计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

生成式隐写研究

周志立[2)] 丁淳[1)] 李进[2)] 彭飞[2)] 张新鹏[3)]

[1)](南京信息工程大学数字取证教育部工程研究中心 南京 210044)

[2)](广州大学人工智能与区块链研究院 广州 510006)

[3)](复旦大学计算机科学技术学院 上海 201203)

摘 要 隐写术通常将秘密信息以不可见的形式隐藏到载体中，从而通过传递含密载体实现隐蔽通信. 嵌入式隐写方案通过修改载体将秘密信息嵌入其中，但会不可避免地改变载体的统计特性，因此难以抵抗各类隐写分析工具的检测. 为了解决此问题，生成式隐写方案以秘密信息为驱动直接生成含密载体. 相比于嵌入式隐写方案，生成式隐写方案针对现有基于统计特征的隐写分析方法具有较好的抗检测性能，因此逐渐成为信息隐藏领域的研究热点. 本文首先对四类生成式隐写方案进行详细地描述和分析，包括: 1) 图像生成式隐写方案，2) 文本生成式隐写方案，3) 音频生成式隐写方案，4) 社交网络行为生成式隐写方案；其次，通过实验详细分析和对比了各种图像和文本生成式隐写方法的性能；最后，本文分析了当前生成式隐写方案仍然存在的问题，并提出相应的解决方案和展望未来的发展方向.

关键词 生成式隐写；信息隐藏；数字水印；隐蔽通信；隐写分析

中图法分类号 TP18

Research on Generative Steganography

ZHOU Zhi-Li[2)] DING Chun[1)] LI Jin[2)] PENG Fei[2)] ZHANG Xin-Peng[3)]

[1)](Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044)

[2)](Institute of Artificial Intelligence and Blockchain, Guangzhou University, Guangzhou 51006)

[3)](School of Computer Science, Fudan University, Shanghai 201203)

Abstract Steganography is a technology that generally hides secret information into a cover invisibly to obtain the stego, so that covert communication can be realized by transmitting the stego. Compared to the technology of encryption, steganography can not only secure the content of secret information but also protect the behavior of covert communication itself. In the past two decades, many multimedia steganographic schemes have been proposed, in which the information embedding-based steganographic methods are the most popular ones. The information embedding-based steganographic schemes usually embed the secret information into the cover with slight modification. However, the modification will inevitably change the statistical features of the cover, which makes it difficult to resist the detection of various steganalysis tools. To address this issue, a new kind of steganographic schemes, *i.e.*, generative steganographic schemes, has been proposed. The generative steganographic schemes usually generate a new multimedia cover as the stego driven by the secret information. Compared with the embedding-based steganographic schemes, the generative steganographic schemes achieve promising anti-detectability to steganalysis, and thus they have become one of hottest research topics in the field

———————————————

本课题得到国家重点研发计划基金资助项目(No. 2022YFB3103100)、国家自然科学基金(No. 61972205, No. U1936218, No. 62122032)和江苏省大气环境与装备技术协同创新中心资助项目(CICAEET)资助. 周志立，博士，教授，硕士生导师，计算机学会(CCF)会员，主要研究领域为信息隐藏、数字取证、多媒体安全、人工智能安全、区块链. E-mail: zhou_zhili@163.com. 丁淳，硕士，主要研究领域为信息隐藏与计算机视觉. E-mail: dylandingc@gmail.com. 李进，博士，教授，主要研究领域为人工智能安全、人工智能博弈、隐私计算和区块链. E-mail: jinli71@gmail.com. 彭飞，博士，教授，主要研究领域为人工智能的安全应用、多媒体安全与保密、工业互联网安全与保密. E-mail: eepengf@gmail.com. 张新鹏(通信作者)，博士，教授，主要研究领域为多媒体信息安全、人工智能安全、图像处理. E-mail: zhangxinpeng@fudan.edu.cn.

of information hiding recently. In this paper, according to the categories of the generated stegos, four kinds of generative steganographic schemes are described, which are: 1) image generative steganography, 2) text generative steganography, 3) audio generative steganography, and 4) social network behavior generative steganography, and the advantages and disadvantages of the four kinds of generative steganographic schemes are also analyzed in detail; Then, by extensive experiments, the performances of recent image and text generative steganographic schemes are analyzed and compared in the aspects of secret information extraction, the robustness against common image attacks, hiding capacity. We also experimentally demonstrate that the stability of stegos generated by the existing generative models is still not good enough; Finally, based on the sufficient analysis of existing generative steganographic schemes, the problems of existing generative steganographic schemes are concluded, and then the corresponding solutions and future research directions are also provided.

Key words generative steganography; information hiding; digital watermarking; covert communication; steganalysis;

1 引言

隐写术(Steganography)通常将秘密信息以不可见的形式隐藏到载体中以实现隐蔽通信的目的. 作为另一种常用的保障数据安全的技术, 数据加密(Encryption)通常将秘密信息加密为无意义的密文形式, 但这也暴露了这些数据的重要性, 使得秘密信息容易遭到第三方的怀疑和拦截. 与加密技术相比, 隐写术不仅可以保护秘密信息的内容安全, 而且可以确保隐蔽通信的行为安全.

最早的隐写术可以追溯到波希战争时期, 一位波斯贵族将消息刻在奴隶的头皮上, 待奴隶的头发长出来后, 把奴隶送到指定地点以隐蔽地传递消息[1]. 在中国古代也有很多关于隐写术的记载, “藏头诗”是其中比较著名的方法, 其将秘密信息写在诗句的特定位置以实现隐蔽通信[2]. 1983 年 Simmons 等[3]提出了经典的“囚徒模型”, 从而奠定了现代隐写术的基础. 如图 1 所示, 该模型将隐写术描述为: 在狱警 Wendy 的监视下, 囚犯 Alice 和 Bob 为谋划越狱建立的一种隐蔽通信方式. 以 Alice 向 Bob 传递消息为例, 根据约定的密钥 K, 在载体 C 中嵌入秘密信息 M 获得含密载体 S, 并通过 Wendy 监控的信道传递给 Bob, 其目的是使得传递消息的行为不会引起 Wendy 的怀疑. 最终, Bob 使用密钥 K 从 S 中恢复出 M.

随着数字多媒体数据(如图像、视频、音频、文本等)的广泛使用, 基于多媒体数据的隐写方案受到了广泛关注. 近二十多年来, 研究者们提出了大量的多媒体隐写方法.

在早期的隐写研究中, 研究者们大多聚焦于嵌入式隐写方案, 其中图像是嵌入式隐写方案中最常

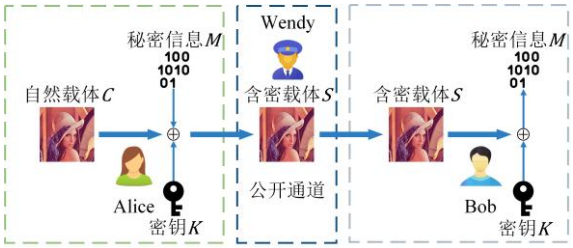


图 1 囚徒模型示意图

用的载体. 作为图像嵌入式隐写的经典方案, 最低有效位(Least Significant Bit, LSB)隐写方法[4-6]通过修改图像载体中每个像素的最低有效位嵌入秘密信息. 然而, 此类方法通常以相同的概率修改载体的每一个像素, 容易导致含密载体大量失真, 使得含密载体与自然载体在统计特性上的差异较大. 为了减少含密载体的失真, 一些研究者通过编码的方式压缩秘密信息以减少需要修改的像素数量[7-9]; 一些研究者手工设计关于图像像素修改的失真函数, 以最小化嵌入失真为目标, 从而自适应地选择载体修改的位置[10-17]; 一些研究者使用神经网络学习出失真函数来代替手工方式设计的失真函数, 进一步降低了含密载体的失真度[18, 19].

受图像嵌入式隐写方法启发, 研究者们随后提出了基于文本和音频的嵌入式隐写方法. 根据这些方法秘密信息嵌入域的不同, 这些方法可以分为: 基于文本格式的隐写方法[20-22], 基于文本内容的隐写方法[23-25], 基于音频时域的隐写方法[26-28], 基于音频频域的隐写方法[29-31]等.

作为隐写术的天生敌手, 隐写分析(Steganalysis)技术旨在检测载体中是否存在隐藏的

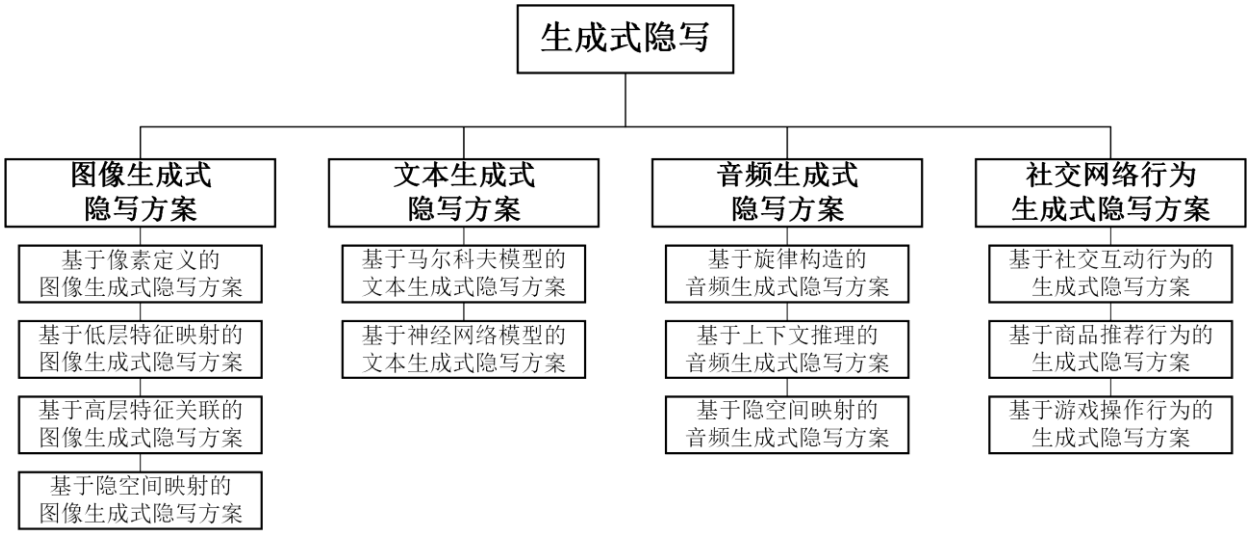


图 2 生成式隐写方案分类

秘密信息. 早期的隐写分析器从含密载体中提取统计特征[32-34], 然后将其输入到分类器中进行训练, 利用训练后的分类器判断载体中是否存在秘密信息. 另一些研究者使用卷积神经网络(Convolutional Neural Networks, CNN)[35, 36]从载体中自动学习出更加有效的隐写分析特征[37]. 由于现有的嵌入式隐写方案不可避免地改变了载体的统计特性, 为隐写分析提供了重要的检测依据, 从而难以逃避隐写分析工具的成功检测.

为了抵抗隐写分析, 研究者们提出了全新的隐写思路, 即“生成式隐写”, 具体是指: 以秘密信息为驱动, 直接“构造”或“生成”含密载体. 近年来, 研究者们提出了大量的多媒体生成网络模型, 包括: 循环神经网络(Recurrent Neural Network, RNN)[38, 39]、变分自编码器(Variational Auto-Encoder, VAE)[40, 41]、生成对抗网络(Generative Adversarial Network, GAN)[42, 43]、流模型(Flow-based Model)[44-46]等. 这些模型生成的多媒体数据足以达到“以假乱真”的程度. 这些性能强大的生成模型为生成式隐写的发展提供了良好的基础.

相比嵌入式隐写方案, 生成式隐写方案没有对现有载体进行任何修改, 因此可以较好地抵抗各类基于统计特性的隐写分析工具的检测; 同时, 生成的多媒体数据在网络中占比快速增长. 根据Gartner①发布的消息, 预计 2025 年生成多媒体数据在网络中占比将达到 10%. 因此采用生成的多媒体数据作为含密载体不容易引起怀疑与攻击. 由于以上因素, 生成式隐写逐渐成为信息隐藏领域的研究热点, 引起了研究者们大量关注.

根据生成的含密载体类型, 本文将现有的生成式隐写方案分为以下四大类: 1) 图像生成式隐写方案, 2) 文本生成式隐写方案, 3) 音频生成式隐写方案, 4) 社交网络行为生成式隐写方案; 根据隐写方式的不同, 每类生成式隐写方案又可以进一步细分为若干个子类. 生成式隐写分类如图 2.

本文在 2-5 章详细描述了每类方案中的各种隐写方法, 总结和分析这些方法的优缺点; 第 6 章通过实验比较了各种图像和文本生成式隐写方法的性能; 第 7 章总结了目前生成式隐写方法存在的问题; 第 8 章展望未来的发展方向.

2 图像生成式隐写方案

根据秘密信息表达方式的不同, 本文将现有的图像生成式隐写方案分为: 基于像素定义的图像生成式隐写方案, 基于低层特征映射的图像生成式隐写方案, 基于高层特征关联的图像生成式隐写方案, 以及基于隐空间映射的图像生成式隐写方案, 如表 1 所示.

2.1 基于像素定义的图像生成式隐写方案

基于像素定义的图像生成式隐写方案主要有两类. 一类方法将秘密信息编码为图像的像素值从而构造整幅图像; 另一类方法先将秘密信息映射图像中部分指定位置像素值, 然后利用图像生成技术补全图像的剩余部分.

① Gartner 发布 2022 年重要战略技术趋势. https://www.gartner.com/cn/newsroom/press-releases/gartner-top-strategic-technology-trends-for-2022

表 1　图像生成式隐写方案的分类对比

方案类别	主要思路	抗隐写分析性能	隐写容量	提取率	鲁棒性
基于像素定义的图像生成式隐写方案	将秘密编码为图像的像素值从而构造整幅图像；或者先映射为图像中部分指定位置像素的值，再补全图像中剩余部分	高	高 (9.80×10^{-3} ~$4.30\pm0.85bpp$)	高 (58%~100%)	低
基于低层特征映射的图像生成式隐写方案	将秘密信息映射为图像低层特征，生成具有此特征的含密图像	高	中 (2×10^{-3}~3.28×10^{-2} bpp)	高 (70%~100%)	中
基于高层特征关联的图像生成式隐写方案	将秘密信息与图像高层特征相关联，生成具有此特征的含密图像	高	低 (5.07×10^{-5}~ 6×10^{-2} bpp)	高 (97.64%~100%)	高
基于隐空间映射的图像生成式隐写方案	通过可逆神经网络学习隐空间和图像空间之间的可逆映射函数，然后将秘密信息编码到隐空间向量，利用映射函数生成含密图像	高	高 (7.81×10^{-3}~$8bpp$)	高 (84%~100%)	中

Yang 等[47]将秘密信息编码为像素值并构成整幅含密图像. 该方法使用像素卷积神经网络(Pixel Convolutional Neural Networks, PixelCNN)[48]对像素之间的依赖关系建模，从而可以根据已经生成像素获得当前待生成像素在[0, 255]范围的取值概率分布. PixelCNN 通过每个像素的条件概率分布的乘积对图像进行建模:

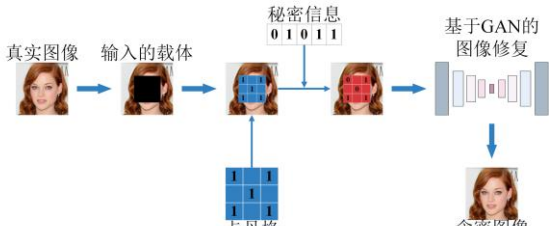
$$P(X)=\prod_{i=1}^{n^2}P(x_i\mid x_1,x_2,...,x_{i-1}) \tag{1}$$

其中 $P(x_i\mid x_1,x_2,...,x_{i-1})$是第 i 个像素 x_i 的条件概率分布；然后，使用拒绝采样的策略[49]，在该像素的概率分布中进行重复采样直到采样像素值满足以秘密信息为约束条件的值，并将该值作为待生成像素的值，这样使得秘密信息编码到该像素中；最后，按以上图像像素生成方法，从空白图像的以左上角到右下角为顺序逐个生成像素，最终构造出整幅含密图像.

Zhang 等[50]在 Yang 等[47]的基础上提出了 Pixel-Stega 方法. Pixel-Stega 方法使用性能更为先进的 PixelCNN++[51]代替 PixelCNN，并通过基于算术编码的采样策略，根据秘密信息的内容和当前待生成像素的概率分布，确定对应的像素值. Pixel-Stega 方法的隐写模型如图 3 所示. 由于使用了基于算术编码的采样策略，Pixel-Stega 方法能够在含密图像熵值较大的区域隐藏较多的信息，而在熵值较小的区域隐藏较少的信息，即能够自适应地隐藏秘密信息，从而在给定的隐写荷载(期望的隐藏信息量)下，一定程度上提升了图像的生成质量.

以上两种方法对像素之间的依赖关系建模，它们根据已生成的像素生成当前像素来构造含密图像，但忽略了整幅图像的上下文信息，即图像的全局信息，使得含密图像难以保持语义的合理性. 因此，以上两种方法生成的含密图像质量较低.

卡丹格(Cardan Grille)是一种西方早期用于隐藏信息的简单网格. 发送者以卡丹格为掩模, 将秘密信息预先写到书信等文本载体上的卡丹格网格指定位置, 然后根据已写好的秘密信息将其余文本补充完整, 从而在传递载体时不引起他人的怀疑, 接收者可以利用卡丹格为掩模从文本指定位置提取出秘密信息. 受卡丹格隐藏信息思路的启发, Liu 等[52]提出了一种利用卡丹格实现基于像素定义的图像生成式隐写方法. 如图 4 所示, 该方案将含有受损区域的图像作为载体并设计了一个与图像受损区域相同形状的卡丹格掩膜, 将秘密信息填充到卡丹格掩膜中标记为“1”的区域; 然后将填充秘密信息后的受损图像输入到 DCGAN 中, 以填充的秘密信息保持不变为条件, 自动生成和恢复出受损区域以获得含密图像. 在秘密信息提取时, 将含密图像与卡丹格掩膜结合, 直接从卡丹格掩膜中的标记为“1”的区域中提取秘密信息. 该方法在信息隐藏过程中, 受损区域的恢复受到秘密信息需保持不变以及与未受损区域需保持语义一致的双重条件约束. 因此, 如果缩小未受损区域面积, 将弱化恢复受损区域时需与未受损区域保持语义一致性的约束, 有利于恢复受损区域时秘密信息保持不变, 从而提高秘密信息提取率. 按照以上思路, 一些研究者缩小未受损区域面积, 然后利用卡丹格将秘密信息隐藏在相应的位置, 并通过 DCGAN 补全受损区域, 从而达到了理想的秘密信息提取率[53, 54]. 以上方法以图像未受损区域作为参考信息, 能够使得补全的受损区域与未受损区域语义保持一致. 因此, 与 Yang 等[47]和 Pixel-Stega 隐写方法相比, 该方法所生成的含密图像质量有所提高.



基于像素定义的图像生成式隐写各方法的对比如表 2 所示. 由于隐写过程中没有对载体进行修改, 基于像素定义的图像生成式隐写方案能够有效抵抗现有基于统计特征的隐写分析工具的检测. 此外, 该方案将秘密信息映射为像素, 图像中的像素所能承载的信息量较大, 因此该方案的隐写容量相对较高. 然而, 图像在传输过程中有可能受到各种各样的攻击(如重压缩、添加噪声等, 其中添加噪声攻击包括添加 3%的椒盐噪声[55]、强度值为 25 的椒盐噪声[56]、方差为 0.1 和 0.01 的高斯噪声[57, 58]等攻击), 这些常见攻击会对图像像素值的影响较大, 因此基于像素定义方法的鲁棒性(即含密载体受到攻击后的秘密信息提取率)普遍较低.

2.2 基于低层特征映射的图像生成式隐写方案

为了解决基于像素定义的图像生成式隐写的鲁棒性较低问题, 部分研究者提出了基于低层特征映射的图像生成式隐写方案. 该类方案将秘密信息映射为图像的低层特征(如纹理、轮廓等), 然后生成具有此特征的含密图像.

