Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(4), 967-976
Published Online July 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.144092

Hans汉斯

面向AIGC生成图像的水印发展进程

耿新晨

北京印刷学院信息工程学院，北京

收稿日期：2025年5月21日；录用日期：2025年7月21日；发布日期：2025年7月30日

摘　要

随着人工智能生成内容(AIGC)技术的迅速发展，扩散模型(Diffusion Models, DMs)在图像生成和编辑领域展现了巨大的潜力。然而，扩散模型生成内容的真实性难以辨别，滥用生成模型可能引发隐私泄露、版权侵权等社会问题。为了解决这一问题，数字水印技术逐渐成为保护AIGC模型版权、追踪生成内容的重要手段。本文从图像生成技术的发展、传统和最新的数字图像水印算法以及针对AIGC的水印方法三个核心领域进行了综述。此外，我们还研究了该领域中使用的常见性能评估指标。最后，对研究中存在的问题进行了讨论，并提出了今后的研究方向。

关键词

版权保护，水印，人工智能生成内容

The Development Progress of Watermarking for AIGC-Generated Images

Xinchen Geng

Department of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: May 21st, 2025; accepted: Jul. 21st, 2025; published: Jul. 30th, 2025

Abstract

With the rapid advancement of Artificial Intelligence Generated Content (AIGC) technologies, diffusion models (DMs) have demonstrated tremendous potential in image generation and editing. However, it is often difficult to distinguish the authenticity of content generated by diffusion models, and their misuse may lead to issues such as privacy breaches and copyright infringement. To address these challenges, digital watermarking has emerged as a crucial approach for protecting the copyright of AIGC models and tracing generated content. This paper reviews the development of image generation technologies, traditional and cutting-edge digital watermarking algorithms, and water-

文章引用: 耿新晨. 面向 AIGC 生成图像的水印发展进程[J]. 人工智能与机器人研究, 2025, 14(4): 967-976.
DOI: 10.12677/airr.2025.144092

marking methods specifically designed for AIGC. Furthermore, commonly used performance evaluation metrics in this field are examined. Finally, the paper discusses current research challenges and outlines future research directions.

Keywords

Copyright Protection, Watermarking, Artificial Intelligence Generated Content

1. 引言

随着人工智能(AI)技术的迅猛发展，尤其是生成式人工智能(AIGC, Artificial Intelligence Generated Content)技术的突破性进展，多个行业正经历着前所未有的变革。AIGC 作为 AI 技术的一项重要应用，已在语言生成、图像生成、音频生成、视频生成等领域取得了显著成就，并极大地推动了创作力和生产力的提升。在图像生成方面，像 DALL-E、Midjourney、Stable Diffusion 等基于扩散模型的图像生成技术，不仅能够根据用户输入的文本描述生成高质量的艺术作品，还极大地扩展了创意行业的生产力和创新空间。这些技术的广泛应用，不仅为传统创作方式带来了革命性变革，也为各行各业带来了更高效、更低成本的生产手段。

然而，尽管 AIGC 技术为社会发展带来了诸多便利和创新，其带来的隐私安全、版权保护、伦理风险等问题也日益突出。随着 AIGC 技术的商业化应用越来越广泛，内容的真实性、所有权以及滥用问题成为亟待解决的关键问题。生成式模型能够快速生成虚假信息、伪造身份、甚至制造深度伪造内容，这些内容可能被恶意用作政治操控、社会舆论误导、甚至身份盗窃等活动。因此，如何有效地检测、追溯生成内容的来源，确保生成内容的合法性和安全性，已成为学术界和产业界关注的焦点。

同时，随着 AIGC 技术的发展，相关的法律和伦理问题也成为研究的重点。AIGC 生成内容的版权归属问题一直没有明确的法律规定。根据国际知识产权组织(WIPO)的最新研究，人工智能生成作品的版权归属仍处于探索阶段，特别是在 AIGC 应用中生成内容的所有权问题，引发了大量关于版权的争论。不同国家和地区对于 AIGC 内容的监管政策也逐步出台，如美国白宫在 2023 年发布的 AIGC 监管规定，要求对生成内容的安全性、公平性和民权进行严格评估；中国也出台了《网络安全标准实践指南——生成式人工智能服务内容标识方法》，明确要求 AIGC 服务提供商在生成内容中嵌入可识别的标识或水印，以确保内容的追溯和法律责任的落实。

因此，在 AIGC 技术不断发展的背景下，针对生成图像水印的研究显得尤为重要。这不仅关系到生成内容的版权保护和合法性，还与防范 AIGC 技术滥用、保护用户隐私、规范生成内容的使用有着密切的联系。本文旨在综述 AIGC 生成图水印的现状与挑战，探讨当前水印嵌入技术的研究进展，并展望未来可能的研究方向。特别是，在基于扩散模型(DMs)新兴生成模型中，如何结合水印技术提高内容的可追溯性和安全性，将是本文的重点讨论内容。

