基于条件可逆网络的生成式图像隐写算法

刘 婷, 任延珍, 王丽娜

武汉大学国家网络安全学院 武汉 中国 430072

摘要 近年来, 随着生成模型的广泛使用, 生成式隐写领域得到了快速发展。生成式隐写是在图像合成过程中隐藏信息的技术。它无需真实图像参与, 只需秘密消息驱动生成模型即可合成载密图像。然而, 现有方法无法控制生成的图像内容, 因此不能保证隐蔽通信行为的安全性。针对上述问题, 本文提出了基于条件可逆网络(Conditional Invertible Neural Network, cINN)的生成式图像隐写术 steg-Cinn。在本文中, 我们将信息隐藏建模为图像着色问题, 并将秘密信息嵌入到灰度图像的颜色信息中。首先, 我们使用映射模块将二进制秘密信息转换为服从标准正态分布的隐变量。而后, 我们以灰度图像作为先验来指导着色过程, 使用条件可逆网络来将隐变量映射为颜色信息。其中 steg-Cinn 生成的彩色图像匹配灰度图像的语义内容, 从而保证了隐蔽通信的行为安全。对比实验结果表明, 本文方法能够控制生成的图像内容并且使得合成颜色真实自然, 在视觉隐蔽性方面表现良好。在统计安全性方面, 本文方法的隐写分析检测正确率为 56.28 %, 说明它能够抵御隐写分析检测。此外, 本文方法在比特消息提取方面可以实现 100%正确提取, 这种情况下的隐藏容量是 2.00 bpp。因此, 与现有方法相比, 本文方法在图像质量、统计安全性、比特提取正确率和隐藏容量方面取得了良好的综合性能表现。迄今为止, 本文方法是在图像隐写术中首次使用 cINN 的工作。考虑到任何信息都可以转换为二进制形式, 我们可以在图像中隐藏任意类型的数据, 因此本文方法在现实世界里也具备实用价值。

关键词 图像隐写; 可逆网络; 内容可控; 行为安全

中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2023.07.02

Generative Image Steganography Via Conditional Invertible Neural Network

LIU Ting, REN Yanzhen, WANG Lina

School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Abstract In recent years, the field of generative steganography has developed rapidly with the widespread use of generative models. Generative steganography is a technology of generating stego images directly from secret messages without real images. However, existing works can't control the generated image content, thus can't guarantee behavioral security during covert communication. To address the above issue, this paper proposes steg-Cinn, a generative image steganography based on Conditional Invertible Neural Network(cINN). In this paper, we formulate the data hiding as an image colorization problem and the secret data is embedded into the color information for a gray-scale host image. First, the binary secret data is transformed into latent variable that follows standard normal distribution using a mapping module. Second, we use a conditional invertible neural network which uses gray-scale image as prior to guide the colorization process, where the latent variable is mapped into color information. The colored images generated by steg-Cinn can match semantic content of gray-scale images, thus ensuring behavioral security in covert communication. The comparative experimental results show that the proposed method is able to control the generated image content and generate realism colors, indicating good performance in terms of visual concealment. For statistical security, the proposed method can resist the detection of steganalysis successfully, where the detection rate calculated by steganalyzer is 56.28 %. In addition, the proposed method can achieve 100% bit extraction accuracy with the hiding capacity of 2.00 bits per pixel(bpp). Therefore, comparing with the existing methods, the proposed method achieves good comprehensive performance in terms of image quality, statistical security, bit extraction accuracy and hiding capacity. As far as we know, the proposed method is the first work to use cINN in image steganography. Since any information can be binarized, we can embed data with arbitrary types into images, thus bringing practical utility in the real-world.

Key words image steganography; invertible neural network; content controllability; behavioral security

通讯作者: 任延珍, 博士, 武汉大学教授, Email: renyz@whu.edu.cn。

本文受到国家自然科学基金项目支持(No. 62172306) 以及湖北省重点研发计划项目(No. 2021BAA034)支持。

收稿日期: 2021-12-18; 修改日期: 2022-02-19; 定稿日期: 2023-04-17

1 引言

隐写术是一种隐蔽通信技术[1], 目的在于隐藏秘密消息的存在性。其中, 携带有秘密消息的图像被称为载密图像(stego image), 不含秘密消息的图像被称为载体图像(cover image)。对于传统的隐写方法而言, 基于校验格码[2](Syndrome-Trellis Codes, STC)的自适应隐写编码方法是最常用的思路。其中, 隐写编码需要的失真代价可以是人工设计的[3-8], 也可以是神经网络设计的[9-12]。传统方法能够在视觉隐蔽性、统计安全性以及隐藏容量方面取得很好的权衡, 但是, 它们依赖于大量的专家知识和手工规则, 隐写编码设计过于复杂。

随着深度学习(Deep learning)的发展, 研究者们开始着手使用深度神经网络技术来设计隐写算法[13], 即深度隐写(Deep steganography)。深度隐写可以分为两类: 嵌入式隐写(Embedding based steganography)和生成式隐写(Generative steganography)。嵌入式隐写的思路是以真实图像为载体图像(cover image), 将秘密消息嵌入载体中, 从而得到载密图像(stego image); 而生成式隐写的思路是在没有真实自然图像参与的情况下, 由秘密消息驱动生成模型(Generative models)即可合成载密图像。

基于深度学习的嵌入式隐写算法主要包括基于自编码器[14](Auto-Encoder, AE)的方法[15-19]和基于可逆网络[20](Invertible Neural Network, INN)的方法[21-22]。基于 Auto-Encoder 的嵌入式隐写[15-19]的通用做法是使用一个编码器网络来嵌入消息, 使用一个解码器网络来提取消息, 使用图像重构损失和消息重构损失来联合训练编码器网络和解码器网络; 而基于 INN 的嵌入式隐写[21-22]只需一个网络模型就可以完成消息嵌入和消息提取任务, 即编码器网络和解码器网络是共享结构、共享参数的。这些方法合成的图像质量良好, 即在视觉隐蔽性方面具有出色的表现。在嵌入式隐写当中, 人们以真实图像作为载体, 希望载密图像和真实自然图像的分布足够接近。然而, 在传统数据环境中, 想要对真实自然图像的分布做准确建模是很困难的。如今, 随着生成模型(Generative models)的发展, AI 合成应用的普及给隐写研究者们提供了新的数据生态和伪装环境: 我们不必局限于使用真实自然图像做载体, 还可以使用 AI 合成的图像做载体。因此, 生成式隐写enerative steganography)具有良好的应用价值。

根据生成模型类别的不同, 现有的生成式图像隐写可以分为: 基于生成对抗网络[23](Generative Adversarial Network, GAN)的方法[24-27]、基于自回归模型[28](Autoregressive models)的方法[29]以及基于可逆网络[20](Invertible Neural Network, INN)的方法[30]。这些方法在统计安全性方面表现良好, 但是, 它们的共同问题是: 生成的图像语义内容不可控, 因此无法保证隐蔽通信的行为安全。为了解决这个问题, 本文提出了基于条件可逆网络(Conditional Invertible Neural Network, cINN)的生成式图像隐写框架 steg-Cinn。在灰度图像的条件指导下, steg-Cinn 可以将秘密消息隐藏在颜色信息里。并且, 这些颜色信息和灰度图像的语义内容是匹配的, 以确保生成图像的视觉隐蔽性。

本文组织结构如下。第 2 节介绍与本文工作相关的研究现状, 包括基于深度学习的隐写技术、可逆网络技术和条件可逆网络技术; 第 3 节描述本文方法 steg-Cinn 的隐写技术框架, 包括信息嵌入过程、提取过程, 映射过程和损失函数设计; 第 4 节进行实验分析, 包括实验准备、实验设计和实验结果评估; 第 5 节总结全文并展望未来方向。

2 相关工作

2.1 基于深度学习的隐写算法

根据隐写算法的实现技术以及算法特性, 本文对目前主流的深度学习隐写算法进行了归纳分析, 将其分为嵌入式隐写和生成式隐写两大类, 如表 1 所示。

基于深度学习的嵌入式隐写算法代表作有基于自编码器的方法[15-17]和基于可逆网络(Invertible Neural Network, INN)的方法 ISN[21]。这些隐写算法在图像视觉隐蔽性方面表现良好。其中, 基于自编码器的方法[15-17]在设计网络训练损失的时候没有考虑到统计安全性这个因素, 因此不能抵抗通用隐写分析器的检测。而基于可逆网络的方法 ISN[21]则具有良好的抗检测性能, 可抵抗伪造检测工具[31]以及针对 LSB 替换(Least Significant Bit Substitution)隐写的专用隐写分析器 stegExpose[32]的检测。对于能否抵抗通用隐写分析器[33-36]的检测, 文中无相关实验分析。

