区块链发展与应用

区块链驱动的 AIGC 图片

隐式标识完整性校验与传播溯源方法研究

学号：23121677

姓名：范舒舰

学院：计算机工程与科学学院

专业：网络空间安全

2026 年 5 月

目　录

1 引言

生成式人工智能图片已经成为网络传播中的常见内容。目前 ChatGPT 的图像生成与编辑能力已经能够达到以假乱真的地步，这类图片如果缺少明确的来源说明，就容易引发内容真实属性不清、生成责任边界模糊以及后续传播追溯困难等问题[4, 6]。

图 1: 怀旧复古风 AIGC 图片

问题在于，图片的生成与图片的传播并不是同一个阶段。平台在生成时可以记录来源声明、元数据、隐式标识或文件摘要，但图片一旦经历压缩、缩放裁剪或截图重保存，原始文件结构和附属信息都可能变化。这样一来，平台虽然保存了来源记录，传播后的图片却未必还能被重新关联回最初的登记对象。

本文关注的正是这一问题。研究对象被限定为已生成、已标识、已登记的 AIGC 图片，讨论当原始文件发生变化后，图片是否仍能够依靠隐式标识、感知哈希和区块链登记记录，被重新关联回原始来源对象。

2 问题分析与总体方案

AIGC 图片在传播场景中面临的关键困难，并不在于生成阶段缺少来源信息，而在于这些信息在后续传播中很难稳定保留[4]。原始平台通常可以为图片附加来源声明、文件摘要、隐式标识或其他描述字段，但图片一旦进入下载、转发、压缩、缩放、裁剪和截图重保存等过程，文件结构与附属信息都可能发生变化。接收端看到的虽然仍然是同一内容的传播版本，却未必还能与最初的登记对象保持直接对应关系。

这种断裂最直接地体现在精确绑定关系上。以 SHA-256 为代表的精确哈希适合确认原始文件是否保持不变，只要图片在传播中发生字节级变化，哈希值就会完全不同，原先建立的文件级对应关系也会立即失效。除了精确哈希之外，依赖文件结构存在的元数据、嵌入式来源说明和外部引用同样会受到传播过程影响。它们在生成阶段能够提供较完整的背景信息，但并不保证在跨平台传播后仍被稳定保留。因此，原图拥有来源记录，并不等于传播后的图片仍然能够被继续验证。

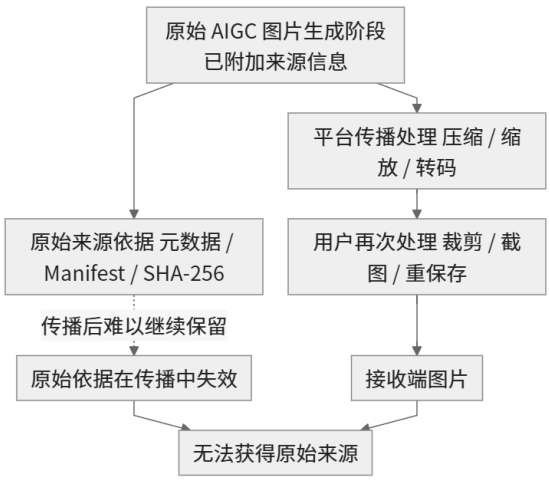

图 2: AIGC 图片在传播中的来源证明断裂示意

单独依赖一种机制，很难覆盖这一传播断裂场景。仅依赖元数据或来源声明，只能处理原始文件仍被完整保留的情况；仅依赖隐式标识，虽然能够在内容内部保留线索，但它本身并不是完整来源对象；仅依赖相似性判断，也只能说明当前图片与某条记录在视觉上接近，难以直接构成正式来源证明。传播后的来源验证更适合由多层机制共同完成，而不是由单点技术替代。

围绕这一问题，整体方案被组织为四个彼此衔接的层次。隐式标识层负责在图片内部保留不可见来源线索，使内容在传播后仍可能携带可恢复的标识；感知哈希层负责在精确哈希失效后，根据视觉结构完成近似找回；来源对象层负责组织平台、时间、摘要和标识等关键信息，使候选结果能够对应到可描述的记录；区块链登记层负责保存这些关键字段，并将其转化为可查询、可追溯的正式登记对象[1]。

图 3: 传播后来源验证的总体方案

这一结构并不是简单叠加。隐式标识解决的是“线索是否仍在”，感知哈希解决的是“传播后是否还能找回”，来源对象解决的是“找回之后记录如何表达”，区块链登记层解决的是“这些记录如何成为外部可验证依据”。在这一链路中，区块链并不承担图像处理功能，也不负责判断图片之间是否相似；它所提供的是来源记录的登记能力和后续验证时的正式依据。