2. 图像生成技术的发展历程

本节将简要回顾图像生成技术的发展历程，并详细阐述基于扩散模型与潜在扩散模型的基本原理，最后总结当前 AIGC 图像生成技术的主要方法。

2.1. 图像生成技术的发展历程

早期的图像生成模型主要包括自回归模型(如 PixelCNN) [1]、变分自编码器(Variational Autoencoder, VAE) [2]和生成对抗网络(Generative Adversarial Network, GAN) [3]等。VAE 由 Kingma 和 Welling 于 2013 年提出，通过神经网络编码器将图像映射到潜在高斯分布，再由解码器重构图像。由于其训练依赖于像素级别的似然最大化，VAE 往往输出较为模糊的图像。GAN 则在 2014 年由 Goodfellow 等引入，其核心思想是使用生成器和判别器对抗训练：生成器从随机噪声中合成图像，判别器则区分真伪样本，两者交替优化直到判别器对真假样本无法准确区分。然而 GAN 训练不稳定，易出现模式崩溃，需要大量数据和计算资源才能获得满意效果。与此同时，自回归模型可以生成像素级别精细图像，但计算效率有限。总体而言，这些早期模型在生成质量上取得了突破，但也存在训练困难或生成效果有限等问题。

2015 年，Sohl-Dickstein 等[4]从物理学视角提出基于扩散过程的生成模型理念，2020 年 Ho 等正式提出 DDPM(Denoising Diffusion Probabilistic Model) [5]，通过逐步向图像加入高斯噪声再逆向去噪，实现高质量图像合成。近年又出现改进，如 2020 年的 DDIM [6]加速采样等。2022 年，Rombach 等提出潜在扩散模型(Latent Diffusion Model，如 Stable Diffusion) [7]，即先用预训练 VAE 将图像映射到低维潜在空间，再在潜在空间中做扩散去噪，大幅降低计算量。Stable Diffusion 综合了 VAE、UNet 和跨注意力机制，能以相对高效的方式生成高分辨率图像。图 1 中标出了上述重要模型及发表年份。总之，从早期的自编码结构(VAE)和对抗训练(GAN)到最新的扩散模型和潜在扩散模型，图像生成技术不断演进，质量和可控性也随之提升。如图 1 所示，简单总结了图像生成的相关技术。

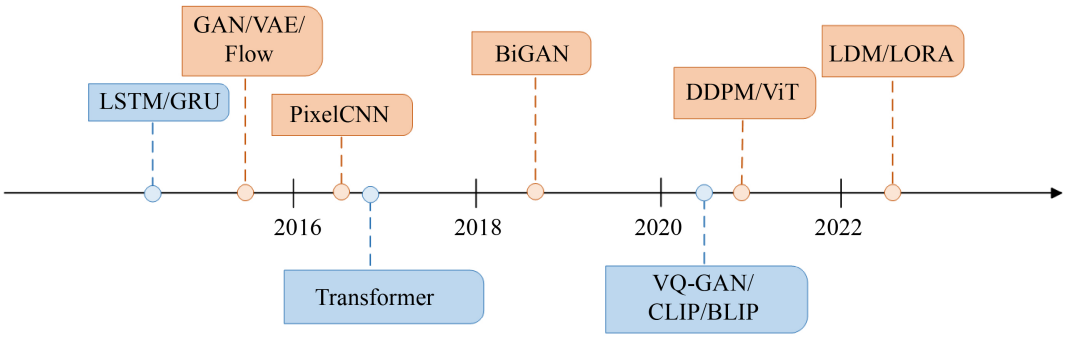

Figure 1. Timeline of generative models
图 1. 生成模型的时间线

2.2. 扩散模型与潜在扩散模型原理

2015 年，首次提出基于热力学思想的扩散式生成方法，将非平衡扩散理论引入深度生成模型。其基本思想借鉴物理扩散过程，将像素看作墨水分子在水中扩散的模拟：如图 2 所示，在正向过程中，将高斯噪声逐渐加到训练图像上，使其逐步“扩散”成纯噪声；在逆向过程中，模型学习逐步去噪还原图像。训练完成后，只需从纯噪声开始，多步迭代地执行逆扩散即可生成新的样本，其原理与去噪自编码器类似。相较于 GAN 等模型，扩散模型训练更为稳定，因为它无须对抗训练，也不易出现模式崩溃；同时它可以产生多样性高且质量优异的图像样本。Ho 等人提出了去噪扩散概率模型(DDPM)采用变分推断的方法对扩散模型进行训练，实验表明 DDPM 可以生成与当时 GAN 相媲美的高质量图像。此外，Song 等人提出了去噪扩散隐式模型(DDIM)，该模型设计了非马尔可夫扩散过程，在保持与 DDPM 相同训练目标的前提下极大加快了采样速度，据报道可使生成速度比 DDPM 快 10~50 倍。