近年来, 随着生成模型(Generative models)的广泛落地, 生成式隐写领域得到了快速发展。与嵌入式隐写不同, 生成式隐写不需要真实自然图像的参与, 只需秘密消息来驱动生成模型即可合成载密图像, 如图 1 所示。其中, 对于由噪声驱动合成的图像被定义为载体图像。生成式隐写和生成模型密切相关, 早期的工作包括基于 GAN 的方法[24-27]和基于自回归模型的方法[29]。这些方法在统计安全性方面表现良好,

表 1 深度学习隐写算法领域的相关工作

Table 1 Related work about steganography based on deep learning

隐写分类		应用	媒介	优势比较		
				内容可控	感知隐蔽性	统计安全性
嵌入式隐写	基于自编码器的方法	文献[18], HiDDeN[19], ddh[15-16], udh[17]	图像	√	√	×
	基于可逆网络(INN)的方法	ISN[21]	图像	√	√	未知
生成式隐写	基于生成对抗网络(GAN)的方法	SGAN[24], SSGAN[25], 文献[26], SSteGAN[27]	图像	×	√	√
	基于自回归模型的方法	文献[29]	图像	×	√	√
	基于可逆网络(INN)的方法	steg-glow[30]	图像	×	√	√
	基于条件可逆网络(cINN)的方法	音频藏视频[37]	音频	√	√	未知
		steg-TTS[38]	音频	√	√	√

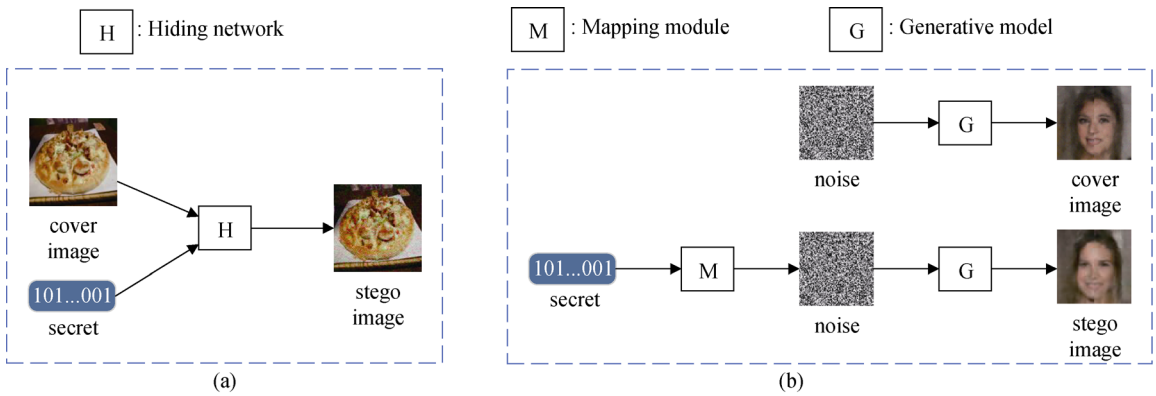

图 1 嵌入式隐写和生成式隐写的区别。(a)嵌入式隐写(b)生成式隐写。

Figure 1 Difference between embedding based steganography and generative steganography.(a) Embedding based steganography.(b) Generative steganography.

但是合成的图像质量较为模糊, 而且图像语义内容不可控。随着生成模型的发展, 可逆网络逐渐引起了研究者们的关注: 其结构天然可逆, 与隐写任务的目标一致, 因此研究者们试图将可逆网络引入到隐写任务当中来。对于基于可逆网络的生成式隐写, 目前现有方法已经在图像领域取得了初步探索[30]。另外, 基于可逆网络的生成式隐写在音频领域也有所应用, 比如音频藏视频[37]、基于文本转语音(Text-to-Speech, TTS)系统的分布保持隐写[38](下文记作steg-TTS)。二者的工作基础选择的都是基于条件可逆网络的声码器 waveGlow[39]: 以 Mel 谱图作为条件指导, 因此它们合成语音的语义内容是可控的。其中, 音频藏视频[37]工作在抗隐写分析性方面没有相关实验分析, 统计安全性未知; 而方法 steg-TTS[38]则能够抵抗通用隐写分析器的检测, 并且从理论角度证明了该方法的隐写分布保持性, 在统计安全性方面具有优良的表现。

由表 1 可知, 现有的生成式图像隐写方法在合成语义内容方面不可控。针对这个问题, 本文提出基于条件可逆网络(Conditional Invertible Neural Network, cINN)[40]的生成式图像隐写算法思路。本文方法一方面能够利用 INN 网络结构的可逆性, 实现对秘密消息数据的可逆提取, 提高隐写算法的提取准确率; 另一方面, 本文方法能够利用 cINN 的条件指导性, 使得所生成的图像语义信息可以调控, 确保载密图像数据内容的可控性, 从而保证了隐写算法的行为安全。

2.2 可逆网络与条件可逆网络

可逆网络(Invertible Neural Network, INN)[20, 41-43]是一种能够对生成数据做显式建模的可逆生成模型。它的基本原理是借助一个隐变量(latent variable)z的简单分布(比如标准正态分布)来拟合一个现实世界数据 x(比如图像数据)的复杂分布 $q(x)$。在结构方面, INN 由一系列的可逆变换组成。受益于数据加密标准[44](Data Encryption Standard, DES)的结构基础 Feistel 网络[45]的启发, 可逆变换一般被设计为仿射

耦合层(affine coupling layer)。在训练方面, 根据极大似然估计的思想, INN 采用负对数似然损失, 即训练目标是使得 z 足够接近标准正态分布: 由于 INN 结构天然可逆, 只要 z 越接近标准正态分布, 那么 $q(x)$就越接近现实数据分布。在应用方面, 最为广泛使用的 INN 是流模型(flow based models), 比如 Nice[41], RealNVP[42], Glow[43]。

条件可逆网络(Conditional Invertible Neural Network, cINN)[40]是增加了条件指导的 INN 模型。它的基本原理是借助一个隐变量 z 的简单分布(比如标准正态分布)来拟合一个现实世界数据 x(比如图像数据)的条件分布 $q(x;c)$, 其中, c 是条件(condition), 在本文中指的是指导 cINN 合成彩色图像的灰度图像。cINN 的推理(inference)过程分为两个步骤。第一, 从标准正态分布(均值为 0, 方差为 1)中采样, 得到隐变量 z。第二, 计算 $x = f^{-1}(z;c)$。其中, x 是我们想要合成的数据, f 是 cINN 前馈方向的映射, f^{-1} 是 cINN 生成方向的映射。z 是驱动 cINN 模型做生成任务的隐变量(latent variable)。值得一提的是, f^{-1} 或 f 是由一系列的可逆变换 f_i 组成的, 即 $z=f(x;c)$可写成 $z = f_k \cdot f_{k-1} \cdots\cdots f_2 \cdot f_1(x;c)$。

在结构方面, 仿射耦合层(affine coupling layer)是条件可逆网络的基本组成部分, 如图 2 所示。它的作用是实现可逆变换 f_i。如果 f 由单个的 f_i 组成, 则 $z=f(x;c)$经历的变换如算法 1 所示, $x = f^{-1}(z;c)$经历的变换如算法 2 所示。其中, s 和 t 可以是任意复杂的深度神经网络, *cat* 表示将两部分数据连接(concatenate)成一个张量数据(tensor), ⊙表示逐元素的按位乘法(element-wise multiplication)。

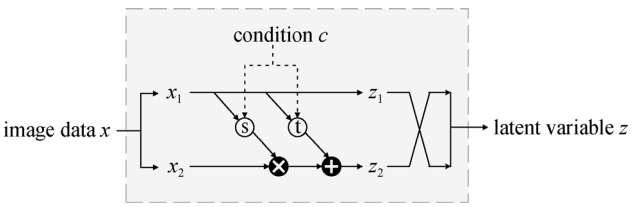

图 2　仿射耦合层的结构

Figure 2　Framework of affine coupling layer

算法 1. 可逆变换 f_i 的正向运算

输入: 图像数据 x, 条件 c

输出: 隐变量 z

1. 将 x 切分为 x_1 和 x_2 两部分
2. 使用恒等映射: $z_1 = x_1$
3. 把 x_1 和 c 连接起来, 即 $x_1 = cat(x_1, c)$
4. 计算 $z_2 = x_2 \odot s(x_1)+t(x_1)$
5. 把 z_1 和 z_2 连接起来, 即 $z = cat(z_1, z_2)$
6. 沿着通道维度对 z 进行置乱

