重新思考基于扩散的生成式隐写术的安全性

朱家豪，陈子轩中山大学计算机科学与工程学院

刘佳丽中山大学中山医学院

杨凌霄中山大学系统科学与工程学院

周毅
中山医学院
中山大学

Weiqi Luo、谢晓华中山大学计算机科学与工程学院

arXiv:2602.10219v1 [cs.MM] 10 Feb 2026

摘要

生成式图像隐写术是一种将秘密消息隐藏在生成图像中的技术，无需依赖预先存在的载体图像。近年来，人们提出了多种基于扩散模型的生成式图像隐写术（DM-GIS）方法，这些方法能够有效对抗传统隐写分析技术。本文指出了影响 DM-GIS 安全性的关键因素，并重新审视了现有方法的安全性。具体而言，我们首先概述了当前 DM-GIS 方法的通用流程，并发现扩散模型的噪声空间是主要的嵌入域。此外，我们分析了 DM-GIS 安全性与扩散模型噪声分布之间的关系，从理论上证明，任何会扰乱噪声分布的隐写操作都会损害 DM-GIS 的安全性。基于这一发现，我们提出了一种基于噪声空间的扩散隐写分析器（NS-DSer）——一个简单而有效的隐写分析框架，可在扩散模型噪声空间中检测由 DM-GIS 生成的图像。我们使用 NS-DSer，在难度逐步增加的检测场景中重新评估了现有 DM-GIS 方法的安全性。实验结果验证了我们对 DM-GIS 安全性的理论分析，并表明 NS-DSer 在各种检测场景中均有效。

关键词 生成式图像隐写术 · 隐写分析 · 扩散模型

1 引言

图像隐写将秘密消息隐藏在看似无害的图像中，以实现隐蔽通信 [1, 2]。传统隐写方法将秘密比特嵌入自然图像的空间域 [3, 4, 5, 6, 7] 或频域 [8, 9, 10] 中（又称 载体图像）。然而，这一过程不可避免地会改变载体图像，使其容易被检测。此外，这些算法的嵌入容量有限。为克服这些局限，一种称为生成式隐写术的新范式应运而生。它利用生成模型（如生成对抗网络 [11, 12, 13]和流模型 [14, 15], ）直接根据秘密消息生成隐写图像，无需预先准备载体图像。

近年来，越来越多基于扩散模型的生成式图像隐写术方法（DM-GIS）相继提出 [16, 17, 18, 19, 20, 21]。与基于生成对抗网络和基于流模型的方法相比，DM-GIS 方法具有无需训练的优势。此外，得益于扩散模型强大的生成能力 [22, 23], ，其生成的隐写图像在视觉上更加逼真。随着人工智能生成内容的快速发展，

预训练扩散模型在网上越来越容易获取，这极大地促进了 DM-GIS 的实现。因此，对有效检测方法的需求日益迫切。

然而，近期研究 [21, 16, 20, 17, 18] 表明，当前的 DM-GIS 方法能够有效规避通用图像隐写分析，达到近乎完美的经验安全性。这一现象可归因于两个关键因素：**（1）独特的嵌入空间**。传统图像隐写会将秘密消息直接嵌入图像域，而 DM-GIS 则利用扩散模型独特的结构属性，将消息嵌入其噪声空间。因此，专注于图像域的传统隐写分析方法 [24, 25, 26, 10]对 DM-GIS 技术的效果显著较差。**（2）异构数据源**。在 DM-GIS 场景中，载体图像不仅包括自然图像，还包括扩散模型生成的图像；在实践中，这些图像属于不同的数据分布。此外，载体图像和隐写图像都可以使用各种扩散模型、采样步数、采样器和引导尺度生成，这进一步增加了区分载体图像与隐写图像的难度。

本文重新思考了 DM-GIS 的隐写安全性，并重新评估了现有方法。我们首先回顾现有的 DM-GIS 流程，发现秘密消息被嵌入扩散模型的初始噪声向量或中间噪声向量中。这一观察促使我们将扩散模型的噪声空间与 DM-GIS 安全性联系起来。具体而言，我们从理论上证明，任何扰乱扩散模型噪声分布的隐写操作都会危及 DM-GIS 的安全性。我们还考察了 DM-GIS 安全性与消息提取准确率之间的关系，并证明在不修改消息嵌入方案的情况下，提取准确率较高的 DM-GIS 方法更容易被检测。基于这些发现，我们提出了基于 **N**oise**S**pace 的**D**iffusion**S**teganalyser（NS-DSer），这是一种简单而强大的隐写分析框架，用于重新评估现有的 DM-GIS 方法。该框架在扩散模型的噪声空间中执行检测，并且无需了解隐写者所使用的具体扩散模型配置。该框架具有两项优势：**(1)** 通过将区分载体图像与隐写图像这一复杂任务转化为更易处理的统计分布区分问题，简化了检测过程，使我们的方法既简单又有效。**(2)**通过使隐写分析不受异构数据源影响，提升了其实用性。NS-DSer 由两个阶段组成：确定性无条件扩散过程和统计特征提取。第一阶段中，我们采用确定性采样，在无条件引导的情况下恢复给定图像的初始噪声向量。第二阶段中，我们从噪声空间及其对应的变换域中提取统计特征，用于分类。
本文的贡献总结如下：

- 我们概述了现有 DM-GIS 方法的一般流程，并提供了理论分析，揭示了影响隐写安全性的关键因素。
- 我们提出了 NS-DSer，这是一种简单而有效的 DM-GIS隐写分析框架，用于重新评估现有 DM-GIS 方法。它有效解决了嵌入空间独特和数据源异构所带来的挑战。
- 我们设计了四种难度逐步提升的场景，用于评估当前的 DM-GIS 方法。实验表明，在所有四种场景中，NS-DSer 的表现均优于传统隐写分析器。

本文结构安排如下。第2节简要回顾现有 DM-GIS 方法和传统隐写分析。第3节介绍扩散模型以及隐写安全性的定义。第4节对 DM-GIS 安全性进行理论分析。第5节介绍我们提出的隐写分析器 NS-DSer。第6节报告实验结果。最后，第7节总结全文并展望未来可能的研究方向。

2 相关工作

2.1 基于扩散模型的生成式隐写

扩散模型广泛应用于生成式隐写术。Peng 等人 [21] 提出了 StegaDDPM，利用中间状态与生成图像之间的概率分布嵌入秘密比特。Jois等人 [20] 提出了 Pulsar，在去噪过程中利用方差噪声隐藏消息。Peng 等人 [27] 通过加入截断高斯编码机制，并将潜在扩散模型（LDMs） [23]用作骨干网络，改进了 StegaDDPM。Kim 等人 [16] 提出了一个两阶段框架，采用四种比特投影方法——消息到噪声（MN）、消息到二进制（MB）、消息到中心化二进制（MC）和多比特——以平衡提取准确率、安全性和图像质量。Yang 等人 [19] 提出了高斯着色，用于潜在扩散模型中的版权保护。Hu 等人[17] 通过在潜在扩散模型的初始噪声变换域中嵌入比特来提升鲁棒性，而Zhou 等人 [18] 则采用离散余弦变换域。Yu 等人 [28] 提出了 CRoSS，利用可控的可逆密钥系统隐藏图像；Yang 等人 [29] 则通过密码依赖的密钥管理解决了 CRoSS 中的密钥泄露问题。