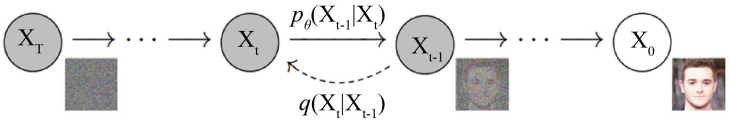

Figure 2. Diffusion process [5]
图 2. 扩散流程[5]

为了在保持生成质量的同时降低计算复杂度，研究者提出了潜在扩散模型(LDM)。2022 年，Rombach 等人在论文《High-Resolution Image Synthesis with Latent Diffusion Models》中描述了这一方法：首先使用预训练的变分自编码器(VAE)将高分辨率图像映射到低维潜在空间，然后在这个潜在空间上训练扩散模型。相比在像素空间直接扩散，在潜在空间建模可以大幅减少数据维度，降低训练和采样成本，而同时保持细节信息，显著提升视觉质量。该论文还在网络中引入跨注意力层，使模型能够根据文本或区域标记等条件进行控制，从而灵活地实现文本到图像、区域到图像等多种条件图像合成。实验结果表明，LDM 在图像修复、类别条件生成等任务上达到了新的最优水平，并在文本到图像等任务上取得了高度竞争力的结果，同时所需计算资源远低于传统的像素级扩散模型。如图 3 所示，LDM 的基本结构。

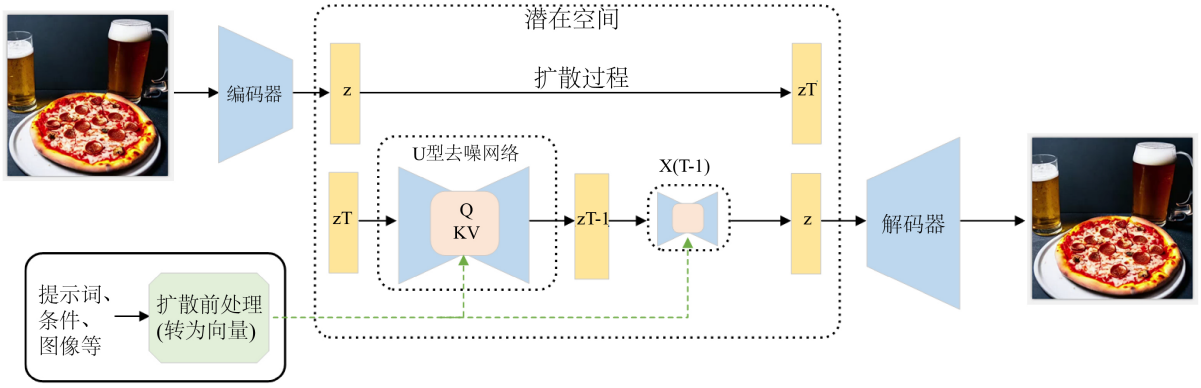

Figure 3. Architecture of LDM (Latent Diffusion Model)
图 3. LDM (潜在扩散模型)结构

2.3. AIGC 图像生成技术方法总结

当前主流的 AIGC 图像生成方法可以粗略分为四大类：自回归模型、VAE 类模型、GAN 类模型和扩散模型。自回归模型按像素顺序逐步生成图像，理论上可产生高质量样本，但需要序列化采样，生成速度很慢，难以并行。VAE 模型使用编码器－解码器结构学习潜在分布，能以端到端方式生成样本，但因最大似然损失通常导致结果模糊。GAN 模型通过对抗训练产生非常清晰的图像，在面对大量无标签图像时表现出色，但训练不稳定且缺少明确生成概率。扩散模型结合了前两者的优点，通过逐步添加、去除噪声生成样本，能够达到目前最高的图像质量和多样性，但其生成过程计算量较大，需要多次迭代采样。选择合适的模型需根据场景需求：当需求最大逼真度时 GAN、扩散模型优先；当要求生成速度或多样性时 VAE 或条件自回归可考虑。

3. 传统与最新数字图像水印算法及其性能评估

本节将介绍相关的数字水印技术，随着图像生成技术的发展，数字水印技术也得到了快速的进步。根据图像生成技术的时间线，数字水印技术可以将其分为三大类。第一类是传统的数字水印技术，主要通过修改像素或是修改频域值实现。第二类是基于深度学习的水印技术自适应的将水印信息嵌入到载体

图像中，使其具有更高的鲁棒性。第三类是在生成过程中加入水印的方法，即基于扩散模型的水印技术，专门保护 AIGC 的生成内容，并在生成质量和水印鲁棒性二者间取得良好的平衡。最后，本节还简述水印技术需要的评估指标。