算法 2. 可逆变换 f_i 的反向运算

输入: 隐变量 z, 条件 c

输出: 图像数据 x

1. 将 z 切分为 z_1 和 z_2 两部分
2. 使用恒等映射: $x_1 = z_1$
3. 把 x_1 和 c 连接起来, 即 $x_1 = cat(x_1, c)$
4. 计算 $x_2 = (z_2 - t(x_1)) \div s(x_1)$
5. 把 x_1 和 x_2 连接起来, 即 $x = cat(x_1, x_2)$
6. 沿着通道维度对 x 进行置乱

在训练方面, 条件可逆网络的训练目标是使得隐变量 z 接近标准正态分布。由于 cINN 在结构上天然可逆, 因此只要 z 接近标准正态分布, 那么 $x=f^{1}(z;c)$就会接近真实世界里数据 x 的分布 $q(x;c)$。根据极大似然估计[46](Maximum Likelihood Estimation, MLE)的思想, 训练损失可以记作负对数似然(Negative log-likelihood, NLL), 如公式(1)所示。

$$\begin{aligned} loss_{NLL} &= -\log\ q(x;c) \\ &= -\log\ \pi(z)\left|\det\frac{\partial z}{\partial x}\right| \end{aligned} \tag{1}$$

其中, $\left|\det\frac{\partial z}{\partial x}\right|$ 是雅可比矩阵的行列式, π 是标准正态分布, z 是模型 cINN 的输出, 如公式(2)所示。

$$\pi(z) = N(z;0,I), z = f(x;c) \tag{2}$$

在生成式隐写应用方面, 条件可逆网络 cINN 已经成功应用于音频藏视频[37]和基于文本转语音系统的隐写 steg-TTS[38]。二者的工作基础选择的都是基于条件可逆网络的声码器 waveGlow[39]: 以 Mel 谱图作为条件指导, 因此它们合成语音的语义内容是可控的。此外, INN 在生成式隐写领域也有应用, 比如基于可逆网络 Glow[43]的生成式图像隐写算法[30](下文记作 steg-glow)。方法 steg-glow[30]在统计安全性方面表现良好, 但是不能控制合成的语义内容, 因此不能保证隐蔽通信行为的安全性。这也是现有的生成式图像隐写方法面临的共同问题。为了解决这个问题, 本文提出基于条件可逆网络的生成式图像隐写算法思路。本文方法一方面能够利用 INN 网络结构的可逆性, 实现对秘密消息数据的可逆提取, 提高隐写算法的提取准确率; 另一方面, 本文方法能够利用 cINN 的条件指导性, 使得所生成的图像语义信息可以调控, 确保载密图像数据内容的可控性, 从而保证了隐写算法的行为安全。

3 基于条件可逆网络的生成式图像隐写

本文提出的基于条件可逆网络的生成式图像隐写框架 steg-Cinn 如图 3 所示, 主要包括秘密消息的嵌入过程和提取过程。对于嵌入过程, 发送方以灰度图像为条件指导, 在涂色过程中隐藏秘密消息, 合成彩色载密图像; 对于提取过程, 接收方先分解载密图像得到灰度图像和颜色信息, 继而从颜色信息当中提取秘密消息。详情如下所述。

3.1 嵌入过程

秘密消息的嵌入过程(Embedding process)如图 3(a)所示, 包括消息映射、基于 cINN 的图像合成这两个环节。嵌入过程的输入是秘密消息 m(二进制比特流)和作为条件(condition)的灰度图像 L, 输出是彩色载密图像(stego image)。其中, 消息映射环节的作用是, 使用映射模块 M 把秘密消息 m 转换为服从标准正态分布的隐变量(latent variable)z。基于 cINN 的图像合成环节的作用是, 在灰度图像 L 的指导下, 条件可逆网络 cINN 将隐变量 z 转换成颜色信息 ab。灰度图像 L 和颜色信息 ab 连接(concatenate)起来即为 Lab 色彩空间上的图像。经过色彩空间转换的计算, RGB 色彩空间上的载密图像便顺利生成。

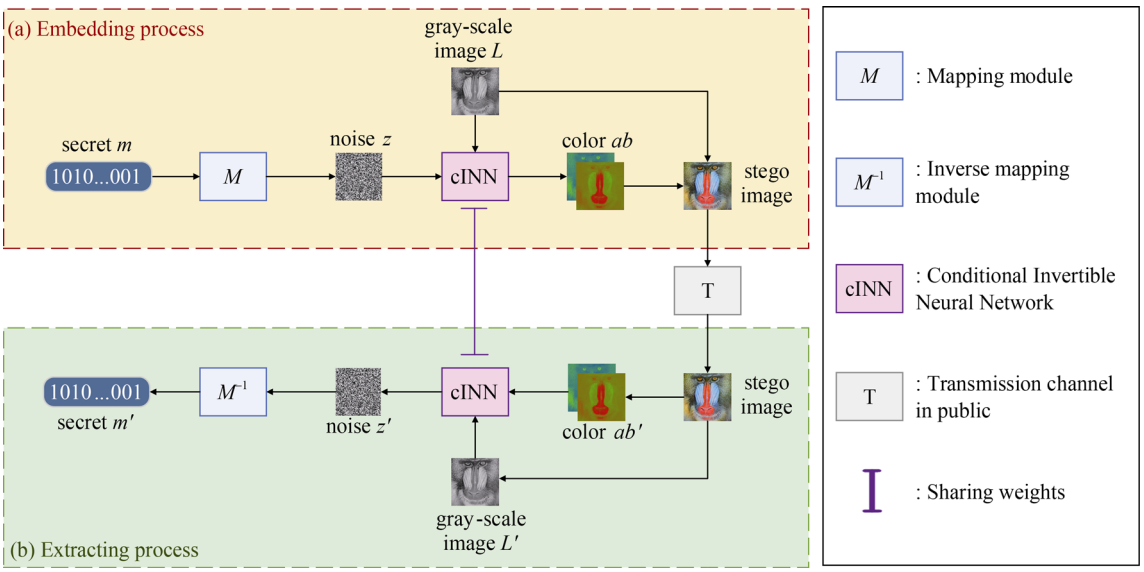

图 3 本文提出的生成式图像隐写框架 steg-Cinn。(a)嵌入过程; (b)提取过程。其中(a)和(b)都用到了映射模块(如算法 3-4 所示)和条件可逆网络 cINN(如图 5 所示)。

Figure 3 Proposed generative image steganography framework steg-Cinn.(a)Embedding process; (b)Extracting process. Note that (a) and (b) both use mapping module (shown in algorithm 3-4) and conditional invertible neural network(shown in Fig.5).

