基于轮廓自动生成的构造式图像隐写方法

周志立[1,2]，王美民[1,2]，杨高波[3]，朱剑宇[1,2]，孙星明[1,2]

（1. 南京信息工程大学数字取证教育部工程研究中心，江苏 南京 210044；
2. 南京信息工程大学计算机与软件学院，江苏 南京 210044；3. 湖南大学信息科学与工程学院，湖南 长沙 410082）

摘 要：为解决现有构造式隐写方法隐藏容量小和秘密信息难以提取的问题，提出一种基于轮廓自动生成的构造式图像隐写方法，具体包括以秘密信息为驱动的轮廓线生成和从轮廓线到图像变换 2 个过程。首先，建立基于长短期记忆网络（LSTM）的轮廓自动生成模型，实现以秘密信息为驱动的图像轮廓线生成；然后，建立基于 pix2pix 模型的轮廓−图像可逆变换模型，将轮廓线变换为含密图像。该模型也支持含密图像到轮廓的逆变换，从而实现秘密信息提取。实验结果表明，所提方法不仅能有效地抵抗隐写分析攻击，还能实现较高的隐藏容量和准确的秘密信息提取，性能明显优于现有的同类构造式图像隐写方法。

关键词：构造式图像隐写；无载体信息隐藏；深度学习；生成对抗网络

中图分类号：TN911.73

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021174

Generative steganography method based on auto-generation of contours

ZHOU Zhili[1,2], WANG Meimin[1,2], YANG Gaobo[3], ZHU Jianyu[1,2], SUN Xingming[1,2]

1. Engineering Research Center of Digital Forensics, Ministry of Education,
Nanjing University of Information Science and Technology, Nanjing 210044, China
2. School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
3. College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

Abstract: To address the problems of limited hiding capacity and inaccurate information extraction in the existing generative steganography methods, a novel generative steganography method was proposed based on auto-generation of contours, which consisted of two main stages, such as the contour generation driven by secret information and the contour-to-image transformation. Firstly, the contour generation model was built based on long short term memory (LSTM) for secret information-driven auto-generation of object contours. Then, a contour-to-image reversible transformation model was constructed based on pix2pix network to obtain the stego-image, and the model also supported the reversible transformations from the stego-image to contours for secret information extraction. Experimental results demonstrate that the proposed method not only achieves high hiding capacity and accurate information extraction simultaneously, but also effectively resists the attacks by steganalysis tools. It performs much better than the state-of-the-art generative steganographic methods.

Keywords: generative steganography, coverless information hiding, deep learning, generative adversarial network

1 引言

信息隐写技术通常将秘密信息以不可见的方式嵌入多媒体数据（包括文本、图片、视频、音频等），然后通过传递含密多媒体数据以实现网络中的隐蔽通信任务[1-2]。相比于加密技术，信息隐写技

收稿日期：2021−05−18；**修回日期**：2021−08−09

基金项目：国家重点研发计划基金资助项目（No.2018YFB1003205）；国家自然科学基金资助项目（No.61972205, No.61972143）

Foundation Items: The National Key Research and Development Program of China (No.2018YFB1003205), The National Natural Science Foundation of China (No.61972205, No.61972143)

术不仅可以保护隐藏数据，并且可以隐藏“秘密通信”事件的本身。

由于数字图像具有获取容易、冗余信息量大的特点，因而被广泛用作信息隐写的载体[3]。传统的图像隐写方法通过对一个现有的载体图像进行轻微的修改，如根据秘密信息修改图像像素最低有效位（LSB, least significant bit）的值，从而实现秘密信息嵌入[4]。为了减少修改图像信息带来的图像失真，研究者进一步提出了一些自适应图像隐写方法，如 S-UNIWARD[5]、HILL[6]、HUGO[7]、WOW[8]等。这些方法先手工设计关于图像像素修改的失真函数，然后以最小化图像总体失真为目标将秘密信息自适应地嵌入图像。文献[9]提出了基于生成对抗网络（GAN, generative adversarial network）[10]的失真函数学习框架，自动地学习并设计适用于图像隐写任务的失真函数。文献[11]利用强化学习（RL, reinforcement learning）优化失真函数。相较于手工设计的失真函数，在相同的隐写负载下，神经网络学习到的失真函数能更进一步地减少图像失真。但是，上述图像隐写方法仍然对载体图像进行了不同程度的修改，从而在含密图像中不可避免地留下一些修改痕迹[12]。残留的修改痕迹使信息隐藏的事实很容易被 Rich Model[13]、XuNet[14]等强大的隐写分析工具成功检测。

