# 深度神经网络模型水印攻击研究

王雯，杨奎武，仝松松，魏江宏，薛岩，周荣魁
（中国人民解放军网络空间部队信息工程大学数据与目标工程学院，河南 郑州 450001）

**摘　要：**模型知识产权保护已成为模型安全中不可忽视的问题，水印技术作为模型溯源的核心手段，通过将特殊标识嵌入模型参数或生成内容中，为版权验证提供技术支撑。然而，训练完成的含水印模型非常容易被复制并扩散，这使得攻击者能够通过微调、剪枝或对抗样本攻击等特定技术手段，破坏或去除深度神经网络（DNN）模型中嵌入的水印，使得模型所有权无法验证。为了更深入地了解模型水印攻击方法，首先对模型水印攻击进行介绍，然后对模型水印攻击方法进行分类，根据攻击者对目标模型的访问权限和信息获取能力，分为白盒水印攻击和黑盒水印攻击两类，对 DNN 模型水印攻击的动因、危害、攻击原理和具体实施手段进行梳理和分析，接着对现有模型水印攻击研究从攻击者能力及性能影响等方面进行比较与总结，最后探讨了神经网络模型水印攻击在未来研究中的潜在积极作用，为模型安全和知识产权保护领域的深入研究提供建议。

**关键词：**深度学习；模型安全；水印技术；人工智能（AI）安全；版权保护
**中图分类号：**TP309　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0252743

# Research on Watermarking Attack of Deep Neural Network Models

WANG Wen, YANG Kuiwu, TONG Songsong, WEI Jianghong, XUE Yan, ZHOU Rongkui
(School of Data and Target Engineering, The PLA Information Engineering University, Zhengzhou 450001, Henan, China)

**【Abstract】** Model intellectual property protection is an issue that cannot be ignored in model security. Watermarking technology, as the core means of model traceability, provides technical support for copyright verification by embedding special identifiers into model parameters or generated content. However, trained watermarked models can easily be copied and spread, which enables attackers to destroy or remove the watermarks embedded in Deep Neural Network (DNN) models using specific technical means such as fine-tuning, pruning, or adversarial sample attacks, making the verification of model ownership impossible. To gain a deeper understanding of model watermarking attack methods, this study begins by introducing model watermarking attacks and proceeds to classify these methods into two categories, white-box watermarking attacks and black-box watermarking attacks, based on the attacker′s access rights and information acquisition capabilities regarding the target model. It also sorts and analyzes the motives, hazards, attack principles, and specific implementation methods of DNN model watermarking attacks. Moreover, it compares and summarizes existing research on model watermarking attacks from the perspectives of attacker capabilities and performance impacts. Finally, it explores the potential positive roles of neural network model watermarking attacks in future research and provides suggestions for in-depth research in the fields of model security and intellectual property protection.

**【Key words】** deep learning; model security; watermarking technology; Artificial Intelligence (AI) security; copyright protection

## 0　引言

深度神经网络（DNN）在计算机视觉[1]、自然语言处理[2]等领域的突破性进展推动其成为智慧城市、医疗诊断等关键场景的核心技术底座。训练高性能的 DNN 模型需要大规模的数据、高性能的计算硬件设施及专业的人员，这使得 DNN 模型有着较高的商业价值，其版权保护也自然成为人们关注的焦点。

为实现 DNN 模型的版权保护，UCHIDA 等[3]在 2017 年首次提出了一个 DNN 模型水印方法，该方法在验证模型所有权的同时又不会影响模型性能，为模型版权保护提供了新的思路。2019 年，WANG 等[4]发现在 UCHIDA 等[3]的水印嵌入过程中，模型参数分布的方差会显著增加，导致水印很容易被检测出来，甚至可以得出嵌入的水印长度，利用这些信息攻击者便可使水印验证失败。这一漏洞

**基金项目：**国家自然科学基金（62172434）；河南省高等教育教学改革研究与实践项目（2024SJGLX0095）。
**作者简介：**王雯（CCF 学生会员），女，硕士研究生，主研方向为人工智能安全、模型水印；杨奎武（通信作者），副教授；仝松松，硕士研究生；魏江宏，讲师、博士后；薛岩，本科生；周荣魁，博士研究生。
**收稿日期：**2025-07-10　　**修回日期：**2025-10-09　**E-mail：**yangkw@aliyun.com

的公开为模型水印攻击提供了新的研究方向，DNN 模型水印攻击技术成为模型安全领域的热点话题。然而，对水印攻击技术的深入研究的价值并不仅限于揭示潜在风险，更在于其为模型安全提供的正向驱动作用。首先，系统性的攻击研究为模型水印方案的鲁棒性提供了评估基准与验证手段，是衡量水印技术有效性和实用性的关键环节。其次，攻击方法的不断创新能够暴露现有水印技术的脆弱性，从而为构建更完善的水印防御体系指明方向，并最终为模型版权保护标准的制定提供关键的科学依据。因此，深入分析模型水印攻击技术无论对于推动隐蔽性、鲁棒性更强的版权标识方案，还是对于构建安全可靠的深度学习模型生态都具有重要意义。

已有模型水印攻击的相关研究成果[5-8]主要集中于传统模型水印攻击方法，更多面向图像分类、目标检测等判别式模型，缺乏对扩散模型、生成对抗网络等生成式模型的水印安全威胁的系统分析。针对这一情况，本文围绕人工智能（AI）安全领域 DNN 模型水印攻击的最新研究进展进行了归纳总结。引言介绍了 DNN 模型所面临的安全风险并由此引出模型水印攻击的研究背景和意义；第 1 章深入分析了同类综述研究现状，并阐明本文主要贡献；第 2 章介绍了模型水印及其典型嵌入方式，引入模型水印攻击，并系统对比了传统的数字水印攻击与模型水印攻击之间的本质区别；第 3 章将现有 DNN 模型水印攻击方法分为针对白盒水印攻击和针对黑盒水印攻击，前者从基于模型修改攻击和基于模型提取攻击进行分类，后者从基于后门去除、基于对抗训练、基于迁移学习、基于生成数据的水印攻击进行分类；第 4 章对模型水印攻击研究中常用到的数据集和模型进行总结，以及对现有水印攻击方法进行对比和分析；第 5 章从多个角度对未来的研究方向进行分析和展望。

## 1 相关研究

在数字版权保护与 AI 安全领域，DNN 模型水印技术作为保障模型知识产权的重要技术途径，近年来受到学术界与工业界的广泛关注。研究背景表明：相关成果在顶级期刊及会议上的发表数量呈现显著增长态势，例如在网络安全领域的多个权威会议与期刊，主要包括 USENIX Security、NDSS、ACM CCS、IEEE S&P 等四大顶级会议，并扩展检索了 AI、信息安全、多媒体等相关领域近 3—5 年发表的研究成果。已有学者尝试对现有工作进行了系统性梳理，其中针对 DNN 模型水印攻击技术的综述研究主要存在两种分类方式：1）按攻击目标分类，何春辉等[5]将攻击行为划分为非技术攻击和技术攻击两类，在此基础上，技术攻击被进一步区分为无意识攻击和有意识攻击，谢宸琪等[6]归纳的模型剪枝、微调与水印覆盖同样适用于此分类；2）按攻击手段分类，夏道勋等[7]提出的分类体系是该分类的典型代表，将攻击技术划分为查询修改攻击、水印移除攻击和逃逸/伪造攻击，吴汉舟等[8]从水印验证的场景角度对攻击和防御机制进行了梳理，并特别强调了攻击手段与模型结构、参数及输出之间的关联性。

相较于现有综述多从攻击者视角或单一技术维度展开分类，本文根据攻击者能力与攻击对象多视角对模型水印攻击方法进行分析，主要贡献如下：

1）根据攻击的直接对象（即是模型本身还是其输出数据）进行细分，构建了一个系统化的分类框架。该框架不仅涵盖传统的模型修改攻击（如微调和剪枝），还系统性地归纳了新兴的基于数据生成的攻击（如利用扩散模型进行水印去除与规避），为理解复杂水印攻击技术提供了新的视角。

2）深入分析了基于生成式 AI，尤其是扩散模型的水印攻击这一新兴研究方向，不仅详细介绍了 DiffWA、CtrlRegen 等代表性方法，而且指出了此类方法的核心机制，即利用生成模型的强大再生能力，从带水印内容中重构视觉无损的无水印版本，或直接操纵水印检测器的解码输出。

3）结合现有 DNN 模型水印攻击研究工作，从攻击者所需数据、模型及水印知识以及性能影响等多个维度对各类攻击方法的有效性、局限性及使用场景进行综合对比分析。研究结果表明：水印攻击研究为评估水印方案鲁棒性提供关键基准，揭示现有水印技术的脆弱性，进而为构建更强的防御体系和推进模型版权保护标准制定提供了科学依据。

## 2 DNN 模型水印相关概念

### 2.1 DNN 模型水印及其典型嵌入方式

模型水印[9]指将数字水印技术应用于 DNN 模型版权保护的一种方法，其通过特定嵌入框架在模型中植入水印信息，以检测训练好的模型是否遭受侵权。现有水印嵌入方法主要包括基于参数、基于后门、基于对抗样本和基于特征 4 类。

基于参数的模型水印方法通过在损失函数中引入约束项，微调模型参数将水印编码嵌入决策边界，不影响模型性能。在水印验证阶段，模型使用者需掌握模型参数的全部信息。

基于后门的模型水印方法适用于黑盒场景，即模型使用者只能通过 API 查询模型输出，并通过训

练引入特定触发样本及其指定标签建立后门行为，验证时依据触发样本的响应判断水印存在性。

基于对抗样本的模型水印方法利用对抗扰动构建水印样本，并通过对抗训练调整决策边界，使模型仅对特定扰动样本具有鲁棒性，验证时以预设对抗样本的分类准确性为依据。

基于特征的模型水印方法通过保存模型训练过程中的数据特征或参数特征等使模型使用者能够借助特征比较来判别模型版权。

尽管具体实现方式多样，但是模型水印的流程可统一概括为水印生成、嵌入与提取 3 个阶段，其整体框架如图 1 所示（彩色效果见《计算机工程》官网 HTML 版，下同）。

### 2.2 DNN 模型水印攻击

模型水印攻击是指通过参数修正、对抗训练等特定技术手段破坏或移除 DNN 模型中嵌入的水印信息，以达到非法使用或否认模型所有权的行为。模型水印攻击的主要流程如图 2 所示。


图 1 DNN 模型水印方法的整体框架

Fig. 1 Overall framework of watermarking methods for DNN models

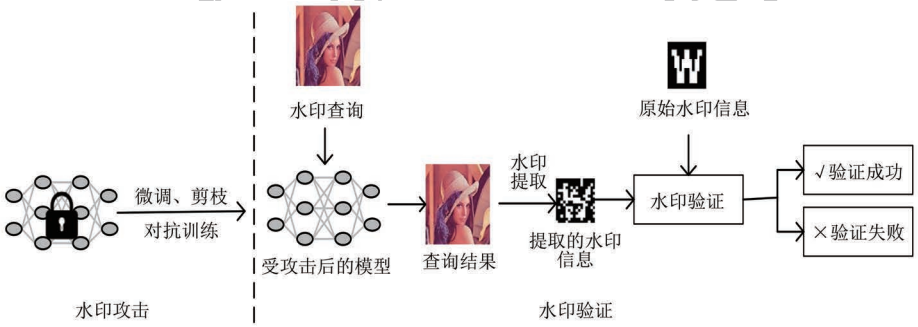


图 2 DNN 模型水印攻击的主要流程

Fig. 2 Main process of watermarking attack for DNN models

### 2.3 传统数字水印攻击与模型水印攻击

传统数字水印技术最初被设计用于在多媒体数据中嵌入不可感知的标识信息来保护数据版权。数字水印攻击技术是指对水印信息发起攻击，破坏水印信息的可验证性，从而使版权保护失效[5]。