3.2 提取过程

秘密消息的提取过程(Extracting process)如图 3(b)所示, 包括基于 cINN 的前馈计算、消息逆映射这两个环节。提取过程的输入是彩色载密图像, 输出是提取出来的秘密消息 m'(二进制比特流)。其中, 基于 cINN 的前馈计算环节的作用是, 在接收方将 RGB 载密图像进行色彩空间转换并将其分解为灰度图像 L'和颜色信息 ab'之后, 条件可逆网络 cINN 以灰度图像 L'为指导, 将颜色信息 ab'映射为服从标准正态分布的隐变量 z'。消息逆映射环节的作用是, 使用映射模块 M 把隐变量 z'转换为二进制比特流 m', m'即为我们想要提取出来的秘密消息。

3.3 消息映射模块

由图 3 可知, 嵌入过程和提取过程都用到了消息映射模块 M。消息映射模块 M 的作用是完成隐变量和秘密消息(二进制比特)之间的映射。映射的主要思想是用隐变量 z 的符号来表示秘密消息 m(比特 1 或比特 0), 伪代码见算法 3-4。其中 m 是服从均匀分布的二进制比特流, 而 z 是服从标准正态分布的隐变量。为了容错起见, 在此定义了一个区间参数 α。参数 α 的值在 0 到 0.5 之间(本文中, α=0.1), 它的作用是调节隐变量 z 的分布, 从而使得秘密消息的提取具有容错功能。在采样 z 的时候, 如果 z 的值落在$-\alpha$ 到 α 之间, 则拒绝接受该值, 重新采样。因此, 在嵌入过程(Embedding process)中, 隐变量 z 的分布存在一个 $2*\alpha$ 宽的间隔, 如图 4(a)所示。考虑到图像存储和传输过程可能会存在浮点精度误差, 因此, 在提取过程(Extracting process)中, 提取出来的隐变量 z'可

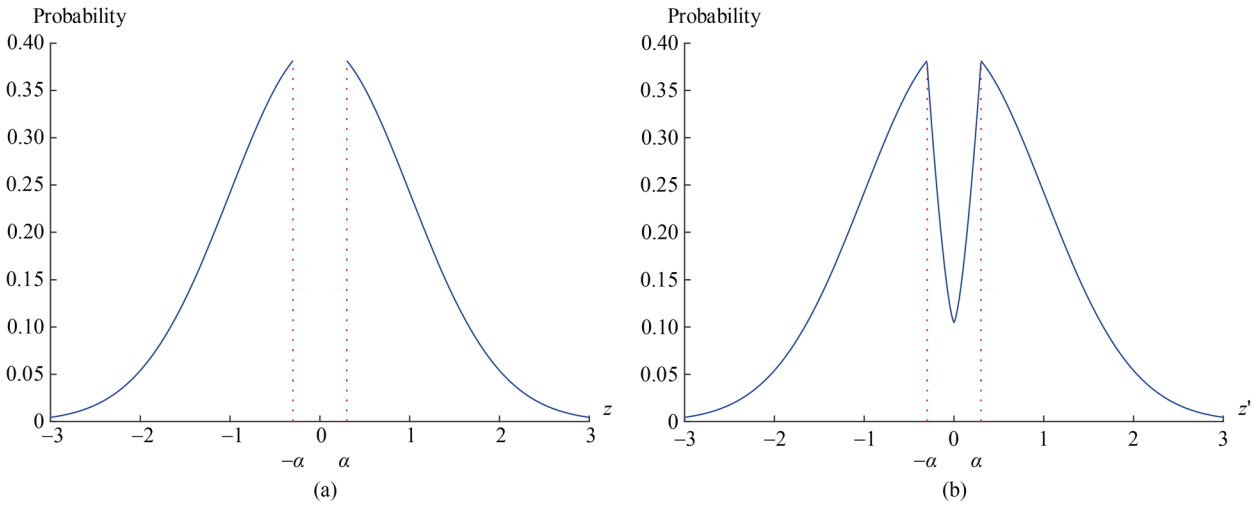

图 4　噪声隐变量的统计分布情况。(a)嵌入过程中噪声 z 的分布; (b)提取过程中噪声 z' 的分布。

Figure 4　Distribution of noise latent variable. (a) Distribution of noise z in embedding process; (b) Distribution of extracted noise z' in extracting process.

能如图 4(b)所示。由此可得, 只要 z 和 z'的符号位保持一致, 就可以准确提取出秘密消息 m。

算法 3. 嵌入过程中的映射模块 M

输入: 消息 msg={0, 1, ..., 0, 1}, 间隔参数 α=0.1

输出: 经过秘密消息映射之后的隐变量 z

1. FOR ALL m IN msg DO
2. IF m==0 THEN
3. 从小于 $-\alpha$ 的标准正态分布里采样 z:
Sample z from N(0, I) until $z < -\alpha$
4. END IF
5. IF m==1 THEN
6. 从大于 α 的标准正态分布里采样 z:
Sample z from N(0, I) until $z > \alpha$
7. END IF
8. END FOR

算法 4. 提取过程中的映射模块 M^{-1}

输入: 提取出来的隐变量 z'

输出: 从隐变量中提取出的比特流 msg'

1. 初始化 msg'
2. FOR ALL $noise'$ IN z' DO
3. IF $noise' < 0$ THEN
4. 提取比特 0 到 msg'
5. ELSE
6. 提取比特 1 到 msg'
7. END IF
8. END FOR

3.4　条件可逆网络与损失函数设计

由图 3 可知, 嵌入过程和提取过程都用到了条件可逆网络 cINN。cINN 的框架如图 5 所示, 它的功能是, 在灰度图像 L 的条件指导下, 完成颜色信息 ab 和隐变量 z 之间的映射。在嵌入过程(Embedding process)中, cINN 将隐变量 z 映射为颜色信息 ab, 映射可以表示为 $ab=x=f^{-1}(z;L)$; 在提取过程(Extracting process)中, 映射可以表示为 $z=f(x;L)$。其中, 对于灰度图像 L, 它是来自于 Lab 颜色空间模型[47]当中的亮度分量(luminance), 取值区间是[0, 100], 表示从纯黑到纯白。由图 5 可知, 灰度图像 L 要经过一个预训练的 VGG 网络[48]的处理, 才能作为条件输入。这个

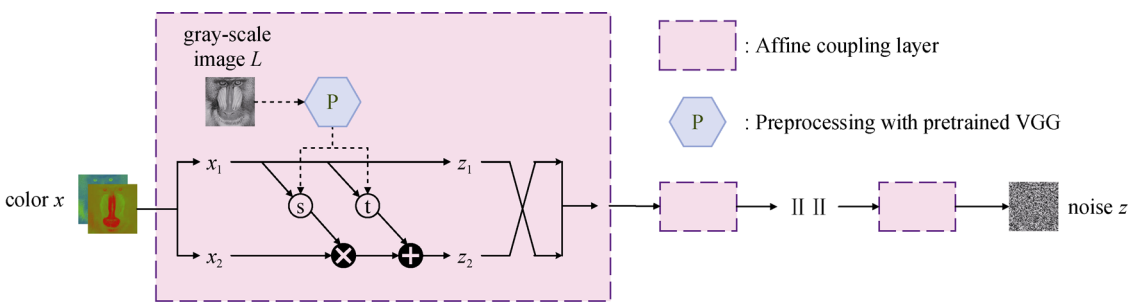

图 5　条件可逆网络(cINN)的框架结构。

Figure 5　Framework of Conditional Invertible Neural Network(cINN) .

预处理操作可以捕获语义信息, 从而使得生成的颜色与图像语义内容相匹配。对于颜色信息 *ab*, 它是来自于 Lab 颜色空间模型[47]当中的色度分量(chroma), 取值区间是[127, –128], 其中 *a* 分量表示从红色到绿色, *b* 分量表示从黄色到蓝色; 对于嵌入过程和提取过程所用到的 cINN, 它们是共享结构、共享权重参数(sharing weights)的, 唯一区别是数据流的方向不同。

本文方法 steg-Cinn 的训练目标是, 通过约束隐变量的统计分布, 从而使得合成图像在颜色方面真实和自然。因此, 根据cINN的损失函数设计思想, 本文方法 steg-Cinn 的损失函数也可以表示为负对数似然(Negative log-likelihood, NLL), 如公式(3)所示。其中, z 是服从标准正态分布 π 的隐变量, L 是作为条件指导的灰度图像, x 是颜色信息 *ab*, f 是 cINN 在前馈(forward)方向的运算, 如公式(4)所示。

$$loss_{NLL} = -\log\ q(x;L) = -\log\ \pi(z)\left|\det\frac{\partial z}{\partial x}\right| \tag{3}$$

$$\pi(z) = N(z;0,I), z = f(x;L), x = ab \tag{4}$$

如图 6 所示, 本文方法 steg-Cinn 在训练的时候, 数据流方向是从图像数据到隐变量。它使用负对数似然损失来约束隐变量 z 的统计分布, 使得 z 越接近标准正态分布越好。只要 z 越接近标准正态分布, 那么做生成(inference)的时候合成的数据 x 就越接近真实世界数据的分布, 合成图像的颜色就越真实和自然。

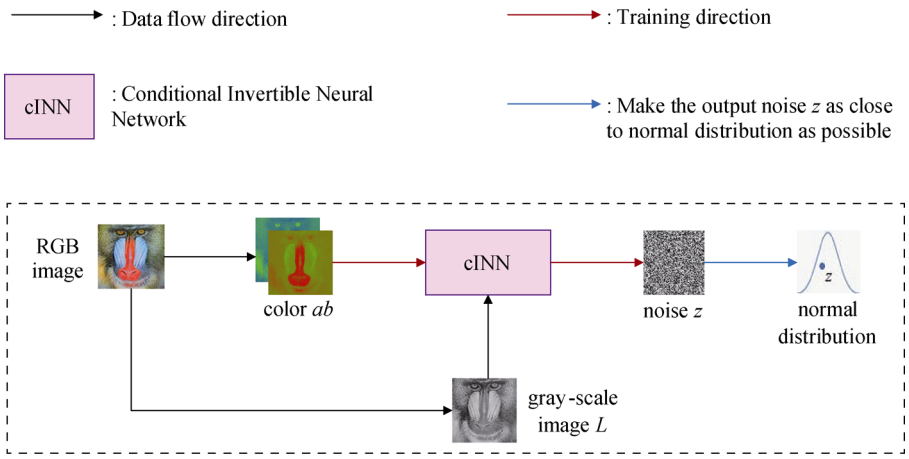

图 6 本文提出的生成式图像隐写方法 steg-Cinn 的训练过程。

Figure 6 The training process of the proposed generative steganography framework steg-Cinn.