为了增强信息隐写的安全性，研究者提出了构造式信息隐写[15-16]这一全新的研究思路。不同于传统的修改式图像隐写方法，构造式图像隐写以秘密信息为驱动直接“构造”出一幅全新的图像作为含密图像，从而实现隐蔽通信[17]。文献[18-19]提出了基于纹理图像构造的隐写方法，即以秘密信息为驱动构造出相应的纹理图像，从而将秘密信息隐写在构造的纹理图像中。此外，文献[20-21]提出了基于指纹图像的构造式隐写方法，对秘密信息进行编码以生成含密的指纹图像。然而，纹理和指纹图像并不是常见的自然图像，它们在某种程度上类似于加密数据，容易引起攻击者的怀疑，从而影响“秘密通信”事件本身的隐蔽性。因此，如何生成“真实自然”的含密图像成为研究的关键。近年来，具有生成“真实自然”图像能力的生成对抗网络的出现为解决以上问题提供了较好的技术基础。文献[22]将秘密信息编码成对应的类别标签，并以标签为驱动采用 ACGAN（auxiliary classifier GAN）生成相应的含密图像。然而标签信息能表达的秘密信息长度非常有限，此类方法通常在一幅图像中仅能够隐藏十几比特信息。文献[23]将秘密信息转换为低维噪声信号作为 DCGAN（deep convolutional GAN）的输入，从而生成相应的含密图像。以上基于 GAN 的构造式隐写方法本质上是训练 GAN 以实现噪声信号与图像隐式特征之间的映射。然而，由于 GAN 的图像生成不是马尔可夫过程，该映射过程难以通过训练 GAN 来实现，从而导致信息隐藏容量仍然非常有限，并且秘密信息提取网络同样难以训练，无法从生成图像中准确提取秘密信息。因此，这些基于 GAN 的构造式隐写方法难以应用于实际的隐写任务。

为了解决现有的构造式信息隐写方法中隐藏容量小以及秘密信息提取困难的问题，本文采用两阶段的图像构造思路：首先将秘密信息映射为图像的显式特征，即轮廓信息，然后将轮廓信息作为 GAN 的输入构造出相应的含密图像。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，因此该方法较容易训练出相应的图像生成网络和秘密信息提取网络，从而有可能实现较高的隐藏容量和秘密信息提取精度。

根据以上思路，本文提出一种基于轮廓自动生成的构造式图像隐写方法。该方法包括轮廓线生成和图像生成 2 个阶段。首先构建基于长短期记忆网络（LSTM, long short-term memory）[24]的轮廓自动生成模型，实现以秘密信息为驱动的轮廓线生成。通过预设合理的损失函数，经过大量训练后的 LSTM 可以自动学习出轮廓点的概率分布，并最终生成轮廓。与传统的马尔可夫模型相比，LSTM 具有更好的适应性，生成的内容更符合客观世界的自然规律[25]。然后建立基于 pix2pix 模型的轮廓−图像可逆变换模型，从而将生成的轮廓线变换为含密图像。接收方利用该轮廓−图像可逆模型，可以将含密图像逆变换为轮廓线，再利用轮廓生成模型从中恢复出秘密信息。

2　相关技术

2.1　循环神经网络

传统的神经网络中，从输入层到隐藏层再到输出层，层与层之间的神经元是全连接的，但每层内的节点是无连接的。由于其结构限制，传统的神经网络只能处理一个一个单独的输入数据，不能处理包含多个具有时序关系数据的序列数据。

循环神经网络（RNN, recurrent neural network）作为一种特殊的神经网络，其隐藏层之间的节点是有连接的，隐藏层的输入不仅包括输入层的输出，还包括上一时刻隐藏层的输出，因此循环神经网络适用于处理序列数据，其模型框架如图 1 所示。在该网络中，一个序列当前时刻的输出与之前时刻的输出有关，即网络会记忆之前时刻的信息并将其应用于当前输出的计算中。t 时刻隐藏层状态 H_t 和输出层的输出 O_t 的计算式分别为

$$\begin{aligned} H_t &= f_H\left(W_{XH}X_t + W_{HH}H_{t-1}\right) \\ O_t &= g_O\left(W_{HO}H_t + b_O\right) \end{aligned} \tag{1}$$