Otori 等[59]提出了一种根据秘密信息生成纹理图像的隐写方法. 该方法首先建立局部二值模式(Local Binary Pattern, LBP)算子[60]的特征值与纹理中彩色点之间的映射规则. 通过该映射规则, 将秘密信息等值的 LBP 特征映射为相应的彩色点, 并绘制在空白图像中的固定位置上; 然后, 根据选定的图像纹理, 利用纹理合成技术对载体图像上的空白区域进行补全, 最终生成含密图像. 由于 LBP 算子

表 2 基于像素定义的图像生成式隐写各方法对比

代表方法	主要思路	优点	缺点
Yang 等[47]	在使用 PixelCNN 生成像素的过程中, 通过拒绝采样的策略, 将秘密信息编码为对应像素值	隐写容量较高(1*bpp*)	图像的生成质量较低, 鲁棒性较低
Pixel-Stega[50]	在使用 PixelCNN++生成像素的过程中, 通过基于算数编码的采样策略, 将秘密信息编码为对应像素值	相比 Yang 等[47]方法, 图像的生成质量有所提高	图像的生成质量仍然较低, 鲁棒性较低
Liu 等[52]	利用卡丹格实现像素定义, 通过 DCGAN 修复受损区域以生成含密图像	相比 Yang 等[47]和 Pixel-Stega[50]方法, 图像的生成质量相对较好	秘密信息提取率较低, 鲁棒性较低
隐写方法[53][54]	在 Liu 等[52]方法基础上, 进一步缩小未受损区域, 实现基于卡丹格的生成式隐写	相比 Liu 等[47]方法, 该方法秘密信息的提取率有所提高(95%~100%)	鲁棒性较低

图 3 Pixel-Stega 隐写模型

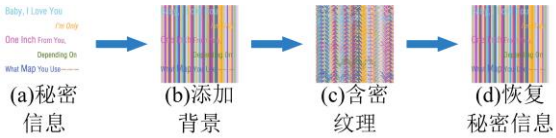


图 5　stego-texture 隐写模型

的特征值和彩色点对各种常见的图像攻击不敏感，因此该方法针对打印扫描和重拍摄等常见图像攻击具有较好的鲁棒性；然而，由于 LBP 特征值维度较低，承载的信息容量受限，因此该方法隐写容量偏低.

Xu 等[61]基于大理石纹理合成技术提出了 stego-texture 方法. 如图 5(a)所示，在隐写过程中，该方法首先将秘密字符显式地绘制在空白图像上；然后根据秘密字符的颜色、曲率、方向、字号等属性添加背景线条得到背景图像，生成的背景线条可以在不遮挡秘密信息的同时，生成视觉自然的纹理图像，生成的背景图像如图 5(b)所示；最后，如图 5(c)所示，通过置乱操作把背景图像映射为最终的含密图像，其中使用的置乱操作由七种可逆几何形变函数排列组合构成，而且置乱操作的参数也会被隐藏到构造的含密图像中；接收者可以提取在含密图像中置乱操作的参数将隐写图像恢复为背景图像，从而恢复出秘密信息，如图 5(d)所示.

为了提升隐写容量，Wu 等[62]根据秘密信息选择相应的纹理图像块，拼接成整幅纹理图像作为含密图像. 该方法选定一个图像纹理块，将此源纹理块分割并拓展为一系列候选图像块；然后按照候选块之间的均方误差(Mean-Square Error, MSE)值排序，根据排序值划分到候选列表对应位置，将每一候选列表映射为不同类型的秘密信息片段(候选列表建立的规则如图 6 所示)；接着根据秘密信息从相应的图像块候选列表中，选择相应的图像块依次填充到载体图像中的空白区域，从而生成整幅含密图像. 此外，基于可逆纹理合成技术，该方法可以从含密图像中准确恢复出源图像块并提取秘密信息.

指纹图像作为一种特殊的纹理图像，被广泛使用于各类实际应用中. Li 等[56]提出了以秘密信息为驱动构造指纹图像的隐写方法. Larkin 等[63]认为，任意一张自然指纹图像都可以被分解为偏置、幅度、全息相位和噪声四个部分，其中全息相位由螺旋相位和连续相位组成. 由于在螺旋相位中螺旋的位置与指纹图像中细节点的位置一致且随机性较强，Li 等[56]提出的方法将秘密信息及其纠错码[64]映射为螺旋相位中螺旋的位置并据此构建出螺旋相位；然后使用基于 Garbor 滤波的指纹生成模型[65]构建连续相位；最后将螺旋相位和连续相位合成为全息相位，并经过加噪、渲染等后期处理得到最终

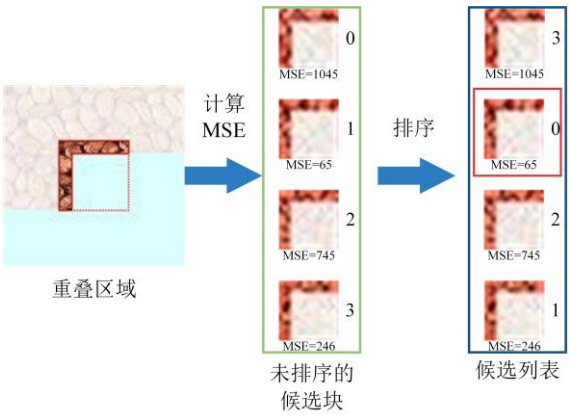


图 6　Wu 等[62]候选图像块列表的建立

的高质量指纹图像. 由于指纹细节点的位置相对比较稳定，接收者可以直接根据细节点的位置反推出秘密信息，实现准确的秘密信息提取.

纹理图像这类非常见的自然图像在网络中的传输易引起攻击者的怀疑，而包含物体轮廓的图像更为常见. 如果将秘密信息隐藏在物体轮廓中并以此构造含密图像，秘密信息的隐蔽性更好. Zhou 等[66]据此设计了基于轮廓自动生成的生成式隐写方法. 该方法将秘密信息映射为轮廓信息，然后将轮廓信息作为 GAN 的约束构造出相应的含密图像. 详细来说，在隐写阶段，首先构建基于长短期记忆网络(Long Short-Term Memory, LSTM)[67]的轮廓自动生成模型，以秘密信息为驱动，生成相应的轮廓线；然后，在原有的 pix2pix 网络[68]基础上增加提取器，建立轮廓−图像可逆变换模型以实现生成轮廓到含密图像的变换. 该模型由生成器 G，判别器 D 和提取器 E 共同组成. 在提取阶段，接收方利用轮廓−图像可逆变换模型提取出轮廓线并对秘密信息进行恢复. 该隐写网络的训练目标函数可以表示为:

$$\begin{aligned} G^*, E^* = \arg\min_G \max_D \min_E \mathrm{L}_{\mathrm{GAN}}(G,D) \\ +\lambda \mathrm{L}_{L_1}(G) + \mu \mathrm{L}_{L_1}(E) \end{aligned} \tag{2}$$

其中，为了对提取器进行训练并优化，该目标函数在原 pix2pix 目标函数的基础上加入了提取器的 L_1 距离损失 $\mathrm{L}_{L_1}(E)$，其表示提取轮廓和生成轮廓之间的 L_1 距离，可通过以下公式计算:

$$\begin{aligned} \mathrm{L}_{L_1}(E) &= \mathrm{E}_{S,x,z}\left[\left\|C_e - C_g\right\|_1\right] \\ C_e &= E\left(G(x,z)\right) \\ C_g &= LSTM\left(S\right) \end{aligned} \tag{3}$$

其中，C_e 表示提取轮廓，C_g 表示生成轮廓，S 表示秘密信息的十进制序列. 该方法虽然能够准确地提取秘密信息，但是由于只能生成一维轮廓，该方法隐写容量相对较小. 该方法的隐写网络如图 7 所示.

表 3　基于低层特征映射的图像生成式隐写各方法对比

代表方法	主要思路	优点	缺点
Otori 等[59]	将秘密信息编码为彩色点并绘制在空白载体图像上，然后补全载体图像剩余部分	具有较好的鲁棒性	隐写容量偏低，生成的纹理图像容易引起攻击者怀疑
stego-texture [61]	根据秘密信息的属性构造背景图像，并通过置乱操作把背景图像映射为含密纹理图像	具有较好的鲁棒性	生成的纹理图像容易引起攻击者怀疑
Wu 等[62]	根据秘密信息选择相应的纹理图像块构造出整幅含密纹理图像	相比 Otori 等[59]和 stego-texture 方法，隐写容量有所提高(3.28×10^{-2} *bpp*)，具有较好的鲁棒性	
Li 等[56]	根据秘密信息构造螺旋相位，并生成相应的指纹图像	具有较好的鲁棒性	生成的指纹图像容易引起攻击者怀疑
Zhou 等[66]	将秘密信息映射为轮廓信息并作为 GAN 的输入构造出相应的含密图像	能够生成自然图像具有较好的隐蔽性，具有较好的鲁棒性	隐写容量较低(4×10^{-3} *bpp*)

基于低层特征映射的图像生成式隐写各方法的对比如表 3 所示. 由于图像的纹理、轮廓等低层特征相对稳定，不容易受到图像攻击的影响. 因此相比基于像素定义的生成式隐写方法，基于低层特征映射的生成式隐写方法显著地提高了鲁棒性. 然而，与像素相比，纹理和轮廓所能承载的信息容量较低. 因此，基于低层特征映射的生成式隐写方案隐写容量普遍低于基于像素定义的生成式隐写方案的隐写容量.

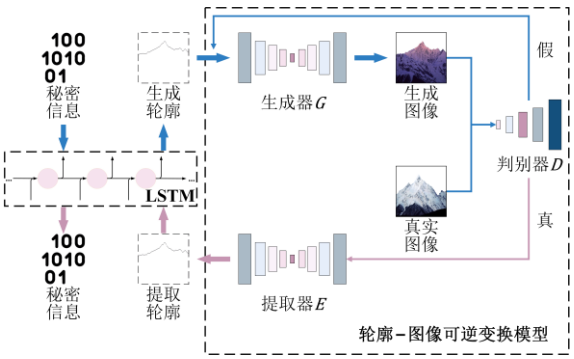


图 7　Zhou 等[66]隐写模型

2.3　基于高层特征关联的图像生成式隐写方案

为了进一步提升含密图像的生成质量，一些研究者提出了基于高层特征关联的图像生成式隐写方案. 该类方案将秘密信息编码为图像的特定高层特征(例如图像的语义信息、图像的风格特征等)，并生成符合该特征的含密图像.

图像的语义信息作为常用的高层特征，被研究者应用于图像生成式隐写任务中. Cao 等[57]提出了一种基于动漫角色生成的图像生成式隐写方法. 该方法主要包括秘密信息与属性标签转换模块、图像生成模块、图像质量评估模块. 首先将秘密信息转换为二进制字符串；然后通过 LSTM 模型将秘密信息转换为动漫角色的属性标签集合(如发型、发色、瞳色等)，以该属性标签为作为 GAN 网络的输入条件生成动漫角色图像；最后，评估生成图像的质量，选择质量较高的图像作为含密图像. 接收者从含密图像中提取动漫角色的标签，并将其转换为秘密信息.

Zhang 等[58]将秘密信息与图像的语义标签信息构建映射规则，在此基础上设计了名为 SSS-GAN(Synthetic Semantics Stego Generative Adversarial Network)的生成式隐写方法. SSS-GAN 方法的隐写模型如图 8 所示. SSS-GAN 方法的生成器 G 根据标签 c 和随机噪声 z 生成图像 X_{fake}，其可表示为 $X_{\text{fake}}=G(c,z)$；判别器 D 用于判别真实图像和生成图像，其可表示为 $D(X)=P(V\mid X)$，$X\in\{X_{\text{real}},X_{\text{fake}}\}$ 表示输入判别器中的图像来源于数据集的真实图像或者由生成器生成的图像，$V\in\{0,1\}$ 表示判别器判断结果的取值，0 表示判断为生成图像，1 表示判断为真实图像；分类器 C 用于识别图像的类别标签，表示为 $C(X)=P(I\mid X)$，$X\in\{X_{\text{real}},X_{\text{fake}}\}$ 表示输入分类器中的图像来源于数据集的真实图像或者由生成器生成的图像，$I\in\{c_1,c_1,\ldots,c_n\}$ 表示分类器识别获得的图像标签. SSS-GAN 的损失函数可定义为:

$$\max_{G\_C}\max_{D\_C}L_{\text{SSS-GAN}}=L_{G\_C}+L_{D\_C} \tag{4}$$

$$\begin{aligned}L_{G\_C}=&(1-\alpha)\cdot\mathrm{E}\left[\log P(V=1\mid X_{\text{fake}})\right]\\&+\alpha\cdot\mathrm{E}\left[\log P(I=c\mid X_{\text{fake}})\right]\end{aligned} \tag{5}$$

$$\begin{aligned}L_{D\_C}=&(1-\alpha)\cdot\{\mathrm{E}\left[\log P(V=1\mid X_{\text{real}})\right]\\&+\mathrm{E}\left[\log P(V=0\mid X_{\text{fake}})\right]\}\\&+\alpha\cdot\{\mathrm{E}\left[\log P(I=c\mid X_{\text{real}})\right]\\&+\mathrm{E}\left[\log P(I=c\mid X_{\text{fake}})\right]\}\end{aligned} \tag{6}$$

其中，L_{G_C} 是生成器与分类器损失的加权求和，L_{D_C} 是判别器与分类器损失的加权求和，α 是控制生成图像质量与信息提取精度权重大小的参数. 在隐写的过程中，SSS-GAN 方法将秘密信息分段并映射为相应的标签，把标签与随机噪声作为生成器的输入从而生成含密图像. 接收者将分类器作为提

取器，从含密图像中提取出标签，再将标签反向映射便可以获取秘密信息. SSS-GAN 方法能够实现接近 100%的提取率且对噪声和压缩攻击具有较好的鲁棒性. 值得注意的是，该方法在实现过程中，秘密信息的隐藏和提取只与图像的标签有关. 因此该方法隐写容量与图像的尺寸无关，而与数据集中标签的数量正相关. 当隐藏的信息稍长时，该方法需要非常庞大的训练数据集的支持，且对生成器和分类器的设计和训练有非常高的要求，因此该方法隐写容量通常受到较大的限制.

Wang 等[69]提出了基于风格迁移的深度隐写模型 STNet(Style Transformation Network for Deep Image Steganography)，在图像风格转换的过程中将秘密信息隐藏到生成图像的风格特征中. 如图 9 所示，STNet 方法由风格转换网络、提取网络和损失网

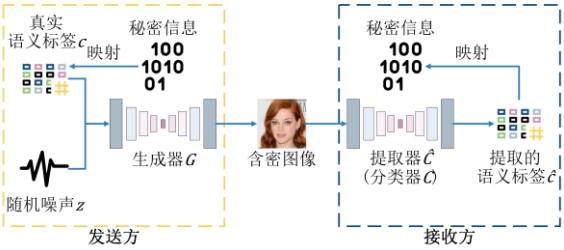


图 8 SSS-GAN 隐写模型

络三部分组成，其中基于 VGG-19 模型[70]设计了编码器、解码器和损失网络. 在风格转换网络中，发送方首先将载体图像和参考图像输入 VGG 编码器以获得载体图像内容特征和参考图像风格特征；然后，使用自适应实例归一化(Adaptive Instance Normalization, AdaIN)层[71]对载体图像内容特征和参考图像风格特征归一化，以获得新的特征；最后，将秘密信息与获得的新特征拼接成为完整的特征，然后输入到解码器中从而获得含密图像；损失网络是预训练的 VGG-19 网络，通过评估含密图像的内容特征和风格特征与输入图像的差异，确保含密图像的内容与风格特征与输入图像一致；提取网络由六个包含批归一化和 Leaky ReLU 激活函数[72]的卷积层构成，接收者将含密图像输入到提取网络便可以提取秘密信息. STNet 整体的损失函数由内容损失 L_C、风格损失 L_S 以及秘密信息重建损失 L_M 经过加权构成，其可表示为:

$$L_{\text{STNet}} = L_C + \lambda L_S + \mu L_M \tag{7}$$

其中，λ 和 μ 分别是风格损失和秘密信息重建损失的权重的系数.

循环一致性生成对抗网络(Cycle-consistency Generative Adversarial Networks, Cycle-GAN)[73]可以实现两个不同图像域的转换，其结构是由两个生成器和两个判别器组成. 受 Cycle-GAN 在风格迁移相关研究的启发，Li 等[74]提出了一种基于风格迁移的图像生成式隐写方法，通过图像风格的迁移网络，将秘密图像与另一幅图像合成为一幅指定风格的含密图像，如图 10 所示. 该方法在隐写阶段和提取阶段各设计了一个生成器、一个提取器以及一个判别器. 其中，这两个阶段的提取器与判别器结构相同，分别采用了 VGG-16[70]和 DCGAN 结构. 在隐写阶段，该方法首先将秘密图像放置在载体图像上组成合成图像，然后将该合成图像和参考图像输入到 Cycle-GAN 的生成器中获得含密图像，这样使得该含密图像的风格与参考图像风格一致；在提取秘密图像的过程中，该方法使用提取器从原载体图像提取特征，将该特征和含密图像作为 Cycle-GAN 的另一生成器的输入，生成和恢复出合成图像. 并且在恢复时需要融合原始载体图像的特征，使恢复的合成图像与原始载体图像风格一致.

基于高层特征关联的图像生成式隐写的各方法的对比如表 4 所示. 相比纹理、轮廓等图像低层特征，图像高层特征更加稳定，更不容易受到图像攻击(如添加噪声等)的影响，因此基于图像高层特征关联的生成式隐写方法鲁棒性更高. 然而，从图像中抽象出来的高层特征所能承载的信息量较低，因此基于高层特征关联的生成式隐写方案的隐写容量会明显低于基于像素定义以及基于低层特征映射的生成式隐写方案.

2.4 基于隐空间映射的图像生成式隐写方案

一些研究者发现自然图像通常服从特殊的复杂分布，通过神经网络能够学习到图像空间分布与某隐空间分布(如高维高斯分布等)的映射规则[44, 46]. 根据以上研究成果启发，基于隐空间映射的图像生成式隐写方案通过构建秘密信息与隐空间向量的映射规则，将秘密信息映射为隐向量并转换为相应的图像. 该图像可以作为含密图像从而实现隐蔽通信.

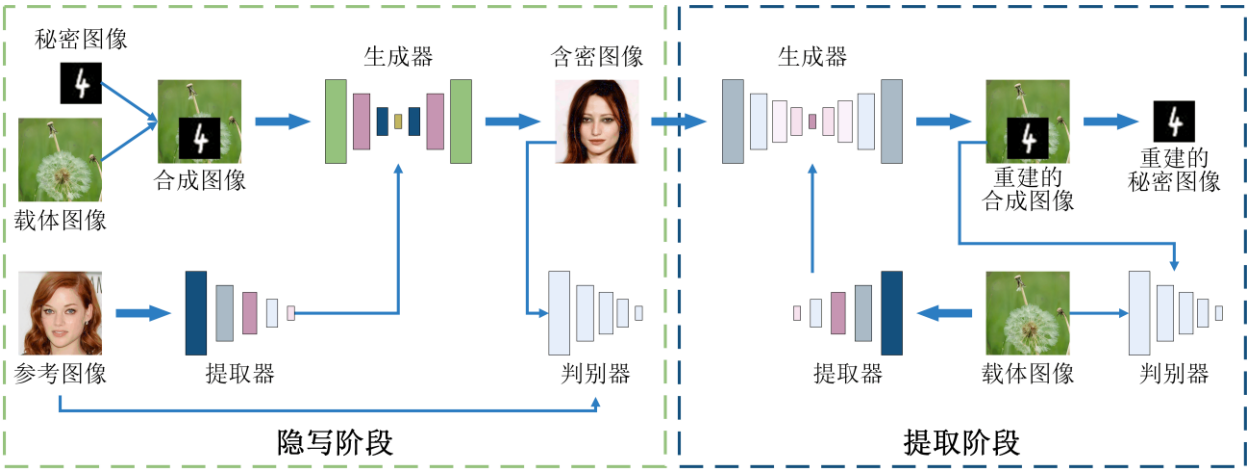


图 10　Li 等[74]隐写模型

图 9　STNet 隐写模型

Hu 等[75]提出了一种将秘密信息映射为隐向量的图像生成式隐写方法. 如图 11 所示, 该方法首先将秘密信息编码为 DCGAN 生成器的输入低维噪声向量, 利用该生成器生成载体图像; 并训练 CNN 模型作为提取器, 确保提取的噪声向量与原始的噪声向量一致, 从而恢复出秘密信息. 训练好的生成器和提取器共同实现了在隐空间和图像空间之间相互映射.

由于 DCGAN 生成的一些图像不够自然, Li 等[76]使用 WGAN-GP(Wasserstein GAN Gradient Penalty)[77]代替 DCGAN 以生成更加真实的含密图像.

