计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009855] [CSTR: 32024.14.csa.009855]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 LoRA 的双阶段扩散模型水印方案①

白少杰, 林立霞, 胡子寒, 袁艺林, 曹　鹏

(北京印刷学院 信息工程学院, 北京 102627)
通信作者: 林立霞, E-mail: linlixia@bigc.edu.cn

摘　要: 扩散模型的发展使得高质量图像生成变得更加便捷, 但同时引发了生成图像的版权保护问题. 现有研究通常在扩散过程中隐秘性地嵌入水印, 以提升水印鲁棒性. 然而, 目前现有基于扩散过程的水印方案集中于嵌入固定水印, 无法满足用户对水印多样化的需求. 此外, 还存在被恶意用户更换解码器规避水印的风险. 为了解决上述问题, 本文提出了基于 LoRA 的双阶段扩散模型水印方案. 首先, 该方案在水印编解码预训练阶段训练出水印编解码器, 保证水印嵌入的稳定性; 然后, 在 U-Net 微调阶段通过 LoRA 和自适应注意力机制, 使 U-Net 在保持生成质量的同时学习到第 1 阶段的水印模式, 实现多用户定制化. 实验表明, 该方案在图像一致性和水印鲁棒性上均优于现有方法. 在图像攻击下, 水印图像的 *FID* 距离提高了 0.61%, 平均提取精度提升了 4.9%.

关键词: 数字水印; 扩散模型; LoRA; 自适应嵌入; 图像攻击

引用格式: 白少杰,林立霞,胡子寒,袁艺林,曹鹏.基于 LoRA 的双阶段扩散模型水印方案.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9855.html

Two-stage Diffusion Model Watermarking Scheme Based on LoRA

BAI Shao-Jie, LIN Li-Xia, HU Zi-Han, YUAN Yi-Lin, CAO Peng

(College of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102627, China)

Abstract: The development of diffusion models has significantly enhanced the generation of high-quality images. However, it has also introduced notable challenges in the copyright protection of generated images. Existing research typically embeds watermarks invisibly during the diffusion process to increase their robustness. However, current watermarking methods based on diffusion processes focus on embedding a fixed watermark, failing to meet the diverse needs of users. In addition, malicious users may bypass watermarking by replacing the decoder, which presents a potential risk. To address these challenges, a two-stage watermarking scheme for diffusion models based on LoRA is proposed. In the first stage, the watermark codec is trained during pre-training to ensure stable watermark embedding. In the subsequent U-Net fine-tuning stage, LoRA and an adaptive attention mechanism are incorporated. This enables U-Net to learn watermark patterns from the first stage while maintaining image quality and supporting multi-user customization. Experimental results demonstrate that the proposed scheme outperforms existing methods in terms of image consistency and watermark robustness. Under image attacks, the *FID* distance of the watermarked images increases by only 0.61%, while the average extraction accuracy improves by 4.9%.

Key words: digital watermark; diffusion model; LoRA; adaptive embedding; image attack

扩散模型的快速发展使得人们可以轻易生成和创建高质量的图像[1–5]. 例如, 游戏设计师 Jason Allen 使用 AI 绘图工具 Mid Journey, 创作了名为《Théâtre D’opéra Spatial》的二维艺术图像. 在美国科罗拉多州举

① 基金项目: 北京市教委科技一般项目 (KM202410015001); 北京印刷学院校级项目 (Ea202302, 27170123033, Ea202301)
收稿时间: 2024-11-01; 修改时间: 2024-11-29; 采用时间: 2024-12-16; csa 在线出版时间: 2025-04-28

办的新兴数字艺术家竞赛中, 该图像获得“数字艺术/数字修饰照片”类别一等奖[6]. 这表明, 人工智能在生成图像方面已经令人难以分辨. 扩散模型的生成能力虽然满足了用户对高质量创意的需求, 但未经同意使用的生成图像以及恶意用户利用模型生成有害内容引发了模型持有者对模型版权保护和图像安全性的严重担忧. 恶意用户利用模型生成有害内容给生成式内容安全带来了严重威胁[7–9].

