

生成式图像版权保护与可信溯源技术综述

王海龙, 谭武征, 张子欣

(暨南大学网络空间安全学院, 广东 广州 511436)

摘要: 生成式图像的版权保护对象已由最终图像扩展至训练数据、模型参数、生成过程及传播凭证, 生成范式的演化进一步改变了版权标识的嵌入位置与证据验证条件。现有综述多围绕特定水印机制或攻防过程展开, 对不同生成范式、保护对象与证据目标之间的联系仍缺少统一比较。为此, 本文以 2023—2026 年的相关研究为重点, 适当回溯早期基础工作, 从生成范式、版权对象和证据机制三个维度梳理生成式图像版权保护与可信溯源技术。比较生成对抗网络、潜在扩散、视觉自回归及混合架构的可利用变量与攻击面, 分析主动水印、训练数据与模型权属保护、被动取证及外部内容凭证的作用机制与适用边界; 结合公开基准报告的结果, 区分共同测试条件与方法配置差异, 讨论水印检测与图像保真度的权衡及篡改定位表现, 并以小规模实测说明多源证据联合保持率的计算条件。现有方法在生成式编辑、跨范式迁移及错误归因控制方面仍存在局限, 单项检测或定位性能不足以表征完整的证据效力。最后, 围绕跨生成范式适配、语义绑定、公开验证与跨媒体证据连续性问题, 讨论可信溯源技术的发展方向。

关键词: 生成式图像; 数字水印; 版权保护; 内容溯源; 训练数据版权; C2PA

中图分类号: TP309

文献标志码: A

doi: 10.11959/j.issn.1000

A Survey of Copyright Protection and Trustworthy Provenance for Generative Images

Wang Hailong, Tan Wuzheng, Zhang Zixin

College of Cybersecurity, Jinan University, Guangzhou, Guangdong 511436, China

Abstract: Copyright protection for generative images now extends beyond final outputs to training data, model parameters, generation processes, and provenance credentials. The evolution of generation paradigms further changes where copyright identifiers can be embedded and under what conditions evidence can be verified. Existing surveys mainly focus on specific watermarking mechanisms or attack and defense processes, while the relationships among generation paradigms, protected objects, and evidence goals remain insufficiently compared. To address this gap, studies from 2023 to 2026 were reviewed from three perspectives: generation paradigms, copyright objects, and evidence mechanisms, with relevant earlier work included for context. Embeddable variables and attack surfaces were compared across generative adversarial networks, latent diffusion models, visual autoregressive models, and hybrid architectures. The mechanisms and applicability of proactive watermarking, training-data and model ownership protection, passive forensics, and external content credentials were examined. Results reported in public benchmarks were analyzed to examine trade-offs between watermark detection performance and image fidelity and to discuss tamper localization performance, while distinguishing shared test conditions from differences in method configurations. A small-scale empirical study was used to illustrate the conditions for calculating the joint retention rate of evidence from multiple sources. Existing methods remain

收稿日期: 2026-09-XX; 修回日期: 2026-XX-XX

通信作者: 谭武征, tanwuzheng@jnu.edu.cn

基金项目: 国家自然科学基金资助项目(No.62272199)

Foundation Items: The National Natural Science Foundation of China (No.62272199)

limited under generative editing, cross-paradigm transfer, and misattribution attacks; performance on an individual detection or localization task is insufficient to establish the validity of a complete evidence set. Finally, future research directions were discussed in terms of cross-paradigm adaptation, semantic binding, public verification, and evidence continuity across media.

Key words: generative images, digital watermarking, copyright protection, content provenance, training-data copyright, Coalition for Content Provenance and Authenticity (C2PA)

0 引言

生成式图像的技术基础已由生成对抗网络 (generative adversarial network, GAN) 扩展至潜在扩散、视觉自回归及 AR-flow 等混合架构^[1-5]。生成机制的变化不仅影响合成质量与效率, 也改变了版权标识的嵌入位置和证据形式: GAN 主要利用模型参数、特征表示与输出指纹; 扩散模型可利用初始噪声、去噪轨迹、潜变量及变分自编码器 (variational autoencoder, VAE) 解码器; 视觉自回归模型则提供词元 (token) 概率、词元选择与词元编码器 (tokenizer) 等操作空间^[6]。

2023 年以前, 图像水印研究已将传播失真纳入训练。HiDDeN 通过编码器、失真层与解码器联合训练提高鲁棒性^[7], StegaStamp 进一步覆盖打印、拍摄等物理传播过程^[8]; GAN 模型指纹与潜在扩散则分别提供模型验证手段和生成时嵌入接口^[1-2,9], 为后续面向生成式编辑的保护奠定基础。

与生成机制演进相伴, 攻击方式由压缩、裁剪和缩放扩展至再生成、局部重绘、视觉释义和图生视频。生成模型可在维持主要语义的同时削弱不可见水印^[10-12]。StableGuard、GenPTW、TAG-WM 和 PAI 等工作进一步将来源追踪、版权验证与篡改定位纳入统一设计^[13-16], 评价目标因而由水印存在性检测扩展至多类证据的保持与验证。

内容来源、权利主体、篡改完整性、训练数据使用和传播历史等证据的验证前提与证明能力并不相同。可信溯源需明确各类证据的适用范围, 并考察身份绑定、内容完整性、抗伪造性、传播鲁棒性和第三方复核能力。

现有工作已从不同视角讨论人工智能生成内容 (artificial intelligence-generated content, AIGC) 的视觉安全与水印技术。刘安安等^[17]梳理视觉内容生成与溯源, 郭钊均等^[18]总结 AIGC 模型水印, 唐晓亮等^[19]综述扩散模型水印, 谢雪等^[20]从“防一查一抗”动态博弈视角讨论主动防御。在已有分类

与攻防分析的基础上, 生成范式演进后嵌入接口如何变化、不同版权对象需要何种证据, 以及各类验证结果能够支持何种判断, 仍需结合比较。

据此, 本文以 2023—2026 年的研究为重点, 并适当回溯基础工作, 从生成范式、版权对象和证据机制三个维度组织现有成果, 如图 1 所示。本文的主要贡献如下:

(1) 建立生成范式、版权对象与证据机制之间的对应关系, 对比 GAN、潜在扩散、视觉自回归及混合架构的嵌入接口与攻击面, 分析保护方法跨范式迁移的适用条件。

(2) 围绕来源追踪、权属验证与篡改定位等目标, 比较主动水印、被动取证和外部内容凭证的验证前提与证明能力, 辨析内容来源识别、权利主体绑定和训练数据使用证明之间的差异。

(3) 依据公开基准报告的结果分析检测与保真度的权衡及定位表现, 结合小规模实测说明多源证据联合保持率的计算及适用条件, 明确单项性能与多项证据同时成立之间的区别。

1 生成范式演进与版权证据载体

1.1 GAN 生成机制与模型指纹

GAN 的单次前向生成缺少扩散模型中的逐步可逆噪声轨迹, 但生成器参数、卷积上采样路径及输出频率模式仍可能形成具有识别价值的模型指纹或来源特征^[1,21]。据此, GAN 场景下的版权保护主要包括参数或激活水印、触发式黑盒验证以及输出指纹归因, 其可靠性容易受到微调、蒸馏、模型提取和跨模型泛化能力不足等因素影响^[9]。

1.2 扩散模型中的噪声、潜变量与生成轨迹

潜在扩散模型 (latent diffusion model, LDM) 将主要生成过程迁移到 VAE 潜空间, 使初始噪声、去噪轨迹、潜变量和解码器等多个层级均可作为水印载体^[2]。Stable Signature 通过微调 VAE 解码器植入模型级签名, Tree-Ring 在初始噪声频域写入环状模式, Gaussian Shading 则通过约束带水印潜变

