计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(7): 1831−1841, 2026

基于双扩散模型的未知源-目标域生成式图像隐写方法

董云云[1,2] 张锦弘[1] 李钒效[1] 朱 波[2] 周 维[3]
[1](云南大学信息学院 昆明 650504)
[2](云南大学软件学院(人工智能学院) 昆明 650504)
[3](云南大学现代工学院 昆明 650500)
(dongyy929@ynu.edu.cn)

A Generative Image Steganography Method Based on Dual Diffusion Models for Unknown Source-Target Domains

Dong Yunyun[1,2], Zhang Jinhong[1], Li Fanxiao[1], Zhu Bo[2], and Zhou Wei[3]
[1](*School of Information Science and Engineering*, *Yunnan University*, *Kunming* 650504)
[2](*School of Software and AI*, *Yunnan University*, *Kunming* 650504)
[3](*School of Engineering*, *Yunnan University*, *Kunming* 650500)

Abstract The rapid expansion of social networks has made data transmission over public channels ubiquitous, raising serious security and privacy concerns. Although encryption can guarantee confidentiality, its distinctive format may arouse suspicion. Steganography offers a covert alternative by hiding secret data within ordinary media, such as images, text, or audio, thereby reducing the risk of detection. Traditional pixel-editing steganographic methods are vulnerable to steganalysis or compression artifacts, while existing generative approaches often depend on tailored prompts or specific data sources, limiting their generalization and flexibility. To address these challenges, we propose a source-agnostic, target-domain image steganography framework based on a dual diffusion model. First, this method employs a latent diffusion model (LDM) to invert and optimize arbitrary secret images, obtaining high-fidelity latent representations. To eliminate residual structural cues from the LDM inversion that could compromise subsequent hiding and extraction, this method introduces a Gaussian structural reordering alignment module which encrypts the secret's latent representation via pseudo-random orthogonal perturbation and spatial shuffling, preserving its norm while enforcing a Gaussian-like distribution that aligns with downstream inputs and enhances security. Finally, we use a denoising diffusion implicit models (DDIM) sampler to denoise and decode the encrypted latent vectors, producing stego images that faithfully reflect the target domain's appearance while carrying the hidden content. Experiments on four public datasets demonstrate that our method requires no source-specific training, achieves strong cross-domain generalization, and significantly outperforms competing techniques in extraction accuracy, imperceptibility, and resistance to steganalysis.

Key words image steganography; image generation; diffusion model; strong generalizability; distribution shift

摘 要 社交网络的快速发展使得通过公共频道的数据传输变得无处不在,带来了显著的安全性和隐私

收稿日期: 2025-05-26; **修回日期**: 2026-03-25
基金项目: 云南省基础研究项目(202401AT070474, 202501AU070059); 国家自然科学基金项目(62562061, 62462067); 云南省专项项目(202403AP140021); 云南省教育厅科学研究项目(2025J0006, 2024J0010, 2025J0007)
This work was supported by the Yunnan Fundamental Research Project (202401AT070474, 202501AU070059), the National Natural Science Foundation of China (62562061, 62462067), the Yunnan Province Special Project (202403AP140021), and the Yunnan Provincial Department of Education Science Research Project (2025J0006, 2024J0010, 2025J0007).
通信作者: 周维(zwei@ynu.edu.cn)

问题。尽管数据加密能确保安全性,但其独特的格式可能引起怀疑。隐写术通过将秘密数据隐藏在普通媒体中,如图像、文本和音频,提供了一种隐蔽的替代方案,降低了被检测的风险。传统基于像素编辑的隐写方法易被隐写分析或压缩干扰所破译,而现有的生成式隐写技术多依赖针对性提示词或特定数据源,难以实现高泛化与灵活隐蔽。针对上述挑战,提出一种基于双扩散模型的未知源-目标域生成式图像隐写方法。该方法首先利用潜在扩散模型(latent diffusion model, LDM)对任意来源的秘密图像进行反演与重构优化,以获取高保真潜在表示。为了破坏 LDM 模型反演过程残留的结构信息对后续图像隐藏和提取的干扰,该方法提出了高斯结构重排对齐模块,通过伪随机正交扰动与空间重排对秘密图像潜在表示进行加密嵌入。该模块在保持秘密潜在表示的模长不变的同时,使其结构上更加服从高斯分布,对齐后续模块的输入,同时提高了安全性。最后,借助去噪扩散隐式模型(denoising diffusion implicit models, DDIM)对加密后的潜在表示进行去噪重采样,生成兼具目标域外观与秘密信息的含密图像。实验证明,所提方法无需针对特定源域训练,具备良好的跨域通用性;在 4 个公开数据集上,与多种对比方法相比,在秘密图像提取精度、含密图像不可感知度及抗隐写分析能力方面均取得了显著提升。

关键词 图像隐写;图像生成;扩散模型;高泛化性;分布迁移

中图法分类号 TP311.13; TP309

DOI: 10.7544/issn1000-1239.202550348 **CSTR:** 32373.14.issn1000-1239.202550348

在当今数字化时代,社交网络呈现出极为迅猛的发展态势。各类社交媒体平台不断涌现,用户规模急剧膨胀,这使得海量的数据持续且高速地通过公共信道进行传输。然而,与此同时,数据传输过程中所引发的安全性问题和隐私问题变得愈发严峻,已然在学术界、企业界以及广大用户群体中引发了广泛而高度的关注。为确保数据安全,数据加密技术成为了一种广泛运用的手段。在数据加密过程中,待加密数据会被转化为密文形式。然而,密文因其特殊的格式,往往与正常数据存在显著差异,这种差异在公共信道中传输时,极易引起恶意窃听者的注意,从而让数据面临被窃取和破解的风险。考虑到加密技术存在的局限性,数字隐写技术作为一种极具创新性的方法应运而生。数字隐写技术巧妙地将秘密数据隐藏在普通媒介之中,包含但不限于图像[1]、文本[2]以及音频[3]等。通过这种方式,秘密数据能够以一种看似普通信息的形态在公共信道中传输,极大地降低了被窃听者怀疑的可能性,进而有效地提高了数据传输的安全性。数字隐写技术在诸如军事通信、商业机密传输以及个人隐私保护等众多对数据安全要求极高的领域,都展现出了巨大的应用潜力和价值。

早期研究主要集中在基于编辑的范式[4],该范式通过修改封面图像的像素值来隐藏秘密数据。此外,随着深度学习的兴起,基于载体图像嵌入的图像隐写方法引起广泛的研究[5]。然而,这些方法生成的隐写图像通常与正常图像存在分布差异,使其容易被隐写分析技术检测。

