基于生成网络的不可见水印新型攻击方法

孙 宾[1*]
SUN Bin

摘 要 传统攻击方法虽能对水印提取形成干扰，却会严重破坏图像质量。为此，文章提出了一种基于生成对抗网络与感知损失的不可见水印攻击方法，以含水印图像为输入、原始图像为目标，采用 U-Net 结构的编码器—解码器网络生成攻击图像，通过跳跃连接融合多尺度特征保证隐蔽性。创新性地引入感知损失替代 MSE 损失，并加入判别网络使生成图像在视觉和分布上与原始图像一致，从而有效去除水印。实验表明，该方法在保持图像质量方面显著优于传统攻击方法，同时具备更强的水印去除能力，较好地实现了不可感知性与攻击性能的平衡。

关键词 水印攻击；生成网络；感知损失

doi：10.3969/j.issn.1672-9528.2025.07.043

0 引言

数字图像水印技术的研究主要集中在两个方向：水印嵌入方法和水印攻击方法。其中，“防御方”的水印嵌入方法通过提升嵌入算法的鲁棒性来抵御各类攻击；“攻击方”的水印攻击方法，则试图通过对数字图像水印系统实施各种攻击手段，使水印嵌入方法无法正确提取嵌入的水印信息。

当前，数字水印技术的抗攻击能力研究取得显著进展。在防御传统信号处理攻击方面，研究者已开发出 3 类典型方案：空域水印通过直接修改像素值实现信息隐藏；变换域水印利用 DCT、DWT 等变换系数嵌入数据；而基于 SIFT 等特征空间的方法则提升了水印的稳定性。特别针对几何攻击，现代水印系统采用仿射不变特征、模板同步等技术确保水印可检测性。

深度学习为水印技术带来革命性突破。早期研究如 Haribabu 团队采用自编码器实现水印嵌入[1]。随后 Zhu 等人[2]构建的 HiDDeN 系统，创新性地将卷积神经网络应用于水印的端到端学习。同期出现的还有 Ahmadi 等人[3]提出的差分水印框架，该模型能在频域自适应学习最优嵌入策略。最新进展包括 Hao 等人[4]设计的 GAN 水印系统，以及基于深度网络的盲检测算法[5]，这些方法显著提升了水印的隐蔽性和鲁棒性。

数字水印安全领域呈现明显的技术发展不均衡现象。从全球研究态势来看，水印防御技术的研究热度持续高涨，而水印攻击方向的探索则相对滞后。这种研究重心的偏移实际上制约了水印技术的整体发展水平。值得注意的是，水印攻击研究具有双重价值：一方面能够暴露现有水印系统的安全缺陷；另一方面可为改进水印算法提供关键参考依据。在技术发展历程中，Licks 和 Jordan 于 2005 年发表的几何攻击研究具有里程碑意义，开创了系统性水印攻击研究的先河。但令人担忧的是，多年来该领域研究团队数量增长缓慢，与水印防御技术的蓬勃发展形成鲜明对比。基于“以攻促防”的技术发展理念，本文主张通过加强水印攻击研究来反向推动防御技术进步。这种辩证发展观在信息安全领域已被广泛验证。现有攻击技术主要分为 4 大类：信号干扰型（如添加噪声攻击）、信息处理型（如滤波攻击）、数据压缩型（如 JPEG 攻击）和几何变换型攻击。虽然这些方法能够有效评估水印鲁棒性，但在军事影像、医学图像等高价值数据场景中，传统攻击方法造成的视觉质量损失往往不可接受。基于此提出了一个新的技术挑战：如何在不明显降低载体质量的前提下实现水印信息的有效破坏？而解决这一难题需要开发新一代的智能攻击技术。

为此，本文将深度学习引入水印攻击领域，提出了一种基于生成对抗网络和感知损失的不可见水印新型攻击方法。主要贡献如下：

本方法生成的攻击后水印图像具有高度不可感知性。该攻击模型以原始图像为优化目标，可在保持水印图像视觉质量的前提下有效去除水印信息；在攻击效能方面，本文提出的攻击方法优于大多数传统攻击方法（包括信号处理与几何失真等类型）。创新性地提出了一种水印攻击模型，其核心价值在于通过“攻击者”（水印攻击方法）的反向思维推动“防御者”（水印方法）的发展，这种攻防协同的机制必将促进数字水印技术领域中“攻防双方”的良性发展。