Zhang 等[78]提出了一种图像生成式隐写网络 GSN(Generative Steganography Network), 不同于现有的大多数方法将秘密信息映射为隐向量然后利用该隐向量生成含密图像, 该方法在通过隐向量生成图像的过程中隐藏秘密信息. GSN 由生成器、判别器、隐写分析器和提取器组成. 在隐写阶段, 发送方首先将隐向量输入生成器获得特征图, 然后将二进制秘密信息通过数据合并操作(Data Merging Operation)添加到特征图中, 从而可以利用添加秘密信息后的特征图生成含密图像; 最后, 发送方同时将生成器生成的载体图像和含密图像通过无损信道传递给接收方. 在提取阶段, 接收方使用提取器从含密图像中提取秘密信息.

GAN 模型训练时, 隐空间的维数通常远低于图像空间的维数, 而提高隐空间的维度将会导致模型难以训练. 因此, Hu 等[75]隐写方法和 Li 等[76]隐写方法的隐写容量受到了限制. 由于 GSN 方法在通过隐向量生成图像的过程中隐藏秘密信息, 避免了隐向量低维度或图像语义标签低信息量的限制, 能够极大地提升隐写容量和含密图像生成质量. 然而, 由于 GAN 模型不能实现隐空间到图像空间的可逆变换, 需要额外训练 CNN 模型提取器, 因此 Hu 等[75]隐写方法、Li 等[76]隐写方法和 GSN 方法的秘密信息的提取准确率需进一步提高.

Liu 等[79]提出了基于图像解耦自编码器(Image Disentanglement Autoencoder for Steganography, IDEAS)的隐写方法. 在隐写过程中, 发送方将秘密信息映射为隐向量, 并将其转换为图像的结构特征; 然后发送方将该结构特征与随机采样获得的图像纹理特征作为生成器的输入以生成含密图像. 在秘密信息提取的过程中, 接收方首先训练编码器从含密图像中恢复结构特征; 然后训练解码器, 根据结构特征恢复隐向量; 最后将该隐向量逆向映射为秘密信息. 由于图像的结构特征较为稳定, 因此 IDEAS 提升了秘密信息提取的准确率. 此外, 由于加入了随机采样得到的纹理特征, IDEAS 可以生成具有多样性的含密图像, 可以避免针对特定的秘密信息只能生成单一的含密图像的问题, 从而提升了隐写的安全性.

表 4　基于高层特征关联的图像生成式隐写各方法对比

代表方法	主要思路	优点	缺点
	利用 GAN	秘密信息提取率较高(100%)，鲁棒性较高	隐写容量非常低(896 bits/carrier)
	射关系	秘密信息提取率较高(100%)，鲁棒性较高	隐写容量非常低(7.30×10^{-4} *bpp*)，且需要庞大的数据库支持
	与秘密信 :图像转换 图像	秘密信息提取率较高(99.80%)，鲁棒性较高	相比 Cao 等[57]和 SSS-GAN [58]方法，隐写容量有所提高但仍然偏低，且训练整个神经网络需要更多额外的资源消耗
	然后转换 图像	秘密信息提取率较高(97.64%)，鲁棒性较高	相比 Cao 等[57]和 SSS-GAN [58]方法，隐写容量有所提高但仍然偏低

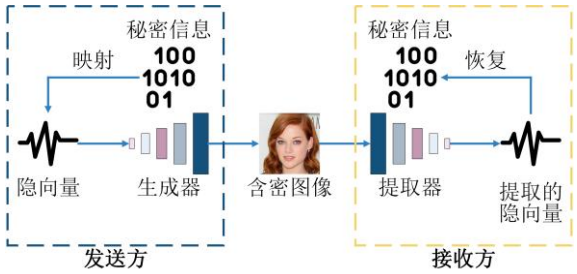


图 11　Hu 等[75]隐写模型

IDEAS 方法采用 Autoencoder 网络实现秘密信息的提取任务. 然而，由于该过程涉及池化和归一化操作，容易丢失图像内部的细节信息. 因此，该方法依旧不能保证秘密信息的精确提取.

与 GAN 和 Autoencoder 网络模型相比，Glow(Generative Flow with Invertible 1x1 Convolutions)[46]模型具有高质量图像的生成能力，并且能够在隐向量和图像之间构成可逆映射，Glow 模型描述如下. 假设 z 是隐空间 Z 中的隐向量，x 是图像空间 X 中的变量，z 和 x 分别服从以下分布:

$$z \sim p_Z(z) \tag{8}$$

$$x \sim p_X(x) \tag{9}$$

其中 z 的维度与 x 相同，$p_Z(z)$ 是一个已知的简单分布，而 $p_X(x)$ 是未知的复杂的分布. 假设 z 的分量 z_d 相互独立，那么 $p_Z(z)$ 可以表示为 $p_Z(z)=\prod_d p_{Z_d}(z_d)$，其中 $p_{Z_d}(z_d)$ 是一个已知分布(如标准高斯分布). 为了学习图像空间分布与该隐空间分布的转换规则, Glow 的训练过程就是学习一个可逆映射函数 $z=f_\theta(x)$ 以及它的反函数 $x=g_\theta(z)=f_\theta^{-1}(z)$. 设 P_{data} 为包含一组图像样本的训练集，可以通过极大对数似然估计对映射函数 $f_\theta(x)$ 中的参数进行拟合. 根据雅克比行列式，该训练过程中目标函数的表达式可表示为:

$$\begin{aligned}&\max_{\theta\in\Theta} \mathrm{E}_{x\sim P_{\text{data}}}\left[\log p_X(x)\right]\\&=\max_{\theta\in\Theta} \mathrm{E}_{x\sim P_{\text{data}}}\left[\log p_Z\left(f_\theta(x)\right)\left|\det\frac{\partial f_\theta(x)}{\partial x}\right|\right]\end{aligned} \tag{10}$$

Zhou 等[55]基于 Glow 模型提出了一种秘密信息到图像的可逆变换(Secret-to-Image Reversible Transformation，S2IRT)隐写方法. 在隐写阶段，为了提升隐写容量，发送方使用位置编码算法将秘密信息编码为高维向量. 即从标准高斯分布中随机采样部分元素(服从标准高斯分布的数值)并进行分组，每组元素在秘密信息的指导下分配到相应的位置以获得位置索引，从而根据得到的位置索引构成高维隐向量；然后通过 Glow 模型将该隐向量映射为含密图像；在提取阶段，接收方可以通过 Glow 模型的逆变换将经过预处理的含密图像恢复为隐向量，对该隐向量进行解码便可获得秘密信息. 由于位置编码具有高效无损性以及 Glow 模型能在隐空间与图像空间之间实现可逆映射，S2IRT 方法可以在极大地提升隐写容量的同时，准确地提取秘密信息.

基于隐空间映射的图像生成式隐写的各方法的对比如表 5 所示. 基于隐空间映射的图像生成式隐写方案根据秘密信息从隐空间中采样，可以获得较高的隐写容量. 常见的图像攻击有可能使得隐空间的向量值受到一定程度的影响，与高层特征关联的图像生成式隐写方案相比，该方案的鲁棒性有所降低. 然而，值得注意的是，基于隐空间映射的生成式隐写方案的模型训练目标是将高斯分布拟合为自然图像在图像空间中的分布，而现有的方法构建的隐向量都是服从高斯分布的，实际上自然图像对应的隐向量分布只是近似的高斯分布，设计隐空间的检测器工具可以根据此现象仍然有可能检测到隐写的存在，将一定程度上影响隐写信息的安全性.

表 5 基于隐空间映射的图像生成式隐写各方法对比

代表方法	主要思路	优点	缺点
Hu 等[75]	将秘密信息映射为噪声隐向量, 分别训练 DCGAN 生成器生成含密图像和提取器提取秘密信息	能生成质量较好的含密图像	生成的含密图像类型比较单一, 且无法实现秘密信息准确提取, 生成的含密图像尺寸较小
Li 等[76]	同时训练生成器和提取器, 并使用 WGAN-GP 代替 DCGAN	能生成质量较好的含密图像	生成的含密图像类型比较单一, 且无法实现秘密信息准确提取
GSN[78]	将隐向量输入生成器获得特征图, 然后将二进制秘密信息张量通过数据合并操作加入到特征图中利用加入秘密信息后的特征图生成含密图像	隐写容量较高(8*bpp*), 能生成质量较好的含密图像	生成的含密图像类型比较单一, 且无法实现秘密信息准确提取
IDEAS[79]	秘密信息映射为隐向量, 将其转换为图像的结构特征并与随机采样的纹理特征输入到 Autoencoder 的编码器生成含密图像, 然后含密图像输入到 Autoencoder 的解码器提取秘密信息	隐写容量较高(7.81×10^{-3} *bpp*), 能够生成种类丰富的含密图像	难以在生成质量较高图像的同时仍然保证秘密信息的准确提取
S2IRT[55]	将秘密信息编码为隐向量, 通过 Glow 模型转换为含密图像, 可以通过 Glow 模型逆变换提取秘密信息	隐写容量较高(4*bpp*), 提取率较高(100%)	生成的含密图像类型比较单一

由于具有较好的抗隐写分析性能, 图像生成式隐写方案已经成为生成式隐写领域的研究热点. 研究者们基于不同的映射规则将秘密信息转换为合适的数据类型从而生成含密图像. 然而, 虽然现有的图像生成式隐写方案所使用的生成模型能够生成较高质量的含密图像, 但也经常生成一些低质量图像. 特别是在隐写荷载较大的情况下, 图像生成质量不高. 因此, 现有的图像生成式隐写方法的图像生成质量仍然不够稳定.

3 文本生成式隐写方案

早期的文本隐写方案大多数是通过修改文本实现信息隐藏的嵌入式隐写方案[25, 80]. 然而, 相比图像载体, 文本载体的数据量小, 存在的冗余修改空间有限, 因此修改文本的嵌入式隐写方式不利于大量秘密信息的隐写. 随着自然语言处理技术的发展, 文本生成模型生成的文本质量越来越高, 这为文本生成式隐写的发展奠定了坚实的基础. 本节根据生成文本所使用生成模型的不同, 将现有的文本生成式隐写方案分为以下两类: 基于马尔科夫模型的文本生成式隐写方案和基于神经网络模型的文本生成式隐写方案. 文本生成式隐写方案的分类对比如表 6 所示.

3.1 基于马尔科夫模型的文本生成式隐写方案

马尔科夫模型可以根据自然文本中邻近单词的出现频次对文本进行建模. 因此, 早期的文本生成式隐写方法大多使用马尔科夫模型生成含密文本. 该类方案首先通过马尔科夫模型对自然文本中每个单词出现的频率进行统计, 近似获得单词的概率, 然后根据秘密信息选择相应概率的单词以生成含密文本, 在生成文本的过程中实现隐写.

Shniperov 等[81]提出了一种基于多阶(Different Order)马尔科夫模型的文本生成式隐写方法. 该方法先采样一组具有相同句式的自然文本, 并对自然文本中前若干个单词已出现时当前单词出现的频率进行统计, 近似作为当前单词的条件概率, 其中“前若干个单词”的数量即为马尔科夫模型的阶数. 以二阶马尔科夫模型为例, 当前单词 w_i 的条件概率可表示为:

$$\begin{aligned} P_{cond}(w_i) &= P(w_i \mid w_1, w_2, ..., w_{i-1}) \\ &\approx P(w_i \mid w_{i-2}, w_{i-1}) \\ &\approx \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})} \end{aligned} \tag{11}$$

其中, $count(w_{i-2}, w_{i-1}, w_i)$ 是短语 $\{w_{i-2}, w_{i-1}, w_i\}$ 在长度为 n 的自然文本中出现的次数; 然后根据秘密信息 M 选择初始单词 w_k 作为密钥 K, 发送方和接收方共享该密钥 K 和每个单词的条件概率 $P_{cond}(w_i)$; 最后, 发送方根据每个单词的条件概率 $P_{cond}(w_i)$ 构建候选池以存储候选单词, 并根据秘密信息 M 从候选池中选择相应概率值的单词, 将选择的单词作为初始单词 w_k 的后续单词, 从而使含密文本 S 中单词的顺序符合自然文本中单词的顺序. 最终, 秘

表 7　基于马尔科夫模型的文本生成式隐写各方法对比

代表方法	主要思路	优点	缺点
Shniperov 等[81]	在使用马尔科夫模型生成文本过程中，根据秘密信息选择相应概率值的单词以生成含密文本	隐写容量较高	含密文本生成质量不高
Yang 等[82]	在使用马尔科夫模型生成文本过程中，对每个单词进行哈夫曼编码，根据秘密信息选择相应的单词以生成含密文本	相比 Shniperov 等[81]方法，进一步提高了隐写容量(每个单词隐藏 1-4 位)	
Cistega[83]	在使用马尔科夫模型生成文本过程中，在预定音律的约束下，根据单词的条件概率选择合适的单词构成含密文本	生成诗歌等特定题材的文本具有更高的生成质量	只能生成特殊题材的文本(诗歌等)，适用性较低；含密文本生成过程中容易出现断链

密信息 M 编码在生成的含密文本 S 中. 在秘密信息提取过程中，接收方根据密钥 K 在含密文本 S 中确定初始单词 w_k，并根据每个单词的条件概率构建候选池，从而可以根据候选池中相应单词的概率值选择合适的单词表达秘密信息.

部分研究者通过对单词进行哈夫曼编码以提高文本的生成质量. Yang 等[82]提出了一种基于马尔科夫模型和哈夫曼编码的文本生成隐写方法. 如图 12 所示，在使用马尔科夫模型生成文本的基础上，对每个单词使用哈夫曼树进行动态编码，根据秘密信息选择相应的单词以生成含密文本. 哈夫曼编码可以依据单词的条件概率分布来构造平均长度最短的编码，使编码概率较高的单词编码长度更短. 该方法在编码过程中充分考虑了每个单词的条件概率分布，从而在一定程度上提高了隐写容量.

部分研究者认为生成某种特殊题材的文本能够使得生成模型生成更高质量的文本. 中国古代宋词作为一种特殊题材，每首词都有固定的音律和句长，生成一首词相当于根据特定的音律选择合适的单词. Luo 等[83]提出了基于词的文本生成式隐写方法(Ci-Based Steganography Methodology, Cistega). Cistega 方法使用马尔科夫模型生成文本，并在预定音律的约束下，选择合适的单词构成含密文本. 在隐写过程中，该方法首先确定初始单词、生成词的音律以及候选池的容量，其中候选池用于存储符合设定音律的当前单词；然后，根据单词的条件概率，选择符合设定音律的单词存储到候选池中；最后对候选池中的单词进行编码，选出编码与秘密信息比特流匹配的单词以构成含密文本，从而生成具有设定音律的含密文本. 生成特殊题材的文本虽然能够获得更高的生成质量，但是由于题材的特殊性，所以难以获取足够大的语料库，可能会导致生成过程中出现断链，即生成某一个单词后没有合适的下一单词能够被选择. 虽然 Cistega 方法通过降低阈值、缩短生成长度、直接选取高频次的策略能在一定程度上缓解以上问题，但不能彻底解决该问题. 除古诗词以外，个人笔记[84]、幽默短文[85]、互联网协议文本[86, 87]等特殊类型的文本同样也被用于生成式隐写.

基于马尔科夫模型的文本生成式隐写方法的对比如表 7 所示. 基于马尔科夫模型的文本生成式隐写方案能够在一定程度上可以保证生成含密文本的质量. 然而，该方案只通过计算前一个或多个单词出现时当前单词出现的频率来计算该单词条件概率，即只考虑当前单词之前的几个单词，而不是所有单词，因此马尔科夫模型不能准确地对自然文本建模. 此外，在文本生成式隐写过程中，马尔科夫模型并没有计算文本中单词实际的条件概率，而是利用统计频率近似作为条件概率. 因此，马尔科夫模型难以获得准确的条件概率以及理想的文本生成模型，从而难以生成质量较高的含密文本.

3.2　基于神经网络模型的文本生成式隐写方案

随着深度学习技术的发展，神经网络模型成为解决文本生成领域各种问题的主流方法，更多的研究者聚焦基于神经网络模型的文本生成式隐写方案. 该方案通常利用神经网络从大量真实文本样本学习文本模型，在生成过程中根据秘密信息选择相应条件概率的单词，从而实现秘密信息的隐写任务.

为了解决基于马尔科夫模型的文本生成式隐写方案难以生成质量较高的含密文本的问题，Yang 等[88]提出了 RNN-Stega 方法，该方法基于 RNN 模型实现文本生成式隐写，如图 13 所示. 在隐写的过程中，发送者首先利用 RNN 模型从大量自然文本样本中学习文本生成模型，从而获得每个单词的条件概率；然后根据条件概率对当前单词进行预测获得候选单词，并将候选单词中概率较大的前 m 个单词组成候选池；接下来根据候选池中的每个单词的条件概率将其编码为二进制比特流，并根据秘密信息的二进制比特流选择相应的单词以生成含密文本；最后将含密文本通过公共网络发送给接收者. 在秘密信息提取过程中，接收者首先使用相同的方法利用 RNN 模型获得候选单词并构建候选池；

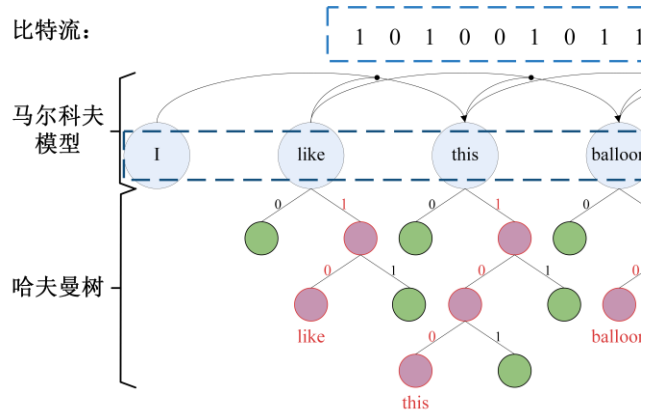


图 12　Yang 等[8

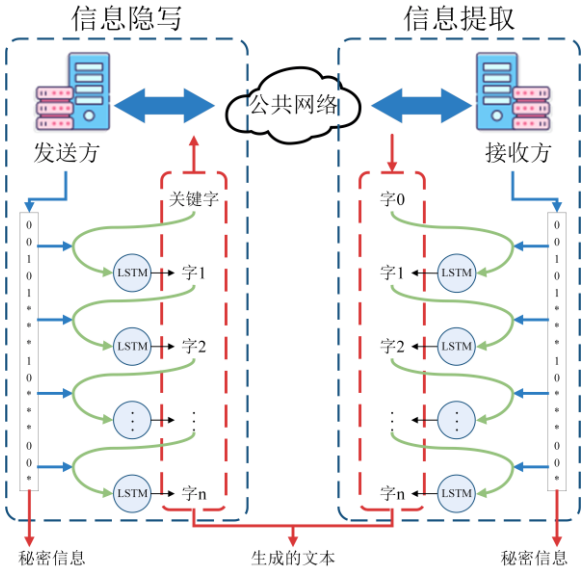


图 13　RNN-Stega 隐写模型

然后，根据含密文本中的单词从候选池中选择具有相应的二进制比特流的单词，从而可以将含密文本中的单词转换为二进制比特流以获取秘密信息．

RNN-Stega 方法将隐写网络的损失函数 $L_{\text{RNN-Stega}}$ 定义如下：

$$\begin{aligned} L_{\text{RNN-Stega}} &= -\log\left(P(S)\right) \\ &= -\log\left(P(w_1, w_2, \ldots, w_n)\right) \\ &= -\log\left(P(w_1)P(w_2 \mid w_1)\ldots\right. \\ &\quad \left.P(w_n \mid w_1, w_2, \ldots, w_{n-1})\right) \\ &= -\sum_{i=1}^{n} \log\left(P(w_i \mid w_1, w_2, \ldots, w_{i-1})\right) \end{aligned} \tag{12}$$

其中，S 表示长度为 n 的整个文本，w_i 表示 S 中的第 i 个单词，$P(S)$ 表示文本 S 的联合概率分布，其由每个单词的条件概率的乘积组成．通过最小化该损失函数以期望获得与训练的自然文本样本尽量分布一致的文本．与马尔科夫模型的相比，RNN 模型能最大程度融合已知单词的信息来估计当前单词条件概率，更加精确计算出的条件概率分布，从而学习到更好的文本生成模型．因此，RNN 模型能够生成更加合理的文本序列，RNN-Stega 方法的隐写容量和含密文本的质量均显著优于基于马尔科夫模型的文本生成式隐写方法．

然而，RNN-Stega 只考虑了文本样本中的单个句子的统计特性，而没有考虑整个文本样本的整体统计特性，导致生成的含密文本质量仍需进一步提高．因此，在 RNN-Staga 的基础上，Yang 等[89]提出 GAN-TStega 方法，首次将 GAN 网络用于文本隐写任务．GAN-TStega 方法利用 GAN 网络在不同类型文本数据集上对抗训练，从而可以生成高质量的含密文本．GAN-TStega 方法使用包含门控循环单元(Gate Recurrent Unit，GRU)的网络作为生成器，使用文本隐写分析模型[90]作为判别器以判断输入的文本是否为真实的文本，该隐写分析模型采用了双向 GRU 增强了判别能力．在隐写过程中，该方法首先对 GAN 网络进行训练并优化生成器，然后利用训练好的生成器生成文本，在生成过程中根据每个单词的条件概率分布对其编码，并根据秘密信息选择相应条件概率分布的单词以隐藏秘密信息．此外，针对传统的 GAN 网络难以生成离散序列化数据的问题，GAN-TStega 设计了新的模型更新策略，即采用强化学习的方法，并引入激励函数作为生成器的损失函数以更新生成器网络．虽然 GAN-TStega 方法的秘密信息编码策略与 RNN-Stega 类似，但得益于 GAN 网络的生成对抗策略，其生成的含密文本质量高于 RNN-Stega 方法生成的含密文本质量．

为了进一步缩减自然文本与生成的含密文本之间的统计分布差异，以提高文本生成质量，Yang 等[91]提出了一种基于变分自编码器的文本生成式隐写方法(VAE-Stega)．VAE-Stega 使用 BERT(Bidirectional Encoder Representations from Transformers)模型[92]作为编码器学习自然文本的统计分布特征，并使用 LSTM 模型作为解码器以生成符合自然文本统计特征的含密文本．VAE-Stega 的损失函数是带有正则化项的负对数似然函数，对于每一个给定的数据样本 x，其损失函数可表示为：

$$\begin{aligned} L_{\text{VAE-Stega}} &= -\mathrm{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] \\ &\quad + D_{KL}\left(q_\phi(z \mid x) \,\|\, p_\theta(z)\right) \end{aligned} \tag{13}$$

该公式的第一项用于激励解码器从隐空间 Z 中学习文本生成模型，其中 $q_\phi(z \mid x)$ 作为编码器将给定的数据样本 x 映射到隐空间 Z 中，$p_\theta(x \mid z)$ 作为解码器从隐空间 Z 中对隐向量 z 进行采样，从而生成新的数据样本；第二项用 $q_\phi(z \mid x)$ 和 $p_\theta(z)$ 的 KL 散度计算正则化项的损失，其中 $p_\theta(z)$ 作为编码器学习文本的整体分布．在隐写阶段，VAE-Stega 根据特定的概率分布从隐空间中采样一个隐向量，解码器在该隐向量的约束下生成文本．该方法在生成文本

的过程中选择条件概率较大的候选单词进行编码，并根据秘密信息选择对应的候选单词形成含密文本. 在提取阶段, VAE-Stega 根据接收的含密文本选择条件概率较大的候选单词进行解码，解码过程使用与隐写阶段相同的隐向量，从而保证准确地提取含密文本中的秘密信息. VAE-Stega 的隐写过程如图 14 所示. 与 RNN-Staga 方法相比，VAE-Stega 进一步缩减自然文本与生成的含密文本之间的统计分布差异，从而提高了文本生成质量.