数字水印是一种被广泛应用于图像、音视频以及深度模型中的版权保护方案. 传统的水印方案通过在图像生成后利用频域变换或者通过训练水印编解码网络来实现图像的版权保护, 这种在生成图像后嵌入水印的方式称为事后嵌入水印方法[10–16]. 然而, 事后水印的工作容易受到水印攻击消除图像中的水印, 例如图像剪切、智能消除等技术手段导致图像中的水印被消除. 因此, 近期研究更关注于在扩散模型的扩散过程中, 将水印无感知地嵌入到扩散模型中, 基于该方法生成的图像会携带水印信息. 这种方式称为扩散过程水印. 其中, Stable Signature[17]提出了一种使用预训练解码器的方法. 具体而言, 将水印嵌入在扩散模型的变分自动编码器 (variational autoencoder, VAE) 解码器中. 但是, 该方案需要对每个不同水印训练不同的 VAE 解码器, 这使得 Stable Signature 难以满足成千上万用户的使用. 此外, 恶意用户可以通过更换未嵌入水印的 VAE 解码器轻松规避水印验证. 此外, Tree-Ring[18,19]提出将水印隐藏在扩散模型初始噪声向量的频域中, 然后通过对生成图像反转, 进而恢复噪声来检测水印信息. 尽管 Tree-Ring 无需训练, 但检测结果极大程度上依赖于恢复过程, 并且在多个用户之间的识别上存在问题. 综上, 现有方案无法有效解决更换 VAE 解码器规避水印的问题, 且现有研究集中于在扩散过程中嵌入固定水印信息, 无法满足用户对嵌入水印多样化的需求.

为了解决上述问题, 本文提出了基于水印 LoRA 的双阶段扩散模型水印方案. 其中, 第 1 阶段为水印编解码预训练阶段, 第 2 阶段为 U-Net 微调阶段. 具体地, 在第 1 阶段, 冻结扩散模型的全部参数, 额外增加水印编码模块及水印提取模块. 这一阶段充分考虑了水印的鲁棒性, 创建了适合 U-Net 学习的水印模式. 在第 2 阶段, 利用 LoRA[20]方法来微调扩散模型的 U-Net 模块使其最小限度地影响扩散模型的图像生成能力并学习到第 1 阶段的水印模式. 此外, 在第 2 阶段增加自适应注意力机制, 通过处理水印特征及 U-Net 特征之间的交叉注意力使得 U-Net 模型能够自适应感受不同水印并适应扩散过程. 所提出算法可将水印作为条件输入并为不同用户生成定制的水印图像.

本文的主要工作总结如下.

(1) 提出了一种基于水印 LoRA 的双阶段扩散模型水印方案, 第 1 阶段关注了水印嵌入的鲁棒性, 第 2 阶段更加注重提升图像的质量.

(2) 针对现有算法仅考虑嵌入单一水印的局限性, 提出了一种自适应交叉注意力机制, 通过处理水印特征以及 U-Net 特征之间的交叉注意力使得算法具有水印灵活性, 并削弱了水印对图像生成的影响.

(3) 大量的实验表明, 与现有的扩散模型水印方案相比, 该方案在图像一致性和水印鲁棒性方面均具有一定程度的提升.

1 相关工作

1.1 事后嵌入水印算法

事后嵌入水印是指在图像生成或处理完成后, 将水印嵌入最终图像的方法. 这类方法主要包括传统的频域变换方法 (如 DWT-DCT[10]、DWT-DCT-SVD[10])、基于每幅图像优化的技术 (如半监督学习 SSL[11]、前馈神经网络合成 FNNS[12]) 以及前向编解码方法 (如 HiDDeN[15]、StegaStamp[13]和 MBRS[14]等). 不同算法侧重点各有不同, FNNS 特别关注如何隐藏更多水印信息, 而 HiDDeN 与 MBRS 则优先考虑在 JPEG 压缩攻击下的鲁棒性. 值得注意的是, Stable diffusion[2]官方库建议采用 DWT-DCT、DWT-DCT-SVD 和 RivaGAN[16]实现水印的嵌入. 然而, 恶意用户只需改变几行代码就可以实现去除生成后的水印. 这些方法在面对高强度的图像攻击时, 如重压缩、剪切、噪声添加等, 仍然无法保证水印图像的安全, 因此最近研究者提出扩散过程水印作为一种新的解决方案. 即在扩散模型生成过程中进行微调, 增强水印方案的鲁棒性.

1.2 扩散过程水印

扩散过程水印是将水印嵌入至扩散模型扩散过程的一种方法. 根据水印嵌入的位置不同, 本文将扩散过程水印分为初始噪声嵌入和潜在空间嵌入两类方案.

