基于分布保持嵌入与正交映射的鲁棒生成式图像隐写方法

袁程胜[1,2]，王昊宇[1]，尹青伟[3]，曹燚[4]，刘庆程[1]，付章杰[1,2]
（1.南京信息工程大学计算机学院、网络空间安全学院，江苏 南京 210044；
2.南京信息工程大学数字取证教育部工程研究中心，江苏 南京 210044；3.江苏省教育厅教育宣传中心，江苏 南京 210036；
4.无锡学院物联网工程学院，江苏 无锡214105）

摘　要：在协同多智能体系统的隐蔽通信中，基于生成式模型的隐写技术能够直接在公开信道中合成载密图像，为安全传输指令或数据提供了新途径。然而，现有方法在抗检测性、鲁棒性等任务上难以实现有效平衡，制约了其在实际场景中的应用。为了解决这一问题，提出了一种基于分布保持嵌入与正交映射的生成式图像隐写方法。首先，设计了一种分布保持的信息嵌入机制，将秘密信息编码至采样的高斯潜向量中。其次，引入向量重构模块，生成满足标准正态分布的向量作为模型输入，以增强对信道干扰的鲁棒性及抗隐写分析能力。最后，接收方基于共享的密钥执行逆向操作以提取信息。实验结果表明，在不同隐写容量下，所提方法生成的载体图像均具有较好的视觉保真度，在遭受JPEG压缩、高斯噪声等常见信道攻击后，信息提取准确率仍保持在97%以上，表现出较强的鲁棒性。此外，面对常见的隐写分析检测器，误检率均保持在0.5左右，表现出优异的抗隐写分析性能。

关键词：分布保持；生成式隐写；正交映射；扩散模型
中图分类号：TP391
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2026063

Generative image steganography method based on density-preserving embedding and orthogonal mapping

Yuan Chengsheng[1,2], Wang Haoyu[1], Yin Qingwei[3], Cao Yi[4], Liu Qingcheng[1], Fu Zhangjie[1,2]
1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China
2. Engineering Research Center of Digital Forensics Ministry of Education, NUIST, Nanjing 210044, China
3. Education Publicity Center, Department of Education of Jiangsu Province, Nanjing 210036, China
4. School of Internet of Things Engineering, Wuxi University, Wuxi 214105, China

Abstract: In covert communication for collaborative multi-agent systems, generative model-based image steganography can directly synthesize stego-images over public channels, offering a novel approach for secure transmission of instructions and data. However, existing methods struggle to balance detectability resistance and robustness, limiting their practical application. To address this, a generative image steganography method based on density-preserving embedding and orthogonal mapping was proposed. Firstly, a distribution-preserving information embedding mechanism was designed to embed the secret information into the sampled Gaussian latent vector. Secondly, a vector reconstruction module was introduced to generate inputs that follow the standard normal distribution, thereby enhancing robustness against channel interference and resistance to steganalysis. Finally, the receiver performed inverse operations based on the shared key to extract the information. Experimental results show that, under different steganographic capacities, the cover images generated by the proposed method consistently achieve superior visual fidelity. After attacks such as JPEG compression and Gaussian noise, the information extraction accuracy remains above 97%, demonstrating strong robustness. Against common steganalysis detectors, the false positive rate is approximately 0.5, highlighting excellent detectability resistance.

Keywords: distribution-preserving, generative steganography, orthogonal mapping, diffusion model

收稿日期：2025–12–30；修回日期：2026–03–01
通信作者：尹青伟，yinqingwei@126.com
基金项目：国家自然科学基金资助项目（No.U22B2062, No.U23B2023, No.62372125）
Foundation Items：The National Natural Science Foundation of China (No.U22B2062, No.U23B2023, No.62372125)

0　引言

在多智能体协同系统中，隐蔽通信是确保协作安全与抗干扰能力的关键技术。在蜂群无人机、分布式传感网络等高动态、强交互的应用场景下，智能体需要通过开放信道可靠传递指令与共享数据，隐蔽通信能够有效应对恶意干扰、窃听与欺骗攻击，保障协同通信任务的安全执行。传统加密技术虽可保护信息内容的机密性，但密文传输本身易揭示通信行为的存在，难以抵御流量分析、元数据探测等深层威胁。相较之下，信息隐藏技术则提供了另一种安全范式，该技术通过将秘密信息嵌入图像[1]、文本[2]、音频[3]等常规载体，可实现通信行为本身的隐蔽化。

现有的隐写方法依据载体图像的生成方式，主要可分为两类[4]：基于修改式的图像隐写方法与基于生成式的图像隐写方法。基于修改式的图像隐写方法通过对原始载体图像进行特定修改以嵌入秘密信息，但此类操作常会留下修改痕迹，改变图像的统计特性，因而易被针对性隐写分析器检测到，存在一定的安全风险。尽管研究人员尝试通过弱化信息嵌入对图像统计分布的影响来提升安全性，但由于该方法本质依赖于对现有载体图像的修改，仍难以彻底避免统计异常的产生。相比之下，基于生成式的图像隐写方法[5-7]因其在抗隐写分析方面的突出优势，近年来受到广泛关注。这类方法基于生成式模型，通过构建从秘密信息到载体图像的直接映射，使发送方能够依据待隐藏信息直接合成通信载体，从而实现更安全和隐蔽的传输。

在各类信息载体中，图像因其传播便利、视觉隐蔽性强等特点，已成为生成式隐写技术中最常用的媒介。现有研究围绕提升生成质量与安全性已取得一系列进展。Wei 等[8]提出一种从秘密数据直接生成含密图像的方法，通过引入互信息约束增强秘密信息与生成图像间的相关性。Kim 等[9]率先将扩散模型引入图像隐写领域，指出模型输入需满足标准正态分布，能生成最优质量的图像。Yang 等[10]提出一种基于生成对抗网络的鲁棒图像隐写框架，在理论层面具备可证明的安全性。然而，当前生成式隐写技术在视觉质量、抗隐写分析能力与信道鲁棒性之间的平衡仍面临挑战。为此，本文提出一种高视觉质量、强鲁棒性的生成式图像隐写方法。该方法以稳定扩散模型为骨干网络，构建端到端的隐写框架，能够在保证信息准确传输的同时显著提升载体图像的视觉保真度，并增强其对常见信道干扰的抵抗能力。本文的主要贡献如下。

