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

谢 雪，戚宇昂，陈可江，张卫明，俞能海
（1. 中国科学技术大学网络空间安全学院，安徽 合肥 230002）

摘 要：数字水印技术是应对人工智能生成内容（AIGC）所引发的视觉信任危机的关键支撑技术。现有数字水印技术综述文献大多局限于静态分类视角，难以有效揭示水印防御机制与攻击手段之间持续升级的动态对抗演化规律。为此，本文从安全博弈论视角出发，面向AIGC内容全生命周期，基于对数字水印技术现状的分析，构建“防-查-抗”三元分析框架。在此框架下，系统梳理并深度剖析了水印技术在三大核心维度上的作用机理：其一，通过阻断非授权数据挖掘实现生成过程的源头管控；其二，面向高维合成伪造内容的版权溯源与身份认证；其三，抵御深度对抗擦除攻击以保障嵌入水印的鲁棒持久性。最后，围绕跨模态泛化能力不足、对抗攻击模型持续演进等核心挑战，展望了主动防御水印技术的未来演进方向。本文旨在为AIGC主动防御体系的理论研究与工程实践提供参考，以期促进可信AIGC生态的安全可持续发展。

关键词：人工智能生成内容；数字水印；主动防御；安全博弈
中图分类号：TP309.2

Proactive Defense Watermarking for AIGC Visual Content: A Perspective of "Prevention-Tracing-Resistance" Dynamic Game

Xie Xue, Qi Yuang, Chen Kejiang, Zhang Weiming, Yu Nenghai
1. School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230002, China

Abstract: Digital watermarking constitutes a fundamental technical pillar for addressing the visual trust crisis engendered by Artificial Intelligence-Generated Content (AIGC). Existing digital watermarking survey literature is predominantly confined to a static classification perspective, rendering it inadequate for elucidating the continuously escalating dynamic adversarial co-evolution between watermarking defenses and attack methodologies. To this end, grounded in the perspective of security game theory, and aiming at analyzing the current status of digital watermarking technologies, this paper proposes a "Prevention-Tracing-Resistance" ternary analytical framework oriented toward the full lifecycle of AIGC. Within this framework, the core mechanisms of watermarking technology are systematically examined across three principal dimensions: (1) interdicting unauthorized data mining to achieve source-level governance of the content generation process; (2) copyright provenance tracing and identity authentication for high-dimensional synthesized forgery content; and (3) resisting deep adversarial erasure attacks to ensure the robust persistence of embedded watermarks. Finally, the critical challenges pertaining to insufficient cross-modal generalization capability and the continuous escalation of adversarial attack models are analyzed, and future evolutionary trajectories of proactive defense watermarking are prospected. This paper aims to provide a systematic reference framework for both theoretical inquiry and engineering practice in next-generation AIGC proactive defense systems, thereby fostering the secure and sustainable development of a trustworthy AIGC ecosystem.

收稿日期：XXXX–XX–XX；修回日期：XXXX–XX–XX
通信作者：戚宇昂，qiyuang@mail.ustc.edu.cn
基金项目：国家自然科学基金资助项目（No.62472398）；
Foundation Items：The National Natural Science Foundation of China (No.62472398)

Keywords: Artificial Intelligence-Generated Content (AIGC), Digital Watermarking, Proactive Defense, Security Games

0 引言

随着以扩散模型[1]和生成式预训练大模型为代表的深度生成范式的持续演进，人工智能生成内容（AIGC）技术在视觉特征的高保真重构与语义级别的可控生成上取得了突破性进展。然而，高维视觉特征的精准拟合能力也衍生出极其严峻的认知安全威胁。深度伪造（Deepfake）生成技术的开源化，使得攻击者能够以极低的成本合成具有高度视觉欺骗性的虚假图像。这种技术滥用不仅从根本上颠覆了传统社会基于视觉感知所构建的媒介信任机制，更诱发了从个体肖像权侵犯、多媒体版权争议乃至国家政治舆论操纵的多维安全危机[2]-[3]。

针对AIGC技术引发的数字信任危机，国际社会与我国相继出台了一系列治理框架，如《生成式人工智能服务管理暂行办法》及《全球人工智能治理倡议》等，试图从法律规制层面规范生成内容的标识与溯源。然而，现有的技术监管体系多依赖于事后的被动检测网络（如基于空域伪影或频域统计特征的二分类器）。由于高质量AIGC模型能够极大地缩小真实自然图像与合成图像在像素域及高频纹理域的分布差异，传统的被动取证模型正面临着严重的跨域泛化能力显著下降与检测决策边界模糊困境。近期的研究进一步印证了这一危机。Yan等[4]构建的DF40基准测试表明，现有基于单一分布训练的Deepfake检测网络在面对未知的高维生成伪造时，泛化能力会出现大幅下降；Lin等[5]也通过实验指出，现有的深伪检测技术在跨域应用时存在严重的公平性问题与性能衰退。不仅如此，针对AIGC视频生成的快速爆发，Lanzino等[6]和Zhao等[7]强调了传统被动检测在面对高维时序伪造时所暴露的算力瓶颈与数据偏见。

在此背景下，将溯源凭证前置化嵌入到内容生成过程中，构建具有主动防御和确定性溯源能力的数字水印体系，已成为维护多媒体数字生态可信可控发展的关键技术路径。然而，现有的附加型隐蔽水印同样面临着极高的被伪造与归因风险。正如Zhou等[8]在2025年指出，与内容无关的水印一旦被逆向，攻击者便可轻易将其转移至其他虚假图像上，导致严重的恶意误归因；近期发表的WM-Copier攻击模型[9]亦从实验层面系统验证了上述安全隐患的现实可行性。此外，基于扩散模型的重生成技术——如Fu等[10]的研究——可在视觉无损的情况下实现对现有嵌入水印的完全去除。上述攻击路径的系统化与低成本化，相关研究开始向全面安全的主动防御过渡。正如Cao等[11]在2025年最新发布的AIGC水印综述中所强调的，未来的数字生态必须将高维重构与对抗安全性提升至核心地位。

数字水印技术通过在视觉载体或模型参数中隐蔽嵌入特异性标识信息，已成为AIGC版权保护与溯源的核心支撑。近年来，针对AIGC水印技术的重构与优化涌现出大量研究，并产生部分综述文献。然而，现有综述在技术体系的划分上普遍受限于静态视角。具体而言，部分研究[12]依据水印嵌入的时间阶段，将其划分为前置嵌入、联合生成嵌入与后置嵌入；另有文献基于保护对象的物理形态，将其归类为面向模型确权的水印、面向生成内容溯源的水印以及面向训练样本保护的水印[13]。此类静态分类法则本质上将数字水印视为一种依附于载体的被动标签，割裂了AIGC生态中攻防双方持续升级的对抗本质，因而难以深刻揭示新一代水印技术在对抗恶意知识蒸馏、高维潜空间伪造及深度网络擦除攻击时的演进机理与博弈规律。

鉴于现有研究框架的上述局限，本文立足于主动取证与安全博弈理论，重塑了AIGC视觉内容安全体系中数字水印技术的演进路线。为此，本文提出面向视觉内容全生命周期的“防-查-抗”动态博弈与主动防御框架。与传统将数字水印视作单向嵌入的被动取证凭证的范式不同，本文的创新之处在于将水印重新定义为防御方在多维特征空间内，针对利用型攻击者与破坏型攻击者所采取的主动防御策略集合。沿着这一动态演进的主线，本文系统展开了三大核心维度的技术剖析：其一，面向AI资产安全的主动防御博弈，揭示特征扰断与后门触发机制在阻断非授权知识提取中的核心作用原理；其二，面向内容取证的溯源鉴伪博弈，阐明潜在空间融合与盲提取技术在高维生成分布中实现同构嵌入的内在机理；其三，面向恶意破坏的主动对抗博弈，系统归纳水印技术在深度网络擦除攻击与跨物理域破坏场景下，基于对抗噪声融合、异构特征解耦及自适应强度调节的鲁棒强化机制体系。