3.1. 传统的水印技术

传统的数字水印方法主要依赖于对图像某些人工设计的可感知或不可感知特征进行分析和操作，例如直接修改像素值或处理其频域系数等。这类方法通常需要人工经验和领域知识来选择合适的特征进行水印嵌入，而不同的图像特征对水印的嵌入效果和鲁棒性也会产生显著影响。根据水印信息嵌入所在的域不同，传统数字图像水印技术大致可以划分为空间域水印和变换域水印。如图 4 所示，描述了水印嵌入、提取的整体流程。

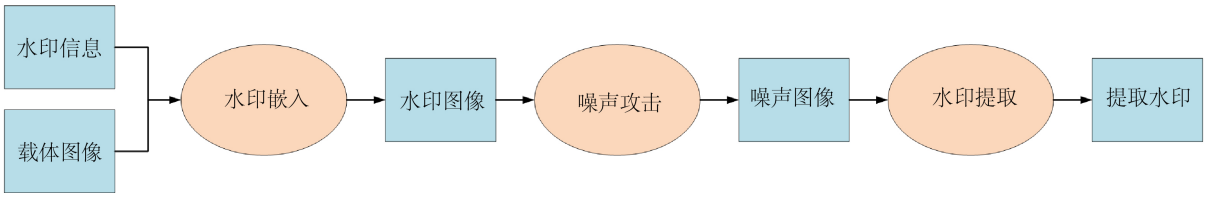

Figure 4. Overall watermarking pipeline
图 4. 水印整体流程图

空间域水印[8]方法是最直观的一类水印技术，其通过直接在图像的像素层面修改数值来嵌入水印。这些方法通常简单易实现，计算成本较低，代表性算法是最低有效位(Least Significant Bit, LSB)替换算法[9]。LSB 等方法虽然可以实现水印的不可见性，但由于修改的是图像最脆弱的部分，因此在面对裁剪、压缩、噪声等攻击时通常表现出较差的鲁棒性。

为了解决空间域方法鲁棒性不足的问题，研究者提出了变换域水印[10]方法。这类方法将图像从空间域变换到频域或其他变换域(如离散余弦变换 DCT、小波变换 DWT、离散傅里叶变换 DFT 等[11]-[16])，并在变换系数中嵌入水印信息，如图 5 所示。由于频域特征更贴近人类视觉系统的感知特性，这使得变换域水印在保持图像质量的同时可以嵌入更多信息，并获得更高的鲁棒性。例如，DWT-DCT 组合方法就广泛用于嵌入高容量水印且能抵御常见图像处理攻击。

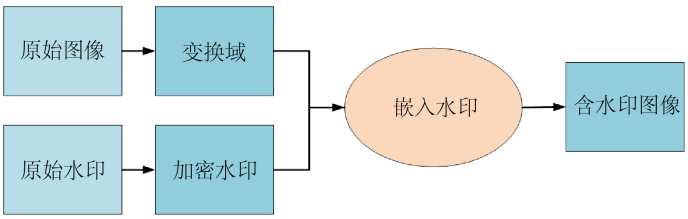

Figure 5. Watermark embedding process in the transform domain
图 5. 变换域水印嵌入流程

3.2. 基于深度学习的水印技术

基于深度学习的水印[17]算法通过神经网络自适应寻找稳定的嵌入空间并对各种攻击展开对抗学习，在嵌入容量、不可见性等方面都优于一些传统方法。2018 年首次提出 HiDDeN [18]，一种基于深度学习的图像水印框架，通过卷积神经网络进行数据隐藏与提取。该框架采用三个主要的卷积网络：编码器、

