AIGC深度伪造主动防御与溯源认证综述

韩禹洋，王志强，陈颖，欧海文
（北京电子科技学院，北京 100070）

摘要：人工智能生成内容（AIGC）深度伪造正在向多模态生成、跨平台传播和再生成攻击扩展，单纯被动检测难以满足来源核验需求。首先，围绕内容生命周期证据链，梳理生成前防护、生成端标识、内容凭证、传播保持和协同验证研究；其次，从功能目标、成立条件、可支持结论和失效边界比较代表方法，分析表明传播扰动会使不同证据保持、衰减、缺失与冲突；最后，指出跨平台证据保持、多模态一致性和动态攻防评测仍是重点方向。

关键词：人工智能生成内容；深度伪造；主动防御；溯源认证；内容凭证；证据链
中图分类号：TP309
文献标志码：A
doi：10.11959/j.issn.1000

Survey on Proactive Defense and Provenance Authentication for AIGC Deepfakes

Han Yuyang, Wang Zhiqiang, Chen Ying, Ou Haiwen
Beijing Electronic Science and Technology Institute, Beijing 100070, China

Abstract: As artificial intelligence-generated content (AIGC) deepfakes expanded toward multimodal generation, cross-platform dissemination, and regeneration attacks, passive detection alone became insufficient for provenance verification. Firstly, studies on pre-generation protection, generation-side identifiers, content credentials, evidence preservation, and collaborative verification were reviewed along a content-lifecycle evidence chain. Secondly, representative methods were compared in terms of functional objectives, validity conditions, supported claims, and failure boundaries. Under dissemination perturbations, different types of evidence were found to be retained, degraded, lost, or placed in conflict. Finally, cross-platform evidence preservation, multimodal consistency, and dynamic adversarial evaluation remain key research directions.

Key words: artificial intelligence-generated content, deepfake, proactive defense, provenance authentication, content credentials, evidence chain

0 引言

人工智能生成内容（artificial intelligence generated content，AIGC）技术正在重塑数字内容的生产和传播方式。以扩散模型[1-3]、生成对抗网络（generative adversarial network，GAN）[4-5]和大语言模型[6]为代表的生成技术，显著降低了高逼真图像、视频、语音和文本内容的制作门槛。深度伪造对象也由早期的人脸替换扩展到身份、声音、动作、语义关系和跨模态一致性等多个维度[7-9]。在网络诈骗、舆情操纵、版权侵权和司法取证等场景中，伪造内容不再只是“真假识别”问题，而是进

收稿日期：XXXX–XX–XX；修回日期：XXXX–XX–XX
通信作者：王志强，wangzq@besti.edu.cn
基金项目：中央高校基本科研业务费专项资金资助项目（No. 3282025044，No. 3282026021）
Foundation Items：The Fundamental Research Funds for the Central Universities (No.3282025044, No.3282026021)

一步演化为来源声明、传播链路、编辑历史和责任主体能否被复核的问题。

被动检测在过去数年取得了大量进展。从手工特征提取[10-12]到深度学习分类器[13-15]，从频域与时序分析[16-18]到域泛化和基础模型适配[19-22]，相关方法能够在若干基准数据集上识别伪造痕迹。然而，AIGC 深度伪造的真实传播环境削弱了单点检测的治理价值。首先，检测器依赖生成过程残留的伪影或统计差异，随着生成质量提升和模型类型扩展，稳定特征不断减少[23]。其次，检测输出多为概率、标签或热力图，难以独立说明内容的生成来源、具体编辑操作及其先后顺序，更不足以验证原始来源声明的完整性[24]。再次，真实平台会引入压缩、缩放、转码、再上传等处理，使跨平台条件下的检测性能出现衰减[25]。同时，训练集伪造类型与测试集生成技术不一致时，检测器在跨数据集、跨生成器条件下仍面临泛化压力[26]。

因此，除事后真假判别外，近年来的深度伪造防护研究也开始关注内容生命周期中的主动证据构建与证据可信性评估。主动防御在素材发布前或生成过程中介入，通过反生成扰动、平台鲁棒扰动和身份特征遮蔽等方式降低身份滥用风险[27-30]。数字水印和生成端标识在内容或模型输出中嵌入可检测信号，为后续验证提供主动证据[31-35]。内容凭证、数字签名和来源认证机制记录生成、编辑和发布过程，使内容完整性和来源声明能够被复核[36-39]。这些技术并非替代被动检测，而是在不同阶段承担证据生成、证据保持、证据验证和冲突复核功能。

已有综述分别梳理视觉深度伪造检测[40-41]、人脸主动防御[42]、AIGC 视觉内容生成与溯源[43]及人脸检测演进[44]，但对三类问题讨论不足：生成前防护、生成端标识与内容凭证如何衔接，证据在生成、发布、传播和验证阶段如何保持或失效，以及再生成、凭证剥离、多次嵌入和跨平台转码如何引发证据冲突。深度伪造防护不能只关注单次真伪判别性能，还需考察证据能否在内容生命周期中持续形成、保持并被验证，还要进一步分析主动防御、生成端标识和来源凭证如何在生成、发布、传播与验证阶段衔接，传播扰动后各类证据如何衰减、缺失或产生冲突，以及不同证据能否在验证端形成互补。针对这些空缺，本文以内容生命周期证据链为主线，比较相关技术的证据形式、成立条件和失效模式，并讨论多源证据协同验证与风险分级。

本文的主要工作如下：1）从内容生命周期视角归纳生成前防护、生成端标识、发布端凭证、传播中保持和验证端协同 5 类研究进展，明确被动检测、生成前主动防御、生成端标识、内容凭证与溯源认证的技术边界及互补关系。2）围绕功能目标、成立条件、可支持结论和失效边界比较代表方法，形成统一的跨机制分析维度。3）归纳压缩转码、截图重传、再生成、凭证剥离和多次嵌入等条件下的证据失效机制，并从综述角度整理多源证据协同验证和风险分级的分析流程。4）结合现有评测资源和部署需求，总结跨平台证据保持、多模态一致性、证据冲突推理、标准化接口和动态攻防评测等后续方向。

本文采用主题检索与引文追溯相结合的文献获取方式。检索范围包括 Web of Science Core Collection、IEEE Xplore、ACM Digital Library 和中国知网等数据库，并补充检索 arXiv、官方标准及平台技术文档。检索时间范围为 2017 年至 2026 年 8 月，关键词由深度伪造与 AIGC、主动防御与身份保护、水印与模型标识、内容凭证与来源认证、传播攻击与协同验证等词组组合设置。文献依次经过去重、题名与摘要初筛、全文复核及前向和后向引文追溯。纳入文献应直接提出相关机制、提供评测结果、形成公开标准或代表实际部署。排除仅涉及一般内容生成、与可信核验无直接关系、缺少方法信息以及已有正式版本的重复预印本。被动检测仅选取用于说明技术演进、泛化边界和协同验证的代表性工作。

图 1 从内容生命周期和证据功能两个维度概括相关研究的分类关系与证据链分析视角，用于说明各类技术在证据生成、保持与验证中的位置关系，并不表示固定系统架构。

1 AIGC 深度伪造防护技术边界

1.1 生成技术演进与防护需求

AIGC 是利用生成式人工智能生成或编辑文本、图像、音频和视频等内容的广义生产范式。本文讨论的深度伪造内容，是其中以人物身份、言行或事件语义为对象进行逼真合成或操纵，并具有身份冒用、事实歪曲、来源伪装或欺骗性传播风险的高风险内容类型[7-9]。因此，AIGC 内容不必然构成