综上所述，本文的贡献可归纳为以下三个方面：

1. 突破传统分类局限，提出主动防御的“防-查-抗”动态博弈框架：本文摒弃了基于嵌入阶段与保护目标的传统静态分类逻辑，将博弈论视角系统性地引入AIGC水印体系的架构设计之中。在该框架下，数字水印从依附于载体的被动身份标识，提升为防御方在多维特征空间内应对非授权知识提取与恶意破坏等异质威胁实体的主动防御工具。

2. 解析多维博弈态势下的水印核心机理与策略演化：本文深入剖析了数字水印在三大对抗维度——面向AI资产安全的主动防御博弈、面向内容取证的溯源鉴伪博弈，以及面向恶意破坏的主动对抗博弈——下的技术作用机理与策略演进规律，为水印系统的对抗性设计提供了理论依据。

3. 提炼了跨模态生成时代的对抗瓶颈与演进路线：在全面梳理多维博弈态势与演进规律的基础上，本文分析了生成大模型时代主动对抗水印在多模态融合与跨域迁移场景下所面临的性能瓶颈，并为构建具备跨域泛化能力与自适应鲁棒性的主动取证防御体系提供了理论参考与方向性探索。

总体而言，本文提出的“防-查-抗”三元框架突破了静态分类的局限，为AIGC全生命周期的攻防演进提供了系统性的理论剖析工具。需要客观指出的是，面对极度复杂的跨模态生成场景，该框架在各博弈阶段的数学安全边界量化上仍面临挑战，这也是当前整个主动防御领域亟待突破的技术局限。

1　面向AIGC视觉内容的三元博弈模型与主动防御框架

在真实的AIGC开放生态中，数字水印全面暴露于具有高度定向性与蓄意对抗性的深度网络攻击之下。为精准刻画攻防双方持续升级的动态对抗本质，本文将博弈论视角引入AIGC水印体系，构建了面向视觉内容全生命周期的“防-查-抗”主动防御框架（如图1所示）。在该框架下，数字水印跃升为防御方与不同威胁实体展开动态博弈的主动防御工具，并据此划分为阻断型博弈（防）、溯源型博弈（查）与反制型博弈（抗）三个核心维度。

1.1　三元博弈角色定义

在AIGC视觉内容的流转生命周期中，攻防双方的策略处于持续的动态对抗之中。本文将该生态中的参与者抽象为以下三个核心博弈实体：

防御方：涵盖数据所有者、模型开发者与监管机构。核心目标是实现AI资产确权、高维内容溯源，并保障水印在极端对抗场景下的鲁棒提取。

利用型攻击者：意图非法窃取高价值的数据与模型资产。具体表现为利用自动化脚本大规模爬取视觉数据，或借助参数微调（如LoRA、知识蒸馏）非法克隆目标大模型的生成能力。

破坏型攻击者：意图篡改或彻底剥离溯源特征。此类实体通常利用生成先验构建高精度伪造内容以规避检测，或定向部署深度擦除网络破坏版权认证机制。

1.2　“防-查-抗”主动防御架构的阶段划分与核心机理

针对上述三元博弈角色所对应的差异化威胁模型，传统被动嵌入策略已难以满足AIGC内容全生命周期的安全防护需求。防御方必须在不同特征空间内，依据攻击者的行为意图与攻击路径动态调整反制策略。基于此，本文将主动防御架构划分为防-查-抗三个核心维度，各维度的技术机理如下。

(1) 面向非授权挖掘的阻断型水印机制(防)

在应对利用型攻击者时，防御方的首要任务是实现AI资产的访问控制。在这一阶段，水印技术从视觉隐蔽标识转变为主动扰断与触发拦截机制。一方面，通过在训练样本中注入对抗性特征，使得未经授权的模型微调行为发生梯度偏离与性能退化；另一方面，将水印构建为模型权重中的后门触发集，在遭遇非授权调用或克隆时主动阻断生成输出，从而实现对底层数据与参数资产的确权保护

(2) 面向高维生成内容的内生溯源与盲提取机制(查)

在对海量生成内容进行溯源确权时，防御方面临的核心挑战是AIGC模型强大的特征重构与覆盖能力。为此，水印的嵌入范式从生成后附加全面转向生成中同构。具体而言，防御方将溯源特征作为条件引导项或初始噪声融入潜在空间，使其与视觉语义在生成轨迹上实现深度耦合。同时，为适应复杂网络环境，防御方引入跨模态与跨媒介的零水印范式，通过提取图像固有的拓扑不变量或边界表征，构建脱离载体像素的溯源凭证，从而实现高置信度的真伪鉴别。

(3)　面向恶意擦除与跨域破坏的主动对抗与鲁棒强化机制(抗)

面对破坏型攻击者时，攻防双方进入了特征空间的直接对抗阶段。在这种抵抗环境下，传统的鲁棒机制向主动反制与自适应强化演进。该阶段的核心机理涵盖三个方面：一是通过对抗表征的联合建模，主动误导并瓦解恶意擦除网络的注意力机制；二是通过异构特征空间的解耦优化，消除对抗噪声与水印分布间的相互干涉；三是引入自适应感知机制，根据载体信道状态动态分配局部嵌入强度，从而在复杂跨域破坏中实现视觉质量与抗重构能力的动态平衡。

2　面向非授权挖掘的阻断型博弈水印技术

在AIGC视觉内容的开放生态中，利用型攻击者常通过自动化工具大规模爬取开源视觉数据，或利用低秩自适应等参数微调技术对高价值生成大模型进行恶意克隆。面对这种以极低成本窃取核心数字资产的行为，传统的被动型水印往往只能在侵权事实发生后提供事后追溯凭证，无法在源头遏制资产流失。因此，阻断型博弈水印技术立足于主动防御视角，通过在视觉数据样本或生成模型参数中前置注入对抗性扰动与隐蔽触发机制，使得非授权的特征提取与模型训练过程发生严重偏离甚至完全失效，从而在源头构建起抵御恶意挖掘的防御机制。本节将从视觉数据的特征扰断与生成模型资产的后门触发两个维度，对现有的阻断型水印技术展开探讨。

2.1　从被动白盒确权到主动黑盒阻断的对抗演进

早期AIGC确权保护的策略主要集中于对模型内部结构进行直接修改的白盒水印技术。例如，Uchida 等[14]首次提出在深度神经网络权重参数中嵌入签名信息的白盒水印方案，奠定了模型版权保护的早期技术基础。随后，Darvish 等[15]提出了DeepSigns框架，通过在各网络层激活图的概率密度函数中编码签名，试图进一步提升水印的抗删除能力。

然而，随着攻防博弈的升级，这类被动机制的局限性逐渐暴露。Wang 等[16]的研究指出，上述白盒方案容易受到针对性的水印检测与覆写攻击。更为关键的是，白盒水印在验证阶段要求验证方对可疑模型的内部结构与参数具备完整的访问权限，而这在现实的黑盒交互或非授权克隆场景中往往不具备可行性。为克服上述博弈劣势，防御方的策略开始向数据源头的特征扰断以及黑盒条件下的后门触发转移，由此推动了具有主动防御性质的阻断型博弈机制的成熟。具体而言，针对防御方不同类型数字资产的保护诉求，该阶段的阻断策略主要演化为两个分支：一是将保护客体聚焦于底层的“视觉训练数据”，通过特征扰断阻断非授权的数据抓取与恶意学习（详见2.2节）；二是将保护客体聚焦于高价值的“生成模型权重”，利用后门触发机制抵御