与传统数字水印技术相比，模型水印及其攻击技术在水印载体、攻击目标以及攻击阶段存在显著差异。首先，在水印载体上，数字水印通常嵌入在图像、音频、视频或者文档等数据中，而模型水印则嵌入在 DNN 模型本身或由模型生成的数据中；其次，在攻击目标上，数字水印攻击是在使水印验证失败的同时保持载体数据的完整性，而模型水印攻击是在使水印验证失败的同时保持模型功能完整性；再次，在攻击阶段上，数字水印攻击多发生在数据传输或存储阶段，而模型水印攻击可发生在模型推理、模型训练和模型部署等多个阶段；最后，在水印的可见性上，数字水印和模型水印均可分为可见和不可见两类。数字可见水印指的是将人眼可察觉的标识嵌入载体数据，不可见水印通常是通过修改载体数据的冗余部分来实现的。模型可见水印直接在输出结果中添加可识别标记[10]，不可见水印通常通过微调模型参数、修改模型结构，或利用特定输入触发特定输出等方式来嵌入水印，以实现其隐蔽性。表 1 多维度地对比了数字水印攻击与模型水印攻击的差异。

表 1 传统数字水印攻击与模型水印攻击

Table 1 Traditional digital watermarking attacks and model watermarking attacks

| 水印类型 | 水印载体 | 攻击目标为破坏水印 | 攻击目标为去除水印 | 攻击目标为伪造水印 | 攻击阶段 | 水印隐蔽性 |
|---|---|---|---|---|---|---|
| 数字水印 | 图像、音频、视频、文档 | 使水印失效或无法检测 | 去除水印且保留载体完整 | 添加虚假水印 | 数据传输/存储阶段 | 可见/不可见水印 |
| 模型水印 | DNN 模型或生成数据 | 干扰模型水印验证功能 | 去除水印且保持模型性能 | 植入伪造水印窃取模型 | 模型推理/训练/部署阶段 | 可见/不可见水印 |

## 3 DNN 模型水印攻击方法分类

根据攻击者对目标模型的访问权限，DNN 模型水印攻击主要分为白盒攻击和黑盒攻击两类。在白盒攻击下攻击者能够完全访问模型的参数与结构，常见的攻击方法包括微调、剪枝、水印覆盖和模型蒸馏等。在黑盒攻击中，攻击者仅能通过模型接口获取输入输出，无法了解模型内部的具体参数与结构，其典型攻击方法有基于后门去除、基于对抗训练、基于迁移学习以及基于生成数据的水印攻击等。图 3 展示了 DNN 模型水印攻击方法的分类。

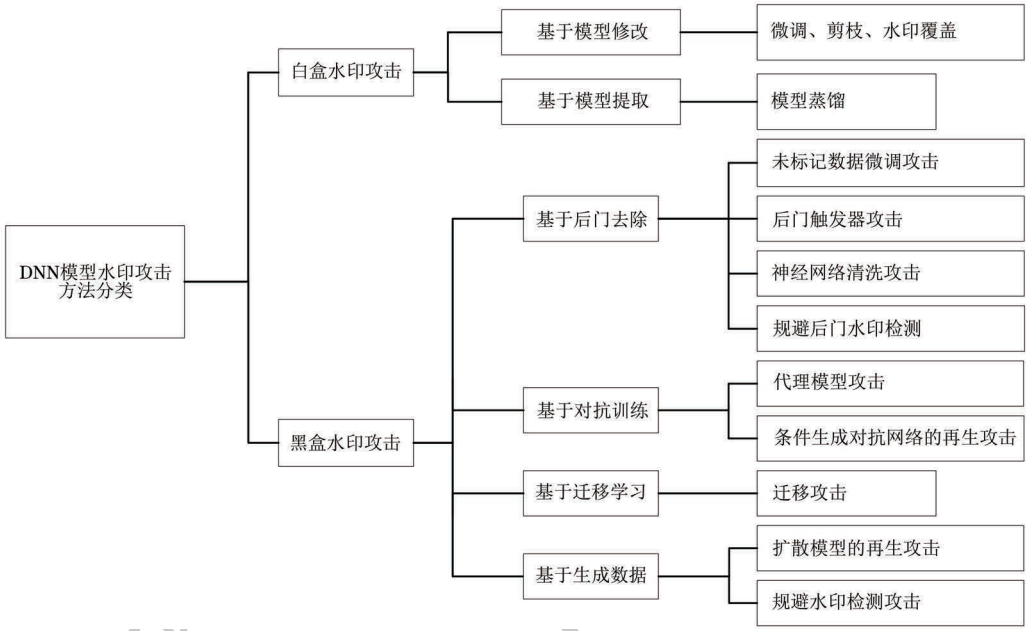


**图 3 DNN 模型水印攻击方法分类**

**Fig. 3 Classification of watermarking attack methods for DNN models**

### 3.1 白盒水印攻击

白盒水印攻击是指攻击者拥有对目标模型的完全访问权限，能够直接分析、修改或移除模型内部嵌入的水印信息，从而破坏水印的有效性。根据攻击手段的不同，白盒水印攻击可分为基于模型修改的攻击和基于模型提取的攻击。

#### 3.1.1 基于模型修改的白盒水印攻击

DNN 模型的可学习参数决定了其从输入数据中提取特征并生成预测的计算过程与最终性能。微调作为一种被广泛研究的水印去除攻击方法，是通过利用少量来自原始训练集或其同分布的数据对已部署模型进行继续训练来实现的。

CHEN 等[11]提出了一个基于微调的通用水印去除框架 REFIT。该框架的设计借鉴了机器学习中的灾难性遗忘现象[12-14]，即后续学习任务会覆盖或削弱先前学习任务的性能表现。如图 4 所示，REFIT 主要结合了弹性权重固化(EWC)[8]和未标记数据增强(AU)[15-16]两种方法。具体而言，首先将未标记数据输入已嵌入水印的模型，利用模型的预测生成伪标签，从而创建带有伪标签的增强训练数据。这些伪标签样本与有限的标记数据混合，并将其作为训练数据输入模型，更新其参数以破坏水印。同时，REFIT 采用 EWC 方法，通过估计对角 Fisher 矩阵并在损失函数中添加正则化项，旨在减轻灾难性遗忘效应，保护水印相关的旧性能不被新训练任务破坏。修正后的损失函数公式可表示如下：

$$\mathcal{L}_{\mathrm{EWC}}(\theta)=\mathcal{L}_{\mathrm{basic}}(\theta)+\frac{\lambda}{2}\sum_i F_i(\theta_i-\theta_i^*)^2 \quad (1)$$

式中：$\mathcal{L}_{\mathrm{basic}}(\theta)$是优化新任务性能的损失；$\lambda$ 是控制正则化强度的参数；$\theta^*$ 是使用上一个任务训练的参数；$F_i$ 是对应于第 $i$ 个参数的对角线项。

与先前灾难性遗忘的研究[12,17]不同的是，尽管攻击者的训练数据与预训练数据[18-20]不同，微调数据集不仅能够去除水印，还能保持模型主要任务性能。

CHEN 等[11]提出的水印攻击方法需要预先知晓水印类型以进行参数微调，这限制了其在复杂水印场景下的适用性。为了克服这一限制，GUO 等[21]提出了一种基于预处理样本变换(PST)和轻量微调的水印攻击方法，该方法可去除水印，且无需事先了解水印方案及水印嵌入过程所用的训练样

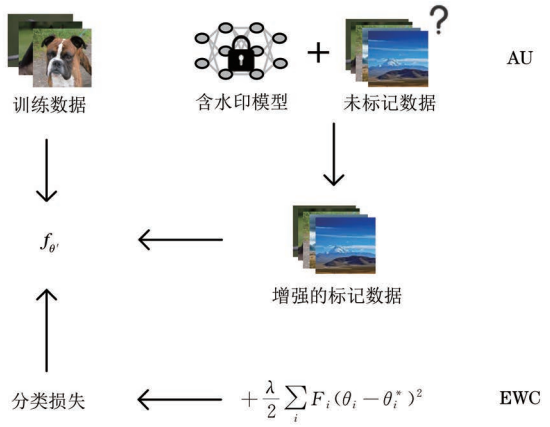


**图 4 REFIT 水印去除框架**

**Fig. 4 REFIT watermarking removal framework**

本。如图 5 所示，该方法首先给定水印模型，攻击者使用 PST 对部分 OOD(Out-Of-Distribution)数据样本进行预处理，并基于这些处理后的样本对目标水印模型进行少量周期的微调。在推理验证阶段，对输入样本统一施加 PST 预处理后再输入模型。此时，未经 PST 处理的样本的模型输出将保持不变，而经 PST 处理的样本的输出会发生显著变化。该机制的本质在于：带水印模型被训练以记忆特定输入样本和对应标签之间的映射关系，而在验证过程中，模型会对输入样本预测其对应水印标签。该标签通常与样本的真实类别不同，从而导致这一记忆关联具有脆弱性。PST 通过预处理破坏水印触发模式，使得水印验证机制失效。

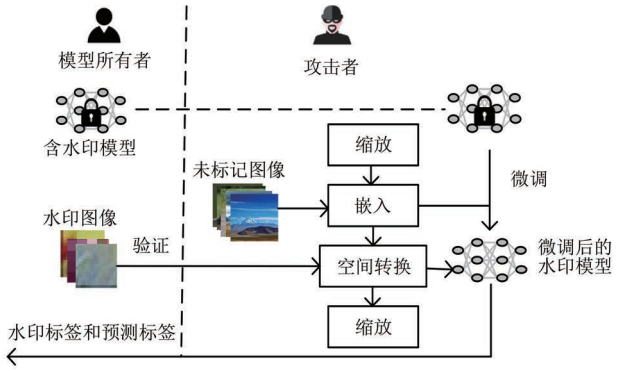


**图 5 基于 PST 的水印攻击流程**

**Fig. 5 Procedure of watermarking attack based on PST**

除了微调之外，模型剪枝通过移除神经网络中不重要的参数或连接，可能影响嵌入的水印信息。若水印位于冗余参数或非关键结构中，剪枝可能导致这些区域的水印信息被移除。若水印的嵌入依赖于模型的特定结构，剪枝通过剪除神经元之间的连接或通道，可能破坏这些结构，从而影响水印验证[22]。水印覆盖攻击是另一种常用的攻击手段。攻击者使用相同的水印嵌入算法，在已嵌入水印的模型中再嵌入水印以覆盖原有水印，导致模型版权归属难以确定或产生歧义[23]。