● 初始噪声嵌入: 2023 年 Tree-Ring[18,19]提出将水印添加到扩散模型初始噪声中, 实现了显著的鲁棒性. 该方法无需训练, 在检测时只需通过反向扩散处理水

印图像恢复原始水印. 然而, Tree-Ring 缺乏多种密钥识别能力, 无法满足大量用户实际应用. 随后的方法采用编码解码器对初始噪声进行改进[21], 结合了随机高斯噪声和特定训练策略. 同样地, 该方法改变了生成图像的布局, 在某些场景下并不适用.

● 潜在空间嵌入: 这类方法将水印嵌入到 VAE 解码器或扩散 U-Net 的潜在空间中. 2023 年 Stable Signature[17]首先提出了一种利用预训练的水印解码器提取扩散图像中水印的方案. 具体而言, Stable Signature 将预设水印通过微调嵌入至 VAE 解码器中, 然后, 将这些定制的解码器分发给个人用户进行检测和识别. 然而, Stable Signature 只能嵌入固定的水印到 VAE 解码器中, 难以满足不同用户的使用. 为了解决这样的问题, FSW[22]利用消息矩阵和消息编码器灵活改变不同水印信息的嵌入, 提高了扩散模型的可扩展性并允许单个架构中携带多个水印信息. 文献[23]提出即插即用水印框架, 是无需对 SD 模型进行重新训练的水印方案. 该方案仅需预训练水印编码器和解码器, 无需对 Stable diffusion 进行微调. 2024 年 AquaLoRA[24]被提出, 这是一种水印 LoRA 方案, 其将水印嵌入扩散模型 U-Net 网络, 通过 LoRA 微调 U-Net 模型并利用 mapper 映射器生成不同水印 LoRA, 达到灵活嵌入水印的效果.

1.3 小结

扩散模型图像生成的广泛应用带来了如未经同意使用的生成图像以及恶意用户利用模型生成有害内容的安全问题. 事后嵌入水印方法[10–16]在面对高强度的图像攻击时仍无法保证水印图像的安全, 因此最近研究者提出了扩散过程水印. 在扩散模型生成过程中进行微调, 增强了水印系统的鲁棒性. 然而, 现有的扩散过程水印方法[17–19,21–23]存在下述问题: 1) 这些方法通常对单个水印进行嵌入与提取, 在面对多个用户使用时, 需要重新训练微调 VAE 解码器模型才能做到多水印嵌入; 2) 单独对 VAE 解码器进行微调可能会被恶意用户通过更换干净的 VAE 解码器轻松规避. 因此, 如何设计出既能满足多用户使用又能规避恶意用户攻击的扩散过程水印方案, 仍然是扩散模型水印领域中一个亟待解决的关键问题.

2 本文模型

基于水印 LoRA 的双阶段扩散模型水印方案整体架构如图 1 所示, 共分为以下两个阶段: 水印编解码预训练阶段和 U-Net 微调阶段. 在水印编解码预训练阶段, 通过冻结扩散模型全部参数, 在不干扰模型原有能力的情况下训练一组水印编码解码器, 确保水印的有效性. 在 U-Net 微调阶段, 冻结第 1 阶段水印编解码器的参数, 图 1 中雪花即表示冻结模型参数. 通过 LoRA 方法有效微调扩散模型 U-Net 去噪部分以确保对图像生成能力的影响最小化, 并使得 U-Net 学习到第 1 阶段的水印模式. 此外, 引入自适应注意力机制, 使得模型可以根据不同的水印特征进行调整, 提高了模型的灵活性和适应性. 水印方案整体架构如图 1 所示. 本文在第 2.1、2.2 节介绍水印编解码预训练阶段, 在第 2.3、2.4 节介绍 U-Net 微调阶段, 在第 2.5 节介绍本文所使用的损失函数.

2.1 水印编解码预训练阶段

(1) 水印编码解码器

如图 1 上半部分所示, 在水印编解码预训练阶段, 水印编码模块 (watermark embedding module, WEM) 的作用是将二进制水印信息 M (如“101011…”) 转换为适合嵌入到图像特征中的形式即图像大小的二维特征图, 以确保水印能够成功嵌入图像特征中. 其中, 图像特征与水印特征的融合是通过简单矩阵加法进行的. 随后, VAE 解码器进一步解码这些特征, 并生成携带水印的图像. 输出的水印图像经过水印解码模块 (watermark decoding module, WDM) 将提取的水印与输入的水印信息进行比对, 确保水印完整性和鲁棒性.