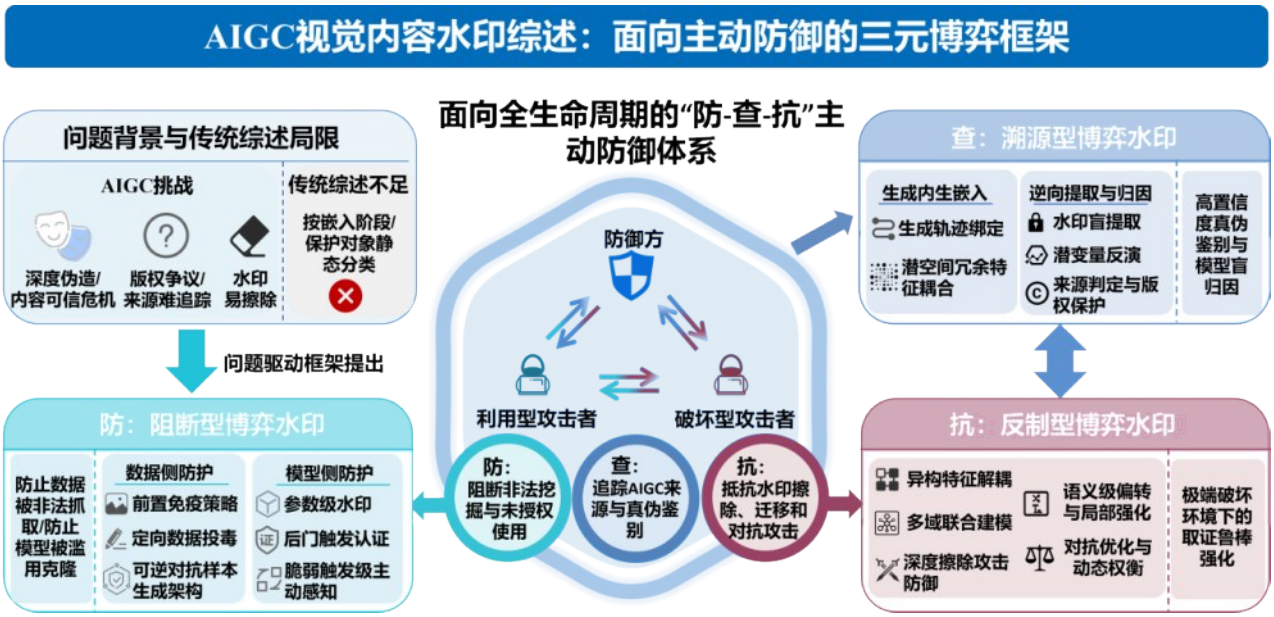

图1　面向AIGC视觉内容的三元博弈模型与主动防御框架

针对大模型资产的非法微调与知识克隆（详见2.3节）。

2.2　视觉训练数据的特征扰断

在防御利用型攻击者对海量视觉数据进行非法爬取与恶意利用的对抗中，水印机制由静态隐蔽标识转变为具有主动对抗能力的防御策略。其核心机理是借助对抗样本生成技术，在视觉载体的像素域施加人眼不可察觉的约束性扰动，使得利用这些数据进行训练的非授权生成网络在潜空间特征映射时发生严重的语义坍塌或风格偏移。

为了应对利用型攻击者对视觉资产的逆向深度解析，Hayes等[17]基于生成式对抗网络设计了端到端的隐蔽特征嵌入架构，通过对抗学习机制使得水印在视觉潜空间中获得极强的隐藏与抗解析能力，为防御方在博弈中构建高维特征防御提供了早期范式。进一步地，面对扩散网络强大的特征重构与恶意篡改能力，Salman等[18]提出了一种针对扩散模型的前置免疫策略，通过向视觉载体中注入隐蔽的对抗性扰动主动干扰扩散网络的前向与反向传播过程，有效遏制了模型对目标图像的非法特征操纵，大幅提升了恶意图像编辑的计算代价。

针对艺术家个人原创视觉作品频遭文生图模型非法风格克隆的挑战，Shan等[19]设计了代表性防御工具Glaze。该工具通过求解特定视觉风格的最小对抗扰动，在严格保障图像感知质量的前提下，迫使图像在深度特征维度向预设的虚假风格发生偏移。当利用型攻击者将这些携带扰动水印的画作输入至微调网络时，生成大模型将被误导并学习到错误的特征映射，从而在模型训练阶段实现了防御性阻断。

此外，为了保障高价值视觉数据在公共网络环境中的安全流通，Chen等[20]构建了图像数据集保护网络。该框架主动将目标图像集转化为包含对抗特征的防御形态，导致恶意攻击者利用此类数据训练出的盗版模型性能急剧退化。针对视觉图像极易被非法收集用于训练模型的技术威胁，Zeng等[21]则引入了一种可逆对抗样本生成架构。该机制通过对原始图像施加隐蔽的特征扰动来构建防御性数据集，使得未经授权的深度网络在进行视觉分类或特征提取时产生严重误判，主动阻断了非法的特征提取；同时，授权方仍可通过逆向映射机制无损恢复原始的干净高维特征。

面对更为隐蔽的提示词级别窃取攻击，Shan等[22]IEEE S&P 2024会议上提出了定向数据投毒工具Nightshade，专门针对文生图大模型的概念学习机制展开攻击。该工具仅需不足100个隐蔽毒化样本，即可彻底破坏目标模型对特定概念的生成能力，有效维护了版权所有者的合法权益。在此基础上，Wang等[23]进一步证实了越强大的生成扩散模型对这类后门注入越敏感，无需改变大模型的微调流程即可诱发其输出侵权图像。为了进一步提升主动防御的威慑力，Souri等[24]利用引导扩散网络，实现了从零基础自动生成具有极强致毒性的抗挖掘样本。此外，针对概念级别的恶意克隆，Lei等[25]在2024年提出了概念水印架构，通过在视觉内容中嵌入抗微调的扰动，主动抵御针对特定艺术风格和人物特征的非法定制化生成。

2.3　生成大模型资产的后门阻断

生成大模型的训练通常依赖庞大的算力集群与海量私有数据，其模型权重本身即为极具商业价值的核心资产。面对利用型攻击者通过模型微调进行的知识蒸馏与克隆提取，阻断型水印通过将特异性标识与模型生成机制深度绑定，构建基于特定触发条件的后门阻断机制。在正常交互下模型输出高质量视觉内容，而一旦遭受非法窃取或触及预设条件，模型将强制输出预设的确权图像或发生功能阻断。

在保护生成对抗网络参数资产的博弈中，Ong等[26]提出了基于正则化约束的主动保护方案，迫使侵权模型在接收特定触发输入时强制输出版权水印，实现了对非法克隆的源头阻断。类似地，Qiao等[27]设计了基于触发集的后门保护策略，将包含水印的校验样本与训练数据混合微调，使版权方可通过特异性密钥样本在黑盒条件下唤醒受控模型的所有权特征。跳出传统白盒水印高度依赖网络内部结构公开的限制，Zeng等[21]将隐写机制与后门注入策略融合，在目标网络参数分布中隐式植入触发后门，使得非授权提取的模型在面对特定输入时主动暴露溯源标识。同时，Li等[28]的研究也证实，将不可见水印深度耦合至神经网络的参数空间内，能够为深度学习模型这一新型资产构建持久且鲁棒的版权保护机制。