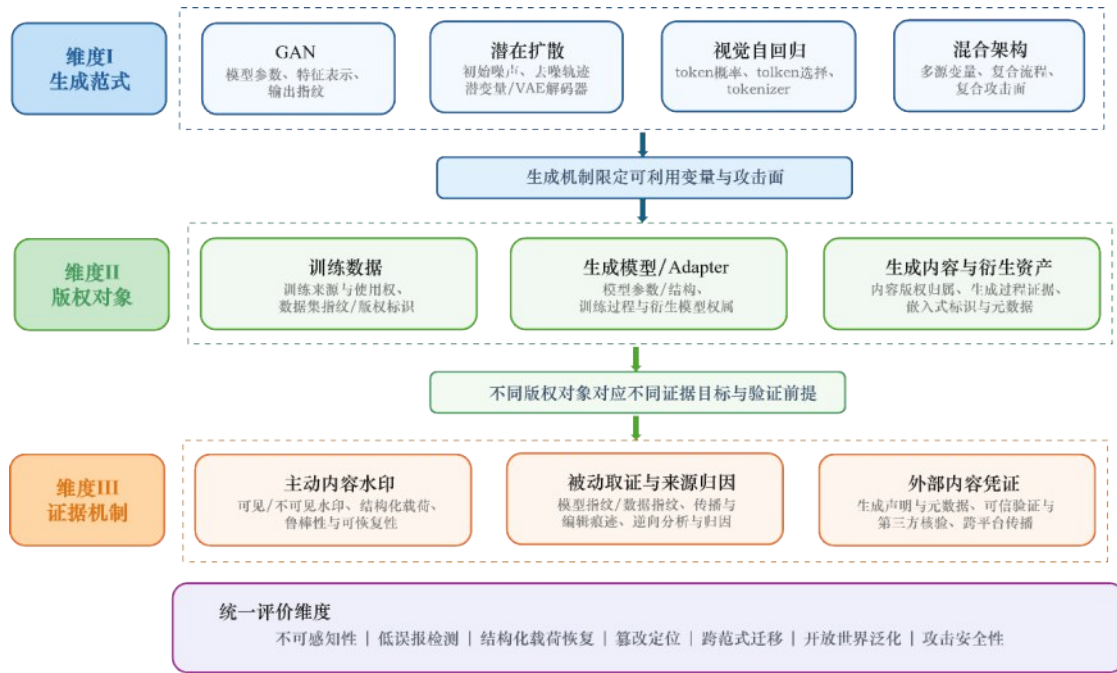


图1 生成式图像版权保护与可信溯源的三维分析框架

量保持高斯先验^[22-24]。这些方法分别对应解码器参数、噪声结构和概率分布三个不同层面的生成时嵌入思路。

扩散模型的去噪能力同时为水印移除提供了新的攻击途径，加噪—去噪、图生图和局部编辑均可在保持主要语义的条件下重构图像像素^[10-11]。受此影响，近期研究除在最终像素中嵌入标识外，还利用生成轨迹、语义表示、密钥条件变量或生成随机性构造可验证信号，Guidance Watermarking、Spherical Watermark、NoisePrints、PAI 和 Gaussian Shannon 体现了这些不同思路^[16,25-28]。

1.3 视觉自回归与混合生成范式下的 token 级证据

视觉自回归模型以离散或连续 token 进行序列建模，VAR、MAR 及后续 AR-flow 混合范式所提供的生成接口与扩散模型存在明显差异^[3-5]。这意

味着依赖 DDIM 反演、固定噪声调度或特定 VAE 结构的水印方案可能受到生成架构演进的限制，因此跨范式适配应作为评价方法长期可用性的重要维度。

NeurIPS 2025 的相关工作将统计水印嵌入自回归 token 生成过程，并通过同步机制与 tokenizer 改进缓解重新 tokenization 引起的失配^[6]；ICLR 2026 提出的 IAR provenance 则表明，自回归模型输出本身也包含可用于来源追踪的被动特征^[29]。表1对不同生成范式的可嵌入变量、适用机制及主要风险进行了归纳。

2 版权与可信溯源的证据目标及评价体系

2.1 可信溯源的证据目标与验证要求

从保护对象和验证功能两个层面看，生成式图

表1 不同生成范式下的版权保护与溯源机制对比

生成范式	典型可嵌入或可观测变量	适合的保护/溯源机制	主要风险
GAN	权重、特征层、生成器输出、频率指纹	模型水印、输出水印、生成器指纹归因	蒸馏、模型提取、再训练及跨模型泛化不足
潜在扩散	初始噪声、去噪轨迹、交叉注意力、潜变量、VAE 解码器	生成时水印、反演检测、轨迹水印、篡改定位	再生成净化、反演漂移、VAE 替换及编辑器迁移
视觉自回归	离散/连续 token、token 概率、tokenizer	token 级统计水印、序列来源证明、被动指纹	重新 tokenization、tokenizer 变化及跨模型同步
AR-flow/混合模型	token 序列、连续流头、解码潜变量	可探索分层水印、adapter 化嵌入、上层统一 Content-ID	多模块耦合、协议碎片化及验证成本增加

像版权与溯源可以进一步细分。保护对象包括训练数据、生成模型或适配器以及最终生成内容；验证功能则涉及来源证明、权利主体绑定、篡改完整性、训练数据使用和传播历史等证据目标。主动水印通常经历嵌入、传播或攻击、解码和验证等环节，不同方法的主要差别在于其作用对象及验证输出，例如水印存在性、Content-ID、来源类别、局部篡改掩码或密码学声明。

2.2 威胁模型与评价指标

威胁模型应同时覆盖传统信号处理、生成式移除、语义编辑、水印伪造或嫁接、模型与密钥攻击以及开放世界分布漂移。WAVES将传统攻击、扩散攻击与对抗攻击纳入统一压力测试^[30]，Diff-Forge进一步表明，攻击者不仅可能移除既有水印，还可能伪造水印并造成错误归因^[31]。因此，鲁棒性结果需要结合攻击者的访问权限解释，并明确其是否能够获得检测器梯度、嵌入算法、在线查询接口或生成模型权重。

在生成式图像场景下，仅采用JPEG压缩、裁剪等传统失真测试难以充分反映水印系统的实际安全边界。再生成、局部扩散编辑、视觉释义和跨模型编辑会直接改变图像生成过程或语义表达，应作为独立的生成式攻击类型纳入评测；同时，还应依据攻击者可获得的算法、模型与验证接口信息，分别报告无盒、黑盒和白盒条件下的攻击效果。

评价指标应与具体证据目标相匹配。不可感知性通常采用峰值信噪比（PSNR）、结构相似性（SSIM）和学习感知图像块相似度（LPIPS）衡量，载荷恢复可报告比特错误率（BER）或标识（ID）准确率，水印存在性检测应报告接收者操作特征曲线下面积（AUC）以及低假阳性率（FPR）条件下的真阳性率（TPR），篡改定位可采用交并比（IoU）和像素级F1分数；来源归因还需考察开放集识别能力与持续学习中的遗忘程度。对于承载结

构化版权信息的方案，还应区分“水印被检测到”与“载荷被准确恢复”两个不同层面的结果^[22,28]。

对于多源证据协同，可预先定义一次来源声明所需的证据集合，并统计传播或攻击后仍满足该集合的样本比例。本文将这一联合通过率称为证据链存活率（Evidence-chain Survival Rate, ESR）。设测试集包含 N 个样本，声明所需证据共 K 项；在固定攻击条件下，第 i 个样本的第 j 项证据通过预设检验时，记 $I_{ij}=1$ ，否则为0，则有

$$ESR = \frac{1}{N} \sum_{i=1}^N \prod_{j=1}^K I_{ij} \quad (1)$$

以来源恢复与局部完整性联合验证为例，必需证据集合可取Content-ID准确恢复、与该ID绑定的预期主体签名验证通过，以及篡改掩码IoU达到预设阈值。模型判定阈值在独立验证集上确定，评价门限预先指定，测试时保持不变，并在同一样本上逐项判定。缺失或无法验证的必需证据记为失败，不适用的任务应单独分组，对无篡改样本还需预先约定空掩码判定规则，避免以缺省满分抬高ESR。