Zhou 等[93]提出了一种基于随机候选池的文本生成式隐写方法. 该方法不选择概率值最大的单词组成候选池，而是随机选择概率值在一定范围的单词组成候选池，有效缓解了生成文本与自然文本的统计偏差. 在隐写过程中，该方法利用 LSTM 对当前单词进行预测，获得当前单词的条件概率分布，然后随机选择概率值在一定范围的单词组成候选池并过滤一些低频单词，最后根据秘密信息从候选池中选择相应的单词生成含密文本. 由于组成候选池是随机选择概率值在一定范围的单词并过滤了低频单词，避免了只选择概率值最大的单词组成候

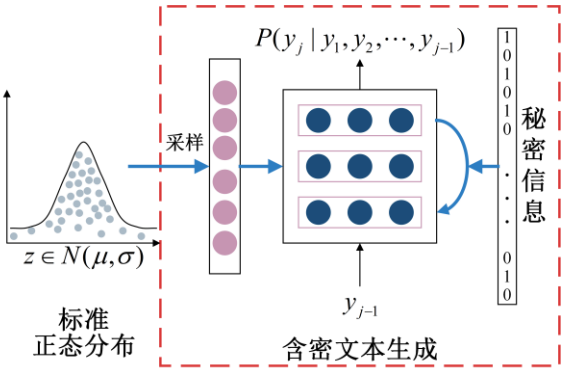


图 14 VAE-Stega 隐写模型

选池所导致生成的含密文本低频单词较多的问题，从而保证了生成的文本分布尽可能接近真实文本分布. 因此该方法一定程度上提高了含密文本的生成质量.

为了解决生成长含密文本时易出现的语义不连贯和语义错误等生成文本质量问题，Cao 等[94]提出了 PPLM-Stega 隐写方法，该方法基于 PPLM 文本生成模型(Plug and Play Language Model, PPLM)[95]. 所使用的 PPLM 能够以较少的计算资源生成基于属性模型的合理可控文本. PPLM-Stega 首先改进了 PPLM，在 PPLM 的输出层之前增加了隐写层以生成含密文本. 在传输秘密信息时，PPLM-Stega 首先根据通信各方之间共享的密钥来确定含密文本的主题. 然后，使用 PPLM 生成符合主题且具有语义一致性的长可读含密文本. 在文本生成过程中，隐写层首先确定具有最高概率的候选词的概率，然后计算其他候选词的概率与最高概率候选词的比例. 如果该比例大于预设的阈值，则将其添加到可嵌入的候选词池(Embeddable Candidate Word Pool, ECWP)，根据 ECWP 的大小来确定可隐藏的秘密信息比特数. 最后，根据编码后的秘密信息，选择当前的单词，从而生成含密文本. PPLM-Stega 能够在生成语义连贯可读性高的长文本同时，具有较高的隐写容量.

Yi 等[96]基于 BERT 生成模型和 Gibbs 采样策略[97]提出了一种文本生成式隐写方法 ALiSa(Acrostic Linguistic Steganography Based on BERT and Gibbs Sampling)，该方法直接将秘密文本的单词隐藏到生成文本的指定位置中以隐蔽地传递秘密文本. 在隐写过程中，发送方首先根据秘密文本中的单词和给定的位置密钥，利用 BERT 生成模型生成初始文本，在生成的初始文本中的特定位置单词使用掩膜代替，并且 BERT 为每个掩膜位置计算了单词的条件分布概率；然后根据每个掩膜位置的条件概率和 Gibbs 采样策略，从秘密文本中选择合适的单词，放置到文本的掩模位置；最后通过多次迭代计算，利用 BERT 模型在保证生成质量的条件下生成含密文本. 在提取过程中，接收者可以根据位置密钥直接从含密文本中提取秘密信息.

最近，一些研究者利用神经网络对古诗词进行建模，提出了基于神经网络的古诗词生成式隐写方法. Qin 等[98]提出了一种基于绝句生成的文本生成式隐写方法. 该方法利用基于注意力机制的 Seq2Seq 模型[99]生成符合设定主题和音律的绝句，并在生成绝句的过程中将秘密信息隐藏到绝句中，从而生成含密文本. 在训练绝句生成模型过程中，该方法将诗句的平仄和音律信息作为约束，使得模型能够生成高质量的文本. 在隐写阶段，发送方将秘密信息分为两段，根据第一段秘密信息确定主题词和音律等参数，将参数输入到生成模型生成候选诗句集；然后，根据第二段秘密信息从候选诗句中选择相应的诗句以构成含密文本. 在提取阶段，接收方使从含密文本中提取相关参数，从而根据参数提取第一段秘密信息，将参数输入到生成模型以生成候选诗句集；然后，从候选诗句集中检索到含密文本的每句候选诗句，根据候选诗句在候选诗句集中的序号提取到第二段秘密信息，最终拼接两段秘

表 8　基于神经网络模型的文本生成式隐写各方法对比

代表方法	主要思路	优点	缺点
RNN-Stega[88]	利用 RNN 对下一个单词进行预测，将预测的单词的条件概率编码为二进制比特流，以此选择对应的单词，将整个秘密信息隐藏到生成的文本中	隐写容量较高(每个单词隐藏 1-5 位)	没有充分考虑自然文本与生成的含密文本之间的总体统计分布差异，文本生成质量相对较低
GAN-TStega[89]	利用 GAN 网络对不同类型文本数据集的对抗训练以生成高质量文本，在文本生成过程中隐藏秘密信息	隐写容量较高(每个单词隐藏 1-5 位)，在 RNN-Stega[88]基础上，进一步提高文本生成质量	
VAE-Stega[91]	从隐空间中采样一个隐向量，在该隐向量的约束下，使用 AutoEncoder 的解码器根据秘密信息选择对应的候选单词生成含密文本	隐写容量较高(每个单词隐藏 1-5 位)，在 RNN-Stega[88]基础上，进一步提高文本生成质量	生成的含密文本的语义是随机的、不可控的
Zhou 等[93]	利用 LSTM 对当前单词进行预测，随机选择条件概率值在一定范围的单词组成候选池，从而根据秘密信息在候选池中选择相应单词生成含密文本	隐写容量较高(每个单词隐藏 1-3 位)，在 RNN-Stega[88]基础上，进一步提高文本生成质量	
PPLM-Stega[94]	计算其他候选词的概率与最高概率候选词的比例，选择该比例大于预设的阈值的单词添加到候选池，并根据秘密信息从候选池中选择相应的单词生成含密文本	隐写容量较高，可以生成长文本	在高隐写荷载的条件下，难以保证含密文本的生成质量，隐蔽通信的安全性有待提高
ALiSa[96]	在使用 BERT 生成文本的过程中，根据生成文本中指定位置的单词分布条件概率和秘密信息中选择相应的单词，隐藏到文本的指定位置中以生成含密文本	隐写容量较高	采样策略需进一步优化
Qin 等[98]	将秘密信息分为两段，首先根据第一段秘密信息生成候选诗句，然后根据第二段秘密信息从候选诗句中选择相应的诗句以构成含密文本	隐写容量较高，可生成高质量的绝句诗	只能生成特殊题材的文本(绝句)，适用性较低
SongNet[100]	将秘密信息分为两段，首先根据第一段秘密信息确定待生成宋词的词牌信息，然后根据第二段秘密信息确定待生成宋词的关键字、韵律和押韵字符，从而生成具有相应音律信息的含密文本	隐写容量较高(每个词句隐藏 18-21 位)，可生成高质量的宋词	只能生成特殊题材的文本(宋词)，适用性较低

密信息获得完整的秘密信息.

Qin 等[100]提出了 SongNet 方法，首先构建秘密信息与宋词的音律信息的映射关系，然后基于改进的 BERT 模型，在利用设定的音律生成宋词的过程中，从而实现秘密信息在宋词中的隐藏. 在隐写阶段，发送者将秘密信息分为两段，首先，发送方根据第一段秘密信息确定待生成宋词的词牌信息和格式；然后，根据第二段秘密信息确定待生成宋词的关键字、韵律和押韵字符，从而利用改进的 BERT 模型根据确定的词牌信息生成具有相应音律信息的含密文本. 在提取阶段，接收方首先提取含密文本中词牌信息，在词牌信息表检索以获得对应的词牌格式，从而可以根据词牌格式模板恢复第一段秘密信息；然后利用提取的词牌格式从含密文本中提取出关键字、韵律以及押韵字符以恢复第二段秘密信息；最后将两段秘密信息拼接获得完整的秘密信息.

基于神经网络模型的文本生成式隐写的各方法的对比如表 8 所示. 随着深度学习技术的快速发展，文本生成式隐写方案的生成质量得到了显著的提升.

相比传统的文本嵌入式隐写方案，文本生成式隐写方案能够抵抗隐写分析工具的检测，从而保证了含密文本传递过程中的隐蔽性. 然而，现有的模型难以对自然文本完美地建模，导致所生成的含密文本与自然文本的统计特性仍然存在一定差距，含密文本生成质量有待进一步提高，尤其当隐写荷载较高时，难以保证含密文本的生成质量.

4 音频生成式隐写方案

随着心理声学模型的发展，研究者们能够用数学模型准确地描述音频，这样为音频生成式隐写技术的发展奠定了基础[30, 101]. 本节根据秘密信息在生成音频载体过程中的隐藏规则，将现有的音频生成式隐写方案分为基于旋律构造的音频生成式隐写方案、基于上下文推理的音频生成式隐写方案、基于隐空间映射的音频生成式隐写方案. 音频生成式隐写方案的分类如表 9 所示.

4.1 基于旋律构造的音频生成式隐写方案

由于旋律是音频的主要要素之一，研究者们以秘密信息为驱动生成音频的旋律，提出了基于旋律构造的音频生成式隐写方案.

Crawford 等[102]提出了一种音频生成式隐写方法. 在预处理阶段，该方法首先将秘密信息中的每个字符都转换为二进制比特流，然后根据二进制“0”和“1”分别选择两种不同的音频片段，并在每两段音频之间加入 0.15 秒的静音，从而生成秘密音频片段. 在隐写阶段，该方法从单声道原始音频中复制声道，并使用 Audacity 音频编辑器[102]将复制的声道与秘密音频片段混合，从而构成含密声道；然后将含密声道与原始声道组合为新的立体音轨，从而得到含密立体音频. 由于秘密信息只隐藏到含密音频的其中一个声道中，含密音频的生成质量较好，收听者听到的含密音频与原始的音频很相似，从而保证了秘密信息的隐蔽性.

表 9　音频生成式隐写各方法对比

方法类别	代表方法	主要思路	优点	缺点
基于旋律构造的音频生成式隐写方案	Crawford 等[102]	将秘密信息转换为秘密音频片段，并使用Audacity音频编辑器将秘密音频片段混合到原始音频的其中一条声道中，从而生成含密立体音频	具有较好的隐蔽性	隐写容量较低，提取秘密信息的准确率受制于环境噪声
	Stegibiza[103]	将秘密信息中的字符转化为摩尔斯码，将相应的摩尔斯码转换为节拍与标准节拍的差值，从而将秘密信息映射为音乐中轻微的节拍变化以重新构造音乐音频	具有较好的隐蔽性	隐写容量较低，秘密信息的提取需要额外的硬件设备
基于上下文推理的音频生成式隐写方案	AAG-Stega[104]	在音频生成过程中根据音符的条件概率分布对每一个音符的进行编码，根据秘密信息的比特流选择相应的音符生成含密音频	含密音频的生成质量较高	隐写容量较低
基于隐空间映射的音频生成式隐写方案	Yang 等[105]	首先将秘密视频转换为二进制比特流并按照 IEEE 754 标准分段编码组成隐向量，然后利用 WaveGlow 将隐向量映射到音频空间生成含密音频	隐写容量较高，提取率较高	隐空间检测器工具可以检测到隐写的存在
	Chen 等[106]	利用拒绝采样的策略将秘密信息映射为服从高斯分布的隐向量，并输入到 WaveGlow 以生成含密音频	隐写容量较高，提取率较高	–

由于人类难以察觉在音乐标准节拍的基础上 1-2bpm(beats per minute，即每分钟的节拍数)的轻微变化，而且大多数音乐的节拍与标准节拍有轻微差异．因此，构造与标准节拍有轻微差异节拍的音频来传递秘密信息具有较好的隐蔽性．Szczypiorski 等[103]介绍了一种名为 Steglbiza 的方法，通过重新构造在音乐中音频的节拍来隐藏信息．Stegibiza 方法首先将秘密信息中的字符转化为摩尔斯码，并根据摩尔斯码与节拍变化的映射规则，将相应的摩尔斯码转换为节拍与标准节拍的差值，从而将秘密信息映射为音乐中轻微的节拍变化；然后根据节拍的变化重新构造音乐，从而达到在音乐音频中隐蔽地传递秘密信息的目的．

基于旋律构造的音频生成式隐写方案在通过构造音频过程中隐藏信息，然而为了保证根据秘密信息构造的新音频与自然音频尽可能相近，基于旋律构造的音频生成式隐写方案的隐写容量普遍较低．

4.2　基于上下文推理的音频生成式隐写方案

部分研究者提出了基于上下文推理的音频生成式隐写方案，利用神经网络学习真实音频的统计分布特性，并根据秘密信息生成符合该统计分布特性的含密音频载体．

Yang 等[104]提出了一种基于音频自动生成的隐写方法 (Automatic Audio Generation-based Steganography, AAG-Stega). AAG-Stega 方法首先利用 RNN 网络将音频建模为每个时刻音符条件概率的乘积；然后在音频生成过程中根据音符的条件概率分布对每一个音符的进行编码；最后根据秘密信息的比特流选择相应的音符构成含密音频．为了提升该网络生成长音频的质量，AAG-Stega 使用结合了门控机制、注意力机制以及回看机制(Lookback Mechanism)的 RNN 网络来生成音频序列．然而，当需要生成较高质量的音频时，AAG-Stega 的隐写容量较低. AAG-Stega 的隐写模型如图 15 所示．

4.3　基于隐空间映射的音频生成式隐写方案

基于隐空间映射的隐写方案将秘密信息编码为隐空间中的隐向量，然后利用可逆网络将该隐向量转换为对应的音频作为含密音频．

Yang 等[105]提出了一种基于流模型的音频隐写方法，将秘密视频隐藏到音频中．该方法首先使用图像压缩网络[107]将秘密视频中的每一帧转换为二进制比特流；然后按照 IEEE 754 标准[108]将秘密视频的比特流分段编码为浮点数并组成隐向量，其中 32 位浮点数中有 12 位为固定值以限制浮点数的值域，其余 20 位用于隐藏秘密视频对应的比特位；最后使用 WaveGlow[109]将隐向量映射到音频空间生成含密音频．提取过程与隐藏过程相反，首先通过 WaveGlow 将含密音频转换为隐向量；然后从隐向量中恢复出视频的比特流；最后使用图像压缩网络的解码器将比特流解压为秘密视频．该方法能有效地将视频隐藏在音频中并确保含密音频的质量．

如果隐写前后音频的分布特性相同，即具有分布保持性，那么理论上该隐写方法具有抗隐写分析性能．因此，隐写方法的分布保持性是隐写安全的基础[49, 110]．受 WaveGlow 的启发，Chen 等[106]设计了一种基于拒绝采样的分布保持的隐写方法．该方法利用拒绝采样的策略将秘密信息映射为服从高斯分布的隐向量，并输入到 WaveGlow 以生成含密音频．在提取秘密信息时，利用 WaveGlow 的反函

数即可将含密音频转换为隐向量, 然后对隐向量进行逆映射便可获得秘密信息. 由于 WaveGlow 的可逆性, 该方法可以保证秘密信息从含密音频中的正确提取. 在该方法中, 服从高斯分布的隐向量是由秘密信息通过拒绝采样的策略映射得到的, 该过程等价于对高斯分布随机采样. 因此在使用同一 WaveGlow 生成网络模型的条件下, 生成的含密音频与生成的非含密音频的分布是接近一致的.

基于隐空间映射的音频生成式隐写方案可以将秘密信息转换为高维的隐向量从而保证了含密音频的隐写容量. 同时, 由于可逆模型的使用, 该方案可以确保秘密信息的准确提取. 然而, 与基于隐空间映射的图像生成式隐写方案存在同样的问题, 大部分方法生成的含密音频与生成的非含密音频的分布仍然存在一定程度的差异, 特定的隐空间检测器工具仍有可能检测到含密音频中隐写信息的存在, 其安全性存在一定的隐患. 目前, 部分研究工作(如 Chen 等[106]的基于 WaveGlow 的生成式隐写方法)已经进行了安全性理论证明, 在理论上能够实现安全的隐写, 然而基于其他生成网络模型的隐写方法的安全性仍然需要进一步进行理论证明.

音频生成式隐写方案通过各种映射规则, 在生成音频的过程中将秘密信息编码到含密音频. 然而, 现有的音频生成式隐写方法的含密音频生成质量仍然有待提高. 此外, 含密音频在传递过程中易受到噪声等攻击, 对秘密信息提取有较大的影响, 因此这些方法的鲁棒性通常较低.

5 社交网络行为生成式隐写方案

社交网络如微信、微博和推特等越来越普及, 极大地影响了人们的生产生活方式, 已经成为人们日常生活种不可或缺的一部分. 社交网络环境下的行为生成式隐写受到了广泛的关注. 根据行为的类型, 现有的社交网络行为生成式隐写方案主要分为: 基于社交互动行为的生成式隐写方案、基于商品推荐行为的生成式隐写方案以及基于游戏操作行为的生成式隐写方案, 如表 10 所示.