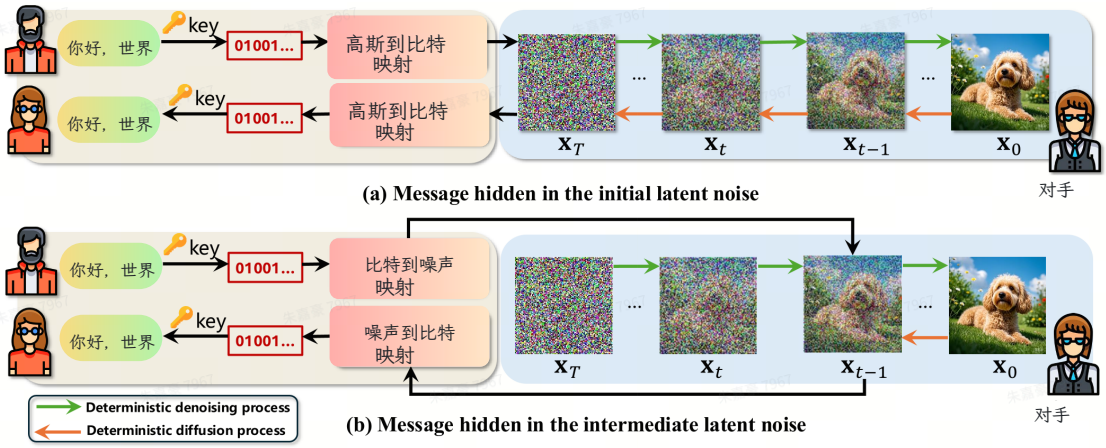

图 1：两种 DM-GIS 框架：(a) 秘密消息嵌入初始噪声 $\mathbf{x}_T$中，该噪声通常为高斯白噪声；(b) 秘密消息隐藏在确定性去噪过程中的某个时间步 t处的中间噪声 $\mathbf{x}_t$ 中。

2.2 图像隐写分析

传统隐写分析依赖SRM [30] 和GFR [31], 等手工设计的特征，但这些方法难以应对先进技术。深度学习，尤其是卷积神经网络，凭借出色的特征提取能力而日益流行。Qian 等人[32] 提出了GNCNN模型，其性能可与经典技术相媲美；而 XuNet [24] 的准确率则超过了SRM。Ye等人 [25] 提出YeNet改进隐写分析，其性能优于SRM和maxSRM [33]。Boroumand等人 [26] 提出了SRNet，这是一种用于更好地检测空间域隐写和JPEG隐写术的深度残差网络。Deng等人 [34] 通过四组卷积神经网络简化了这一方法。Zhang 等人 [35] 提出了ZhuNet，采用深度可分离卷积。近期研究聚焦于彩色图像隐写分析，Zeng 等人 [36] 提出WISERNet以提升准确率。Butora 等人 [37] 评估了经ImageNet预训练的卷积神经网络，发现EfficientNet [38] 适用于JPEG隐写术，但不适用于空间域。You 等人 [39] 提出了孪生卷积神经网络架构，Wei 等人 [40] 则提出了用于先进空间域隐写分析的UCNet。最近，Wei 等人提出了PENet [41], ，这是一种无需重新训练即可检测不同尺寸隐写图像的模型。

3 预备知识

3.1 扩散模型

扩散模型 [42, 43, 44] 由扩散过程和去噪过程组成。扩散过程由具有转移核 $\mathbb{P}(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\mathbf{x}_t;\ \alpha_t\mathbf{x}_0, \sigma_t^2\mathbf{I})$ 的 Itˆo 随机微分方程建模。去噪过程由反向随机微分方程描述：

$$\mathrm{d}\mathbf{x} = \left[f(\mathbf{x},t) - g(t)^2\nabla_{\mathbf{x}}\log\mathbb{P}_t(\mathbf{x})\right]\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{\omega}, \tag{1}$$

其中， $f(\mathbf{x}\ t)$ 和 $g(t)$ 分别为漂移系数和扩散系数， $\boldsymbol{\omega}$ 是布朗运动。神经网络 $\epsilon_\phi(\mathbf{x}\ t)$ 用于近似缩放后的得分函数 $-\sigma_t\nabla_{\mathbf{x}}\log\mathbb{P}_t(\mathbf{x})$。

通过从噪声 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\mathbf{I})$ 开始求解方程 (1)，即可得到图像 $\mathbf{x}_0$。或者，也可以通过由方程 (1) 推导出的概率流常微分方程生成图像：

$$\mathrm{d}\mathbf{x} = \left[f(\mathbf{x},t) - \frac{1}{2}g(t)^2\nabla_{\mathbf{x}}\log\mathbb{P}_t(\mathbf{x})\right]\mathrm{d}t. \tag{2}$$

3.2 基于扩散模型的隐写系统

在隐写系统中，隐写者将秘密消息嵌入图像等载体，并通过公开信道传输，而被动对手 $\mathcal{A}$ 则试图检测隐蔽通信。隐写者的目标是避开$\mathcal{A}$ 并传输编码消息。在基于扩散模型的生成式隐写（DM-GIS）中，隐写者使用密钥加密消息，并采用可逆的比特到噪声映射来隐藏消息。借助扩散模型的可逆性 [43, 22, 23], ，隐写者建立了隐写图像与

秘密消息之间的可逆映射。一些方法 [20, 21, 27] 使用共享随机种子的随机微分方程来实现可逆性，而另一些方法 [16, 19, 17, 18, 28, 29] 使用概率流常微分方程。图 1 展示了两种 DM-GIS 框架。

3.3 隐写安全性的定义

隐写安全性可以通过两种方式定义。Cachin [45]基于信息论定义安全性，并用载体图像分布 $\mathbb{P}_c$ 与隐写分布 $\mathbb{P}_s$之间的 Kullback-Leibler（KL）散度对其进行量化。若满足以下条件，则系统是 ϵ-安全的：

$$D_{KL}(\mathbb{P}_c \parallel \mathbb{P}_s) \leq \epsilon. \tag{3}$$

若满足 $\epsilon = 0$，则系统是完美安全的。Hopper 等人 [46] 从复杂性理论的角度定义安全性。若满足以下条件，则隐写系统在隐藏文本攻击下是（N_q、T_r、ϵ）-安全的：

$$\max_{N_q, T_r} \left| \mathbb{P}\left(\mathcal{A}^{O_S(\mathbf{k},\mathbf{m})} = 1\right) - \mathbb{P}\left(\mathcal{A}^{O_C} = 1\right) \right| < \epsilon. \tag{4}$$

O_C 和 O_S 分别是载体图像分布和隐写分布的预言机。$\mathbf{k}$ 是随机选择的密钥，$\mathbf{m}$ 是秘密消息。对手 $\mathcal{A}$ 最多可以发起 N_q 次查询，且运行时间不超过 T_r。

