Deepfake 检测技术的可信性与可解释性研究进展综述

潘瑶婷 陈澳迪 韦彩虹 唐钰婷 杨小梅（通讯作者）

广西职业师范学院，广西南宁，530007；

摘要：Deepfake 技术的飞速发展对社会信任体系与信息安全构成严峻威胁，推动检测技术成为学术界与工业界的关注焦点。当前，尽管检测模型在受控环境下的识别准确率较高，但其在真实场景中的可信性与可解释性仍存在明显不足，制约了在司法取证、新闻核查等高可信要求领域的实际应用。可信性方面，现有模型面临跨数据集泛化能力弱、对噪声与对抗攻击鲁棒性差、结果可靠性缺乏统计依据等关键问题；可解释性方面，多数模型仅提供二分类输出，缺乏对决策依据与伪造痕迹的直观解释，难以令人信服。本文系统综述了近年来 Deepfake 检测技术在可信性与可解释性方面的研究进展，分别从迁移性优化、鲁棒性增强、可靠性评估，以及基于取证痕迹、模型透明化与多模态解释等角度梳理主要方法，并总结了相关数据集、评估指标与实际应用。最后，本文指出该领域仍面临生成技术快速迭代、极端场景适应性不足、解释细粒度有限等挑战，未来需进一步探索自适应检测架构、统一评估标准、细粒度可解释框架以及可信与可解释协同优化的技术路径，以推动构建可靠、透明、实用的 Deepfake 治理体系。

关键词：Deepfake 检测；可信人工智能；可解释人工智能；多模态融合

DOI：10.69979/3041-0673.26.04.005

引言

Deepfake 技术依托深度学习算法，能够对图像、音频、视频等数字媒体内容进行高度逼真的篡改与合成，实现包括人脸替换、表情操控、语音模仿在内的多种伪造效果。近年来，随着 DeepFaceLab、MidJourney 等开源工具的普及，以及 Stable Diffusion、VITS 等生成模型的优化与轻量化，普通用户已能够以较低成本生成难以肉眼辨别的伪造内容[1]。这一趋势直接助长了虚假信息传播、数字身份盗用、基于 AI 语音的诈骗等恶性事件，例如 2019 年发生的利用 AI 模仿企业高管声音骗取巨额资金的案件，以及 2023 年以来多起涉及政治人物的 Deepfake 视频所引发的社会信任危机，均凸显出该项技术对社会治理与公共安全的潜在威胁。

1 核心概念与技术基础

1.1 检测技术的分类体系与发展脉络

根据检测对象与机理的不同，当前 Deepfake 检测技术已形成多维度发展的格局。视觉检测主要针对人脸图像与视频，通过分析空间域的纹理一致性、频率域的生成伪影以及时间域的动作连贯性来识别伪造痕迹。音频检测则聚焦于合成语音，从声学特征、韵律模式等维度捕捉不自然之处。随着伪造手段的复杂化，多模态检测日益受到重视，其通过融合视觉、音频乃至文本信息，利用跨模态的一致性（如唇音同步）进行综合研判[2]。此外，面对扩散模型等新一代生成技术，不限于特定内容的通用生成内容检测也逐渐兴起，其核心在于学习不同生成模型的固有指纹。这一分类体系不仅反映了技术发展的演进路径，也为理解不同方法的适用范围提供了清晰框架。

1.2 可信性与可解释性的内涵界定

在 Deepfake 检测领域，可信性与可解释性是评估技术实用价值的两个核心维度。可信性指检测系统在真实复杂环境中稳定、可靠运行的能力，具体涵盖三个方面：迁移性强调模型对未知数据分布和新型伪造算法的适应能力；鲁棒性关注系统抵御数据质量退化与恶意对抗干扰的韧性；可靠性评估则旨在通过统计学方法量化输出结果的可信程度，为高风险决策提供依据。可解释性则致力于解决模型决策过程的“黑箱”问题，其目标是以人类可理解的方式（如可视化证据或逻辑阐述）揭示判断依据，核心在于回答“为何判定为伪造”以及“证据何在”等关键问题，从而建立用户对自动化检测结果的技术信任。