5.1 基于社交互动行为的生成式隐写方案

在社交网络中, 如聊天、对动态点赞等互动是最为常见的行为. 研究者们利用社交网络上的互动行为传递秘密信息, 提出了基于社交互动行为的生成式隐写方案.

Liu 等[111]提出了一种基于社交网络聊天软件语音消息编码的生成式隐写方法, 利用社交网络聊天消息中语音消息的时长来表示秘密信息, 从而可以在社交网络中的通信双方之间传递秘密信息. 该方法将秘密信息进行加密和行程编码(Run Length Encoding)并映射为相应的时间长度值. 在社交网络的通信传输过程中, 发送方根据秘密信息的映射规则, 发送具有特定时长的语音消息, 接收方通过约定的映射规则, 根据发送的语音消息中提取出秘密信息. 但该方法的隐写容量较低, 只能传递隐写密钥、密码等较短的秘密信息.

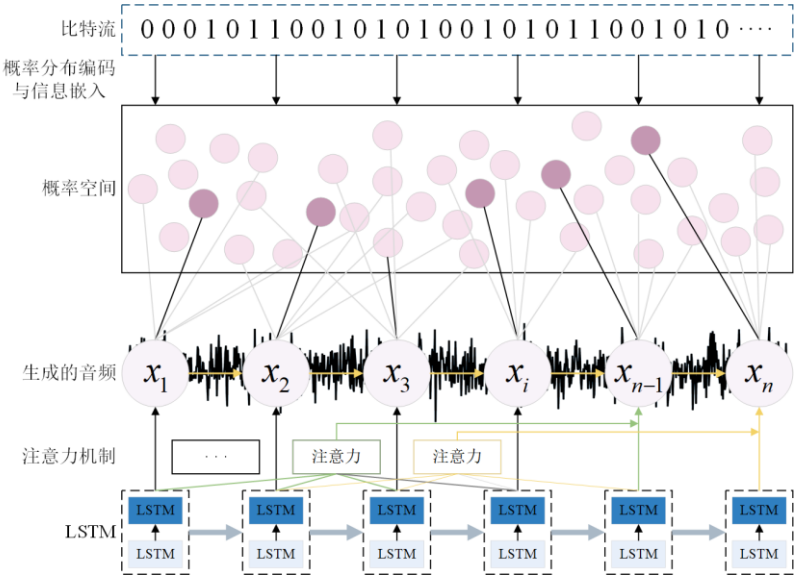


图 15 AAG-Stega 隐写模型

表 10　社交网络行为生成式隐写各方法对比

方法类别	代表方法	主要思路	优点	缺点
基于社交互动行为的生成式隐写方案	Liu 等[111]	将秘密信息进行加密和行程编码，以利用聊天消息中语音消息的时长来表示秘密信息	鲁棒性较高，抗统计分析能力较好	隐写容量较低
	Zhang[112]	发送者按一定的规则根据秘密信息对某些动态进行点赞操作，从而利用点赞行为传递秘密信息	鲁棒性较高	隐写容量较低，无差别的大量点赞易引起第三方怀疑
	Hu 等[113]	建立了一个用户行为相关性模型以预测发送者对每一条动态的点赞概率，根据点赞概率决定点赞行为，并通过点赞行为传递秘密信息	鲁棒性较高，可以减少异常点赞行为	隐写容量较低
基于商品推荐行为的生成式隐写方案	Zhou 等[114]	根据秘密信息从载体数据之间的转移概率图中每次选取一组高度相关的载体数据构成发送序列进行秘密信息的传递	隐写容量较高	生成长序列时，序列生成质量有所下降
基于游戏操作行为的生成式隐写方案	StegoRogue[115]	按一定的步骤创建二维游戏地图，把需要隐藏的信息字符用地图空间中的物体来表示	鲁棒性较高	隐写容量较低，应用场景有限(只限二维游戏地图)
	Hernandez-Castro 等[116]	根据设定的阈值，对给定位置上落子的好坏程度进行排序，并根据秘密信息中相应位置的数据，选择对应的走法	鲁棒性较高	隐写容量较低，应用场景有限(只限围棋游戏)
	Ou 等[117]	根据秘密信息与俄罗斯方块形状的构建映射关系，将秘密信息编码为俄罗斯方块的形状	鲁棒性较高	隐写容量较低，应用场景有限(只限俄罗斯方块游戏)
	Mahato 等[118]	建立扫雷网格中地雷的位置与秘密信息的映射关系，通过地雷的位置信息对秘密信息进行编码	鲁棒性较高	隐写容量较低，应用场景有限(只限扫雷游戏)

在社交网络中，用户可以为好友的动态点赞、评论，好友也可以观测到用户之间的点赞信息或评论内容，因此可以利用社交网络上好友的点赞行为传递秘密信息．Zhang[112]提出了一种基于社交网络中用户点赞行为的隐写方法，将秘密信息编码为社交网络中的点赞行为．发送者和接收者事先约定一定数量的动态作为观测目标，在秘密信息传递过程中，发送者按一定的规则根据秘密信息对某些动态进行点赞操作；接收者可以通过观测发送者对动态的点赞行为，根据约定的规则反推出秘密信息．这种隐写方法抛弃大多数人使用多媒体数据作为载体的方式，将秘密数据转换成社交网络中个人的行为，只需要消耗少量的网络资源就可以创建一个隐蔽的通道来传输秘密信息．然而，由于某些动态不适合点赞，因此用户无差别的大量点赞易引起第三方怀疑．为了解决以上问题，Hu 等[113]对上述方法加以改进，提出了一种用户行为相关性模型以预测发送者对每一条动态的点赞概率，如图 16 所示．该模型规定给某个动态点赞的概率与当前社交网络中其他正常用户的点赞数量成正比．因此，用户可以根据模型对相应动态的点赞概率决定点赞行为，并通过点赞行为传递秘密信息，从而提高了点赞行为隐写的安全性．

基于社交互动行为的生成式隐写方案可以通过正常的社交互动行为传递秘密信息，传统的多媒体攻击难以影响双方互动行为的传输，因此其鲁棒性较高．然而，利用互动行为可传递的秘密信息数量十分有限，因此该方案的隐写容量很低，难以用于实际应用．

5.2　基于商品推荐行为的生成式隐写方案

在复杂的社交网络环境中需要传输的载体种类较多，针对单一类型载体所设计的隐写方案已经不能很好地满足隐写的需求．为了在社交网络环境下实现隐蔽通信，一些研究者提出了基于商品推荐行为的生成式隐写方案．然而，在商品推荐行为应用中，需要传输载体数据是一系列商品编码、订单编号等结构化数据，这类数据冗余空间较小，如果直接修改这些数据来隐藏信息将容易导致秘密信息的暴露．针对这些问题，Zhou 等[114]提出了一种在商品推荐应用中的基于概率图学习的生成式隐写方法．该方法通过学习现有载体内容之间的关系，构造包含秘密信息的载体序列，如图 17 所示．具体来说，首先根据用户和载体数据之间的交互关系学习一个包含网络环境中载体数据之间的转移概率图；然后，根据秘密信息，发送者从转移概率图中每次选取一组高度相关的载体数据构成发送载体序列，并依次发送给接收者；最后，接收者将接收到的序列与转移概率图进行匹配即可恢复秘密信息．实验结果表明，该方法生成的含密发送载体序列与真实环境中的发送载体序列非常接近，同时该方法具有较高的隐写容量，并且可以准确地提取出秘密信息．

5.3　基于游戏操作行为的生成式隐写方案

各种类型的游戏正成为现代的休闲生活中不可或缺的娱乐方式．研究者们通过对游戏中的行为

进行编码来传递秘密信息, 提出了基于游戏操作行为的隐写方案.

Gibbs 等[115]提出了 StegoRogue, 将秘密信息隐藏到电子游戏地图中. 在生成电子游戏地图的过程中, 发送方首先根据设定的规则, 将秘密信息的字符放置在地图的相应位置; 然后, 为了消除有关秘密信息字符的痕迹, 发送方使用符合游戏设定的物体代表字符, 从而将秘密信息编码到该电子游戏地图中.

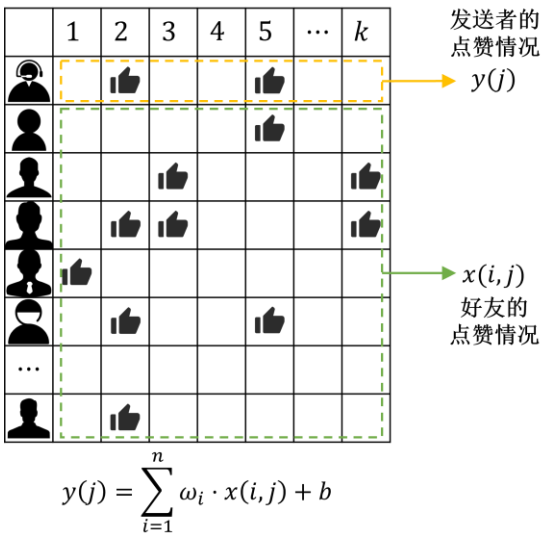


图 16 Hu 等[113]提出的用户行为相关性模型

Hernandez-Castro 等[116]提出了一种在围棋游戏中隐藏秘密信息的方法. 该方法适用于两种不同场景, 即在创建新游戏隐藏秘密信息以及将秘密信息隐藏到已经开局的游戏中. 在第一种场景下, 发送方根据设定的阈值, 对放置在给定位置上落子的好坏程度进行排序, 并根据秘密信息中相应位置的数据, 选择对应的走法, 从而将秘密信息隐藏到围棋开局中. 在第二种场景下, 发送方则需要先结合游戏开始时间和时长, 以及游戏的注释信息决定落子的好坏程度, 然后使用与第一种场景相同的编码方式对秘密信息进行编码.

Ou 等[117]提出了一种基于在线俄罗斯方块游戏操作的隐写方法. 由于在每场游戏中, 每位玩家所操作的在线俄罗斯方块游戏都能够产生一组具有不同形状的俄罗斯方块序列. 因此, 该方法将不同形状的俄罗斯方块编码为不同的数字, 并根据秘密信息与俄罗斯方块形状的构建映射关系, 从而可以将秘密信息编码为俄罗斯方块的形状, 实现在俄罗斯方块游戏中传递秘密信息.

Mahato 等[118]提出了一种将秘密信息隐藏到“扫雷”游戏的隐写方法. 该方法首先将秘密信息转换为二进制字符串, 并将字符串进行分组; 然后根据分组中的字符串定位地雷的位置, 构建每个分组的扫雷网格, 并建立扫雷网格中地雷的位置与秘密信息的映射关系, 从而可以通过地雷的位置信息对秘密信息进行编码. 同时, 该方法可以结合不同的矩阵遍历规则对秘密信息进行编码, 提高秘密信息的安全性.

基于游戏操作行为的生成式隐写方案可以有效抵抗常规多媒体攻击, 具有较高的鲁棒性. 然而, 游戏操作不能对大规模的信息进行编码, 因此基于游戏操作行为的生成式隐写方案的隐写容量较低; 该方案需要通信双方掌握一定门槛的游戏规则, 且部分隐写方法的实现依赖多位玩家同时在线, 限制了该方案的适用场景.

由于社交网络环境下的正常行为隐蔽性强, 难以被隐写分析工具检测到, 因此社交网络行为生成式隐写方案与传统的多媒体生成式隐写方案相比具有更高的鲁棒性. 然而, 大部分的社交网络行为能够承载的信息量十分有限, 导致该方案的隐写容量低于传统的多媒体生成式隐写方案; 此外, 该方案的实现依赖于特定的应用场景, 难以广泛地实际使用.

6 实验对比与分析

本章主要对具有代表性的图像生成式隐写和文本生成式隐写方案的实验结果进行对比与分析.

6.1 图像生成式隐写方案的实验结果分析

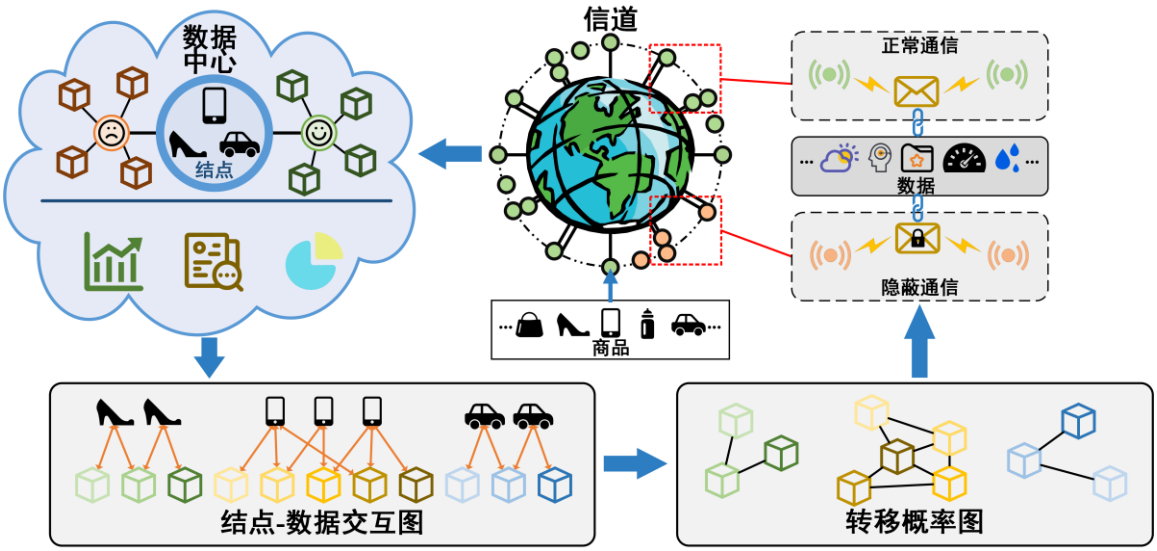


图 17　Zhou 等[114]提出的基于概率图学习的隐写模型

根据各隐写方法的隐写容量、秘密信息的提取率以及在多种噪声(如添加 3%的椒盐噪声[55]、强度值为 25 的椒盐噪声[56]、方差为 0.1 和 0.01 的高斯噪声[57, 58]等攻击)攻击下的提取率，本文对基于像素定义的图像生成式隐写方案、基于低层特征映射的图像生成式隐写方案、基于高层特征关联的图像生成式隐写方案以及基于隐空间映射的图像生成式隐写方案的实验结果进行分析. 此外，本文还对典型的生成模型生成图像载体的稳定性进行评估. 为评估图像隐写方法的隐写容量，我们使用每像素嵌入比特数(bit per pixel, *bpp*)作为隐写容量的指标，其表示每像素隐藏秘密信息的位数，可通过以下公式计算:

$$bpp = \frac{N}{W \times H \times C} \tag{14}$$

其中，N 表示隐藏到图像中的秘密信息位数，W 表示图像的宽度，H 表示图像的高度，C 表示图像的通道数. 表 11-14 为部分图像生成式隐写方案的隐写算法性能对比.

6.1.1　基于像素定义的图像生成式隐写方案

Pixel-Stega 方法使用 MNIST[119]、Frey Faces[50]以及 CIFAR-10[120]数据集与 Yang 等[47]的隐写方法进行了对比实验. 为了进行定量评估，使用 Pixel-Staga 生成了 5,000 幅载体图像和 5,000 幅含密图像并使用 Yang 等[47]生成了 5,000 幅含密图像. 根据实验结果，相比 Yang 等[47]的隐写方法在三个数据集上均为 1.00*bpp* 的隐写容量, Pixel-Stega 方法的隐写容量最高可以达到 4.30*bpp*. 虽然 Pixel-Stega 方法在 MNIST 数据集上的隐写容量低于 Yang 等[47]方法，但从整体上看，Pixel-Stega 方法的隐写容量大于 Yang 等[47]方法的隐写容量. 这是因为 Pixel-Stega 方法能够根据图像像素的熵自适应地隐藏秘密信息. 由于 MNIST 数据集中的图像均为黑白图像，大多数像素的熵较低，而 Frey Faces 和 CIFAR-10 数据集中的图像更加多样化，它们的像素具有较大的熵，Pixel-Stega 方法可以充分隐藏秘密信息. 实验结果表明，Pixel-Stega 方法能够显著提高秘密信息的隐写容量.

针对 Liu 等[52]的隐写方法使用 LFW 数据集[121]进行相关实验. 该方法的受损区域由受损图像中未受损部分和秘密信息填充部分组成,受损区域的面积约占整个图像区域的 12.5%，当隐写容量为 2.50×10^{-3}*bpp* 时，提取率约为 58%; 在隐写方法[53]的实验中，受损区域完全是由秘密信息填充部分组成，受损区域的面积占整个图像区域的 90%，当隐写容量为 5×10^{-1}*bpp* 时，提取率约为 95%; 而隐写方法[54]使用 CelebA 和 LSUN[122]数据集进行相关实验，与隐写方法[53]相同，该方法受损区域完全是由秘密信息填充部分组成，受损区域的面积占整个图像区域的 95%，将 20 字节的秘密信息隐藏到卡丹格中，最大的隐写容量约为 9.80×10^{-3}*bpp*，最大的提取率为 100%. 实验结果表明，对于使用卡丹格对像素进行预定义的图像生成式隐写方法，缩小未受损区域的面积有利于提升秘密信息的提取率.

基于像素定义的图像生成式隐写方案的性能对比如表 11 所示. 在基于像素定义的图像生成式隐写方案的性能对比实验中，各方法均使用添加 3%的椒盐噪声[55]攻击方式进行抵抗噪声攻击的鲁棒性对比. 以上方法的隐写容量整体上处于较高的水平，这是因为基于像素定义的图像生成式隐写方案将秘密信息映射为像素，而图像中的像素所能承载的信息量较大. 然而，由于像素值对噪声攻击的较为敏感，以上方法在受到攻击后的提取率会大大降低. 因此，此类方法的鲁棒性较低，难以抵抗各种噪声攻击.

6.1.2 基于低层特征映射的图像生成式隐写方案

Otori 等[59]的隐写方法用 EPSON PX-G5100 彩色喷墨打印机在超薄的 A4 纸上在 2×2 英寸的正方形区域内打印 200×200 像素的数据编码后的纹理图像, 并打开手机的摄像头以 480×640 像素的微距模式拍摄打印的图像. 通过对 4 种不同纹理的例子进行 10 次测试, 该方法可以隐藏 200-800 位信息, 因此其最大隐写容量约为 $2\times10^{-3}bpp$, 经过打印和拍摄攻击后提取率约为 90%.

Stego-texture 方法使用 7 种大理石纹理图案的样本以隐藏秘密信息, 其将秘密信息转换为纹理图像, 并从 512×512 像素缩小到 180×180 像素, 然后把该图像编码为 777,600 位的二进制字符串以隐藏到大理石纹理图案中. 经过检测, 最大隐写容量约为 $6\times10^{-3}bpp$, 最大提取率为 70%.

Wu 等[62]的隐写方法通过对比四种不同的源纹理图案作为测试图案得到实验结果, 结果显示源纹理图案的分辨率越大, 其能提供的总隐写容量越小. 该方案中每幅图像的隐写容量最大可达 34,398 位, 换算得到容量为 $3.28\times10^{-2}bpp$, 并且能够准确地提取信息. 与 Otori 等[59]和 stego-texture 方法相比, 该方法隐写容量得到了显著提高.

Li 等[56]的隐写方法根据特定的阈值构造了 1,000 幅指纹图像, 分别具有 300×300 和 500×500 像素两种尺寸. 其中每个指纹图像都拥有二值化、稀疏化和灰度三种不同形式. 实验结果表明, 虽然该方法的提取率可以达到 100%, 但指纹图像中细节点的数量有限, 为了保证图像的质量, 隐写容量受到了限制, 约为 $1.70\times10^{-3}bpp$.

Zhou 等[66]的隐写方法在实验中建立了包含 500 幅分辨率为 256×256 像素的彩色真实山脉图像库作为训练集; 使用训练好的轮廓−图像可逆变换模型, 在每个轮廓点中隐藏长度不同秘密信息, 生成 10,000 幅含密图像作为测试集. 具体来说, 在每个轮廓点中分别隐藏 1-8 bit 长度不同的秘密信息, 从而生成对应的 8 类含密图像, 每类含密图像生成 1,250 幅, 一共得到 10,000 幅含密图像. 由于从作为显式特征的轮廓信息到图像的映射过程更易于学习和训练, 因此, 与现有的生成式图像隐写方法相比, Zhou 等[66]的隐写方法很容易训练出相应的图像生成网络和秘密信息提取网络, 从而可以获得较高的隐写容量(当生成图像图像的尺寸为 256×256 时, 隐写容量约为 $4\times10^{-3}bpp$)和秘密信息提取的准确率(98.55%).