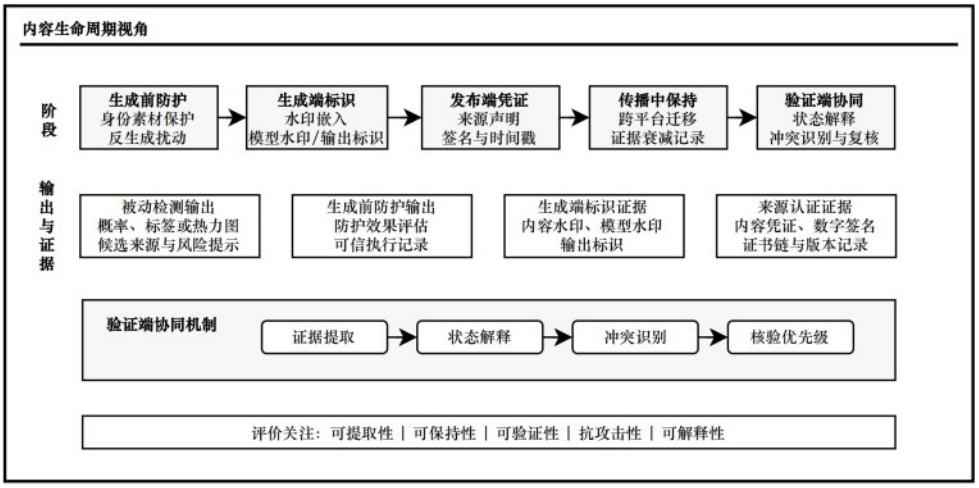

图1　内容生命周期视角下 AIGC 深度伪造防护机制、证据机制及验证关系

深度伪造。本文所称“AIGC 深度伪造”即指该类内容，不讨论与这些风险无关的一般 AIGC 内容。在此范围内，相关技术由 GAN 驱动的人脸合成与属性编辑，发展至扩散模型图像、视频生成以及多模态基础模型支持的跨模态生成[1-5]。生成内容质量的持续提升使得伪造痕迹逐渐从肉眼可辨变为模型也难以稳定检测。

在早期 GAN 阶段，生成的人脸图像在上采样、归一化和特征图处理过程中会引入可观测的棋盘状伪影、频谱异常和局部纹理不一致[10-11]，这些线索为手工特征和浅层检测器提供了依据。2020—2021 年，StyleGAN2[4]等模型通过改进归一化和生成结构减弱了部分可见伪影，FaceShifter 等面部交换方法提高了身份保持和融合质量，使检测方法需要进一步利用融合边界、频域统计和时序不一致等更细粒度特征[13-14]。为降低对特定伪造算法伪影的依赖，Self-Blended Images[45]通过自融合样本模拟融合边界和源/目标统计差异，提升跨数据集和跨操纵泛化能力，重构与分类联合学习与 UCF 分别从真实人脸重构差异、通用伪造特征解耦角度抑制对已知伪造模式的过拟合[46-47]。RealForensics[48]利用真实说话人视频中的音视对应关系学习时序表征，UniForensics[49]进一步引入高层人脸语义表征和动态视频自融合样本，以增强跨操纵、跨数据集和压缩条件下的鲁棒性。2022 年以来，扩散模型[1-3]通过逐步去噪方式生成图像，其在噪声残差、重构误差和频谱统计方面与 GAN 存在差异，使部分基于 GAN 上采样伪影训练的检测器面临泛化压力[23]。2024 年以来，多模态基础模型和扩散生成模型的发展使伪造内容进一步覆盖长时序视频、音视频同步和文本驱动的跨模态生成[50-51]，浅层像素纹理和局部频谱特征上的差异持续减弱[20-22]。

上述演化表明，深度伪造风险已从早期单一模态、局部伪影的形态，发展为多模态、高逼真、可传播、难溯源的复杂威胁。面向 AIGC 内容可信的防护体系不能仅依赖对特定生成器伪影的事后捕捉，而需要在内容生命周期的更早阶段主动介入。

1.2　被动检测与主动证据构建

被动检测是当前深度伪造防护中研究最为充分的技术路线。其基本范式是在内容生成并传播之后，通过分析图像、视频、语音或多模态内容的内部特征判断其是否疑似伪造。该路线在空间伪影检测[13-14]、频域残差分析[16-17]、时序不一致建模[18]、生理信号验证[11]和跨模态关系分析[52-53]等方面形成了大量成果。

然而，在 AIGC 内容大规模生成与跨平台传播的背景下，被动检测难以独立支撑内容可信治理。首先，检测器通常依赖训练数据中覆盖的伪造伪影或统计差异，在未知生成模型、新压缩强度和真实平台处理条件下容易出现性能衰减[25-26]。其次，被动检测多输出概率、标签或热力图，虽能提示伪造风险，却难以独立追溯内容的完整编辑序列——包括生成工具、操作类型、执行顺序及中间状态——因而难以构成来源认证、完整性验证和责任核验所需的闭环证据链[24]。此外，深度学习检测器的决策依据通常不够透明，已有研究尝试通过 LIME 等[54]解释方法定位影响检测器决策的关键区域。然而，这类可视化结果主要反映模型决策敏感区

域，仍不能直接等同于稳定、可复核的取证证据。

深度伪造防护正在由事后判别拓展至可验证信息的预先构建。本文将生成端标识、来源凭证和可信传播记录统称为主动证据，它们在内容生成、发布或传播阶段形成，并可由验证端提取或核验。生成前主动防御主要降低素材滥用和伪造成功风险，其防护效果不直接构成来源证明，执行状态经可信记录并与素材绑定后，可形成审计线索。主动证据可从可提取性、可保持性、可验证性、抗攻击性和可解释性等方面评价。本文所称“证据链”是面向技术综述的分析框架，用于描述证据在生成、发布、传播和验证阶段的形成与衰减关系，并不等同于司法取证中的法定证据保全链条。

1.3　被动检测、主动防御与溯源认证

本文采用二维分类方式组织相关研究。纵向按照内容生命周期划分为生成前防护、生成端标识、发布端凭证、传播中保持和验证端协同5个阶段，横向按照技术在证据链中的作用区分为被动检测、生成前防护、生成端标识和来源认证4类功能。其中，生成前防护主要承担风险抑制作用，其余技术分别形成风险判断、主动标识和来源声明等验证证据。多源证据协同验证不作为与上述技术并列的基础类型，而是验证端对检测结果、水印状态、凭证状态和传播记录进行融合、解释与复核的机制。

不同技术支持的核验命题并不相同。预防与防御用于降低素材滥用或伪造成功率，检测用于判断内容是否具有伪造风险，归因用于推断相关模型、设备或来源主体，认证用于核验来源声明、签发主体和内容绑定关系，溯源则依据版本及处理记录恢复内容的派生链路。

第一类是被动检测机制，其在内容生成或传播后分析内部特征，并输出概率、标签或热力图等统计风险证据。其部署灵活，不要求生成方预先配合，适用于缺少凭证的场景，不足是泛化受限，难以独立回答来源认证问题。

第二类是生成前防护措施，在伪造发生前或素材发布前介入，通过身份保护扰动和反生成扰动降低素材被利用或伪造成功的风险。该类技术的主要输出是防护效果，而不是传播后可直接提取的来源证据。只有防护配置、执行时间和保护状态等信息被可信记录并与素材绑定时，才能形成审计线索。其不足是对跨模型迁移性和平台兼容性要求较高。

第三类是生成端标识证据，典型技术包括内容侧水印、生成端水印、模型水印。其优势在于可在内容或模型输出中形成后续可提取的标识，不足是依赖生成端或嵌入端配合，并会受到压缩、再生成、模型蒸馏和多次嵌入等攻击影响。

第四类是来源认证证据，关注内容来源、编辑历史和完整性的验证，典型技术包括内容凭证、数字签名和可信时间戳。其证据语义较清晰，能支持来源声明和来源核验，不足是依赖信任链和平台配合，且可能受到凭证剥离、重绑定和伪造攻击。

4类技术在内容生命周期中承担不同功能。被动检测提供无凭证场景下的风险判断，生成前防护提高伪造成本，生成端标识形成可提取信号，来源认证记录并验证内容来源。协同验证根据各类技术的功能、输出及成立条件组织，不将风险控制措施与可验证证据归为同一类型。为比较不同技术机制的功能定位、成立条件及保证边界，表1在上述4类功能基础上细分为7类技术机制，从功能目标、信任基础和失效边界等方面进行归纳。

1.4　内容生命周期证据链分析框架

证据链分析以具有派生关系的内容版本为基本单元。每个版本同时关联传播操作、证据单元、声明主体和验证依据。内容生成形成初始版本，编辑、转码、截图和再生成产生派生版本。每个证据单元至少包含证据载荷、绑定对象、产生或签发主体、时间信息和验证方式。检测结果、生成端标识、来源凭证和传播记录与相应版本关联。生成前防护主要承担风险控制功能，其配置、执行时间和保护状态经可信记录与素材绑定后，可作为审计线索。