1) 设计了一种分布保持的秘密信息嵌入机制，使含密向量的分布与正常输入向量的分布一致。为了避免因分布偏移引发噪声预测误差累积进而降低载体图像的质量，本文方法在预采样的高斯噪声向量中嵌入秘密信息。其核心在于通过预先设计的编码方式进行信息嵌入，确保嵌入过程不改变向量的整体概率分布，从而有效抑制误差传播，维持生成图像的视觉质量与统计不可检测性。

2) 为抑制隐式扩散模型中因反演过程带来的数值不稳定性，本文设计了一种隐向量重构模块。该模块能够降低模型对输入噪声的敏感性，并确保输入的潜在向量满足标准正态分布要求，从而提升生成过程的稳定性和载体图像的统计可靠性。

3) 为验证所提方法的性能，本文在 MS-COCO 和 Flicker8K 两个数据集上进行了测试。实验结果表明，所提方法生成的载体图像在两种无参考图像空间质量评估 BRISQUE 和 NIQE 指标上均取得最优数值，具有更好的视觉保真度。在图像压缩、高斯噪声、裁切等多种常见信道干扰下，秘密信息的提取准确率大多保持在 97% 以上。此外，在应对常见隐写分析检测时，其误检率稳定在约 0.5，表现出较好的抗隐写分析性能。

1　相关研究工作

现有图像隐写方法主要可分为两类：基于修改式的图像隐写方法与基于生成式的图像隐写方法[11]，具体介绍如下。

1.1　基于修改式的图像隐写方法

在传统隐写方法中，秘密信息通常通过直接修改载体图像的像素值进行嵌入，如 Mielikainen 等[12]提出的最低有效位替换算法。这类方法对信道噪声较敏感，鲁棒性较弱。同时，由于修改会破坏图像固有的统计特性，使其易被隐写分析工具检测，因此抗隐写分析能力有限。为缓解这一问题，研究者提出了多种自适应嵌入策略，如校验子格编码[13]、信令点编码[14]等，旨在通过优化修改模式来降低信息嵌入对图像统计特性的影响，

从而提升隐蔽性。然而，由于基于修改式的隐写技术本质上依赖于对现有载体的改动，仍难以避免引入可被检测的统计异常[15]。为了提升隐写安全性，Sharp等[16]提出最低有效位匹配方法，通过在嵌入过程中随机增减像素值来降低信息嵌入对图像统计分布的影响。针对隐写嵌入导致载体图像统计特性变化这一核心问题，研究人员进一步提出了一系列自适应隐写框架，系统性地削弱信息嵌入引起的分布偏移，如Wani等[17]提出的利用深度神经网络学习自适应失真函数。然而，此类方法受限于其设计原理，难以保证载密图像与自然图像在统计分布上完全一致，无法实现可证明的安全性。特别是当嵌入容量较大时，这些方法会不可避免地引入视觉可察或统计可检的痕迹，安全性面临挑战。

为了提升隐写鲁棒性，研究人员利用空域与频域的可转换特性，通过修改频域系数实现信息隐藏，再经过逆变换重构载密图像。早期方法主要基于离散余弦变换、离散小波变换或离散傅里叶变换等，由于频域特征对常见噪声的敏感度通常低于空域，这类方法在鲁棒性方面表现更优。为进一步抑制噪声对信息提取的影响，Zhu等[18]提出了一种端到端的训练框架，在训练阶段，显式引入噪声层以增强模型鲁棒性；在提取阶段，则通过计算解码信息与原始信息之间的相似度，基于预设置信度阈值判断信息是否被准确恢复。

基于修改式的图像隐写方法通常使用人为定义的失真函数，以最小化整体失真代价为目标进行信息嵌入。在传统隐写方法中，通过精心设计失真函数，可在一定程度上保障隐藏信息的安全性与可提取性。但信息嵌入操作本质上会改变载体内容，导致其在面对日益复杂的隐写分析算法时，抗检测能力仍存在显著局限性。

1.2 基于生成式的图像隐写方法

基于生成式的图像隐写方法通过构建秘密信息至生成模型输入之间的双向映射关系，直接合成含密图像，不需要对现有载体内容进行任何修改，从源头上避免了因修改操作引入的统计特征异常，提升了隐写的安全性与抗检测能力。

为了降低修改式隐写方法存在的安全风险，研究人员尝试从编码设计与生成方式两方面进行改进。Peng等[19]提出一种基于自适应像素选择的编码方法，通过将图像属性映射为固定长度的哈希序列来对应二进制秘密信息，以减少可检测的修改痕迹。另一条技术路线则直接利用生成式模型的能力来避免对原始载体的修改。例如，Zhou等[20]通过将秘密信息嵌入生成网络的隐向量中，并训练相应的提取器以恢复信息。Holub等[21]提出了一种适用于任意域的通用隐写失真函数，通过利用嵌入修改对图像方向性滤波残差的相对影响来定义嵌入代价，为空间域等内容自适应隐写方法提供了统一的失真建模框架。

基于生成对抗网络及流模型的生成式隐写方法常因图像反演的过程不够稳定导致信息提取准确率受到限制。相比之下，扩散模型凭借其稳定、可控的生成机制，在隐写领域中展现出显著优势。Yang等[22]提出了一种基于图像域的信息隐写方法，通过逆采样过程建立从秘密信息到隐向量的可靠映射，实现高精度的信息恢复。Hu等[23]设计了一种基于稳定扩散模型的文本驱动式隐写方法，不需要对模型进行微调。该方法通过映射模块将二进制秘密信息转换为符合高斯分布的隐向量，并结合文本提示控制生成图像的语义内容，在保证采样效率的同时，显著提升了载体的视觉质量。

在生成式隐写框架中，发送方依据预设规则将秘密信息映射为生成模型所需的潜在向量，进而合成视觉质量较高的载密图像。然而，该过程中存在两方面关键挑战：一是若由秘密信息生成的潜向量与模型期望输入的概率分布不相符，将导致生成的载体图像出现质量下降；二是在实际传输中，载密图像常受到压缩、噪声等有损信道操作的影响，这些干扰会显著降低信息提取的准确性，制约方法的实际可用性。