WANG 等[4]提出了水印覆盖攻击和多嵌入攻击两种方法，旨在通过向模型注入干扰信息来破坏原始水印。在水印覆盖中，基于固定的去除矩阵和固定的新水印，通过训练将新水印嵌入已含水印的模型，以覆盖原有水印信息。此过程通常涉及在损失函数中加入新水印的嵌入损失[24]，计算过程可表示如下：

$$E_R(w)=-\sum_{j=1}^{T}(b_j\log_a(y_j)+(1-b_j)\log_a(1-y_j)) \tag{2}$$

$$E_{\text{remove}}(w)=E_0(w)+\lambda E_R(w)+\lambda_{\text{new}}E_R^{\text{new}}(w) \tag{3}$$

式中：$y_j=\sigma\left(\sum_i \boldsymbol{X}_{ji}w_i\right)$，$\sigma(\cdot)$为 Sigmoid 函数，$\boldsymbol{X}$为固定去除矩阵；$b$ 为固定水印位；$E_0(w)$为原始任务损失；$T$ 为去除维度；$E_R(w)$为原有水印的嵌入损失；$E_R^{\text{new}}(w)$为新水印的嵌入损失；$\lambda$ 为权重系数。

多嵌入移除水印是在模型训练中动态生成不同的去除矩阵和新水印，通过高频随机扰动破坏原有水印结构。在第 $t$ 轮训练的损失函数可表示如下：

$$E_{\text{remove}}^t(w)=E_0(w)+\lambda E_R(w)+\lambda_t E_R^t(w) \tag{4}$$

在第 $t$ 轮训练时，随机生成去除矩阵$\boldsymbol{X}_t\sim\mathcal{N}(0,1)$和随机生成水印 $b_t\in\{0,1\}^{T'}$，每轮训练后丢弃 $\boldsymbol{X}_t$和 $b_t$，下一轮使用新随机值继续训练。这种方式既可以提高嵌入水印的隐蔽性，同时保留模型参数分布并提高水印的不可检测性。

因 WANG 等[4]用水印覆盖的方法去除水印可能会引起嵌入水印时造成的参数扰动，这一局限性促进了可逆水印技术的发展，GUAN 等[25]提出一种白盒场景下的可逆水印方法。该方法首先明确了卷积神经网络中可逆水印的嵌入需求，进而通过模型剪枝来构造宿主序列(即由筛选出的权重参数构成的序列)，并将其作为水印嵌入的载体；接着对所选参数进行数据预处理，以适应经典的可逆数据隐藏方法；最后通过嵌入并提取哈希值，实现对水印信息是否被修改的有效检测。

ZHAO 等[26]提出的白盒水印方案采用了微调攻击和随机参数篡改攻击对水印进行破坏。微调攻击是对已嵌入水印的模型进行再训练，调整权重参数的高有效位，高位比特的变化导致重构的哈希值与原始水印不匹配。随机参数篡改攻击直接修改水印嵌入层的参数低有效位(LSB)，将参数置零或随

机修改，水印信息被覆盖或混淆，导致提取的哈希值无效。如果篡改攻击发生在水印嵌入层，模型输出标签与预期不匹配，直接触发认证失效。

当前白盒水印方案的安全性通常基于一个假设：攻击者无法在不显著影响模型性能的前提下修改模型结构。基于此假设，这类方案常将水印信息嵌入模型的内部参数或利用激活特性。然而，YAN 等[27]针对深度学习模型的白盒水印技术提出了神经网络结构混淆攻击。该方法能在不降低模型性能、无需重新训练且不了解水印具体方案的情况下，使水印验证失效。如图 6 所示，该攻击通过两种主要方式实现：一是构造输入权重相同但输出权重和为 0 的神经元组；二是将原始神经元拆分为多个功能等效的子神经元。此外，结合权重缩放和随机打乱神经元顺序等操作进一步扰乱模型结构，最终导致嵌入的水印信息失效。

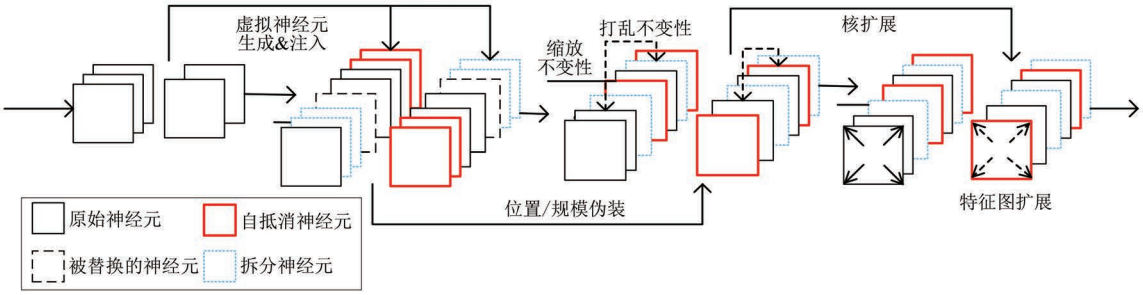


**图 6 神经网络结构混淆的白盒水印攻击**

**Fig. 6 White-box watermarking attack with neural network structure confusion**

3.1.2 基于模型提取的白盒水印攻击

不同于直接修改原始模型参数或结构的微调和剪枝方法，模型蒸馏攻击通过知识蒸馏构建一个轻量化的学生模型，在继承教师模型功能的同时过滤与水印相关的冗余特征，有效移除水印。YANG 等[28]指出，现有水印方案通常将水印信息存储于模型任务相关的冗余中，导致水印与模型核心功能解耦。在蒸馏攻击中，攻击者将原始水印模型作为教师模型，使用真实数据训练一个结构更为紧凑的学生模型，通过应用高温 Softmax 平滑教师的输出概率分布，蒸馏过程促使学生模型专注于学习执行原始任务所需的核心特征，由于水印信息未被整合到这些核心特征中，学生模型在训练过程中会忽略与水印相关的参数或噪声模式，因此从蒸馏得到的学生模型中提取水印的准确率显著下降，而学生模型执行原始任务的精度则基本不受影响。

模型蒸馏攻击相较于微调和剪枝，不直接修改原始模型，避免了结构性损伤的风险，对结构性水印具有良好的去除效果，提供了一种更为安全的水印去除方法。

### 3.2 黑盒水印攻击

黑盒水印攻击是指攻击者在仅能访问目标模型预测接口或有限模型信息的条件下，尝试移除、破坏或规避模型中嵌入的水印，从而削弱模型版权保护机制的攻击行为。根据攻击目标和核心技术手段，黑盒水印攻击可主要分为基于后门去除、基于对抗训练、基于迁移学习和基于生成数据的黑盒水印攻击 4 类。

3.2.1 基于后门去除的黑盒水印攻击

ADI 等[18]提出了一种基于后门植入的黑盒模型水印技术，该技术通过将特定触发样本与固定标签绑定来实现模型版权验证。嵌入后门水印的主要方法包括数据投毒攻击[29-30]和木马攻击[31]。例如，GU 等[32]提出的 BadNets 利用恶意训练，使模型在良性样本上保持正常性能，而在触发样本上表现异常。尽管后门水印已被证明是有效的模型版权保护机制，但其应用也引发了对抗性研究：攻击者试图去除模型中嵌入的后门，从而导致版权验证失效。根据攻击原理和适应场景的不同，基于后门去除的黑盒水印攻击主要分为未标记数据微调攻击、后门触发器攻击、神经网络清洗攻击和规避后门水印检测攻击 4 类。

1）未标记数据微调攻击。

为去除后门水印，CHEN 等[33]提出了基于无标签数据的增强微调（FTAU）方法，该方法利用预训练模型为无标签数据生成伪标签，构建数据集并与少量真实标记数据混合进行微调，实验结果表明：现有水印技术可通过此类微调被有效移除，且攻击者无需知晓水印的具体形式或嵌入方式。类似地，SHAFIEINEJAD 等[34]针对黑盒水印方案，提出了一种黑盒模型窃取攻击，攻击者首先收集一个与原始训练数据同分布且不包含水印触发样本的无标签数据集，并通过应用程序编程接口（API）查询目标水印模型以获取其预测标签，利用该无标签数据集

及对应的伪标签，攻击者训练一个与原始水印模型结构相同的新模型，新模型在保持原有分类性能的同时丧失了对特定水印触发样本的响应能力，即对触发样本表现出随机分类行为，从而实现水印的移除。

2）后门触发器攻击。

后门触发器机制与水印技术的结合利用触发器的隐蔽激活特性实现模型知识产权保护[35-36]，典型的基于后门的水印方案主要有 3 种（如图 7 所示）：基于内容的水印，在干净图像上嵌入有意义的内容（如一小块贴片）；基于噪声的水印，向图像添加特定模式的噪声（如高斯噪声）；基于无关样本的水印，使用来自不同域的图像作为水印载体。

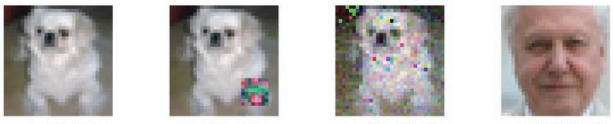


**图 7　3 种基于后门的水印**

**Fig. 7　Three types of watermarks based on backdoors**

在模型训练阶段，研究者将此类嵌入了水印的图像分配一个特定的触发标签，该标签通常与图像的原始内容无关。在训练完成后，在模型验证阶段，向待验证模型输入嵌有相同水印模式的测试图像。若该模型将其分类为预设的触发标签，则可验证该模型的所有权。水印的嵌入与验证过程如图 8 所示。

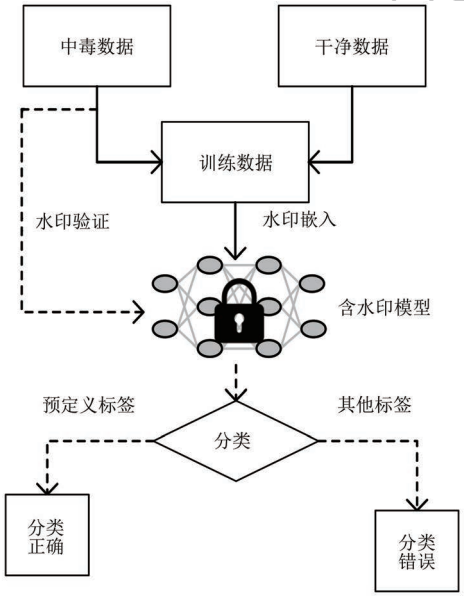


**图 8　基于后门的水印嵌入和验证过程**

**Fig. 8　Watermarking embedding and verification process based on backdoors**

LIU 等[37]评估了现有水印方案的鲁棒性，并提出了一种基于后门的水印去除框架 WILD，WILD 利用有限的干净数据，结合数据增强和特征分布对齐技术，消除 DNN 中的后门触发器，具体而言：WILD 采用随机擦除对干净数据进行增强，在图像中随机选取矩形区域并填充高斯噪声，以模拟水印触发的遮挡模式。随机擦除的增强图像如图 9 所示。特征分布对齐的目标是使增强数据与干净数据在深层特征空间（例如分类层前的倒数第二层）中的分布尽可能相似，尽管它们在输入空间中存在差异。WILD 通过最小化两者在该特征空间中的特征分布距离来实现这一目标。总体损失函数结合了分类损失和特征分布对齐损失，表示如下：

$$\mathcal{L}=\mathcal{L}_{\text{aug}}+\beta\cdot\mathcal{D}(\mathcal{G}(d_{\text{clean}}),\mathcal{G}(d_{\text{aug}})) \tag{5}$$