为了增强安全性和鲁棒性,随着强大生成模型的兴起,研究逐渐转向生成式隐写技术。CRoSS[6]提出了一种基于图像编辑技术的生成式图像隐写方法,该方法首先对秘密图像进行描述,然后在描述中替换目标提示词,并利用新的提示词通过扩散模型生成隐写图像。该方法保证了隐写图像的可控性。然而,该方法依赖于秘密图像提示词的准确性,因此仅适用于部分对象。此外,由于仅替换了提示词,生成的隐写图像与秘密图像仍然具有一定的相似度,具有泄露秘密信息的潜在风险。Yang 等人[7]提出了一种通过结构提示的基于扩散模型的图像隐写方法,该方法相较于 CRoSS 方法不需要精准的提示词,不过局限性在于依然无法处理任意源的图像。

基于上述分析,我们分析出生成式图像隐写技术面临以下 2 个问题:

1)生成式隐写方法如何实现高泛化能力,即不依赖于特定的秘密图像;

2)如何提高含密图像的灵活性,从而提高隐写算法的隐蔽性。

针对上述问题,本文提出了一种基于扩散模型的未知源-目标域的图像隐写方法,能够对未知来源的图像进行目标域的伪装。该方法基于双扩散模型进行图像隐写,通过潜在扩散模型(latent diffusion model, LDM)[8]对秘密图像进行反演处理,从而获得图像的潜在表示。本文提出了一种潜在空间的高斯结构重排对齐模块,该模块通过密钥种子生成高斯

噪声，并对含密噪声的潜在表示进行正交分解、空间打散和随机重排，从而实现对图像潜在信息的安全重嵌。随后，该框架利用去噪扩散隐式模型（denoising diffusion implicit models, DDIM）[9]对该潜在表示进行重采样解码，以生成包含秘密信息的隐写图像。本文的核心贡献点有3个方面：

1）提出了一种无需针对特定数据集进行专门训练且可对未知源输入图像进行目标域隐藏的生成式图像隐写方法；

2）提出了一种潜在空间的高斯结构重排对齐模块来处理基于扩散模型的图像生成式隐写中存在的高斯对齐问题；

3）在4个数据集上进行了广泛的实验，验证了本文所提方法在秘密图像提取精确度、含密图像不可感知度和抵御隐写分析安全性上具有明显优势。

1 相关工作

1.1 基于嵌入式的图像隐写

早期的图像隐写技术主要通过修改载体图像的像素来实现信息隐藏，典型方法如最低有效位替换（least significant bit replacement, LSBR）[4]。该类方法直接操作像素值，因破坏图像一阶统计特性而易于被检测。随着深度学习的发展，生成对抗网络（generative adversarial network, GAN）成为隐写领域的重要工具[5]。Baluja[10]较早提出了一种基于深度神经网络的端到端图像隐写框架，通过同时训练的编码器-解码器对将秘密图像嵌入同尺寸载体中，实现了完整尺寸图像的端到端隐藏。随后，Wu等人[11]提出了StegNet模型，该模型在网络中引入指数线性单元（exponential linear unit, ELU）激活函数和批归一化层，有效克服了早期方法中隐藏图像视觉质量较差的问题。Zhu等人[12]提出HiDDeN，构建了由编码器、解码器和判别器组成的隐藏框架，并在训练阶段模拟JPEG压缩与裁剪噪声，显著提升了在多种失真通道下的鲁棒性及安全性。Jing等人[13]提出HiNet，第1次将可逆神经网络（invertible neural networks, INNs）[14]引入图像隐写，利用网络的无损映射特性，实现了秘密图像的完全无损恢复，同时通过信息瓶颈设计兼顾了容器图像质量与隐写容量。基于图像风格迁移理念，ISTNet[15]被提出用于将载体图像转换为具有完全不同风格的隐写图像，从而进一步提升隐蔽性。此外，研究者还尝试将SqueezeNet深度学习网络应用于空间域隐写，通过轻量化网络结构在保持较高视觉质量的同时实现信息隐藏[16]。Cui等人[17]提出了多任务身份感知图像隐写（multi-task identity-aware image steganography, MIAIS），该方法无需恢复秘密图像，即可对容器图像进行直接身份识别，进一步拓展了隐写技术在图像分析中的应用场景。

Volkhonskiy等人[18]率先提出了基于GAN的隐写方法SGAN，为后续将GAN引入图像隐写领域奠定了基础。为了提升隐写图像的视觉质量，Shi等人[19]基于SGAN构建了SSGAN模型，通过改进判别器结构与损失函数设计，实现了更逼真的隐写效果。在此基础上，王耀杰等人[20]进一步提出了Stego-WGAN，通过Wasserstein距离优化隐写安全性，增强了对抗分析器的鲁棒性。在此之后，Tang等人[21]提出了一种基于嵌入代价学习的新框架SPAR-RL，通过强化学习自适应地优化嵌入代价分布，进一步提升了隐写容量与抗检测性能。

1.2 生成式图像隐写

生成式隐写术[22]作为新兴方向，突破了传统嵌入式方法依赖载体图像的限制，可直接生成与秘密信息相关的隐秘图像，并支持后续信息提取。Tancik等人[23]提出StegaStamp，将任意超链接比特串稳健地编码进照片中，网络在模拟打印、拍照等失真后仍能可靠提取。

随着扩散模型的兴起，研究者开始探索基于扩散模型的生成式隐写框架。Yu等人[6]提出的CRoSS方法是该领域的代表性工作，其借助扩散模型的可逆编辑能力，通过多模态语义模型（如BLIP[24]）提取秘密图像的语义特征，经局部语义修改后，利用DDIM逆过程将图像映射至潜在高斯空间，最终结合文本提示生成可控的隐秘图像。该方法展现了扩散模型在隐写任务中的鲁棒性与抗检测潜力，但其局限在于对生成内容的控制能力有限，仅能对秘密图像的主体内容进行局部调整，难以实现完全脱离原始语义的自由生成，限制了隐写图像的多样性与应用场景。近期，Yang等人[7]提出的Diffstega则通过引入密码学安全的参考图像与“噪声翻转”技术，进一步增强了无载体生成式隐写的通用性。该方法基于ControlNet[25]以秘密图像的结构信息作为提示，生成含密图像。I2IStega[26]方法致力于解决在扩散模型潜空间中秘密图像潜在分布与高斯分布不一致所导致的含密图像生成质量差、秘密图像还原精度低的问题。I2IStega方法利用减均值、除方差的方式将秘密图像潜在分布与高斯分布靠齐，并随机打乱潜在分布从而与高斯分布更加相似。

1.3 扩散模型

扩散模型(diffusion models, DM)是一类生成模型,近年来在图像生成等领域取得了显著成果。其基本思想是通过逐步向数据添加噪声,将数据分布转换为标准正态分布,然后学习去噪过程,实现从噪声生成数据的能力。在训练阶段,DDIM 通常包括加噪和去噪 2 个过程。在加噪过程中,扩散模型逐步地向原始图像$\boldsymbol{x}_0$添加高斯噪声,经过T次迭代后,得到一个加入噪声的图像$\boldsymbol{x}_{\mathrm{T}}$。这个前向扩散过程能够显式地根据前一步计算得到的$\boldsymbol{x}_{t-1}$来计算当前时刻的$\boldsymbol{x}_t$。