表 11 基于像素定义的图像生成式隐写方案的性能对比

代表方法	数据集	隐写容量	提取率	噪声攻击后的提取率
Yang 等[47]	MNIST	$1.00\pm0.00bpp$	100%	41.76%
	Frey Faces	$1.00\pm0.00bpp$	100%	52.96%
	CIFAR-10	$1.00\pm0.00bpp$	100%	48.62%
Pixel-Stega[50]	MNIST	$0.58\pm0.19bpp$	100%	48.07%
	Frey Faces	$4.05\pm0.14bpp$	100%	58.39%
	CIFAR-10	$4.30\pm0.85bpp$	100%	50.81%
Liu 等[52]	LFW	$2.50\times10^{-1}bpp$	58%	32.94%
隐写方法[53]	LFW	$5\times10^{-1}bpp$	95%	57.65%
隐写方法[54]	CelebA、LSUN	$9.80\times10^{-3}bpp$	100%(最大)	60.18%

基于低层特征映射的图像生成式隐写方案的性能对比如表 12 所示. 在基于低层特征映射的图像生成式隐写方案的性能对比实验中, 除 Otori 等[59]的隐写方法不进行抵抗噪声攻击、Li 等[56]的隐写方法添加强度值为 25 的椒盐噪声攻击, 其余方法均添加 3%的椒盐噪声[55]攻击. 根据各类方法在攻击后的最大提取率可以得到以下结论. 图像的纹理、轮廓等特征与像素相比更加稳定, 从而可以提升在噪声攻击下的鲁棒性. 然而, 其所能承载的信息容量却有所降低, 导致该方案的隐写容量普遍低于基于像素定义的生成式隐写方案的隐写容量.

6.1.3 基于高层特征关联的图像生成式隐写方案

Cao 等[57]的隐写方法使用从 Getchu①上采集的动漫头像进行训练, 将 $N\times N$ 个较小的动漫角色组成含密图像. 当 $N=1$ 时, 每个含密图像可以表达 14 位秘密信息. 由于利用 GAN 网络生成的动漫角色以 8×8 的形式输出更为常见, 因此该方法将 N 设

① Getchu: http://www.getchu.com

定为 8, 相应的隐写容量为每个载体可以隐藏 896 位信息(896 bits/carrier).

为了验证方法的可行性, SSS-GAN 方法使用 MNIST[119]、CIFAR-10[120]和 CIFAR-100[120]数据集来训练模型. 该方法等效于将 m 位秘密信息映射到图片中, 在实验中将 m 设置为 6. SSS-GAN 方法的隐写容量取决于含密图像中包含的语义标记的数量, 根据实验结果, 该方法的隐写容量超过 7.30×10^{-4} *bpp*. 并且, 由于构建了秘密信息和图像语义信息之间的映射关系, SSS-GAN 方法可以在不同图像数据集的训练以达到模型的收敛, 从而可以恢复的秘密图像的最大准确率为 97.64%, 其最大的隐写容量为 $\frac{\log_2 10}{256\times256}\approx5.07\times10^{-5}\,bpp$.

基于高层特征关联的图像生成式隐写方案的性能对比如表 13 所示. 在基于高层特征关联的图像生成式隐写方案的性能对比实验中, 各方法均使用添加高斯噪声攻击方式进行抵抗噪声攻击的鲁棒性对比. 其中, Cao 等[57]的隐写方法和 STNet[69]隐写方法使用方差为 0.1 的高斯噪声, SSS-GAN 隐写方法使用方差为 0.01 的高斯噪声, Li 等[74]的隐写方法添加的高斯噪声遵循文献[74]中的设置. 从整体

表 12 基于低层特征映射的图像生成式隐写方案的性能对比

代表方法	数据集	隐写容量	提取率	噪声攻击后的提取率
Otori 等[59]	4 种不同纹理图像	2×10^{-3}*bpp* (打印和拍摄攻击后)	90% (打印和拍摄攻击后)	-
stego-texture[61]	7 种大理石纹理图案的样本	6×10^{-3}*bpp* (最高)	70% (最高)	65.27%
Wu 等[62]	四种不同的源纹理图案	3.28×10^{-2} *bpp* (最高)	100% (最高)	82.48%
Li 等[56]	1,000 幅指纹图像	1.70×10^{-3} *bpp* (最高)	100% (最高)	100%
Zhou 等[66]	500 幅彩色真实山脉图像库	4×10^{-3}*bpp*	98.55%	79.33%

使提取器能够准确提取秘密信息.

STNet 方法使用 COCO 数据集[123]作为载体图像的数据集, 并将从 wikiart.org 获取的图片数据集作为参考图像的数据集. 将载体图像和参考图像的大小调整为 512×512 像素, 在图像中随机裁剪大小为 256×256 像素的区域. 为评估秘密信息的提取率, STNet 方法随机选择了 10,000 幅载体图像和参考图像, 生成 10,000 幅含密图像作为测试图像. 实验结果显示, STNet 方法可以成功提取 99.80%的秘密信息, 并且能够在每个像素隐藏 6×10^{-2} 位信息生成任意大小的含密图像.

Li 等[74]的隐写方法使用 MNIST 数据集[119]中的图像作为秘密图像, 并收集了 30,000 幅包含一朵花的图像作为载体图像, 从 seepretty_anime_face 和 faces_datasets 数据集[74]中选择了 50,000 幅漫画图像作为反差图像. 为了评估隐写方法的性能, 在实验中采用三种不同大小(7×7, 14×14, 28×28 像素)的秘密图像, 并且将载体图像和参考图像调整为 256×256 像素. 实验结果显示, 该方案对噪声和滤波攻击有良好的鲁棒性. 由于 MNIST 数据集是 0-9 数字的集合, 可以表示 10 个数, 当秘密图像的大小为 28×28 像素, 构建的含密图像为 256×256 大小时, 上看, 由于图像高层特征比图像低层特征更加稳定, 不容易受到图像攻击(如添加噪声等)的影响, 因此该方案在受到噪声攻击后的提取率与原提取率差异较小, 从而具备较高的鲁棒性. 然而, 从图像中抽象出的高层特征所能承载的信息量较少, 该隐写方案的隐写容量会明显低于基于像素定义以及基于低层特征映射的生成式隐写方案.

6.1.4 基于隐空间映射的图像生成式隐写方案

Hu 等[75]的隐写方法使用 CelebA ① 和 Food101[124]数据集训练 DCGANs. 在实验中, 设置噪声与秘密信息的映射参数 σ 为 1-3, 从而评估不同情况下秘密信息的提取率. 该方法的提取率随着训练轮次数的增加而增加, 300 次训练后, 当 σ 为 1 时, 隐写容量为 $2.40\times10^{-2}bpp$ 时, 提取率达到约 96%; 当 σ 为 3 时, 隐写容量为 $7.30\times10^{-2}bpp$ 时, 在两个训练集训练下的提取率分别为 89%和 88%. 而 Li 等[76]的隐写方法的实验均基于 CelebA 数据集进行. 与 Hu 等[75]的隐写方法一致, 其设置系数 σ 来

① CelebA 数据集: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

评估秘密信息的提取率. 当σ为 1 时, 提取率为 98%; 当σ为 3 时, 提取率约为 84%, 在$\sigma=1$和$\sigma=3$的条件下, 虽然其隐写容量与Hu等[75]方法相同, 但提取率均高于 Hu 等[75]方法. 实验结果表明, 由于 Li 等[76]方法使用 WGAN-GP 替代了 DCGAN, 秘密信息的提取率得到了有效提升.

GSN 方法使用 CelebA 数据集和 LSUN 中的卧室场景图像子数据集. 实验结果表明, 虽然随着隐写荷载的增加, 秘密信息的提取准确率和生成质量均有所下降, 但是该方法仍然可以提供最大 8*bpp* 的隐写容量以及 97.53%的秘密信息提取准确率.

IDEAS方法使用的数据集分别来自LSUN数据集[122]中的卧室和教堂场景图像子数据集以及FFHQ数据集[125]中的人脸图像子数据集. 每一子集包括 70,000 幅随机选择的图像, 并将这些图像归一化为 256×256 像素的图像. 根据实验结果, 由于 IDEAS 方法可以利用图像结构特征的稳定性提升秘密信息的提取率, 在三个子数据集上, IDEAS 方法均获得最高 100%的提取率以及每幅图像 1536bits(7.81×10^{-3}*bpp*)的隐写容量.

S2IRT 方法在 CelebA-HQ 数据集上进行实验, 该数据集由 30,000 幅从 CelebA 数据集选出的高分辨率人脸图像. 在实验中, 每一幅图像均被缩放为 256×256 像素的大小以训练Glow模型. 实验结果表明, 由于位置编码的使用以及 Glow 模型提供的隐空间与图像空间可逆映射功能, S2IRT 方法的秘密信息提取率在隐写容量从 0.1*bpp* 到 4*bpp* 的范围内均能保持非常高的水平, 最大的提取率为 100%. 此外, 通过改进 S2IRT 方法的编码方式而形成的 SE-S2IRT(Separate Encoding-based S2IRT)方法, 其使用独立编码的策略代替位置编码的策略对秘密信息进行编码, 可以避免隐向量中单一元素的改变而导致的对整体秘密信息提取的影响, 从而可以提升隐写方法的鲁棒性. SE-S2IRT 方法在各种随机噪声攻击下, 秘密信息提取率仍然可以达到 91%.

基于隐空间映射的图像生成式的隐写方法性能对比如表 15 所示. 在基于隐空间映射的图像生成式隐写方案的性能对比实验中, 各方法均使用添加 3%的椒盐噪声[55]攻击方式进行抵抗噪声攻击的鲁棒性对比. 由于该隐写方案的实现基于从隐空间采样的方式, 一方面可以获得较高的隐写容量, 另一方面各种随机噪声会影响隐向量的值, 从而导致该方案与基于高层特征关联的图像生成式隐写方案相比有所降低.

表 13 基于高层特征关联的图像生成式隐写方案的性能对比

代表方法	数据集	隐写容量	提取率	噪声攻击后的提取率
Cao 等[57]	Getchu 上采集的动漫头像	896 bits/carrier	100%	96.38%
SSS-GAN [58]	MNIST、CIFAR-10、CIFAR-100	7.30×10^{-4} *bpp*	100%	99.67%
STNet[69]	COCO	6×10^{-2}*bpp*	99.80%	98.64%
Li 等[74]	MNIST、seepretty_anime_face、faces_datasets	5.07×10^{-5} *bpp* (最大)	97.64%	97.88%

我们基于 BigGAN[126]和 Glow[46]模型, 分别使用 ImageNet[127]和 CelebA[55]数据集训练模型. 每个训练后的模型生成 10,010 幅图像. 我们使用无参考图像空间质量评估器(Blind/Referenceless Image Spatial QUality Evaluator, BRISQUE)[128]以评估生成图像的质量. 根据 Mittal 等[128], BRISQUE 得分越低表示图像的质量越高. 根据实验结果, 虽然 BigGAN 和 Glow 模型能够生成质量较高的图像(即 BRISQUE 得分较低的图像), 但仍然也会生成部分质量较低的图像. 此外, BigGAN 和 Glow 模型生成的图像 BRISQUE 得分的方差分别为 9.31 和 6.71, 这表明这两类生成模型的图像生成质量不稳定. 由此, 我们认为现阶段含密载体生成质量的稳定性仍然不足. 部分实验结果如图 18 所示, 18(a)为 BigGAN 模型生成的部分图像, 18(b)为 Glow 模型生成的部分图像.

6.2 文本生成式隐写方案的实验结果分析

本文通过实验数据对基于马尔科夫模型和基于神经网络模型的文本生成式隐写方案的隐写能力进行评估. 表 15-16 为文本生成式隐写方案的隐写算法性能对比.

6.2.1 基于马尔科夫模型的文本生成式隐写方案

Shniperov 等[81]的隐写方法针对基于 1 阶、2 阶和 3 阶马尔科夫模型的隐写方法进行实验. 在实验中使用大量不同的自然文本序列构建马尔科夫模型, 并使用大量不同长度的秘密信息来实现该隐写

方法. 这些秘密信息由字母、数字和标点符号随机组成. 实验结果表明, Shniperov 等[81]的隐写方法的隐写能力取决于构建马尔科夫模型自然文本序列的字数.

Yang 等[82]的隐写方法在实验部分选择了文本样本. 在基于两个数据集分别训练生成器后, 使用训练好的生成器生成 1,000 个文本, 然后基于随机生成的 01 位流(01 bit stream)生成具有不同隐写容量的含密文本. 根据实验所得到的数据, GAN-TStega 方法能够在每个单词隐藏 1-5 位秘密信息.

表 14 基于隐空间映射的图像生成式隐写方案的性能对比

代表方法	数据集	隐写容量	提取率	噪声攻击后的提取率
Hu 等[75]	CelebA、Food101	2.40×10^{-2} *bpp* ($\sigma=1$)	96%	89.83%
		7.30×10^{-2} *bpp* ($\sigma=3$)	89%	86%
Li 等[76]	CelebA	2.40×10^{-2} *bpp* ($\sigma=1$)	98%	90.17%
		7.30×10^{-2} *bpp* ($\sigma=3$)	84%	82.50%
GSN[78]	CelebA、LSUN	8*bpp*	97.53%	88.62%
IDEAS[79]	LSUN、FFHQ	7.81×10^{-3} *bpp*	100%	90.42%
S2IRT[55]	CelebA-HQ	4*bpp*	100%	91%

Twitter[129]、IMDB[130]和 News①数据集作为训练集以训练模型, 并在构建马尔科夫模型之前对数据进行预处理. 根据实验结果分析, Yang 等[82]的隐写方法的秘密信息隐写能力较高. 该方法的可以在每个单词中隐藏 1-4 位的秘密信息.

为了评估算法的性能, Cistega 方法收集了《全宋词》[83]中的宋词作为算法的语料库, 并根据现代汉语的拼音来确定每个字的音律. Cistega 方法的隐写能力与每个词隐藏的哈夫曼编码的位数相关, 隐写率在 7-10%之间.

6.2.2 基于神经网络模型的文本生成式隐写方案

RNN-Stega 方法使用三种大规模文本数据集作为模型的训练集, 分别为: Twitter[129]、IMDB[130]和 News 数据集, 并在模型训练前对数据集中的文本进行预处理, 包括将所有单词转换为小写、删除特殊符号以及过滤低频单词等. RNN-Stega 方法通过实验分析了生成的文本中可以隐藏的信息容量, 并将其与部分文本隐写算法进行了比较. 实验结果显示, RNN-Stega 方法能够在每个单词中隐藏 1-5 位的秘密信息.

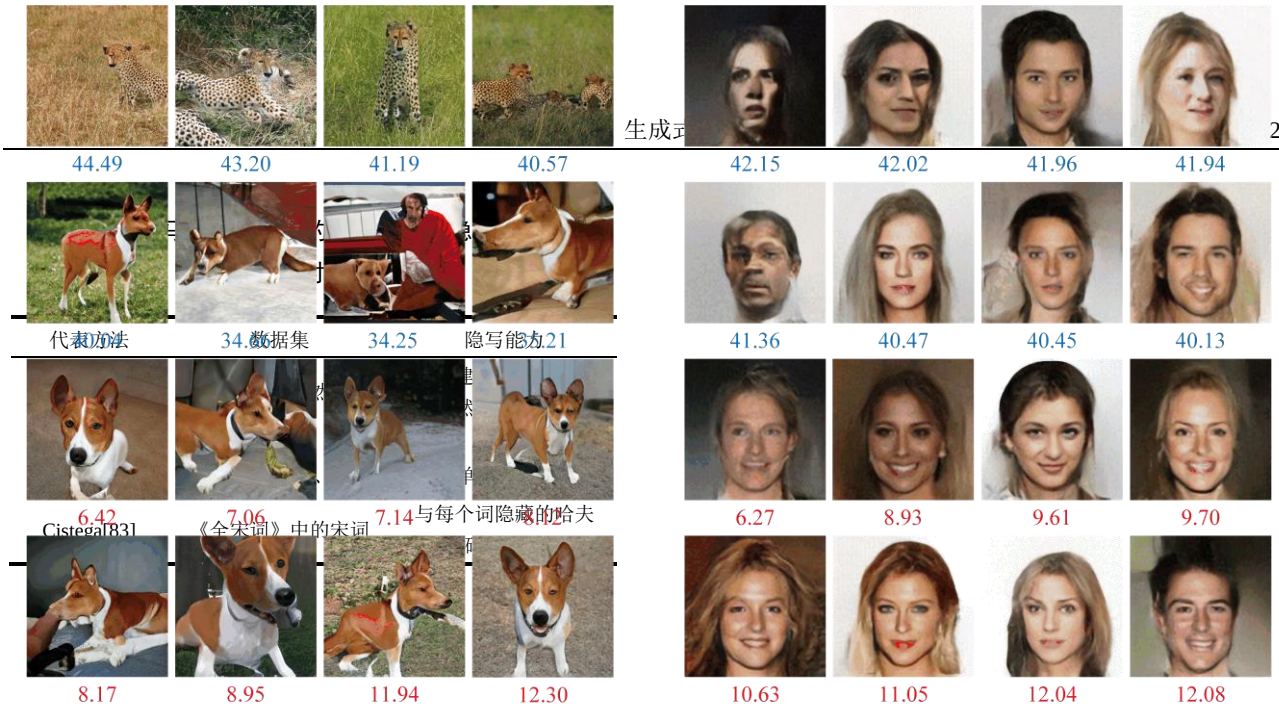
GAN-TStega 方法使用随机初始化的 LSTM 网络生成 10,000 个长度为 20 的文本序列作为数据样本并分别使用最大似然估计和对抗学习策略训练两个生成器. GAN-TStega 方法使用 Image COCO[131]和 EMNLP WMT17[89]数据集作为真实的信息.

VAE-Stega 方法使用 Twitter 和 IMDB 数据集训练模型, 并使用与 RNN-Stega 方法相同的设置来生成文本. 与 RNN-Stega 等方法相比, 虽然隐写能力相同, 但 VAE-Stega 方法在引入 VAE 架构后, 隐写模型可以学习正常自然文本的总体统计分布特征, 并在一定程度上进一步约束生成的含密文本的统计分布特征, 因此 VAE-Stega 方法可以显著减少生成的含密文本的总体统计分布与正常语句的总体统计分布之间的差异, 从而生成较高质量的含密文本.

Zhou 等[93]的隐写方法基于三种不同的数据集进行实验以评估文本隐写模型的性能, 这些数据集包括 Twitter[129]、Microsoft Coco(MSCOCO)[123]以及 IMDB[130]. 通过实验结果可知, Zhou 等[93]的隐写方法的隐写能力为每个单词隐藏 1-3 位秘密信息, 并且对自然文本自适应的采样策略, 该方法生成的含密文本质量较高.

PPLM-Stega 方法在实验中基于 Twitter[129]和 IMDB[130]数据集重点与 RNN-Stega 方法对比. 实验结果显示, 由于在生成过程中对含密文本的主题进行了控制, PPLM-Stega 方法能够生成比 RNN-Stega 方法质量更高的含密文本. PPLM-Stega 方法能够根据 ECWP 的大小动态地确定可隐藏的秘密信息容量, 从而可以根据不同的阈值获得不同的隐写容量.