1.3 关键技术基础

Deepfake 检测技术的发展建立在一系列特征提取与模式识别的基础之上。传统方法依赖于手工设计的特

征描述，而当前主流方法则广泛采用深度学习架构进行端到端的特征学习与分类。卷积神经网络、视觉 Transformer 等模型在空间与时序特征提取方面表现出色；与此同时，针对可解释性的需求，类激活映射、局部可解释模型等事后分析工具，以及注意力机制、胶囊网络等内生可解释结构，共同构成了提升模型透明度的技术工具箱。这些基础技术不仅支撑着检测性能的提升，也为实现可信可解释的检测系统提供了必要的方法论与实现手段。

2 Deepfake 检测技术的可信性研究进展

2.1 迁移性优化：从过拟合到泛化适应

当前检测模型面临的核心可信性挑战之一，在于其性能高度依赖于训练数据所限定的特定伪造特征分布，从而导致在面对未知生成算法或真实开放场景时泛化能力显著下降。为解决此迁移性瓶颈，近期研究呈现出从依赖特定伪影转向学习通用生成指纹的趋势。具体而言，一方面通过构建模态无关的架构，旨在提取如扩散模型噪声模式等超越具体算法类型的本质特征；另一方面，零样本与少样本学习框架受到重视，借助检索增强生成等技术整合先验知识，实现对新兴伪造手段的快速响应[3]。此外，深度融合视觉、音频等多模态线索，利用其内在互补性来构建更为稳健的跨域表示，已成为提升模型适应性的有效途径。这些努力共同推动了检测系统从封闭实验室环境向复杂现实世界的过渡。

2.2 鲁棒性增强：抵御失真与对抗攻击

检测模型在实际部署中必须抵御两类主要干扰：一是内容在传播链中不可避免的质量损耗，如压缩编码、分辨率下降与随机噪声；二是恶意攻击者精心设计的对抗性扰动。针对前者，研究主要通过引入涵盖多种失真类型的数据增强策略，使模型在训练阶段即暴露于模拟的退化环境中，从而学习对质量波动不敏感的鲁棒特征。对于后者，则普遍采用对抗训练将扰动样本纳入优化过程，或从特征工程入手，利用频域变换等对像素级扰动相对不敏感的表征方法。更有前沿工作将检测与溯源相结合，通过识别生成来源进而采用针对性防御策略，形成了“识别-适应”的主动鲁棒性增强范式，显著提升了系统在对抗环境下的生存能力。

2.3 可靠性评估：从性能指标到统计信度

传统基于准确率、召回率等指标的评估体系，难以衡量模型在不确定的真实场景中的统计可靠性，这限制了检测结果在司法等高风险领域的采信度。为此，当前研究正致力于建立更为严谨的、基于统计推断的评估框架。该类方法的核心在于构建一个尽可能逼近真实世界数据分布的总体样本空间，涵盖不同伪造技术、质量等级与应用情境。通过系统的随机采样模拟，计算模型在不同置信水平下的性能边界，从而量化其决策的可信程[4]。这种评估范式不仅揭示了模型在受控测试集与开放环境中的表现差异，更重要的是为检测结果提供了概率化的解释框架，使“可靠”不再停留于抽象概念，而是具备了可计算、可报告的形式化基础，为技术落地提供了关键的信任桥梁。

3 Deepfake 检测技术的可解释性研究进展

3.1 取证痕迹分析：揭示物理与算法指纹

可解释性的核心诉求在于为检测结论提供客观、可验证的证据。基于取证痕迹的分析方法通过揭示数字内容中违背真实世界物理规律或残留的生成算法特定指纹，直接回应了这一需求。在视觉层面，该方法系统性地检测光照方向不一致、面部生物力学运动异常、以及生成对抗网络所引入的网格状纹理伪影等内在矛盾[5]。音频取证则专注于分析合成语音中不自然的声学特征，例如过于平坦的韵律轮廓、频谱不连续点或特定文本到语音模型遗留的共振峰分布模式。随着检测任务从二分类判断向精细化定位发展，例如要求在音频中标记伪造片段的起止时间，取证痕迹分析已从辅助性解释工具演进为支撑检测决策的关键技术分支，其输出结果因其客观性与可验证性，在司法鉴定等严肃场景中具有重要价值。

3.2 模型透明化技术：从“黑箱”到可理解的决策过程

为直接打开模型决策的“黑箱”，研究者发展了旨在提升过程透明度的模型中心化方法。这类方法可分为内嵌透明性与事后解释两类路径。内嵌透明性指在模型设计阶段即集成可解释模块，例如利用注意力机制自动生成凸显关键判别区域的热力图，或通过胶囊网络动态路由机制直观展示特征组合的逻辑。事后解释则不改变模型内部结构，而是借助外部工具进行逆向分析，如采用梯度加权类激活映射可视化深层神经网络的激活区域，或运用基于博弈论的 Shapley 值等特征归因方法，