解码器和判别器，如图 6 所示。

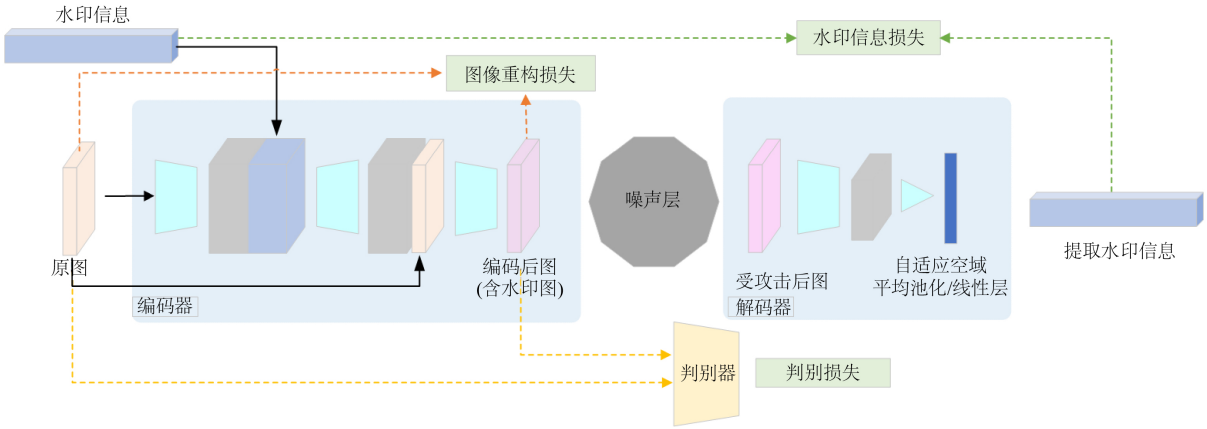

Figure 6. Overall framework of HiDDeN
图 6. HiDDeN 整体框架

编码器接收两种输入：原始图像和消息。编码器网络的目标是将原始图像和消息进行融合，并输出一个包含水印信息的含水印图像。这个图像包含了对原图的微小变化，但足够隐蔽，以至于人眼无法察觉。解码器接收含水印图像，并从中提取出隐藏的消息。解码器的目的是尽可能准确提取嵌入的水印消息。通过训练，解器能够识别和还原经过编码器处理后的图像中嵌入的信息。判别器的作用是在训练过程中对含水印图像和原始图像的区别进行评估。它学习如何区分包含水印信息的图像和原始图像，从而促使编码器生成更为自然、隐蔽的水印图像。通过判别器的反馈，模型不断调整，优化水印图像的不可感知性。

基于以上水印技术，可以为图像等多媒体数据进行保护和溯源，但针对生成式的数据，对其进行后处理，会导致图像内容受损并增加计算量。基于后处理的水印[19] [20]调整鲁棒图像特征以嵌入水印，从而直接改变图像并降低其质量。为了减轻这种担忧，最近的研究提出了基于微调的方法[21]-[24]，将水印嵌入过程与图像生成过程合并在一起。直观地说，这些方法需要修改模型参数，引入了额外的计算开销。由此，为了减少后处理的操作并结合扩散模型的特性[25]，设计不影响模型性能的情况下嵌入水印，使图像在生成后就蕴含水印信息。

3.3. 基于扩散模型的水印技术

为进一步明确本文所采用扩散水印方法的定位，我们对传统空域水印、频域水印以及深度学习水印方法的典型特性进行了总结与对比。空域方法结构简单、嵌入速度快，但在鲁棒性和隐蔽性方面相对较弱；频域方法在鲁棒性上有所提升，尤其在抗压缩方面表现更优，但面对复杂多变的攻击仍存在一定局限性，例如在 JPEG 高压缩率条件下易丢失水印信息，因此在实际应用中仍面临挑战。深度学习水印则是通过神经网络自适应嵌入水印信息，并设计噪声层以模拟实际可能遇到的攻击，从而提升水印的鲁棒性。

随着 AIGC 技术的快速发展，扩散模型在图像生成任务中得到广泛应用，同时也引发了诸如虚假新闻、图像伪造等潜在安全风险。为实现对生成内容的有效溯源，近年来提出了基于扩散模型的水印方法。该类方法不同于传统的后处理方式，能够在图像生成过程中直接嵌入水印信息，将其与图像生成所需的潜在特征深度融合，在保持较高视觉质量的同时显著提升了水印的鲁棒性，为生成式内容的管理与溯源等新兴应用场景提供了有力支持。

开源的图像生成模型 LDM [7]通过传统数字图像水印算法为生成的图像添加水印，虽然后处理的方

法在应用中相对灵活，但这类方法并未涉及图像生成过程，水印容易被移除。针对生成式图像的特点，研究人员提出在图像生成过程中添加水印，直接生成具有不可见水印的可追踪图像。LDM [22]通过微调模型将水印与模型耦合，这种水印一般难以移除，但由于水印与模型耦合在一起，若要对另一个模型嵌入水印，则需要对该模型微调，复杂度较高。Wen 等[25]从另一个角度提出了一种灵活的树环(Tree-ring)水印方法，该方法不依赖于模型微调，通过在初始噪声向量中添加水印图案，影响整个采样过程，产生人类无法察觉的模型水印。通过潜在空间的优化，突破了传统水印技术在视觉质量和鲁棒性之间的平衡。从理论角度推动了生成模型的多重功能扩展，不仅提升了潜在空间在图像生成中的表现，也拓宽了生成模型在信息保护、版权保护等领域的应用范围。

3.4. 评估指标

水印技术需要注重视觉质量、鲁棒性和嵌入容量三者的权衡，下面介绍实验中所需的评价指标。

3.4.1. 视觉质量

图像的视觉质量是判断图像编码器优劣的重要指标，最常用的评价标准是峰值信噪比(Peak signal-to-noise Ratio，简称 PSNR)和结构相似度(structural similarity，简称 SSIM)。嵌入操作不可避免地影响视觉质量。PSNR 表示原始图像与重构图像在像素域的差值，定义为

$$\mathrm{PSNR}=10\cdot\log\frac{\mathrm{MAX}^2}{\mathrm{MSE}} \tag{1}$$

$$\mathrm{MSE}=\frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(a(i,j)-b(i,j)\right)^2 \tag{2}$$