证据链的连续性与可验证性由绑定、继承、时间和信任四类关系共同决定。绑定关系用于确认内容版本和声明主体，继承关系说明派生版本保留、迁移或重新签发的证据，时间关系用于核对生成、编辑、签发和验证顺序，信任关系则涉及签名密钥、证书路径、时间戳和吊销状态。编辑与传播操作可能使这些关系保持、衰减或断裂。证据无法提取时记录为缺失，签名验证通过时记录凭证完整和密钥关联状态[39]。

多源证据的联合解释需满足三个条件。各项证据应指向同一内容版本，针对同一待验证命题，并具有有效的验证依据。被动检测、水印和内容凭证

分别反映统计风险、标识存在和来源声明。两项可得证据对同一命题给出不能同时成立的判断时，进入冲突核验。表 1 比较各类机制的功能与成立条件，表 3 归纳传播操作引起的状态变化，图 4 据此组织状态解释、冲突核验和复核优先级。

2　生成前主动防御与身份保护

2.1　生成前主动防御方法

本文将生成前主动防御限定为在伪造发生前或素材发布前，主动改变攻击者可利用的素材、身份表征或训练过程，以降低伪造成功率。其直接目标是阻断或削弱生成，而非为成品附加来源标识。只有防护配置、执行时间和保护状态被可信记录并与素材绑定时，相关信息才可形成审计线索。

按照作用对象，生成前主动防御可分为输入扰动、身份表征遮蔽和训练链路干预。Anti-Forgery[27]、DF-RAP[28]和 RDDA[29]均在输入侧构造扰动，但分别侧重感知隐蔽性、社交平台处理后的保持性以及对输入变换的泛化，NullSwap[30]则通过遮蔽身份表征，降低换脸模型对目标身份的利用。CMUA-Watermark[36]通过跨模型通用对抗扰动干扰换脸生成。

面向扩散模型，Salman 等[55]干扰图像编码与去噪过程，MetaCloak[56]利用元学习和变换采样增强个性化训练场景中的黑盒迁移性，但 Zhao 等[57]在包含网络图像变换和扩散净化的现实威胁模型下发现，多类保护扰动经净化后可能失效。Silencer[58]进一步通过音频控制无效化和抗净化约束，将肖像防护扩展到音频驱动的说话人视频生成，TSDF[59]则关注攻击者重训练后的防护持续性。上述方法的差异主要来自保护对象、攻击者能力和传播处理假设，不能依据单一模型上的生成失败率直接排序。

2.2　局限性与证据链定位

生成前主动防御的比较应同时考察防护有效性、原始内容可用性和现实攻击鲁棒性。前者包括伪造成功率或身份相似度下降，内容可用性关注感知质量和正常使用，现实鲁棒性则包括跨模型迁移、压缩转码后保持以及对净化、自适应攻击和重训练的持续性。现有方法普遍依赖防御者能够在发布前处理素材，其效果还受代理模型、平台处理和攻击者知识影响。生成前干预能够降低身份滥用风险，却不能独立证明传播后内容的来源和编辑历史，仍需与生成端标识和内容凭证衔接。

3　生成端水印与模型标识

3.1　水印要求与模态适配

数字水印通过在内容或生成过程中嵌入可检测信号，为标识核验提供主动信息。只有当标识与可信签发主体、密钥或来源登记可靠绑定时，水印才可进一步支持来源核验。评价维度包括不可感知性、鲁棒性、容量、安全性以及误检率和漏检率[60-62]。按证据形成方式，可分为内容侧水印、生成端水印、模型水印。不同模态的嵌入对象与失效条件不同，图像水印作用于空间域、频域或潜在空间，视频和音频水印还需适应帧间编码、压缩和常

表 1　AIGC 深度伪造防护、证据与验证机制比较

技术机制	功能目标	成立前提与信任基础	可支持的结论	攻击假设与失效边界
被动检测	评估伪造风险	模型、阈值和测试分布适配	概率、标签或热力图所示统计风险	未知模型、域偏移和传播处理可能导致误判
生成前防护	降低素材利用率或伪造成功率	发布前可处理素材，代理模型与实际威胁相近	给定条件下的防护效果及可信执行记录	净化、跨模型迁移、平台处理和重训练可能削弱防护
内容水印与输出标识	附加可检测标识	嵌入端配合，检测器或密钥可用	标识检出状态或载荷恢复结果	压缩、裁剪、再生成、改写和覆盖嵌入可能破坏标识
模型水印	核验模型所有权或指定模型输出	可控制模型或解码器，验证密钥可用	模型或输出与指定水印相关	微调、蒸馏、剪枝、模型替换和有损传播可能导致失效
模型指纹与来源归因	推断生成模型或来源类别	归因模型和参考数据覆盖候选来源	真实与生成分类或候选模型归因	未知模型、域偏移、压缩和再生成可能改变指纹
内容凭证与数字签名	核验来源声明、版本绑定和派生记录	密钥、证书、时间戳、信任锚及验证服务可靠	签名、绑定、信任和派生状态	剥离、重放、重绑定、密钥泄露和跨平台丢失
验证端协同	解释证据状态并安排复核	证据对应同一版本和待验证命题	有效、缺失、冲突或未知状态及优先级	错误绑定、共同失效和信任假设不一致可能导致误判

规编辑。文本输出水印在生成时调制词元（token）分布形成统计标识[6]，其检测受文本长度、改写和翻译影响。已有理论研究表明，在攻击者能够评价输出质量并实施质量保持扰动等明确假设下，强文本水印可能被移除，即使采用私有检测密钥也不能保证水印不可消除[63]。Tree-Rings[33]和 Stable Signature[34]分别在扩散模型初始噪声和解码阶段嵌入标识，StreamMark[35]、音色水印[64]和 AudioMarkNet[65]代表面向语音深度伪造的主动标识方法。

3.2 生成端水印与输出标识

生成端水印将标识嵌入与内容生成同步完成。Pluggable Watermarking[66]通过改造生成器解码器嵌入可提取水印，适用于平台可控模型。这类方法回答“输出是否携带指定标识”，依赖生成方配合，

并可能受改写、再生成和模型蒸馏影响，但其不能独立证明实际调用的模型版本、提示输入、工具操作和中间编辑过程，后者需结合生成日志、内容凭证或可验证推理机制。

3.3 模型水印、模型指纹与主动取证

模型水印在训练阶段主动嵌入标识，但可能在蒸馏和微调后减弱。模型指纹利用不同生成模型残留的统计特征进行被动来源归因[37]，FakeTagger[38]进一步面向传播过程提供来源追踪。二者与内容水印互补：当水印因传播处理失效时指纹可能残留，而再生成则可能覆盖指纹。近期主动取证水印进一步从单一水印提取转向检测、篡改定位与来源追踪的联合取证。DiffMark[67]、WaveGuard[68]和 KAD-Net[69]通过鲁棒嵌入增强提升水印可提取性，LIDMark[70]和 SegFaceMark[71]面向检测、定位与追踪一体化，MEA[73]则分析多次嵌入攻击对原始水印的削弱作用。相较于只验证标识是否存在，这些研究开始处理篡改位置、来源归属和标识冲突，但不同任务的评价指标尚未统一。同时，研究也推进生