① News 数据集: https://www.kaggle.com/snapcrack/all-the-news/data



(a) BigGAN模型生成的部分图像　　(b) Glow模型生成的部分图像

图 18　图像载体生成质量的稳定性评估实验结果

ALiSa 方法在实验中从 BookCorpus[132]中随机选择了 10,000 份自然文本作为数据集，根据相应的比例将数据集随机划分为训练集、验证集和测试集，并生成了 10,000 份含密文本．根据实验结果，该方法可以在每份生成文本中最高隐藏 4 个单词．假设生成文本的平均长度为 24，则该方法的隐藏能力为每 24 个单词隐藏 4 个单词．

Qin 等[98]的隐写方法根据 25,000 首绝句诗[133]划分数据集，包括 23,000 首绝句诗组成的训练集、1,000 首绝句诗组成的测试集以及 1,000 首绝句诗组成的验证集．每首绝句诗中每一句的前两个字被选为主题词并随机选取 1-4 个主题词作为模型的输入以训练生成模型．实验结果表明，Qin 等[98]的隐写方法的隐写能力主要与四种参数相关(候选诗句的数量、绝句诗模板、韵律的种类以及主题词的个数)．

SongNet 方法共使用 21,053 首宋词作为数据集，其中 19,905 首词作为训练集、661 首词作为测试集以及 487 首词作为验证集，并在模型的预训练阶段使用 27,681 个字符构建词汇．根据实验的最终结果，SongNet 方法将秘密信息以词牌格式中每个句子为单位进行隐藏，其隐写能力主要与 SongNet 方法的多种参数相关(例如词牌信息、关键字、韵律和押韵字符信息等)，可在每个词句中隐藏 18-21 位秘密信息．假设宋词的词句平均长度为 5 个汉字，则该方法的隐藏能力为每 5 个汉字隐藏 18-21 位信息．

表 16　基于神经网络模型的文本生成式隐写方案的性能对比

代表方法	数据集	隐写能力
RNN-Stega[88]	Twitter、IMDB、News	每个单词隐藏 1-5 位
GAN-TStega[89]	Image COCO、EMNLP WMT17	每个单词隐藏 1-5 位
VAE-Stega[91]	Twitter、IMDB	每个单词隐藏 1-5 位
Zhou 等[93]	Twitter、MSCOCO、IMDB	每个单词隐藏 1-3 位
PPLM-Stega[94]	Twitter、IMDB	取决于 ECWP 的大小
ALiSa[96]	BookCorpus	每 24 个单词隐藏 4 个单词
Qin 等[98]	25,000 首绝句诗	与四种参数相关
SongNet[100]	21,053 首宋词	每 5 个字隐藏 18-21 位

7　存在的问题

相比传统的嵌入式隐写方案，生成式隐写方案针对现有基于统计特征的隐写分析方法具有较好的抗检测性能，然而仍然存在以下问题．

(1) 在高隐写荷载条件下，秘密信息的提取难以达到完全无损．为了从含密载体中提取秘密信息，基于 GAN 模型的隐空间映射的生成式隐写方法通常利用全卷积神经网络结构来设计和训练秘

密信息提取器[75, 76]. 然而, 随着隐写的荷载增加, 提取器的网络模型在训练的过程中难以收敛得到全局最优解, 导致提取器提取秘密信息的准确性大大降低. 为了解决该问题, 另一些研究者基于可逆模型如 Glow 等流模型设计了隐写算法[55], 这类模型不仅可以实现秘密信息的隐写, 也支持秘密信息的直接提取. 然而, 以基于隐空间映射的图像生成式隐写方案为例, 虽然流模型理论上是可逆的, 但是将隐空间向量映射到图像空间再重新映射回隐空间后, 所得到的隐向量与原隐向量存在一定的差异. 这是因为流模型是将连续的隐空间和离散的图像数据建立映射关系, 隐向量映射为图像数据时会产生超出固定范围的异常数据元素并有所损失, 这样会显著影响秘密信息的准确提取.

(2) 含密载体生成质量的稳定性较低. 生成式隐写方案通过模型生成含密载体, 与嵌入式隐写方案相比, 含密载体的质量较差, 而且并不能保证每个含密载体都有较高质量, 即生成的含密载体的质量存在稳定性较差的问题. 其原因主要包括以下两点. 首先, 目前大多数的生成式隐写方案是基于 GAN 和流模型等生成网络模型实现的. GAN 模型存在训练过程不稳定的问题如梯度消失和模式崩溃; 流模型是在高维度的连续隐空间和连续数据空间之间建立可逆映射关系, 而多媒体数据通常是离散, 将连续的隐向量映射为离散的多媒体数据将会存在映射误差. 因此, 这两类模型均难以保证含密载体生成质量的稳定性. 其次, 现有的生成网络模型的生成含密载体能力是建立在庞大数据集上充分训练为基础, 但实际用于生成网络模型训练的数据集大小通常受限, 影响生成网络模型的训练效果, 进而影响含密载体的生成质量的稳定性.

(3) 非空间域的抗隐写分析性能有待提高. 现有的生成式隐写方案在多媒体空间域上具有较好的抵抗隐写分析能力, 但在隐空间域和通信上下文域的抗隐写分析性能有待提高. 1) 隐空间域: 现有的生成网络在训练阶段, 通常将服从高斯分布的隐向量映射为多媒体数据. 基于隐空间映射的生成式隐写方案, 首先将秘密信息映射为隐空间的隐向量, 然后将该隐向量输入到训练后的生成网络模型以生成含密载体. 然而, 在秘密信息映射过程中, 由秘密信息映射得到的隐向量难以保持高斯分布, 因此攻击者可以将含密载体逆变换为隐向量, 在隐空间域可以通过检测该隐向量是否符合高斯分布来判断载体是否存在隐写行为; 2) 通信上下文语义域: 当发送方与接收方进行多次隐蔽通信时, 通常需要多次传递含密载体数据. 然而, 现有的隐写方法大多数没有考虑通信数据上下文语义关联性, 那么攻击者可以通过分析通信上下文语义关联从而轻易地检测出通信载体数据是否有可能包含秘密信息. 因此为了保证含密载体图像通信的安全性, 不仅需要考虑多媒体空间域的抗隐写分析性能, 还需要考虑非空间域包括隐空间域、通信上下文域等其他域的抗隐写分析性能. 如何同时保证生成式隐写方案在各个域同时保证具有较好的抗隐写分析性能, 仍然是一个具有挑战性的问题.

8 总结与展望

相比于嵌入式隐写方案, 生成式隐写方案不需要对现有的载体进行修改, 而是以秘密信息为驱动直接生成含密载体, 因此针对现有基于统计特征的隐写分析方法具有较好的抗检测性能, 成为信息隐藏领域具有前景发展方向 . 本文根据隐藏秘密信息载体的类别, 将生成式隐写分类为图像生成式隐写方案、文本生成式隐写方案、音频生成式隐写方案和社交网络行为生成式隐写方案, 分别对其中每一类方案进行了细分, 并对其中的方法进行分析和总结. 然后, 本文通过大量实验, 着重对图像生成式隐写方案和文本生成式隐写方案的性能进行了对比和分析. 此外, 本文总结了现有的生成式隐写存在的问题, 包括秘密信息的提取难以达到完全无损、含密载体生成质量的稳定性较低、非空间域的抗隐写分析性能有待提高.

针对第 7 章所提出的问题, 提出相应的解决方案和展望未来的发展方向.

(1) 针对在高隐写荷载条件下秘密信息的难以精确提取问题, 拟采用基于流模型映射误差校正的生成式隐写方案. 由于流模型在构建训练隐空间和多媒体空间的可逆映射时存在大量的映射误差, 拟在原始流模型的映射中增加一对校正函数. 一方面, 在连续隐向量映射到离散多媒体数据过程中, 学习一个可逆的校正函数对生成的多媒体数据进行校正; 另一方面, 在离散多媒体数据映射到连续隐向量映射反过程中, 使用校正函数的反函数对原始生成的多媒体数据进行恢复. 以上校正函数可以确保超出固定范围的异常多媒体数据校正到固定的范围内, 而对其余数据元素进行微调, 利用其反函数可以准确地恢复原始图像数据和隐向量, 从而

保证在高隐写荷载条件下秘密信息的精确提取.

(2) 针对含密载体生成质量的稳定性不足的问题, 拟采用基于深度自注意力变换(Transformer)网络的生成式隐写方案. 与卷积网络和循环网络等网络类型相比, 基于 Transformer 的模型由于引入了自注意力模块, 可以自动地捕获用于多媒体内容的全局依赖关系[134, 135], Transformer 网络尤其是视觉自注意力变换(Vision Transformer, ViT)网络[136]在各种计算机视觉领域上表现出强大的性能. 由于目前用于含密载体生成的生成器大多利用深度卷积网络来实现含密载体的生成, 而较小的卷积核很难捕获多媒体数据的有效特征[137]. 采用 ViT 网络设计生成式隐写模型的生成器, 可以更加有效捕获多媒体数据的全局相关性, 从而提高多媒体数据生成质量和稳定性. 为了进一步提高多媒体数据生成质量和稳定性, 拟采用自监督学习方法和训练数据自动增强方法, 以高效的方式来解决现有生成式隐写方法的训练数据集不足和训练不充分的问题.

(3) 针对非空间域的抗隐写分析性能不足的问题, 拟采用以下解决方案: 1) 为了提高生成式隐写方案在隐空间域的抗隐写分析性能, 应确保秘密信息映射的隐向量仍是服从高斯分布的. 为此, 并非将秘密信息直接编码为隐向量的元素值, 而是拟将秘密信息编码为隐向量元素的位置排列顺序, 而元素排列位置的变化将不会改变隐向量的高斯分布特性, 因此能够有效保持隐向量的高斯分布特性, 从根本上提高了生成式隐写方法在隐空间域的抗隐写分析能力. 2) 为了解决生成的含密多媒体载体在通信上下文环境中语义合理性的问题, 拟从已经传递的多媒体数据(图像、文本、语音等)序列中提取语义信息, 并用 LSTM 建立语义序列自动生成模型, 将可以得到与真实语义序列统计特性基本一致的语义序列自动生成模型. 然后利用该生成模型, 根据已经传递的多媒体数据序列预测当前待传递的多媒体语义信息, 然后生成相应语义的含密载体多媒体数据用于隐蔽通信, 有效保证了含密多媒体载体在通信上下文环境中语义合理性, 提高了生成式隐写方法在上下文语义域的抗隐写分析能力.

(4) 为了进一步提高现有的生成式隐写方案的隐写容量, 拟设计秘密信息到多媒体数据的高效可逆转换方式. 例如: 由于 Zhou 等[66]的隐写方法在图像的一维轮廓上进行隐写的隐写容量有限, 而图像的二维轮廓相比一维轮廓信息承载量更大. 因此, 在后续的研究中, 可以将秘密信息转换为图像的二维轮廓, 将该二维轮廓输入到生成网络模型中生成相应的含密载体图像, 从而提高隐写容量.

(5) 为了进一步验证生成式隐写方法的安全性, 将研究面向生成式隐写的安全证明模型. 随着各类隐写分析工具的发展, 研究者们针对隐写的安全性通常以大量实验的方式进行评价. 然而, 这些实验数据来说明隐写方法针对某一种或几种隐写分析工具具有抵抗能力, 难以从实验上验证对现有其他隐写分析工具和未知的隐写分析工具具有较好的抵抗能力. 因此, 需要研究如何从理论的角度证明隐写方法的安全性. 可证明安全隐写指通过一定的理论推导的方式证明隐写方法是具有安全性的. 随着生成数据的越来越普及, 其分布特性可以用规范的分布如高维高斯分布拟合并表达, 这样为可证明安全隐写的发展提供了数学基础. 因此, 面向生成式隐写的安全证明模型是信息隐藏领域值得关注的研究方向.

参 考 文 献

[1] Jamil Tariq. Steganography: The Art of Hiding Information in Plain Sight. IEEE potentials, 1999,18(1): 10-12

[2] Wang Shuo-Zhong, Zhang Xin-Peng, Zhang Kai-Wen. Steganography and steganalysis: Information warfare technology in the internet age. Beijing:Tsinghua University Press, 2005
(王朔中, 张新鹏, 张开文. 数字密写和密写分析: 互联网时代的信息战技术. 北京：清华大学出版社有限公司, 2005)

[3] Simmons Gustavus J. The Prisoners' Problem and the Subliminal Channel//Advances in Cryptology. Boston, USA, 1984: 51-67

[4] Van Schyndel Ron G, Tirkel Andrew Z, Osborne Charles F. A Digital Watermark//Proceedings of 1st international conference on image processing. Austin, USA, 1994: 86-90

[5] Bender Walter, Gruhl Daniel, Morimoto Norishige, Lu Anthony. Techniques for Data Hiding. IBM systems journal, 1996,35(3.4): 313-336

[6] Mielikainen Jarno. Lsb Matching Revisited. IEEE signal processing letters, 2006,13(5): 285-287

[7] Fridrich Jessica, Soukal David. Matrix Embedding for Large Payloads. IEEE Transactions on Information Forensics and Security, 2006,1(3): 390-395

[8] Zhang Xinpeng, Wang Shuozhong. Dynamical Running Coding in Digital Steganography. IEEE signal processing letters, 2006,13(3): 165-168

[9] Willems Frans Mj, Van Dijk Marten. Capacity and Codes for Embedding Information in Gray-Scale Signals. IEEE Transactions

on Information Theory, 2005,51(3): 1209-1214

[10] Filler Tomáš, Judas Jan, Fridrich Jessica. Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes. IEEE Transactions on Information Forensics and Security, 2011,6(3): 920-935

[11] Pevný Tomáš, Filler Tomáš, Bas Patrick. Using High-Dimensional Image Models to Perform Highly Undetectable Steganography//International workshop on information hiding. 2010: 161-177

[12] Holub Vojtěch, Fridrich Jessica. Digital Image Steganography Using Universal Distortion//Proceedings of the first ACM workshop on Information hiding and multimedia security. Montpellier, France, 2013: 59-68

[13] Li Bin, Wang Ming, Huang Jiwu, Li Xiaolong. A New Cost Function for Spatial Image Steganography//2014 IEEE International Conference on Image Processing (ICIP). Paris, France, 2014: 4206-4210

[14] Sedighi Vahid, Cogranne Rémi, Fridrich Jessica. Content-Adaptive Steganography by Minimizing Statistical Detectability. IEEE Transactions on Information Forensics and Security, 2015,11(2): 221-234

[15] Zhang Weiming, Zhang Zhuo, Zhang Lili, Li Hanyi, Yu Nenghai. Decomposing Joint Distortion for Adaptive Steganography. IEEE Transactions on Circuits and Systems for Video Technology, 2016,27(10): 2274-2280

[16] Liao Xin, Yu Yingbo, Li Bin, Li Zhongpeng, Qin Zheng. A New Payload Partition Strategy in Color Image Steganography. IEEE Transactions on Circuits and Systems for Video Technology, 2019,30(3): 685-696

[17] Su Wenkang, Ni Jiangqun, Hu Xianglei, Fridrich Jessica. Image Steganography with Symmetric Embedding Using Gaussian Markov Random Field Model. IEEE Transactions on Circuits and Systems for Video Technology, 2020,31(3): 1001-1015

[18] Tang Weixuan, Tan Shunquan, Li Bin, Huang Jiwu. Automatic Steganographic Distortion Learning Using a Generative Adversarial Network. IEEE signal processing letters, 2017,24(10): 1547-1551

[19] Yang Jianhua, Ruan Danyang, Huang Jiwu, Kang Xiangui, Shi Yun-Qing. An Embedding Cost Learning Framework Using Gan. IEEE Transactions on Information Forensics and Security, 2019,15839-851

[20] Low Steven H, Maxemchuk Nicholas F, Brassil Jack T, O'gorman Lawrence. Document Marking and Identification Using Both Line and Word Shifting//Proceedings of INFOCOM'95. Boston, MA, USA, 1995: 853-860

[21] Fu Bing. Research on text information hiding algorithms based on Unicode coding parity. Fujian Computer, 2008, 24(12):66-66
(付兵. 基于字符 Unicode 编码奇偶性的文本信息隐藏算法研究. 福建电脑, 2008,24(12): 66-66)

[22] Yang De-Ming, Guo Sheng. Data hiding method based on word document. Computer Applications and Software, 2015, 32(5):314-318
(杨德明, 郭盛. 基于 Word 文档的数据隐藏方法. 计算机应用与软件, 2015,32(5): 314-318)

[23] Chang Ching-Yun, Clark Stephen. Practical Linguistic Steganography Using Contextual Synonym Substitution and a Novel Vertex Coding Method. Computational linguistics, 2014,40(2): 403-448

[24] Zhang Jianjun, Wang Lucai, Lin Haijun. Coverless Text Information Hiding Method Based on the Rank Map. Journal of Internet Technology, 2017,18(2): 427-434

[25] Yang Xiao, Li Feng, Xiang Ling-Yun. Synonym substitution-based steganographic algorithm with matrix coding. Journal of Chinese Computer Systems, 2015, 36(6):1296-1300
(杨潇, 李峰, 向凌云. 基于矩阵编码的同义词替换隐写算法. 小型微型计算机系统, 2015,36(6): 1296-1300)

[26] Gopalan Kaliappan. Audio Steganography Using Bit Modification//2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698). Baltimore, USA, 2003: I-629

[27] Gruhl Daniel, Lu Anthony, Bender Walter. Echo Hiding//International Workshop on Information Hiding. Cambridge, UK, 1996: 295-315

[28] Erfani Yousof, Siahpoush Shadi. Robust Audio Watermarking Using Improved Ts Echo Hiding. Digital Signal Processing, 2009,19(5): 809-814

[29] Paillard Bruno, Mabilleau Philippe, Morissette Sarto, Soumagne Joël. Perceval: Perceptual Evaluation of the Quality of Audio Signals. Journal of the Audio Engineering Society, 1992,40(1/2): 21-31

[30] Gang Litao, Akansu Ali N, Ramkumar Mahalingam. Mp3 Resistant Oblivious Steganography//2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221). 2001: 1365-1368

[31] Djebbar Fatiha, Hamam Habib, Abed-Meraim Karim, Guerchi Driss. Controlled Distortion for High Capacity Data-in-Speech Spectrum Steganography//2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing.

Darmstadt, Germany, 2010: 212-215

[32] Fridrich Jessica, Goljan Miroslav. On Estimation of Secret Message Length in Lsb Steganography in Spatial Domain//Security, steganography, and watermarking of multimedia contents VI. 2004: 23-34

[33] Chen Chunhua, Shi Yun Q. Jpeg Image Steganalysis Utilizing Both Intrablock and Interblock Correlations//2008 IEEE International Symposium on Circuits and Systems. Seattle, USA, 2008: 3029-3032

[34] Pevny Tomáš, Bas Patrick, Fridrich Jessica. Steganalysis by Subtractive Pixel Adjacency Matrix. IEEE Transactions on Information Forensics and Security, 2010,5(2): 215-224

[35] Qian Yinlong, Dong Jing, Wang Wei, Tan Tieniu. Learning and Transferring Representations for Image Steganalysis Using Convolutional Neural Network//2016 IEEE international conference on image processing (ICIP). Phoenix, Arizona, USA, 2016: 2752-2756

[36] Xu Guanshuo, Wu Han-Zhou, Shi Yun-Qing. Structural Design of Convolutional Neural Networks for Steganalysis. IEEE signal processing letters, 2016,23(5): 708-712

[37] Xu Guanshuo, Wu Han-Zhou, Shi Yun Q. Ensemble of Cnns for Steganalysis: An Empirical Study//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. Vigo, Spain, 2016: 103-107

[38] Zaremba Wojciech, Sutskever Ilya, Vinyals Oriol. Recurrent Neural Network Regularization. arXiv preprint arXiv:1409.2329, 2014

[39] Lipton Zachary C, Kale David C, Elkan Charles, Wetzel Randall. Learning to Diagnose with Lstm Recurrent Neural Networks. arXiv preprint arXiv:1511.03677, 2015

[40] Kingma Diederik P, Welling Max. Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114, 2013

[41] Doersch Carl. Tutorial on Variational Autoencoders. arXiv preprint arXiv:1606.05908, 2016

[42] Goodfellow Ian J, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, Courville Aaron,Bengio Yoshua. Generative Adversarial Nets//Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. Montreal, Canada, 2014: 2672-2680

[43] Creswell Antonia, White Tom, Dumoulin Vincent, Arulkumaran Kai, Sengupta Biswa, Bharath Anil A. Generative Adversarial Networks: An Overview. IEEE signal processing magazine, 2018,35(1): 53-65

[44] Dinh Laurent, Krueger David, Bengio Yoshua. Nice: Non-Linear Independent Components Estimation. arXiv preprint arXiv:1410.8516, 2014

[45] Dinh Laurent, Sohl-Dickstein Jascha, Bengio Samy. Density Estimation Using Real Nvp. arXiv preprint arXiv:1605.08803, 2016

[46] Kingma Diederik P,Dhariwal Prafulla. Glow: Generative Flow with Invertible 1× 1 Convolutions//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 10236-10245

[47] Yang Kuan, Chen Kejiang, Zhang Weiming, Yu Nenghai. Provably Secure Generative Steganography Based on Autoregressive Model//International Workshop on Digital Watermarking. Jeju Island, Korea, 2018: 55-68

[48] Van Oord Aaron, Kalchbrenner Nal, Kavukcuoglu Koray. Pixel Recurrent Neural Networks//International conference on machine learning. New York, USA, 2016: 1747-1756

[49] Hopper Nicholas J, Langford John, Ahn Luis Von. Provably Secure Steganography//Annual International Cryptology Conference. Santa Barbara, USA, 2002: 77-92

[50] Zhang Siyu, Yang Zhongliang, Tu Haoqin, Yang Jinshuai, Huang Yongfeng. Pixel-Stega: Generative Image Steganography Based on Autoregressive Models. arXiv preprint arXiv:2112.10945, 2021

[51] Salimans Tim, Karpathy Andrej, Chen Xi, Kingma Diederik P. Pixelcnn++: Improving the Pixelcnn with Discretized Logistic Mixture Likelihood and Other Modifications. arXiv preprint arXiv:1701.05517, 2017