2 相关基础理论

2.1 隐写安全性定义

从信息论角度出发，Cachin等[24]给出了隐写安全性的形式化定义，以图像的自然分布与含密图像的分布之间的相对熵作为安全性度量的依据。当两者分布一致时，表明该隐写方法满足可证明安全性。相对熵的计算式为

$$D(P_e\|P_s)=\sum_{x\in C}P_e(x)\log\frac{P_e(x)}{P_s(x)} \tag{1}$$

其中，P_e指正常图像集合的分布，P_s指载体图像的分布。当相对熵的数值为0时，可认为隐写系统是完全安全的。

对于图像隐写术来说，可证明安全的隐写算法可以分为以下三类。

1) 基于拒绝采样的方法。该方法的核心是使用映射函数采样，生成载体图像。为了得到隐藏有秘密信息的图像，发送方与接收方共享从集合$\mathcal{K}$中获取密钥k，结合预先训练的采样器从采样空间$\mathcal{R}$中进行采样，直到采样结果完全满足映射函数，将其作为秘密信息的载体，以保证隐写方法的安全性。尽管基于拒绝采样的方法能够实现可证明安全，但其嵌入容量较低。

2) 基于逆采样的方法。基于逆采样的算法通过设计可逆映射函数，将秘密信息转换为服从特定目标分布的数据，并在接收端使用逆变换从数据中恢复原始信息。在目标数据分布已知的情况下，可通过设计映射函数，使信息载体与正常媒介分布一致，实现可证明安全。然而，传统载体的真实数据分布通常难以准确建模。

3)基于生成式模型的隐写方法。生成模型能够将固定先验分布（如高斯分布）映射为多样化的载体数据，从而有效解决分布建模的难题。在此框架下，载体图像与正常图像分别由含密向量和普通潜在向量通过模型生成。为满足载体图像的不可区分性，要求二者的输入分布保持一致，即相对熵为零。此外，得益于生成模型强大的合成能力，由其生成的图像具有高度的视觉自然性，在通信过程中不易引起攻击者的警觉，因而更贴合实际隐蔽通信场景对隐匿性的要求。

2.2　去噪扩散隐式模型

稳定扩散模型的工作机制主要包含前向扩散与反向生成两个过程。在前向扩散过程中，模型通过逐步向原始数据添加高斯噪声，使其逐渐转化为近似纯高斯噪声的向量；在反向生成过程中，模型通过迭代去噪从噪声分布中重建出原始数据。图1展示了该模型生成载体图像的完整流程，其中，第1行、第3行和第5行表示从隐写图像中恢复潜在向量的增量加噪过程，第2行、第4行和第6行表示从潜在向量生成隐写图像的渐进去噪过程。生成式隐写方法正是依托于该模型的可逆结构，实现秘密信息在潜空间中的嵌入与可靠提取。

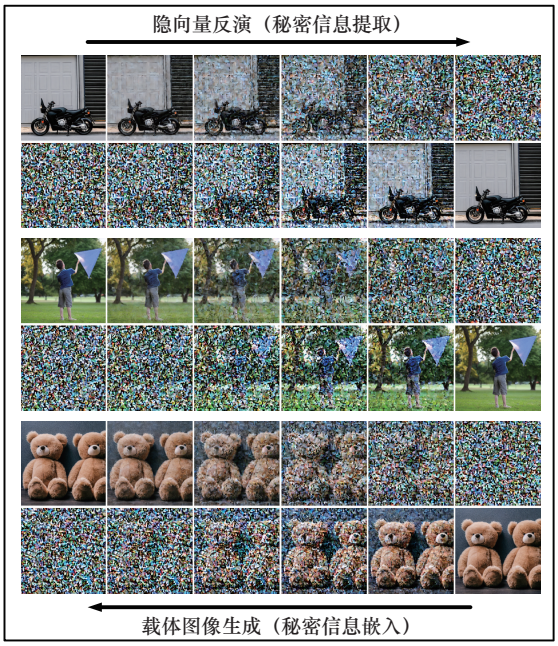

图1　载体图像生成与秘密信息提取示例

在去噪扩散概率模型中，前向扩散过程基于图像的边缘分布$q(x_t|x_0)$，去噪扩散隐式模型则在此基础上进一步改进，将扩散过程重构为非马尔可夫链过程，降低了计算复杂度，并且确保反演操作的可逆性。在模型迭代过程中，每一时间步的向量不仅依赖于前一步，还受到初始潜在向量的影响。其概率分布的计算式为

$$q_\sigma(x_{1:T}|x_0)=q_\sigma(x_T|x_0)\prod_{t=2}^{T}q_\sigma(x_{t-1}|x_t,x_0) \tag{2}$$

其中，$\boldsymbol{X_0}$为初始输入，$\boldsymbol{X_t}$表示扩散过程进行到第t步时的向量，参数σ为扩散过程中的随机因子，q_σ代表扩散过程中由σ控制的前向扩散转移概率。在前向过程中，第t步的含噪向量与初始向量满足如式(3)所示的关系。

$$q_\sigma(x_T|x_0)=\mathcal{N}(x_T;\sqrt{\bar{\alpha}_T}x_0,(1-\bar{\alpha}_T)I) \tag{3}$$

其中，I表示单位矩阵。当$t>1$时，含噪向量与初始向量的具体关系如式(4)所示。

$$q_\sigma(x_{t-1}|x_t,x_0)=\mathcal{N}(x_{t-1};\sqrt{\bar{\alpha}_{t-1}}x_0+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\frac{x_t-\sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}},\sigma_t^2I) \tag{4}$$

其中，$\bar{\alpha}_t$是关于时间步t的超参数。

前向扩散过程的噪声添加可形式化描述为式(5)所示的贝叶斯递推过程。

$$q_\sigma(x_t|x_{t-1},x_0)=\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)} \tag{5}$$

在训练阶段，当初始潜在向量和随机生成的噪声ε_t满足标准正态分布时，由前向扩散过程的性质可进一步推导初始向量出任意时间步向量的表达式，即：

$$x_t=\sqrt{\bar{\alpha}_t}\,x_0+\sqrt{1-\bar{\alpha}_t}\,\epsilon_t \tag{6}$$

经过训练后，神经网络可以预测每一步的噪声ε_t，表示为$\varepsilon_t(x_t,t)$。由ε_t引导每一步的去噪过程，最终反演得到原始干净向量的预测值，计算式为：

$$x'_0=\frac{x_t-\sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}} \tag{7}$$

将预测出的向量代入式(2)中，得到反演过程中每个时间步的含噪向量预测值，如式（8）所示：

$$p_\theta(x_{t-1}|x_t)\approx q(x_{t-1}|x_t,x_0=x'_0)=\mathcal{N}\left(x_{t-1};\sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t-\sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\epsilon_\theta(x_t,t),\sigma_t^2 I\right) \tag{8}$$