式中，a 和 b 分别为原图像素值和重构像素值，尺寸为 $W\times H$ 。MSE 是原图像素值和重构像素值的均方误差，MAX 表示当前图像的最大像素值。PSNR 的单位是 dB，数值越大表示失真越小。因为数值越大代表 MSE 越小。MSE 越小代表两张图片越接近，失真就越小。这个指标很容易计算，但不能反映人眼的视觉特性。

另一个度量 SSIM 将失真建模为三个不同的因素：亮度、对比度和结构。它被定义为

$$\mathrm{SSIM}=\frac{\left(2\mu_a\mu_b+c_a\right)\left(2\sigma_{ad}+c_b\right)}{\left(\mu_a^2+\mu_b^2+c_a\right)\left(\sigma_a^2+\sigma_b^2+c_b\right)} \tag{3}$$

其中，μ_a 和 μ_b 为原图像素值和重构像素的均值，σ_a^2、σ_b^2、σ_{ab} 为原图像素值和重构像素的方差及其协方差。SSIM 取值范围为 $[-1,1]$，值越大表示视觉质量越好，能反映人眼的视觉特性。

3.4.2. 比特准确率

采用位精度(Bit accuracy)来评估水印的鲁棒性，比特准确率由水印序列中正确解码的位所占的比例定义。图像的比特准确率是一个非常重要的指标，其定义如下：

$$\text{Bit ACC}=1-\frac{\sum_{i=1}^{L}\left|m_i-m_i'\right|}{L} \tag{4}$$

其中：m_i 和 m_i' 分别表示第i位原始水印和提取水印信息，L 代表水印序列长度。

3.4.3. 其他评价指标

FID 通过计算真实图像和生成图像特征分布之间的 Fréchet 距离，来量化生成模型输出的质量和多样性，分数越低表示越接近于真实图像。

嵌入容量：在利用扩散模型生成图像的过程中，理论上可以将水印信息嵌入至与潜空间向量相同维

度的区域，实现较大的信息容量。然而，若嵌入的信息量过大，可能会扰乱初始高斯噪声的分布特性，使其偏离原始设计的标准高斯分布，从而对扩散过程中的图像生成质量产生负面影响。水印容量是衡量水印系统性能的一个关键指标，指在可接受的失真范围内，能够嵌入的水印信息的最大数量或比例。换句话说，水印容量决定了可嵌入的有效信息量，在确保水印不可察觉性与鲁棒性的前提下，容量越大，系统的实用性与信息表达能力越强。

提高水印容量需要在嵌入强度与不可察觉性之间进行平衡：容量越大，可能引入的视觉或统计失真也越明显，因此在设计水印算法时，需根据应用需求合理设置容量目标，兼顾安全性、鲁棒性与隐蔽性。

4. 针对 AIGC 的水印方法研究

根据以上文献综述，数字水印技术随着图像生成技术的发展也在更新迭代，现已经取得了显著的进展。但是，由于 AIGC 生成模型的复杂性，基于扩散模型的数字水印技术仍然存在许多值得进一步探索的问题。

4.1. 考虑轻量的水印嵌入模块设计

为了保持原始生成模型的结构以及生成性能，同时减少计算资源的消耗，需要设计合理的水印嵌入模块。水印嵌入模块不仅包含水印的嵌入方式同时还要考量水印的预处理。水印的鲁棒性需要从图像中的冗余水印信息获得，所以水印嵌入模块的设计需要考虑更多的因素。但是，各因素之间的相互作用是什么、对水印嵌入的设计有什么影响也应得到充分考虑。

其次，为了保障水印信息的冗余性，目前的方法多进行多次嵌入以实现较高的鲁棒性，但是多次嵌入的方式会导致生成图像的视觉质量大大降低。同时针对一些攻击，多次嵌入的方式并没有提升水印的鲁棒性，反而降低图像视觉质量、增加计算消耗。可以看出水印信息和原始内容之间的融合是水印嵌入的一大挑战，如何充分利用它们之间的相互作用来设计一个轻量级水印嵌入模块是一个值得探索的关键问题。

4.2. 基于修改扩散模型潜在变量的水印算法

方法利用扩散模型的特性将水印信息无缝集成到扩散流程中，通常会带来较大的资源消耗和较长的训练时间。例如，需要对 LDM 中的 U-Net 结构进行微调，或重新训练一个具备水印嵌入与提取能力的编解码器。为缓解上述问题，本文提出在冻结的预训练 LDM 层中集成水印特征，在不改动原始扩散模型结构的前提下，实现水印信息与潜在特征的融合。