针对模型流转中可能遭遇的恶意微调威胁，Yin等[29]提出了基于脆弱触发集的主动防御架构，

一旦大模型参数被攻击者非法篡改，后置验证网络中触发样本的准确率将显著下降，建立起对模型恶意修改的高度敏感感知机制。随着大规模文生图模型的快速迭代，基于特定提示词的定制化微调为非授权的模型克隆提供了便利，成为大模型资产流失的主要途径，Zhao 等[30]提出了一种黑盒确权防御范式，利用微调策略将版权信息编码至分布先验中，防御方仅需利用罕见的文本提示即可强制触发侵权网络输出确权图像。最后，为遏制利用开源模型模仿特定视觉艺术的侵权行为，Luo 等[31]构建了基于验证的反制体系，将不可见水印深度耦合至艺术作品中，使得水印特征在被恶意爬取微调时隐式迁移至侵权模型深层权重，为大模型滥用与视觉风格窃取提供了坚实的取证依据与维权支撑。

综上所述，阻断型博弈水印技术在 AIGC 视觉内容生命周期的早期构成了第一道安全防线。无论是针对训练数据的特征扰断，还是面向模型权重的后门阻断，其核心逻辑均在于通过增加非授权挖掘的计算代价与特征偏离度来实现主动防御。然而，在开放世界中，防御方无法完全杜绝所有非授权生成行为的发生。当未经授权的高维视觉内容已经生成并进入网络传播渠道时，前置的阻断机制便触及了其防御边界。因此，如何在脱离了原始模型环境的海量视觉生成物中，以高置信度提取溯源凭证并完成真伪鉴别，构成了攻防博弈的下一个核心命题。这一需求驱动水印技术从源头阻断的“防”向生成过程中内生溯源的“查”跨越演进，进入面向高维生成重构的同构型博弈阶段。

3 面向高维伪造生成的溯源型博弈水印技术

当未经授权的模型绕过前置阻断机制，或利用型攻击者借助开源大模型生成海量虚假视觉内容并发布至公开网络时，防御策略必须从源头的访问控制转向生成后的精准确权与鉴伪。在这一溯源博弈阶段，防御方面临的主要对手是那些企图隐匿内容真实来源、利用高维伪造技术逃避版权监管与真伪追溯的利用型攻击者。面对这一威胁，防御方需要解决的核心问题是：如何在脱离原始生成环境的条件下，从高维视觉载体中高置信度地提取溯源凭证。本节将首先梳理溯源水印从后置附加向内生同构的博弈演进逻辑，进而从潜在空间的特征融合与逆向反演盲提取两个维度，详细阐述面向高维伪造生成的溯源型博弈机理。

3.1 从后置附加到内生同构的对抗演进

早期 AIGC 视觉内容溯源的工作主要沿用传统多媒体水印的后处理范式，即在视觉内容生成之后，通过外置模块的形式将标识嵌入空间域或变换域。例如，在面向扩散模型等生成图像的保护中，早期普遍采用基于离散小波变换与离散余弦变换的 DWT-DCT[32] 以及 DWT-DCT-SVD[33] 算法；同时，Sivananthamaitrey 等[34] 与 Kumar 等[35] 也验证了利用 SWT、DWT 及 SVD 等多域联合变换可以在视觉载体中嵌入多重取证信息。此外，工业界（如 Google DeepMind 发布的 SynthID 工具）以及基于深度学习的 RivaGAN[36] 等机制，均试图在不显著影响视觉感知质量的前提下，将溯源信息隐蔽地附加于生成图像之上。然而，在动态对抗环境中，这类外置水印机制表现出明显的滞后性与脆弱性。由于附加式水印独立于 AIGC 模型的内在生成机制，其在特征空间中仅呈现为浅层的附加高频噪声。Jiang 等[37] 的研究表明，攻击者能够通过添加微小的对抗性扰动或常规的图像后处理操作，在视觉无损的状态下有效规避自动检测系统并剥离水印。传统的防御策略在面对高维生成的对抗重构时逐渐失效。正是为了突破这一局限性，防御方的溯源策略开始向生成过程内部转移，旨在通过将溯源凭证与高维视觉语义深度绑定，提升提取的可靠性与抗破坏能力。

3.2 潜在空间的特征融合与内生嵌入

在面向高维图像生成的同构型博弈中，防御方不再依赖于图像生成后的像素级修改，而是将水印特征直接内生于扩散模型或生成对抗网络的特征空间中。这种内生嵌入机制使得水印分布与视觉图像的核心特征紧密耦合，显著提升了内容溯源的可靠性。

为了实现无损且隐蔽的特征融合，部分研究致力于挖掘大模型潜空间自身的特征冗余度。Bui 等[38] 研究指出，扩散模型等生成模型的潜在表示空间允许一定强度的抗噪扰动。利用这一网络特性，他们借助自编码器架构在潜空间中实现了高容量水印特征的冗余嵌入，从而在不降低视觉生成保真度的前提下，构建了稳健的内生确权凭证。沿着这一维度，Xiong 等[39] 设计了一种端到端的特征融

合架构，在基于扩散模型的图像生成过程中，将预定义的消息矩阵与中间特征输出进行联合编码，从网络深层实现了高维视觉特征与溯源信息的深度同构演化。

除了利用潜空间冗余，通过网络微调直接干预生成轨迹成为另一种主流策略。针对扩散模型的生成确权，Fernandez等[40]通过微调图像解码器植入模型签名，迫使模型在解码阶段隐式地将水印信息固化。在此基础上，Cui等[41]设计的Diffusion-Shield架构进一步优化了融合策略，将特定的版权编码转换为视觉空间扰动并在微调期间注入，使得大模型在迭代去噪中主动学习并重现这些标识。为了克服上述微调策略带来的算力开销与生成质量受损问题，近期的研究开始向无训练和无损耗方向演进。例如，Yang等[42]提出了Gaussian Shading机制，将水印信息映射至标准高斯分布的特征空间中，从数学理论上严格证明了其在实现盲提取的同时，能够做到对模型生成质量的零损耗。为了提升在复杂大模型架构中的适配性，Zhang等[43]提出了MarkPlugger框架，通过正交空间的附加融合策略，在完全不重新训练潜扩散模型的约束下，实现了高泛化性的即插即用水印嵌入。同时，为平衡水印的隐蔽性与提取召回率，T2SMark机制[44]引入了基于尾部截断采样的两阶段融合方案，有效规避了单域映射的局限。值得一提的是，这种在高维空间寻找不变特征的同构思想，不仅是对传统零水印架构[45]非破坏性认证理念的继承，更是在AIGC时代将其跃升至生成语义空间的高阶演进。

3.3　逆向扩散反演与盲提取机制

同构型博弈在溯源阶段的另一核心优势在于其盲提取能力。在真实的鉴伪溯源场景中，监管方通常无法获取原始的无水印图像或确切的模型输入参数。因此，利用生成模型本身的数学可逆性或对偶特征空间，实现无需原始载体辅助的水印盲提取，成为鉴伪博弈的关键突破点。