根据式(8)推导出从x_t采样出x_{t-1}的公式

$$x_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t-\sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\varepsilon_\theta(x_t,t)+\sigma_t\varepsilon_t \tag{9}$$

其中，ε_t是服从均值为0、方差为1的正态分布的随机噪声。

本文提出一种面向盲传输信道的鲁棒生成式图像隐写方法，以潜在向量为秘密信息的载体，将秘密信息嵌入作为模型输入的潜在向量中，利用扩散模型的反演能力恢复潜在向量，从而提取秘密信息。

3 本文方法

3.1 扩散模型中的噪声偏移

去噪扩散隐式模型可在初始向量经过 t 步迭代后的隐向量之间建立可逆的双射映射关系。然而，受模型内在数值不稳定性及扰动敏感性的影响，其在反演过程中容易引入迭代误差。即使秘密信息嵌入仅对初始向量产生微小分布偏移，该误差也会随着迭代步长的增加而逐渐积累，最终被显著放大。初始向量与去噪过程完成后的向量关系如式(10)所示。

$$\mu_0(x_T,T)=\frac{1}{\sqrt{\alpha_T}}\left(x_T-\frac{\beta_T}{\sqrt{1-\bar{\alpha}_T}}\epsilon_\theta(x_T,T)\right) \tag{10}$$

其中，x_T为去噪过程中的初始噪声，服从标准正态分布，即$x_T=\epsilon_T,\epsilon_T\sim\mathcal{N}(0,I)$；$T$为扩散模型的总迭代步数，$\epsilon_\theta(x_T,T)$为模型预测的噪声，用于恢复向量；$\alpha_T$表示模型去噪过程中的缩放因子。若初始的噪声偏离标准正态分布，即$x_t=\delta+x_T$，此时模型对噪声的预测值为

$$\mu_0(x_T,T)=\frac{1}{\sqrt{\alpha_T}}\left(\delta+x_T-\frac{\beta_T}{\sqrt{1-\bar{\alpha}_T}}\epsilon_\theta(x_T,T)\right) \tag{11}$$

由此可知，初始输入的微小分布偏移会在去噪迭代过程中被模型中的非线性项逐级放大，通过误差累积效应，最终导致输出分布严重偏离预期。这种分布偏移最终在视觉上表现为生成图像的高频细节模糊、语义结构扭曲或出现伪影等问题，严重影响了载体图像的视觉质量与隐蔽性。

3.2 算法框架

本文提出的基于分布保持嵌入与正交映射的算法框架如图2所示。该算法结合含密噪声生成模块与预训练的扩散模型，将秘密信息转换为高质量的载密图像，在保证高安全性与强鲁棒性的同时，显著降低了图像生成的计算复杂度。整个流程可概括为3个阶段：含密噪声构造阶段、载体图像生成阶段和秘密信息提取阶段。在含密噪声构造阶段，发送方结合二进制秘密信息生成高斯潜向量；在载体图像生成阶段，将含密潜向量输入模型，借助文本提示词作为辅助控制信号，将秘密信息转化为载体图像；在信息提取阶段，接收方依据共享密钥与可逆运算，准确恢复原始信息。

3.3 信息嵌入与提取方法

现有方法在将秘密信息映射为潜在向量时，生成的向量往往偏离标准高斯分布，进而导致载密图像质量下降。为提高视觉质量并增强抗隐写分析能力，本文设计了一种含密噪声生成方法，确保生成的高斯向量同时满足随机性、独立性及标准正态分布3个条件，从而改善生成图像的视觉保真度。假设秘密信息服从均匀二项分布，其表示为

$$M=\left\{m_1,m_2,m_3,\cdots,m_j\right\},M\in\{0,1\}^L \tag{12}$$

其中，m_i表示每一位秘密信息，L表示秘密信息的总长度。

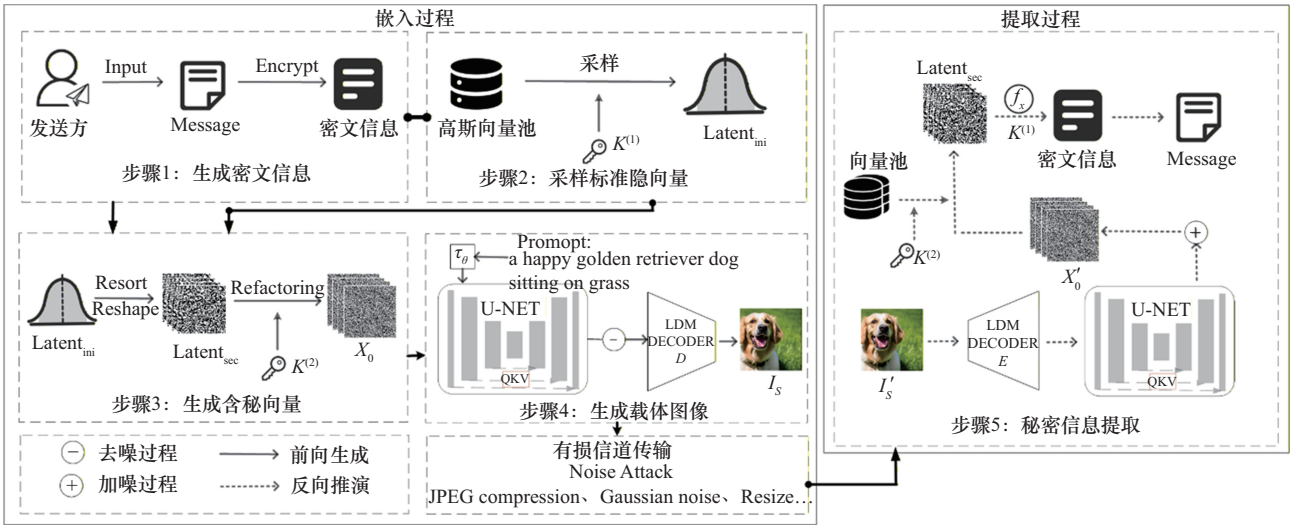

图2　本文提出的基于分布保持嵌入与正交映射的算法框架

为确保嵌入过程使用的高斯向量具备独立随机性，首先使用采样函数通过密钥K_1从高斯向量池中采样，得到初始随机向量$\text{Latent}_{\text{ini}}$，如式(13)所示。

$$\text{Latent}_{\text{ini}} = \text{Rand}\left(K_1\right) \tag{13}$$

其中，Rand (·)是高斯随机采样函数，发送方和接收方使用相同的密钥K_1生成同样的高斯向量。

随后，在密文信息的引导下，对向量元素位置进行更改，根据需要嵌入的秘密比特该系数对中两个元素的位置是否交换。由于反演误差与噪声干扰，用于存储秘密信息的系数越接近，对数值变化越敏感。与之相对地，系数对的差值较大时，其区分性较强，能有效降低信息提取的错误率。因此，通过选取差值较大的系数对进行信息嵌入，能够提高算法的鲁棒性。