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})=\mathcal{N}\left(\boldsymbol{x}_t;\sqrt{1-\beta_t}\boldsymbol{x}_{t-1},\beta_t\boldsymbol{I}\right),\tag{1}$$

其中$\beta_t\in(0,1)$表示协方差,$\sqrt{1-\beta_t}$表示所加高斯分布的均值项。当$\lim\limits_{T\to\infty}$时,最终的$\boldsymbol{x}_{\mathrm{T}}$将近似服从一个各向同性的高斯分布。

去噪过程的目标是系统性地去除$\boldsymbol{x}_t\in\mathcal{N}(0,\boldsymbol{I})$中的噪声,以生成清晰的图像。在 DDPM 中,这一过程构成一个马尔可夫过程。相反,DDIM 将去噪过程转化为非马尔可夫形式,以加速采样。潜向量的下一个时刻$\boldsymbol{x}_{t-1}$可通过式(2)计算为

$$\boldsymbol{x}_{t-1}=\sqrt{\alpha_{t-1}}\left(\frac{\boldsymbol{x}_t-\sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{x}_t,t)}{\sqrt{\alpha_t}}\right)+\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(\boldsymbol{x}_t,t)+\sigma_t\epsilon_t,\tag{2}$$

其中$\sigma_t^2=\eta\cdot\sqrt{\dfrac{(1-\alpha_{t-1})\left(1-\dfrac{\alpha_t}{\alpha_{t-1}}\right)}{1-\alpha_t}}$,$\epsilon_\theta$表示一个经过充分训练的神经网络,$\alpha_t=1-\beta_t$为单步系数。显然,当$\eta=0$时,整个生成过程是一个不含随机噪声的确定性过程。当$\boldsymbol{x}_{\mathrm{T}}$是确定性的,反向过程同样也是确定性的。

由于 DDIM 的确定性生成过程可以进一步视为对常微分方程(ordinary differential equations, ODE)求解的欧拉积分,因此 DDIM 的反向过程可以将 1 张真实图像$\boldsymbol{x}_0$反向映射为其对应的$\boldsymbol{x}_{\mathrm{T}}$。也就是说,DDIM 的反向过程本质上是“反向扩散过程的反过程”。给定一张真实图像$\boldsymbol{x}_0$,我们可以通过 DDIM 的反向过程以迭代方式获得对应的$\boldsymbol{x}_{t+1}$为

$$\boldsymbol{x}_{t+1}=\sqrt{\bar{\alpha}_{t+1}}\left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}}-\sqrt{\frac{1}{\bar{\alpha}_{t+1}}}\right)\boldsymbol{x}_t+\left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}}-1}-\sqrt{\frac{1}{\bar{\alpha}_t}-1}\right)\epsilon_\theta(\boldsymbol{x}_t,t)\right]+\boldsymbol{x}_t,\tag{3}$$

其中$\bar{\alpha}_t=\prod\limits_{s=1}^{t}\alpha_s$为累积系数。

2 高泛化性生成式图像隐写方法

2.1 方法概述

本文提出的方法框架图如图 1 所示,描述了隐写框架中秘密图像发送方和接收方的整体流程。其中,该框架包含了 LDM 和 DDIM 的 2 个扩散模型用于秘密图像的隐藏和提取。

具体而言,我们在 LDM 的反演过程中引入了一个重构优化模块,通过迭代优化潜在表示$\boldsymbol{x}_{\mathrm{latent}}$,使得经过变分自编码器(variational autoencoder, VAE)重建后的秘密图像$\boldsymbol{x}_{\mathrm{extract}}$在像素空间中尽可能逼近原始秘密图像$\boldsymbol{x}_{\mathrm{sec}}$,从而有效降低由 VAE 重建误差所引起的秘密图像提取失真。在完成潜在向量的反演与优化后,本文进一步设计了一个高斯结构重排对齐模块,用于提升潜在噪声的安全性和嵌入的稳健性。该模块基于密钥种子生成伪随机噪声,并通过正交

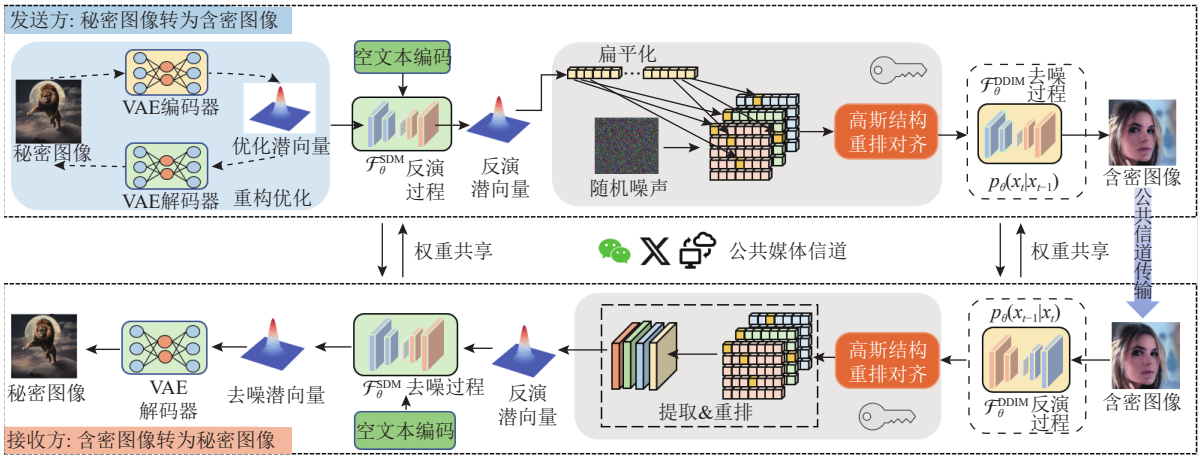


Fig. 1 Overall framework for generative image steganography with unknown source-target domains

图 1 未知源-目标域生成式图像隐写方法整体框架

矩阵分解与空间重排操作，对潜在表示进行打乱和加密，以增强系统的安全性。

在完成上述信息嵌入后，使用 DDIM 的去噪过程对加密后的潜在表示进行采样，最终生成具有指定视觉内容的含密图像。该图像不仅保留了目标域图像的外观风格，同时隐式地嵌入了秘密图像信息。在接收端，含密图像通过共享权重的扩散模型进行潜在反演与解码，借助对称的重排机制和重构优化模块，最终还原出原始的秘密图像$\boldsymbol{x}_{\text{sec}}$。