水印编码模块 WEM 模型架构如图 2 所示, 由线性层, 卷积层和非线性激活层构成. 线性层将长度 48 bit 的一维水印信息转化为 64×64 长度的一维信息, 接着利用 Reshape 将一维信息转换成二维特征, 再通过 SiLU 激活函数增强特征的非线性表示, 最后通过卷积层获取水印特征图, 并随后与图像特征进行融合. 其中, 卷积模块只改变特征图的通道数不改变特征图的尺寸大小, 用于增强模型提取水印特征的能力.

水印解码模块 WDM, 使用 StegaStamp[13]的水印解码器模型结构重新训练, 完成对水印信息解码的目的.

(2) 图像攻击层

由于在图像使用的实际场景中存在各种类型的图像攻击, 因此在水印编解码预训练阶段, 将水印图像输入到水印解码器 WDM 之前, 利用攻击层处理水印图像. 具体而言, 引入了多种噪声层, 包括 JPEG 压缩、裁剪与缩放、高斯模糊、高斯噪声和颜色抖动. 在前向

传播过程中, 图像攻击层根据预设概率从各个噪声层中随机选择一个应用于输入图像. 通过设计使得每次训练时, 输入数据的扰动都具有随机性, 从而提高了模型的鲁棒性.

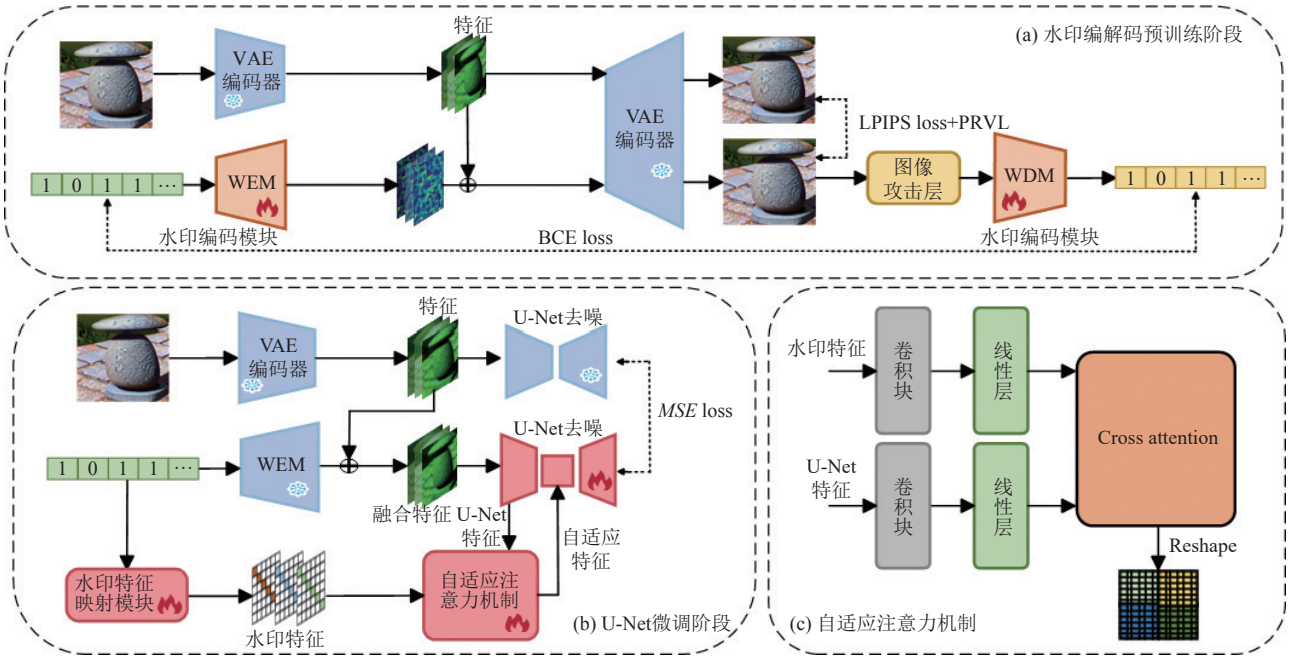

图 1 方案整体架构图

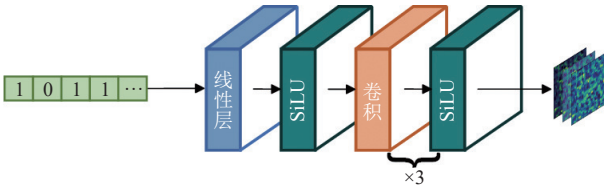

图 2 水印编码模块 (WEM)