4 实验分析

为了评估本文方法 steg-Cinn 的性能表现, 本文分别对视觉质量(visual quality)、统计安全性(statistical security)、隐藏容量(hiding capacity)和提取准确率(extraction accuracy)四个方面的性能进行综合和对比性实验分析。实验情况如下。

4.1 实验准备

4.1.1 实验数据

实验中使用的图像数据集来源于 Coco[49]数据集和 BossBase[50]数据集。Coco[49]数据集包含来自人类日常生活的 91 种对象类型的照片, 包括人物、动物、植物、建筑物、自然景物, 等等。BossBase[50]数据集包含来自 7 个不同相机的 10, 000 张未压缩图像, 主要包括自然景观、建筑物、室内外静物。为了体现泛化性能并使实验结果令人信服, 训练数据选自于 Coco[49]数据集, 测试数据选自 BossBase[50]数据集。其中, Coco[49]数据集的分类预处理操作视情况而定。以对比方法 steg-glow[30]为例, 如果使用未经分类的数据来训练 steg-glow[30]模型, 训练好的模型会输出一些极其抽象的、人类无法理解的无意义图像。因此, steg-glow[30]的训练数据采用的是分类之后的 Coco[49]数据集, 比如分类依据是人物。为了保证训练集里数据分布的平衡性, 分类之后的人物图像的人脸部分要经过识别[51]和裁剪(crop)。而对于本文方法(steg-Cinn)和其它两个对比方法(ddh[16]和 udh[17]), 使用 Coco[49]数据集的时候则不需要分类预处理操作。此外, 所有被用于训练或者测试的图像均被转换为 PNG 格式, 并调整为 128×128 的分辨率。

4.1.2 对比方法

本文选择了三种基于深度神经网络(Deep Neural Network, DNN)的隐写方法作为对比方法(baseline), 它们分别是 steg-glow[30]、ddh[16]和 udh[17]。其中,

(1)steg-glow[30]属于生成式隐写, 它的结构源自于最常用的可逆网络 Glow[43], 主要思想是在图像合成的过程当中隐藏秘密消息比特, 这个过程不需要真实自然图像的参与。对于 steg-glow[30]来说, 载体图像是由噪声驱动 Glow[43]直接合成的图像, 而载密图像则是经过秘密消息映射之后的噪声驱动Glow[43]合成的图像。由于 Glow[43]结构天然可逆, 因此 steg-glow[30]自身既可以是嵌入器, 也可以是提取器, 从而可以无损恢复秘密消息。(2)而 ddh[16]和 udh[17]属于嵌入式隐写, 它们以自编码器(Auto-Encoder)为结构基础, 以真实自然图像为载体图像, 以 encoder 输出的图像为载密图像。方法 ddh[16]和 udh[17]面向的原始任务是在图像中隐藏图像, 即秘密消息本身也是图像, 隐藏容量高达 24 bpp。但是, 它们放松了对秘密消息准确提取的约束。严格来说, 不能准确提取是一种有悖于隐写术基本原则的做法, 这是不合理的。严谨的做法理应像 steg-glow[30]和本文方法 steg-Cinn 这样, 把比特流作为秘密消息。因此, 为了实验对比的公平性, 在此将 ddh[16]和 udh[17]都改造为图像藏比特流消息。除了这一点改动, 其它实验设置继续严格遵守原文。

4.1.3 术语定义

本文实验中使用的术语主要有载体(cover)图像和载密(stego)图像。其中, 由于 ddh[16]和 udh[17]属于嵌入式隐写, 所以这两种方法对应的载体是真实自然图像, 对应载密图像是模型 ddh[16]或 udh[17]输出的图像; 而 steg-Cinn 和 steg-glow[30]都属于生成式隐写, 所以这两种方法对应的载体是噪声驱动 steg-Cinn 或 steg-glow[30]合成的图像, 对应载密图像是经过秘密消息映射之后的噪声驱动 steg-Cinn 或 steg-glow[30]合成的图像。

4.1.4 评价指标

实验使用了四个指标(metric)对本文方法 steg-Cinn 和对比方法(baseline)进行评估: 视觉质量(visual quality)、统计安全性(statistical security)、隐藏容量(hiding capacity)和提取准确率(extraction accuracy)。每个指标的描述如下。

1) 视觉质量代表着一个隐写算法的感知隐蔽性。本文中, 为了实验对比的公平性, 视觉质量由无参考图像质量评价 (No-reference image quality assessment, NR-IQA)工具[52-53]来衡量, 而不是基于相似性度量的通用工具(如 PNSR、SSIM[54])来衡量。这是因为, 作为对比方法(baseline)之一, steg-glow[30]是一种在数据生成过程中隐藏秘密消息的方法, 无需真实图像的参与。即, steg-glow[30]没有任何基准(Ground truth, GT)图像作为参考图像, 所以它不能使用基于相似性度量的通用工具来评价合成图像的质量。

因此, 在本文中, 视觉质量由两个无参考图像质量评价 (No-reference image quality assessment, NR-IQA)工具来衡量: Brisque[52]和 hyperIQA[53]。Brisque[52]是一种经典的 NR-IQA 工具, 它的原理是, 根据自然场景统计规律, 来评估图像自然度方面可能出现的失真。而另一种 NR-IQA 工具 hyperIQA[53]则很新, 它使用了一种超网络(hyper network), 以 ResNet[55]为骨干(backbone)提取语义特征, 从而对人类视觉系统(Human visual system, HVS)的特征进行建模。Brisque[52]和 hyperIQA[53]的分数通常落在[0, 100]范围内。对于 Brisque[52], 分数越低表示图像质量越好; 对于 hyperIQA[53], 分数越高表示图像质量越好。

2) 统计安全性是指一个隐写算法使得载体的统计分布特征得到保持的能力。本文实验采用隐写分析网络 Ke-Net[56], 通过算法的抗隐写分析能力来评估统计安全性。Ke-Net[56]是一个基于孪生网络(siamese network)结构的隐写分析模型, 可以应用于任意尺寸的图像并且无需重新训练参数。它的灵活性和优秀的性能使其得到了广泛使用, 因此, 本文选择 Ke-Net[56]作为隐写分析器, 来评估本文方法和对比方法的统计安全性。

评估统计安全性的指标是隐写分析检测正确率, 记作 P_{detect}。隐写分析检测正确率 P_{detect} 采用真正例率(True Positive Rate, TPR)和真反例率(True Negative Rate, TNR)的均值来表示, 计算方式见公式(5-7)。其中, 真正例率表示在所有载密样本中, 有多少样本被预测正确了; 真反例率表示在所有载体样本中, 有多少样本被预测正确了。因此, 隐写分析检测正确率 P_{detect} 越低, 说明隐写算法越安全, 抗检测性越强。

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$TNR = \frac{TN}{FP + TN} \tag{6}$$

$$P_{detect} = \frac{TPR + TNR}{2} \tag{7}$$

3) 隐藏容量是指在秘密消息可以被准确提取的前提下, 可以隐藏到载体图像当中的秘密消息的数量。衡量隐藏容量的指标是每一个三通道像素能够隐藏的比特个数, 因此, 其单位是比特每像素(bit per pixel, bpp)。