2.2 信息隐藏

信息隐藏阶段，本文提出基于 LDM 构建的图像隐写机制，以实现秘密图像在公开媒介图像中的无感知嵌入。该阶段的目标是将秘密图像$\boldsymbol{x}_{\text{sec}} \in \mathbb{R}^{H\times W\times 3}$嵌入至扩散模型的潜在空间中，最终生成视觉自然的含密图像$\boldsymbol{x}_{\text{stego}} \in \mathbb{R}^{H\times W\times 3}$。

首先，使用训练好的 VAE 编码器$Enc(\cdot)$对秘密图像进行隐空间编码，得到其潜空间表示为

$$\boldsymbol{x}_{\text{latent}} = Enc(\boldsymbol{x}_{\text{sec}}) \text{。} \tag{4}$$

接着，将该潜在表示$\boldsymbol{x}_{\text{latent}}$输入至 LDM 中的反向扩散路径中，以构建与之对应的潜在噪声表示$\boldsymbol{x}_{\text{sec}}$。我们采用空文本条件$c=\varnothing$并通过 LDM 的反向过程$\mathcal{F}_{\text{SDM}}^{-1}(\cdot)$获得

$$\boldsymbol{x}_{\text{latent}} = \mathcal{F}_{\text{SDM}}^{-1}(\boldsymbol{x}_{\text{sec}}, c) \text{。} \tag{5}$$

为了提升潜变量在扩散空间的嵌入安全性并消除潜空间非高斯结构关系，我们进一步将$\boldsymbol{x}_{\text{sec}}$与一组高斯噪声样本$\boldsymbol{n}_{\text{rand}} \sim \mathcal{N}(0, \boldsymbol{I})$进行联合重排与扰动。该扰动过程结合了密钥种子生成的正交矩阵来构造高斯结构扰动，并在空间上打散潜变量（详见 2.5 节）。

扰动操作可形式化为

$$\tilde{\boldsymbol{x}}_{\text{latent}} = \mathcal{R}(\boldsymbol{x}_{\text{latent}}, \boldsymbol{n}_{\text{rand}}; K) \text{，} \tag{6}$$

其中$\mathcal{R}(\cdot)$表示高斯结构扰动与空间重排操作，K为密钥控制的种子参数。

最终，我们利用 DDIM 模型的去噪过程将扰动后的潜变量$\tilde{\boldsymbol{x}}_{\text{latent}}$还原为图像域中的含密图像$\boldsymbol{x}_{\text{stego}}$：

$$\boldsymbol{x}_{\text{stego}} = \mathcal{F}_{\text{DDIM}}(\tilde{\boldsymbol{x}}_{\text{latent}}) \text{。} \tag{7}$$

通过上述过程，秘密图像信息被安全地嵌入至$\boldsymbol{x}_{\text{stego}}$中，并具备良好的图像自然性与目标域适应性。该方法无需依赖目标域图像的训练过程，因此具备较强的跨域通用性，能够在不同图像内容和风格下实现稳健的隐写操作。

2.3 信息提取

在接收端，信息提取过程的目标是从含密图像$\boldsymbol{x}_{\text{stego}}$中准确提取出秘密图像$\boldsymbol{x}_{\text{extract}}$。提取过程的第 1 步是对含密图像进行去噪反演，通过 DDIM 模型$\mathcal{F}_{\text{DDIM}}^{-1}(\cdot)$执行反向扩散操作，得到其潜空间隐表示为

$$\tilde{\boldsymbol{x}}_{\text{latent}} = \mathcal{F}_{\text{DDIM}}^{-1}(\boldsymbol{x}_{\text{stego}}) \text{。} \tag{8}$$

该过程基于与信息隐藏阶段共享的噪声预测器和调度参数，确保解码路径与编码路径的一致性。随后，与前文相似的利用嵌入阶段使用的高斯结构重排对齐模块的逆操作，依据共享的密钥种子（key seed）和重排矩阵恢复原始特征的空间顺序，逆向还原为初始隐空间特征$\boldsymbol{x}_{\text{latent}}$。为了确保顺序还原正确，我们设计了如下恢复函数为

$$\boldsymbol{x}_{\text{latent}} = \mathcal{R}^{-1}(\tilde{\boldsymbol{x}}_{\text{latent}}, \boldsymbol{n}_{\text{rand}}; K) \text{，} \tag{9}$$

其中，密钥种子K控制高斯采样与特征打散顺序，确保接收方与发送方在没有通信的情况下共享特征重排规则。

最后，隐空间特征$\boldsymbol{x}_{\text{latent}}$被输入到 LDM 的去噪过程及 VAE 解码器$Dec(\cdot)$以重建出秘密图像为

$$\boldsymbol{x}_{\text{extract}} = Dec(\mathcal{F}_{\text{DDIM}}(\boldsymbol{x}_{\text{latent}}, c)) \text{，} \tag{10}$$

其中文本引导条件$c=\varnothing$。

2.4 重构优化模块

由于 Stable Diffusion 中的 VAE 在编码秘密图像$\boldsymbol{x}_{\text{sec}}$到潜空间表示$\boldsymbol{x}_{\text{latent}}$，再由解码器重建图像的过程中，往往会引入重构误差，造成最终提取图像$\boldsymbol{x}_{\text{extract}}$与原始秘密图像之间存在差异。因此，本文提出重构优化模块，以显式优化潜空间表示$\boldsymbol{x}_{\text{latent}}$，使其在扩散解码后能更好地重建出高保真的秘密图像。

给定原始秘密图像$\boldsymbol{x}_{\text{sec}}$，我们的目标是寻找一个优化后的隐空间表示$\boldsymbol{x}_{\text{latent}}$，使得经过 VAE 解码器$Dec(\cdot)$解码后的图像与$\boldsymbol{x}_{\text{sec}}$的差异最小。我们定义以下重构优化目标函数为

$$\tilde{\boldsymbol{x}}_{\text{latent}} = \arg\min_{\boldsymbol{x}_{\text{latent}}} \mathcal{L}_{\text{rec}}(Dec(\boldsymbol{x}_{\text{latent}}), \boldsymbol{x}_{\text{sec}}) \text{，} \tag{11}$$

其中$\mathcal{L}_{\text{rec}}$为重构损失，目的是计算解码图像和原始图像之间的距离。在本文中，我们采用均方误差为

$$\mathcal{L}_{\text{rec}} = \|Dec(\boldsymbol{x}_{\text{latent}}) - \boldsymbol{x}_{\text{sec}}\|_2^2 \text{。} \tag{12}$$