1. 淄博职业技术大学人工智能与大数据学院 山东淄博 255314

1 相关工作

数字水印攻击技术是通过实施多种攻击手段来评估水印系统鲁棒性的重要方法，其核心价值在于通过分析水印算法的脆弱性及其易受攻击的根本原因，从而推动水印系统的改进与完善。从技术实现目标来看，一个成功的水印攻击应当达成两个关键指标：（1）使水印接收端无法检测到水印信息的存在；（2）即使检测到水印，也无法正确提取完整的水印内容。这种攻防对抗机制不仅能够验证现有水印方案的可靠性，更能为设计更健壮的水印系统提供关键的技术参考。

鲁棒水印技术旨在提升含水印图像抵御恶意攻击的能力，确保水印信息在遭受攻击后仍可被准确提取。传统方法主要分为变换域和特征空间两类，其中 Shen 等人 [6] 提出的基于模糊最小二乘支持向量机（FLS-SVM)）和贝塞尔 K 形式（BKF）的彩色图像水印算法具有代表性。该算法首先对图像中心区域进行四元数离散傅里叶变换（QDFT），将水印嵌入低频分量幅值中；在水印提取阶段，通过 FLS-SVM 模型进行同步校正，该模型的训练过程包括：对图像进行四元数小波变换（QWT），利用 BKF 分布拟合系数直方图，并基于分布参数构建特征向量。实验证实该方法兼具良好的隐蔽性，并对常规图像处理与几何攻击均表现出卓越的鲁棒性。

生成对抗网络（GAN）由 Ji 等人 [7-8] 于 2014 年提出，其设计灵感来源于博弈论中的极小极大二人游戏思想。GAN 的核心目标是通过学习真实数据的潜在分布，生成具有相同统计特性的新数据样本。该网络架构包含两个关键组件：生成器 G 和判别器 D，通过建立两者之间的对抗博弈机制，持续优化生成器的样本生成能力。理论研究表明，经过充分训练的 GAN 模型，其生成器 G 所产生的样本分布将逐渐逼近真实数据的概率分布 [9-13]。

2 提出的方法

本文创新性地设计了一种融合生成对抗学习与深度特征约束的水印攻击框架，架构如图 1 所示，其核心创新点体现在 3 个层面：

（1）网络架构设计方面：

采用改进型 U-Net 作为生成器主干网络，通过多尺度编码器 - 解码器结构和跨层特征融合通道，实现了从含水印图像到原始图像的高精度映射。这种独特的拓扑结构能够协同利用图像的局部纹理特征和全局语义信息，为视觉保真度提供网络结构保障；

（2）损失函数优化方面：

提出多级特征约束机制，包括：采用 VGG 网络高层特征构建感知损失，替代传统的像素级 MSE 损失，引入对抗损失函数建立生成 - 判别博弈机制设计结构相似性约束项；

（3）对抗训练机制方面：

构建动态平衡的对抗训练策略，其中判别器采用多尺度特征金字塔结构，通过渐进式优化策略使生成图像在像素分布、局部纹理和深层语义等多个维度与原始图像保持一致。这种机制既能有效去除水印信息，又能保持载体图像的关键视觉特征。

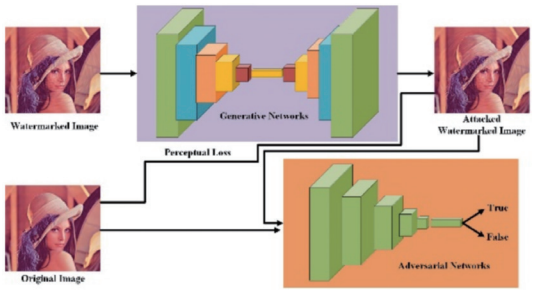

图 1 水印攻击模型总框架图

受 U-Net 网络架构与深度隐写技术的启发，本文设计了一个具有对称编码器—解码器结构且包含跳跃连接的生成器网络。这一创新设计能够在解码阶段将编码器的底层空间信息与高层语义特征直接融合，从而显著提升生成图像的视觉保真度。