式中：$\mathcal{L}_{\text{aug}}$ 是使用增强数据（$d_{\text{aug}}$）及其正确标签计算的分类损失；$\mathcal{D}(\mathcal{G}(d_{\text{clean}}),\mathcal{G}(d_{\text{aug}}))$ 是特征分布距离损失项，$d_{\text{clean}}$ 是干净样本；$\beta$ 是超参数，用于调节分布对齐损失的强度。

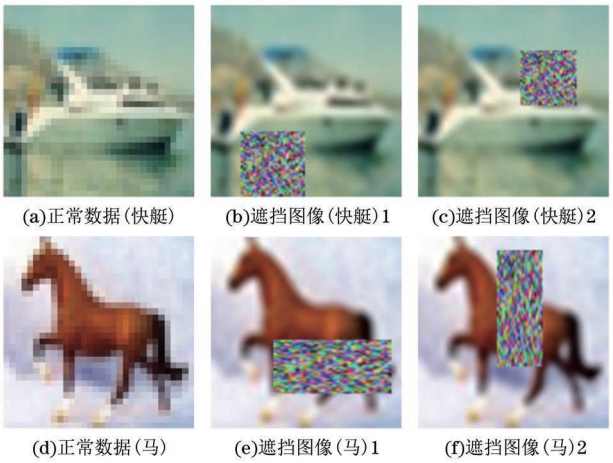


**图 9　随机擦除的增强图像**

**Fig. 9　Randomly erased enhanced images**

与利用数据增强的水印去除方法不同，PUAH 等[38]观察到检测并修改触发样本可导致所有权验证失败，揭示了基于后门触发器的水印方案的脆弱性，提出 BlockDoor 模型封装架构，旨在阻断后门触发式水印验证。该架构针对 3 类典型触发器设计了专用处理模块：（1）对抗样本触发器：检测对抗噪声并通过自编码器净化扰动；（2）分布外样本触发器：训练二分类器区分正常样本与异常分布样本；（3）随机标签触发器：重构特征提取器及分类器修正标签映射。BlockDoor 通过选择性屏蔽水印触发功能，在保留水印结构完整性的同时阻断其验证能力，实现主动防御。

3）神经网络清洗攻击。

AIKEN 等[39]提出一种黑盒水印攻击方法，用于在无需先验知识的情况下清洗 DNN 模型中的后门水印。该方法包含 3 个步骤：水印重构，黑盒清洗

和对抗性再训练。首先，水印重构阶段通过后门重建算法求解优化问题，以重构潜在的触发模式：

$$\min_{m,\Delta}\ell(y_t,f(A(x,m,\Delta)))+\lambda\cdot|m| \tag{6}$$

式中：$A(\cdot)$是水印函数；$x$ 是原始图像；$\Delta$ 是触发器图像；$m$ 是触发器掩码；$f(\cdot)$是模型预测函数；$\ell(\cdot)$是损失函数；$\lambda$ 是控制触发器图案大小的超参数。

该优化输出重构的触发样本 $W_k$。然后，黑盒清洗阶段定位并剪裁与水印相关的神经元。计算特定神经元或通道在正常样本集和重构触发样本集上的平均激活绝对差值：

$$A_j^{\text{diff}}=W_j^{\text{avg}}-N_j^{\text{avg}} \tag{7}$$

式中：$W_j^{\text{avg}}$ 表示神经元或通道在触发样本上的平均激活值；$N_j^{\text{avg}}$ 表示神经元或通道在正常样本上的平均激活值。

若某个神经元 $A_j^{\text{diff}}$ 大于预设阈值，则剪裁或归零该神经元权重；若通道级 $A_j^{\text{diff}}$ 超过阈值，则剪裁或归零整个通道的权重。

最后，执行对抗性再训练以消除残留水印并恢复模型性能。

4）规避后门水印检测攻击。

HITAJ 等[40]指出：当无法有效去除水印时攻击者可在黑盒场景下规避基于后门的水印检测，从而削弱版权保护机制。该团队提出两种攻击策略：其一为集成模型攻击，即攻击者通过窃取多个模型构建集成模型，在处理查询输入时该模型通过集成多数投票机制汇总各子模型的预测结果以决定最终输出类别，若不存在绝对多数类别，则随机返回一个结果，每个模型的水印触发器均独立生成，版权方使用特定水印触发器查询该集成模型时无法稳定获得正确的水印标签，导致验证失败；其二为水印检测器攻击，在仅窃取单一模型的情况下攻击者训练一个二分类检测器区分正常样本与水印样本，对于正常输入，检测器输出其原始预测类别，而对含有水印触发器的输入，则输出随机类别而非对应水印标签，从而显著降低版权验证成功率。

3.2.2 基于对抗训练的黑盒水印攻击

在基于对抗训练[41-42]的黑盒水印攻击中，根据攻击发生的阶段不同，主要可分为基于代理模型攻击和基于条件生成对抗网络（CGAN）的再生攻击两类。

1）基于代理模型攻击。

在水印验证过程中，攻击者可利用代理模型模拟真实水印检测器的行为，生成对抗样本以破坏水印检测或身份识别功能。

AN 等[43]提出训练一个二分类神经网络作为代理水印检测器，学习区分含水印图像与非含水印图像，并对目标含水印图像应用投影梯度下降（PGD）攻击生成对抗样本，可表示如下：

$$\min_{x_{\text{adv}}}\mathcal{L}(f(x_{\text{adv}}),y_{\text{target}}),\ \text{s.t.}\ \|x_{\text{adv}}-x\|_\infty\leqslant\varepsilon \tag{8}$$

式中：$f$ 为代理检测器；$x_{\text{adv}}$ 为生成的对抗样本；$y_{\text{target}}$ 为攻击目标标签；$\mathcal{L}$ 是交叉熵损失函数；$\varepsilon$ 为扰动阈值。

实验结果表明，扰动后的图像 $x_{\text{adv}}$ 能显著降低原始水印检测器的检测率。

受到训练代理检测器去除水印的启发，LIN 等[44]提出一种针对 Tree-Ring 水印[45]的新型攻击方法，该方法的核心思想是利用公开可用的变分自编码器（VAE）来近似目标扩散模型的潜在空间，从而在潜在空间中训练一个代理检测器，并通过 PGD 生成对抗样本，使得 Tree-Ring 水印检测器失效，同时保持图像质量，实验结果表明，即使仅具备黑盒查询能力，攻击者仍可借助公开 VAE 实施有效的潜在空间攻击，从而挑战 Tree-Ring 水印的鲁棒性假设。

QUIRING 等[46-47]进一步研究了基于代理模型的黑盒水印攻击。在文献[46]中，该团队训练一个代理 DNN 模型来逼近目标检测器的二分类决策边界，但采用了小波域高频系数作为模型输入特征，基于该代理模型使用梯度下降方法优化扰动，生成对抗性高频系数，再与原始低频系数重构得到空域对抗样本，成功欺骗原始检测器。文献[47]在文献[46]的基础上进行了重要扩展：(1)采用了更复杂的代理模型结构；(2)将优化目标改进为双距离优化；(3)提出了一个黑盒水印攻击的统一框架。该框架的核心为攻击者利用黑盒访问权限训练代理模型近似目标检测器，再求解以下优化问题生成对抗扰动[48]：

$$\text{minimize}\ c\|\tilde{z}+\boldsymbol{t}\|_2+\hat{F}(\tilde{z}+\boldsymbol{t}) \tag{9}$$

式中：$\tilde{z}$ 表示目标含水印图像的高频系数；$\hat{F}$ 为代理模型相关的目标函数；$\|\tilde{z}+\boldsymbol{t}\|_2$ 控制扰动向量 $\boldsymbol{t}$ 的幅度；$c$ 为控制扰动幅度的权衡参数。

优化得到的扰动 $\boldsymbol{t}$ 作用于重构的图像能使原始水印检测器输出错误结果，即水印被成功去除。

2）基于 CGAN 的再生攻击。

CHENG 等[49]将水印检测建模为目标检测任务，利用大规模多样化数据集提出一个端到端框架，同时进行水印检测与去除。如图 10 所示，该检测框

架以带水印图像为输入，输出图像中不同位置、尺度和宽高比的水印覆盖区域的概率估计。

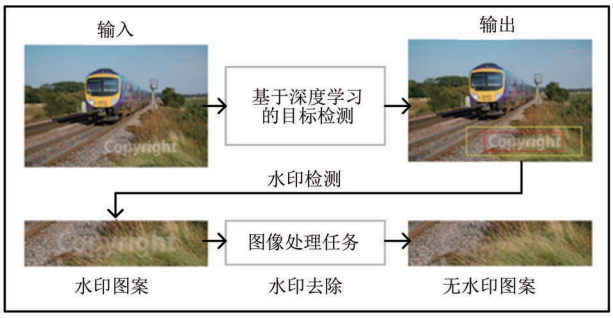


**图 10 基于图像处理任务的水印去除框架**

**Fig. 10 Watermarking removal framework based on image processing tasks**

然而，LI 等[50]指出：虽然 CHENG 等[49]将水印去除视为图像处理任务，其结果仍存在残留水印痕迹。为此，该团队提出了一种基于 CGAN 的水印去除方法。生成器 $G$ 学习从带水印图像到真实无水印图像的映射，判别器 $D$ 评估生成图像是否与真实图像一致。模型通过最小化组合损失函数进行训练。

对抗损失 $\mathcal{L}_{adv}$ 可表示如下：

$$E_{x,y}[\log_a D(x,y)]+E_x[\log_a(1-D(x,G(x)))] \tag{10}$$

内容损失 $\mathcal{L}_{L1}$ 可表示如下：

$$\mathcal{L}_{L1}=\|G(x)-y\|_1 \tag{11}$$

感知损失 $\mathcal{L}_{per}$ 可表示如下：

$$\mathcal{L}_{per}=\|\Phi_j(G(x))-\Phi_j(y)\|_2^2 \tag{12}$$

总目标损失函数 $\mathcal{L}$ 可表示如下：

$$\mathcal{L}=\mathcal{L}_{adv}+\alpha\mathcal{L}_{L1}+\beta\mathcal{L}_{per} \tag{13}$$

式中：$x$ 为带水印图像；$y$ 为无水印真实图像；$G(x)$ 为生成图像；$D(x,y)$ 为判别器 $D$ 判定真实图像对 $(x,y)$ 为真的概率；$D(x,G(x))$ 为判定生成图像对 $(x,G(x))$ 为真的概率。