具体实现中，我们采用梯度下降优化策略，在冻结 VAE 解码器$Dec(\cdot)$权重的前提下，仅对$\boldsymbol{x}_{\text{latent}}$进行迭代更新为

$$\boldsymbol{x}_{\text{latent}}^{(t+1)} = \boldsymbol{x}_{\text{latent}}^{(t)} - \eta \cdot \nabla_{\boldsymbol{x}_{\text{latent}}^{(t)}} \mathcal{L}_{\text{rec}}\left(Dec\left(\boldsymbol{x}_{\text{latent}}^{(t)}\right), \boldsymbol{x}_{\text{sec}}\right) \text{，} \tag{13}$$

其中η为学习率，最终优化得到的$\tilde{\boldsymbol{x}}_{\text{latent}}$将替代原始潜表示用于后续信息嵌入过程，从而增强图像隐写系统的保真性和提取鲁棒性。

2.5 高斯结构重排对齐模块

为消除 LDM 反演残留的原始结构特征对后续

DDIM 去噪的干扰, 并在模型参数泄露时依然保证隐写通信安全, 本文设计了高斯结构重排对齐模块, 如图 2 所示。该模块通过密钥驱动的可逆正交扰动与向量重排, 在严格保持潜在表示的模长与高斯分布性质不变的前提下, 对隐空间特征进行加密式处理。

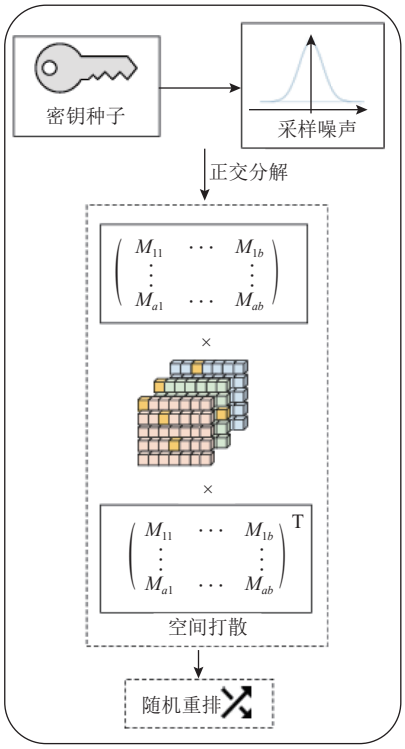


Fig. 2 Gaussian structure reordering alignment module

图 2 高斯结构重排对齐模块

给定 VAE 编码得到的潜在表示 $\boldsymbol{x}_{\text{latent}} \in \mathbb{R}^{H\times W\times 4}$, 依据共享的密钥种子 K 生成正交扰动核与重排顺序。以 K 为种子构造高斯矩阵 $\boldsymbol{H}$ 并进行 QR 分解为

$$\boldsymbol{H} \sim \mathcal{N}(0,\boldsymbol{I})^{H\times W},\ \boldsymbol{H}=\boldsymbol{Q}\boldsymbol{R},\ \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}=\boldsymbol{I}\ 。 \tag{14}$$

之后, 对每个通道 $\boldsymbol{x}_{\text{latent}}$ 在空间维度上应用正交变换以重排残留结构而保持每个通道特征的模长与高斯性为

$$\boldsymbol{x}'=\boldsymbol{Q}\boldsymbol{x}_{\text{latent}}\boldsymbol{Q}^{\mathrm{T}}\ 。 \tag{15}$$

为了进一步破坏原始图像在潜在空间中的结构连贯性, 本文在正交扰动后引入基于密钥的向量级随机重排操作。将经双向正交扰动后的潜在特征 $\boldsymbol{x}'\in\mathbb{R}^{H\times W\times 4}$ 展平为向量形式 $\boldsymbol{v}'=Flatten(\boldsymbol{x}')$。随后, 基于密钥种子 K, 确定一组置换索引 $\pi=Perm(K)$, π 是 $\{1,2,\cdots,D\}$ 的一个全排列。利用该置换索引对 $\boldsymbol{v}'$ 进行重排, 得到打乱后的潜在向量 $\boldsymbol{v}_{\text{shuffled}}$, 并对该潜向量进行重新排序, 恢复为图像潜在表示的空间形状为

$$\boldsymbol{v}_{\text{shuffled}}=\boldsymbol{v}'[\pi]=\left(\boldsymbol{v}'_{\pi(1)},\boldsymbol{v}'_{\pi(2)},\cdots,\boldsymbol{v}'_{\pi(D)}\right)^{\mathrm{T}}, \tag{16}$$

$$\tilde{\boldsymbol{x}}_{\text{latent}}=reshape(\boldsymbol{v}_{\text{shuffled}},C,H,W)\ 。 \tag{17}$$

在接收端, 可按相同顺序对 $\tilde{\boldsymbol{x}}_{\text{latent}}$ 执行逆置换与逆正交的过程 $\mathcal{R}^{-1}$, 从而得到重排前的潜空间表示 $\boldsymbol{x}_{\text{latent}}$。该模块确保重排后的 $\tilde{\boldsymbol{x}}_{\text{latent}}$ 在模长服从高斯分布的同时, 重排打散了 LDM 所保留的结构特征, 对齐了 DDIM 的输入分布, 保证后续 DDIM 去噪和反演过程的准确度。

3 实验结果

3.1 实验设置

1)实验细节

本实验基于预训练模型 Stable Diffusion v1.5 与 DDIM 构建隐写框架, 设置扩散步长 T=50, 采样器为 DPM-Solver Multi-step Scheduler [27], 调度器参数 β 开始值为 0.000 85, 结束值为 0.012, 调度方法为“scaled_linear”。重构优化中, 采用 Adam 优化器, 初始学习率为 0.000 7, 每个批次图像的优化轮次为 200, 每个轮次用时约 0.215 s。对比的基线方法包括近 2 年最先进的以图像为秘密信息载体的生成式图像隐写方法 CRoSS, Diffstega, I2IStega。所有实验在单卡 RTX 4090 GPU 上完成, 无需额外模型参数训练。

2)评价指标

图像质量指标包括峰值信噪比(peak signal-to-noise ratio, PSNR); 学习感知图像块相似度(learned perceptual image patch similarity, LPIPS)用来评估语义级一致性(值越低越好); 自然图像质量评价器(naturalness image quality evaluator, NIQE)用来衡量生成图像的自然性(值越低越接近真实分布)。

安全性指标为基于 Yedroudj-Net[28], XuNet[29], DFNet[30] 的检测错误率(接近 50% 表明安全性高), 包括隐写分析检测率。

3)数据集

我们的评估在 4 个标准数据集上进行: Bedroom[31], CelebA[32], Church[31], ImageNet[33]。为了验证方法的泛化性, 我们对比了所有方法在不经过任何训练的情况下, 测试不同域(秘密图像与含密图像所属域不同)下的隐写性能。我们在每一个数据集中随机挑选了 1 000 张样本作为秘密图像来评估隐写方法, 以充分评估隐写方法的隐藏能力。