2.2 U-Net 微调阶段

(1) 基于 LoRA 的 U-Net 微调

LoRA[20]是一种用于微调大模型的低秩适应技术. 它通过训练低秩矩阵将微调参数注入原始模型, 实现对模型的修改. 这种方法不仅降低了计算需求, 还显著减少了训练资源, 相较于直接训练原始模型而言, 资源消耗要小得多. 具体而言, LoRA 微调表达式如下:

$$W_{\text{watermark}} = W + \alpha \Delta W_{\text{lora}} \tag{1}$$

其中, ΔW_{lora}表示微调后的 LoRA 层, W表示原扩散模型参数, α表示权重参数, 本文中α设置为 1.

在 U-Net 微调阶段, 本文的目标是将先前生成的水印模式集成嵌入 U-Net 中. 为了实现这个目标, 使用 LoRA 技术完成水印的嵌入. 总体上, 原图像经过 VAE 编码器生成特征, 与水印编码器 WEM 提取出的水印特征合并. 此时, 特征包含了水印嵌入的信息, 形成融合特征表示. 为了进一步增强水印特征的表达和嵌入, 通过引入自适应注意力机制, 模型自适应地处理水印映射特征和 U-Net 去噪的特征. 自适应注意力机制为解决单一水印嵌入问题提供了有效的方案. 具体地, 根据输入水印的不同, 动态调整 U-Net 模型关于不同水印特征的注意力, 这一机制确保模型在生成图像时能够更好地保留水印特征的完整性和鲁棒性. 生成的图像在各种条件下都能有效地维护水印的识别性, 从而提升整体水印效果, 满足了不同水印的灵活性.

(2) 自适应过程

自适应过程通过引入灵活的自适应注意力机制, 使模型具备应对不同水印特征之间分布差异的能力, 而无需为每个新的水印重新训练模型. 在面对多样化水印特征时, 模型仍能保持较高的嵌入和解码准确率, 进而解决了多个用户之间的识别问题.

自适应过程包含水印特征映射模块与自适应注意力机制. 水印特征映射模块是自适应过程中至关重要的一部分, 用于将不同分布的水印特征映射到统一的特征空间中. 具体地, 将长度为N的水印序列转换为长度为 r 的向量. 对于水印的第i位, 使用嵌入向量I_i和 0 来表示二进制状态 0 和 1, 其中I_i由标准高斯分布初始化. 水印特征映射模块的计算过程如下:

$$f_i(M_i)=\begin{cases} I_i, & \text{if } M_i=0 \\ 0, & \text{else} \end{cases} \tag{2}$$

$$S=\text{diag}\left(1+\frac{1}{N}\sum_{i=1}^{N}f_i(M_i)\right) \tag{3}$$

其中, $f_i(M_i)$表示第i位水印映射后得到的序列, M_i表示第i位水印序列值, N为水印序列长度. S表示水印映射特征矩阵, diag(·) 表示用于创建对角矩阵的函数. 最终, 得到映射后的水印序列的对角矩阵作为水印映射特征.

图 1(c) 展示了自适应注意力机制模块的结构. 自适应注意力机制是通过交叉注意力模块实现的, 它通过联合 U-Net 提取的特征和水印特征, 进一步对生成过程中的特征进行权重调整. 该模块首先将来自 U-Net 的特征和水印特征通过线性层得到降维序列特征, 再在交叉注意力中计算特征间的相关性, 从而生成加权特征表示. 最后, 将自适应特征输入给下一级 U-Net 网络. 具体而言, 自适应过程修改了 U-Net 的下采样层. 这一过程允许 U-Net 在生成过程中更加精确地将水印特征嵌入到目标图像中并实现算法在处理不同水印时的灵活性.

2.3 损失函数

(1) 水印编解码预训练阶段

在第 1 阶段训练过程中, 本文使用了二值交叉熵损失 (binary cross entropy loss, BCE loss) 来评估水印解码的准确性, 并通过 learned perceptual image patch similarity (LPIPS)[25]和峰值区域变化损失 (peak regional variation loss, PRVL)[24]感知特征之间的距离增强图像质量, 保证模型生成的图像与原始图像的相似性. 损失函数表达式如下:

$$Loss=5\times Loss_{\text{LPIPS}}+1\times Loss_{\text{BCE}}+1.5\times Loss_{\text{PRVL}} \tag{4}$$

(2) U-Net 微调阶段

在 U-Net 微调阶段, 本文使用了均方误差损失函数 (*MSE* loss) 用于衡量原 U-Net 生成特征与 LoRA 微调特征之间的像素级差异. 通过最小化 *MSE* loss, 模型能够逐步调整生成的特征表示, 使得嵌入的水印信息更加准确地出现在生成图像中, 从而提升水印的质量和鲁棒性.