ESR反映必需证据同时成立的比例，各项检验须针对同一来源声明，证据集合与判定条件一致时方可比较。若仅已知同一样本集上各项证据的通过率 p_j ，则ESR不高于其中的最小值，且不低于 $\max(0, \sum_j p_j - K + 1)$ ；一般不能由各项成功率的平均或乘积确定其值。本文第7.1节中图3所依据的汇总表不含逐样本联合记录，同时给出小规模测试数据。评测应同时报告证据集合、攻击参数、样本量、各单项通过率和联合通过率。对于水印嫁接等攻击，还须在伪造样本上另报错误接受率，以免将错误主体的有效签名计作来源验证成功。表2归纳了不同证据目标的典型输出、评价指标及局限。

3 主动水印与内容级版权保护

典型主动水印方法的技术流程如图2所示。各

表 2 不同证据目标的典型输出、评价指标及局限

证据目标	典型输出	关键指标	单独使用的局限
内容来源	水印存在性、模型/平台ID	TPR@低FPR、BER、ID准确率	难以单独建立法律权利主体绑定
权利主体	Content-ID、账户/作者签名	验签率、错误归因率、抗伪造率	依赖密钥管理与身份治理机制
篡改完整性	区域掩码、局部消息	IoU、F1、局部BER	通常缺少上游来源信息
训练数据使用	统计检验、归因分数	检出率、效应量、因果稳定性	与法律意义上的复制关系仍需进一步解释
传播链	manifest、签名、编辑历史	验签成功率、链路完整率、ESR	元数据可能被剥离，需内容内证据补充

路径依次给出嵌入、生成或传播、信号恢复及验证步骤,嵌入位置与验证方式的差异决定了其部署条件和主要失效环节。

3.1 后处理水印与局部证据保持

后处理水印不依赖具体生成器,因而具有较好的部署通用性。WAM通过局部消息与区域分割支持拼接和局部编辑场景,VINE借助生成先验提高水印在再生成、全局或局部编辑以及图生视频条件下的保持能力^[11,32]。相关研究的评价重点已由单纯关注视觉质量,扩展到生成式编辑条件下局部版权证据的保持与恢复。

图2中的后处理路径先将图像与消息共同输入编码器,再对传播或编辑后的图像解码。VINE利用生成先验改善含水印图像质量,并以编辑相关失真增强训练;WAM还输出区域掩码,将消息恢复限制在对应区域。二者分别侧重编辑后的全局消息保持和局部消息归属,不能仅凭比特准确率评价区域定位能力^[11,32]。

SynthID-Image代表了平台级水印部署的一类实现思路,其公开研究同时将有效性、保真度、鲁棒性和安全性作为互联网规模应用的基本要求^[33]。这类工程实践表明,实验室条件下的bit accuracy只能反映系统性能的一部分,实际部署还需要同时

考虑密钥轮换、版本兼容、误报基线和平台再编码等因素。

3.2 扩散生成过程中的内生水印

扩散生成过程中的内生水印直接利用模型内部变量形成标识。Stable Signature、Tree-Ring 和 Gaussian Shading 分别在 VAE 解码器、初始噪声频域和高斯潜变量中嵌入版权证据^[22-24],从而使水印机制与生成过程建立更加直接的联系。

这几类方法的验证过程也有所不同。Stable Signature 先离线微调 VAE 解码器,生成时写入固定签名,检测时直接从图像提取比特并进行统计检验;Tree-Ring 等初始噪声方法先按密钥构造噪声,再经去噪与解码生成图像,检测时则需编码和反演以恢复噪声特征。前者承担模型适配成本,后者的验证开销与反演误差随生成配置变化,因此免训练并不意味着低验证成本。DeWM 通过编码器与解码器特征共享训练水印网络,再微调扩散模型解码器,使生成图像携带特定水印^[34]。

2026 年的相关研究进一步关注生成轨迹、分布保持和结构化通信。Guidance Watermarking 利用解码器梯度引导采样,Spherical Watermark 强调带水印变量与高斯分布的一致性,NoisePrints 将生成随机种子(seed)用于生成来源证明并结合零知识

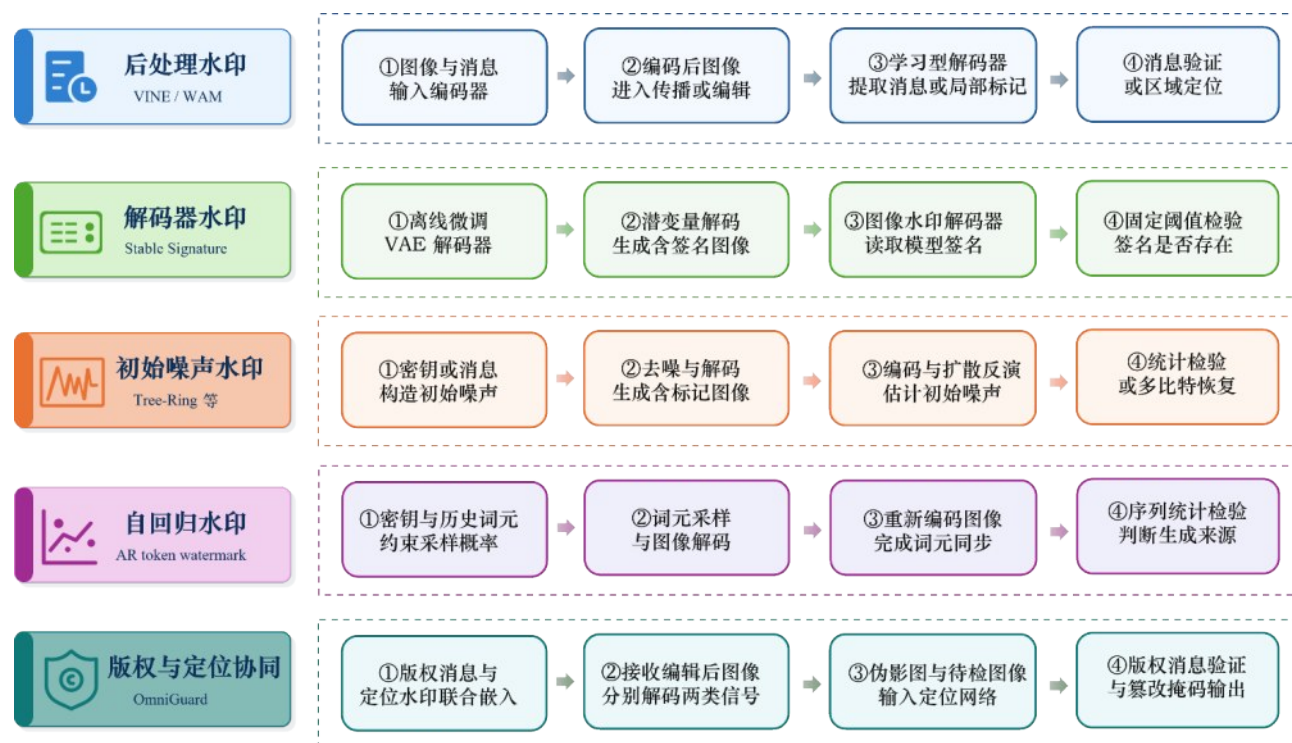


图2 典型主动水印方法的技术流程

验证, Gaussian Shannon 则把生成、反演与攻击过程建模为通信信道, 以支持多比特信息恢复^[25-28]。

除嵌入机制外, 验证效率和迁移成本也开始成为方法比较的重要因素。FARI、SERUM 和 Diff-Mark 分别从快速反演、多用户标记以及跨冻结扩散模型迁移等角度降低验证或适配开销^[35-37]。因此, 对生成时水印的评价还应纳入验证开销、跨模型检查点 (checkpoint) 迁移能力和密钥扩展能力等工程指标。

3.3 版权验证与篡改定位的协同设计

版权验证与篡改定位的协同设计是近两年的重要发展方向。EditGuard 将鲁棒版权消息恢复与篡改定位结合, OmniGuard 进一步融合主动嵌入和被动定位, TAG-WM 则利用扩散反演对局部编辑的敏感性同时承载版权信息与定位信息^[15,38-39]。