4 DM-GIS 安全性的理论分析

在第4.1节中，我们对基于扩散模型的生成式隐写（DM-GIS）进行隐写安全性分析，揭示 DM-GIS 安全性、扩散模型的噪声分布与消息提取准确率之间的关系。

4.1 将扩散模型噪声分布与 DM-GIS 安全性联系起来

先前的扩散模型 [43, 42] 指出，理论上可以构造一个可逆映射，将潜变量$\mathbf{x}_T$映射到由方程（11）驱动的图像$\mathbf{x}_0$。对于扩散模型，生成过程可定义为$\mathbf{x}_0 = F(\mathbf{x}_T)$，它可以分解为 N_s 分段积分 [47, 43],

$$\mathbf{x}_t = \mathbf{x}_{t+1} + \int_{t+1}^{t} \left[f(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log \mathbb{P}_t(\mathbf{x}) \right] \mathrm{d}t. \tag{5}$$

令 F^{-1} 表示相应的反向过程；下面给出以下定理：

定理1。令 $\mathbb{Q}_c$ 和 $\mathbb{Q}_s$ 分别表示普通扩散模型噪声分布和包含秘密消息的噪声分布。我们分别用 $\mathbb{P}_s$ 和 $\mathbb{P}_c$ 表示隐写图像分布和正常生成图像分布。理论上，以下等式成立： $D_{KL}(\mathbb{P}_c \| \mathbb{P}_s) = D_{KL}(\mathbb{Q}_c \| \mathbb{Q}_s)$。

证明。 假设 $\mathcal{X}$ 表示整个图像空间，则有

$$D_{KL}(\mathbb{P}_c(\mathbf{x}) || \mathbb{P}_s(\mathbf{x})) = \sum_{\mathbf{x} \in \mathcal{X}} \mathbb{P}_c(\mathbf{x}) \log \frac{\mathbb{P}_c(\mathbf{x})}{\mathbb{P}_s(\mathbf{x})} = \sum_{\mathbf{g} \in F^{-1}(\mathcal{X})} \sum_{\mathbf{x} \in F(\mathbf{g})} \mathbb{P}_c(\mathbf{x}) \log \frac{\mathbb{P}_c(\mathbf{x})}{\mathbb{P}_s(\mathbf{x})}. \tag{6}$$

令 $F(\mathbf{g}) = \{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$，则有

$$\sum_{\mathbf{x} \in F(\mathbf{g})} \mathbb{P}_c(\mathbf{x}) \log \frac{\mathbb{P}_c(\mathbf{x})}{\mathbb{P}_s(\mathbf{x})} = \sum_{i=1}^{k} \mathbb{P}_c(\mathbf{x}_i) \log \frac{\mathbb{P}_c(\mathbf{x}_i)}{\mathbb{P}_s(\mathbf{x}_i)}. \tag{7}$$

根据对数和不等式，我们可以很容易地得到

$$\sum_{i=1}^{k} \mathbb{P}_c(\mathbf{x}_i) \log \frac{\mathbb{P}_c(\mathbf{x}_i)}{\mathbb{P}_s(\mathbf{x}_i)} \geq \sum_{i=1}^{k} \mathbb{P}_c(\mathbf{x}_i) \log \frac{\sum_{j=1}^{k} \mathbb{P}_c(\mathbf{x}_j)}{\sum_{j=1}^{k} \mathbb{P}_s(\mathbf{x}_j)}. \tag{8}$$

由于$\mathbf{g} \in F^{-1}(\mathcal{X})$， $\mathbb{Q}_c(\mathbf{g}) = \sum_{\mathbf{x} \in F^{-1}(\mathbf{g})} \mathbb{P}_c(\mathbf{x})$，同理，$\mathbb{Q}_s(\mathbf{g}) = \sum_{\mathbf{x} \in F^{-1}(\mathbf{g})} \mathbb{P}_s(\mathbf{x})$。因此，我们有

$$D_{KL}(\mathbb{P}_c || \mathbb{P}_s) \geq \sum_{\mathbf{g} \in F^{-1}(\mathcal{X})} \sum_{i=1}^{k} \mathbb{P}_c(\mathbf{x}_i) \log \frac{\sum_{j=1}^{k} \mathbb{P}_c(\mathbf{x}_j)}{\sum_{j=1}^{k} \mathbb{P}_s(\mathbf{x}_j)} = \sum_{\mathbf{g} \in F^{-1}(\mathcal{X})} \mathbb{Q}_c(\mathbf{g}) \log \frac{\mathbb{Q}_c(\mathbf{g})}{\mathbb{Q}_s(\mathbf{g})} = D_{KL}(\mathbb{Q}_c || \mathbb{Q}_s). \tag{9}$$

由于去噪函 函数 F 是确定且可逆的，我们可以使用上述相同的方法来获 n

$$D_{KL}(\mathbb{Q}_c || \mathbb{Q}_s) \geq D_{KL}(\mathbb{P}_c || \mathbb{P}_s). \tag{10}$$

结合公式 (9) 和公式 (10)，我们有 $D_{KL}(\mathbb{P}_c \| \mathbb{P}_s) = D_{KL}(\mathbb{Q}_c \| \mathbb{Q}_s)$ 成立。证明完毕。 □

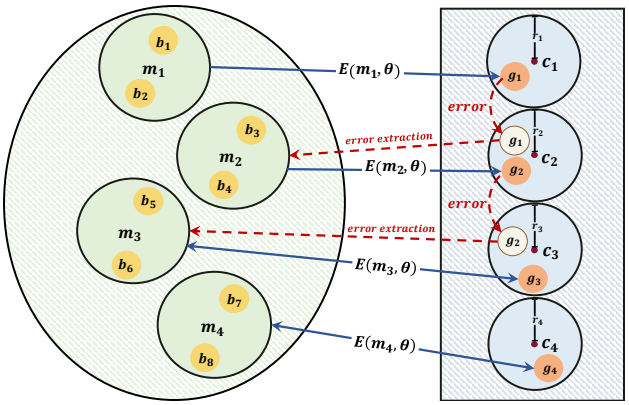

图 2： $E(\mathbf{m}\ \theta) = \mathbf{g}$ 的示意示例，其中展示了 DM-GIS 的可逆比特到噪声映射，同时指出了一种会出现提取错误的情形。

定理1表明，正常图像与隐写图像分布之间的 KL 散度在扩散模型的噪声空间中理论上得以保持。这意味着，任何破坏扩散模型原始噪声分布的隐写操作，都不可避免地会在隐写图像与正常生成图像之间引入分布散度。根据公式（3）中的隐写安全性定义，此类操作会增加$D_{KL}(\mathbb{P}_c \parallel \mathbb{P}_s)$，从而降低安全性。换言之，只有当 DM-GIS 的消息映射方案不改变扩散模型的自然噪声分布时，其在理论上才可以实现不可检测。

4.2 将消息提取准确率与 DM-GIS 安全性联系起来

除了隐写安全性之外，消息提取准确率也是隐写者考虑的另一项关键指标。下面我们探讨 DM-GIS 安全性与消息提取准确率之间的关系。

给定加密秘密比特序列 $\mathbf{b} = b_1 b_2 \cdots b_n$，序列 $\mathbf{b}$ 还可以进一步划分为多个长度为 $l\,(l \geq 1)$[19] 的子序列。这些子序列共同构成一个整数序列 $\mathbf{m} = m_1 m_2 \cdots m_k$，其中 $m_i \in \{0, 1, \cdots, 2^l - 1\}$。