该损失组合旨在提升生成图像的视觉质量并保留细节。

图 11 对比了文献[49-50]研究工作的结果，图 12 展示了基于 CGAN 的水印去除框架。

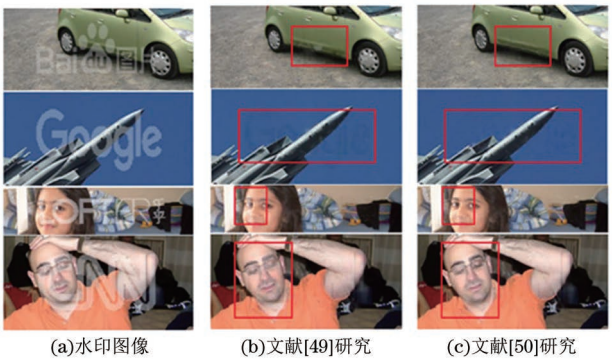


**图 11 研究工作对比**

**Fig. 11 Comparison of research works**

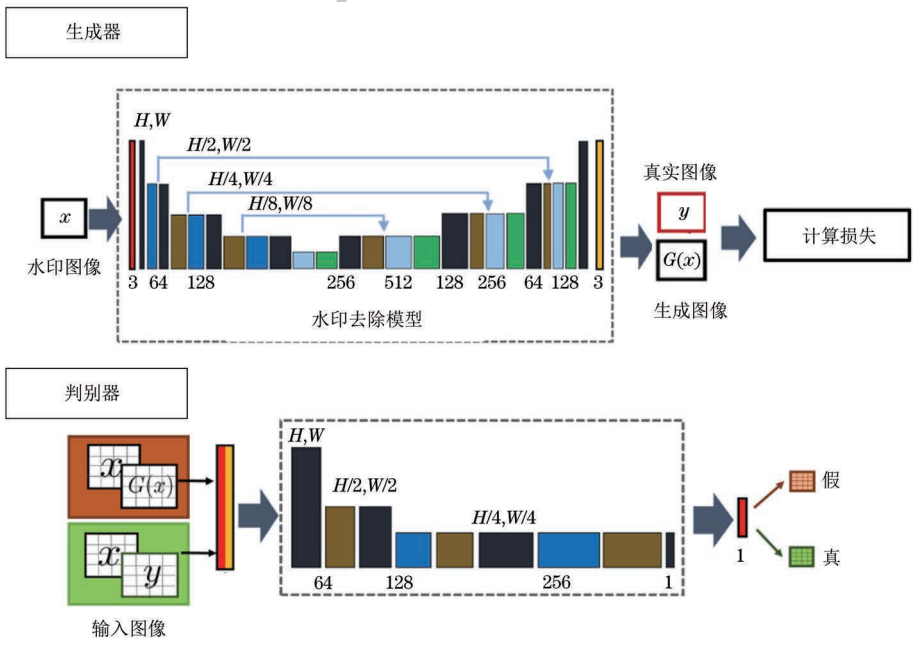


**图 12 基于 CGAN 的水印去除框架**

**Fig. 12 Watermarking removal framework based on CGAN**

在 CGAN 框架的基础上，CAO 等[51]进一步提出 VWGAN，它是一种结合生成对抗网络和自注意力机制的可见水印去除方法。VWGAN 的核心思想是将水印视为局部噪声，利用自注意力层自动学习水印区域的关键特征，无需预先定位水印位置。与使用传统 $\mathcal{L}_{L1}$ 损失导致模糊的生成对抗

网络不同，VWGAN 采用结构相似性（SSIM）损失 $\mathcal{L}_{\mathrm{SSIM}}$：

$$\mathcal{L}_{\mathrm{SSIM}}(G)=E_{x,y,z}[\|y-G(x,z)\|_{s}] \tag{14}$$

式中：$x$ 表示带水印的输入图像；$y$ 表示真实无水印图像；$G(x,z)$表示生成器输出的预测图像；$\|\cdot\|_{s}$ 表示 SSIM 距离度量。

该损失函数通过约束生成图像与真实无水印图像在亮度、对比度和结构上的感知一致性，有效解决了像素级损失导致的模糊问题。

3.2.3 基于迁移学习的黑盒水印攻击

迁移学习通过将源领域获得的知识迁移到相关但不同的目标领域，旨在提升目标任务性能。该方法尤其适用于目标领域数据稀缺或计算资源受限的场景，通常涉及对预训练模型进行微调以适应新任务。LUKAS 等[52]将迁移学习应用于 DNN 模型水印去除。该方法采用在无关领域数据集上预训练的模型，替换其最后一层（分类层）以匹配目标任务的类别数。训练过程分为两个阶段：第一，冻结除新分类层以外的所有权重，使用目标数据集及其真实标签训练该分类层；第二，解冻所有权重并微调整个模型，同时逐步降低学习率。水印去除成功的判定标准为：模型输出的水印信息与原始嵌入水印的匹配率低于预设阈值。研究表明：迁移学习能有效去除多种水印，其有效性源于微调过程修改了模型的底层权重与特征表示，而这些特征正是水印嵌入所依赖的基础。

3.2.4 基于生成数据的黑盒水印攻击

近年来，扩散技术在图像生成、图像修复和超分辨率等领域取得了显著进展。受此启发，研究者将其引入水印攻击领域，旨在通过修改模型生成过程或参数，在生成的图像中嵌入或去除水印信息，同时保持视觉质量。在基于生成数据的黑盒水印攻击中，根据其攻击目标的不同主要分为基于扩散模型的再生攻击和规避水印检测攻击两类。

1）基于扩散模型的再生攻击。

基于扩散模型的再生攻击旨在通过生成模型的强大再生能力，从带水印内容中重构出视觉无损的无水印版本。

LI[53]提出的 DiffWA 首次将扩散模型应用于水印去除。该方法采用基于距离引导的条件生成框架，以水印图像为输入，训练扩散模型生成无水印图像。采样过程中，通过距离损失函数约束生成图像与水印图像的相似性（可能结合水印嵌入特征，如低频/高频分量）指导模型重构。此外，LI 等[53]提出一种估计器以加速推理，并结合不同水印攻击模型以提升水印去除效果。

不同于上述基于距离引导的条件扩散模型方法，ZHAO 等[54]提出用于图像水印去除的再生攻击，该方法分为破坏和重建两阶段：在破坏阶段，将水印图像通过嵌入函数映射到特征空间，注入高斯噪声破坏水印信息；在重建阶段，使用生成模型从噪声嵌入中重建图像，重建后的图像保持高视觉质量且水印被有效去除。计算公式可表示如下：

$$\hat{x}=\mathcal{A}(\varphi(x_{w})+\mathcal{N}(0,\sigma^{2}I_{d})) \tag{15}$$

式中：$\varphi(x_{w})$是水印图像 $x_{w}$ 的特征嵌入函数；$\mathcal{N}(0,\sigma^{2}I_{d})$为破坏水印的高斯噪声，$\sigma$ 表示噪声强度控制参数，$d$ 表示特征空间维度。

该方法的攻击框架如图 13 所示。

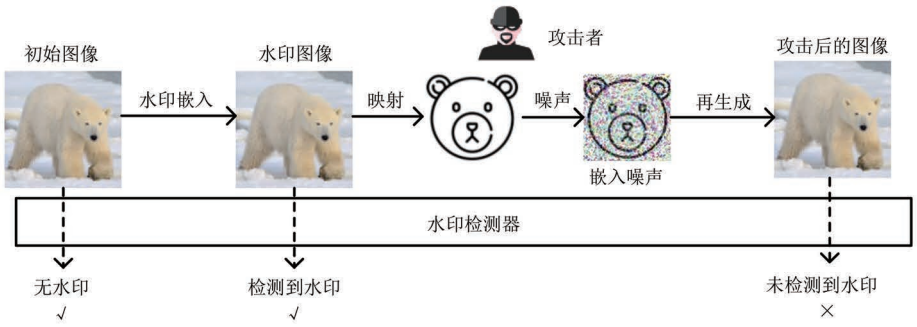


**图 13 再生攻击的水印去除框架**

**Fig. 13 Watermarking removal framework for regenerative attacks**

与 ZHAO 等[54]通过噪声注入破坏特征空间的思路不同，HU 等[55]直接针对水印的嵌入结构提出解码器微调攻击。该方法聚焦水印编解码框架的脆弱性，通过修改扩散模型解码器参数破坏水印完整性，生成视觉无损的无水印图像。

其攻击流程分为两个关键阶段，如图 14 所示。首先，进行目标潜在变量估计：当水印系统的编码器已知时，直接通过编码器将原始无水印图像转换为对应的潜在表示；当编码器未知时，通过优化算法调整潜在变量，使水印解码器的输出逼近原始无水印图像。随后，执行解码器参数微调：基于前阶段获得的潜在表示，调整水印解码器的模型参数。优化过程

联合 3 种约束:像素级重建约束保证图像结构一致性,感知质量约束维持视觉逼真度,对抗训练约束增强水印去除效果。最后,使微调后的解码器能够利用该潜在表示生成与原始无水印图像高度一致的结果。

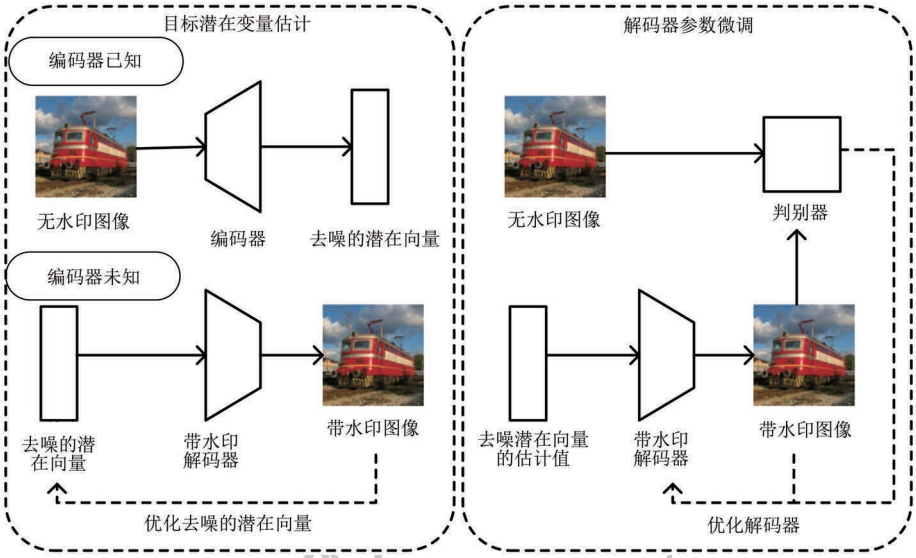


**图 14 微调扩散模型解码器的水印攻击框架**

**Fig. 14 Watermarking attack framework for fine-tuning diffusion model decoder**

2)规避水印检测攻击。

规避水印检测攻击旨在操纵水印检测器的解码输出,使其无法正确识别水印,而非移除水印本身。

由于生成式 AI 技术的快速发展和水印检测在内容真实性验证中的关键作用,JIANG 等[56]利用水印技术对对抗扰动的脆弱性,提出 WEvade 框架,如图 15 所示。该框架通过向水印图像注入视觉不可察的微小扰动,操纵检测器解码输出至随机猜测水平以规避检测:白盒场景下利用已知解码器添加扰动,使水印图像解码准确率低于阈值 $\theta$ 而被误判为非 AI 生成;黑盒场景下适配 HopSkipJump[57],被识别为非 AI 生成的随机图像,迭代优化扰动使其持续规避检测并逼近原水印图像。