在图2的协同路径中, OmniGuard 联合嵌入版权消息和定位水印, 验证时分别恢复两类信号, 再将定位水印的恢复伪影与待检图像输入定位网络。版权分支要求消息抵抗编辑, 定位分支则需要对局部修改保持敏感; 对两类信号分别建模, 有助于减轻全局退化与局部篡改之间的混淆^[39]。

StableGuard 将水印直接嵌入潜在扩散生成过程, 并通过取证网络联合建模全局水印、局部篡改痕迹和频域线索^[13]。与仅在取证端独立识别的方案相比, 该方法更强调嵌入机制与后续取证任务之间的协同优化。

GenPTW 尝试在潜空间统一生成前与生成后的水印嵌入, 并进一步扩展到潜在扩散模型和视觉自回归模型; PAI 则通过密钥条件下的轨迹偏转增强身份信息与内容语义之间的关联, 同时执行攻击检测和语义级定位^[14,16]。这类工作表明, 水印所承载的证据正由固定全局消息扩展到来源身份、内容完整性和局部篡改证据等多个层面。

3.4 视觉自回归水印与跨范式适配

视觉自回归水印可将图像水印转化为 token 序列上的统计检测问题, 其优势在于能够利用 token 采样概率构造检测统计量, 而重新 tokenization 以及 tokenizer 结构变化则可能破坏序列一致性, 从而影响水印检测稳定性^[6]。

其实现过程通常包括密钥约束下的 token 采样、图像解码、待检图像重新编码和序列统计检验。与扩散反演依赖噪声重建不同, 该路径首先需要恢复

可对齐的 token 序列; 若编辑或编码器变化破坏同步, 即使生成时的统计偏置仍部分保留, 检测能力也可能下降。

从长期兼容性看, 一种可行思路是将上层证据语义与底层生成器实现解耦: 上层统一定义 Content-ID、时间戳或主体签名, 底层则针对扩散模型、IAR 或混合模型分别设计适配机制。表3按照嵌入层级、主要能力和局限对 2023—2026 年的代表性主动水印方法进行了比较。

4 训练数据与模型权属保护

4.1 训练数据水印与统计使用证据

训练数据水印关注标记信号能否被模型学习并在输出中检出。DiffusionShield、ProMark 和 Copr-Guard 分别从训练数据标记、主动因果归因和频谱统计等角度构造数据使用证据^[40-42]。由于此类证据容易受到数据构成和训练配置影响, 实验通常需要同时报告标记比例、阴性对照以及不同模型架构和训练设置下的效应稳定性, 以降低自然统计偏差造成的混淆。

4.2 保护性扰动与训练阶段干预

Glaze 和 Nightshade 分别通过风格遮蔽与提示词特定投毒提高未经授权训练或风格模仿的成本^[43-44], Watermark-embedded Adversarial Examples 则尝试将训练干扰与可验证标识结合^[45]。这类方法主要作用于训练前的预防阶段, 其技术效果与训练后证明特定数据已被使用属于不同的证据问题, 综述与评价中应予以区分。

因此, 训练数据版权保护可以按照授权控制、保护性扰动、水印追踪和争议审计等环节组织, 不同技术对应的作用阶段、成立条件和证据含义需要分别说明。

4.3 数据归因与贡献解释

数据归因旨在没有预置水印的条件下分析训练样本与特定输出的联系, 可用于历史数据的核查, 其模型访问需求和计算成本随方法而异。Wang 等通过定制生成样本构建归因评测集, 比较不同特征空间下的样本关联能力^[46]。归因分数不能直接等同于法律意义上的复制或侵权认定, 更适合作为训练日志、水印检验和局部视觉对应关系的补充证据。

表 32023—2026 年代表性生成式图像主动水印方法对比

方法	发表时间/状态	嵌入层级	主要能力	关键局限或研究问题
Stable Signature	ICCV 2023	VAE 解码器	模型级签名、低 FPR 检测	强生成式编辑条件下鲁棒性仍受限
Tree-Ring	NeurIPS 2023	初始噪声频域	免训练、反演检测	依赖扩散反演与特定生成机制
Gaussian Shading	CVPR 2024	高斯潜变量	概率保持、多比特	密钥与反演管理复杂，强编辑鲁棒性仍需验证
EditGuard	CVPR 2024	后处理深度水印	版权恢复+篡改定位	与具体生成过程耦合较弱
WAM	ICLR 2025	后处理局部水印	多区域、多消息、拼接定位	强语义重生条件下仍可能失效
VINE	ICLR 2025	生成先验辅助后处理	面向再生成与编辑的鲁棒性增强	模型体量与跨编辑器泛化成本
OmniGuard	CVPR 2025	主动与被动混合	混合编辑定位与版权恢复	系统复杂度较高
TAG-WM	ICCV 2025	扩散潜空间	版权验证与篡改定位	依赖反演敏感性
StableGuard	NeurIPS 2025	VAE/生成过程内	联合版权验证+篡改定位	当前主要面向 LDM
GenPTW	AAAI 2026	潜空间	兼容生成前后嵌入、追踪与定位，并支持 LDM/VAR	开放世界条件下的验证仍需扩展
Guidance Water-marking	ICLR 2026	去噪轨迹	现有解码器转为生成时嵌入	生成开销及梯度引导安全性仍需评估
Spherical Water-mark	ICLR 2026	初始高斯噪声	分布保持、无损标记思路	复杂编辑与跨架构适应性仍需验证
NoisePrints	ICLR 2026	随机 seed	生成来源证明、零知识验证	载荷表达能力有限
PAI	ICLR 2026	密钥条件轨迹	权属验证、攻击检测与语义定位	对更强自适应攻击的鲁棒性仍需评估
Gaussian Shannon	CVPR 2026 Findings	初始高斯噪声	结构化多比特精确恢复	依赖反演与通信信道假设
FARI	ICLR 2026	快速反演+轻量 LoRA	一步验证、外部失真优化	仍以扩散反演为核心
SERUM	ICLR 2026	初始噪声+轻量检测器	高效、多用户、解耦验证	跨生成范式适配能力仍需扩展
AR token water-mark	NeurIPS 2025	视觉 token	自回归生成来源检测	受重新 tokenization 与 tokenizer 迁移影响

4.4 模型所有权与可迁移组件标识

AquaLoRA 将水印能力封装在 LoRA 模块中，使可公开访问参数的定制扩散模型生成带标识图像^[47]。在模型即服务环境下，模型身份还会受到蒸馏、微调、模型提取和版本迁移等因素影响，因

此可考虑按照“基础模型—adapter—服务”的层级管理模型标识。表 4 归纳了训练数据水印、保护性扰动、数据归因和模型水印对应的证据类型及其特点。

表 4训练数据与模型权属保护方法的证据类型及特点

路线	代表工作	目标证据	优势	核心不足
训练数据水印	DiffusionShield、ProMark、Copr-Guard	数据曾被模型学习的统计证据	支持黑盒条件下的事后追踪	低掺入率、训练配置变化与因果解释困难
保护性扰动	Glaze、Nightshade	降低未授权训练/风格模仿效果	具备较强的事前干预能力	不能直接证明数据已被使用，且可能受预处理与数据稀释影响
数据归因	训练监控、扩散归因、区域级归因	特定样本对输出的影响	无需预先标记	计算成本较高，贡献分数与侵权认定并不等价
模型水印	AquaLoRA、GAN/DNN 模型水印	模型身份或来源标识	为模型资产权属核验提供线索	蒸馏、微调、模型提取与版本迁移仍是难点

5 被动生成图像取证与来源归因

5.1 从生成器指纹到扩散伪影

被动取证不要求生成方预先嵌入标识，因此可用于恶意、开源或历史生成内容的分析。UniversalFakeDetect借助预训练视觉特征提高跨生成器泛化能力，DIRE和AEROBLADE分别利用扩散重建误差或自编码器重建误差识别生成图像，GenImage则通过大规模多生成器数据推动统一基准评测^[21,48-50]。