对于本文方法 steg-Cinn 和对比方法的隐藏容量, 若无特殊说明, steg-Cinn 的隐藏容量是 2.00 bpp, 对

比方法 steg-glow[30]的隐藏容量是 3.00 bpp, ddh[16]和 udh[17]的隐藏容量是 1.00 bpp。具体地, 由于 steg-Cinn 和 steg-glow[30]的工作基础都是可逆网络, 而可逆网络的输入输出总维数是保持一致的, 因此, steg-Cinn 可以隐藏的秘密消息比特个数和颜色信息 *ab* 的总维数($n*n*c_1$)一致; steg-glow[30]可以隐藏的秘密消息比特个数和 RGB 像素总维数($n*n*c_2$)一致。其中, n 是图像的分辨率, c_1 是颜色 *ab* 的通道数, c_2 是 RGB 像素的通道数。本文实验中 n=128, c_1=2, c_2=3。因此, steg-Cinn 的隐藏容量是 2.00 bpp, 对比方法 steg-glow[30]的隐藏容量是 3.00 bpp。而对于 ddh[16]和 udh[17]来说, 一张彩色图像中的每个 $p*p*3$ 区块(patch)可以隐藏一个比特。其中, p 是 2 的整数次幂, 在此 p 默认值是 1, 即隐藏容量为 1.00 bpp。

4) 提取准确率是指能够正确恢复的秘密消息比特占据全部隐藏消息比特的百分比。因此, 评估提取准确率的指标是比特恢复准确率(bit recovery accuracy)。

4.2 实验设计

为了评估本文方法 steg-Cinn 的性能表现, 本小节设计了三个实验, 分别在视觉质量、统计安全性和提取准确率方面进行了综合的评估分析。实验设计如下。

4.2.1 实验一: 视觉质量评估

本实验通过计算 Brisque[52]和 hyperIQA[53]分数, 来评估和比较本文方法 steg-Cinn、steg-glow[30]、ddh[16]和 udh[17]的视觉质量, 从而评价感知隐蔽性, 以及分析本文方法是否能够使得生成的图像内容是可控的。

4.2.2 实验二: 统计安全性评估

本实验采用隐写分析器 Ke-Net[56]来计算隐写分析检测正确率, 以便于评估和比较本文方法 steg-Cinn、steg-glow[30]、ddh[16]和 udh[17]这四种方法的统计安全性。

考虑到映射模块 M 的间隔参数 α 会影响载密图像的分布, 因此, 本实验还分析了 α 对 steg-Cinn 统计安全性的影响。从理论分析上来说, 只要载体图像和载密图像由相同的 cINN 模型合成, 并且载体和载密图像使用的驱动噪声的分布足够相似, 则隐写算法 steg-Cinn 是抗检测的。其中, 载体图像使用的驱动噪声记作 z_c, 载密图像使用的驱动噪声记作 z_s。由分析可知, 噪声 z_s 的分布由间隔参数 α 控制, α 越小, 噪声 z_c 和噪声 z_s 的分布就越接近, 载体和载密图像的分布也越接近, 抗检测性能就越强, 统计安全性越强。

4.2.3 实验三: 提取准确率评估

本实验通过计算比特恢复准确率(bit recovery accuracy), 来评估和比较本文方法 steg-Cinn、steg-glow[30]、ddh[16]和 udh[17]的消息提取准确率。此外, 本实验还分析了在秘密消息被准确提取的情况下, 不同隐写方法对应的隐藏容量。

4.3 实验结果

4.3.1 实验一: 视觉质量评估

表 2 展示的是对比方法(baseline)和本文方法的 Brisque[52]分数以及 hyperIQA[53]分数。由表格数据可知, 在 Brisque[52]评价指标下, 表现最好的是 udh[17]; 在 hyperIQA[53]评价指标下, 表现最好的是 steg-Cinn。横向观察可知, steg-Cinn 在两个指标下都比 steg-glow[30]表现更好。

表 2 在无参考图像质量评价下, 不同隐写方法的视觉质量分数

Table 2 Visual quality scores of different steganography methods measured by two no-reference image quality assessment (NR-IQA)

隐写方法	无参考图像质量评价(No-reference image quality assessment, NR-IQA)	
	↓ Brisque[52]	↑ hyperIQA[53]
ddh[16]	49.9708	20.27
udh[17]	**18.3516**(Top 1)	36.11(Top 2)
steg-glow[30]	33.2320	27.31
本文方法 steg-Cinn	21.7449(Top 2)	**38.46**(Top 1)

(注: “↑”表示对应分数值越大, 图像质量越好, “↓”表示对应分数值越小, 图像质量越好。)

图 7 展示的是对比方法(baseline)和本文方法在可视化方面的视觉效果。从图 7(b, c)可以看出, ddh[16]和 udh[17]合成的载密图像与载体图像高度相似。其中, 由于 ddh[16]和 udh[17]属于嵌入式隐写, 所以这两种方法对应的载体是真实自然图像, 对应载密图像是模型 ddh[16]或 udh[17]输出的图像; 而 steg-Cinn 和 steg-glow[30]都属于生成式隐写, 所以这两种方法对应的载体是噪声驱动 steg-Cinn 或 steg-glow[30]合成的图像, 对应载密图像是经过秘密消息映射之后的噪声驱动 steg-Cinn 或 steg-glow[30]合成的图像。从图 7(g,

h)可以看出, steg-glow[30]合成的图像在语义内容方面不可控, 即发送方不能控制合成图像的人脸属性、表情等等。这是因为, steg-glow[30]在合成图像的时候仅仅由单独的噪声驱动, 没有其它的条件信息来辅助合成过程。从图 7(e, f)可以看出, 本文方法 steg-Cinn 合成的图像颜色看起来真实自然, 并且, 在语义内容方面和灰度图像是匹配的。

因此, 与 steg-glow[30] 相比, 本文提出的 steg-Cinn 在视觉质量方面表现良好, 并且可以控制生成图像的语义内容, 从而保证了隐蔽通信的行为安全性。

4.3.2 实验二: 统计安全性评估

评估统计安全性的结果显示在表 3 中。由表格数据可知, ddh[16]和 udh[17]的隐写分析检测正确率都在 98%以上, 这意味着隐写分析器可以准确区分载体图像和载密图像的统计分布差异。而 steg-glow[30]和本文方法 steg-Cinn 的隐写分析检测正确率都在 56.5% 左右, 这意味着隐写分析器对于载体和载密图像统计分布的区分能力很低, 说明隐写方法的抗检测性能好, 可以保证统计安全性。这是因为, 它们的目标不同: steg-glow[30]和本文方法 steg-Cinn 追求的是载体和载密图像在统计分布上是接近的: 只要

图 7 不同隐写方法在合成图像方面的视觉效果。

Figure 7 Visual effects of synthesized images for different steganography methods.

表 3 不同隐写方法的统计安全性(隐写分析检测正确率)

Table 3 Statistical security(steganalyzer's detection rate) of different steganography methods

隐写方法	隐藏容量(bit per pixel, bpp)	↓ 由隐写分析器 Ke-Net[56]计算而来的检测正确率
ddh[16]	1.00	100.0 %
	0.25	100.0 %
udh[17]	1.00	99.04 %
	0.25	98.80%
steg-glow[30]	3.00	57.09 %
本文方法 steg-Cinn	2.00	**56.28 %**

(注: "↓"表示隐写分析检测正确率越低, 隐写算法抵抗隐写分析检测的能力就越强, 统计安全性就越好。)

合成载体和载密图像所用到的生成模型一致, 并且驱动载体合成的噪声和驱动载密图像合成的经过秘密消息映射的噪声在分布上是相似的, 那么载体和载密图像的统计分布也是相似的, 从而可以抵抗隐写分析器的检测。而 ddh[16]和 udh[17]追求的是载体和载密图像在像素级的意义上接近, 而不是在统计意义上接近, 因此它们很容易被隐写分析器检测到。由此观之, 与 ddh[16]和 udh[17]相比, 本文方法 steg-Cinn 的隐写分析检测正确率较低, 在统计分布保持方面更加安全。