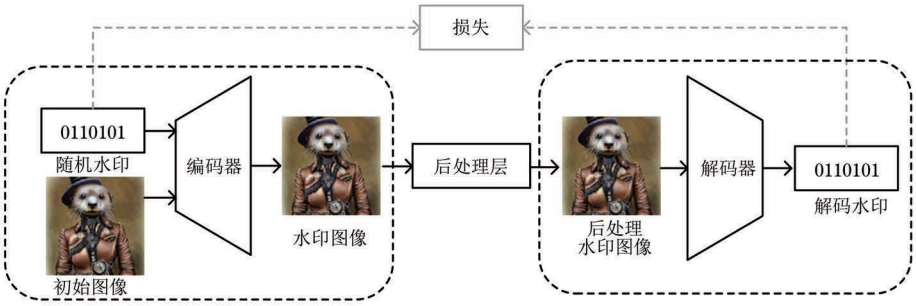


**图 15 基于生成数据的水印规避检测框架**

**Fig. 15 Watermarking avoidance detection framework based on generated data**

基于 JIANG 等[56]提出的水印对抗攻击框架,HU 等[58]提出了一种针对图像水印检测系统的迁移攻击,旨在通过生成微小扰动,使水印图像能够有效规避检测。该方法的核心在于攻击者利用自主训练的一组多样化代理水印模型,通过集成优化方式协同生成对抗样本,从而实现对未知目标水印模型的高效攻击。具体而言,该方法提出了基于多代理模型集成的扰动生成机制,通过优化使所有代理模型对扰动图像解码出预设目标水印,显著增强了攻击的迁移能力,其中所采用的逆解码策略通过反转代理模型对原图像的解码结果生成目标水印,进一步提高了攻击效率。实验结果表明,该方法对 HiDDeN[59]、StegaStamp[60]等多种主流图像水印方案的规避成功率超过 85%,且所引入的扰动远小于传统图像处理方法造成的失真。

为了克服 HU 等[58]提出的迁移攻击在真实场景下面临的替代模型训练成本高、扰动痕迹可能过大的局限性,ZHAO 等[61]提出了一种 CtrlRegen 的可控再生水印去除方法,该方法无需依赖任何替代模型或检测器知识,基于纯净高斯噪声进行可控图像再生,实现了扰动水印的高效、高质量去除。该方法的实验结果表明,即使不进行模型特

定训练，仅凭先进的生成式 AI 技术也能实现对水印的有效攻击，为评估水印方案的鲁棒性设立了新的标杆，并强调了生成式 AI 进步对水印技术带来的挑战。

与这种统一的生成式攻击思路不同，SABERI 等[62]提出分级攻击框架：对弱鲁棒性水印采用扩散净化攻击，通过注入高斯噪声破坏统计特征后经扩散模型去噪生成视觉一致的无水印图像；对强鲁棒性水印实施代理对抗攻击，通过训练代理检测器模拟目标决策边界、生成可迁移对抗样本实现检测规避。

本章主要对 DNN 水印攻击方法进行了详细的梳理与总结。白盒水印攻击凭借对模型内部状态的完全访问，通常具备更高的攻击效率与成功率，然而该优势也严格受限于其对模型参数的依赖性，因而在实际应用中普适性较低。相较之下，黑盒水印攻击仅依赖模型输入输出交互，适用性更广，但其攻击效果往往受到数据质量、代理模型准确性以及模型生成能力等多种因素的制约。因此，在实际攻击场景中，需综合考虑模型访问权限、水印嵌入机制、可用数据资源等关键因素，以选择合适的攻击方法实现水印的有效破坏或去除。

## 4 DNN 水印攻击方法对比分析

AI 安全领域的模型水印攻击研究不仅揭示了版权保护的系统性漏洞与防御瓶颈，也为模型安全提供了重要的评估参考。表 2 对现有 DNN 模型水印攻击方法中常用到的数据集进行总结，数据集的选择需匹配水印攻击的目标模型任务，不同任务直接影响攻击方法的验证维度[63]。图像分类数据集中，CIFAR-10/100 数据规模适中，适用于轻量级水印攻击实验，ImageNet 则用于验证大规模模型攻击的扩展性。人脸识别数据集（如 YouTube Aligned Face 和 VGG-FACE）适用于测试对身份认证水印的攻击效果。在评估针对扩散模型的水印攻击时，领域专用数据集[64-66]（如 DiffusionDB）则发挥着关键作用。当前的研究正朝着 3 个主要数据集发展趋势进行 DNN 模型水印攻击方法的设计与评估：一是从通用数据集向领域专用数据扩展；二是从静态图像向多模态动态数据演进[67-68]；三是日益重视生成式数据集。通过遵循这些趋势，研究者能够开发出更有效的攻击策略，并更准确地评估水印技术在日益复杂的数据环境下的鲁棒性，从而间接促进水印系统的安全性设计。

**表 2 DNN 模型水印攻击中的常用数据集**

**Table 2 Common datasets of watermarking attacks for DNN models**

| 数据集 | 数据集类别 | 相关文献 |
|---|---|---|
| CIFAR-10 | 图像分类 | 文献[4,11,18,21,26-28,33-34,37-39,53] |
| CIFAR-100 | 图像分类 | 文献[11,18,21,27,33-34,38] |
| ImageNet | 图像分类、目标检测 | 文献[11,18,25,40,50-51,55-56,58] |
| STL-10 | 迁移学习 | 文献[11,33] |
| MNIST | 手写数字识别 | 文献[26,28,32,34,37,39,40] |
| DiffusionDB | 生成模型训练 | 文献[43,55-56] |
| MS COCO | 目标检测 | 文献[27,43,52,54-56] |
| YouTube Aligned Face | 人脸识别 | 文献[40,51] |
| VGG-FACE | 人脸识别 | 文献[43,51] |
| Dresden Image Database | 图像取证 | 文献[46-47] |

表 3 总结了现有 DNN 水印攻击研究中常用的模型，旨在为研究者根据具体场景选择合适的攻击策略提供参考。模型结构特性的差异（包括参数冗余度、架构设计和任务专一性等）是影响攻击效果的核心因素[69]。这直接体现在：ResNet 凭借其高参数冗余度常被用于测试水印鲁棒性[70]；Stable Diffusion的兴起推动了生成式模型版权攻防研究

**表 3 DNN 模型水印攻击中的常用模型**

**Table 3 Common models of watermarking attacks for DNN models**

| 模型 | 模型任务 | 相关文献 |
|---|---|---|
| 宽残差网络 | 图像分类 | 文献[4,24] |
| ResNet | 图像分类 | 文献[11,18,21,25-27,33-34,38,40,49-50,52-53] |
| VGG | 图像分类 | 文献[11,25-26,34,37-38,50-51] |
| AlexNet | 图像分类 | 文献[25-26,39] |
| DenseNet | 图像分类 | 文献[25] |
| MobileNet | 图像分类 | 文献[25,38] |
| LeNet | 手写数字识别 | 文献[26,34,37] |
| 简单卷积网络 | 图像分类 | 文献[40,46-47] |
| DCGAN | 图像生成 | 文献[27] |
| DeepID | 人脸识别 | 文献[39] |
| Transformer(ViT) | 图像分类 | 文献[38] |
| Stable Diffusion | 图像生成 | 文献[43,54-56] |
| Tree-Ring | 生成模型水印 | 文献[43,52,62] |
| StegaStamp | 水印嵌入 | 文献[43,62] |
| DiffWA | 水印去除 | 文献[53] |
| HiDDeN | 水印嵌入 | 文献[53,56] |

的深入[71]；对 Tree-Ring 等模型的研究则揭示了频域水印的结构性弱点[72]。这些研究表明：当前模型选择的一个重要趋势是紧跟 AI 前沿发展，尤其在生成式 AI 崛起背景下，针对那些具备多样化、特定化和独特结构特性的目标模型展开攻击研究。更重要的是，此类研究不仅验证了攻击方法的有效性，其发现反过来直接驱动了防御技术与水印嵌入方案的创新，显著提升了水印系统的整体安全性。

模型水印技术是 DNN 知识产权的重要手段，其有效性直接取决于水印的鲁棒性。随着对水印进行系统性安全评估需求的日益增长，针对模型水印的攻击技术研究也取得了显著进展。表4针对

**表 4 DNN 模型水印攻击方法对比**

**Table 4 Comparison of watermarking attack methods for DNN models**

| 攻击方法 | 白盒/黑盒攻击 | 攻击分类 | 攻击者能力 | | | 性能影响 |
|---|---|---|---|---|---|---|
| | | | 数据 | 模型 | 水印知识 | |
| 文献[4]方法 | 白盒水印攻击 | 基于水印覆盖攻击 | × | √ | × | 模型原水印准确率下降至 50%，水印无法被有效提取或验证 |
| 文献[11]方法 | | 基于微调攻击 | √ | √ | √ | 有限标注数据下，在 CIFAR-10 和 CIFAR-100 上分类任务准确率下降 2～5 百分点 |
| 文献[21]方法 | | 基于微调攻击 | √ | √ | × | 模型任务性能损失小于等于 5 百分点，对模型功能影响较小 |
| 文献[25]方法 | | 基于剪枝攻击 | × | √ | × | 模型任务性能损失小于等于 0.5 百分点，对模型功能影响较小 |
| 文献[26]方法 | | 基于微调攻击 | × | √ | × | 水印嵌入在 LSB 中，不影响模型原始准确率 |
| 文献[27]方法 | | 基于神经网络结构混淆的攻击 | × | √ | × | 平均水印误码率(BER)大于等于 52%，水印无法被有效提取和验证 |
| 文献[28]方法 | | 基于模型蒸馏攻击 | √ | √ | × | 平均水印 BER 大于等于 50%，水印无法被有效提取和验证 |
| 文献[33]方法 | 黑盒水印攻击 | 基于未标记数据微调攻击 | × | × | × | 模型任务性能损失小于等于 1 百分点，对模型功能影响较小 |
| 文献[36]方法 | | 基于未标记数据微调攻击 | × | × | × | 在 CIFAR-10 和 ImageNet 上准确率损失 3.7～4.1 百分点 |
| 文献[37]方法 | | 基于后门触发器攻击 | √ | × | × | 在 MNIST 和 CIFAR-10 上准确率平均下降 0.3～1.7 百分点 |
| 文献[38]方法 | | 基于后门触发器攻击 | √ | × | √ | 水印保留率小于等于 15%，水印验证失效 |
| 文献[39]方法 | | 基于神经网络清洗攻击 | √ | × | × | 清洗后主任务准确率约从 99%降至 97%，损失约 2 百分点 |
| 文献[42]方法 | | 基于规避后门水印检测攻击 | × | × | × | 原始水印验证率 100%，攻击后验证率小于等于 34%，水印验证失效 |
| 文献[43]方法 | | 基于代理模型攻击 | × | × | × | 对嵌入攻击敏感，再生攻击可完全破坏水印 |
| 文献[46]方法 | | 基于代理模型攻击 | √ | × | × | 峰值信噪比(PSNR)大于等于 35.6 dB，输出图像视觉无损，攻击成功率 100% |
| 文献[47]方法 | | 基于代理模型攻击 | √ | × | × | PSNR 大于等于 39 dB，输出图像视觉无损，攻击成功率 100% |
| 文献[49]方法 | | 基于 CGAN 的再生攻击 | √ | × | × | PSNR 为 30.86 dB，输出图像视觉无损 |
| 文献[50]方法 | | 基于 CGAN 的再生攻击 | √ | √ | × | PSNR 为 30.69 dB，DSSIM 为 0.045，输出图像视觉无损，且与原始图像相似 |
| 文献[51]方法 | | 基于 CGAN 的再生攻击 | √ | √ | × | PSNR 大于 35 dB，SSIM 大于 0.96，输出图像视觉无损，且与原始图像相似 |
| 文献[52]方法 | | 基于迁移攻击 | √ | √ | × | 攻击后水印检测准确率降至小于等于 6.3% |
| 文献[53]方法 | | 基于扩散模型的再生攻击 | √ | √ | √ | BER 大于 0.4，PSNR 大于 33 dB，SSIM 大于 0.98，水印被有效提取，输出图像视觉无损，且与原始图像相似 |
| 文献[54]方法 | | 基于扩散模型的再生攻击 | √ | × | × | PSNR 大于 30 dB，水印移除率大于 98%，输出图像视觉无损，水印移除率高 |
| 文献[55]方法 | | 基于扩散模型的再生攻击 | √ | √ | × | 攻击成功率大于 94%，FID(Fréchet Inception Distance)小于 26，水印验证失效且生成图像质量高 |
| 文献[56]方法 | | 基于规避水印检测攻击 | √ | √ | × | 攻击成功率大于 94.5%，水印验证失效 |
| 文献[62]方法 | | 基于规避水印检测攻击 | √ | × | × | 低扰动水印和 Tree-Ring 水印的受试者工作特征曲线下面积(AUROC)小于 0.65，水印检测器功能失效 |