围绕这一链路，后续讨论的重点落在两个位置。感知哈希决定传播后的图片能否重新回到原始记录附近，直接影响近似找回的有效范围；区块链登记层决定找回结果能否进一步对应到正式来源对象，并形成可验证、可审计的记录。

3 传播后近似找回机制：pHash

传播后的图片通常已经不再保持原始文件形态，因此无法继续通过 SHA-256 与链上 `contentId` 直接匹配。JPEG 压缩、缩放和裁剪都会改变文件字节，精确哈希会随之失效。传播后的来源验证因此需要一种新的匹配方式，用于判断当前图片是否仍然与原始登记对象保持同源关系。

pHash 采用的是视觉结构比较，而不是文件字节比较。处理时先将图片灰度化并统一缩放，再通过离散余弦变换提取低频特征，对低频块进行均值二值化，最终得到定长哈希值[5]。两张图片之间的差异由汉明距离表示。距离越小，说明视觉结构越接近；距离越大，说明图像内容差异越明显。基于这一特点，pHash 更适合处理压缩、缩放等轻中度传播失真，而对大幅裁剪更敏感。

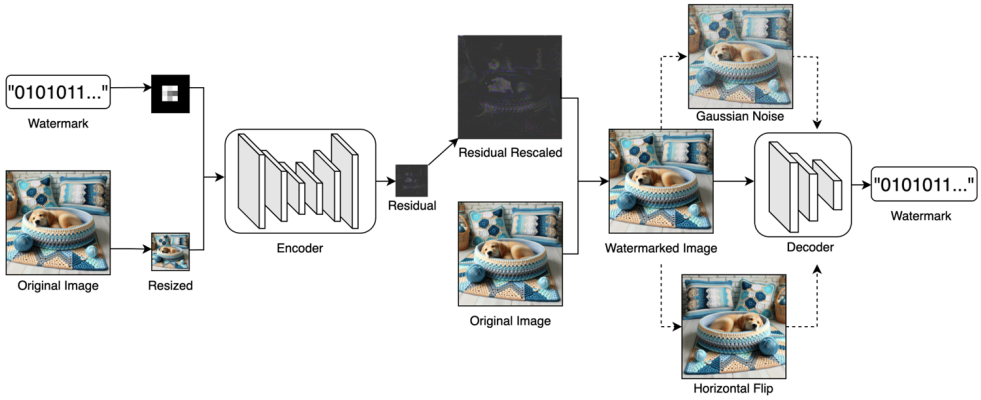

图 4: pHash 的基本处理流程

pHash 通过提取图像的低频结构特征完成感知哈希计算。设输入图片为 I，其处理流程为：先将图片灰度化并统一缩放为固定尺寸，再对缩放后的图像执行二维离散余弦变换，得到频域系数矩阵 F。由于图像的整体结构信息主要集中在低频区域，因此从 F 中截取左上角的低频子块，通常取 8×8 区域。随后对该低频块求均值，并按“大于等于均值记为 1，小于均值记为 0”的规则进行二值化，最终得到定长二进制序列作为感知哈希值[5]。

若两张图片的 pHash 分别为 p 和 q，则它们之间的相似程度由汉明距离衡量。距离越小，说明两张图片在整体视觉结构上越接近。由于 JPEG 压缩和缩放对图像低频分布影响较小，pHash 对这类传播处理通常更稳定；而裁剪会直接改变图像构图，因此更容易导致距离增大。基于这一特性，pHash 更适合作为传播后的近似找回机制，而不是单独承担最终认证功能[1]。

图 5: 代表样本及其传播处理结果

样本类型	pHash	汉明距离
原图	e0da822b05f0c20d	0
JPEG 压缩	e0da822b05f0c20d	0
缩放重采样	e0da822b05f0c20d	0
裁剪样本	f8d1d4022af08052	25

表 1: 代表样本的 pHash 值与汉明距离

这一点在实验结果中表现得很清楚。图 5 给出了代表样本及其传播处理结果，表 1 列出了对应的 pHash 距离。JPEG 压缩图与原图之间的距离为 0，缩放图与原图之间的距离也为 0，说明这两类传播处理虽然破坏了文件级一致性，但没有显著改变整体视觉结构。裁剪样本与原图之间的距离上升到 26，这表明 pHash 可以稳定处理轻中度传播失真，但对大幅裁剪的适应性明显下降。

尽管 pHash 对裁剪较为敏感，本文仍采用它作为传播后的近似找回机制，因为实际传播中更常见的是压缩、缩放和截图式重保存，而 pHash 对这类轻中度失真保持了较好稳定性。它在这里承担的不是最终认证功能，而是候选记录搜索功能。对于裁剪带来的结构破坏，后续可通过更强的隐式标识、局部特征匹配或深度特征表示进一步增强抗裁剪能力。