3 实验结果与分析

3.1 数据集

为验证所提方法中水印的鲁棒性以及图像的保真度, 实验利用 COCO2017test 数据集进行训练和评估. 并使用 AquaLoRA[24]提供的 metadata 文件选择 10 000 张图像以及 prompt 进行训练. 在双阶段训练时对于每次前向传播, 均嵌入不同的水印信息M, 该水印信息M从二进制分布$M\sim\{0,1\}^N$中随机采样.

3.2 评价指标

为客观评估所提出的两阶段水印方案的水印鲁棒性和图像保真度, 本文采用了一系列客观的量化评估标准. 为验证鲁棒性, 我们对照原始水印信息M_i, 评估了提取水印信息M_i'的精度. 对于每个输入图像I_m, *Acc* 代表水印提取准确率, 计算提取的水印信息M_i'和其对应真实水印信息的M_i位准确率, 遍历每个位计算相等位的比例. 此外, 通过真实阳性率 (true positive rate, *TPR*) 在另一个角度描述水印信息的鲁棒性, 公式如下:

$$Acc=\frac{1}{N}\sum_{i=1}^{N}1(M_i=M_i') \tag{5}$$

$$TPR=\frac{TP}{TP+FN} \tag{6}$$

其中, *TP* (true positive) 表示解码后准确率高于阈值的样本数, *FN* (false negative) 是解码后准确率低于阈值的样本数.

对于水印图像I_m'的保真度, 本文采用 *FID* (Fréchet inception distance)[26]以及峰值信噪比 (peak signal-to-noise ratio, *PSNR*) 进行评价, 其中 *FID* 描述图像在特征空间的相似性, 而 *PSNR* 关注图像在视觉上的相似程度. 公式如下:

$$FID=\|\mu_{\text{im}}-\mu_{\text{im}'}\|_2^2+\text{tr}\left(\Sigma_{\text{im}}+\Sigma_{\text{im}'}-2\times(\Sigma_{\text{im}}\Sigma_{\text{im}'})^{\frac{1}{2}}\right) \tag{7}$$

$$PSNR(I_m,I_m')=20\times\log_{10}\frac{MAX(I_m,I_m')-1}{MSE(I_m,I_m')} \tag{8}$$

其中, μ_{im}和Σ_{im}表示训练图像的特征均值和协方差矩阵, $\mu_{\text{im}'}$和$\Sigma_{\text{im}'}$表示水印图像的特征均值和协方差矩阵. tr(·) 表示矩阵的迹运算. 式 (8) 中, I_m与I_m'分别表示为原始图像与生成的图像, *MAX*(·) 为图像的最大像素值, *MSE*(·) 为均方误差.

3.3 实验结果

为体现本文水印方案的有效性, 在 Stable diffusion v2.1 和 Guided diffusion[3]两种扩散模型上进行实验. 所有模型的图像大小都调整为 512×512, 水印长度为 48 bit. 本方案使用 RTX 3090 显卡进行训练, 批大小设置

为 4, 采用默认超参数的 Adam 优化器. 关于扩散模型参数, 本文使用 DDIM[27]采样器, 去噪步数为 50 步, Seed 随机种子默认设置为 42.

在具体实现中, 采用 JPEG 压缩、裁剪与缩放、高斯模糊、高斯噪声和颜色抖动 5 种图像攻击方法用于评估所提出水印方案的有效性.

表 1 为本方法在上述 5 种图像攻击下水印的鲁棒性测试. 本文实验在 Stable diffusion 与 Guided diffusion 两种扩散模型方法上进行, 可以发现所提出方法在面对常见图像攻击时水印几乎不受影响. 这表明, 所提出水印方案在实际场景下有效.