DNN 模型水印攻击方法进行总结对比，从攻击方法、攻击分类、攻击者能力及性能影响 4 个维度对比现有工作。其中：攻击者能力是评估攻击是否具备所需资源（即攻击者对数据、模型和水印知识等）的访问权限；性能影响是指水印攻击方法对模型多项关键指标所产生的作用，主要包括被攻击模型的任务准确率、水印提取率、攻击成功率以及图像质量评估指标[73-75]（如 SSIM、FID、BER、SSIM、AUROC 等）的变化。这些指标是评估攻击方法综合效果的核心依据。

在 DNN 水印攻击方法对比中，表 4 所提供的详细数据为进一步理解不同攻击方法的有效性和局限性提供了重要依据。首先，白盒攻击方法（如 CHEN 等[11]和 GUAN 等[25]提出的基于微调和剪枝的方法）由于可直接访问并修改模型参数，在 CIFAR-10 和 ImageNet 等通用图像分类数据集上能够以极低的主任务性能损失（普遍下降 2～5 百分点）实现高效的水印去除效果。这表明白盒水印攻击方法在保持模型功能完整性的同时能够精准破坏水印结构，尤其适用于冗余参数较多、结构开放的模型。相比之下，黑盒水印攻击方法更依赖于间接手段与推理策略，其效果受数据质量与代理模型性能影响显著。例如：CHEN 等[33]基于未标记数据微调的方法虽在一定程度上维持了主任务准确率（损失约 0.3～4.1 百分点），但其水印去除成功率波动较大；QUIRING 等[46-47]基于代理模型攻击的方法尽管在黑盒场景下实现了高攻击成功率（100%）与良好的输出质量（PSNR 大于等于 35.6 dB），但其性能高度依赖于数据的代表性与代理模型的性能。此外，针对不同类型的模型，其评估准则也存在差异。对于生成模型的水印攻击，评估指标从分类准确率转向感知质量指标[76]。例如，LI 等[53]和 ZHAO 等[54]所采用的基于扩散模型的再生攻击虽不直接影响分类精度，但可通过高 SSIM、低 FID 等指标表明其能在维持视觉质量的同时有效去除水印，而在 SABERI 等[62]的方法中 AUROC 降至 0.14 表明水印检测系统被破坏的程度，进一步体现出生成任务中水印攻击的特殊性。

## 5 总结与展望

DNN 模型水印技术在 AI 安全领域日益应用广泛，更加凸显了模型版权保护研究的重要性。但是，模型本身易受窃取与恶意篡改攻击，带来了显著安全风险，因此深入理解水印攻击方法十分必要。本文首先将水印攻击方法宏观地划分为白盒与黑盒两大类，随后进一步细化：白盒攻击按实现方式分为基于模型修改和基于模型提取两类；黑盒攻击分为基于后门去除、对抗训练、迁移学习及生成数据 4 类。在此基础上，从攻击者能力和性能影响等维度，对比分析了现有各类水印攻击方法，并归纳了其各自的优势与局限性。

随着深度模型应用的日益普及，保护模型知识产权的需求愈发迫切，但深度模型水印技术尚处在初级阶段，在鲁棒性[77]、安全性[78]等方面仍存在许多关键问题有待进一步研究。在未来的 DNN 模型版权保护方面，有 3 个方面可以重点关注：

1）多用户部署的模型水印冲突解决机制。随着模型即服务（MaaS）[79]的广泛应用，同一平台需为众多用户提供模型定制与部署服务，如何确保不同用户水印的唯一性与可区分性已成为亟待突破的挑战。未来研究应致力于构建抗冲突的水印生成与验证架构，例如将水印与模型指纹相耦合[80-81]，在为用户嵌入水印时，将其密钥与模型的固有特征进行绑定，从而防止水印冲突。

2）拓展生成模型水印研究[82]。人工智能生成内容（AIGC）技术[83]正生成大量文本、图像、视频和音频内容，其广泛传播引发了严峻的版权问题。与此同时，水印技术面临的挑战也在持续升级，特别是基于扩散模型的生成式内容重建技术已成为去除水印的有效手段，这标志着水印攻防步入以生成式 AI 为核心的新阶段。然而，当前针对生成模型水印技术的研究仍处于早期探索阶段，成果尚显不足。因此，未来应加强对生成模型水印技术的研究，探索有效保护 AI 生成内容版权的方案，以应对新兴的版权挑战。

3）构建通用评估体系。当前 DNN 模型水印技术缺乏通用的评估指标体系，难以系统地衡量不同水印嵌入方案在隐蔽性、抗攻击性等关键指标上的表现，尤其在面对多样化水印技术时，现有评估方法往往局限于单一场景或特定攻击类型，导致水印方案之间对比性不足。未来的水印技术研究可以着手构建通用的评估体系，以确保对不同类型的水印技术都能进行全面评估。这将有助于推动版权保护的标准化与 DNN 模型水印技术的共同发展。

### 参考文献

[1] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA:

IEEE Press, 2016: 779-788.
[2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Philadelphia, USA: ACL Press, 2019: 4171-4186.
[3] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. New York, USA: ACM Press, 2017: 269-277.
[4] WANG T H, KERSCHBAUM F. Attacks on digital watermarks for deep neural networks[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D.C., USA: IEEE Press, 2019: 2622-2626.
[5] 何春辉, 葛斌, 徐浩, 等. 版权保护视角下的图像水印处理技术综述[J]. 计算机技术与发展, 2025, 35(8): 1-9.
HE C H, GE B, XU H, et al. Survey of image watermarking processing technology from perspective of copyright protection [J]. Computer Technology and Development, 2025, 35(8): 1-9. (in Chinese)
[6] 谢宸琪, 张保稳, 易平. 人工智能模型水印研究综述[J]. 计算机科学, 2021, 48(7): 9-16.
XIE C Q, ZHANG B W, YI P. Survey on artificial intelligence model watermarking [J]. Computer Science, 2021, 48(7): 9-16. (in Chinese)
[7] 夏道勋, 王林娜, 宋允飞, 等. 深度神经网络模型数字水印技术研究进展综述[J]. 科学技术与工程, 2023, 23(5): 1799-1811.
XIA D X, WANG L N, SONG Y F, et al. Review of deep neural network model digital watermarking technology[J]. Science Technology and Engineering, 2023, 23(5): 1799-1811. (in Chinese)
[8] 吴汉舟, 张杰, 李越, 等. 人工智能模型水印研究进展[J]. 中国图象图形学报, 2023, 28(6): 1792-1810.
WU H Z, ZHANG J, LI Y, et al. Overview of artificial intelligence model watermarking[J]. Journal of Image and Graphics, 2023, 28(6): 1792-1810. (in Chinese)
[9] 金彪, 林翔, 熊金波, 等. 基于水印技术的深度神经网络模型知识产权保护[J]. 计算机研究与发展, 2024, 61(10): 2587-2606.
JIN B, LIN X, XIONG J B, et al. Intellectual property protection of deep netural network models based on watermarking technology[J]. Journal of Computer Research and Development, 2024, 61(10): 2587-2606. (in Chinese)
[10] 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 (CVPR). Washington D.C., USA: IEEE Press, 2025: 3008-3018.
[11] CHEN X Y, WANG W X, BENDER C, et al. REFIT: a unified watermark removal framework for deep learning systems with limited data[C]//Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. New York, USA: ACM Press, 2021: 321-335.
[12] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences, 2017, 114(13): 3521-3526.
[13] MIYATO T, MAEDA S I, KOYAMA M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979-1993.
[14] GRANDVALET Y, BENGIO Y. Semi-supervised learning by entropy minimization [C] // Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2004: 529-536.
[15] GOODFELLOW I J, MIRZA M, XIAO D, et al. An empirical investigation of catastrophic forgetting in gradient-based neural networks [EB/OL]. [2025-06-04]. https://arxiv.org/abs/1312.6211.
[16] KEMKER R, MCCLURE M, ABITINO A, et al. Measuring catastrophic forgetting in neural networks [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2018, 32(1): 1-8.
[17] COOP R, MISHTAL A, AREL I. Ensemble learning in fixed expansion layer networks for mitigating catastrophic forgetting[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(10): 1623-1634.
[18] ADI Y, BAUM C, CISSE M, et al. Turning your weakness into a strength: watermarking deep neural networks by backdooring[C]//Proceedings of the 27th USENIX Security Symposium (USENIX Security 18). Austin, USA: USENIX Association, 2018: 1615-1631.
[19] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[C]//Proceedings of International Symposium on Research in Attacks, Intrusions, and Defenses. Berlin, Germany: Springer International Publishing, 2018: 273-294.
[20] ZHANG J L, GU Z S, JANG J, et al. Protecting intellectual property of deep neural networks with watermarking[C]// Proceedings of the 2018 on Asia Conference on Computer and Communications Security. New York, USA: ACM Press, 2018: 159-172.
[21] GUO S W, ZHANG T W, QIU H, et al. Fine-tuning is not enough: a simple yet effective watermark removal attack for DNN models[C]//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 2021: 3635-3641.
[22] ROUHANI B D, CHEN H L, KOUSHANFAR F. DeepSigns: an end-to-end watermarking framework for ownership protection of deep neural networks [C] // Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems. New York, USA: ACM Press, 2019: 485-497.
[23] 李珮玄, 黄土, 罗书卿, 等. 深度学习模型版权保护技术研究综述[J]. 信息安全学报, 2025, 10(1): 17-35.
LI P X, HUANG T, LUO S Q, et al. A survey on copyright protection technology of deep learning model[J]. Journal of Cyber Security, 2025, 10(1): 17-35. (in Chinese)
[24] KROGH A, HERTZ J A. A simple weight decay can improve generalization [C] // Proceedings of the 5th International Conference on Neural Information Processing Systems. San Francisco, USA: Morgan Kaufmann Publishers Inc., 1991: 950-957.
[25] GUAN X Q, FENG H M, ZHANG W M, et al. Reversible watermarking in deep convolutional neural networks for integrity authentication[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: ACM Press, 2020: 2273-2280.
[26] ZHAO G J, QIN C, YAO H, et al. DNN self-embedding watermarking: towards tampering detection and parameter recovery for deep neural network[J]. Pattern Recognition Letters, 2022, 164: 16-22.
[27] YAN Y, PAN X, ZHANG M, et al. Rethinking white-box watermarks on deep learning models under neural structural obfuscation[C] // Proceedings the 32nd USENIX Security