具体而言，LDM 首先通过编码器将图像下采样为潜在变量 z，并在潜在空间中执行扩散过程，随后再通过解码器将潜在变量上采样以重建图像。在此框架下，水印嵌入被设计在编码之后、解码之前，从而避免对多个扩散阶段中 U-Net 模块的微调，显著降低了训练资源的消耗。更具体地说，设待嵌入的水印消息为 m，首先通过全连接(FC)层对其进行预处理，并结合重塑或上采样插值操作，生成与潜在变量 z 维度一致的水印特征 δ。随后将其以残差形式嵌入到潜在表示中，嵌入过程可表示如下公式：

$$z = z + \delta \tag{5}$$

该方法通过在预训练扩散模型中冻结主体结构，仅在潜在空间中引入水印特征，以残差方式实现嵌入，避免了对 U-Net 等扩散核心组件的微调，从而大幅降低了训练成本与计算资源消耗。此外，由于水印与潜在特征在压缩空间中深度融合，该策略在保持图像生成质量的同时具备良好的隐蔽性和鲁棒性。

4.3. 扩散模型水印算法存在的问题和未来可能的研究方向

随着深度生成模型的迅速发展，扩散模型在图像生成与编辑中的表现逐渐超过传统生成框架，其在

图像水印与版权保护中的应用也引起了广泛关注。当前，基于扩散模型的图像水印技术仍处于探索阶段，尚未形成系统而通用的方法体系。本文认为，这主要受限于两个方面：一方面，扩散模型本身具有较高的计算复杂度，将水印嵌入过程融入扩散推理或训练流程，会显著增加整体资源消耗；另一方面，针对通用性和鲁棒性需求，水印技术往往需要在生成质量和嵌入稳定性之间取得平衡，而这在扩散模型中仍是挑战。

目前，基于扩散模型的图像水印方法主要可分为两类研究路径：其一是基于解码器微调的水印嵌入，即在保留扩散模型主干结构的基础上，通过微调 LDM 解码器，在解码阶段注入特定水印信息。这种方式具有较强的稳定性和适应性，更适合实际应用场景，但其训练过程较为复杂，计算资源消耗大；其二是基于初始潜在表示的水印嵌入，即直接修改扩散过程的起始噪声或潜在表示，使其在生成图像中隐式携带语义水印信息。这种方法无需修改模型参数，适合专用场景快速部署，但缺乏通用性和鲁棒性，且对扩散流程的控制要求较高。

此外，未来研究还可探索生成式图像水印的设计思路，通过构建水印与潜在变量之间的直接映射，实现无需显式载体修改的“无载体”隐写机制，从而在保持原始图像结构的前提下增强抗分析能力。这一方向有望突破现有基于像素或频域修改的水印方法的局限，为 AIGC 时代下的版权确权与内容追溯提供更高效的解决方案。

5. 研究中的挑战与未来发展方向

在当今数字化迅猛发展的背景下，人工智能生成内容(AIGC)技术已经广泛应用于文本、图像、音频和视频等多个领域。这些内容愈发逼真，不仅极大推动了内容创作效率，也对传统的版权保护机制和内容管理方式带来了深远影响。然而，AIGC 在带来机遇的同时，也伴随着诸多挑战，尤其集中在版权归属、内容真实性、伦理责任以及技术滥用等方面。

首先，版权保护问题日益严峻。由于 AIGC 生成内容的原创性难以界定，传统的版权确权和追踪手段难以适用。虽然已有如腾瑞云推出的 CPSP 数字版权资产服务平台等尝试借助数字水印来实现内容确权，但当前水印技术在跨平台应用和鲁棒性方面仍存在局限，如何确保其长期有效仍是待解难题。

其次，内容真实性面临新挑战。随着生成模型的质量不断提升，AIGC 产出的文本和图像与人类创作的差异愈发难辨。这不仅影响新闻报道、司法证据等严肃场景中的真实性判断，也使得虚假信息在社交平台上的传播风险倍增。

第三，伦理和责任问题日趋复杂。AIGC 可能被用于伪造身份、编造虚假内容、侵犯隐私等，甚至对公共安全构成威胁。当前的监管体系尚未能有效厘清 AIGC 生成内容中各方的责任归属问题。

第四，技术滥用风险不可忽视。AIGC 作为一项通用生成技术，其“中性”特征意味着既能用于艺术创作，也可能被滥用于诈骗、造假、网络攻击等不法行为。因此，亟需在技术研发之初融入风险防控机制。

面对这些挑战，未来的应对方向应包括：提升版权识别与确权水印技术的适用性与鲁棒性，制定更严格的法律法规，推动全球范围内跨平台协同机制的建立，以及加强公众 AI 素养教育，提升其对 AIGC 生成内容的识别与辨别能力，从而实现 AIGC 技术的可持续健康发展。本文阐述了未来可能的研究方向，希望能为读者提供一些参考。

参考文献

[1] Van den Oord, A., Kalchbrenner, N., Espeholt, L., *et al*. (2016) Conditional Image Generation with PixelCNN Decoders. *Advances in Neural Information Processing Systems*, **29**.