3.2 图像提取精确度对比实验结果

表 1 展示了分别以 Church, CelebA, Bedroom 数据集中的样本作为秘密图像, 并分别将其隐藏为 Church, CelebA, Bedroom, ImageNet 目标域图像的隐

Table 1 Experimental Results of Open Image Target Domain Steganography

表 1 开放图像目标域隐写实验结果

方法	指标	Bedroom			CelebA			Church		
		ImageNet	Church	CelebA	ImageNet	Bedroom	Church	ImageNet	Bedroom	CelebA
CRoSS	PSNR/dB	15.971 0	15.273 8	17.383 4	16.546 2	14.109 9	14.537 0	14.519 5	15.125 8	15.782 3
Diffstega		21.640 1	17.595 0	**29.314 0**	21.936 0	20.173 3	17.640 9	21.839 7	20.139 4	**29.288 5**
I2IStega		21.116 8	23.223 6	24.199 8	21.902 4	24.595 9	**25.129 9**	22.393 6	24.669 8	25.838 4
本文		**27.139 4**	**28.113 2**	28.692 1	**25.793 6**	**25.280 0**	24.858 0	**25.294 6**	**27.120 0**	28.505 8
CRoSS	LPIPS	0.305 5	0.279 2	0.251 9	0.345 4	0.359 5	0.369 8	0.372 9	0.353 4	0.321 0
Diffstega		0.142 8	0.198 9	**0.060 5**	0.132 4	0.164 0	0.196 7	0.133 1	0.163 1	**0.058 2**
I2IStega		0.144 8	0.093 8	0.145 7	**0.122 9**	0.090 7	**0.059 3**	**0.110 8**	**0.090 0**	0.105 0
本文		**0.110 3**	**0.090 3**	0.074 6	0.128 2	**0.160 0**	0.129 4	0.168 0	0.153 8	0.101 5

注：Bedroom，CelebA，Church 为实验所用的来源域图像数据集；ImageNet，Church，CelebA，Bedroom 为对应目标域图像数据集。黑体数值表示最优值。

写结果。其中 PSNR 值越大表示原始秘密图像和提取的秘密图像之间的差异越小，LPIPS 值越小表示提取到的秘密图像和原始图像在特征上的差异越小。在图像级别，由于 CRoSS 方法所采用的文本与图像一致性较差，导致在公开数据集上的可逆性较差，秘密图像的还原精度较低。其次，尽管 Diffstega 方法使用了额外的 ControlNet 和提示进行图像的提取，其在秘密图像来源较为复杂的数据集（如 ImageNet, Church, Bedroom）上的可逆性依然不理想。在图像纹理和内容较为复杂的数据集上，本文提出的方法具有更高的像素相似性和特征相似性。

然而，在 CelebA 数据集上，本文方法的表现与 Diffstega 方法相比，相对不是那么理想。原因是 CelebA 数据集主要由人脸图像组成，这些图像的纹理结构较为规整且包含丰富的人脸特征。相比之下，Diffstega 方法通过引入 ControlNet 和特定的人脸提示信息，能够更好地提取秘密图像，因此在该数据集上具有天然优势。与此不同，本文方法并未使用人脸先验或关键点约束，导致在该数据集上的恢复精度略低于 Diffstega。

3.3 含密图像抗隐写分析安全性对比实验结果

为了评估隐写方法的安全性，我们利用 3 个隐写分析器 XuNet[29], YeNet[28], DFNet[30] 来对封面和隐写图像进行分类。对于每一个目标域数据集，我们分别基于每一种隐写方法生成 1 000 张的含密图像和 1 000 张非含密图像。3 个目标域数据集上的安全性能如表 2 所示。检测错误率（probability of error, PE）波动在 0.5 左右，表明隐写分析器无法正确分类载体和隐写图像载体图像。结果表明，我们的高斯结构重排对齐模块的设计理论上确保生成含密图像的输入非常接近原始高斯分布，从而生成的含密图像与正常的图像具有极其一致的视觉效果。此外，我们绘制了 ROC 曲线进行检测。在隐写术中，最佳 ROC 曲线与反对角线对齐（如图中的虚线所示），理想的 AUC 值为 0.5。如图 3 所示，我们的 ROC 曲线的 AUC 值非常接近理想结果，这表明我们的 CDIS 方案表现出高安全性。

3.4 含密图像无参考不可感知性对比实验结果

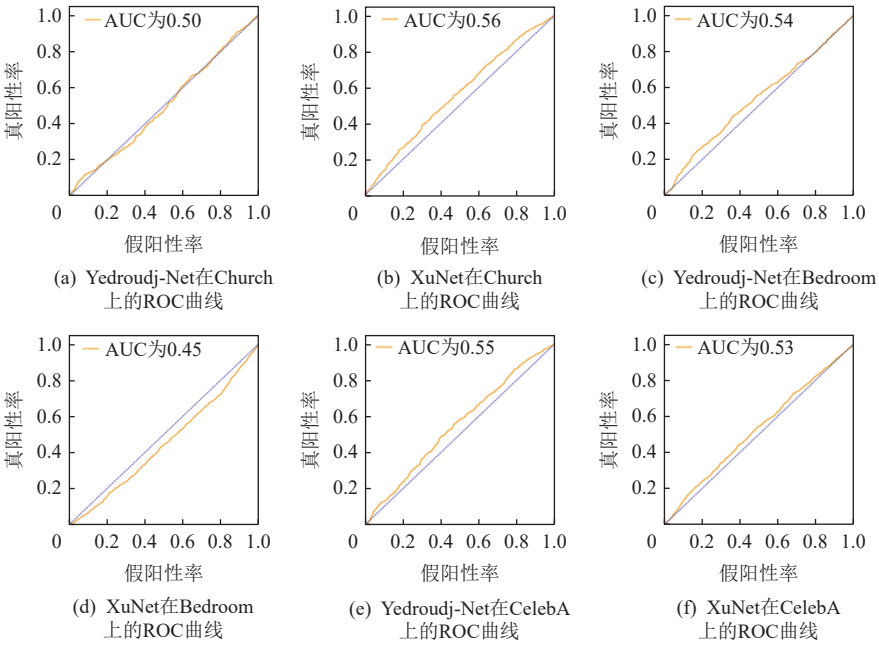
为验证不同隐写方法生成的含密图像与自然图像的统计特性一致性，我们基于 NIQE 指标对本文方法和基线方法在 Church, Bedroom, CelebA 数据集上的表现进行量化对比。如表 3 所示，NIQE 值越低，表明含密图像的统计分布越接近自然图像，其视觉隐蔽性越强。其中，Vanilla 是以标准的高斯噪声进