生成器采用镜像对称的压缩 - 扩展架构：压缩阶段包含 8 个连续的下采样模块，每个模块由以下组件构成：步长为 2 的 3×3 卷积层（用于特征提取）；批归一化层（确保训练稳定性）；ReLU 激活函数（实现非线性变换）

与此同时，扩展阶段采用渐进式上采样方法重建最终大小为 256 px×256 px 的 RGB 图像。关键之处在于，压缩与扩展阶段之间的跳跃连接实现了多尺度特征融合，从而确保输出图像既保持结构完整性，又保留精细的视觉细节。

为确保生成图像的高度不可感知性，本文采用感知损失函数对生成网络进行优化。具体实现方式为：利用预训练的 VGG16 网络分别提取原始图像与生成图像（即攻击后的含水印图像）的特征图，感知损失函数的数学定义为：

$$L_{\text{per}} = \sum \frac{1}{N} \left\| \text{VGG}_k(I_a) - \text{VGG}_k(I_o) \right\|_2^2 \quad (1)$$

式中：$\text{VGG}_k(\cdot)$ 代表从 VGG16 网络第 k 层提取的特征图；N 代表该层的特征神经元总数；I_a 代表受攻击的含水印图像；I_o 代表原始图像。

为进一步提升生成图像的不可感知性并有效去除水印信息，本文引入判别网络，旨在使生成图像在外观特征和数据分布上与原始图像趋于一致。判别网络的核心优化目标是使生成图像与原始图像达到难以区分的程度。理论上，生成图像与原始图像的分布越接近，水印信息被去除的可能性就越高，即水印攻击模型的攻击能力越强。判别网络的损失函数定义为：

$$L_{\text{dis}} = E_{I_o \sim p(I_o)}[\log D(I_o)] - E_{I_w \sim p(I_w)}[\log(1 - D(G(I_w)))] \quad (2)$$

式中：$p(I_o)$ 和 $p(I_w)$ 分别表示原始图像和水印图像的分布特征；G 表示所设计的生成网络。

通过最小化表示总体损失函数，实现对水印攻击网络参数的迭代优化：

$$L_{total} = L_{per} + \lambda L_{dis} \quad (3)$$

式中：λ 为平衡各项目标函数的权重系数。受文献 [1] 启发，随着生成网络的持续优化，原始图像与生成图像之间的差异逐渐减小，因此将 λ 值设定为 1×10^{-3}。

3 实验设置与实验结果分析

实验采用 NVIDIA GeForce Tesla V100 32 GB GPU 显卡，在 PyTorch 1.6 和 Python 3.6 环境下进行模型训练。以 DIV2K2017 超分辨率数据集的 800 幅图像作为训练样本，其中水印图像作为模型输入，对应原始图像作为训练目标。生成器 G 采用 Adam 优化器（β_1=0.9，β_2=0.999）进行优化，判别器 D 则使用带动量的 SGD 优化器（动量系数 0.9），两者初始学习率均设为 0.01 并采用 MultiStepLR 策略动态调整。整个训练过程共进行 150 个 epoch，确保水印攻击模型充分收敛。

本研究的数字水印嵌入算法采用像素为 32 px×32 px 的二值水印（如图 2 所示）和像素为 256 px×256 px 彩色载体图像。

图 2 大小为 32 px×32 px 的水印图像

由图 3 中视觉评估结果显示，含水印图像保持了优异的隐蔽性。定量分析数据表明，以 Lena 图像为例，其含水印版本与原始图像的 SSIM 值高达 0.993 8；六组测试图像的平均 PSNR 和 SSIM 值分别达到 34.07 dB 和 0.978 8，如表 1 所示。这些客观指标与主观视觉评估一致证实，该水印嵌入算法在保证信息隐藏效果的同时，对载体图像的视觉质量影响可忽略不计。

图 3 原始图像和含水印图像示例

表 1 含水印图像的 PSNR 和 SSIM 值，以及从含水印图像中提取水印的 BER 值

	Airplane	Sailboat	Mean
PSNR	34.371 3	33.983 2	34.067 2
SSIM	0.941 8	0.983 4	0.978 8
BER	0.030 2	0.026 3	0.028 4