本文通过设置阈值参数d（表示被允许用于存储信息的系数对中两系数的最小距离），以保证其鲁棒性。发送方使用密钥K_1采样初始随机高斯向量$\text{Latent}_{\text{ini}}$后，首先搜索满足条件的系数对用于嵌入信息，其余不满足条件的元素不作为载体单元使用。通过预先共享的数对匹配方法和共享的信息长度参数L，执行秘密信息的嵌入。

嵌入信息的数对的匹配方法如下：首先将初始高斯向量按数值升序排序，随后采用贪心策略进行系数对选择：每一步从当前最小的未使用元素出发，在其右侧寻找第一个同样未使用且与其差值大于预设阈值d的元素作为配对者；配对成功，则将该两元素标记为已使用，并将该系数对作为一个载体单元。重复上述过程，直到不存在新的满足条件的未使用元素对为止。对于得到的每一个系数对，根据秘密比特决定是否交换元素位置，从而实现每个系数对存储1位密文信息，嵌入方式如式(14)所示。

$$P\left(s \middle| m_i\right) = \begin{cases} (s_i, s_j), & m_i = 0 \\ \text{swap}(s_i, s_j), & m_i = 1 \end{cases} \tag{14}$$

其中，S_i表示向量$\boldsymbol{Latent}_{\text{sec}}$中的各个元素。

为提升方案对信道噪声的鲁棒性，引入正交映射模块对嵌入后的含密向量进行重构。该模块基于共享密钥K_2采样生成随机矩阵，并通过施密特正交化构造相应的正交矩阵$\boldsymbol{M}_{\text{init}}$，利用式(15)对向量$\boldsymbol{Latent}_{\text{sec}}$进行变换，得到满足模型输入要求的潜在向量。最终，结合文本提示词经由生成模型合成视觉质量良好的载密图像。

$$\boldsymbol{X}_0 = \boldsymbol{Latent}_{\text{sec}} \cdot \boldsymbol{M}_{\text{init}} \tag{15}$$

由于$\boldsymbol{Latent}_{\text{ini}}$由密钥从高斯向量池中采样获得，其本身服从标准正态分布，而嵌入阶段仅根据秘密比特对选定系数对的两个坐标进行交换，不改变向量元素取值集合，该操作本质上是对初始高斯样本的坐标置换，没有改变分布。在后续向量重构阶段，由于矩阵$\boldsymbol{M}_{\text{init}}$的正交性，经过正交映射模块后生成的潜向量仍服从标准正态分布，与扩散模型输入先验一致，该方法能够使分布保持不变。

载体图像生成与秘密信息提取的伪代码如算法1所示。在信息提取阶段，接收方利用与发送方相同的采样器及预训练扩散模型，并基于共享的参数，包括采样密钥及信息长度L，执行逆向恢复，最终还原出秘密信息。

算法1 载体图像生成与秘密信息提取

输入 预训练扩散模型G_{dif}，确定性采样器$D(\cdot)$，秘密信息嵌入模块$F_1(\cdot)$，向量重构模块$F_2(\cdot)$，秘密信息m，密钥K_1与K_2

输出 载体图像I_s，秘密信息m'

1)使用密钥K_1采样高斯向量:$\boldsymbol{Latent}_{ini} \leftarrow D(K_2)$

2)嵌入秘密信息：$\boldsymbol{Latent}_{sec} \leftarrow F_1(\boldsymbol{Latent}_{ini})$

3)使用密钥K_2生成随机矩阵：$\boldsymbol{M} \leftarrow D(K_2)$

4)对向量执行正交化处理：$\boldsymbol{M}_{init} \leftarrow F(\boldsymbol{M})$

5)生成含密向量：$\boldsymbol{X}_0 \leftarrow F_2(\boldsymbol{Latent}_{sec},\boldsymbol{M}_{init})$

6)由预训练模型生成载体图像：$I_s \leftarrow G_{dif}(\boldsymbol{X}_0)$

7)接收方逆向提取含密向量：$\boldsymbol{Latent}'_{sec} \leftarrow G_{dif}^{-1}(I_s')$

8)逆向提取秘密信息：$m' \leftarrow F^{-1}(\boldsymbol{Latent}',K)$

通过调节隐写参数，即设置不同的距离阈值，能够调节本文算法的隐写容量。需要指出的是，即使选择相同的阈值d，随机采样得到的高斯向量中满足条件的数值对的数量也不会相同。因此，对于给定阈值d，本文以多个独立密钥生成的合格数对平均数量，作为该阈值对应的平均隐藏容量。在实际应用中，通过选取合适的隐写参数以实现容量和鲁棒性的平衡。

4 实验结果与分析

为分析不同采样器对提取精度的影响以确定最优采样器，本文在MS-COCO数据集上进行了消融实验。实验中保持其他配置不变，仅替换采样器，并根据秘密信息提取精度对各采样器的性能进行对比评估，实验结果如表1所示。实验表明，高阶采样器的数据预测机制能更有效地抵抗数据失真，在数值精度与稳定性方面表现更优，因此提取准确率更高。此外，使用高阶采样器生成的载密图像在视觉质量方面也优于其他采样器。因此，本文选用DPM-Solver作为最终的采样器。

表1 采用不同采样器时提取秘密信息的准确率

攻击方法	DDIM	PNDM	DPM-Solver
Lossless	99.12%	99.11%	99.16%
AWGN 0.01	96.53%	97.34%	98.17%
JPEG90	97.88%	97.67%	98.30%
MBlur3×3	96.04%	96.84%	97.78%
Gblur3×3	94.10%	94.42%	96.01%