近期研究进一步关注跨生成器家族和未知生成器条件下的泛化能力，其中FIRE采用频率引导的重建误差，VIB-Net利用信息瓶颈抑制与真伪判别无关的表面偏差^[51-52]。此类方法有助于提高通用检测性能，但其输出通常仍是生成概率或来源类别，难以直接恢复作者身份、授权状态或Content-ID。

5.2 开放集、持续学习与语义偏差

Forensic Self-Descriptions利用自监督取证表征统一零样本检测、开放集来源归因和聚类任务^[53]，IAR provenance则表明视觉自回归模型输出同样具有可用于来源识别的特征^[29]。国内研究近年来也开始系统关注持续检测、扩散伪影和语义对齐问题，并分别从灾难性遗忘、跨域增强和数据集捷径等角度改进传统闭集评测方式^[54-57]。

5.3 主动标识与被动取证的互补验证

主动水印与被动取证更适合作为相互补充的证据来源。当内容内水印指向某一平台或Content-ID，而被动取证结果明显指向另一生成模型家族时，二者之间的不一致可作为水印嫁接或伪造的风险信号；当传播过程中元数据被剥离且水印部分受损时，被动特征又可用于缩小可能的来源范围。表5汇总了代表性被动检测与来源归因方法的主要线索、泛化目标和局限。

6 外部内容凭证与监管标识

6.1 C2PA与内容内水印的协同

C2PA通过签名的内容凭证清单（manifest）记录创建者、编辑动作、上游来源资产（ingredients）、软件和声明信息，并利用哈希与数字签名验证相关记录的完整性；2026年发布的2.4版本进一步扩展了资产类型和声明能力^[58]。与内容内水印相比，C2PA更适合表达结构化编辑历史和身份声明，但内容凭证清单在截图、转码或不兼容平台传播过程中仍可能丢失。

从证据互补角度看，C2PA与内容内水印可以采用协同方式：内容凭证清单保存较完整的权利声明和编辑记录，图像内部写入相同的Content-ID或其摘要；当元数据被剥离时，可通过水印恢复索引并查询受信凭证库；凭证签名及其与当前内容的绑定仍需分别验证。

6.2 国内标识规范与开放验证

我国GB 45438—2025和《人工智能生成合成内容标识办法》已于2025年9月1日起实施，对图片等生成合成内容提出显式标识与隐式标识要求^[59-60]。受此影响，相关技术的评价对象不宜仅停留在水印恢复率，还应考虑服务提供者信息、内容编号、文件元数据、用户端展示以及平台传播后的标识完整性。

工程部署还需要处理开放验证与密钥安全之间的平衡。封闭检测器有利于保护密钥，但会限制第三方审计；验证机制完全公开，则可能增加自适应伪造攻击的风险。NoisePrints利用零知识证明保护生成seed的设计^[27]表明，可进一步研究Content-ID、公钥签名、内容水印与C2PA凭证的协同，使跨平台验证在不直接暴露秘密材料的条件下完成。

表 5 代表性生成图像被动检测与来源归因方法对比

方法	主要线索	目标	泛化重点	主要局限
UniversalFakeDetect	预训练视觉特征	真/假检测	跨GAN/扩散生成器	难以恢复权利主体与Content-ID
DIRE / AEROBLADE	扩散/自编码器重建误差	生成图像检测	跨扩散模型、低训练成本	对非扩散生成范式的适用性受限
FIRE / VIB-Net	频率重建/信息瓶颈	通用AIGI检测	未知模型、跨家族	对开放世界新生成范式的适应性仍受限
Forensic Self-Descriptions	自监督取证表征	零样本检测/开放集归因	未见生成器	主体绑定与结果解释能力有限
CLDA / 持续检测基准	扩散伪影、增量学习	通用检测/新生成器适应	域迁移、灾难性遗忘	仍需真实平台数据与严格语义控制
IAR provenance	自回归输出特征	来源归因	跨IAR模型	仍需与主动身份信息结合

7 攻击评测、关键瓶颈与发展方向

7.1 生成式攻击与标准化评测

生成式攻击已成为现代水印鲁棒性评测中不可忽视的因素。WAVES、再生成攻击以及 VINE/W-Bench 分别考察了扩散净化、对抗攻击、全局或局部编辑以及图生视频等操作对现代水印的影响^[10-11,30]。这些结果表明,仅报告 JPEG 压缩、裁剪和噪声等传统失真已不足以评价生成式版权水印在复杂传播与编辑环境中的适用性。

除水印移除外,近期研究还关注水印伪造与语义级绕过。DiffForge 在较低访问权限下估计水印分布并将相应标记嫁接到无水印内容中,从而造成潜在错误归因^[31];视觉释义攻击则通过“图像描述—重新生成”方式,在保持主要语义的同时切断原有像素标识与元数据^[12]。因此,生成式水印评测还应同时报告错误归因率、语义保持程度和攻击者访问权限。

为比较相同载荷与编辑条件下的方法表现,图 3(a)选取 W-Bench 中载荷均为 100 bit 的 4 种方法^[11]。测试分辨率为 512×512,画质指标基于 10000 幅图像统计,TPR 沿用原文在 FPR=0.1% 下对各编辑难

度的汇总结果。VINE-Base 在随机再生成下的 TPR 为 91.03%,高于 TrustMark 的 9.22%,两者 PSNR 分别为 40.51 dB 和 41.27 dB。VINE-Robust 的对应 TPR 达到 99.66%,PSNR 为 37.34 dB;两种 VINE 配置呈现了编辑鲁棒性与图像保真度的权衡。

图 3(a)也表明,性能提升具有攻击依赖性。在图生视频条件下,4 种方法的 TPR 仅为 25.44%~39.60%,与其在多种图像编辑条件下的结果存在明显差距。图 3(b)汇总 OmniGuard 论文中基于 COCO 测试集 1000 个样本、在共同编辑条件下获得的定位结果^[39],保留各方法原有配置;EditGuard 与 OmniGuard 的版权载荷分别为 64 bit 和 100 bit。在 Stable Diffusion 局部重绘后叠加退化时,EditGuard 与 OmniGuard 的 F1 分别由 0.951 和 0.961 降至 0.740 和 0.947;但在无附加退化的拼接条件下,EditGuard 为 0.928,高于 OmniGuard 的 0.913。可见,抗退化表现与无附加退化时的定位精度应分别考察。