Table 2 The PE Results of Target Steganographic Security Experiment

表 2 目标域隐写安全性实验 PE 结果 %

方法	Yedroudj-Net			XuNet			DFNet		
	Church	Bedroom	CelebA	Church	Bedroom	CelebA	Church	Bedroom	CelebA
CRoSS	82.80	60.10	92.83	73.50	72.01	90.30	93.60	91.88	96.10
Diffstega	58.10	65.20	78.20	62.15	81.20	55.60	95.30	84.90	91.50
I2IStega	73.23	63.81	82.33	80.51	83.01	86.30	89.62	92.25	88.49
ablation	54.95	47.20	43.15	61.60	92.15	87.65	91.60	89.85	83.80
本文	50.20	49.70	49.95	50.6	45.15	50.04	58.15	53.90	55.90

注：ablation 代表我们的方法消融高斯结构重排对齐模块的结果。



(a) Yedroudj-Net在Church上的ROC曲线 (b) XuNet在Church上的ROC曲线 (c) Yedroudj-Net在Bedroom上的ROC曲线

(d) XuNet在Bedroom上的ROC曲线 (e) Yedroudj-Net在CelebA上的ROC曲线 (f) XuNet在CelebA上的ROC曲线

Fig. 3 ROC curve of steganalysis

图 3 抗隐写分析 ROC 曲线

行采样来生成的非含密图像。可以看出，由于本文提出的高斯结构重排对齐模块能够将包含秘密信息的隐空间向量很好地隐藏进高斯噪声中，本文方法生成的含密图像与正常生成的图像具有很高的统计特性一致性，确保了隐写方法不被轻易地察觉。

图 4 展示了以不同数据集的图像作为秘密图像源并将其隐藏进不同目标域的可视化结果。该结果展示了本文方法的良好泛化性，能够将不同的图像伪装成目标域的图像。同时，灵活地选择隐藏的目标域可以提高隐写方法的安全性。例如，将秘密图



Fig. 4 Visualization results of the proposed method on different source domain and target domain

图 4 本文所提方法在不同来源域和目标域的可视化结果

Table 3 Naturalness Metrics NIQE of the Stego Images

表 3 含密图像自然性指标 NIQE

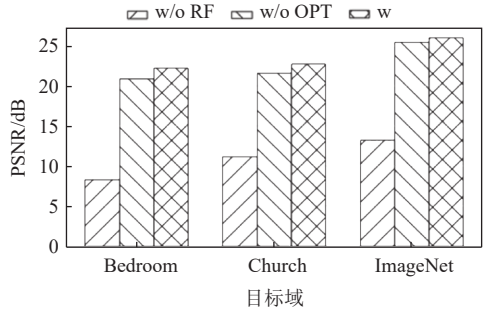
方法	含密图像自然性指标 NIQE		
	Church	Bedroom	CelebA
CRoSS	5.866 1	4.692 8	4.635 5
Diffstega	5.137 3	5.074 7	4.553 4
I2IStega	**4.134 3**	4.473 1	5.600 0
Vanilla	4.767 5	**4.426 8**	**3.533 9**
本文	4.603 2	4.584 8	3.697 3

注：黑体数值表示最优值。

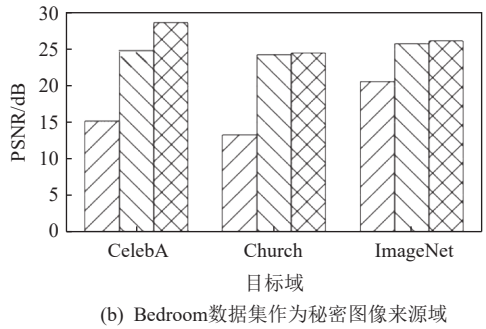
像伪装成人脸、卧室图像和教堂图像，能够不引起第三方的怀疑，从而使含密图像能够顺利地在公共信道上传输，进一步提高了隐写方法的安全性。

3.5 消融实验结果

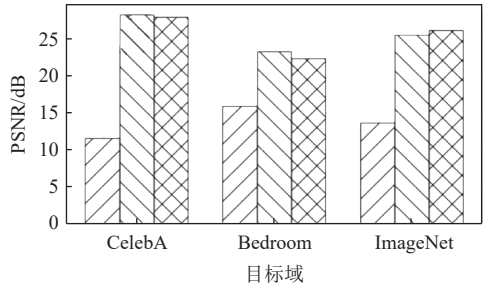
图 5 展示了对本文所提框架每个部分消融的提取秘密图像 PSNR 实验结果。其中，w/o RF 为消融高斯结构重排对齐模块的结果，w/o OPT 为消融重构误差的结果，w 为整体本文所提的框架。可以看出，高斯结构重排对齐模块极大地提高了秘密图像提取的准确度，重构误差能够减少图像在 VAE 中存在的误差，从而进一步提高提取图像和秘密图像的 PSNR 值。



(a) CelebA数据集作为秘密图像来源域



(b) Bedroom数据集作为秘密图像来源域



(c) Church数据集作为秘密图像来源域

Fig. 5 The ablation study results on different source domain and target domain

图 5 在不同来源域和目标域上的消融实验结果

4 结 论

本文针对生成式图像隐写技术中存在的依赖特定提示与源域的局限，提出了一种无需专门训练即可对未知输入图像进行目标域伪装的未知源-目标域生成式隐写框架。通过重构优化模块与高斯结构重排对齐模块的协同作用，实现了潜在空间信息的高保真、安全嵌入与隐蔽传输。大量实验结果表明，本文方法在跨域隐写的泛化能力、含密图像的自然性以及对抗隐写分析的鲁棒性方面均优于现有技术。未来工作可在进一步降低计算复杂度、扩展多模态密文嵌入以及强化隐写容量等方向开展，以推动生成式隐写技术在更广泛的安全通信应用中落地。

作者贡献声明：董云云提出本文的生成式图像隐写方法，完成了该方法的相关实验，撰写论文；张锦弘参与了部分论文写作，进行了 CRoSS 相关对比实验与消融实验；李钒效进行了隐写安全性实验、含密图像自然性分析实验；朱波进行了 I2IStega, Diffstega 相关对比实验；周维修改论文。

参 考 文 献

[1] Zhang Xuanbo, Zhang Xinpeng, Feng Guorui. Image steganalysis network based on dual-attention mechanism[J]. IEEE Signal Processing Letters, 2023, 30: 1287−1291

[2] Li Fanxiao, Wu Sixing, Yu Jiong, et al. Rewriting-Stego: Generating natural and controllable steganographic text with pre-trained language model[C]//Proc of the 28th Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2023: 617−626

[3] Wu Junqi, Chen Bolin, Luo Weiqi, et al. Audio steganography based on iterative adversarial attacks against convolutional neural networks[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 2282−2294

[4] Chan C K, Cheng L M. Hiding data in images by simple LSB substitution[J]. Pattern Recognition, 2004, 37(3): 469−474

[5] Liu Jia, Ke Yan, Zhang Zhou, et al. Recent advances of image steganography with generative adversarial networks[J]. IEEE Access, 2020, 8: 60575−60597

