Freqblender: enhancing deepfake detection by blending frequency knowledge[C]//38th Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2024: 1−24.

[9] GUO Bin, DING Yasan, YAO Lina, et al. The future of false information detection on social media[J]. *ACM Computing Surveys*, 2021, 53(4): 1−36.

[10] PAN Deng, SUN Lixian, WANG Rui, et al. Deepfake detection through deep learning[C]//*2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. Piscataway: IEEE, 2020: 134−143.

[11] WANG Chengrui, DENG Weihong. Representative forgery mining for fake face detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 14923−14932.

[12] TAK H, KAMBLE M, PATINO J, et al. Rawboost: a raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 6382−6386.

[13] TAN Chuangchuang, ZHAO Yao, WEI Shikui, et al. Frequency-aware deepfake detection: improving generalizability through frequency space domain learning[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(5): 5052−5060.

[14] YAN Zhiyuan, ZHANG Yong, FAN Yanbo, et al. UCF: uncovering common features for generalizable deepfake detection[C]//*2023 IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 2023: 22355−22366.

[15] WU Zhizheng, YAMAGISHI J, KINNUNEN T, et al. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2017, 11(4): 588−604.

[16] KANG W H, ALAM J, FATHAN A. CRIM's system description for the ASVSpoof2021 challenge[C]//2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge: Virtual. ISCA, 2021: 100−106.https://www.isca-archive.org/asvspoof_2021/kang21b_asvspoof.html.

[17] CUCCOVILLO L, GERHARDT M, AICHROTH P. Audio transformer for synthetic speech detection via formant magnitude and phase analysis[C]//ICASSP 2024−2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul: IEEE, 2024: 4805−4809.

[18] DAS R K. Known-unknown data augmentation strategies for

detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021[C]//2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge:Virtual. ISCA, 2021: 29−36. https://www.isca-archive.org/asvspoof_2021/das21_asvspoof.pdf.
[19] RAVANELLI M, BENGIO Y. Speaker recognition from raw waveform with SincNet[C]//2018 IEEE Spoken Language Technology Workshop. Piscataway: IEEE, 2018: 1021−1028.
[20] ZEGHIDOUR N, TEBOUL O, QUITRY F D C, et al. LEAF: a learnable frontend for audio classification[C]// International Conference on Learning Representations. Vienna: ICLR. 2021: 1−16.
[21] YANG Wenyuan, ZHOU Xiaoyu, CHEN Zhikai, et al. AVoiD-DF: audio-visual joint learning for detecting deepfake[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 2015−2029.
[22] YAMAGISHI J, WANG Xin, TODISCO M, et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection[EB/OL]. (2021−09−01)[2025−05−14]. https://arxiv.org/abs/2109.00537v1.
[23] The Asvspoof 2024 Challenge[EB/OL]. (2024−07−31)[2025−03−02]. https://www.asvspoof.org/.
[24] KHALID H, TARIQ S, KIM M, et al. FakeAVCeleb: a novel audio-video multimodal deepfake dataset[EB/OL]. (2021−08−11)[2025−03−02]. https://arxiv.org/abs/2108.05080v4.
[25] ZHOU Yipin, LIM S N. Joint audio-visual deepfake detection[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 14780−14789.
[26] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. *Computer Science*, 2015, 14(7): 38−39.
[27] ALBADAWY E A, LYU S, FARID H. Detecting AI-synthesized speech using bispectral analysis[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Piscataway: IEEE, 2019: 1−8.
[28] JACOB B, KLIGYS S, CHEN Bo, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2704−2713.
[29] LIM Y, LEE C, KIM A, et al. DistilDIRE: a small, fast, cheap and lightweight diffusion synthesized deepfake detection[EB/OL]. (2024−06−02)[2025−03−02]. https://arxiv.org/abs/2406.00856v1.

Deepfake detection: key challenges and technical approaches

REN Kui, LIN Feng, BA Zhongjie, LIU Zhenguang, CHENG Peng
Zhejiang University

Abstract: The remarkable advancements in generative artificial intelligence have ushered in deepfake technology. This technology enables high-fidelity manipulation and synthesis of multimedia content with unprecedented ease, thereby posing significant threats to digital society security. In this regard, this study systematically identifies four core challenges hindering deepfake detection: insufficient robustness in complex scenarios, limited generalization capability for novel forgery techniques, scarcity of standardized databases and evaluation benchmarks, and the inherent trade-off between computational efficiency and detection accuracy. In response to these challenges, this paper proposes a multifaceted technical framework, including frequency-domain decoupled dynamic feature extraction, multimodal contrastive learning, adversarial robustness augmentation, and lightweight model deployment. Lastly, this paper explores the establishment of a synergistic governance paradigm that facilitates the coordinated evolution of legal norms and detection technologies.

Keywords: generative artificial intelligence; deepfake detection; multimedia content security; artificial intelligence security; face forgery detection; audio forgery detection; adversarial robustness enhancement; multimodal contrastive learning

摘　要: 生成式人工智能的突破性发展催生了深度伪造技术，其对多媒体内容轻易且高质量地操纵与合成已严重威胁数字社会安全。对此，系统剖析了深度伪造检测面临的核心挑战，包括复杂场景下的鲁棒性不足、面对新型伪造的泛化性受限、数据底座与评估基准缺失、效率与精度失衡。针对这些挑战，本文进一步分析了基于频域解耦的动态特征提取、多模态对比学习、对抗鲁棒增强及模型轻量化部署等可行技术路径。在最后的展望中探讨了构建法律规范与检测技术协同演进的治理体系。

关键词: 生成式人工智能；深度伪造检测；多媒体内容安全；人工智能安全；人脸伪造检测；音频伪造检测；对抗鲁棒增强；多模态对比学习

中图分类号: TP309.2

中文引用格式: 任奎, 林峰, 巴钟杰, 等. 深度伪造内容检测的挑战及可行技术路径 [J]. 计算, 2025, 1(2): 8−15.

英文引用格式: REN Kui, LIN Feng, BA Zhongjie, et al. Deepfake detection: key challenges and technical approaches[J]. *Computing Magazine of the CCF*, 2025, 1(2): 8−15.