命题 1。令 $E(\mathbf{m}; \theta) = \mathbf{g}$ 表示 *DM-GIS* 中使用的隐写编码器，其中 $\mathbf{g} = (g_1, g_2, \ldots, g_k) \sim \mathbb{Q}_s$，$\theta$ 表示一个可调参数。*DM-GIS* 的安全性由$D_{KL}(\mathbb{Q}_c \,\|\, \mathbb{Q}_s) = \epsilon$量化。若将 θ 修改为 θ' 以提高消息提取准确率，从而得到 $E(\mathbf{m}; \theta') = \mathbf{g}'$，则 $D_{KL}(\mathbb{Q}_c \,\|\, \mathbb{Q}_s) > \epsilon$。

证明。 该证明可在编码理论框架下表述。具体而言，隐写编码器可表示为 $E(m_i, \theta) = g_i = c_j + \eta_j$，其中 $j \in \{1, \ldots, 2^l - 1\}$定义了一个以 c_j为中心、半径为 r的汉明球。为了提取秘密比特， E^{-1} 必须至少是满射，但不一定是单射。图 2 给出了一个示例：所有元素 g_i 都被投影到蓝色互不相交区域中，确保任何满足 $\|\eta_j\| \leq r$的 g_i 都能解码为唯一的比特序列。然而，当发生扰动时，例如 g_1 落入以 c_2为中心的区域，就会出现消息提取错误。为了降低错误率，可以采用两种策略：增大中心点 c_j之间的距离，或缩小每个蓝色区域的 r。形式上，可以选择另一编码器参数 θ'，使得 $E(\mathbf{m}\ \theta') = \mathbf{g}'$，从而改善消息提取。然而，根据定理1，这不可避免地会改变$\mathbf{g}$的分布，导致 $D_{KL}(\mathbb{Q}_c \| \mathbb{Q}_s) > \epsilon$。证毕。 □

根据定理1，命题 1表明，对于给定的隐写编码器 $E(\cdot; \theta)$，任何能够提高消息提取准确率的θ 调整，都必然会损害 DM-GIS 的安全性。换言之，一旦固定了 $E(\cdot; \theta)$，DM-GIS 就无法同时实现高提取准确率和强安全性。这一发现使命题 1特别适用于分析可调式 DM-GIS 方法 [29, 27]的安全性。要克服这一固有的权衡，因此需要对编码器 E 进行更精细的设计。

5 基于噪声空间的扩散隐写分析器

定理1表明，载体图像与隐写图像分布之间的 KL 散度在扩散模型的噪声空间中得以保留。这意味着理论上可以在噪声空间中进行隐写分析。该过程

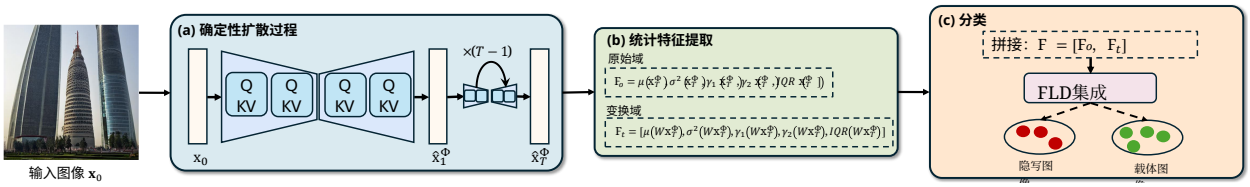

图3：**所提出的 NS-DSer 的检测流程**。输入图像$\mathbf{x}_0$ 首先经过由常微分方程求解器 Φ实现的无条件确定性扩散过程，得到噪声 $\hat{\mathbf{x}}_T^{\Phi}$。从 $\hat{\mathbf{x}}_T^{\Phi}$中提取五种统计特征——均值、方差、偏度、峰度和四分位距——以构成特征 $\mathcal{F}_o$。此外，还从 $\hat{\mathbf{x}}_T^{\Phi}$中提取变换域特征 $\mathcal{F}_t$。将特征 $\mathcal{F}_0$ 和 $\mathcal{F}_t$ 拼接后，输入二分类器。

比在图像空间中进行隐写分析容易得多，因为它可以简化为区分隐写噪声与高斯噪声的问题。然而，实际上，由于公式（5）中的数值误差和图像量化，F只能近似地进行逆变换。此外，定理1仅在已知 F 的假设下成立，而这在现实世界的检测场景中并不切实际。尽管如此，近期研究 [19] 通过实验证明，即使使用不同于隐写图像生成过程中所用的采样器和条件引导，仍能实现较高的消息提取准确率，这表明基于噪声空间的隐写分析有望应用于 DM-GIS 方法。受此启发，我们提出了 NS-DSer，它包含三个关键步骤。

确定性扩散过程。我们选择与去噪过程起始时间步 T 对应的噪声空间进行隐写分析。具体而言，给定图像 $\mathbf{x}_0$,，我们首先使用常微分方程求解器 Φ，经过 N_s 个采样步数，迭代地将其扩散为噪声 $\hat{\mathbf{x}}_T^{\Phi}$：

$$\hat{\mathbf{x}}_{t+1}^{\Phi} = \mathbf{x}_t + \int_t^{t+1} \left(f(t)\mathbf{x}_t + \frac{g^2(t)}{2\sigma_t} \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, \mathbf{c}) \right) \mathrm{d}t \approx \Phi(\mathbf{x}_t, t, t+1, \mathbf{c}; \phi), \tag{11}$$

其中$\mathbf{c}$ 是给定的条件（例如文本）。本文为兼顾计算效率和数值精度，采用 2nd 阶常微分方程求解器，例如 DPM-Solver-2 [47] 和 Heun求解器 [44],，以实现 $\Phi(\cdot,\ \cdot,\ \cdot,\ \cdot;\ \phi)$。

Wh当应用无分类器引导（CFG） [48] 时，公式（11）中的 $\boldsymbol{\epsilon}_\phi(\mathbf{x}_t,\ t,\ \mathbf{c})$ 将转换为 to

$$\tilde{\boldsymbol{\epsilon}}_\phi(\mathbf{x}_t, t, \mathbf{c}) = \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, \emptyset) + \omega(\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, \mathbf{c}) + \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, \emptyset)), \tag{12}$$

其中，ω 表示 CFG 尺度。在实际检测场景中，隐写者所使用的具体条件 $\mathbf{c}$ 通常是未知的。因此，我们可以利用现成的视觉语言模型（VLM），例如 ChatGPT [49],，只需向 VLM 提示“请简要描述这张图像。”，即可从图像中提取 $\mathbf{c}$。或者，我们也可以选择在确定性扩散过程中不应用 $\mathbf{c}$。这两种方法的比较见第6节.2。

统计特征提取。在目标噪声空间中，$\hat{\mathbf{x}}_T^{\Phi}$近似于随机噪声，这意味着它几乎不包含原始图像 $\mathbf{x}_0$中的语义信息。因此，以往图像隐写分析方法中常用的深度网络架构并不适用于这种情形。此外，我们发现，在噪声空间中直接应用这些网络很容易导致过拟合，进而造成训练不稳定。为应对这些挑战，我们提出使用 $\hat{\mathbf{x}}_T^{\Phi}$的统计特征进行隐写分析。具体而言，我们发现，当所选扩散模型 ϕ 为方差保持型 [42, 44]时，$\hat{\mathbf{x}}_T^{\Phi}$近似于标准高斯噪声。因此，区分载体图像和隐写图像可视为区分隐写噪声分布与已知的载体噪声分布。基于这一观察，我们首先将 $\hat{\mathbf{x}}_T^{\Phi}$展平，并提取五种统计特征——均值、方差、偏度、峰度和四分位距（IQR）——以如下方式表示 $\hat{\mathbf{x}}_T^{\Phi}$：

$$\mathcal{F}_o = [\mu(\hat{\mathbf{x}}_T^{\Phi}), \sigma^2(\hat{\mathbf{x}}_T^{\Phi}), \gamma_1(\hat{\mathbf{x}}_T^{\Phi}), \gamma_2(\hat{\mathbf{x}}_T^{\Phi}), IQR(\hat{\mathbf{x}}_T^{\Phi})] \tag{13}$$