表 2 AIGC 内容标识、主动取证与来源归因代表性方法比较

方法	对象/模态	介入位置与验证前提	可支持结论	已报告鲁棒性与边界	主要依赖
Tree-Rings[33]	图像	扩散初始噪声频域嵌入，验证时执行反演	检出指定生成标识	可抵抗裁剪、旋转和缩放，其他再生成条件待验证	扩散模型、反演过程和密钥
Stable Signature[34]	图像	微调潜在扩散解码器并提取二进制签名	识别带标识模型的输出	可抵抗压缩和大比例裁剪，模型修改攻击覆盖有限	解码器、签名和提取器
DiffMark[67]	人脸图像	条件扩散嵌入并训练水印提取器	支持真实性核验和来源追踪	覆盖典型人脸伪造，跨平台条件验证有限	扩散模型、提取器和训练攻击模型
FaceSigns[31]	人脸图像	发布前嵌入半脆弱签名	核验完整性并发现人脸伪造	对常规处理鲁棒，对测试的伪造操作敏感	秘密签名、嵌入器和提取器
WaveGuard[68]	人脸图像	小波高频域嵌入并利用图网络提取	支持来源追踪和伪造检测	覆盖换脸、表情重演和常规失真	编码器、提取器和图网络
KAD-Net[69]	人脸图像	水印嵌入及追踪、检测双分支验证	输出来源和局部伪造风险	覆盖常规失真及多类人脸操纵	水印和检测网络
LIDMark[70]	人脸图像	嵌入关键点与身份水印	支持检测、定位和来源追踪	可抵抗测试的失真和篡改，限于人脸内容	身份编码及水印编解码器
SegFaceMark[71]	人脸图像	分区嵌入并使用身份动态密钥	支持检测、追踪和区域解释	面向复杂人脸伪造，跨模态攻击未覆盖	身份特征、动态密钥和验证器
SSD[72]	人脸图像	将人脸自隐写嵌入原图并联合恢复检测	核验完整性和伪造风险	覆盖常规处理及 8 类人脸伪造	嵌入、恢复和检测网络
Pluggable Watermarking[66]	模型输出	在生成器解码器中插入水印模块	识别指定模型输出并辅助所有权核验	覆盖有损信道和自适应攻击	解码器、水印模块和 BCH 编码
StreamMark[35]	语音	频域嵌入半脆弱水印	区分正常处理与语音伪造	覆盖压缩、噪声、语音转换和编辑	音频编解码器和水印消息
SynthID-Text[6]	文本	生成采样阶段调整词元选择	提示文本可能来自带水印模型	覆盖编辑和改写，深度重写可能削弱检测	生成服务和统计检测器
GAN 指纹[37]	图像	学习模型特有统计指纹	判断生成内容并归因候选 GAN	覆盖多种 GAN，不能直接推广至未知扩散模型	归因模型和参考数据

成端水印从内容标识向多任务主动取证扩展。

为便于比较和查阅，表2进一步选取不同模态和机制下的代表性方法，比较其介入位置、验证前提、可支持结论及原文报告的鲁棒性边界。

3.4　水印攻击与再生成威胁

水印的可靠性取决于其所面临的威胁模型，常见的攻击方式包括压缩、裁剪、转码、加噪、滤波、去水印、覆盖水印和再生成。再生成攻击利用扩散模型或图像编辑模型对带水印内容重新渲染，可能在保持主要语义内容的同时显著削弱或移除像素级水印信号，尤其当水印与低层纹理、频域残差或局部像素绑定较强时，其鲁棒性更易受到影响[74]。覆盖水印和多次嵌入攻击[73]可能导致来源混淆和原始水印被削弱，需在训练阶段模拟多嵌入干扰。水印与被动检测之间也可能相互影响。水印信号可能被检测器学习为伪捷径[75]，水印嵌入也可能改变图像统计特征导致检测器误判，因此水印系统应与检测系统联合评估。数字水印作为信号级认证难以完整描述创建主体和编辑过程，需进一步与数字签名、可信时间戳和内容凭证衔接，形成语义更明确的溯源认证机制。

4　内容凭证与来源认证

4.1　内容凭证与来源声明

内容凭证旨在为数字内容附加可验证的来源声明和处理记录。生成式AI平台、拍摄设备和编辑工具可分别签发或追加凭证，用于声明AI来源、光学采集来源和编辑类型。从机制分析角度看，典型凭证链路通常包括内容哈希、凭证清单、来源声明或处理断言、数字签名、可信时间戳、证书链和信任锚等要素。内容哈希建立凭证与资产的硬绑定，数字签名将凭证清单与签发密钥关联，时间戳和证书链用于判断签名时点与签发者信任状态。三者属于不同验证维度，签名验证通过仅表明已签声明与相应密钥关联且签发后未被篡改，不等于签发者受信，也不等于声明语义真实[39]。

对于大模型生成，来源声明还应区分输出标识、过程记录与计算验证，过程记录用于绑定模型或服务版本、提示摘要、工具调用和派生关系。zkLLM等可验证推理机制可在不公开模型参数的条件下核验声明模型是否执行了相应计算[76]，但不能证明输出事实或语义真实。Chimera[77]进一步表明，即使采集设备生成有效硬件签名，也不能排除镜头前已经呈现伪造内容的场景。内容经历多次编辑时，凭证链应记录各版本之间的引用关系，平台转码、截图重传、换脸和拼接等操作可能造成哈希不一致、凭证链断裂或凭证与实际内容不一致。

C2PA规范[39]定义了内容凭证中的断言、声明、清单、签名绑定与验证流程，为跨平台来源声明、完整性校验和凭证互操作提供了标准化基础。Amazon Titan Image Generator部署内容凭证的实践说明，该机制已用于生成式AI平台的实际输出流程[78]，但这一案例不能代表跨平台保持机制和信任策略已经统一。

内容凭证从签发、传播到验证的基本链路及主要攻击面如图2所示。验证端应分项检查凭证可得性、清单结构、签名完整性、资产硬绑定、签发凭据与吊销状态以及派生关系。核验结果应保留具体状态，不能压缩为单一真假结论。图中将凭证伪造、重绑定、信任链攻击和跨平台认证失败分别映射到签发、绑定、信任与传播环节。

4.2　凭证攻击与跨平台认证失效

溯源认证的安全分析应区分传播处理导致的鲁棒性下降和攻击者主动破坏。保护对象包括内容与凭证的版本绑定、签发者身份、处理顺序和派生关系。攻击者可能编辑内容和元数据、调用生成模型、剥离或重放凭证、阻断或操纵外置清单检索，更强攻击还可能滥用签名服务、获取签名密钥或操纵信任列表。验证所需安全属性包括内容完整性、版本绑定性、签名完整性、签发者身份可验证性、时间有效性和凭证可获得性。其成立依赖密钥保护、证书与吊销服务可用、验证器正确实现以及时间戳和信任锚可靠。满足这些条件能够证明声明的签发及版本绑定关系，但不能证明拍摄场景或声明语义真实[39,77,79]。

C2PA 2.4《安全考虑》以非规范性方式归纳上述威胁和缓解措施，但不评价威胁发生概率或风险等级[79]。凭证重绑定是指合法凭证或声明被关联至另一资产或未记录的派生版本。硬绑定在内容变化后通常表现为哈希失配，软绑定用于发现转码版本或找回外置清单时，还需防范碰撞、相似内容误关联和清单检索被操纵。

信任链攻击针对证书路径、信任锚、信任列表、时间戳或吊销信息，密钥泄露和过期信任信息

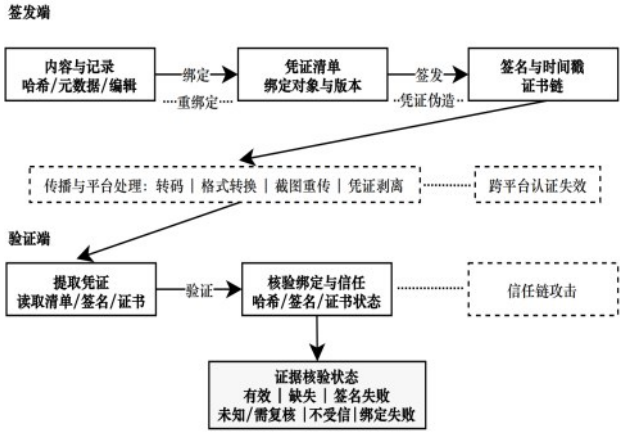

图2　内容凭证签发、传播与验证链路及主要攻击与失效面

既可能使恶意内容被错误接受，也可能使合法内容被拒绝[79]。跨平台认证失败既可能由恶意剥离引起，也可能由转码、格式转换、元数据清理、外置清单不可达、格式不支持或信任策略差异造成。验证端因此应分别报告凭证缺失、签名失败、绑定失配、签发者不受信、状态未知和派生关系不完整，验证报告应分别保留上述状态，供后续协同核验和人工复核使用。

5　传播扰动下的证据保持与失效

5.1　平台压缩与转码

压缩和转码是导致主动证据衰减的常见因素。社交平台和内容分发网络（content delivery network，CDN）通常会对上传内容进行重新压缩。JPEG压缩通过离散余弦变换（discrete cosine transform，DCT）量化和高频去除削弱频域水印和精细纹理扰动，H.264/H.265编码中的运动补偿和帧间预测可能破坏帧间水印同步，MP3、AAC和Opus等有损音频编码会去除掩蔽阈值以下的频率成分。DF-RAP[28]实验表明，在其设定的真实平台处理条件下，相关防护信号仍表现出一定保持能力，但不同平台的压缩、转码和重采样策略差异较大，鲁棒性评测仍需覆盖主流平台的实际处理链路。