4.1 数据集介绍

为评估本文方法的性能，本节在MS-COCO和Flicker8K两个数据集上进行相关实验测试。两个数据集的基本信息如下。

1) MS-COCO数据集。该数据集包含5000幅图像样本，每幅图像均配有对应的文本描述。实验中，随机抽取文本提示词生成载体图像。

2) Flicker8K数据集。该数据集包含8092幅图像样本，每幅图像均附有文本描述。实验中，随机抽取提示词生成载体图像。

4.2 性能评价指标

1) 隐藏容量。隐藏容量采用每像素比特数进行量化，定义为载体图像中每个像素平均能嵌入的秘密信息比特数，具体计算式如式(16)所示。

$$\mathrm{bpp} = \frac{L}{H \times W} \tag{16}$$

其中，L表示秘密消息的最大长度，H和W分别表示图像的高度和宽度。

2) 提取准确率。采用正确率（Acc）作为秘密信息提取准确性的量化指标。通过逐比特比对原始秘密信息与提取出的恢复信息，统计二者完全一致的比特数占总比特数的比例，具体计算式如式(17)所示。

$$\mathrm{Acc} = \frac{F(\oplus m)m'_L}{L} \times 100\% \tag{17}$$

3) 鲁棒性。本实验旨在评估所提方法在含噪信道中维持信息提取准确性的能力。为验证所提隐写方法在有损传输环境中的鲁棒性，对含密图像施加了如下典型噪声干扰。

① 缩放（Resize）。该操作通过改变数字图像的分辨率评估鲁棒性，实验中分别将图像按0.75、1.25和1.50的尺度比例进行尺寸调整。

② JPEG压缩。该操作旨在评估不同程度图像退化对隐写性能的影响，实验分别在质量因子Q=90和Q=70的条件下进行测试。

③ 中值滤波（Median Blur）。该操作通过用相邻像素的中值代替每个像素的值来平滑图像，带来噪声干扰。

④ 高斯模糊（Gaussian Blur）。该操作使用不同强度的高斯核对图像局部区域进行处理。

⑤ 高斯噪声（Gaussian Noise）：该操作通过向图像中随机添加服从高斯分布的噪声实现，噪声

标准差的大小表示噪声的强度。

4) 抗检测性。隐写方法的抗隐写分析能力指其面对隐写分析检测器时能够避免被正确识别的性能。为评估该性能，本文选用检测错误率作为评价指标，其计算式如式(18)所示。

$$P_E = \frac{P_{\mathrm{FA}} + P_{\mathrm{MD}}}{2} \tag{18}$$

其中，P_{FA}为虚警率，P_{MD}为漏检率。当误检率接近0.5时，说明检测器不能有效区分载体图像与正常图像。

5) 视觉质量。在生成式图像隐写中，由于缺乏原始载体图像作为参考基准，本文采用两种无参考图像质量评价指标BRISQUE和NIQE对生成图像的视觉质量进行度量。

4.3　实验结果

4.3.1　载体图像质量

在固定相同秘密信息的条件下，本文通过输入不同的文本提示词生成多组载密图像，并对这些生成图像的视觉质量进行系统评估，部分生成示例如图3所示。

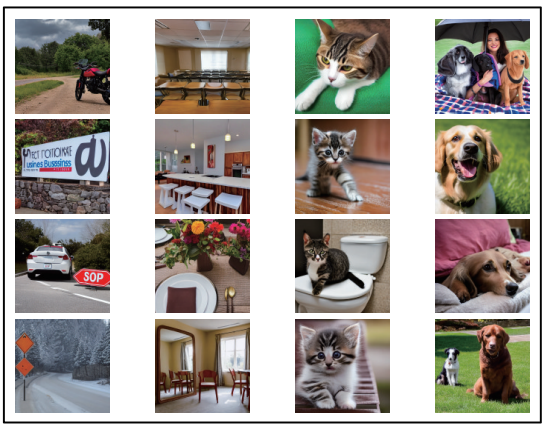

图3　不同隐写方法生成的载体图像示例

在视觉质量评估方面，由于生成式隐写方法不需要原始载体图像，本文采用无参考图像质量评价指标BRISQUE和NIQE对生成图像的真实感进行量化评估。实验在MS-COCO数据集上进行。BRISQUE与NIQE的得分越低，表明生成图像的视觉质量越高，越接近自然图像分布。

实验结果如表2所示。由表2可知，本文方法在BRISQUE和NIQE两项质量评价指标上均取得最低数值，即所生成的载密图像在视觉质量上优于对比的方法，具有更好的视觉保真度。此外，本文方法不需要对预训练生成模型进行微调，有效保留了其原有的生成能力。同时，通过引入文本提示作为控制信号，实现了对生成内容的语义引导，从而将秘密信息自然地嵌入多样化的图像内容中，进一步增强了隐写的隐蔽性。

表2　不同隐写方法生成的载体图像质量的对比

指标	S2IRT[25]	GRDH[23]	LAGDE[26]	本文
BRISQUE	19.17	18.21	17.69	**17.27**
NIQE	4.21	4.03	3.91	**3.87**

4.3.2　提取准确率与鲁棒性

为评估算法的鲁棒性，本文在MS-COCO数据集上与多个基线算法在不同攻击场景下进行了对比。不同隐写算法在多种图像攻击下的秘密信息提取准确率如表3所示。

实验结果表明，随着攻击强度增加，隐写算法的提取精度均呈现逐渐下降的趋势。在无损传输条件下，本文方法的秘密信息提取准确率高于所有对比方法。在大多数常见后处理攻击下，本文方法仅表现出轻微的性能下降。尤其对于传输过程中常见的JPEG压缩、裁剪等失真操作，本文方法的性能几乎不受影响，体现了良好的实用性与稳定性。此外，在绝大多数攻击场景下，本文方法的提取准确率仍能保持在97%以上，显示出较强的抗干扰鲁棒性。

4.3.3　隐写容量的可调性

为评估不同隐写容量下的提取准确率，验证所提方法的性能一致性，本节对不同容量设置下的载体图像质量进行了对比分析。

表4展示了不同隐写容量下，本文方法在MS-COCO数据集上生成载体图像的视觉质量对比情况。表4数据表明，随着隐写容量参数的变化，视觉质量指标保持稳定且与数据集中真实图像接近，表明载体视觉质量基本不受本文隐写容量设定的影响。

表5给出了本文方法在MS-COCO数据集上的鲁棒性测试结果。实验表明，本文方法在4种不同的隐写容量配置下均表现优异，即便遭受多种干扰攻击，在绝大多数测试场景中仍能稳定维持在97%以上的准确率，充分验证了该方案的强鲁棒性。