上述两组结果来自不同基准,分别对应水印检测和像素定位,不能合并为统一性能排名。TPR 依赖检测阈值和误报约束,F1 依赖篡改掩码及定位

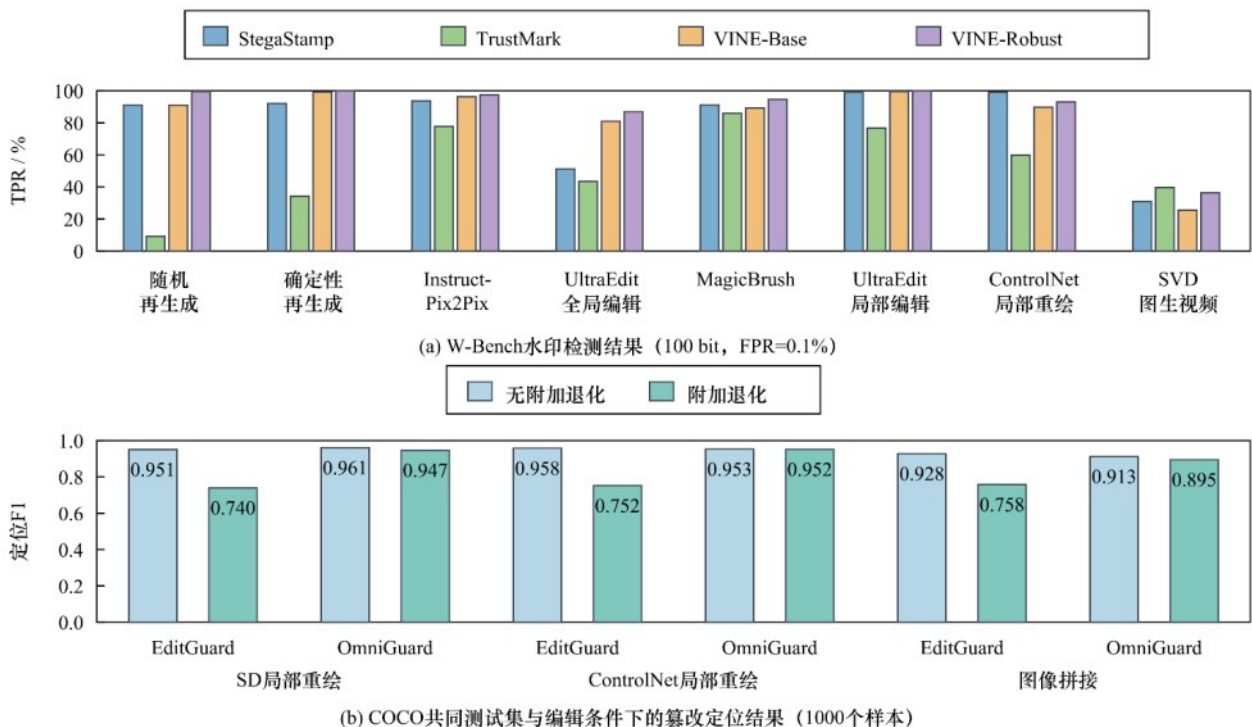


图 3 公开基准报告的水印检测与篡改定位结果

图(a)据文献^[11]表 1 绘制,各编辑项按原文不同难度汇总;图(b)据文献^[39]表 2 绘制,附加退化为随机选取 JPEG 压缩、颜色抖动、高斯噪声或椒盐噪声。两组结果分别用于检测与定位任务内的比较。

阈值；较高的 TPR 或 F1 均不等同于完整载荷恢复。为了说明式(1)的实际计算，本文进一步采用 Edit-Guard 的两种公开权重进行联合验证^[38]。

从 COCO2017 验证集固定抽取 40 幅图像，其中 8 幅用于阈值选择、32 幅用于测试，均缩放为 512×512 并嵌入 64 bit Content-ID。攻击前以 Ed25519 签署含标识、主体及发布图像哈希的记录，采用预置主体公钥验签；该记录验证不要求编辑后图像与原始字节一致。局部篡改改为 128×128 异图块拼接，再分别施加质量因子 90 和 70 的 JPEG 压缩。clean 与 degrade 权重的像素判定阈值经校准分别取 0.3 和 0.5；定位以 IoU≥0.5 为通过，无篡改图则要求预测掩码为空。异常凭证校验包括缺证、签名损坏、错误主体和标识不匹配 4 类。

表 6 中，degrade 权重在 JPEG Q90 和 Q70 下均有 32 幅图像通过定位判定，完整标识却分别仅恢复 3 幅和 0 幅，ESR 为 9.38% 和 0。相较于 clean 权重，degrade 权重在本组压缩图像上的定位结果有所改善，但标识恢复仍限制了联合验证成功率。因此，仅凭定位结果评价该任务，会遗漏来源标识失效的情况。

局部拼接组移除 8 条凭证后，两种权重的 ESR 分别由 90.63%、93.75% 降至 65.63%、68.75%。此时图像与水印恢复结果未变，联合通过率的下降来自凭证缺失。无攻击组的 ESR 低于拼接组，则与前者要求预测掩码全空有关，不能据此认为拼接改善了其性能，每种权重下的 4 类异常凭证各 32 条均

未通过校验。上述结果反映了本组图像在固定拼接与压缩条件下，标识恢复、凭证验证和定位判定之间的差异。

7.2 跨范式适配、结构化载荷与语义绑定

2023—2024 年的代表性方法主要依赖 LDM 解码器和初始噪声，而 token 级水印、IAR 来源追踪以及 GenPTW 的跨范式设计表明，生成架构差异本身已成为影响水印生命周期的重要因素^[6,14,29]。为降低底层模型迭代带来的影响，后续研究可进一步分离上层载荷与验证协议和底层生成器适配机制，使 Content-ID 与主体签名在模型架构更换后仍保持一致的证据语义。

与此同时，版权水印研究更加重视结构化载荷、多比特传输和精确恢复。Gaussian Shannon、FARI 和 SERUM 分别从通信建模、快速验证和多用户标记等方面补充相关工程能力^[28,35-36]，SEAL、StableGuard、TAG-WM 和 PAI 则将语义信息或局部完整性纳入水印验证^[13,15-16,61]。由此产生的关键问题是如何在载荷容量、低误报、篡改定位和语义绑定之间形成可解释、可复现的性能权衡。

对此，可先固定 Content-ID 及签名载荷格式，仅替换不同生成器的嵌入与提取模块，在相同载荷长度、画质约束和 FPR 下考察迁移损失；再引入来源正确而局部内容被修改、水印被嫁接但主体声明不匹配等对照，区分消息恢复与身份绑定能力。语义绑定的关键还在于划定允许编辑范围：阈值过严会拒绝正常衍生内容，过宽则可能放过语义替换，

表 6 两种公开权重下的 ESR 小规模复核结果

注：每组均为同一批 32 幅测试图像；通过数按样本计。预设缺证指移除 8/32 条签名记录。clean 与 degrade 分别为作者公开的干净条件与抗退化权重。

权重	测试条件	ID 准确恢复	签名通过	定位通过	联合通过	ESR/%
clean	无攻击	31	32	29	28	87.50
clean	局部拼接	29	32	32	29	90.63
clean	拼接+JPEG Q90	4	32	0	0	0.00
clean	拼接+JPEG Q70	1	32	0	0	0.00
clean	拼接+预设缺证	29	24	32	21	65.63
degrade	无攻击	32	32	25	25	78.13
degrade	局部拼接	30	32	32	30	93.75
degrade	拼接+JPEG Q90	3	32	32	3	9.38
degrade	拼接+JPEG Q70	0	32	32	0	0.00
degrade	拼接+预设缺证	30	24	32	22	68.75

需要结合内容相似度与错误归因率共同确定。

7.3 开放世界取证与可验证证据

被动取证的评价正在由闭集条件下的高准确率扩展到未知生成器、持续学习和语义对齐等开放世界问题。未来基准除严格分离训练与测试生成器外,还应控制真实图像与生成图像之间的语义差异,并纳入平台压缩、截图和二次编辑等常见传播操作,以降低检测器利用数据集捷径获得虚高性能的可能^[56-57]。

训练数据证据还面临从统计相关到因果使用、再到争议证明的解释跨度。更可靠的实验设计需要包含阴性对照、多种训练配置、不同数据掺入比例和统计功效分析,并与训练授权记录结合。对于公开验证机制,则需要在验证透明度与密钥泄露风险之间进行权衡,可考虑结合零知识证明或受信执行环境,使第三方在不直接获得秘密材料的情况下验证相关权利声明^[27,40-42]。

7.4 跨媒体证据连续性

生成式编辑链已由单一图像扩展到视频和3D资产等多种媒体形态。VINE已将image-to-video纳入鲁棒性基准,C2PA也在持续扩展多资产支持能力^[11,58]。后续研究需要进一步明确同一Content-ID在衍生媒体中的继承规则,包括何时沿用原有标识、何时生成新版本,以及原始资产与衍生资产的来源关系如何记录。

图3(a)中图生视频后的检测下降说明,图像域内的鲁棒性难以直接延伸到时序媒体。可从关键帧标识、跨帧冗余和父子资产标识关联入手,将逐帧检出率与视频级来源恢复率分别报告,并在裁帧、重编码和时序重排后检验关联是否仍能成立。