5.2　截图重传、二次编辑与多模态不一致

截图重传会重新采样并重构像素分布，通常导致文件层元数据和嵌入式凭证丢失。裁剪可能切除关键标识区域，视频帧率变化和抽帧同时破坏帧内水印和帧间一致性。二次编辑（滤镜、美颜、配音和字幕等）进一步增加证据保持复杂性，半脆弱水印区分良性编辑与恶意篡改是核心难点[31]。多模态内容的图像、音频、文本和凭证可能经历不同处理路径，除证据可提取性分化外，还可能出现音画时序、主体身份或语义声明不一致。验证端应分别报告各模态证据状态并检查跨模态绑定关系，现有研究仍缺少统一基准。

5.3　再生成与去水印攻击

再生成攻击利用生成式AI模型从语义层面清洗主动证据，可能显著削弱或移除嵌入的像素级水印与扰动信号[74]。攻击者可利用扩散模型的图像到图像生成（image-to-image，Img2Img）能力，在保持语义内容基本不变的情况下重新生成像素，也可借助大语言模型改写或翻译，削弱文本水印的统计可检测性[63]。Liang等[80]对10类文本水印和12类代表性攻击的统一评测进一步表明，不同水印设计在对抗环境下的鲁棒性存在明显差异。语音转换则可能通过改变声学特征削弱音频标识。模型净化通过微调、蒸馏或剪枝改变模型输出分布，使模型级标识失效。与传统压缩、加噪等攻击不同，再生成和模型净化针对的是证据与内容语义的绑定关系。

5.4　证据失效模式比较

传播扰动既包括平台压缩、格式转换、截图转发和常规编辑等传播处理，也包括去水印、覆盖水印、再生成和凭证剥离等对抗性操作。前者通常不以破坏证据为直接目的，后者则主要针对证据嵌入位置或信任链，但操作动机与证据后果并非一一对应：常规处理也可能造成证据缺失，对抗性操作也可能仅导致局部衰减。因此，后续比较主要依据操作机制及其证据后果，而不单纯按实施意图区分[28,31,65,73-74]。多次嵌入或覆盖水印也可能削弱原始水印[73]。表3基于相关水印、主动防御、内容凭证和再生成攻击研究中的典型设置，将证据影响归纳为保持、衰减、缺失和冲突风险4类状态后果，用于比较不同传播操作对多类证据的影响。表中结论为机制层面的定性归纳，不表示同一数据集和统一平台链路下的量化排序。

截图重传通常使嵌入式凭证和元数据缺失，但部分鲁棒水印和检测特征仍可能保留。再生成会重新构造像素和生成痕迹，也可能同时削弱水印、改变检测特征并破坏凭证与当前版本的绑定。多源证据能否形成补偿，取决于其产生机制是否相对独立、是否与当前版本正确绑定，以及至少一类证据

是否仍然有效。压缩转码主要削弱水印和检测特征，但对内容凭证的影响取决于平台是否保留、迁移或重新签发凭证，凭证剥离对水印和检测特征影响较小，却会破坏来源链完整性。图3进一步概括了典型传播扰动对被动检测、数字水印和内容凭证的影响程度，以及三类证据之间的可补偿关系。图中符号表示基于文献报告和机制分析形成的定性判断，不表示同一数据集、同一平台链路下的量化实验结果。该结果说明，主动防御与溯源认证的评测应从单点鲁棒性测试扩展为传播链路级证据保持评测，即同时考察水印、凭证和被动检测在上传、转码、截图、再上传等真实链路中的保持率、冲突率和互补效果。

6　被动检测与主动证据的协同验证

6.1　被动检测与主动证据互补关系

当内容缺少水印和凭证，或主动证据已在传播过程中失效时，被动检测仍是对已传播内容进行独立审查的重要工具。已有文献[40-41,44]较充分地梳理检测模型本身，本文更关注被动检测在证据链中的功能定位。被动检测提供统计风险信息，数字水印提供主动标识，模型指纹提供被动来源归因线索，内容凭证提供来源声明及版本关系。三类证据的可信含义不同，不能直接相互替代，但可在验证端互为补充。

从介入时机看，生成前防护主要在素材被滥用之前降低攻击成功率，传播后的验证端通常只能间

表3　典型传播操作下多类证据状态变化与失效模式

传播扰动	对被动检测的影响	对水印的影响	对内容凭证的影响	状态后果与处置
JPEG压缩（轻/重）	轻压缩影响较小，重压缩可能显著减弱	频域提取率下降，空间域耐轻压缩	影响取决于平台是否保留、迁移或重新签发凭证	**（保持/衰减/绑定失配）**重压缩致哈希不一致或凭证验证失败，开展多因素鲁棒性测试并结合多源证据
视频转码	帧间线索改变，频域伪影受量化影响	帧间同步可能破坏	凭证元数据轨可能被保留、迁移或清理	**（衰减/缺失/绑定失配）**码流重编致哈希不一致或凭证验证失败，测试码率档次并结合关键帧水印
截图重传	像素统计重构，文件线索丢失	鲁棒水印部分保持	嵌入式凭证通常丢失，外置凭证取决于绑定方式	**（衰减/缺失）**溯源链中断，依赖内容水印和被动检测
裁剪/二次编辑	局部特征改变，全局不稳定	半脆弱水印可能失效	原始凭证哈希失效	**（衰减/绑定失配）**检查编辑链连续性，结合半脆弱水印与凭证链比对
扩散模型再生成	伪影和指纹可能被覆盖或改变	多数像素级水印可能失效	原凭证通常无法直接绑定当前派生内容	**（衰减/缺失/绑定失配）**提高核验优先级，并结合被动检测和人工复核
凭证剥离	对内容特征影响较小	对内容水印影响较小	凭证缺失	**（缺失）**水印和被动检测补充
覆盖水印/多次嵌入	可能引入新特征致误判	原始水印被削弱或覆盖	凭证链本身可保持，若未重新签发派生凭证，内容哈希可能不一致	**（衰减/冲突风险）**来源标识冲突，检查嵌入顺序和多次嵌入痕迹

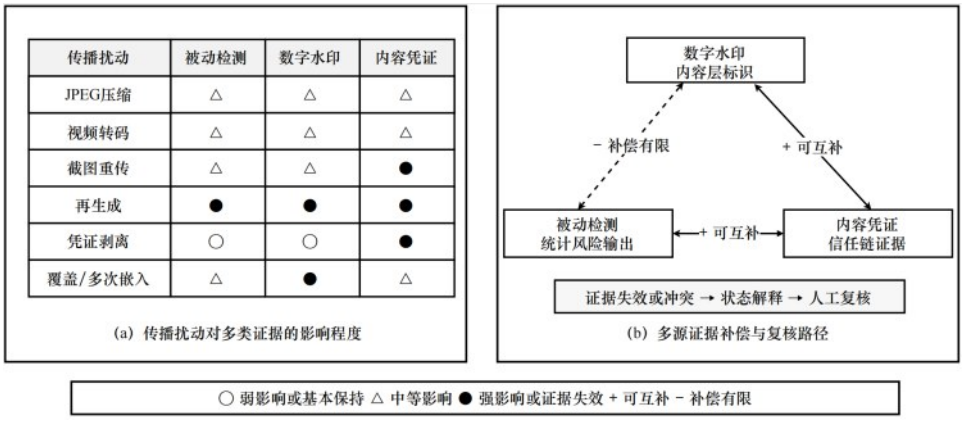

图3　典型传播扰动下多源证据失效与补偿关系

接观察其保护结果。水印在生成或发布阶段主动嵌入，模型指纹由生成过程留下并在验证阶段提取。内容凭证在发布或编辑阶段签发，被动检测在传播后独立运行。水印和凭证可为传播后的风险判断提供独立验证依据，被动检测则可在主动证据缺失或失效时提供补充信息。若三类证据具有相对独立的产生、绑定和验证机制，联合核验可能提高同时规避的难度，但一次再生成也可能同时破坏水印、凭证绑定和检测特征，因此证据数量增加并不必然提高安全性。

6.2　证据缺失、冲突与风险分级

依照证据状态判据，证据缺失仅表示相关证据不可提取或不可访问，不应直接等同于内容伪造。内容可能因未接入凭证体系、传播平台清理元数据、截图重传或来自不生成水印的渠道而缺少主动证据。协同验证系统应将证据缺失记录为一种可解释状态，并在授权、合规和最小必要原则下保留缺失原因、传播路径、上传主体和平台处理信息。从审核记录角度看，证据缺失更适合作为提高核验优先级的信号，而不是作为自动判定依据。