Zhu等[46]提出的HiDDeN模型首次确立了端到端深度学习水印盲提取的范式，证明了通过神经网络隐式拟合能够摆脱对原始视觉载体的依赖。在早期的生成对抗网络（GAN）溯源博弈中，Albright与McCloskey[47]率先探索了基于反演的生成源盲归因机制。该方法通过将生成的人脸图像逆向映射回生成器的潜在编码空间，寻找最匹配的潜在特征，在无需任何附加输入信息的条件下，实现了对伪造图像的精准模型盲归因。随着生成架构向更深层演进，Asnani等[48]进一步将逆向工程引入盲提取中，通过对生成图像进行指纹估计与逆向解析，成功盲推断出生成模型的网络架构与超参数，为在开放世界中提取未知大模型的高维指纹凭证提供了全新路径。此外，针对潜空间特征的高效提取问题，Bui等[38]设计了一种专注于生成潜空间的自编码器盲提取架构，将溯源信息作为隐变量的特征偏移量进行同构编码，该模型参数量小、计算代价低，在面对复杂对抗时展现出极强的特征恢复与盲解码能力。

在扩散生成机制的主流架构下，常微分方程的可逆性为盲提取提供了更为坚实的数学基础。Wen等[49]利用了去噪扩散隐式模型的反演可逆性提出了一种隐蔽指纹方案Tree-Ring。为进一步扩展盲提取的适用边界与精度，Li等[50]从对偶映射的视角出发提出了镜像扩散网络模型。该方法在由镜像映射构建的对偶空间内学习视觉数据的扩散与去噪分布，在严格的凸约束集合上实现了生成图像中高维隐蔽水印的精确量化与盲提取。李莉等[51]通过引入基于耦合变换的精确反向扩散，可以更加准确地重建初始噪声向量，提升水印提取的准确性。然而，不可忽视的是，尽管基于扩散反演（如DDIM反演）的盲提取机制在理论上摆脱了对原始视觉载体的依赖，但在实际工程落地中，其多步迭代的反演过程不可避免地面临着极高的算力开销与时间延迟。这种高昂的计算成本构成了该类技术在要求高并发、实时取证场景下的核心应用瓶颈。

综上所述，同构型博弈水印技术通过潜在特征融合与逆向反演提取，在AIGC内容的生成与传播过程中建立了隐蔽且稳固的溯源机制。这种将水印与生成分布深度同构的策略，不仅有效抵御了常规的信道噪声，更赋予了防御方在脱离原始生成环境条件下的盲提取确权能力。然而，随着深度学习对抗技术的持续演进，破坏型攻击者开始部署更为专业的对抗性扰动与深度擦除网络，试图在不破坏原始图像语义的约束条件下强制干扰或剥离这些内生水印。当攻防对抗从生成阶段的特征同构升级为针对溯源凭证的恶意剔除与破坏时，水印技术便进入了反制型博弈阶段。这就要求水印防御体系在“查”的基础上进一步向“抗”演进，以在极端恶

意破坏的环境下维持取证凭证的鲁棒性。

4　面向深度擦除的反制型博弈水印技术

随着同构型水印在AIGC溯源中的广泛应用，攻防博弈进入了更为复杂的阶段。破坏型攻击者不再局限于添加随机信道噪声或进行简单的几何变换，而是开始利用强大的生成先验部署深度擦除与重构网络。在这一极端约束下，水印技术必须从单一的真伪鉴别向主动的对抗与强化演进，以在恶意破坏的环境下维持取证线索的鲁棒性。本节首先梳理了从静态鲁棒到动态对抗的演进逻辑，进而从异构特征解耦与多维联合建模、语义级偏转与自适应对抗强化两个维度展开面向深度擦除的反制型水印的探讨。

4.1　从静态鲁棒到动态对抗的演进

传统数字水印的鲁棒性设计多针对JPEG压缩、高斯滤波等静态的数字信号处理操作。然而，随着扩散模型图像生成与编辑能力的跃升，攻击者开始利用生成式重构或深度去噪技术实施定向擦除。Guo等[52]在2026年的研究中证明，基于扩散模型的图像重生成攻击能够在保持视觉语义无损的前提下，将多种最先进的鲁棒水印检测率降至接近随机猜测的水平。此外，Shamshad等[53]在NeurIPS的隐蔽水印擦除挑战赛中，进一步展示了结合扩散先验与对抗性优化的黑盒攻击能够实现近乎完美的水印剥离。

面对此类掌握先进高维重构能力的破坏型实体，传统的静态防御策略面临失效。这种技术代差迫使防御方必须将对抗学习与特征解耦机制引入网络设计中，促成了水印技术向反制型博弈的范式演进。该演进旨在通过主动瓦解恶意擦除网络的注意力机制或进行深层语义绑定，实现极端破坏条件下的高置信度提取。

4.2　异构特征解耦与多维联合建模

面对破坏型攻击者利用深度网络对水印特征进行的高效剥离，反制型博弈的核心策略之一是实现水印信号与视觉内容的深度解耦，消除两者在高维分布上的相互干涉，从而避免擦除网络在重构图像时将水印一并滤除。

为了突破深层擦除网络的破坏，Sun等[54]提出了一种基于扩散模型的鲁棒水印框架DiffMark。该方法引入冻结的预训练自编码器来模拟破坏性操作，并设计了跨信息融合模块，对抗性损失与水印保真度的联合约束下实现了复杂图像背景与水印特征的高效解耦，有效提升了水印在高维空间中的抗干扰能力。同时，针对大模型在特征映射阶段的脆弱性，Rezaei等[55]提出了LaWa框架，通过在预训练自编码器的潜在空间中嵌入多尺度编码模块，实现了水印信息与视觉特征在生成过程中的内生解耦。针对生成大模型在遭受恶意微调时容易丢失内生水印的缺陷，Wang等[56]提出的SleeperMark框架通过显式引导文本到图像扩散模型，将水印信息与视觉语义概念进行强制分离。该机制使得模型在适应新任务或遭受下游微调擦除时，依然能够保留嵌入的标识信息，为抵御模型级与图像级的双重破坏提供了全新的网络设计思路。此外，Zhu等[57]提出了一种针对扩散模型的对抗性水印验证范式，通过定向优化生成携带版权信息的对抗样本，主动阻止非授权网络对底层图像特征的提取与擦除。然而，面对破坏型攻击者日益复杂的联合擦除手段，仅依赖单一维度的解耦往往容易被瓦解。为此，研究者们进一步引入了多域联合建模与大容量对抗机制。席祖平等[58]提出了基于模拟对抗的鲁棒模型水印架构，通过在训练阶段引入针对性的对抗攻击模拟，使模型在生成过程中能够自适应地强化水印特征的抗擦除边界。Li等[59]提出了双域扩散模型联合水印框架GaussMarker，并引入了独立于模型的学习型噪声恢复器，大幅提升了视觉内容在遭受深度重构后的高置信度召回。此外，针对极端恶意环境下的标识碰撞与信息丢失问题，诸如MaXsive[60]等大容量免训练防御方案，以及MarkDiffusion[61]这一涵盖多种对抗测评基准的开源工具包相继问世，进一步将AIGC水印的特征解耦与抗深度擦除能力推向了新的博弈高度。

4.3　语义级偏转与自适应对抗强化

在反制跨物理域破坏与高维语义编辑攻击的博弈中，防御方逐渐摒弃了脆弱的像素域嵌入，转而寻求更为隐蔽且动态的语义级对抗强化策略。针对AIGC驱动的深度语义篡改，Liu等[62]提出了PAI框架，该框架引入了一种无需训练的语义级偏转机制，跳出了仅在噪声初始化阶段嵌入水印的传统限制。通过基于密钥的轨迹偏转，在扩散去噪的深层语义空间中隐蔽植入特征，不仅有效抵御了多种生成式破坏攻击，还实现了针对局部篡改的精准定