8 结论

本文从生成范式、版权对象和证据机制三个维度梳理相关研究。水印载体由最终像素扩展至噪声、潜变量、生成轨迹和token,验证目标逐渐涵盖结构化载荷恢复、来源归因和篡改定位,攻击模型也从传统信号失真扩展至再生成、语义编辑和水印伪造。不同技术的验证前提与证明能力各有边界,来源识别、主体绑定和训练数据使用证明不能相互替代,亦不能直接等同于内容真实或法律权属成立。

公开基准报告结果显示,所比较方法在图生视

频后的水印检测能力明显下降,定位方法在无附加退化与叠加退化条件下的相对表现也有所不同。抗退化性与定位精度需分别考察,单项指标不足以反映多项证据同时成立的情况。小规模实测中,判定判定通过的图像仍存在标识恢复失败,凭证缺失也会使图像内证据保留的样本无法完成联合验证。ESR可补充反映这类单项结果与联合结果之间的差异。后续研究应重点解决跨范式适配、语义绑定及跨媒体证据连续性问题,在内容水印、被动取证与外部凭证之间建立可独立核验的协同机制,同时考察错误归因控制与第三方复核能力。

参考文献:

- [1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets[C]//Advances in Neural Information Processing Systems 27 (NIPS 2014). Red Hook, NY: Curran Associates, 2014: 2672-2680.
- [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. Piscataway: IEEE Press, 2022: 10684-10695.
- [3] Tian K Y, Jiang Y, Yuan Z H, et al. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction[C]//Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Red Hook, NY: Curran Associates, 2024: 84839-84865. DOI: 10.52202/079017-2694.
- [4] Li T H, Tian Y L, Li H, et al. Autoregressive Image Generation without Vector Quantization[C]//Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Red Hook, NY: Curran Associates, 2024: 56424-56445. DOI: 10.52202/079017-1797.
- [5] Han C R, Li G P, Wu J W, et al. NextStep-1: Toward Autoregressive Image Generation with Continuous Tokens at Scale[C]//International Conference on Learning Representations, 2026: 93867-93889.
- [6] Jovanović N, Labiad I, Souček T, et al. Watermarking Autoregressive Image Generation[C]//Advances in Neural Information Processing Systems 38 (NeurIPS 2025). Red Hook, NY: Curran Associates, 2025: 71801-71848. DOI: 10.52202/085713-2411.
- [7] Zhu J R, Kaplan R, Johnson J, et al. HiDDeN: Hiding Data With Deep Networks[C]//Computer Vision - ECCV 2018. Cham: Springer, 2018: 682-697. DOI: 10.1007/978-3-030-01267-0_40.
- [8] Tancik M, Mildenhall B, Ng R. StegaStamp: Invisible Hyperlinks in Physical Photographs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2117-2126.
- [9] Liu X Y, Xu R H. A comprehensive survey of watermarking techniques for copyright protection and integrity verification on DNNs and generative models[J]. Discover Applied Sciences, 2026, 8: 563. DOI: 10.1007/s42452-026-08576-3.
- [10] Zhao X D, Zhang K X, Su Z H, et al. Invisible Image Watermarks Are Provably Removable Using Generative AI[C]//Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Red Hook, NY:

- Curran Associates, 2024: 8643-8672. DOI: 10.52202/079017-0276.
- [11] Lu S L, Zhou Z H, Lu J Y, et al. Robust Watermarking Using Generative Priors Against Image Editing: From Benchmarking to Advances [C]//International Conference on Learning Representations, 2025: 83902-83936.
 - [12] Dixit S, Aziz A, Bajpai S, et al. PECCVAI: Overcoming the Brittleness of AI Image Watermarking Under Visual Paraphrasing Attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2026: 24471-24480.
 - [13] Yang H X, Liu B Z, Xu X M, et al. StableGuard: Towards Unified Copyright Protection and Tamper Localization in Latent Diffusion Models[C]//Advances in Neural Information Processing Systems 38 (NeurIPS 2025). Red Hook, NY: Curran Associates, 2025: 14692-14720. DOI: 10.52202/085713-0495.
 - [14] Gan Z L, Liu C Y, Tang Y C, et al. GenPTW: Latent Image Watermarking for Provenance Tracing and Tamper Localization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2026, 40(5): 4085-4093. DOI: 10.1609/aaai.v40i5.37412.
 - [15] Chen Y Z, Ma Z H, Fang H, et al. TAG-WM: Tamper-Aware Generative Image Watermarking via Diffusion Inversion Sensitivity[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2025: 16723-16732.
 - [16] Liu Q Y, Zhang Y T, Ba Z J, et al. Attack-Resistant Watermarking for AIGC Image Forensics via Diffusion-based Semantic Deflection[C]//International Conference on Learning Representations, 2026: 49787-49813.
 - [17] 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. DOI: 10.11834/jig.240003.
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. DOI: 10.11834/jig.240003.
 - [18] 郭钊钧, 李美玲, 周杨铭, 等. 人工智能生成内容模型的数字水印技术研究进展[J]. 网络空间安全科学学报, 2024, 2(1): 13-39. DOI: 10.20172/j.issn.2097-3136.240102.
Guo Z J, Li M L, Zhou Y M, et al. Survey on digital watermarking technology for artificial intelligence generated content models[J]. Journal of Cybersecurity, 2024, 2(1): 13-39. DOI: 10.20172/j.issn.2097-3136.240102.
 - [19] 唐晓亮, 孙振. 面向扩散模型的图像水印方法研究综述[J]. 计算机应用研究, 2026, 43(9): 2575-2588. DOI: 10.19734/j.issn.1001-3695.2025.12.0525.
Tang X L, Sun Z. Survey of image watermarking methods for diffusion models[J]. Application Research of Computers, 2026, 43(9): 2575-2588. DOI: 10.19734/j.issn.1001-3695.2025.12.0525.
 - [20] 谢雪, 戚宇昂, 陈可江, 等. 面向 AIGC 视觉内容的主动防御水印综述: “防-查-抗”动态博弈视角[J]. 通信学报, 2026, 47(7): 240-252. DOI: 10.11959/j.issn.1000-436x.TXXB260122.
Xie X, Qi Y A, Chen K J, et al. Overview of active defense watermarking for AIGC visual content: a perspective of “prevention-tracing-resistance” dynamic game[J]. Journal on Communications, 2026, 47(7): 240-252. DOI: 10.11959/j.issn.1000-436x.TXXB260122.
 - [21] Ojha U, Li Y H, Lee Y J. Towards Universal Fake Image Detectors that Generalize Across Generative Models[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 24480-24489. DOI: 10.1109/CVPR52729.2023.02345.
 - [22] 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. Piscataway: IEEE Press, 2023: 22466-22477.
 - [23] Wen Y X, Kirchenbauer J, Geiping J, et al. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust[PP]. V3. (2023-07-04)[2026-09-24]. arXiv: arXiv.2305.20030.
 - [24] Yang Z J, Zeng K, Chen K J, et al. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 12162-12171.
 - [25] Gesny E, Giboulot E, Furon T, et al. Guidance Watermarking for Diffusion Models[C]//International Conference on Learning Representations, 2026: 113957-113988.
 - [26] Hu X X, Jin J Q, Li S, et al. Spherical Watermark: Encryption-Free, Lossless Watermarking for Diffusion Models[C]//International Conference on Learning Representations, 2026: 115345-115377.
 - [27] Goren N, Katzir O, Nakarmi A, et al. NoisePrints: Distortion-Free Watermarks for Authorship in Private Diffusion Models[C]//International Conference on Learning Representations, 2026: 99698-99727.
 - [28] Zhang Y, Huang H B, Zhang L J. Gaussian Shannon: High-Precision Diffusion Model Watermarking Based on Communication[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings. Piscataway: IEEE Press, 2026: 3924-3932.
 - [29] Zhao B H, Kerner L, Meintz M, et al. Data Provenance for Image Auto-Regressive Generation[C]//International Conference on Learning Representations, 2026: 54973-55009.
 - [30] An B, Ding M C, Rabbani T, et al. WAVES: Benchmarking the Robustness of Image Watermarks[C]//Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024, 235: 1456-1492.
 - [31] Dong Z P, Shuai C, Ba Z J, et al. Imperceptible but Forgeable: Practical Invisible Watermark Forgery via Diffusion Models[PP]. V1. (2025-03-28)[2026-09-24]. arXiv: arXiv.2503.22330.
 - [32] Sander T, Fernandez P, Durmus A, et al. Watermark Anything with Localized Messages[C]//International Conference on Learning Representations, 2025: 79569-79599.
 - [33] Goyal S, Bunel R, Stimberg F, et al. SynthID-Image: Image Watermarking at Internet Scale[PP]. V1. (2025-10-10)[2026-09-24]. arXiv: arXiv.2510.09263.
 - [34] 贾源, 袁得崙, 潘语泉, 等. 面向扩散模型输出的水印方法[J]. 计算机应用, 2026, 46(1): 161-168. DOI: 10.11772/j.issn.1001-9081.2025010006.
Jia Y, Yuan D Y, Pan Y Q, et al. Watermarking method for diffusion model output[J]. Journal of Computer Applications, 2026, 46(1): 161-168. DOI: 10.11772/j.issn.1001-9081.2025010006.
 - [35] Yang J D, Fang H, Zhang W M, et al. FARI: Robust One-Step Inversion for Watermarking in Diffusion Models[C]//International Conference on Learning Representations, 2026: 135290-135311.
 - [36] Kociszewski J, Jastrzębski H, Stępkowski T, et al. SERUM: Simple, Efficient, Robust, and Unifying Marking for Diffusion-based Image Generation[C]//International Conference on Learning Representations, 2026: 71380-71405.
 - [37] Nguyen-Le H H, Tran V T, Nguyen T D, et al. Transferable Multi-Bit