除证据缺失外，多源证据在联合核验中还可能呈现不一致。检测结果与水印或凭证状态不一致时，不能仅根据输出差异直接认定证据冲突，而应先判断各项证据是否针对同一内容版本和同一待验证命题。如检测结果为高风险而水印可检出或凭证签名有效，可能源于检测误报、凭证签发后的局部编辑或证据覆盖范围不同。检测结果为低风险而主动证据缺失，则只能说明现有证据不足。只有水印来源、凭证声明或编辑记录针对同一对象给出不能同时成立的结论时，才构成实质冲突。此时需结合凭证链、签名时间、编辑历史和水印嵌入顺序，区分合法转载、平台处理与主动攻击造成的不一致。

基于上述证据状态及其不一致程度，可将核验优先级划分为低、中、高三类。低优先级内容具有完整主动证据、连续传播链路和一致来源声明，可采用简化核验；中优先级内容存在部分证据缺失或轻微不一致，应补充上下文和传播链路验证；高优先级内容表现为凭证链断裂、关键证据冲突、高影响主体或跨模态不一致，应进入完整验证和人工复核流程。该分级用于配置核验资源和复核深度。

6.3　多源证据协同验证流程

多源证据协同验证首先提取被动检测结果、水印状态、凭证状态和传播记录，再根据内容版本、待验证命题、绑定关系和时效状态解释各项证据。图4将验证过程归纳为快速筛查、证据核验、证据状态识别、核验优先级安排和人工复核5个环节。

审核首先通过快速筛查发现疑似内容，再核验水印、凭证签名链、证书信任及资产绑定关系。发现证据不一致后，根据不一致程度及冲突强度确定复核深度，高影响主体和强冲突样本交由人工复核。该流程保留证据状态及失效原因，不把多源结果压缩为单一真假标签。

7　评测体系与研究方向

7.1　威胁模型与评测资源

威胁模型是评价主动防御与溯源认证技术有效性的前提。相较于仅测试单一模型在给定压缩参数下的准确率，证据链评测还需明确攻击者知识、

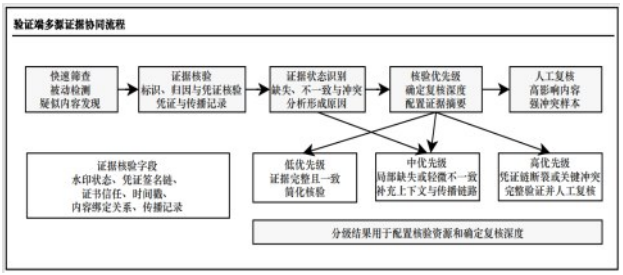

图4　多源证据协同验证流程与核验优先级安排

操作能力、攻击目标和传播链路。攻击者知识可分为白盒、灰盒和黑盒，操作能力可按内容编辑权限、生成模型访问权限、模型训练能力以及密钥或平台服务控制权限分级，攻击目标则包括质量保持优先、隐匿性优先、去证据优先和诬陷优先。只有明确上述条件，水印保持率、凭证完整率和检测准确率才具有可比较意义。评测还应固定签名密钥状态、信任列表和吊销服务可用性、验证服务可信性以及平台凭证迁移规则等信任假设。

现有评测资源在不同方向上的成熟度并不均衡。被动检测已形成覆盖跨生成器、真实平台压缩和多模态内容的基准体系[13,26,81-87]，Omni-Fake[88]进一步统一图像、音频、视频和音视频说话人内容，并设置检测、定位与解释任务。图像水印方向也出现了覆盖传统失真、扩散和对抗攻击的WAVES基准[89]，文本水印则可借助Liang等[80]的统一评测比较多类水印与移除攻击。相比之下，生成前防护和内容凭证仍多依赖作者自建协议，现有

基准也尚未统一传播链路、跨证据冲突和人工复核任务。

评测资源应同时覆盖内容、传播链路和复核结果。内容数据需包含图像、视频、语音、文本及其多模态组合，传播链路数据应记录压缩、转码、截图、裁剪、再上传和再生成过程，复核数据则用于标注证据缺失、冲突、风险等级和处置结果。现有基准主要评价单项检测或单项水印在给定扰动下的性能，尚难覆盖“证据生成—传播处理—证据缺失/冲突—人工复核”的链路级验证过程。缺少第三层数据时，协同验证只能评估单项证据能否被提取，难以评估证据冲突是否被正确解释。

7.2　评价指标与动态攻防

主动防御与溯源认证评测可分为单项证据、链路保持和决策效用3个层面。单项证据指标包括水印检测率、误检率、漏检率、比特错误率及凭证签名和证书链验证成功率。链路保持指标依据第1.4节的状态判据，记录证据经过压缩、转码、截图、二次编辑和再生成后的保持、衰减、缺失与冲突比例。决策效用则关注复核排序、误处置成本和证据摘要可理解性。由于目前缺少统一标注协议，这些内容更适合作为待建立的评价维度，而非已经成熟的量化指标。

现有评测多采用静态单点测试，即在预定义压缩参数、噪声类型或裁剪比例下测量性能。真实传播环境更接近动态链路，不同平台使用不同转码策略，同一平台策略也会随时间调整，攻击者还可能根据公开防护机制自适应修改攻击方式。因此，评测体系有必要由静态单点测试扩展至动态攻防场景。动态评测应先复现“上传—转码—分发—截图—再上传”等传播链路，记录各阶段的证据衰减。随后在灰盒条件下实施针对水印、凭证和检测特征的自适应攻击，并通过多轮攻防检验防御更新后的迁移性。人工复核实验则用于判断证据摘要和冲突报告能否支持专家决策。

7.3　存在的问题及研究方向

图5依据相关文献归纳了2017年以来生成范式、防护检测、主动证据和认证部署的演进脉络。GAN与StyleGAN阶段推动了人脸伪造检测，扩散模型和多模态生成提高了内容逼真度，再生成和跨平台传播则进一步放大证据保持问题。

7.3.1　跨平台证据保持与多模态一致性

现有主动证据多针对单一模型、单一平台或单一模态设计，而真实内容通常跨平台传播并经历多种处理。平台转码、截图重传和异步编辑不仅会削弱证据，还会造成不同证据类型之间的差异化衰减。如水印可能在转码后仍可提取，但文件层凭证已因格式转换而丢失。被动检测可能给出低风险结果，但凭证链已断裂。后续研究应建立跨平台证据保持评测集，记录不同平台处理链路中的水印、凭证、哈希、元数据和检测响应变化，并按证据类型报告保持情况。

深度伪造生成与防护技术演进脉络

维度	2017-2019	2020-2021	2022-2023	2024-2026	发展方向
生成范式	GAN人脸合成 属性编辑	视频换脸发展	扩散模型 跨模态生成	多模态大模型 实时音视频伪造	可控生成 再生成攻击
防护与检测	手工特征 频域伪影	深度检测器 域泛化	基础模型适配 多模态检测	主动取证 跨平台验证	动态攻防评测
主动证据	初始水印 素材扰动	模型指纹 平台水印	扩散水印 内容凭证	主动取证水印 多源证据补偿	证据冲突推理
认证部署	传统签名 平台内声明	平台侧声明	C2PA凭证 签名链	生成平台接入 分级验证	接口互操作 分级核验

图5　深度伪造生成与防护技术演进脉络示意

7.3.2　证据冲突推理与动态评测

被动检测、水印和凭证因介入时机、作用机制和适用条件不同，在复杂传播中可能出现证据缺失、不一致或冲突。再生成及水印移除研究[74]与多次嵌入攻击研究[73]表明，攻击者可在保持语义或覆盖来源标识的同时破坏原有证据。证据冲突推理的关键问题，不只是判断某项证据是否存在，还要解释证据为何缺失、冲突来自平台处理还是主动攻击，以及哪些内容需要更高复核优先级。动态评测应覆盖攻击者迭代适应、防护策略更新和平台处理策略变化，避免仅在固定扰动参数下评价单点鲁棒性。

7.3.3　标准化接口与分级验证

主动水印、内容凭证和被动检测采用不同的嵌入方式、验证接口和证据描述格式，第三方验证端因而难以完成联合核验。C2PA规范[39]已为内容凭证的数据结构、签名绑定和验证流程提供标准化基础，但水印检测结果、被动检测置信度、跨证据失效原因和多源证据摘要仍缺少统一表达。面向智能体生成内容，过程记录还需关联模型调用、提示、工具操作、中间步骤与派生输出，并明确这些记录与最终内容之间的绑定关系。

在部署层面，分级验证有助于平衡准确性、成本和吞吐量。其阈值应综合内容影响、证据冲突强度和人工复核预算确定，并保留可审计的升级理由。分级结果用于将复核资源配置至高影响、强冲突和自动系统难以解释的样本，并保留可审计的升级依据。

8 结束语

在本文讨论的传播与威胁条件下，深度伪造核验不宜依赖单一证据。被动检测、水印、模型指纹和内容凭证分别支持统计风险判断、标识核验、来源归因和来源声明，其有效性取决于内容版本绑定、传播保持和相应信任条件。本文围绕内容生命周期梳理了生成前防护、生成端标识、内容凭证、传播保持和验证端协同研究，并从功能目标、成立条件、可支持结论和失效边界比较相关方法。后续研究应重点解决跨平台证据保持、多模态一致性、冲突解释和动态攻防评测问题，为AIGC治理中的深度伪造核验提供技术支撑。

参考文献：

[1] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [C]//Advances in Neural Information Processing Systems. 2020, 33: 6840-6851.