定量评估不同输入特征（如特定像素区域或音频帧）对最终“伪造”判断的贡献度。这些技术将模型的内部计算过程转化为人类可感知的形式，不仅增强了用户信任，也为模型自身的优化与调试提供了至关重要的洞察。

3.3 多模态解释生成：构建面向用户的综合证据体系

最终的可解释性需服务于不同背景的用户，因此，将技术证据转化为直观易懂的综合报告成为当前的研究前沿。多模态解释生成框架旨在融合视觉标记、音频对比与自然语言描述，形成一个层次化的证据展示体系。其典型实现方式是，首先通过检测模型提取关键的可视化证据（如伪造区域热力图）与量化特征；随后，借助大型语言模型的语义生成与推理能力，将这些技术性特征组织成逻辑连贯、语言自然的叙述性报告，明确指出“何处存疑”与“为何存疑”。更先进的系统则引入检索增强生成技术，动态关联外部伪造知识库，使生成的解释不仅描述现象，更能关联到潜在的生成方法与技术特征，从而显著提升了解释的深度、准确性与用户的可理解度，推动了检测技术从专家工具向公共服务的关键转变。

4 总结

本文系统梳理了近年来 Deepfake 检测技术在可信性与可解释性方面的主要进展。研究表明，可信性研究的核心突破体现在通过通用特征学习、零样本适应及多模态融合以提升模型迁移性，借助抗失真训练与对抗鲁棒设计增强系统鲁棒性，并引入基于统计推断的可靠性评估方法为检测结果提供量化依据。在可解释性方面，取证痕迹分析已从理论走向实用，能够识别生成伪影与物理矛盾；模型中心方法通过可视化与特征归因揭示决策依据；而融合多模态信息与自然语言生成的解释框架，进一步降低了专业门槛，提升了结果的可信度与可接受性。当前该领域仍面临诸多挑战，包括生成技术快速迭代导致的检测滞后、复杂真实场景下的鲁棒性不足、细粒度解释与多模态融合的困难，以及可信与可解释目标之间的内在权衡。展望未来，研究应致力于发展自适应与持续学习的检测架构，构建标准化评估体系，推动细粒度、因果性与用户自适应的解释方法，并探索与数字水印、区块链等技术的跨域融合，最终为构建可靠、透明、实用的 Deepfake 治理体系提供理论支撑与技术路径。

参考文献

[1]时超轶.面向跨域场景的 Deepfake 检测方法研究[D].齐鲁工业大学,2025.DOI:10.27278/d.cnki.gsdqc.2025.000142.

[2]许裕雄,李斌,谭舜泉,等.语音深度伪造及其检测技术研究进展[J].中国图象图形学报,2024,29(08):2236-2268.

[3]陈梦秋.深度伪造信息的特征识别、行为交互与反制研究[D].南昌大学,2024.DOI:10.27232/d.cnki.gnchu.2024.004592.

[4]李彝利,姚洁彤,郎健,等.视频虚假新闻检测：方法、挑战与可解释性研究[J/OL].计算机科学,1-28[2025-12-07].https://link.cnki.net/urlid/50.1075.TP.20251203.1315.002.

[5]丁峰,匡仁盛,周越,等.深度伪造及其取证技术综述[J].中国图象图形学报,2024,29(02):295-317.

作者简介：1）潘瑶婷（2004 年-），女，壮族，广西贵港，本科，研究方向：人工智能；

2）陈澳迪（2002 年-），女，汉族，广西钦州，本科，研究方向：人工智能；

3）韦彩虹（2002 年-），女，壮族，广西都安瑶族自治县，本科，研究方向：人工智能；

4）唐钰婷（2003 年-），女，汉族，广西容县，本科，研究方向：人工智能；

通讯作者：杨小梅（1981 年-），女，汉族，山东青岛，博士研究生，研究方向：计算语言学，人工智能。

项目信息：本项目由国家级大学生创新创业训练计划项目资助，项目名称：广西职业师范学院 2025 年大学生创业训练计划项目《生成式人工智能 Deepfake 技术的网络舆论风险防控与治理研究》，项目级别：国家级，项目类别：一般项目，项目编号：202514684005；