[52] Liu Jia, Zhou Tanping, Zhang Zhuo, Ke Yan, Lei Yu, Zhang Minqing. Digital Cardan Grille: A Modern Approach for Information Hiding//Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence. Shenzhen, China, 2018: 441-446

[53] Liu Jia, Ke Yan, Lei Yu, Li Jun, Wang Yaojie, Han Yiliang, Zhang Minqing, Yang Xiaoyuan. The Reincarnation of Grille Cipher: A Generative Approach. arXiv preprint arXiv:1804.06514, 2018

[54] Wang Yaojie, Yang Xiaoyuan, Liu Wenchao. Generative Image Steganography Based on Digital Cardan Grille//International Conference on Security and Privacy in New Computing Environments. Lyngby, Denmark, 2020: 343-355

[55] Zhou Zhili, Su Yuecheng, Li Jin, Yu Keping, Wu Qm Jonathan, Fu Zhangjie, Shi Yunqing. Secret-to-Image Reversible Transformation for Generative Steganography. IEEE Transactions on Dependable and Secure Computing, 2022

[56] Li Sheng, Zhang Xinpeng. Toward Construction-Based Data Hiding: From Secrets to Fingerprint Images. IEEE Transactions on

Image Processing, 2018,28(3): 1482-1497

[57] Cao Yi, Zhou Zhili, Wu Qm, Yuan Chengsheng, Sun Xingming. Coverless Information Hiding Based on the Generation of Anime Characters. EURASIP Journal on Image and Video Processing, 2020,2020(1): 1-15

[58] Zhang Zhuo, Fu Guangyuan, Ni Rongrong, Liu Jia, Yang Xiaoyuan. A Generative Method for Steganography by Cover Synthesis with Auxiliary Semantics. Tsinghua Science and Technology, 2020,25(4): 516-527

[59] Otori Hirofumi, Kuriyama Shigeru. Texture Synthesis for Mobile Data Communications. IEEE Computer graphics and applications, 2009,29(6): 74-81

[60] Mäenpää Topi, Pietikäinen Matti. Texture Analysis with Local Binary Patterns. Handbook of Pattern Recognition and Computer Vision. Singapore: World Scientific,2005:197-216

[61] Xu Jiayi, Mao Xiaoyang, Jin Xiaogang, Jaffer Aubrey, Lu Shufang, Li Li, Toyoura Masahiro. Hidden Message in a Deformation-Based Texture. The Visual Computer, 2015,31(12): 1653-1669

[62] Wu Kuo-Chen, Wang Chung-Ming. Steganography Using Reversible Texture Synthesis. IEEE Transactions on Image Processing, 2014,24(1): 130-139

[63] Larkin Kieran G, Fletcher Peter A. A Coherent Framework for Fingerprint Analysis: Are Fingerprints Holograms? Optics express, 2007,15(14): 8667-8677

[64] Reed Irving S, Solomon Gustave. Polynomial Codes over Certain Finite Fields. Journal of the society for industrial and applied mathematics, 1960,8(2): 300-304

[65] Cappelli Raffaele, Erol A, Maio D, Maltoni D. Synthetic Fingerprint-Image Generation//Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. Barcelona, Spain, 2000: 471-474

[66] Zhou Zhi-Li, Wang Mei-Min, Yang Gao-Bo, Zhu Jian-Yu, Sun Xing-Ming. Generative steganography method based on auto-generation of contours. Journal on Communications, 2021, 42(9):144-154
(周志立, 王美民, 杨高波, 朱剑宇, 孙星明. 基于轮廓自动生成的构造式图像隐写方法. 通信学报, 2021,42(9): 144-154)

[67] Hochreiter Sepp, Schmidhuber Jürgen. Long Short-Term Memory. Neural computation, 1997,9(8): 1735-1780

[68] Isola Phillip, Zhu Jun-Yan, Zhou Tinghui, Efros Alexei A. Image-to-Image Translation with Conditional Adversarial Networks//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, Hawaii, USA, 2017: 1125-1134

[69] Wang Zihan, Gao Neng, Wang Xin, Xiang Ji, Liu Guanqun. Stnet: A Style Transformation Network for Deep Image Steganography//International Conference on Neural Information Processing. Sydney, Australia, 2019: 3-14

[70] Simonyan Karen, Zisserman Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556, 2014

[71] Huang Xun, Belongie Serge. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization//Proceedings of the IEEE international conference on computer vision. Venice, Italy, 2017: 1501-1510

[72] Gu Jiuxiang, Wang Zhenhua, Kuen Jason, Ma Lianyang, Shahroudy Amir, Shuai Bing, Liu Ting, Wang Xingxing, Wang Gang, Cai Jianfei. Recent Advances in Convolutional Neural Networks. Pattern recognition, 2018,77：354-377

[73] Zhu Jun-Yan, Park Taesung, Isola Phillip, Efros Alexei A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks//Proceedings of the IEEE international conference on computer vision. Venice, Italy, 2017: 2223-2232

[74] Li Qi, Wang Xingyuan, Wang Xiaoyu, Ma Bin, Wang Chunpeng, Shi Yunqing. An Encrypted Coverless Information Hiding Method Based on Generative Models. Information Sciences, 2021,（553）：19-30

[75] Hu Donghui, Wang Liang, Jiang Wenjie, Zheng Shuli, Li Bin. A Novel Image Steganography Method Via Deep Convolutional Generative Adversarial Networks. IEEE Access, 2018,（6）：38303-38314

[76] Li Jun, Niu Ke, Liao Liwei, Wang Lijie, Liu Jia, Lei Yu, Zhang Minqing. A Generative Steganography Method Based on Wgan-Gp//International Conference on Artificial Intelligence and Security. Hohhot, China, 2020: 386-397

[77] Gulrajani Ishaan, Ahmed Faruk, Arjovsky Martin, Dumoulin Vincent,Courville Aaron. Improved Training of Wasserstein Gans//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA, 2017: 5769-5779

[78] Wei Ping, Li Sheng, Zhang Xinpeng, Luo Ge, Qian Zhenxing, Zhou Qing. Generative Steganography Network. arXiv preprint arXiv:2207.13867, 2022

[79] Liu Xiyao, Ma Ziping, Ma Junxing, Zhang Jian, Schaefer Gerald, Fang Hui. Image Disentanglement Autoencoder for Steganography without Embedding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition. New Orleans, USA, 2022: 2303-2312

[80] Huang Ding, Yan Hong. Interword Distance Changes Represented by Sine Waves for Watermarking Text Images. IEEE Transactions on Circuits and Systems for Video Technology, 2001,11(12): 1237-1245

[81] Shniperov Alexey Nikolaevich, Nikitina Ka. A Text Steganography Method Based on Markov Chains. Automatic Control and Computer Sciences, 2016,50(8): 802-808

[82] Yang Zhongliang, Jin Shuyu, Huang Yongfeng, Zhang Yujin, Li Hui. Automatically Generate Steganographic Text Based on Markov Model and Huffman Coding. arXiv preprint arXiv:1811.04720, 2018

[83] Luo Yubo, Huang Yongfeng, Li Fufang, Chang Chinchen. Text Steganography Based on Ci-Poetry Generation Using Markov Chain Model. KSII Transactions on Internet and Information Systems (TIIS), 2016,10(9): 4568-4584

[84] Desoky Abdelrahman. Notestega: Notes-Based Steganography Methodology. Information Security Journal: A Global Perspective, 2009,18(4): 178-193

[85] Desoky Abdelrahman. Jokestega: Automatic Joke Generation–Based Steganography Methodology. International Journal of Security and Networks, 2012,7(3): 148-160

[86] Huang Yongfeng, Liu Chenghao, Tang Shanyu, Bai Sen. Steganography Integration into a Low-Bit Rate Speech Codec. IEEE Transactions on Information Forensics and Security, 2012,7(6): 1865-1875

[87] Huang Yong Feng, Tang Shanyu, Yuan Jian. Steganography in Inactive Frames of Voip Streams Encoded by Source Codec. IEEE Transactions on Information Forensics and Security, 2011,6(2): 296-306

[88] Yang Zhong-Liang, Guo Xiao-Qing, Chen Zi-Ming, Huang Yong-Feng, Zhang Yu-Jin. Rnn-Stega: Linguistic Steganography Based on Recurrent Neural Networks. IEEE Transactions on Information Forensics and Security, 2018,14(5): 1280-1295

[89] Yang Zhongliang, Wei Nan, Liu Qinghe, Huang Yongfeng, Zhang Yujin. Gan-Tstega: Text Steganography Based on Generative Adversarial Networks//International Workshop on Digital Watermarking. Chengdu, China, 2019: 18-31

[90] Yang Zhongliang, Wang Ke, Li Jian, Huang Yongfeng, Zhang Yu-Jin. Ts-Rnn: Text Steganalysis Based on Recurrent Neural Networks. IEEE signal processing letters, 2019,26(12): 1743-1747

[91] Yang Zhong-Liang, Zhang Si-Yu, Hu Yu-Ting, Hu Zhi-Wen, Huang Yong-Feng. Vae-Stega: Linguistic Steganography Based on Variational Auto-Encoder. IEEE Transactions on Information Forensics and Security, 2020,16：880-895

[92] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2018

[93] Zhou Xuejing, Peng Wanli, Yang Boya, Wen Juan, Xue Yiming, Zhong Ping. Linguistic Steganography Based on Adaptive Probability Distribution. IEEE Transactions on Dependable and Secure Computing, 2022, 19(5): 2982-2997

[94] Cao Yi, Zhou Zhili, Chakraborty Chinmay, Wang Meimin, Wu Qm Jonathan, Sun Xingming, Yu Keping. Generative Steganography Based on Long Readable Text Generation. IEEE Transactions on Computational Social Systems, to appear

[95] Dathathri Sumanth, Madotto Andrea, Lan Janice, Hung Jane, Frank Eric, Molino Piero, Yosinski Jason, Liu Rosanne. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. arXiv preprint arXiv:1912.02164, 2019

[96] Yi Biao, Wu Hanzhou, Feng Guorui, Zhang Xinpeng. Alisa: Acrostic Linguistic Steganography Based on Bert and Gibbs Sampling. IEEE signal processing letters, 2022, 29: 687-691

[97] Gelfand Alan E, Smith Adrian Fm. Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American statistical association, 1990,85(410): 398-409

[98] Qin Chuan, Wang Meng, Si Guang-Wen, Yao Heng. Constructive information hiding with chinese quatrain generation. Chinese Journal of Computers, 2021, 44(4):773-785
(秦川，王萌，司广文，姚恒. 基于绝句生成的构造式信息隐藏算法. 计算机学报, 2021,44(4): 773-785)

[99] Sutskever Ilya, Vinyals Oriol,Le Quoc V. Sequence to Sequence Learning with Neural Networks//Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. Montreal, Canada, 2014: 3104-3112

[100] Qin Chuan, Li Rong-Shou, Qian Zhen-Xing, Zhang Xin-Peng. Large-capacity constructive information hiding based on song ci generation. Chinese Journal of Computers, 2023,46(1):17-30
(秦川，李蓉受，钱振兴，张新鹏. 基于宋词生成的大容量构造式信息隐藏算法. 计算机学报, 2023,46(1):17-30)

[101] Gopalan Kaliappan, Wenndt Stanley. Audio Steganography for Covert Data Transmission by Imperceptible Tone Insertion//Proc. The IASTED International Conference on Communication Systems And Applications (CSA 2004), Banff, Canada. 2004: 262-266

[102] Crawford Heather, Aycock John. Supraliminal Audio Steganography: Audio Files Tricking Audiophiles//International Workshop on Information Hiding. Darmstadt, Germany, 2009:

1-14

[103] Szczypiorski Krzysztof. Stegibiza: New Method for Information Hiding in Club Music//2016 2nd International Conference on Frontiers of Signal Processing (ICFSP). Warsaw, Poland, 2016: 20-24

[104] Yang Zhongliang, Du Xingjian, Tan Yilin, Huang Yongfeng, Zhang Yu-Jin. Aag-Stega: Automatic Audio Generation-Based Steganography. arXiv preprint arXiv:1809.03463, 2018

[105] Yang Hyukryul, Ouyang Hao, Koltun Vladlen, Chen Qifeng. Hiding Video in Audio Via Reversible Generative Models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 1100-1109

[106] Chen Kejiang, Zhou Hang, Zhao Hanqing, Chen Dongdong, Zhang Weiming, Yu Nenghai. Distribution-Preserving Steganography Based on Text-to-Speech Generative Models. IEEE Transactions on Dependable and Secure Computing, 2021, 19(5): 3343-3356

[107] Toderici George, Vincent Damien, Johnston Nick, Jin Hwang Sung, Minnen David, Shor Joel, Covell Michele. Full Resolution Image Compression with Recurrent Neural Networks//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA, 2017: 5306-5314

[108] Ping Wei, Peng Kainan, Chen Jitong. Clarinet: Parallel Wave Generation in End-to-End Text-to-Speech. arXiv preprint arXiv:1807.07281, 2018

[109] Prenger Ryan, Valle Rafael, Catanzaro Bryan. Waveglow: A Flow-Based Generative Network for Speech Synthesis//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK, 2019: 3617-3621

[110] Cachin Christian. An Information-Theoretic Model for Steganography//International Workshop on Information Hiding. Portland, USA, 1998: 306-318

[111] Liu Hanlin, Liu Jingju, Yan Xuehu. Social Network Behavior-Oriented Audio Steganography Scheme//2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC). Harbin, China, 2018: 417-420

[112] Zhang Xinpeng. Behavior Steganography in Social Network[C]//Advances in Intelligent Information Hiding and Multimedia Signal Processing.Springer,2017:21-23.

[113] Hu Yinghong, Wang Zichi, Zhang Xinpeng. Steganography in Social Networks Based on Behavioral Correlation. IETE Technical Review, 2021,38(1): 93-99

[114] Zhou Zhili, Su Yuecheng, Zhang Yulan, Xia Zhihua, Du Shan, Gupta Brij B, Qi Lianyong. Coverless Information Hiding Based on Probability Graph Learning for Secure Communication in Iot Environment. IEEE Internet of Things Journal, 2022, 9(12): 9332-9341

[115] Gibbs Chance, Shashidhar Narasimha. Stegorogue: Steganography in Two-Dimensional Video Game Maps. Advances in Computer Science, 2015,4(3): 141-146

[116] Hernandez-Castro Julio C, Blasco-Lopez Ignacio, Estevez-Tapiador Juan M, Ribagorda-Garnacho Arturo. Steganography in Games: A General Methodology and Its Application to the Game of Go. computers & security, 2006,25(1): 64-71

[117] Ou Zhan-He, Chen Ling-Hwei. A Steganographic Method Based on Tetris Games. Information Sciences, 2014,276: 343-353

[118] Mahato Susmita, Yadav Dilip Kumar, Khan Danish Ali. A Minesweeper Game-Based Steganography Scheme. Journal of information security and applications, 2017,32: 1-14

[119] Lecun Yann, Bottou Léon, Bengio Yoshua, Haffner Patrick. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 1998,86(11): 2278-2324

[120] Krizhevsky Alex, Hinton Geoffrey. Learning Multiple Layers of Features from Tiny Images. University of Toronto, Toronto, Canada, 2009

[121] Learned-Miller Erik, Huang Gary B, Roychowdhury Aruni, Li Haoxiang, Hua Gang. Labeled Faces in the Wild: A Survey//Advances in Face Detection and Facial Image Analysis.Springer,2016:189-248.

[122] Yu Fisher, Seff Ari, Zhang Yinda, Song Shuran, Funkhouser Thomas, Xiao Jianxiong. Lsun: Construction of a Large-Scale Image Dataset Using Deep Learning with Humans in the Loop. arXiv preprint arXiv:1506.03365, 2015

[123] Lin Tsung-Yi, Maire Michael, Belongie Serge, Hays James, Perona Pietro, Ramanan Deva, Dollár Piotr, Zitnick C Lawrence. Microsoft Coco: Common Objects in Context//European conference on computer vision. Zurich, Switzerland, 2014: 740-755

[124] Bossard Lukas, Guillaumin Matthieu, Gool Luc Van. Food-101–Mining Discriminative Components with Random Forests//European conference on computer vision. Zurich, Switzerland, 2014: 446-461

[125] Karras Tero, Laine Samuli, Aila Timo. A Style-Based Generator Architecture for Generative Adversarial Networks//Proceedings of the IEEE/CVF conference on computer vision and pattern

recognition. Long Beach, USA, 2019: 4401-4410

[126] Brock Andrew, Donahue Jeff, Simonyan Karen. Large Scale Gan Training for High Fidelity Natural Image Synthesis. arXiv preprint arXiv:1809.11096, 2018

[127] Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Li Kai, Fei-Fei Li. Imagenet: A Large-Scale Hierarchical Image Database//2009 IEEE conference on computer vision and pattern recognition. Miami, USA, 2009: 248-255

[128] Mittal Anish, Moorthy Anush Krishna, Bovik Alan Conrad. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing, 2012,21(12): 4695-4708

[129] Go Alec, Bhayani Richa, Huang Lei. Twitter Sentiment Classification Using Distant Supervision. CS224N project report, Stanford, 2009,1(12): 2009

[130] Maas Andrew, Daly Raymond E, Pham Peter T, Huang Dan, Ng Andrew Y, Potts Christopher. Learning Word Vectors for Sentiment Analysis//Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. Portland, USA, 2011: 142-150

[131] Chen Xinlei, Fang Hao, Lin Tsung-Yi, Vedantam Ramakrishna, Gupta Saurabh, Dollár Piotr, Zitnick C Lawrence. Microsoft Coco Captions: Data Collection and Evaluation Server. arXiv preprint arXiv:1504.00325, 2015

[132] Zhu Yukun, Kiros Ryan, Zemel Rich, Salakhutdinov Ruslan, Urtasun Raquel, Torralba Antonio, Fidler Sanja. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books//Proceedings of the IEEE international conference on computer vision. Santiago, Chile, 2015: 19-27

[133] Zhang Jiyuan, Feng Yang, Wang Dong, Wang Yang, Abel Andrew, Zhang Shiyue, Zhang Andi. Flexible and Creative Chinese Poetry Generation Using Neural Memory. arXiv preprint arXiv:1705.03773, 2017

[134] Zhang Han, Goodfellow Ian, Metaxas Dimitris, Odena Augustus. Self-Attention Generative Adversarial Networks//International conference on machine learning. Long Beach, USA, 2019: 7354-7363

[135] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz,Polosukhin Illia. Attention Is All You Need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6000-6010

[136] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929, 2020

[137] Yu Cong, Hu Donghui, Zheng Shuli, Jiang Wenjie, Li Meng, Zhao Zhong-Qiu. An Improved Steganography without Embedding Based on Attention Gan. Peer-to-Peer Networking and Applications, 2021,14(3): 1446-1457

ZHOU Zhi-Li, Ph.D., professor. His research interests include information hiding, digital forensics, cybersecurity, multimedia security, blockchain and secret sharing.

DING Chun, M.S. His research interests include information hiding and computer vision.

LI Jin, Ph.D., professor. His research interests include artificial intelligence security, AI gaming, privacy computing and blockchain.

PENG Fei, Ph.D., professor. His research interests include security applications of artificial intelligence, multimedia security and confidentiality, industrial internet security and confidentiality.

ZHANG Xin-Peng, Ph.D., professor. His research interests include multimedia information security, AI security and image processing

Background

This research is a survey of generative steganographic schemes in the field of information security. Steganography is generally used to hide secret information into cover as stego in an invisible form so that covert communication can be achieved by transmitting the stego. The information embedding-based steganographic schemes are difficult to resist the detection of steganalysis tools since the modification of the cover inevitably leads to changes in the statistical properties of the cover. To address the problem, a lot of generative steganographic schemes are proposed and they usually generate a new multimedia cover as a stego driven by secret information, thus the generative steganographic schemes can achieve promising anti-detectability to steganalysis. As a result, the generative steganography has become one of the hottest research topics in the field of information hiding in recent years. The generative steganographic schemes are classified into four categories in this paper, which are described and analyzed in detail; Then, the performances of some generative steganographic schemes are analyzed and compared experimentally; Finally, the problem of existing generative steganographic schemes is concluded, and then the corresponding solutions and future research directions are provided.

This work is supported by the National Key Research and Development Program of China (No. 2022YFB3103100), in part by the National Natural Science Foundation of China (No. 61972205), in part by National Natural Science Foundation of China under Grant for Joint Fund Project (No. U1936218), in part by National Natural Science Foundation of China under Grant for Outstanding Youth Foundation (No. 62122032), and in part by the Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET) fund, which aim to study information hiding and digital forensics.