[2] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10684-10695.

[3] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]//Advances in Neural Information Processing Systems. 2021, 34: 8780-8794.

[4] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8110-8119.

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

[6] DATHATHRI S, SEE A, GHAISAS S, et al. Scalable watermarking for identifying large language model outputs[J]. Nature, 2024, 634(8035): 818-823.

[7] KORSHUNOV P, MARCEL S. Deepfakes: a new threat to face recognition? Assessment and detection[J]. arXiv preprint arXiv: 1812.08685, 2018.

[8] CROITORU F A, HIJI A I, HONDRU V, et al. Deepfake media generation and detection in the generative ai era: A survey and outlook[J]. arXiv preprint arXiv:2411.19537, 2024.

[9] LIU P, TAO Q, ZHOU J T. Evolving from single-modal to multi-modal facial deepfake detection: A survey[J]. arXiv preprint arXiv: 2406.06965, 2024, 1.

[10] DURALL R, KEUPER M, KEUPER J. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7890-7899.

[11] LI Y, CHANG M C, LYU S. In ictu oculi: Exposing AI created fake videos by detecting eye blinking[C]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018: 1-7.

[12] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019: 83-92.

[13] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: Learning to detect manipulated facial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1-11.

[14] LI L, BAO J, ZHANG T, et al. Face X-Ray for more general face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5001-5010.

[15] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.

[16] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 86-103.

[17] LIU H, LI X, ZHOU W, et al. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 772-781.

[18] GU Z, CHEN Y, YAO T, et al. Spatiotemporal inconsistency learning for deepfake video detection[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 3473-3481.

[19] LV Q, LI Y, DONG J, et al. DomainForensics: Exposing face forgery across domains via bi-directional adaptation[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 7275-7289.

[20] CUI X, LI Y, LUO A, et al. Forensics adapter: Adapting CLIP for generalizable face forgery detection[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 19207-19217.

[21] PENG C, YAN T, LIU D, et al. Face forgery detection with CLIP-enhanced multi-encoder distillation[J]. IEEE Transactions on Image Processing, 2025, 34: 8474-8484.

[22] LI X, XU J, CHEN J. VIGIL: Part-grounded structured reasoning for generalizable deepfake detection[J]. arXiv preprint arXiv:2603.21526, 2026.

[23] CORVI R, COZZOLINO D, ZINGARINI G, et al. On the detection of synthetic images generated by diffusion models[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[24] HOI M, ZHENG Z, LIU P, et al. SEED: A Large-Scale Benchmark for Provenance Tracing in Sequential Deepfake Facial Edits[J]. arXiv preprint arXiv:2604.10522, 2026.

[25] MONTIBELLER A, SHULLANI D, BARACCHI D, et al. Bridging the Gap: A Framework for Real-World Video Deepfake Detection via Social Network Compression Emulation[C]//Proceedings of the 1st on

Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media. 2025: 29-36.

[26] YAN Z, YAO T, CHEN S, et al. Df40: Toward next-generation deepfake detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 29387-29434.

[27] WANG R, HUANG Z, CHEN Z, et al. Anti-Forgery: Towards a stealthy and robust DeepFake disruption attack via adversarial perceptual-aware perturbations[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. 2022: 761-767.

[28] QU Z, XI Z, LU W, et al. DF-RAP: A robust adversarial perturbation for defending against deepfakes in real-world social network scenarios [J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 3943-3957.

[29] GUO Q, PANG S, CHEN Z, et al. Towards robust DeepFake distortion attack via adversarial autoaugment[J]. Neurocomputing, 2025, 617: 129011.

[30] WANG T, NIU S, CHENG H, et al. NullSwap: Proactive identity cloaking against deepfake face swapping[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 9945-9954.

[31] NEEKHARA P, HUSSAIN S, ZHANG X, et al. FaceSigns: Semi-fragile neural watermarks for media authentication and countering deepfakes[J]. arXiv preprint arXiv:2204.01960, 2022.

[32] ZHAO Y, LIU B, ZHU T, et al. Proactive image manipulation detection via deep semi-fragile watermark[J]. Neurocomputing, 2024, 585: 127593.

[33] WEN Y, KIRCHENBAUER J, GEIPING J, et al. Tree-Rings watermarks: Invisible fingerprints for diffusion images[C]//Advances in Neural Information Processing Systems. 2023, 36: 58047-58063.

[34] FERNANDEZ P, COUAIRON G, JÉGOU H, et al. The Stable Signature: Rooting watermarks in latent diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 22466-22477.

[35] LIU Z, CERNAK M. StreamMark: A deep learning-based semi-fragile audio watermarking for proactive deepfake detection[C]//ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026: 14107-14111.

[36] HUANG H, WANG Y, CHEN Z, et al. CMUA-Watermark: A cross-model universal adversarial watermark for combating deepfakes[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(1): 989-997.

[37] YU N, DAVIS L S, FRITZ M. Attributing fake images to GANs: Learning and analyzing GAN fingerprints[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7556-7566.

[38] WANG R, JUEFEI-XU F, LUO M, et al. FakeTagger: Robust safeguards against deepfake dissemination via provenance tracking[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 3546-3555.

[39] Coalition for Content Provenance and Authenticity. Content Credentials: C2PA technical specification 2.4[S/OL]. (2026-04) [2026-06-29]. https://spec. c2pa. org/specifications/specifications/2.4/specs/C2PA_Specification.html.

[40] 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(2): 295-317.
DING F, KUANG R S, ZHOU Y, et al. A survey of Deepfake and related digital forensics[J]. Journal of Image and Graphics, 2024, 29(2): 295-317.

[41] 暴雨轩, 芦天亮, 杜彦辉. 深度伪造视频检测技术综述[J]. 计算机科学, 2020, 47(9): 283-292.
BAO Y X, LU T L, DU Y H, et al. Overview of Deepfake Video Detection Technology[J]. Computer Science, 2020, 47(9): 283-292.

[42] 瞿左珉, 殷琪林, 盛紫琦, 等. 人脸深度伪造主动防御技术综述[J]. 中国图象图形学报, 2024, 29(2): 318-342.
QU Z M, YIN Q L, SHENG Z Q, et al. Overview of Deepfake proactive defense techniques[J]. Journal of Image and Graphics, 2024, 29 (2): 318-342.

[43] 刘安安, 苏育挺, 王岚君, 等. AIGC 视觉内容生成与溯源研究进展[J]. 中国图象图形学报, 2024, 29(6): 1535-1554.
LIU A A, SU Y T, WANG L J, et al. Review on the progress of the AIGC visual content generation and traceability[J]. Journal of Image and Graphics, 2024, 29(6): 1535-1554.

[44] 李卫斌, 冯雨婷, 侯彪, 等. 深度伪造人脸检测技术发展与应用综述[J]. 中国图象图形学报, 2026, 31(6): 2260-2278.
LI W B, FENG Y T, HOU B, et al. Review of development and application of deepfake face detection technology[J]. Journal of Image and Graphics, 2026, 31(6): 2260-2278.

[45] SHIOHARA K, YAMASAKI T. Detecting Deepfakes with Self-Blended Images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18720-18729.

[46] CAO J, MA C, YAO T, et al. End-to-end reconstruction-classification learning for face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4113-4122.