与传统图像隐写类似，传统图像隐写将秘密消息嵌入频域，一些 DM-GIS 方法 [18, 17] 将秘密比特隐藏在 $\mathbf{x}_T$ 的变换域中。本文中，我们选择离散余弦变换，并将其应用于 $\hat{\mathbf{x}}_T^{\Phi}$ 的每个通道，简记为 $W\hat{\mathbf{x}}_T^{\Phi}$。为了捕捉变换域中载体图像与隐写样本之间统计分布的差异，我们同样提取五个统计特征，使用

$$\mathcal{F}_t = [\mu(W\hat{\mathbf{x}}_T^{\Phi}), \sigma^2(W\hat{\mathbf{x}}_T^{\Phi}), \gamma_1(W\hat{\mathbf{x}}_T^{\Phi}), \gamma_2(W\hat{\mathbf{x}}_T^{\Phi}), IQR(W\hat{\mathbf{x}}_T^{\Phi})]. \tag{14}$$

最终特征由 =　　$_o$,　$_t$ 表示d b　$\mathcal{F}$　[$\mathcal{F}$　$\mathcal{F}$]

分类 对于分类，我们采用费 Fisher线性判别（FLD）集成方法 [50]，因为它在性能和计算效率之间取得了出色的平衡。NS-DSer 的训练流程见算法1和图3。

算法1NS-DSer 的训练流程

输入：载体图像 图像 $\mathbf{x}_c$、隐写图像 $\mathbf{x}_s$、采样步数 N_s、引导尺度 ω、常微分方程求解器 $\Phi(\cdot,\ \cdot,\ \cdot,\ \cdot;\ \phi,\ \omega\ )$.

输出：最优 FLD 集成。

1：对于 $i=1$ 到 N_s，执行 2：$\hat{\mathbf{x}}^{\Phi}_{c,t+1}=\Phi$ ($\hat{\mathbf{x}}^{\Phi}_{c,t}, t, t+1, \varnothing;\ \phi, \omega$) 3：$\hat{\mathbf{x}}^{\Phi}_{s,t+1}=\Phi(\hat{\mathbf{x}}^{\Phi}_{s,t}, t,$ $t+1, \varnothing;\ \phi, \omega$) 4：展平 $\hat{\mathbf{x}}^{\Phi}_{c,T}$ 和 $\hat{\mathbf{x}}^{\Phi}_{s,T}$5：使用公式（13）提取 $\hat{\mathbf{x}}^{\Phi}_{c,T}$ 的特征 $\mathcal{F}^c_o$ 6：使用公式（13）提取 $\hat{\mathbf{x}}^{\Phi}_{s,T}$ 的特征 $\mathcal{F}^s_o$ 7：使用公式（14）提取 $W_q\hat{\mathbf{x}}^{\Phi}_{c,T}$ 的特征 $\mathcal{F}^c_t$ 8：使用公式（14）提取 $W_q\hat{\mathbf{x}}^{\Phi}_{s,T}$ 的特征 $\mathcal{F}^s_t$ 9：$\mathcal{F}_c=[\mathcal{F}^c_o, \mathcal{F}^c_t]$ 和 $\mathcal{F}_s=[\mathcal{F}^s_o, \mathcal{F}^s_t]$10：基于特征 $\mathcal{F}_c$ 和 $\mathcal{F}_s$训练 FLD 集成 11：返回最优 FLD 集成

表 1：场景 *#1*∼*#4*的设置。r_g 和 r_n 分别表示载体（或隐写图像）数据集中生成图像与自然图像的比例。SD1.5、SD2.1 和 DS7 分别是 Stable Diffusion 1.5 、 Stable Diffusion 2.1 和 Dreamshaper 7 的简称。 1st 阶采样器采用 50 步采样，其中 2nd 阶采样器采用 20 步采样。
†表示 CRoSS 的设置。

		r_g	r_n	骨干网络	采样步数	ω
#1	载体图像	100%	0%	SD1.5	20, 50	7.5, 1†
	隐写图像	100%	0%	SD1.5	20, 50	7.5, 1†
#2	载体图像	50%	50%	SD1.5	20, 50	7.5, 1†
	隐写图像	100%	0%	SD1.5	20, 50	7.5, 1†
#3	载体图像	50%	50%	SD1.5,SD2.1,DS7	[15,25],[45,55]	[4.5,10.5],[1,9]†
#4	隐写图像	100%	0%	SD1.5,SD2.1,DS7	[15,25],[45,55]	[4.5,10.5],[1,9]†

6 实验

我们通过五种隐写分析器重新评估了九种现有 DM-GIS 方法的安全性：MN [16], MC [16], MB [16], 高斯阴影（G-S）[19], StegaDDPM [21], LDStega [27], GSD [18], mas-GRDH [17], 和 CRoSS [28], 以及 XuNet[24], SRNet [26], SuaStegBet [39], UCNet [40],和 NS-DSer。

DM-GIS设置。对于 MN、MC、MB、G-S、GSD 和 mas-GRDH，我们在生成隐写图像时使用 2nd 阶常微分方程求解器 [44]，其他方法则使用 1st 采样器 [43]。LDStega 的截断区间设为 0.3。除非另有说明，其他方法均采用默认设置。

隐写分析器设置。前四种隐写分析器分别使用 6,000 对载体-隐写图像对进行训练，并在 1,000 对图像上测试，共训练 100 个周期。相比之下，由于 NS-DSer 提取的特征空间仅有 10 维，因此使用较小的数据集进行训练，其中包含 1,800 对载体-隐写图像对，并在 200 对图像上测试。所有训练和测试图像均根据 Flickr8K 数据集中的提示生成 [1]。鉴于几乎所有现有的 DM-GIS 方法都利用基于 U-Net 的扩散模型进行隐写，NS-DSer 也采用这一架构进行特征提取。除非另有说明，我们采用 Stable Diffusion 2.1 作为基础生成模型，并使用 20 步 Heun 采样 [44]，不进行提示引导。
NS-DSer 使用 PyTorch 和 MATLAB 实现，所有实验均使用 8 个 A100 GPU 和 8 个 L40 GPU 开展。