位。此外，You 等[63]设计的 PatchSeal 架构采用多目标分散嵌入策略，将冗余的水印比特分布于多个具有独立语义的图像区域中，并利用注意力机制动态调节局部嵌入强度。这种自适应感知机制确保了水印在经历高强度的语义转换或二次编辑后，仍能维持极高的提取置信度。为了进一步平衡对抗鲁棒性与生成保真度，Huang 等[64]提出的 ROBIN 框架将水印嵌入转化为潜在空间中的对抗优化问题，通过自适应寻找抗几何变换与深度擦除的最优特征解，显著强化了水印在开放网络分发与跨域传播中的安全边界。

5　主流 AIGC 水印技术的评价体系与对比分析

5.1　评价指标与主流评测基准

在 AIGC 水印的“防-查-抗”动态博弈中，科学严谨的定量评估体系是衡量防御策略有效性与攻击手段破坏力的关键标尺。当前，面向 AIGC 视觉内容的主动防御水印评价体系已逐渐从传统的数字信号处理标准，向适应高维语义生成与深度对抗环境的高阶指标演进，主要涵盖视觉保真度、提取可靠性以及综合评测基准三个核心维度。

（1）视觉保真度与生成质量指标

为了评估水印的隐蔽嵌入对 AIGC 高维语义与视觉质量的影响，最基础的指标是峰值信噪比（PSNR）和结构相似性（SSIM）。其中，PSNR 用于衡量像素级的绝对误差

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{1}$$

式中，MAX_I 表示图像颜色的最大数值，MSE 为原始图像与含水印图像间的均方误差。SSIM 则从亮度、对比度和结构三个维度衡量图像的感知相似度

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2}$$

此外，针对 AIGC 高维重构特性，学习感知图像块相似度被广泛用于精确衡量生成图像在人类视觉感知上的细微差异。更为关键的是，Fréchet Inception 距离（FID）通过计算生成图像与真实自然图像在深层特征空间中的分布距离，已成为评估内生同构水印是否破坏大模型原始生成先验的核心标准

$$FID = \| \mu_r - \mu_g \|^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{3}$$

式中，μ_r,Σ_r 与 μ_g，Σ_g 分别代表真实图像和生成图像在特征空间中的均值与协方差矩阵。

（2）提取可靠性与鲁棒性指标

在面向高维伪造与深度擦除的“查”与“抗”博弈阶段，溯源机制的可靠性评估分为零比特鉴伪与多比特提取两大任务。针对零比特检测，通常采用真正率（TPR）、假正率（FPR）以及 ROC 曲线下面积（AUC）来综合衡量模型的分类边界。对于多比特信息溯源任务，则主要依赖比特准确率（Bit Accuracy, BACC）来精确量化经历网络擦除后的溯源信息残留量

$$BACC = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(w_i = w_i') \tag{4}$$

式中，N 为水印序列长度，w_i 与 w_i' 分别代表原始嵌入水印与提取出的水印比特。同时，归一化相关系数（NC）亦常被用于衡量提取凭证与原始水印特征间的拓扑相似度

$$NC = \frac{\sum_{i,j} W(i,j)W'(i,j)}{\sqrt{\sum_{i,j} W(i,j)^2}\sqrt{\sum_{i,j} W'(i,j)^2}} \tag{5}$$

（3）主流评测基准与开源对抗平台

随着破坏型攻击手段的持续迭代，单纯依赖 MS-COCO 或 ImageNet 等常规数据集已难以全面反映防御机制在极端 AIGC 生态中的真实表现。针对这一评测痛点，学界相继推出了一系列综合性基准。例如，DF40 基准测试集[4]提供了涵盖多类先进生成架构的伪造数据；NeurIPS 隐蔽水印擦除挑战赛[53]构建了贴近真实黑盒擦除环境的对抗性测试标准，极大地推动了抗擦除模型的演进。此外，以 MarkDiffusion 为代表的开源评测工具包，[61]集成了针对潜扩散模型的多种内生生成与破坏攻击管线，正逐步成为主动防御水印技术标准化测评的核心平台。

5.2　代表性防御水印算法对比

为了更直观地展现本文提出的“防-查-抗”三元分析框架下各阶段算法的技术演进脉络与优劣势，本节对现有代表性的主动防御水印策略进行了系统性的归纳与对比。如表 1 所示，本文紧扣阻断型（防）、溯源型（查）与反制型（抗）三大博弈阶段，从核心机理、适用场景、核心优势与现有局限等多个关键维度，对文献进行了横向梳理。通过

对比可以发现，AIGC 水印技术正逐步从外置的静态标识向内生的动态对抗演进，但在应对高维语义编辑优化计算量以及跨模态特征泛化方面，仍面临着共同的技术挑战。

6 挑战与展望

随着生成式基础大模型向多模态加速演进，基于防-查-抗动态博弈框架的主动防御体系在应对极端对抗环境时，仍面临一系列挑战。本节将集中剖析当前水印技术的核心挑战，并据此提出未来的关键演进路线。

（1）多模态与跨域重构带来的失效风险：当前的同构型与反制型水印多局限于单一的视觉模态。然而，随着 AIGC 生态向文本-图像-视频的多模态联合生成演进，不同物理媒介间存在显著的特征异构性。隐蔽水印在经历跨模态转换与特征重采样时，极易发生信息丢失与结构坍塌，难以维持溯源特征的一致性。

（2）强对抗场景下的理论安全边界缺失：目前的防御机制多依赖于经验性的网络设计与启发式的对抗训练。在面对攻击者利用充沛算力进行高频次黑盒查询与深度重构试探时，现有技术的容量、视觉保真度与鲁棒性之间缺乏严格的数学极限论证。这导致防御方难以在复杂的攻击场景下提供具有确定性保证的提取置信度。

（3）模型算力开销高昂与评估标准碎片化：在工程层面，将水印深度融入百亿参数级的大模型往往会带来显著的计算延迟与显存开销。此外，当前的测评维度多依赖传统的像素级攻击测试，缺乏统一的大型视觉-语言模型级别的自动化评估基准，不同水印算法的跨平台兼容性与抗高维语义编辑能力难以得到客观衡量。

针对上述挑战，结合现有主动防御体系的演进规律，本文认为未来 AIGC 数字水印技术的研究可重点探索以下三个方向：

（1）构建跨模态统一的拓扑级内生水印：未来的防御策略需减少对单一像素域或特定生成架构的依赖，转向基础模型的多模态联合表示空间。通过

表 1 面向 AIGC 视觉内容的主动防御水印算法对比总结表

博弈阶段	策略分类	代表性工作	核心机理	适用场景	核心优势	现有局限
防(阻断型)	视觉数据特征扰断	Glaze[19], Nightshade[22], Invertible Protection [20]	像素域注入对抗扰动，误导特征映射与概念学习	保护开源数据/个人作品，抵御恶意爬虫与风格克隆	无需修改模型；源头前置阻断	微小画质损耗；面对更强模型易失效
防(阻断型)	模型参数后门触发	Zeng 等[21], Zhao 等[30], Luo 等[31]	水印绑定触发集，隐式植入模型深层权重	保护大模型参数，追踪非法微调/知识蒸馏	隐蔽性强；黑盒凭提示词触发验证	需干预训练流程；部署成本较高
查(溯源型)	潜在空间特征融合	Bui 等[38], Tree-Ring[49], T2SMark[44]	利用潜空间冗余，深度耦合水印与视觉语义	追踪海量 AIGC 内容，防范深伪泛滥	内生同构，高隐蔽；抗常规噪声	依赖特定模型架构；跨模态易丢失
查(溯源型)	逆向扩散反演提取	镜像扩散网络[50], Asnani 等[48]	利用生成模型可逆性，在对偶空间盲推断指纹	无原图辅助的开放网络盲提取确权	零载体依赖，盲取证能力强	多步反演算力开销大、耗时长
抗(反制型)	异构特征解耦建模	DiffMark[54], Sleeper-Mark[56], GaussMarker[59]	引入对抗约束，强制分离水印与视觉概念	应对破坏型攻击者的深度去噪与恶意擦除	显著提升抗深度擦除与微调破坏能力	网络复杂；可能牺牲部分生成保真度
抗(反制型)	语义偏转与自适应强化	PAI[62], PatchSeal[63], ROBIN[64]	动态调节局部强度，深层语义空间植入特征	抵御跨物理域破坏、几何变换及局部篡改	自适应信道状态；抗高频语义编辑	优化计算量大；缺乏数学下界证明