Symposium (USENIX Security 23). Austin, USA: USENIX Association 2023: 2347-2364.
[28] YANG Z, DANG H, CHANG E C. Effectiveness of distillation attack and countermeasure on neural network watermarking[EB/OL]. [2025-06-04]. https://arxiv.org/abs/1906.06046.
[29] WANG B L, YAO Y S, SHAN S, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks[C]//Proceedings of the IEEE Symposium on Security and Privacy (SP). Washington D.C., USA: IEEE Press, 2019: 707-723.
[30] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[EB/OL]. [2025-06-04]. https://arxiv.org/abs/1712.05526.
[31] LIU Y Q, MA S Q, AAFER Y, et al. Trojaning attack on neural networks[C]//Proceedings of 2018 Network and Distributed System Security Symposium. San Diego, USA: Internet Society, 2018: 27-41.
[32] GU T, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain[EB/OL]. [2025-06-04]. https://arxiv.org/abs/1708.06733.
[33] CHEN X, WANG W, DING Y, et al. Leveraging unlabeled data for watermark removal of deep neural networks[EB/OL]. [2025-06-04]. https://wangwenxiao.github.io/files/watermark_removal_icml19_workshop.pdf.
[34] SHAFIEINEJAD M, LUKAS N, WANG J Q, et al. On the robustness of backdoor-based watermarking in deep neural networks[C]//Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. New York, USA: ACM Press, 2021: 177-188.
[35] CHEN H, ROUHANI B D, KOUSHANFAR F. Blackmarks: blackbox multibit watermarking for deep neural networks[EB/OL]. [2025-06-04]. https://arxiv.org/abs/1904.00344.
[36] BANSAL A, CHIANG P, CURRY M J, et al. Certified neural network watermarks with randomized smoothing[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2022: 1450-1465.
[37] LIU X K, LI F T, WEN B H, et al. Removing backdoor-based watermarks in neural networks with limited data[C]//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Washington D.C., USA: IEEE Press, 2021: 10149-10156.
[38] PUAH Y H, NGO T A, CHATTOPADHYAY N, et al. BlockDoor: blocking backdoor based watermarks in deep neural networks [EB/OL]. [2025-06-04]. https://arxiv.org/abs/2412.12194.
[39] AIKEN W, KIM H, WOO S, et al. Neural network laundering: removing black-box backdoor watermarks from deep neural networks[J]. Computers & Security, 2021, 106: 102277.
[40] HITAJ D, MANCINI L. Have you stolen my model? Evasion attacks against deep neural network watermarking techniques[EB/OL]. [2025-06-04]. https://arxiv.org/abs/1809.00615.
[41] 张学军, 席阿友, 加小红, 等. 基于深度学习的指纹室内定位对抗样本攻击研究[J]. 计算机工程, 2024, 50(10): 228-239.
ZHANG X J, XI A Y, JIA X H, et al. Study on adversarial sample attacks on deep learning based fingerprinting indoor localization[J]. Computer Engineering, 2024, 50(10): 228-239. (in Chinese)
[42] 刘帅威, 李智, 王国美, 等. 基于 Transformer 和 GAN 的对抗样本生成算法[J]. 计算机工程, 2024, 50(2): 180-187.
LIU S W, LI Z, WANG G M, et al. Adversarial example generation algorithm based on Transformer and GAN[J]. Computer Engineering, 2024, 50(2): 180-187. (in Chinese)
[43] 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 (ICML'24). Washington D.C., USA: IEEE Press, 2024. 1456-1492.
[44] LIN J, JUAREZ M. A crack in the bark: leveraging public knowledge to remove Tree-Ring watermarks [EB/OL]. [2025-06-04]. https://arxiv.org/abs/2506.10502.
[45] WEN Y, KIRCHENBAUER J, GEIPING J, et al. Tree-Rings watermarks: Invisible fingerprints for diffusion images [J]. Advances in Neural Information Processing Systems, 2023, 36: 58047-58063.
[46] QUIRING E, RIECK K. Adversarial machine learning against digital watermarking[C]//Proceedings of the 26th European Signal Processing Conference (EUSIPCO). Washington D.C., USA: IEEE Press, 2018: 519-523.
[47] QUIRING E, ARP D, RIECK K. Forgotten siblings: unifying attacks on machine learning and digital watermarking[C]//Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P). Washington D.C., USA: IEEE Press, 2018: 488-502.
[48] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. [2025-06-04]. https://arxiv.org/abs/1312.6199.
[49] CHENG D N, LI X, LI W H, et al. Large-scale visible watermark detection and removal with deep convolutional networks[C]//Proceedings of the 1st Chinese Conference on Pattern Recognition and Computer Vision. Berlin, Germany: Springer International Publishing, 2018: 27-40.
[50] LI X, LU C, CHENG D N, et al. Towards photo-realistic visible watermark removal with conditional generative adversarial networks [C]// Proceedings of the 10th International Conference on Image and Graphics. Berlin, Germany: Springer International Publishing, 2019: 345-356.
[51] CAO Z Y, NIU S Z, ZHANG J W, et al. Generative adversarial networks model for visible watermark removal [J]. IET Image Processing, 2019, 13(10): 1783-1789.
[52] LUKAS N, DIAA A, FENAUX L, et al. Leveraging optimization for adaptive attacks on image watermarks[EB/OL]. [2025-06-04]. https://arxiv.org/abs/2309.16952.
[53] LI X Y. DiffWA: diffusion models for watermark attack[C]//Proceedings of the International Conference on Integrated Intelligence and Communication Systems (ICIICS). Washington D.C., USA: IEEE Press, 2023: 1-8.
[54] 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.
[55] HU Y, JIANG Z, GUO M, et al. Stable signature is unstable: removing image watermark from diffusion models [EB/OL]. [2025-06-04]. https://arxiv.org/abs/2405.07145.
[56] JIANG Z Y, ZHANG J H, GONG N Z. Evading watermark based detection of AI-generated content[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM Press, 2023: 1168-1181.
[57] CHEN J B, JORDAN M I, WAINWRIGHT M J. HopSkipJumpAttack: a query-efficient decision-based attack[C]//Proceedings of the IEEE Symposium on Security and Privacy (SP). Washington D.C., USA: IEEE Press, 2020: 1277-1294.
[58] HU Y, JIANG Z, GUO M, et al. A transfer attack to image watermarks[EB/OL]. [2025-06-04]. https://www.semanticscholar.org/paper/A-Transfer-Attack-to-Image-Watermarks-Hu-Jiang/fe8e1c1765bc1edda1100de281224892f4197f70/figure/0.
[59] ZHU J R, KAPLAN R, JOHNSON J, et al. HiDDeN:

hiding data with deep networks[C]//Proceedings of the European conference on computer vision (ECCV). Berlin, Germany: Springer, 2018: 682-697.
[60] 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 (CVPR). Washington D. C., USA: IEEE Press, 2020: 2117-2126.
[61] ZHAO X D, ZHANG K X, SU Z B, et al. Invisible image watermarks are provably removable using generative AI[C]// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc., 2024: 1-12.
[62] SABERI M, SADASIVAN V S, REZAEI K, et al. Robustness of AI-image detectors: fundamental limits and practical attacks[EB/OL]. [2025-06-04]. https://arxiv.org/abs/2310.00076.
[63] 汪旭童, 尹捷, 刘潮歌, 等. 神经网络后门攻击与防御综述[J]. 计算机学报, 2024, 47(8): 1713-1743.
WANG X T, YIN J, LIU C G, et al. A survey of backdoor attacks and defenses on neural networks[J]. Chinese Journal of Computers, 2024, 47(8): 1713-1743. (in Chinese)
[64] FAN Z K, GUAN Y P. A deep learning framework for face verification without alignment[J]. Journal of Real-Time Image Processing, 2021, 18(4): 999-1009.
[65] HUANG Y G, PAN L, LUO W, et al. Machine learning-based online source identification for image forensics[M]// CHEN X F, SUSILO W, BERTINO E. Cyber security meets machine learning. Singapore: Springer, 2021: 27-56.
[66] 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 (CVPR). Washington D. C., USA: IEEE Press, 2022: 10674-10685.
[67] XUE Z H, MARCULESCU R. Dynamic multimodal fusion[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Washington D. C., USA: IEEE Press, 2023: 2575-2584.
[68] ZHANG Q, WU H, ZHANG C, et al. Provable dynamic fusion for low-quality multimodal data[C]//Proceedings of the International Conference on Machine Learning. [S. l.]: PMLR, 2023: 41753-41769.
[69] LIU A W, PAN L Y, LU Y J, et al. A survey of text watermarking in the era of large language models[J]. ACM Computing Surveys, 2025, 57(2): 1-36.
[70] WANG B W, WU Y F, WANG G L. Adaptor: improving the robustness and imperceptibility of watermarking by the adaptive strength factor[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6260-6272.
[71] ZHANG L, LIU X, MARTIN A V, et al. Attack-resilient image watermarking using stable diffusion[J]. Advances in Neural Information Processing Systems, 2024, 37: 38480-38507.
[72] 闫培玲, 刘俊娟, 高志宇. 基于多模态深度神经网络的 Web 网页攻击重定向混淆检测[J]. 吉林大学学报(理学版), 2025, 63(6): 1731-1736.
YAN P L, LIU J J, GAO Z Y. Web page attack redirection confusion detection based on multimodal deep neural network[J]. Journal of Jilin University (Science Edition), 2025, 63(6): 1731-1736. (in Chinese)
[73] AGGARWAL A, MITTAL M, BATTINENI G. Generative adversarial network: an overview of theory and applications[J]. International Journal of Information Management Data Insights, 2021, 1(1): 100004.
[74] ZHAI G T, MIN X K. Perceptual image quality assessment: a survey[J]. Science China Information Sciences, 2020, 63(11): 211301.
[75] HANCOCK J T, KHOSHGOFTAAR T M, JOHNSON J M. Evaluating classifier performance with highly imbalanced big data[J]. Journal of Big Data, 2023, 10(1): 42.
[76] BETZALEL E, PENSO C, FETAYA E. Evaluation metrics for generative models: an empirical study[J]. Machine Learning and Knowledge Extraction, 2024, 6(3): 1531-1544.
[77] FEI J W, XIA Z H, TONDI B, et al. Wide flat minimum watermarking for robust ownership verification of GANs[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 8322-8337.
[78] OGUNDOKUN R O, ABIKOYE C O, SAHU A K, et al. Enhancing security and ownership protection of neural networks using watermarking techniques: a systematic literature review using PRISMA[M]//SAHU A K. Multimedia watermarking. Singapore: Springer, 2024: 1-28.
[79] TRAMER F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]//Proceedings of the 25th USENIX Security Symposium (USENIX Security 16). Austin, USA: USENIX Association, 2016: 601-618.
[80] MO M K, WANG C T, GUO Q W, et al. A novel robust black-box fingerprinting scheme for deep classification neural networks[J]. Expert Systems with Applications, 2024, 252: 124201.
[81] ZONG W, CHOW Y W, SUSILO W, et al. IPRemover: a generative model inversion attack against deep neural network fingerprinting and watermarking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2024: 7837-7845.
[82] 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 (ICCV). Washington D.C., USA: IEEE Press, 2023: 22409-22420.
[83] ZHANG G K, WANG L J, SU Y T, et al. MarkPlugger: generalizable watermark framework for latent diffusion models without retraining[J]. IEEE Transactions on Multimedia, 2025, 27: 6211-6220.

文字编辑 陆燕菲
栏目编辑 赖玉玲