[2] Kingma, D.P. and Welling, M. (2013) Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114

[3] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, **27**.

[4] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., *et al*. (2015) Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *International Conference on Machine Learning*, Lille, 6-11 July 2015, 2256-2265.

[5] Ho, J., Jain, A. and Abbeel, P. (2020) Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, **33**, 6840-6851.

[6] Song, J., Meng, C. and Ermon, S. (2020) Denoising Diffusion Implicit Models. arXiv: 2010.02502.

[7] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) High-Resolution Image Synthesis with Latent Diffusion Models. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 10674-10685. https://doi.org/10.1109/cvpr52688.2022.01042

[8] 王树梅. 数字图像水印技术综述[J]. 湖南理工学院学报(自然科学版), 2022, 35(1): 31-36+68.

[9] 冯乐, 朱仁杰, 吴汉舟, 等. 神经网络水印综述[J]. 应用科学学报, 2021, 39(6): 881-892.

[10] 赵蕾, 桂小林, 邵屹杨, 等. 数字图像多功能水印综述[J]. 计算机辅助设计与图形学学报, 2024, 36(2): 195-222.

[11] Barni, M., Bartolini, F., Cappellini, V. and Piva, A. (1998) A DCT-Domain System for Robust Image Watermarking. *Signal Processing*, **66**, 357-372. https://doi.org/10.1016/s0165-1684(98)00015-2

[12] Piva, A., Barni, M., Bartolini, F. and Cappellini, V. (1997) DCT-Based Watermark Recovering without Resorting to the Uncorrupted Original Image. *Proceedings of International Conference on Image Processing*, Santa Barbara, 26-29 October 1997, 520-523. https://doi.org/10.1109/icip.1997.647964

[13] Daren, H., Jiufen, L., Jiwu, H., *et al*. (2001) A DWT-Based Image Watermarking Algorithm. *IEEE International Conference on Multimedia and Expo*, 2001. Tokyo, 22-25 August 2001, 313-316.

[14] Jiansheng, M., Sukang, L. and Xiaomei, T. (2009) A Digital Watermarking Algorithm Based on DCT and DWT. *The* 2009 *International Symposium on Web Information Systems and Applications* (*WISA* 2009), Academy Publisher, 104.

[15] Urvoy, M., Goudia, D. and Autrusseau, F. (2014) Perceptual DFT Watermarking with Improved Detection and Robustness to Geometrical Distortions. *IEEE Transactions on Information Forensics and Security*, **9**, 1108-1119. https://doi.org/10.1109/tifs.2014.2322497

[16] Kang, X., Huang, J., Shi, Y.Q., *et al*. (2003) A DWT-DFT Composite Watermarking Scheme Robust to Both Affine Transform and JPEG Compression. *IEEE Transactions on Circuits and Systems for Video Technology*, **13**, 776-786.

[17] 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.

[18] Zhu, J. (2018) HiDDeN: Hiding Data with Deep Networks. arXiv: 1807. 09937.

[19] Cox, I., Miller, M., Bloom, J., *et al*. (2007) Digital Watermarking and Steganography. Morgan Kaufmann.

[20] Zhang, K.A., Xu, L., Cuesta-Infante, A., *et al*. (2019) Robust Invisible Video Watermarking with Attention. arXiv: 1909. 01285.

[21] Cui, Y., Ren, J., Xu, H., *et al*. (2023) Diffusionshield: A Watermark for Copyright Protection against Generative Diffusion Models. arXiv: 2306. 04642.

[22] Fernandez, P., Couairon, G., Jégou, H., Douze, M. and Furon, T. (2023) The Stable Signature: Rooting Watermarks in Latent Diffusion Models. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 22409-22420. https://doi.org/10.1109/iccv51070.2023.02053

[23] Liu, Y., Li, Z., Backes, M., *et al*. (2023) Watermarking Diffusion Model. arXiv: 2305. 12502.

[24] Xiong, C., Qin, C., Feng, G. and Zhang, X. (2023) Flexible and Secure Watermarking for Latent Diffusion Model. *Proceedings of the* 31*st ACM International Conference on Multimedia*, Ottawa, 29 October 2023-3 November 2023, 1668-1676. https://doi.org/10.1145/3581783.3612448

[25] Wen, Y., Kirchenbauer, J., Geiping, J., *et al*. (2023) Tree-Ring Watermarks: Fingerprints for Diffusion Images that Are Invisible and Robust. arXiv: 2305. 20030.