[47] YAN Z, ZHANG Y, FAN Y, et al. UCF: Uncovering common features for generalizable deepfake detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 22412-22423.

[48] HALIASSOS A, MIRA R, PETRIDIS S, et al. Leveraging real talking faces via self-supervision for robust forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14950-14962.

[49] FANG Z, ZHAO H, WEI T, et al. Uniforensics: Face forgery detection via general facial representation[J]. IEEE Transactions on Dependable and Secure Computing, 2025.

[50] PEEBLES W S, XIE S. Scalable diffusion models with transformers [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 4195-4205.

[51] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[C]//Advances in Neural Information Processing Systems. 2022, 35: 36479-36494.

[52] WANG J, WU Z, OUYANG W, et al. M2TR: Multi-modal multi-scale transformers for deepfake detection[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval. 2022: 615-623.

[53] YANG W, ZHOU X, CHEN Z, et al. AVoiD-DF: Audio-visual joint learning for detecting deepfake[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2015-2029.

[54] TSIGOS K, APOSTOLIDIS E, BAXEVANAKIS S, et al. Towards quantitative evaluation of explainable AI methods for deepfake detection [C]//Proceedings of the 3rd ACM International Workshop on Mul-

timedia AI against Disinformation. 2024: 37-45.

[55] SALMAN H, KHADDAJ A, LECLERC G, et al. Raising the cost of malicious AI-powered image editing[C]//Proceedings of the 40th International Conference on Machine Learning. 2023, 202: 29894-29918.

[56] LIU Y, FAN C, DAI Y, et al. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 24219-24228.

[57] ZHAO Z, DUAN J, XU K, et al. Can protective perturbation safeguard personal data from being exploited by stable diffusion?[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 24398-24407.

[58] GAN Y, MIAO J, WANG Y, et al. Silence is golden: Leveraging adversarial examples to nullify audio control in ldm-based talking-head generation[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2025: 13434-13444.

[59] ZHENG H, LI Y, WANG L, et al. Boosting active defense persistence: A two-stage defense framework combining interruption and poisoning against deepfake[J]. IEEE Transactions on Information Forensics and Security, 2026.

[60] WANG Z, BYRNES O, WANG H, et al. Data hiding with deep learning: A survey unifying digital watermarking and steganography[J]. IEEE Transactions on Computational Social Systems, 2023, 10(6): 2985-2999.

[61] 刘瑞祯,谭铁牛. 数字图像水印研究综述[J]. 通信学报, 2000,(8): 39-48.
LIU R Z, TAN T N. Survey of watermarking for digital images[J]. Journal on Communications, 2000,(8):39-48.

[62] 李伟,袁一群,李晓强,等. 数字音频水印技术综述[J]. 通信学报, 2005,(2):100-111.
LI W, YUAN Y Q, LI X Q, et al. Overview of digital audio watermarking[J]. Journal on Communications, 2005,(2):100-111.

[63] ZHANG H, EDELMAN B L, FRANCATI D, et al. Watermarks in the sand: impossibility of strong watermarking for language models[C]//Proceedings of the 41st International Conference on Machine Learning. 2024, 235: 58851-58880.

[64] LIU C, ZHANG J, ZHANG T, et al. Detecting voice cloning attacks via timbre watermarking[J]. arXiv preprint arXiv:2312.03410, 2023.

[65] ZONG W, CHOW Y W, SUSILO W, et al. AudioMarkNet: Audio watermarking for deepfake speech detection[C]//34th USENIX Security Symposium (USENIX Security 25). 2025: 4663-4682.

[66] BAO H, ZHANG X, WANG Q, et al. Pluggable watermarking of deepfake models for deepfake detection[C]//AI for Critical Infrastructure Workshop@ IJCAI-24. 2024.

[67] SUN C, SUN H, GUO Z, et al. DiffMark: Diffusion-based robust watermark against deepfakes[J]. Information Fusion, 2025: 103801.

[68] HE Z, GUO Z, WANG L, et al. WaveGuard: Robust deepfake detection and source tracing via dual-tree complex wavelet and graph neural networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025.

[69] HE S, DIAO Y, LI Y, et al. KAD-Net: Kolmogorov-Arnold and Differential-Aware Networks for Robust and Sensitive Proactive Deepfake Forensics[J]. Knowledge-Based Systems, 2025: 114692.

[70] WU J, WANG L, GUO Z. All in one: Unifying deepfake detection, tampering localization, and source tracing with a robust landmark-identity watermark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026: 14106-14115.

[71] HE Z, GUO Z, WANG L, et al. One-identity-one-key: Region-aware watermark for proactive deepfake detection and robust source tracing [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2026.

[72] XIA R, ZHOU D, YUAN L, et al. SSD: Making face forgery clues evident again with self-steganographic detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

[73] JIA L, SUN H, GUO Z, et al. Uncovering and mitigating destructive multi-embedding attacks in deepfake proactive forensics[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2026, 40(1): 471-479.

[74] ZHAO X, ZHANG K, SU Z, et al. Invisible image watermarks are provably removable using generative ai[J]. Advances in neural information processing systems, 2024, 37: 8643-8672.

[75] WU X, LIAO X, OU B, et al. Are watermarks bugs for deepfake detectors? rethinking proactive forensics[J]. arXiv preprint arXiv: 2404.17867, 2024.

[76] SUN H, LI J, ZHANG H. zkLLM: zero knowledge proofs for large language models[C]//Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2024: 4405-4419.

[77] PARK S, VILESOV A, ZHANG J, et al. Chimera: Creating digitally signed fake photos by fooling image recapture and deepfake detectors [C]//34th USENIX Security Symposium (USENIX Security 25). 2025: 4305-4324.

[78] Amazon Web Services. Announcing Content Credentials for Amazon Titan Image Generator[EB/OL]. (2024-09-20) [2026-06-29]. https://aws. amazon. com/about-aws/whats-new/2024/09/content-credentials-amazon-titan-image-generator/.

[79] Coalition for Content Provenance and Authenticity. C2security considerationsPA, version 2.4[EB/OL]. (2026-08) [2026-08-15]. https://spec.c2pa. org/specifications/specifications/2.4/security/Security_Considerations.html.

[80] LIANG J, WANG Z, HONG S, et al. Watermark under Fire: A Robustness Evaluation of LLM Watermarking[C]//EMNLP (Findings). 2025: 21050-21074.

[81] LI Y, YANG X, SUN P, et al. Celeb-DF: A large-scale challenging dataset for DeepFake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3207-3216.

[82] CHANDRA N A, LEE H, MURTFELDT R, et al. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026: 10668-10678.

[83] JIANG L, LI R, WU W, et al. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2889-2898.

[84] HE Y, GAN B, CHEN S, et al. ForgeryNet: A versatile benchmark for comprehensive forgery analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4360-4369.

[85] ZI B, CHANG M, CHEN J, et al. WildDeepfake: A challenging real-world dataset for deepfake detection[C]//Proceedings of the 28th ACM

International Conference on Multimedia. 2020: 2382-2390.

[86] ZHU M, CHEN H, YAN Q, et al. Genimage: A million-scale benchmark for detecting ai-generated image[J]. Advances in neural information processing systems, 2023, 36: 77771-77782.

[87] CAI Z, GHOSH S, ADATIA A P, et al. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset[C]//Proceedings of the 32nd ACM international conference on multimedia. 2024: 7414-7423.

[88] LI T, HUANG Z, WEN H, et al. OMNI-fake: benchmarking unified multimodal social media deepfake detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2026: 30299-30311.

[89] AN B, DING M, RABBANI T, et al. WAVES: Benchmarking the robustness of image watermarks[C]//Proceedings of the 41st International Conference on Machine Learning. 2024, 235: 1456-1492.

韩禹洋（1995− ），男，陕西西安人，北京电子科技学院博士研究生，主要研究方向为人工智能安全、深度伪造检测和 AI 安全。

王志强（1985− ），男，安徽宿州人，博士，北京电子科技学院副教授，硕士生导师，主要研究方向为人工智能安全、漏洞发现、恶意软件检测和 AI 安全。

陈颖（1977− ），男，博士，北京电子科技学院教授，硕士生导师，主要研究方向为数据挖掘、人工智能和图像处理。

耿海文（1963− ），男，黑龙江海伦人，博士，北京电子科技学院教授，主要研究方向为密码编码、算法应用及密码技术应用。