[6] Yu Jiwen, Zhang Xuanyu, Xu Youmin, et al. CRoSS: Diffusion model makes controllable, robust and secure image steganography[C]//

Advances in Neural Information Processing Systems 36. Cambridge, MA: MIT, 2023: 80730−80743

[7] Yang Yiwei, Liu Zheyuan, Jia Jun, et al. Diffstega: Towards universal training-free coverless image steganography with diffusion models[C]//Proc of the 33rd Int Joint Conf on Artificial Intelligence (IJCAI−24). New York: IJCAI Organization, 2024: 1579−1587

[8] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR' 2022). Piscataway, NJ: IEEE, 2022: 10684−10695

[9] Song Jiaming, Meng Chenlin, Ermon S. Denoising diffusion implicit models[C]//Proc of the 9th Int Conf on Learning Representations. Washington: OpenReview. net, 2021: 1−20

[10] Baluja S. Hiding images within images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(7): 1685−1697

[11] Wu Pin, Yang Yang, Li Xiaoqiang. StegNet: Mega image steganography capacity with deep convolutional network[J]. Future Internet, 2018, 10(6): 54

[12] Zhu Jiren, Kaplan R, Johnson J, et al. HiDDeN: Hiding data with deep networks[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 657−672

[13] Jing Junpeng, Deng Xin, Xu Mai, et al. HiNet: Deep image hiding by invertible network[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 4713−4722

[14] Gomez A N, Ren M, Urtasun R, et al. The reversible residual network: Backpropagation without storing activations[C]//Advances in Neural Information Processing Systems 31. Cambridge, MA: MIT, 2017: 2214−2224

[15] Bi Xinliang, Yang Xiaoyuan, Wang Chao, et al. High-capacity image steganography algorithm based on image style transfer[J]. Security and Communication Networks, 2021, 2021(1): 4179340

[16] Hamid N, Sumait B S, Bakri B I, et al. Enhancing visual quality of spatial image steganography using SqueezeNet deep learning network[J]. Multimedia Tools and Applications, 2021, 80(28): 36093−36109

[17] Cui Jiabao, Zhang Pengyi, Li Songyuan, et al. Multitask identity-aware image steganography via minimax optimization[J]. IEEE Transactions on Image Processing, 2021, 30: 8567−8579

[18] Volkhonskiy D, Nazarov I, Burnaev E. Steganographic generative adversarial networks[C]//Proc of the 12th Int Conf on Machine Vision (ICMV' 2019). Bellingham, WA: SPIE, 2020, 11433: 991−1005

[19] Shi Haichao, Dong Jing, Wang Wei, et al. SSGAN: Secure steganography based on generative adversarial networks[C]//Proc of the 18th Pacific Rim Conf on Multimedia. Berlin: Springer, 2018: 534−544

[20] Wang Yaojie, Niu Ke, Yang Xiaoyuan. Information hiding scheme based on generative adversarial network[J]. Journal of Computer Applications, 2018, 38(10): 2923−2928 (in Chinese)
(王耀杰, 钮可, 杨晓元. 基于生成对抗网络的信息隐藏方案[J]. 计算机应用, 2018, 38(10): 2923−2928)

[21] Tang Weixuan, Li Bin, Barni M, et al. An automatic cost learning framework for image steganography using deep reinforcement learning[J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 952−967

[22] Fu Zhangjie, Li Enlu, Cheng Xu, et al. Recent advances in image steganography based on deep learning[J]. Journal of Computer Research and Development, 2021, 58(3): 548−568 (in Chinese)
(付章杰, 李恩露, 程旭, 等. 基于深度学习的图像隐写研究进展[J]. 计算机研究与发展, 2021, 58(3): 548−568)

[23] Tancik M, Mildenhall B, Ng R. StegaStamp: Invisible hyperlinks in physical photographs[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 2117−2126

[24] Li Junnan, Li Dongxu, Xiong Caiming, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proc of the 39th Machine Learning Research. New York: PMLR, 2022: 12888−12900

[25] Zhang L M, Rao Anyi, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2023: 3836−3847

[26] Jiang Jingyuan, Wang Zichi, Zhang Xinpeng. Image-to-image steganography based on multimodal generative model[J]. Signal Processing, 2026, 238: 110106

[27] Lu Cheng, Zhou Yuhao, Bao Fan, et al. DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps[C]//Advances in Neural Information Proc Systems 36. Cambridge, MA: MIT, 2022: 5775−5787

[28] Yedroudj M, Comby F, Chaumont M. Yedroudj-Net: An efficient CNN for spatial steganalysis[C]//Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2018: 2092−2096

[29] Xu Guanshuo, Wu Hanzhou, Shi Yunqing. Structural design of convolutional neural networks for steganalysis[J]. IEEE Signal Processing Letters, 2016, 23(5): 708−712

[30] Fu Tong, Chen Liquan, Jiang Yinghua, et al. Image steganalysis based on dual-path enhancement and fractal downsampling[J]. IEEE Transactions on Information Forensics and Security, 2024, 20: 1−16

[31] Yu F, Seff A, Zhang Yinda, et al. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop[J]. arXiv preprint, arXiv: 1506.03365, 2015

[32] Liu Ziwei, Luo Ping, Wang Xiaogang, et al. Deep learning face attributes in the wild[C]//Proc of the IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2015: 3730−3738

[33] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//Proc of the 22nd IEEE Conf Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2009: 248−255



Dong Yunyun, born in 1989. PhD, lecturer. Member of CCF. Her main research interests include image steganography and large model security.

董云云, 1989 年生。博士, 讲师。CCF 会员。主要研究方向为图像隐藏、大模型安全。



Zhang Jinhong, born in 1999. PhD candidate. Student member of CCF. His main research interests include AI model security and multimedia data privacy protection.

张锦弘, 1999 年生。博士研究生。CCF 学生会员。主要研究方向为人工智能模型安全、多媒体数据隐私保护。



Li Fanxiao, born in 1998. PhD candidate. Student member of CCF and IEEE. His main research interests include misinformation detection and multi modal large language models.

李钒效, 1998 年生。博士研究生。CCF 和 IEEE 学生会员。主要研究方向为虚假信息检测、多模态大模型。



Zhu Bo, born in 2001.Master candidate. His main research interests include privacy protection and image steganography.

朱　波, 2001 年生。硕士研究生。主要研究方向为隐私保护、图像隐写。



Zhou Wei, born in 1974. Professor, PhD supervisor. Senior member of CCF. His main research interests include cyberspace security, distributed cloud computing, intelligent information processing, and interdisciplinary research.

周　维, 1974 年生。教授，博士生导师。CCF 高级会员。主要研究方向为网络空间安全、分布式云计算、智能信息处理、交叉学科研究。