- Watermarking Across Frozen Diffusion Models via Latent Consistency Bridges[PP]. V1. (2026-03-19)[2026-09-24]. arXiv: arXiv.2603.20304.
- [38] Zhang X Y, Li R Y, Yu J W, et al. EditGuard: Versatile Image Watermarking for Tamper Localization and Copyright Protection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 11964-11974.
- [39] Zhang X Y, Tang Z C, Xu Z P, et al. OmniGuard: Hybrid Manipulation Localization via Augmented Versatile Deep Image Watermarking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2025: 3008-3018.
- [40] Cui Y Q, Ren J, Xu H, et al. DiffusionShield: A Watermark for Copyright Protection against Generative Diffusion Models[PP]. V1. (2023-05-25)[2026-09-24]. arXiv: arXiv.2306.04642.
- [41] Asnani V, Collomosse J, Bui T, et al. ProMark: Proactive Diffusion Watermarking for Causal Attribution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 10802-10811.
- [42] Liu Z G, Shuai C, Fan S J, et al. Harnessing Frequency Spectrum Insights for Image Copyright Protection Against Diffusion Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2025: 18653-18662.
- [43] Shan S, Cryan J, Wenger E, et al. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models[C]//32nd USENIX Security Symposium. Berkeley, CA: USENIX Association, 2023: 2187-2204.
- [44] Shan S, Ding W X, Passananti J, et al. Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models[C]//2024 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2024: 807-825.
- [45] Zhu P F, Takahashi T, Kataoka H. Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 24420-24430.
- [46] Wang S Y, Efros A A, Zhu J Y, et al. Evaluating Data Attribution for Text-to-Image Models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 7192-7203.
- [47] Feng W T, Zhou W B, He J Y, et al. AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA[C]//Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024, 235: 13423-13444.
- [48] Wang Z D, Bao J M, Zhou W G, et al. DIRE for Diffusion-Generated Image Detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 22445-22455.
- [49] Ricker J, Lukovnikov D, Fischer A. AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 9130-9140.
- [50] Zhu M J, Chen H T, Yan Q Y, et al. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image[C]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Red Hook, NY: Curran Associates, 2023: 77771-77782. DOI: 10.52202/075280-3398.
- [51] Chu B L, Xu X, Wang X, et al. FIRE: Robust Detection of Diffusion-Generated Images via Frequency-Guided Reconstruction Error[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2025: 12830-12839.
- [52] Zhang H F, He Q H, Bi X L, et al. Towards Universal AI-Generated Image Detection by Variational Information Bottleneck Network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2025: 23828-23837. DOI: 10.1109/CVPR52734.2025.02219.
- [53] Nguyen T D, Azizpour A, Stamm M C. Forensic Self-Descriptions Are All You Need for Zero-Shot Detection, Open-Set Source Attribution, and Clustering of AI-generated Images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2025: 3040-3050.
- [54] 李美玲, 钱振兴, 张新鹏. 人工智能生成图像检测技术综述[J]. 中国图象图形学报, 2026, 31(1): 13-44. DOI: 10.11834/jig.250053.
- Li M L, Qian Z X, Zhang X P. Survey on artificial intelligence-generated image detection[J]. Journal of Image and Graphics, 2026, 31(1): 13-44. DOI: 10.11834/jig.250053.
- [55] 袁程胜, 陈金瑞, 曹巍, 等. 基于扩散伪影对比学习的生成式图像检测方法[J]. 电子学报, 2026, 54(1): 248-261. DOI: 10.12263/DZXB.20250663.
- Yuan C S, Chen J R, Cao Y, et al. Generative Image Detection Based on Diffusion Artifact Contrast Learning[J]. Acta Electronica Sinica, 2026, 54(1): 248-261. DOI: 10.12263/DZXB.20250663.
- [56] 王亚斌, 洪晓鹏, 黄智武. 面向 AI 生成图像持续检测基准数据集与框架研究[J]. 中国图象图形学报, 2025, 30(11): 3438-3450. DOI: 10.11834/jig.250167.
- Wang Y B, Hong X P, Huang Z W. Benchmark dataset and framework for continual AI-generated image detection[J]. Journal of Image and Graphics, 2025, 30(11): 3438-3450. DOI: 10.11834/jig.250167.
- [57] 杨锦旭, 王志远, 赵鹏铨, 等. 面向人工智能生成图像检测的语义对齐数据集与基准研究[J/OL]. 中国图象图形学报, 2026: 1-21[2026-09-24]. <https://doi.org/10.11834/jig.260297>.
- Yang J X, Wang Z Y, Zhao P C, et al. Semantic alignment dataset and benchmark for AI-generated image detection[J/OL]. Journal of Image and Graphics, 2026: 1-21[2026-09-24]. <https://doi.org/10.11834/jig.260297>.
- [58] Coalition for Content Provenance and Authenticity. C2PA Technical Specification, Version 2.4[EB/OL]. (2026-04) [2026-09-24]. <https://spec.c2pa.org/specifications/specifications/2.4/index.html>.
- [59] 国家市场监督管理总局, 国家标准化管理委员会. GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法[S]. 北京: 中国标准出版社, 2025.
- State Administration for Market Regulation, Standardization Administration of China. GB 45438—2025 Cybersecurity technology—Labelling method for content generated by artificial intelligence[S]. Beijing: Standards Press of China, 2025.
- [60] 国家互联网信息办公室, 工业和信息化部, 公安部, 等. 人工智能生成合成内容标识办法[Z]. 2025-03-14 发布, 2025-09-01 施行.
- Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, et al. Measures for the Labelling of AI-Generated Synthetic Content[Z]. Issued on 2025-03-14, effective on 2025-09-01.
- [61] Arabi K, Witter R T, Hegde C, et al. SEAL: Semantic Aware Image Watermarking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2025: 16196-

16205. DOI: 10.1109/ICCV51701.2025.01503.

[作者简介]



王海龙 (1995-), 男, 博士生, 暨南大学在读博士生, 主要研究方向为数字图像处理、计算机视觉、人工智能安全, 尤其是后门攻击、对抗学习以及图像数据版权保护。



谭武征 (1976-), 男, 博士, 暨南大学研究员, 主要研究方向为数字资产安全、AI 安全、工业控制安全以及农业物联网安全。



张子欣 (2002-), 女, 硕士生, 暨南大学在读硕士生, 主要研究方向为应用密码学、工业物联网安全、分布式身份认证以及访问控制。