6.1 安全性与检测评估

为全面评估 DM-GIS 方法的安全性并比较当前隐写分析器的性能，我们设置了以下四种难度逐渐增加的检测场景。

[1]https://www.kaggle.com/datasets/adityajn105/flickr 8k

表 2：各隐写分析器在场景 *#1*∼*#3*中对九种 DM-GIS 方法的检测准确率（%）及总体评估。

	场景 *#1*	MN	MC	MB	G-S	StegaDDPM	LDStega	GSD	mas-GRDH	CRoSS	总体
#1	XuNet	49.17	52.25	58.86	51.38	50.67	89.41	50.79	50.88	83.63	0.8255
	SRNet	49.71	80.95	97.29	50.21	51.92	98.75	49.25	51.88	88.40	0.9039
	SiaStegNet	50.33	91.52	98.71	50.75	50.96	99.90	50.38	50.13	93.20	0.9283
	UCNet	49.67	92.90	99.60	51.50	50.54	99.95	49.67	50.70	**93.65**	0.9281
	NS-DSer（我们的方法）					50.05	**99.99**	**99.58**	51.13	85.70	**0.9742**
#2	XuNet	59.68	61.65	70.21	60.18	60.35	91.25	60.21	60.19	81.13	0.7635
	SRNet	70.28	86.95	97.90	70.80	70.25	99.96	70.25	70.00	87.50	0.7392
	SiaStegNet	74.10	91.80	98.55	74.20	73.95	99.95	73.00	74.00	91.45	0.7141
	UCNet	73.10	90.55	99.15	74.05	73.50	99.90	73.80	74.10	**93.50**	0.7200
	NS-DSer（本文）					62.83	**99.99**	**99.60**	62.98	90.28	**0.8633**
#3	XuNet	61.14	62.94	61.96	58.50	60.19	89.73	59.39	59.68	72.40	0.7476
	SRNet	69.05	76.60	92.80	69.06	69.30	99.55	69.00	69.20	77.90	0.7243
	SiaStegNet	73.75	83.40	94.40	73.10	73.95	99.90	73.50	73.00	77.40	0.6941
	UCNet	73.85	86.35	96.20	74.35	74.15	99.80	74.65	74.45	79.85	0.6949
	NS-DSer（我们的方法）					59.57	**99.99**	**96.15**	59.53	**83.48**	**0.8817**

表 3：各隐写分析器在场景 *#4*下的检测准确率（%）。

#4	XuNet	SRNet	SiaStegNet	UCNet	NS-DSer（我们的方法）
	57.96	74.80	75.80	73.00	**86.26**

场景 ***#1***。在此场景中，隐写者使用相同的 DM-GIS 方法进行隐蔽通信。载体图像由标准正态分布生成，采用与生成隐写图像相同的一组采样配置。具体参数详见表 1。被动对手 $\mathcal{A}$ 使用各个隐写分析器进行隐写图像检测。

场景 ***#2***。在此场景中，隐写者也采用相同的 DM-GIS 方法进行隐蔽通信。不过，载体图像集有所扩展，除生成图像外，还包括来自 Flickr8K 数据集的自然图像，具体详情见表 1。

场景 ***#3***。在此场景中，隐写者继续采用相同的 DM-GIS 方法进行隐蔽通信，载体图像既包括生成图像，也包括自然图像。关键区别在于，隐写图像和生成的载体图像采用不同的生成骨干网络、采样步数和 CFG 缩放系数 ω 生成，如表 1 所示。

场景 ***#4***。在此场景中，隐写者可以采用不同的 DM-GIS 方法进行隐蔽通信。载体图像既包括生成图像，也包括自然图像；而隐写图像和生成的载体图像采用不同的生成骨干网络、采样步数和 CFG 缩放系数 ω，如表 1 所示。具体而言，我们假设隐写者使用四种 DM-GIS 方法生成隐写图像：MN、MC、MB 和 GSD。

场景 #1 的结果场景 ***#1***。据我们所知，在评估 DM-GIS 方法时，MN、G-S、StegaDDPM 和 mas-GRDH 均属于保持分布方法。从理论上讲，已证明它们与生成的载体图像在统计上不可区分。简单来说，检测这些方法归结为区分加密的 0-1 序列和伪随机 0-1 序列。显然，这类问题无法在多项式时间内解决；因此，根据公式 (4)，它们是计算安全的。表 2 中的实验结果也与这一结论一致。至于其他 DM-GIS 方法，至少有一种隐写分析器能够检测出它们，这表明它们在场景 *#1* 下的安全性并非如其所宣称的那样。

在隐写分析器方面，除 LDStega 和 CRoSS 外，XuNet 无法检测所有 DM-GIS 方法。SRNet、SiaStegNet 和 UCNet 能够检测的 DM-GIS 方法也十分有限。相比之下，NS-DSer 能有效检测几乎所有 DM-GIS 方法，唯独无法检测保持分布方法。这凸显了以往隐写分析器的一项关键局限：它们依赖图像空间，因而不足以检测 DM-GIS 方法。NS-DSer 通过在噪声空间中进行隐写分析解决了这一问题。不过，我们发现，在场景 *#1*下，NS-DSer 检测 CRoSS 的表现相对较差。这是因为它更关注分布差异，而非细微扰动。尽管如此，正如表 7 所示，这种设计使训练速度更快。

表 4：跨场景 # N_s 的采样步数消融研究*1* ∼*#3*。

		MC	MB	LDStega	GSD	CRoSS
#1	$N_s = 15$	99.78	99.99	99.99	99.43	85.72
	$N_s = 20$	99.78	99.99	99.99	99.58	85.70
	$N_s = 25$	99.82	99.99	99.99	99.42	85.52
#2	$N_s = 15$	99.67	99.99	99.99	99.52	89.30
	$N_s = 20$	99.73	99.99	99.99	99.60	90.28
	$N_s = 25$	99.73	99.99	99.99	99.77	90.28
#3	$N_s = 15$	98.43	99.12	99.99	96.65	83.08
	$N_s = 20$	98.18	99.05	99.99	96.15	83.48
	$N_s = 25$	98.20	99.20	99.99	97.53	83.22

表 5：跨场景 *#1* ∼对有无提示引导的检测进行消融研究 *#3*.

		MC	MB	LDStega	GSD	CRoSS
#1	无条件	99.78	99.99	99.99	99.58	85.70
	有条件	99.60	99.99	99.99	99.50	75.18
#2	无条件	99.73	99.99	99.99	99.60	90.28
	有条件	99.67	99.88	99.99	99.48	78.00
#3	无条件	98.18	99.05	99.99	96.15	83.48
	有条件	99.35	99.92	99.99	99.60	77.12

表6：场景 *#4*下采样步数 N_s 与提示引导的消融研究。

#4	$N_s = 15$	$N_s = 20$	$N_s = 25$	无条件	有条件
	86.25	86.26	86.00	86.26	86.90

场景 ***#2***的结果。场景 *#2*在训练和测试中均使用自然图像作为载体图像。对于 MC、MB、LDStega、GSD 和 CRoSS 的检测评估结果，与场景 *#1*中的结果相当。然而，在检测保持分布方法时，所有隐写分析器的检测准确率都显著提升。正如场景 *#1*中所述，保持分布方法在统计上与生成的载体图像无法区分。**这些方法的检测准确率提高，本质上说明这五种隐写分析器并未学会区分载体图像和隐写图像，而是区分自然图像与生成图像**。理想情况下，表现良好的隐写分析器在处理保持分布方法时应取得较差的结果（检测准确率接近 50% ），而在其他方法上则应达到较高的准确率（接近 100% ）。基于这一原则，我们对五种隐写分析器在九种 DM-GIS 方法上的表现进行如下综合评估：对于保持分布方法的检测结果（记为 D_{dp}），我们首先使用公式 $1 - \frac{|D_{dp}-50|}{50}$ 对其进行归一化。对于其他方法的结果（记为 D_{ndp}），则按 $1 - \frac{|D_{ndp}-100|}{100}$ 进行归一化。随后，我们计算所有归一化结果的平均值。如表 2 所示，NS-DSer 的总体检测性能优于其他四种隐写分析器。