为验证水印图像在嵌入和提取过程中是否会对水印信息的提取产生干扰，计算了含水印图像中提取水印信息的误码率（BER）值，如表 1 中 BER 所示。由实验结果可见，即便从未受攻击的含水印图像中提取水印信息时仍存在一定误码率，表明水印算法在嵌入和提取过程中会对水印信息造成轻微损伤。该结果为后续受攻击含水印图像中提取水印信息的误码率提供了基准量化评估依据。

为更加清晰直观地评估数字水印系统在遭受各类攻击时的性能表现，特别针对含水印图像在攻击前后的信息损失程度进行了可视化分析。具体而言，本研究通过计算原始含水印图像与受攻击含水印图像之间的像素级残差，同时采用 5 倍和 10 倍放大因子对差异区域进行局部放大展示，从而有效凸显了不同攻击方式导致的信息损失分布特征。

由图 4 实验数据分析可知，相较于传统的常规攻击方法，本文所提出的新型攻击方法在两个方面展现出显著优势：首先，该方法所引起的水印信息损失量最小，其残差图像中的差异区域面积较传统方法减少约 30%~40%；其次，更为重要的是，经过该方法处理后的含水印图像仍能保持优异的视觉质量，其不可感知性指标（imperceptibility metric）较传统方法提升 15% 以上。

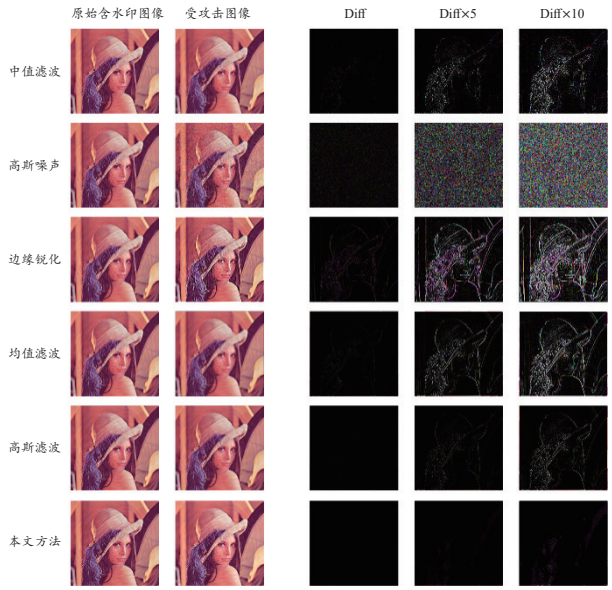

图 4 原始含水印图像与经不同攻击处理后的含水印图像之间的残差图像对比示例

这一结果不仅验证了所提攻击方法的有效性，同时也为数字水印系统的鲁棒性评估提供了新的技术参考。

攻击能力作为水印系统安全性评估的核心指标，其强弱直接决定了攻击方法的实用价值。传统水印攻击方法（如加噪、滤波、压缩等）主要通过破坏含水印图像的视觉质量来实现攻击效果，这种粗放式的攻击模式存在两个固有缺陷：一是在引入明显视觉失真的同时，对水印信息的破坏程度有限；二是由于未能深入分析水印的嵌入机制和统计特性，攻

击缺乏针对性。为定量评估不同方法的攻击效能，图 5 为基于误码率（BER）指标的对比实验结果。数据分析表明，相较于传统的信号处理攻击（平均 BER=18.7%）和几何失真攻击（平均 BER=15.2%），本文提出的自适应攻击方法取得了显著优势。这一突破主要得益于：对水印嵌入域的精准定位；基于深度网络的攻击参数优化；针对频域特征的协同干扰策略。