表 1 不同图像攻击下的水印鲁棒性结果

攻击类型	*TPR*	*Acc* (%)
JPEG压缩	0.998/0.997	94.93/94.21
裁剪与缩放	0.923/0.929	91.33/90.12
颜色抖动	0.951/0.942	93.67/93.39
高斯模糊	0.996/0.997	95.98/94.33
高斯噪声	0.963/0.959	93.35/94.11
平均值	**0.9662/0.9468**	**93.852/93.232**

注: /前后分别为Stable diffusion和Guided diffusion上的实验结果

表 2 展示了将本方法分别与 DWT-DCT-SVD[10]、RivaGAN[16]、Tree-Ring[18]以及 Stable Signature[17]根据 *FID* 距离及 *PSNR* 峰值信噪比进行图像保真度对比的测试结果, 从表 2 中可以看出, 所提出方法在图像保真度方面表现优于其他对比方法. 具体而言, 本方法在 *FID* 距离和 *PSNR* 上均有提升, 表明所提出方案在水印不可感知性方面有效.

表 3 提供了不同水印方案与本方案在水印鲁棒性方面的对比, 其中 *Acc* 与 *TPR* 是表 1 中 5 种图像攻击的均值. 从表 3 的结果可以看出, 本方法在应对图像攻击时的水印鲁棒性表现优于其他对比方法. 在各种攻击场景 (如高斯噪声、JPEG 压缩、裁剪、旋转等) 下, 本方法均展现出更高的水印提取精度, 说明其在多种攻击条件下能够有效保持水印信息的完整性. 此外, 图 3 展示了本方法与其他方法在水印提取方面的两个示例. 采用了标准差为 0.2 的高斯噪声作为攻击手段干扰水印. 图 3 从上到下依次为: Tree-Ring[18]、Stable Signature[17]以及本文方法. 通过这两个示例可以观察到, 本文方法生成的水印图像在图像质量方面表现优于其他方法. 同时, 在水印提取时, 针对图 3 中的两个具体示例, 本文方法在具体随机高斯噪声下的提取准确率分别达到了 100% 与 97.91%, 显示出其在噪声干扰下的显著鲁棒性.

表 2 不同水印方案图像保真度对比结果

方案	Stable diffusion		Guided diffusion	
	FID↓	*PSNR* (dB)↑	*FID*↓	*PSNR* (dB)↑
DWT-DCT-SVD[10]	25.21	26.43	24.98	26.87
RivaGAN[16]	24.53	27.92	24.77	28.21
Tree-Ring[18]	25.93	26.79	25.82	27.03
Stable Signature[17]	24.77	27.61	**24.85**	28.53
Ours	**24.62**	**28.56**	24.96	**28.78**

表 3 不同水印方案的水印鲁棒性对比结果

方案	Stable diffusion		Guided diffusion	
	Acc (%)	*TPR*	*Acc* (%)	*TPR*
DWT-DCT-SVD[10]	53.48	0.521	59.48	0603
RivaGAN[16]	92.21	0.877	91.24	0.871
Stable Signature[17]	89.39	0.823	90.63	0.838
Ours	**93.852**	**0.9662**	**93.232**	**0.9468**

图 3 所提出方法与其他方法水印提取示例

为了证明所提出模块的有效性, 本文进行消融实验系统地评估了水印编码模块以及自适应过程对整个算法的影响, 详细结果如表 4 与表 5 所示. 通过更换 StegaStamp 以及 HiDDeN 中提出的编码器, 证明了水印编码模块的有效性. 此外, 通过移除自适应过程分支, 水印信息直接通过编码及去噪, 验证了该模块在整个方法中的有效性. 上述所有的消融实验均在以 Stable diffusion 作为生成模型的背景下进行.

表 4 关于水印编码模块的消融实验

方法	*Acc* (%)	*TPR*	*FID*	*PSNR* (dB)
HiDDeN[16]	91.53	0.866	25.33	28.08
StegaStamp[14]	92.91	0.903	25.14	28.17
Ours	**93.852**	**0.9662**	**24.62**	**28.56**

表 5 关于自适应过程的消融实验

方法	*Acc* (%)	*TPR*	*FID*	*PSNR* (dB)
无自适应过程	92.323	0.9023	27.23	26.93
Ours	**93.852**	**0.9662**	**24.62**	**28.56**

4 结语

本文提出了基于水印 LoRA 的双阶段扩散模型水印方案, 有效地提高了水印对图像攻击时的不可感知性和鲁棒性. 此外, 所提出方法能够根据不同水印信息自适应选择最优的扩散过程, 生成高质量的水印图像. 该方法解决了多个用户之间的识别问题, 并有效避免恶意用户绕过水印. 最后, 在 COCO2017test 数据集上使用 Stable diffusion 以及 Guided diffusion 两种不同的扩散模型进行大量实验. 实验表明, 此方法有效提升了扩散模型水印的不可感知性和鲁棒性, 并且解决了多种扩散模型的普适性问题.