4 区块链登记实验设计与实现

区块链在这一方案中承担的是来源登记层。图像相似性判断仍在链下完成，链上的作用是保存关键摘要与标识，使来源记录能够被查询和追溯。链上只登记与验证相关的字段，包括 `contentId`、`fileName`、`platformName`、`watermarkId`、`pHashValue`、`timestamp` 和 `exists`。其中，`contentId` 对应嵌入隐式标识后的正式图片 SHA-256，`watermarkId` 用于保存内容线索，`pHashValue` 用于保存传播后的近似匹配摘要。

实验环境采用 Ganache 本地测试链，并通过 Python 的 Web3 接口完成合约部署、记录写入和链上查询。链上合约采用 `ProvenanceRegistryV3`，以 `contentId` 为主索引保存完整记录，同时建立 `watermarkId` 到 `contentId` 的映射关系，使链上对象既能支持精确查询，也能配合隐式标识完成来源回溯。

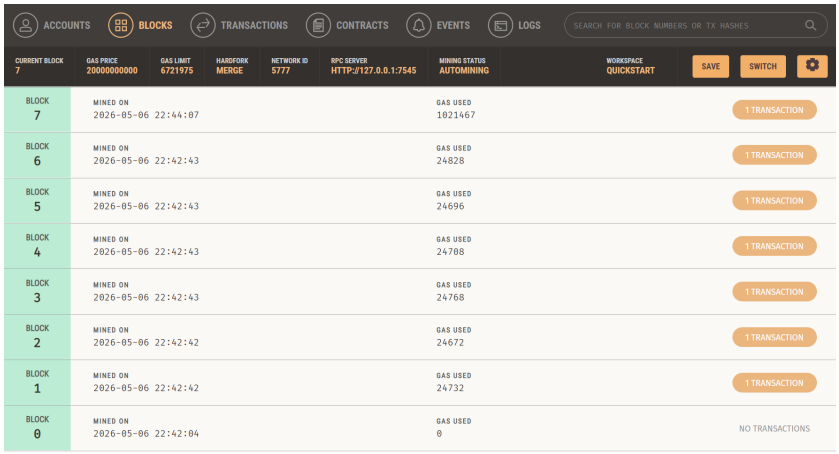

图 6: Ganache 本地测试链环境

登记过程围绕嵌入隐式标识后的正式图片展开。脚本先为图片生成 `watermark_id`，

再计算对应的 SHA-256 与 pHash。其中，SHA-256 作为链上主键，pHash 作为传播后的近似匹配依据。完成双摘要计算后，脚本调用 `registerRecord` 接口，将文件名、来源平台、隐式标识、感知哈希和时间戳写入链上。这样，图片的来源信息就从本地处理结果转化为正式登记对象。

查询阶段以 `contentId` 为输入，通过 `getRecordByContentId` 接口读取链上记录，返回文件名、平台标识、隐式标识、感知哈希、时间戳和存在标志。这样，传播后的图片在完成近似找回后，就可以进一步对应到链上正式对象，而不只是停留在“视觉上接近”的层面。

5 实验结果与可行性分析

5.1 链上登记与查询成本

链上成本记录包括合约部署、单样本登记和查询估算三部分。实验合约部署成本为 1,021,467 gas，说明该合约虽然承担了来源字段登记和索引映射功能，但整体结构仍保持在可控范围内。代表样本登记成本为 396,789 gas，对应的 `contentId` 为 `69bcbdd9d31426fde1dcece2d39349e326e439e6b80e99bedea14cfd583e1e87`，其 `pHashValue` 为 `e0da822f05f1da1f`。

查询阶段采用只读调用方式，不产生真实写链费用。但对 `getRecordByContentId` 的估算开销约为 48,999 gas。

查询阶段通过 `getRecordByContentId` 读取链上正式记录。返回结果包含文件名、来源平台、隐式标识、感知哈希、时间戳和存在标志。验证链上对象能够被后续验证阶段稳定取回。

传播后的图片在完成近似找回之后，可以进一步对应到正式来源对象，而不只是停留在“视觉上接近”的层面。

表 2: 链上查询成功返回的数据

字段	返回值
contentId	69bcbdd9...583e1e87
fileName	zuixuanminzufeng.png
platformName	local_aigc_platform
watermarkId	WM900001
pHashValue	e0da822f05f1da1f
exists	True

5.2 可行性与边界分析

从实验结果看，这条验证链路具备可行性：区块链登记层能够稳定保存原始来源对象，精确哈希能够处理 exact 场景，pHash 能够在 JPEG 压缩、缩放和截图式传播下继续支持近似找回，隐式标识和链上记录则为后续正式关联提供了依据。