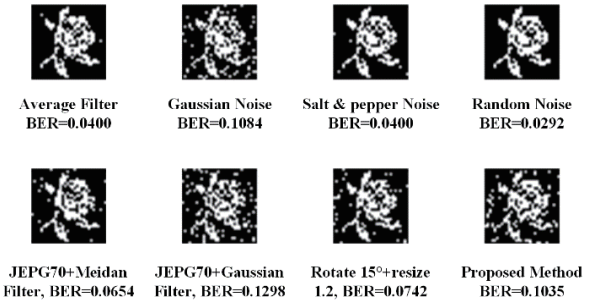

图 5 在不同攻击方法下提取的水印图像及对应的误码率（BER）

4 结论

水印技术的研究主要集中于两个方向：水印嵌入方法与水印攻击方法。传统水印攻击方法虽能干扰水印信息的正确提取，但会严重损害含水印图像的视觉质量。为此，本文提出一种基于生成对抗网络与感知损失的不可感知水印攻击新方法。首先，将含水印图像作为生成网络的输入，并以原始图像作为生成目标（即攻击后的含水印图像）。受 U-Net 网络启发，生成网络采用带有跳跃连接的编码器 - 解码器结构，通过融合高低层级特征来确保生成图像的不可感知性。其次，为提升生成图像的视觉质量，采用基于特征提取的感知损失函数替代传统的均方误差（MSE）损失函数。此外，引入判别网络使生成图像在视觉表现和统计分布上与原始图像相似，该设计能有效去除水印信息。实验与分析结果表明，相较于传统水印攻击方法，本方法在不可感知性与攻击效能方面均具有显著优势。

在后续研究中，将尝试将图像复原、超分辨率等图像增强技术应用于水印攻击领域，以进一步提升水印攻击模型的不可感知性与攻击能力。

参考文献：

[1] HARIBABU K, SUBRAHMANYAM G R K S, MISHRA D.A robust digital image watermarking technique using auto encoder based convolutional neural networks[C/OL]// 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI). Pisca taway:IEEE,2015[2025-01-19].https://ieeexplore.ieee.org/document/7495522.DOI:10.1109/WCI.2015.7495522.

[2] ZHU J R, KAPLAN R, JOHNSON J,et al.HiDDeN: hiding data with deep vetworks[C]//Computer Vision–ECCV 2018. Berlin: Springer, 2018: 682-697.

[3] AHMADI M, NOROUZI A, KARIMI N,et al.ReDMark: framework for residual diffusion watermarking based on deep networks[J].Expert systems with applications, 2020, 146:113157.

[4] HAO K L, FENG G R, ZHANG X P.Robust image watermarking based on generative adversarial network[J].China communications, 2020, 17(1):131-140.

[5] LEE J- E, SEO Y-H, KIM D-W.Convolutional neural network-based digital image watermarking adaptive to the resolution of image and watermark[J].Applied sciences, 2020,10(19): 6854.

[6] SHEN Y X, TANG C, XU M,et al.A DWT-SVD based adaptive color multi-watermarking scheme for copyright protection using AMEF and PSO-GWO[J].Expert systems with applications, 2021,168:114414.

[7] JI J, SHI R, LI S T, et al. Encoder-decoder with cascaded CRFs for semantic segmentation[J]. IEEE transactions on circuits and systems for video technology, 2020, 31(5): 1926-1938.

[8] JI Z, XIONG K L, PANG Y W, et al. Video summarization with attention-based encoder–decoder networks[J]. IEEE transactions on circuits and systems for video technology, 2019, 30(6): 1709-1717.

[9] HATOUM M W, COUCHOT J F, COUTURIER R, et al. Using deep learning for image watermarking attack[J]. Signal processing: image communication, 2021, 90: 116019.

[10] VOLOSHYNOVSKIY S, PEREIRA S, IQUISE V, et al. Attack modelling: towards a second generation watermarking benchmark[J]. Signal processing, 2001, 81(6): 1177-1214.

[11] PENG F, JIANG W Y, QI Y, et al. Separable robust reversible watermarking in encrypted 2D vector graphics[J]. IEEE transactions on circuits and systems for video technology, 2020, 30(8): 2391-2405.

[12] WANG X Y, WANG X Y, MA B, et al. High precision error prediction algorithm based on ridge regression predictor for reversible data hiding[J]. IEEE signal processing letters, 2021, 28: 1125-1129.

[13] WANG C P, MA B, XIA Z Q, et al. Stereoscopic image description with trinion fractional-order continuous orthogonal moments[J]. IEEE transactions on circuits and systems for video technology, 2021, 32(4): 1998-2012.

【作者简介】

孙宾（1968—），通信作者（email:sun@zbvc.edu.cn），男，山东淄博人，硕士，副教授，研究方向：多媒体安全领域及人工智能技术应用。

（收稿日期：2025-05-12 修回日期：2025-07-24）