在高维上寻找抗干扰的鲁棒子空间，使溯源凭证能够在不同媒介形态转换中实现自适应映射，进而达成跨模态的泛化取证。

（2）探索动态对抗下的可证明鲁棒性体系：为应对高强度的网络擦除，水印技术需从经验设计向理论可证明方向深化。未来可引入不完全信息动态博弈等理论，量化不同生成架构在遭受破坏时的信息残留量的下界，从数学理论层面确立水印防篡改的安全边界。

（3）推进轻量化部署与全链路标准化监管：在工业级应用中，探索基于参数高效微调的轻量化水印注入范式将是重要趋势，以期在极低算力开销下完成特征前置同构。同时，学术界与工业界需共同探索建立解耦的嵌入-提取零知识证明协议，在保护模型权重与数据隐私的前提下，构建覆盖AIGC全生命周期的标准化溯源接口。

7　结束语

AIGC视觉生成技术的跨越式演进在重塑多媒体内容生产模式的同时，也引发了深远的数字信任危机。面对高保真重构与深度网络擦除带来的多维安全威胁，传统的被动取证与静态水印分类框架已难以适应持续升级的攻防态势。本文突破了将数字水印视为依附性凭证的传统视角，将动态博弈理论引入AIGC确权与溯源领域，构建了覆盖视觉内容全生命周期的“防-查-抗”主动防御框架。依托该框架，本文系统梳理了数字水印从被动标识向主动防御跃升的演进逻辑，系统揭示了防御方在应对非授权特征挖掘、高维潜空间伪造以及网络级恶意破坏时的核心博弈机制。尽管该框架为AIGC水印技术的演进提供了系统性的理论指引，但其在实际应用中仍然存在局限性：现有的三元博弈模型仍难以对多域联合破坏场景下的信息残留量给出绝对的理论下界证明。在未来的数字生态治理中，面对多模态融合与跨域生成的复杂趋势，数字水印技术必将超越单纯的附加溯源属性，逐步演化为具备可证明安全性与跨域泛化能力的内生架构，成为维护AI资产权益与构建可信多媒体的核心策略之一。

参考文献：

[1] . Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]// Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc., 2020, 33: 6840-6851.

[2] Howe N P, Thompson B. This isn’t the Nature Podcast—how deepfakes are distorting reality[J]. Nature, 2023, 1.

[3] . Xu D, Fan S, Kankanhalli M. Combating misinformation in the era of generative AI models[C]//Proceedings of the 31st ACM International Conference on Multimedia (MM ’23). New York: ACM Press, 2023: 9291-9298.

[4] . Yan Z, Yao T, Chen S, et al. DF40: toward next-generation deepfake detection[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc., 2024, 37: 29387-29434.

[5] . Lin L, He X, Ju Y, et al. Preserving fairness generalization in deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 16815-16825.

[6] . Lanzino R, Fontana F, Diko A, et al. Faster than lies: real-time deepfake detection using binary neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 3771-3780.

[7] . Zhao Z, Jin H, Guo Q, et al. SIGMark: scalable in-generation watermark with blind extraction for video diffusion[C]//The Fourteenth International Conference on Learning Representations. 2026.

[8] Zhou T, Ding R, Liu G, et al. A content-dependent watermark for safeguarding image attribution[EB/OL]. arXiv:2509.10766, 2025.

[9] . Dong Z, Shuai C, Ba Z, et al. WMCopier: forging invisible image watermarks on arbitrary images[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc., 2025.

[10] Fu W, Carter F, Wang Y, et al. Diffusion-based image editing: an unforeseen adversary to robust invisible watermarks[EB/OL]. arXiv: 2511.05598, 2025.

[11] Cao J, Li Q, Zhang Z, et al. Secure and robust watermarking for AI-generated images: a comprehensive survey[EB/OL]. arXiv: 2510.02384, 2025.

[12] 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.

[13] 郭钊均, 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.
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.

[14] . Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM International Conference on Multimedia Retrieval (ICMR). New York: ACM Press, 2017: 269-277.

[15] . Darvish rouhani B, Chen H, Koushanfar F. Deepsigns: an end-to-end watermarking framework for ownership protection of deep neural networks[C]//Proceedings of the Twenty-Fourth International Con-ference on Architectural Support for Programming Languages and Operating Systems,2019: 485-497.

[16] . Wang T, Kerschbaum F. RIGA: covert and robust white-box watermarking of deep neural networks[C]//Proceedings of the Web Conference. New York: ACM Press, 2021: 993-1004.

[17] . Hayes J, Danezis G. Generating steganographic images via adver-

sarial training[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 1951-1960.

[18] Salman H, Khaddaj A, Leclerc G, et al. Raising the cost of malicious AI-powered image editing[EB/OL]. arXiv:2302.06588, 2023.

[19] Shan S, Cryan J, Wenger E, et al. Glaze: Protecting artists from style mimicry by text-to-image models[EB/OL]. arXiv:2302.04222, 2023.

[20] . Chen K, Zeng X, Ying Q, et al. Invertible image dataset protection [C]//2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE Press, 2022: 1-6.

[21] . Zeng Y, Tan J, You Z, et al. Watermarks for generative adversarial network based on steganographic invisible backdoor[C]//2023 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE Press, 2023: 1211-1216.

[22] . Shan S, Ding W, Passananti J, et al. Nightshade: prompt-specific poisoning attacks on text-to-image generative models[C]//2024 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2024: 807-825.

[23] Wang H, Shen Q, Tong Y, et al. The stronger the diffusion model, the easier the backdoor: data poisoning to induce copyright breaches without adjusting finetuning pipeline[EB/OL]. arXiv:2401.04136, 2024.

[24] Souri H, Bansal A, Kazemi H, et al. Generating potent poisons and backdoors from scratch with guided diffusion[EB/OL]. arXiv: 2403.16365, 2024.

[25] Lei L, Gai K, Yu J, et al. Watermarking visual concepts for diffusion models[EB/OL]. arXiv:2411.11688, 2024.

[26] . Ong D S, Chan C S, Ng K W, et al. Protecting intellectual property of generative adversarial networks from ambiguity attacks[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 3629-3638. DOI: 10.1109/CVPR46437.2021.00363.

[27] Qiao T, Ma Y Y, Zheng N, et al. A novel model watermarking for protecting generative adversarial network[J]. Computers & Security, 2023, 127: 103102. DOI: 10.1016/j.cose.2023.103102.

[28] . Li F Q, Wang S L. Persistent watermark for image classification neural networks by penetrating the autoencoder[C]//Proceedings of the IEEE International Conference on Image Processing (ICIP). Los Alamitos: IEEE Computer Society Press, 2021: 3063-3067.