结果： 场景 ***#3***。如表 2 所示，传统隐写分析器的表现明显更差，表明其难以应对异构数据。相比之下，我们的 NS-DSer 性能下降极小，在 MC、MB、LDStega 和 GSD 上仍保持 95% 以上的准确率，从而验证了其对这类数据的鲁棒性。

场景结果 场景 ***#4***。场景 *#4* 模拟了一个更接近真实条件的隐写场景。如表 3 所示，传统图像隐写分析器的表现再次显著落后，而 NS-DSer 的表现则远超它们，进一步验证了其对异构数据的适应能力。这些结果彰显了 NS-DSer 在真实场景中的稳健性和实用性。

表 7：各隐写分析器的训练时长。

时间（小时）	XuNet	SRNet	SiaStegNet	UCNet	NS-DSer（我们的方法）
	1.84	5.52	7.90	8.25	1.11

6.2 消融研究

采样步数 N_s。我们在 NS-DSer 的确定性扩散采样中使用三个 N_s 值，以评估它们对检测的影响。正如表 4 和表6所示，在场景 *#1*∼*#4*中的结果一致表明，NS-DSer 的检测性能不受 N_s的影响。

提示引导。我们使用豆包 1.5 视觉提取图像描述，作为确定性扩散采样的提示引导。如表 5 所示，提示引导几乎不会影响 NS-DSer 对 MC、MB、GSD 和 LDStega 的检测，但会显著影响对 CRoSS 的检测。不过在实践中，隐写者使用的引导提示通常无法获取，因此这种性能下降不在我们的责任范围内。

6.3 计算效率

表 7 展示了在单张 L40 GPU 上训练各隐写分析器所需的时间。具体而言，XuNet、SRNet、SiaStegNet 和 UCNet 均训练 100 个周期。如图所示，训练 NS-DSer 所需时间最短，这得益于我们新颖的隐写分析框架。该框架避免在复杂的高维图像空间中进行隐写分析，而是将隐写分析转移到低维噪声空间，从而实现高效训练。此外，NS-DSer 采用无条件确定性采样进行特征提取（如图 1 所示），进一步加快了训练过程。

7 结论

本文重新审视了 DM-GIS 的安全性，并从理论和实验两个角度分析了改变扩散模型中的噪声分布对 DM-GIS 安全性的影响。我们提出了 NS-DSer，这是首个专门用于检测 DM-GIS 的隐写分析框架。通过将隐写分析从复杂的高维图像空间转移到扩散模型的低维噪声空间，NS-DSer 实现了高效的训练和测试。我们在四种难度逐渐增加的检测场景中评估了 NS-DSer，实验结果表明，总体而言，它在所有场景中的表现均优于传统图像隐写分析器，使实用隐写分析更近一步。未来，我们将改进特征提取流程，以进一步提升检测性能。

参考文献

[1] Weiqi Luo, Kangkang Wei, Qiushi Li, Miaoxin Ye, Shunquan Tan, Weixuan Tang, Jiwu Huang 等. 数字图像隐写与隐写分析综合综述. *APSIPA*信号与信息处理汇刊, 13(1), 2024.[2] Pratap Chandra Mandal, Imon Mukherjee, Goutam Paul 和 B.N. Chatterji. 数字图像隐写：文献综述. 信息科学, 609:1451–1488, 2022.[3] Bin Li, Ming Wang, Xiaolong Li, Shunquan Tan 和 Jiwu Huang. 空间图像隐写中聚类修改方向的策略. *IEEE*信息取证与安全汇刊, 10(9):1905–1917, 2015.[4] Vahid Sedighi、Rémi Cogranne 和 Jessica Fridrich. 通过最小化统计可检测性实现内容自适应隐写. *IEEE*信息取证与安全汇刊, 11(2):221–234, 2015.[5] Weiming Zhang、Zhuo Zhang、Lili Zhang、Hanyi Li 和 Nenghai Yu. 自适应隐写的联合失真分解. *IEEE* 视频技术电路与系统汇刊, 27(10):2274–2280, 2016.[6] Dongxia Huang、Weiqi Luo、Peijia Zheng 和 Jiwu Huang. 基于生成对抗网络的自动非对称嵌入代价学习. 收录于第*31*届 *ACM* 多媒体国际会议论文集，第8316–8326页，2023.[7] Dongxia Huang、Weiqi Luo、Minglin Liu、Weixuan Tang 和 Jiwu Huang. 基于生成式多重对抗网络的隐写嵌入代价学习. *IEEE*信息取证与安全汇刊, 19:15–29, 2023.