表3 不同隐写方法在有损信道传输下的鲁棒性测试

攻击方法	失真参数	GRDH[23] (bpp=0.0625)	S2IRT[25] (bpp=0.0625)	LAGDE[26] (bpp=0.0625)	SNAD[27] (bpp=0.0625)	GSAGSC[28] (bpp=0.0234)	本文 (bpp=0.025)
Lossless	—	98.35%	81.67%	—	98.97%	98.39%	**99.16%**
Resize	0.75	97.41%	79.68%	—	97.81%	—	**98.29%**
	1.25	97.77%	81.02%	—	98.33%	—	**99.03%**
	1.5	97.94%	82.31%	—	98.94%	—	**99.14%**
JPEG压缩	90	96.73%	79.16%	**98.37%**	97.85%	—	98.30%
	70	93.53%	76.63%	95.42%	95.41%	—	**95.82%**
Gaussian Noise	0.001	97.82%	81.24%	98.87%	98.96%	—	**99.13%**
	0.01	97.56%	80.35%	98.32%	**98.72%**	—	98.17%
Gaussian Blur	3×3	97.42%	80.22%	97.72%	97.52%	—	**98.12%**
	5×5	95.81%	76.64%	93.25%	94.02%	—	**97.14%**
	7×7	92.99%	73.21%	88.03%	91.27%	—	**94.21%**
Median Blur	3×3	95.82%	78.37%	—	—	—	**97.21%**
	5×5	90.13%	75.11%	—	—	—	**92.31%**
	7×7	82.31%	72.51%	—	—	—	**85.72%**

表4 不同隐写参数下的视觉质量比较

d	容量/bit	MS-COCO	Flicker8K
1.65	6 553	17.27	18.47
1.75	6 144	17.25	18.32
1.85	5 734	17.23	18.38
real		14.03	16.21

表5 不同隐写容量下的鲁棒性

攻击方法	参数	*d*=1.65	*d*=1.75	*d*=1.85	*d*=1.95
Lossless	—	99.16%	99.37%	99.70%	99.73%
Resize	0.75	98.59%	99.07%	99.41%	99.49%
JPEG压缩	90	98.30%	98.89%	99.17%	99.21%
Gaussian Blur	3	97.78%	98.39%	98.79%	98.93%
Median Blur	3	96.01%	97.00%	97.62%	97.79%
Gaussian Noise	0.01	98.17%	98.79%	99.12%	99.23%

4.3.4 安全性分析

在抗隐写分析性能评估中，本节在包括SRNet、XuNet在内的主流隐写分析工具上对各方法进行了对比测试，以评估其抗检测能力。

得益于所设计的分布保持嵌入机制，含密向量的统计特性与正常向量高度一致，从而显著增强了方法的隐蔽性。实验结果如表6所示。由表6可知，本文方法在多种主流检测器下的误检率均稳定在0.500附近，表明检测器难以有效区分正常图像与载体图像，验证了其优越的抗隐写分析性能。

表6 抗隐写分析性能比较

数据集	方法	SRNet	XuNet	YeNet
MS-COCO	GRDH[23]	0.507	0.501	0.506
	LAGDE[26]	0.502	0.503	0.501
	SNAD[27]	0.495	0.492	0.498
	本文方法	0.502	0.501	0.510
Flicker8K	GRDH[23]	0.502	0.503	0.512
	LAGDE[26]	0.500	0.501	0.500
	SNAD[27]	0.497	0.500	0.495
	本文方法	0.501	0.501	0.509

4.3.5 潜在向量分布

为验证本文方法在分布保持方面的性能，本节对生成的含密噪声向量进行了均值与标准差的统计分析。含密向量的分布与理论高斯分布越接近，表明其统计特性越稳定，生成的载体图像视觉质量也相应越高。

实验结果如表7所示，其中各容量下的均值与标准差均与标准正态分布高度吻合，验证了所提机制在分布保持上的优越性。实验结果表明，本文方法生成的隐写向量均值在10^{-4}与10^{-3}量级内波动，

其标准差与标准正态分布向量的差异约为10^{-5}，几乎可忽略。上述结果说明，所生成潜向量在均值和标准差上与标准正态分布高度相近，能够更稳定地适配扩散模型的生成过程，从而合成出视觉质量更高的载密图像。

表7　　潜在向量的均值与标准差

组别	均值	标准差
1	2.1×10^{-4}	$1+9.7\times10^{-6}$
2	-2.4×10^{-3}	$1-1.1\times10^{-5}$
3	3.6×10^{-4}	$1-1.4\times10^{-4}$
4	1.2×10^{-3}	$1+3.6\times10^{-4}$
5	1.5×10^{-3}	$1-6.1\times10^{-5}$

4.3.6　消融实验

为评估向量重构模块对信息提取准确率的提升效果，本文在MS-COCO数据集上进行了消融实验，对比了算法在包含与不包含重构模块时在多种噪声攻击下的提取性能，实验结果如图4所示。

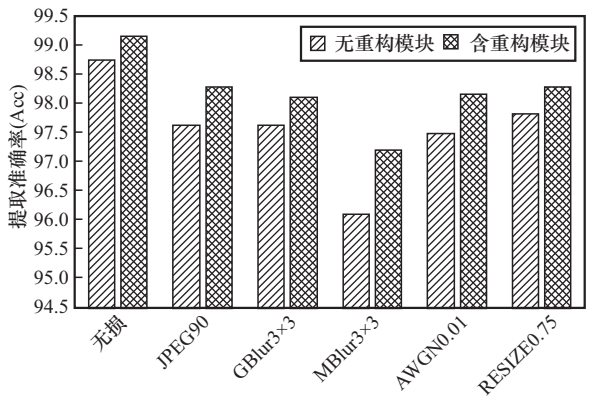

图4　消融实验

实验数据表明，在各类噪声攻击下，无重构模块的算法信息提取准确率偏低。这表明本文设计的映射机制虽可部分抑制误差传递，但在较强噪声干扰下，迭代过程中仍会累积数值误差，最终导致提取失败。相比之下，引入的向量重构模块能将施加于局部数值的扰动分散至整个向量空间，从而缓解噪声对信息提取过程的影响。该模块在保持隐写容量不变的前提下，显著提升了系统的提取稳定性与准确率。