参考文献

1 Ramesh A, Dhariwal P, Nichol A, *et al*. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125, 2022.

2 Rombach R, Blattmann A, Lorenz D, *et al*. High-resolution image synthesis with latent diffusion models. arXiv:2112.10752, 2022.

3 Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021. 672.

4 刘安安, 苏育挺, 王岚君, 等. AIGC 视觉内容生成与溯源研究进展. 中国图象图形学报, 2024, 29(6): 1535–1554. [doi: 10.11834/jig.240003]

5 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. 网络空间安全科学学报, 2024, 2(1): 13–39. [doi: 10.20172/j.issn.2097-3136.240102]

6 Gault M. An AI-generated artwork won first place at a state fair fine arts competition, and artists are pissed. https://www.vice.com/en/article/bvmvqm/an-ai-generated-artwork-won-first-place-at-a-state-fair-fine-arts-competition-and-artists-are-pissed. [2024-11-01].

7 Brundage M, Avin S, Clark J, *et al*. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv:1802.07228, 2024.

8 Srinivasan R, Parikh D. Building bridges: Generative artworks to explore AI ethics. arXiv:2106.13901, 2021.

9 Nightingale SJ, Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the 2019 National Academy of Sciences of the United States of America, 2022, 119(8): e2120481119. [doi: 10.1073/pnas.2120481119]

10 Cox IJ, Miller ML, Bloom JA, *et al*. Digital watermarking and steganography. Amsterdam: Morgan Kaufmann, 2008.

11 Fernandez P, Sablayrolles A, Furon T, *et al*. Watermarking images in self-supervised latent spaces. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 3054–3058. [doi: 10.1109/ICASSP43922.2022.9746058]

12 Kishore V, Chen XY, Wang Y, *et al*. Fixed neural network steganography: Train the images, not the network. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

13 Tancik M, Mildenhall B, Ng R. StegaStamp: Invisible hyperlinks in physical photographs. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2114–2123. [doi: 10.1109/CVPR42600.2020.00219]

14 Jia ZY, Fang H, Zhang WM. MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression. Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021. 41–49. [doi: 10.1145/3474085.3475324]

15 Zhu JR, Kaplan R, Johnson J, *et al*. HiDDeN: Hiding data with deep networks. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 682–697.

16 Zhang KA, Xu L, Cuesta-Infante A, *et al*. Robust invisible

video watermarking with attention. arXiv:1909.01285, 2019.

17 Fernandez P, Couairon G, Jégou H, *et al*. The stable signature: Rooting watermarks in latent diffusion models. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 22409–22420.

18 Wen YX, Kirchenbauer J, Geiping J, *et al*. Tree-Ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv:2305.20030, 2023.

19 Ci H, Yang P, Song YR, *et al*. RingID: Rethinking Tree-Ring watermarking for enhanced multi-key identification. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 338–354. [doi: 10.1007/978-3-031-73390-1_20]

20 Hu EJ, Shen YL, Wallis P, *et al*. LoRA: Low-rank adaptation of large language models. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

21 Lei LQ, Gai KK, Yu J, *et al*. Diffusetrace: A transparent and flexible watermarking scheme for latent diffusion model. arXiv:2405.02696, 2024.

22 Xiong C, Qin C, Feng GR, *et al*. Flexible and secure watermarking for latent diffusion model. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023. 1668–1676. [doi: 10.1145/3581783.3612448]

23 Zhang GK, Wang LJ, Su YT, *et al*. A training-free plug-and-play watermark framework for stable diffusion. arXiv:2404.05607, 2024.

24 Feng WT, Zhou WB, He JY, *et al*. AquaLoRA: Toward white-box protection for customized stable diffusion models via watermark LoRA. Proceedings of the 41st International Conference on Machine Learning. Vienna: OpenReview.net, 2024. 11135.

25 Zhang R, Isola P, Efros AA, *et al*. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595.

26 Heusel M, Ramsauer H, Unterthiner T, *et al*. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.

27 Song JM, Meng CL, Ermon S. Denoising diffusion implicit models. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021. 02502.

(校对责编: 王欣欣)