这套方案的边界也同样清楚。pHash 对大幅裁剪较为敏感，说明单一全局感知哈希并不能覆盖所有传播变形；隐式标识在强攻击下也可能失去稳定恢复能力；区块链则始终只承担登记和查询角色，并不直接提高图像匹配能力。因此，这一方案更适合作为传播后的来源验证机制，而不是通用图像鉴别器。通过“精确摘要 + 隐式标识 + pHash + 区块链登记层”的组合，已登记 AIGC 图片在轻中度传播场景下仍能够被重新关联回原始来源对象，同时又保留了对强结构破坏的敏感性。

6 扩展方向

6.1 更细粒度的分区式视觉摘要

当前实验采用的是单一全局 pHash。后续更合理的扩展方向，是将图片划分为多个局部区域并分别计算感知摘要，再将这些局部结果组合为区域化特征。裁剪通常只破坏部分区域，局部区域仍有保留，系统仍可能通过局部摘要与原始记录建立对应关系[1, 4]。

6.2 链上索引与来源对象扩展

链上记录目前保存的是 `contentId`、`watermarkId` 和 `pHashValue` 等核心字段。若后续引入分区式视觉摘要或更丰富的局部特征，可以在链下保存完整描述，在链上登记对应的摘要、索引或压缩表示。已有的 provenance 研究表明，来源记录还可以被组织为多阶段对象，而不只是一条静态登记记录[3]。

6.3 多平台传播链重建

当前实验针对的是单张图片和单阶段传播过程。进一步扩展时，可以将不同平台上的传播状态组织为阶段性记录，并将这些记录与链上登记对象关联起来。这样，系统不仅能够判断图片对应哪条原始记录，还可以描述图片在传播过程中经历了哪些处理阶段。

7 现有产业实践与项目意义

AIGC 内容来源标识已逐步从研究探索走向实际应用。OpenAI 已在 ChatGPT 图像生成中引入 C2PA 元数据，用于标识图片来源，并将其纳入内容透明度机制之中；相关系统说明也将 provenance 工具视为图像生成安全方案的重要组成部分[6]。Google 则在 Imagen、Google Photos 的 Reimagine 等产品中采用 SynthID，将不可见数字水印嵌入 AI 生成或编辑后的内容，并进一步提供了 SynthID Detector 等识别工具[7]。由此可见，当前产业实践已初步形成两类较为明确的技术路径：一类侧重于基于 C2PA 的来源声明与可验证元数据，另一类侧重于以不可见水印为代表的内容内嵌标识。

本项目将“精确摘要、隐式标识、感知哈希与区块链登记层”组织为一条可运行的验证链路，并通过真实样本表明，轻中度传播失真条件下的来源回溯具备可验证性；进一步将来源记录由本地脚本中的临时结果转化为链上正式对象，使图片在完成近似找回之后，仍能够继续对应到带有时间戳与状态信息的登记记录。相较于单纯的相似性判断，这种结构更接近实际场景中对“来源回溯”与“可验证依据”同时提出要求的应用需求[1, 2]。

在更完整的应用场景中，区块链还可以承担跨平台来源记录同步、阶段性传播状态登记以及多主体协同验证等功能。若不同平台能够围绕统一字段结构或统一来源对象标准进行登记，则传播后的图片不仅可以对应到单一原始记录，还可以进一步形成跨平台可追溯的来源链条。这样，区块链的作用将不再局限于静态存证，而能够延伸到来源记录共享、传播路径重建和验证责任划分等更完整的治理场景[3]。

参考文献

[1] Mohit A, Aggarwal B, Gondhalekar C. Provenance Verification of AI-Generated Images via a Perceptual Hash Registry Anchored on Blockchain. arXiv preprint arXiv:2602.02412, 2026.

[2] Xu R, Hu M, Lei D, et al. InvisMark: Invisible and Robust Watermarking for AI-Generated Image Provenance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025: 909–918.

[3] England P, Malvar H S, Horvitz E, et al. AMP: Authentication of Media via Provenance. In: Proceedings of the 12th ACM Multimedia Systems Conference, 2021: 108–121.

[4] 刘安安，苏育挺，王岚君，等. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报，2024，29(06): 1535–1554. DOI: 10.11834/jig.240003.

[5] Zauner C. Implementation and Benchmarking of Perceptual Image Hash Functions. Master’s thesis, Upper Austria University of Applied Sciences, Hagenberg Campus, 2010.

[6] OpenAI. C2PA in ChatGPT Images. OpenAI Help Center, 2025.

[7] Google DeepMind. SynthID. Google DeepMind, 2025.