5　结束语

针对现有生成式图像隐写方法在载体图像视觉质量、抗干扰鲁棒性方面存在的不足，本文提出一种基于分布保持嵌入与正交映射的图像隐写方法。该方法通过双系数嵌入机制将秘密信息高效编码至潜空间，并结合基于向量重构模块，生成服从标准正态分布的向量作为扩散模型的输入，显著提升了载密图像的视觉质量。实验结果表明，本文方法在JPEG压缩、滤波及噪声等多种有损信道条件下均保持较好的提取准确率。此外，所生成的载密图像在视觉质量与统计特性上与自然图像高度一致，且能有效抵抗主流隐写分析攻击。

然而，受限于当前的嵌入机制，隐写容量仍有提升空间。未来的研究工作可从以下两方面展开：一是在保持生成质量的前提下，提升方法对复杂真实场景中多模态复合攻击的鲁棒性；二是针对参数规模持续增长的生成式基础模型，系统探索隐写算法在计算效率、隐蔽容量与抗检测性之间的平衡机制，以推动生成式隐写技术向实用化与规模化方向发展。

参考文献：

[1] Zhou X J, Peng W L, Yang B Y, et al. Linguistic steganography based on adaptive probability distribution[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(5): 2982-2997.

[2] Deng Y X, Tan Y, Mao L, et al. A robust generative coverless co-steganography method of text and image[J]. International Journal of Autonomous and Adaptive Communications Systems, 2025, 18(6): 485-508.

[3] Gopalan K. Audio steganography using bit modification[C]//Proceedings of the 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings. Piscataway: IEEE Press, 2003: I-629.

[4] 周志立, 丁淳, 李进, 等. 生成式隐写研究[J]. 计算机学报, 2023, 46(9): 1855-1887.
Zhou Z L, Ding C, Li J, et al. Research on generative steganography[J]. Chinese Journal of Computers, 2023, 46(9): 1855-1887.

[5] Duan X T, Song H X, Qin C, et al. Coverless steganography for digital images based on a generative model[J]. Computers, Materials & Continua, 2018, 55(3): 483-493.

[6] Lee W K, Ong S, Wong K, et al. A novel coverless information hiding technique using pattern image synthesis[C]//Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway: IEEE Press, 2018: 1122-1127.

[7] Cao Y, Zhou Z L, Wu Q M J, et al. Coverless information hiding based on the generation of anime characters[J]. EURASIP Journal on Image and Video Processing, 2020, 2020: 36.

[8] Wei P, Li S, Zhang X P, et al. Generative steganography network[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 1621-1629.

[9] Kim D, Shin C, Choi J, et al. Diffusion-Stego: Training-free diffusion generative steganography via message projection[J]. Information Sciences, 2025, 718: 122358.

[10] Yang Z J, Chen K J, Zeng K, et al. Provably secure robust image steganography[J]. IEEE Transactions on Multimedia, 2024, 26: 5040-5053.
[11] 吴槟, 薛瑞. 基于深度学习特征分布优化的无载体图像隐写方法[J]. 信息安全学报, 2024: doi.10.19363/J.cnki.cn10-1380/tn.2024.04.05.
Wu B, Xue R. A coverless image steganography method using deep-learning with feature distribution optimization [J]. Journal of Cyber Security, 2024: doi.10.19363/J.cnki.cn10-1380/tn.2024.04.05.
[12] Mielikainen J. LSB matching revisited[J]. IEEE Signal Processing Letters, 2006, 13(5): 285-287.
[13] Filler T, Judas J, Fridrich J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935.
[14] Li W X, Zhang W M, Li L, et al. Designing near-optimal steganographic codes in practice based on polar codes[J]. IEEE Transactions on Communications, 2020, 68(7): 3948-3962.
[15] Sharifzadeh M, Aloraini M, Schonfeld D. Adaptive batch size image merging steganography and quantized Gaussian image steganography[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 867-879.
[16] Sharp T. An implementation of key-based digital signal steganography [C]//Information Hiding. Berlin, Springer, 2001: 13-26.
[17] Wani M A, Sultan B. Deep learning based image steganography: a review[J]. WIREs Data Mining and Knowledge Discovery, 2023, 13(3): e1481.
[18] Zhu J R, Kaplan R, Johnson J, et al. HiDDeN: hiding data with deep networks[C]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 682-697.
[19] Peng Y Y, Hu D H, Wang Y F, et al. StegaDDPM: generative image steganography based on denoising diffusion probabilistic model[C]// Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 7143-7151.
[20] Zhou Z L, Sun H Y, Harit R, et al. Coverless image steganography without embedding[C]//Cloud Computing and Security. Cham: Springer International Publishing, 2015: 123-132.
[21] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, 2014: 1.
[22] Yang X, Wu T Y, Huang F J. Reversible data hiding in JPEG images based on coefficient-first selection[J]. Signal Processing, 2022, 200: 108639.
[23] Hu X X, Li S, Ying Q C, et al. Establishing robust generative image steganography via popular stable diffusion[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 8094-8108.
[24] Cachin C. An information-theoretic model for steganography[J]. Information and Computation, 2004, 192(1): 41-56.
[25] Zhou Z L, Su Y C, Li J, et al. Secret-to-image reversible transformation for generative steganography[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(5): 4118-4134.
[26] Xu Y H, Sun W, Tang C P, et al. Security-robustness trade-offs in diffusion steganography: a comparative analysis of pixel-space and vae-based architectures[PP]. V2. (2025-10-08) [2025-12-30]. arXiv: arXiv.2510.07219.
[27] Zhang X, Song T H, Peng F, et al. Denoising diffusion probabilistic steganography based on standardized secret noise[J]. IEEE Signal Processing Letters, 2025, 32: 4124-4128.
[28] Zhou Z L, Dong X H, Meng R H, et al. Generative steganography via auto-generation of semantic object contours[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2751-2765.

[作者简介]

袁程胜（1989– ），男，江苏东海人，南京信息工程大学副教授，主要研究方向为人工智能安全、信息隐藏。

王昊宇（2002– ），男，江苏扬州人，南京信息工程大学硕士生，主要研究方向为信息隐藏。

尹青伟（1989– ），男，江苏淮安人，江苏省教育厅教育宣传中心中级工程师，主要研究方向为网络态势感知、网络空间治理等。

曹燚（1994– ），男，江苏宿迁人，无锡学院副教授，主要研究方向为信息隐藏。

刘庆程（1987– ），男，安徽滁州人，南京信息工程大学博士生，主要研究方向为多媒体内容安全、电力大数据挖掘与分析。

付章杰（1983– ），男，河南南阳人，南京信息工程大学教授，主要研究方向为信息安全。