[8] Fangjun Huang、Jiwu Huang 和 Yun-Qing Shi. JPEG 隐写术的新通道选择规则. IEEE信息取证与安全汇刊, 7(4):1181–1191, 2012.[9] Weixiang Li、Kejiang Chen、Weiming Zhang、Hang Zhou、Yaofei Wang 和 Nenghai Yu. 利用估计辅助信息的 JPEG 隐写术. IEEE 视频技术电路与系统汇刊, 30(7):2288–2294, 2019.[10] Zichi Wang、Guorui Feng、Zhenxing Qian 和 Xinpeng Zhang. 采用内容相似度评估的 JPEG 隐写术. IEEE 控制论汇刊, 53(8):5082–5093, 2022.[11] Ming-ming Liu、Min-qing Zhang、Jia Liu、Ying-nan Zhang 和 Yan Ke. 基于生成对抗网络的无载体信息隐藏. arXiv预印本 arXiv:1712.06951, 2017.[12] Zihan Wang、Neng Gao、Xin Wang、Xuexin Qu 和 Linghui Li. Sstegan：基于生成对抗网络的自学习隐写术. 收录于《神经信息处理：第25届国际会议，ICONIP 2018》，柬埔寨暹粒，2018年12月13日至16日，会议录，第二部分25，第253–264页. 施普林格，2018.[13] Ping Wei、Sheng Li、Xinpeng Zhang、Ge Luo、Zhenxing Qian 和 Qing Zhou. 生成式隐写网络. 收录于第30届 ACM 多媒体国际会议论文集，第1621–1629页，2022.[14] Ping Wei、Ge Luo、Qi Song、Xinpeng Zhang、Zhenxing Qian 和 Sheng Li. 生成式隐写流. 收录于2022年 IEEE 多媒体与博览会国际会议（ICME），第1–6页. IEEE，2022.[15] Zhili Zhou、Yuecheng Su、Jin Li、Keping Yu、QM Jonathan Wu、Zhangjie Fu 和 Yunqing Shi. 面向生成式隐写术的秘密信息到图像的可逆变换. IEEE 可靠与安全计算汇刊, 20(5):4118–4134, 2022.[16] Daegyu Kim、Chaehun Shin、Jooyoung Choi、Dahuin Jung 和 Sungroh Yoon. Diffusion-stego：通过消息投影实现无需训练的扩散生成式隐写术. 信息科学, 第122358页，2025.[17] Xiaoxiao Hu、Sheng Li、Qichao Ying、Wanli Peng、Xinpeng Zhang 和 Zhenxing Qian. 基于流行的 Stable Diffusion 实现鲁棒生成式图像隐写术. IEEE信息取证与安全汇刊, 2024.[18] Qing Zhou、Ping Wei、Zhenxing Qian、Xinpeng Zhang 和 Sheng Li. 基于扩散模型的改进生成式隐写术. IEEE 视频技术电路与系统汇刊，2025.[19] Zijin Yang、Kai Zeng、Kejiang Chen、Han Fang、Weiming Zhang 和 Nenghai Yu. 高斯着色：针对扩散模型、性能损失可证明无损的图像水印. 收录于 IEEE/CVF 计算机视觉与模式识别会议论文集，第12162–12171页，2024.[20] Tushar M Jois、Gabrielle Beck 和 Gabriel Kaptchuk. Pulsar：面向扩散模型的安全隐写术. 收录于2024年 ACM SIGSAC 计算机与通信安全会议论文集，第4703–4717页，2024.[21] Yinyin Peng、Donghui Hu、Yaofei Wang、Kejiang Chen、Gang Pei 和 Weiming Zhang. StegaDDPM：基于去噪扩散概率模型的生成式图像隐写术. 收录于第31届 ACM 多媒体国际会议论文集，第7143–7151页，2023.[22] Jonathan Ho、Ajay Jain 和 Pieter Abbeel. 去噪扩散概率模型. 神经信息处理系统进展, 33:6840–6851, 2020.[23] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer. 利用潜在扩散模型进行高分辨率图像合成. 收录于 IEEE/CVF 计算机视觉与模式识别会议论文集，第10684–10695页, 2022.[24] Guanshuo Xu、Han-Zhou Wu 和 Yun-Qing Shi. 用于隐写分析的卷积神经网络结构设计. IEEE 信号处理快报, 23(5):708–712, 2016.[25] Jian Ye、Jiangqun Ni 和 Yang Yi. 用于图像隐写分析的深度学习层次化表示. IEEE信息取证与安全汇刊, 12(11):2545–2557, 2017.[26] Mehdi Boroumand、Mo Chen 和 Jessica Fridrich. 用于数字图像隐写分析的深度残差网络. IEEE信息取证与安全汇刊, 14(5):1181–1193, 2018.[27] Yinyin Peng、Yaofei Wang、Donghui Hu、Kejiang Chen、Xianjin Rong 和 Weiming Zhang. LDStega：基于潜在扩散模型的实用且鲁棒的生成式图像隐写术. 收录于第32届 ACM 多媒体国际会议论文集，第3001–3009页，2024.[28] Jiwen Yu、Xuanyu Zhang、Youmin Xu 和 Jian Zhang. CRoSS：扩散模型实现可控、鲁棒且安全的图像隐写术. 神经信息处理系统进展, 36:80730–80743, 2023.

[29] Yiwei Yang、Zheyuan Liu、Jun Jia、Zhongpai Gao、Yunhao Li、Wei Sun、Xiaohong Liu 和 Guangtao Zhai。DiffStega：借助扩散模型实现通用的免训练无载体图像隐写。arXiv 预印本 arXiv:2407.10459，2024。[30] Jessica Fridrich 和 Jan Kodovsky。用于数字图像隐写分析的富模型。IEEE信息取证与安全汇刊，7(3):868–882，2012。[31] Xiaofeng Song、Fenlin Liu、Chunfang Yang、Xiangyang Luo 和 Yi Zhang。使用二维Gabor滤波器对自适应JPEG隐写进行隐写分析。发表于第3届 ACM 信息隐藏与多媒体安全研讨会论文集，第 15–23 页，2015。[32] Yinlong Qian、Jing Dong、Wei Wang 和谭铁牛。通过卷积神经网络进行隐写分析的深度学习方法。发表于《2015 年媒体水印、安全与取证》，第 9409 卷，第 171–180 页。SPIE，2015。[33] Tomas Denemark、Vahid Sedighi、Vojtech Holub、Rémi Cogranne 和 Jessica Fridrich。用于数字图像隐写分析的选择信道感知富模型。发表于《2014 年 IEEE 信息取证与安全国际研讨会》，第 48–53 页。IEEE，2014。[34] Xiaoqing Deng、Bolin Chen、Weiqi Luo 和 Da Luo。用于图像隐写分析的快速高效全局协方差池化网络。发表于 ACM 信息隐藏与多媒体安全研讨会论文集，第 230–234 页，2019。[35] Ru Zhang、Feng Zhu、Jianyi Liu 和 Gongshen Liu。深度可分离卷积与多级池化：用于高效空间 CNN 隐写分析。IEEE信息取证与安全汇刊，15:1138–1150，2019。[36] Jishen Zeng、Shunquan Tan、Guangqing Liu、Bin Li 和 Jiwu Huang。WISERNet：用于彩色图像隐写分析的加宽先分离后汇合网络。IEEE信息取证与安全汇刊，14(10):2735–2748，2019。[37] Jan Butora、Yassine Yousfi 和 Jessica Fridrich。如何为隐写分析进行预训练。发表于《2021 年 ACM 信息隐藏与多媒体安全研讨会论文集》，第 143–148 页，2021。[38] Mingxing Tan 和 Quoc Le。EfficientNet：重新思考卷积神经网络的模型缩放。发表于《国际机器学习会议》，第 6105–6114 页。PMLR，2019。[39] Weike You、Hong Zhang 和 Xianfeng Zhao。用于图像隐写分析的孪生 CNN。IEEE信息取证与安全汇刊，16:291–306，2020。[40] Kangkang Wei、Weiqi Luo、Shunquan Tan 和 Jiwu Huang。基于通道表示的彩色图像隐写分析通用深度网络。IEEE信息取证与安全汇刊，17:3022–3036，2022。[41] Kangkang Wei、Weiqi Luo 和 Jiwu Huang。基于像素差分卷积和选择性池化增强型 Transformer 的彩色图像隐写分析。IEEE信息取证与安全汇刊，2024。[42] Yang Song、Jascha Sohl-Dickstein、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程进行基于分数的生成建模。arXiv 预印本 arXiv:2011.13456，2020。[43] Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。arXiv 预印本 arXiv:2010.02502，2020。[44] Tero Karras、Miika Aittala、Timo Aila 和 Samuli Laine。阐明基于扩散的生成模型的设计空间。神经信息处理系统进展，35:26565–26577，2022。[45] Christian Cachin。隐写的信息论模型。发表于《信息隐藏国际研讨会》，第 306–318 页。施普林格，1998。[46] Nicholas J Hopper、John Langford 和 Luis von Ahn。可证明安全隐写术。发表于《年度国际密码学会议》，第 77–92 页。施普林格，2002。[47] Cheng Lu、Yuhao Zhou、Fan Bao、Jianfei Chen、Chongxuan Li 和 Jun Zhu。DPM-Solver：用于扩散概率模型采样的快速常微分方程求解器，仅需约 10 步。神经信息处理系统进展，35:5775–5787，2022。[48] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv 预印本 arXiv:2207.12598，2022。[49] OpenAI。你好，GPT-4o，2024。[50] Jan Kodovsky、Jessica Fridrich 和 Vojtˇech Holub。用于数字媒体隐写分析的集成分类器。IEEE信息取证与安全汇刊，7(2):432–444，2011。