另外, 本小节还进行了一个基于控制变量法的实验分析: 探索映射模块 M 的间隔参数 α 对于 steg-Cinn 统计安全性的影响, 如表 4。由表格数据可知, 随着 α 的增大, steg-Cinn 的安全性下降(隐写分析检测正确率上升)。这是因为, 驱动载体合成的噪声(记为 z_c)严格服从标准正态分布, 而驱动载密图像合成的噪声(记为 z_s)是“空心”的标准正态分布(使用间隔参数 α 调节) , 如图 8 所示。间隔参数 α 越大, z_c 和 z_s 在统计分布上的差异越大, 载体和载密图像在统计分布上的差异也越大, 就越容易被隐写分析器检测。由此分析, α 取 0 为最好。但是, 考虑到图像存储和传输过程可能会存在精度误差, 因此, 为了兼顾容错功能和统计安全性, 本文选取间隔参数 α 为 0.1。即, 除了本小节的控制变量实验以外, 其它实验均在 α=0.1 设置下进行。

表 4　不同 α(映射模块 M 的间隔参数)情况下, 本文方法 steg-Cinn 的统计安全性(隐写分析检测正确率)

Table 4　Statistical security(steganalyzer's detection rate) of the proposed steg-Cinn when α(interval parameter of mapping module M) changes

隐写方法	间隔参数 α	↓ 由隐写分析器 Ke-Net[56]计算而来的检测正确率
本文方法 steg-Cinn	0	51.09%
	0.10	56.28%
	0.15	64.02%
	0.20	71.82%

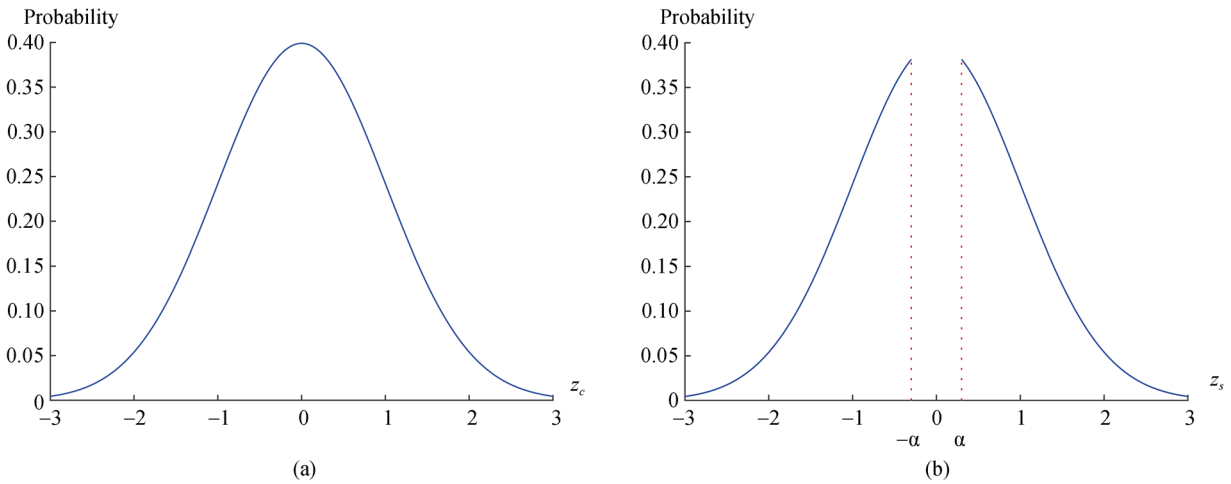

图 8　噪声隐变量的统计分布情况。(a)驱动载体图像合成的噪声 z_c 的分布; (b)驱动载密图像合成的噪声 z_s 的分布。

Figure 8　Distribution of noise latent variable. (a) Distribution of noise z_c which drives generating cover; (b) Distribution of noise z_s which drives generating stego.

4.3.3　实验三: 提取准确率评估

评估秘密消息的提取准确率的结果如表 5 所示, 其中, 提取准确率是通过逐比特位地提取秘密消息从而计算出来的, 单位是比特恢复准确率。

由表格数据可知, steg-glow[30] 和本文方法 steg-Cinn 都可以实现 100%正确提取。这是因为它们的结构基础是天然可逆的, 训练好以后的模型既可以作为嵌入器也可以作为提取器, 与隐写任务的目标一致。由于 steg-Cinn 和 steg-glow[30]的工作基础都是可逆网络, 而可逆网络输入和输出的总维数是相同的, 因此, steg-Cinn 可以隐藏的秘密消息比特个数和颜色信息 ab 的总维数(128*128*2)相同; steg-glow[30]可以隐藏的秘密消息比特个数和 RGB 像素总维数(128*128*3)相同。因此, steg-Cinn 的隐藏容量是 2.00 bpp, 对比方法 steg-glow[30] 的隐藏容量是 3.00 bpp。此外, 方法 ddh[16]和 udh[17]也能做到 100%

表 5 不同隐写方法的消息提取准确率(比特恢复准确率)

Table 5 The message extraction accuracy(bit recovery accuracy) of different steganography methods

隐写方法	隐藏容量(bit per pixel, bpp)	消息提取准确率(比特恢复准确率)
ddh[16]	1.00	98.21 %
	0.25	**100.00 %**
udh[17]	1.00	89.95 %
	0.25	**100.00 %**
steg-glow[30]	3.00	**100.00 %**
本文方法 steg-Cinn	2.00	**100.00 %**

正确提取, 但是这种情况下的隐藏容量只有 0.25 bpp。当提取准确率在 90%左右的时候, 隐藏容量依然不够高(1.00 bpp)。由此观之, 与 ddh[16]和 udh[17]相比, 本文方法具有更高的隐藏容量和更高的提取准确率。

5 结论

本文提出了一种基于条件可逆网络的生成式图像隐写算法 steg-Cinn: 在灰度图像的指导下, steg-Cinn 可以在图像上色过程中隐藏秘密消息, 使得合成图像的内容可控, 从而保证了隐蔽通信的行为安全。实验结果表明, steg-Cinn 生成的彩色图像可以很好地匹配灰度图像中的语义信息, 在视觉隐蔽性方面能够得到保证。此外, 在统计安全性、消息提取准确率和隐藏容量方面, 本文方法具有良好的综合性能表现。未来工作将会继续研究在有噪音隐蔽通信环境中的可逆网络隐写算法, 提高本方法的实用性。

参考文献

[1] Fridrich J. Steganography in digital media: principles, algorithms, and applications[M]. Cambridge University Press, 2009.

[2] Filler T, Judas J, Fridrich J. Minimizing Embedding Impact in Steganography Using Trellis-Coded Quantization[C]. *Proc SPIE 7541, Media Forensics and Security II*, 2010, 7541: 38-51.

[3] Holub V, Fridrich J. Designing Steganographic Distortion Using Directional Filters[C]. *2012 IEEE International Workshop on Information Forensics and Security*, 2013: 234-239.

[4] Pevný T, Filler T, Bas P. Using High-Dimensional Image Models to Perform Highly Undetectable Steganography[C]. *International Workshop on Information Hiding*, 2010: 161-177.

[5] Li B, Wang M, Huang J W, et al. A New Cost Function for Spatial Image Steganography[C]. *2014 IEEE International Conference on Image Processing*, 2015: 4206-4210.

[6] Holub V, Fridrich J, Denemark T. Universal Distortion Function for Steganography in an Arbitrary Domain[J]. *EURASIP J Information Security*, 2014, 2014(1): 1-13.

[7] Guo L J, Ni J Q, Shi Y Q. Uniform Embedding for Efficient JPEG Steganography[J]. *IEEE Transactions on Information Forensics and Security*, 2014, 9(5): 814-825.

[8] Guo L J, Ni J Q, Su W K, et al. Using Statistical Image Model for JPEG Steganography: Uniform Embedding Revisited[J]. *IEEE Transactions on Information Forensics and Security*, 2015, 10(12): 2669-2680.

[9] Tang W X, Tan S Q, Li B, et al. Automatic Steganographic Distortion Learning Using a Generative Adversarial Network[J]. *IEEE Signal Processing Letters*, 2017, 24(10): 1547-1551.

[10] Yang J H, Liu K, Kang X G, et al. Spatial Image Steganography Based on Generative Adversarial Network[EB/OL]. 2018: arXiv: 1804.07939. https://arxiv.org/abs/1804.07939

[11] Yang J H, Ruan D Y, Huang J W, et al. An Embedding Cost Learning Framework Using GAN[J]. *IEEE Transactions on Information Forensics and Security*, 2020, 15: 839-851.

[12] Tang W X, Li B, Barni M, et al. An Automatic Cost Learning Framework for Image Steganography Using Deep Reinforcement Learning[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 952-967.