[29] . Yin Z X, Yin H, Zhang X P. Neural network fragile watermarking with no model performance degradation[C]//Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2022: 3958-3962. DOI: 10.1109/ICIP46576.2022.9897413.

[30] Zhao Y, Pang T, Du C, et al. A recipe for watermarking diffusion models[EB/OL]. arXiv:2303.10137, 2023.

[31] Luo G, Huang J, Zhang M, et al. Steal my artworks for fine-tuning? A watermarking framework for detecting art theft mimicry in text-to-image models[EB/OL]. arXiv:2311.13619, 2023.

[32] Al-Haj A. Combined DWT-DCT digital image watermarking[J]. Journal of Computer Science, 2007, 3(9): 740-746.

[33] . He Y, Hu Y. A proposed digital image watermarking based on DWT-DCT-SVD[C]//2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). Piscataway: IEEE Press, 2018: 1214-1218.

[34] Sivananthamaitrey P, Murthy P S N, Kumar P R. Multifaceted watermarking of medical images using SWT and SVD[J]. International Journal of Advanced Science and Technology, 2019, 28(7): 1-14.

[35] Kumar S, Singh B K. DWT based color image watermarking using maximum entropy[J]. Multimedia Tools and Applications, 2021, 80 (10): 15487-15510.

[36] Zhang K A, Xu L, Cuesta-Infante A, et al. Robust invisible video watermarking with attention[EB/OL]. arXiv:1909.01285, 2019.

[37] Jiang Z, Zhang J, Gong N Z. Evading watermark-based detection of AI-generated content[EB/OL]. arXiv:2305.03807, 2023.

[38] . Bui T, Agarwal S, Yu N, et al. RoSteALS: robust steganography using autoencoder latent space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 933-942.

[39] . Xiong C, Qin C, Feng G, et al. Flexible and secure watermarking for latent diffusion model[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM Press, 2023: 1668-1676.

[40] Fernandez P, Couairon G, Jégou H, et al. The stable signature: Rooting watermarks in latent diffusion models[EB/OL]. arXiv: 2303.15435, 2023.

[41] Cui Y, Ren J, Xu H, et al. DiffusionShield: A watermark for copyright protection against generative diffusion models[EB/OL]. arXiv: 2306.04642, 2023.

[42] . Yang Z, Zeng K, Chen K, 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 (CVPR). Piscataway: IEEE Press, 2024: 12162-12171.

[43] Zhang G, et al. MarkPlugger: generalizable watermark framework for latent diffusion models without retraining[J]. IEEE Transactions on Multimedia, 2025.

[44] Yang J, et al. T2SMark: balancing robustness and diversity in noise-as-watermark for diffusion models[EB/OL]. arXiv:2510.22366, 2025.

[45] Hurrah N N, Parah S A, Loan N A, et al. Dual watermarking framework for privacy protection and content authentication of multimedia [J]. Future Generation Computer Systems, 2019, 94: 654-673.

[46] . Zhu J R, Kaplan R, Johnson J, et al. HiDDeN: hiding data with deep networks[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Heidelberg: Springer, 2018: 657-672.

[47] . Albright M, McCloskey S. Source generator attribution via inversion [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2019: 8. Art. no. 3.

[48] Asnani V, Yin X, Hassner T, Liu X M. Reverse engineering of generative models: inferring model hyperparameters from generated images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15477-15493. DOI: 10.1109/TPAMI.2023.3301451.

[49] Wen Y, Kirchenbauer J, Geiping J, et al. Tree-ring watermarks: fingerprints for diffusion images that are invisible and robust[EB/OL]. arXiv: 2305.20030, 2023.

[50] Li G, Chen Y, Zhang J, et al. Towards the vulnerability of watermarking artificial intelligence generated content[EB/OL]. arXiv: 2310.07726, 2023.

[51] 李莉,张新鹏,王子驰,等. 基于精确扩散反演的生成式图像内生水印方法 [J]. 网络空间安全科学学报, 2024, 2 (01): 92-100. DOI: 10.20172/j.issn.2097-3136.240108.

Li L, Zhang X P, Wang Z C, et al. Generative image endogenous watermarking method based on exact diffusion inversion[J]. Journal of Cybersecurity, 2024, 2(1): 92-100.

[52] Guo F, Kang J, Ming Q, et al. Vanishing watermarks: diffusion-based image editing undermines robust invisible watermarking[EB/OL]. arXiv:2602.20680, 2026.

[53] . Shamshad F, Bakr T, Shaaban Y S, et al. First-place solution to NeurIPS 2024 invisible watermark removal challenge[C]//The 1st Workshop on GenAI Watermarking. 2025.

[54] Sun C, Sun H, Guo Z, et al. DiffMark: diffusion-based robust watermark against deepfakes[J]. Information Fusion, 2025: 103801.

[55] . Rezaei A, Akbari M, Alvar S R, et al. LAWA: using latent space for in-generation image watermarking[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 118-136.

[56] . Wang Z, Guo J, Zhu J, et al. SleeperMark: towards robust watermark against fine-tuning text-to-image diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2025: 8213-8224.

[57] . Zhu P, 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 (CVPR). Piscataway: IEEE Press, 2024: 24420-24430.

[58] 席祖平, 瞿左珉, 卢伟, 等. 基于模拟对抗的鲁棒模型水印[J]. 网络空间安全科学学报, 2024, 2(5): 67-77.
Xi Z P, Qu Z M, Lu W, et al. Robust model watermarking based on adversarial simulation[J]. Journal of Cybersecurity, 2024, 2(5): 67-77.

[59] Li K, Huang Z, Hou X, et al. GaussMarker: robust dual-domain watermark for diffusion models[EB/OL]. arXiv:2506.11444, 2025.

[60] . Mao P Y, Tsai C C, Lu C S. MaXsive: high-capacity and robust training-free generative image watermarking in diffusion models[C]//Proceedings of the 33rd ACM International Conference on Multimedia. New York: ACM Press, 2025: 11443-11452.

[61] Pan L, Guan S, Fu Z, et al. MarkDiffusion: an open-source toolkit for generative watermarking of latent diffusion models[EB/OL]. arXiv: 2509.10569, 2025.

[62] Liu Q, Zhang Y, Ba Z, et al. Attack-resistant watermarking for AIGC image forensics via diffusion-based semantic deflection[EB/OL]. arXiv:2601.06639, 2026.

[63] You T, Zheng H, Wang Z, Chen Y. PatchSeal: a robust and intangible image watermarking framework for AIGC[J]. Mathematics, 2026, 14 (4): 679.

[64] . Huang H, Wu Y, Wang Q. Robin: robust and invisible watermarks for diffusion models with adversarial optimization[C]//Advances in Neural Information Processing Systems, 2024, 37: 3937-3963.

谢雪（1989− ），男，吉林长春人，中国科学技术大学博士研究生，主要研究方向为信息隐藏、人工智能安全。

戚宇昂（1999− ），男，河北廊坊人，中国科学技术大学博士研究生，主要研究方向为信息隐藏、人工智能安全。

陈可江（1994− ），男，浙江温州人，博士，中国科学技术大学副教授，主要研究方向为信息隐藏、人工智能安全。

张卫明（1976− ），男，河北保定人，博士，中国科学技术大学教授，主要研究方向为信息隐藏和多媒体安全。

俞能海（1964− ），男，安徽芜湖人，博士，中国科学技术大学教授，主要研究方向为多媒体安全、多媒体信息检索、视频处理和信息隐藏。