[13] Zhang C N, Lin C G, Benz P, et al. A Brief Survey on Deep Learning Based Data Hiding[EB/OL]. 2021: arXiv: 2103.01607. https://arxiv.org/abs/2103.01607

[14] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. *Science*, 2006, 313(5786): 504-507.

[15] Baluja S. Hiding Images in Plain Sight: Deep Steganography[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 2066-2076.

[16] Baluja S. Hiding Images within Images[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(7): 1685-1697.

[17] Zhang C N, Benz P, Karjauv A, et al. UDH: Universal Deep Hiding for Steganography, Watermarking, and Light Field Messaging[C]. *The 34th International Conference on Neural Information Processing Systems*, 2020: 10223-10234.

[18] Hayes J, Danezis G. Generating Steganographic Images via Adversarial Training[EB/OL]. 2017: arXiv: 1703.00371. https://arxiv.org/abs/1703.00371

[19] Zhu J R, Kaplan R, Johnson J, et al. HiDDeN: Hiding Data with Deep Networks[EB/OL]. 2018: arXiv: 1807.09937. https://arxiv.org/abs/1807.09937

[20] Ardizzone L, Kruse J, Wirkert S, et al. Analyzing Inverse Problems with Invertible Neural Networks[EB/OL]. 2018: arXiv: 1808.04730. https://arxiv.org/abs/1808.04730

[21] Lu S P, Wang R, Zhong T, et al. Large-Capacity Image Steganog-

raphy Based on Invertible Neural Networks[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 10811-10820.

[22] Jing J P, Deng X, Xu M, et al. HiNet: Deep Image Hiding by Invertible Network[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2022: 4713-4722.

[23] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets[C]. *The 27th International Conference on Neural Information Processing Systems - Volume 2*, 2014: 2672-2680.

[24] Volkhonskiy D, Nazarov I, Burnaev E. Steganographic generative adversarial networks[C].*Twelfth International Conference on Machine Vision. International Society for Optics and Photonics*, 2020, 11433: 114333M.

[25] Shi H C, Dong J, Wang W, et al. SSGAN: Secure Steganography Based on Generative Adversarial Networks[C]. *Pacific Rim Conference on Multimedia*, 2018: 534-544.

[26] Li J, Niu K, Liao L W, et al. A Generative Steganography Method Based on WGAN-GP[C]. *International Conference on Artificial Intelligence and Security*, 2020: 386-397.

[27] Wang Z, Gao N, Wang X, et al. SSteGAN: self-learning steganography based on generative adversarial networks[C]. *International Conference on Neural Information Processing*, 2018: 253-264.

[28] Van Den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel Recurrent Neural Networks[C]. *The 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016: 1747-1756.

[29] Yang K, Chen K, Zhang W, et al. Provably secure generative steganography based on autoregressive model[C]. *International Workshop on Digital Watermarking*, 2018: 55-68.

[30] Chen K J, Zhou H, Hou D D, et al. Provably Secure Steganography on Generative Media[EB/OL]. 2018: arXiv: 1811.03732. https://arxiv.org/abs/1811.03732

[31] Wu Y, AbdAlmageed W, Natarajan P. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries with Anomalous Features[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 9535-9544.

[32] Boehm B. StegExpose - a Tool for Detecting LSB Steganography[EB/OL]. 2014: arXiv: 1410.6656. https://arxiv.org/abs/1410.6656

[33] Xu G S, Wu H Z, Shi Y Q. Structural Design of Convolutional Neural Networks for Steganalysis[J]. *IEEE Signal Processing Letters*, 2016, 23(5): 708-712.

[34] Ye J, Ni J Q, Yi Y. Deep Learning Hierarchical Representations for Image Steganalysis[J]. *IEEE Transactions on Information Forensics and Security*, 2017, 12(11): 2545-2557.

[35] Yedroudj M, Comby F, Chaumont M. Yedroudj-Net: An Efficient CNN for Spatial Steganalysis[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 2092-2096.

[36] Boroumand M, Chen M, Fridrich J. Deep Residual Network for Steganalysis of Digital Images[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 14(5): 1181-1193.

[37] Yang H, Ouyang H, Koltun V, et al. Hiding Video in Audio via Reversible Generative Models[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 1100-1109.

[38] Chen K J, Zhou H, Zhao H Q, et al. Distribution-Preserving Steganography Based on Text-to-Speech Generative Models[J]. *IEEE Transactions on Dependable and Secure Computing*, 2022, 19(5): 3343-3356.

[39] Prenger R, Valle R, Catanzaro B. Waveglow: A Flow-Based Generative Network for Speech Synthesis[C]. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2019: 3617-3621.

[40] Ardizzone L, Lüth C, Kruse J, et al. Guided Image Generation with Conditional Invertible Neural Networks[EB/OL]. 2019: arXiv: 1907.02392. https://arxiv.org/abs/1907.02392

[41] Dinh L, Krueger D, Bengio Y. NICE: Non-Linear Independent Components Estimation[EB/OL]. 2014: arXiv: 1410.8516. https://arxiv.org/abs/1410.8516

[42] Dinh L, Sohl-Dickstein J, Bengio S. Density Estimation Using Real NVP[EB/OL]. 2016: arXiv: 1605.08803. https://arxiv.org/abs/1605.08803

[43] Kingma D P, Dhariwal P. Glow: Generative Flow with Invertible 1x1 Convolutions[EB/OL]. 2018: arXiv: 1807.03039. https://arxiv.org/abs/1807.03039

[44] Tuchman W. A brief history of the data encryption standard[M].Internet besieged: countering cyberspace scofflaws. 1997: 275-280.

[45] Luby M, Rackoff C. How to Construct Pseudorandom Permutations from Pseudorandom Functions[J]. *SIAM Journal on Computing*, 1988, 17(2): 373-386.

[46] Fisher R A. Theory of Statistical Estimation[J]. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1925, 22(5): 700-725.

[47] Zhang X, Wandell B A. A Spatial Extension of CIELAB for Digital Color-Image Reproduction[J]. *Journal of the Society for Information Display*, 1997, 5(1): 61-63.

[48] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL]. 2014: arXiv: 1409.1556. https://arxiv.org/abs/1409.1556

[49] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[C]. *European Conference on Computer Vision*, 2014: 740-755.

[50] Bas P, Filler T, Pevný T. “Break our steganographic system”: the ins and outs of organizing BOSS[C]. *International workshop on information hiding*, 2011: 59-70.

[51] Zhang S F, Zhu X Y, Lei Z, et al. FaceBoxes: A CPU Real-Time Face Detector with High Accuracy[C]. *2017 IEEE International Joint Conference on Biometrics*, 2018: 1-9.

[52] Mittal A, Moorthy A K, Bovik A C. No-Reference Image Quality Assessment in the Spatial Domain[J]. *IEEE Transactions on Image Processing*: *a Publication of the IEEE Signal Processing Society*, 2012, 21(12): 4695-4708.

[53] Su S L, Yan Q S, Zhu Y, et al. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 3664-3673.

[54] Wang Z, Bovik A C, Sheikh H R, et al. Image Quality Assessment:

From Error Visibility to Structural Similarity[J]. *IEEE Transactions on Image Processing*: *a Publication of the IEEE Signal Processing Society*, 2004, 13(4): 600-612.

[55] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[56] You W K, Zhang H, Zhao X F. A Siamese CNN for Image Steganalysis[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 291-306.

刘婷 于 2015 年在中国地质大学(武汉)计算机学院获得学士学位。现在于武汉大学国家网络安全学院攻读硕士学位。研究领域为多媒体内容安全。研究兴趣包括: 隐写术、AI 安全。Email: leeeliu@whu.edu.cn。

王丽娜 于 2001 年在东北大学获得博士学位。现任武汉大学国家网络安全学院教授。研究领域为多媒体安全、云计算安全、网络安全。研究兴趣包括: 隐写术、信隐写分析、虚拟化、数字信号处理与识别等。Email: lnwang@whu.edu.cn

任延珍 于 2009 年在武汉大学通信与信息系统专业获得博士学位。现任武汉大学国家网络安全学院教授。研究领域为多媒体内容安全。研究兴趣包括: AI 交互安全, 多媒体取证, 多媒体伪造检测, 多媒体信息隐藏及隐写分析, 多媒体特征表示学习等。Email: renyz@whu.edu.cn