其中，f_H 和 g_O 分别表示隐藏层和输出层的激活函数，X_t 表示 t 时刻的输入，H_{t-1} 表示 $t-1$ 时刻隐藏层的状态，W_{XH}、W_{HH} 和 W_{HO} 分别表示输入层到隐藏层、隐藏层神经元之间和隐藏层到输出层的权重参数，b_O 表示输出层中的偏置。RNN 具有较好的处理序列数据的能力，但简单的 RNN 在处理较长的序列时，由于梯度返回困难，容易产生梯度消失和梯度爆炸问题，导致模型难以收敛。LSTM 在传统 RNN 的基础上添加了门控机制，即输入门、遗忘门和输出门。遗忘门根据上一时刻隐藏层状态 H_{t-1} 和当前时刻输入 X_t，通过 sigmoid 激活函数舍弃部分旧的信息；输入门根据上一时刻隐藏层状态 H_{t-1} 和当前时刻输入 X_t，通过 tanh 激活函数保留新的信息；输出层根据上一时刻隐藏层状态 H_{t-1} 和当前时刻输入 X_t，通过 tanh 激活函数计算当前时刻的隐藏层状态 H_t，因此能够处理序列较长的数据，有效避免梯度消失和梯度爆炸的问题。具有门控结构的 LSTM 更适于各种实际的应用，如自然语言处理、语音识别、商品推荐等。

2.2 生成对抗网络

GAN 是一种深度学习模型，基本结构包括 2 个相互对抗的模块：生成模型 G 和判别模型 D。生成模型 G 的目的在于通过输入的一个噪声信号 z 生成逼真的样本，使判别器无法判断真伪；判别模型 D 的目的是尽量判别一个样本是来自真实样本还是生成样本。G 和 D 使用以下损失函数进行训练。

$$\mathcal{L}_{\text{GAN}}(G,D) = \mathbb{E}_y\left[\log D(y)\right] + \mathbb{E}_z\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{2}$$

其中，y 表示真实图像的分布，z 表示随机噪声的分布。网络训练的最终目的是通过 G 和 D 的相互对抗学习从而获得强大的生成器 G。由于 G 和 D 是在相互博弈中共同进步的关系，训练时通常固定其中一个模型参数对另一个模型进行交替训练。在一次对抗训练中，首先固定 G 中的参数，按照梯度上升算法最大化 $\mathcal{L}_{\text{GAN}}(G,D)$，以更新 D 的参数；然后固定 D 中的参数，最小化 $\max\limits_D \mathcal{L}_{\text{GAN}}(G,D)$，以更新 G 的参数。整体训练目标可以表示为式(3)，通过多次迭代训练，最后得到生成器 G^*。

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{GAN}}(G,D) \tag{3}$$

由于经典 GAN 仅以随机噪声 z 为输入来生成数据，因此生成器的生成过程是不可控的，无法根据预先设定的约束条件生成相应的图像。因此，Mirza 等[26]在 GAN 的基础上进行改进提出了 cGAN（conditional GAN），通过对生成模型添加约束条件 x 来解决这个问题，其损失函数为

$$\begin{aligned} \mathcal{L}_{\text{cGAN}}(G,D) = {} & \mathbb{E}_{x,y}\left[\log D(x,y)\right] + \\ & \mathbb{E}_{x,z}\left[\log\left(1-D\left(x,G(x,z)\right)\right)\right] \end{aligned} \tag{4}$$

Isola 等[27]提出了一种特殊的 cGAN 即 pix2pix 模型。该网络在生成模型中添加输入图像（如轮廓图）为约束条件，以生成符合具有该轮廓信息的图像，其结构如图 2 所示。首先将轮廓图输入生成器中以得到生成图像，然后将生成图像和轮廓图以及真实图像和轮廓图成对地输入判别

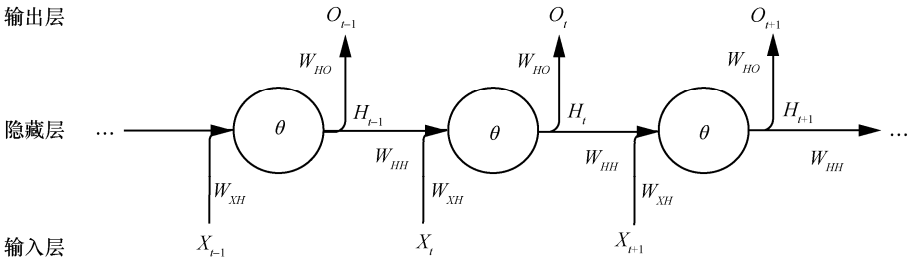


图 1 RNN 模型框架

器中，判别器来判断这些图像对的真伪性。同时，为了降低生成图像的模糊程度，提高视觉质量，pix2pix 在 cGAN 的基础上加入 L1 距离损失，L1 距离的计算式为

$$\mathcal{L}_{\mathrm{L1}}(G)=\mathbb{E}_{x,y,z}\left[\left\|y-G(x,z)\right\|_1\right] \tag{5}$$

其中，y 表示真实图像的分布，$G(x,z)$ 表示生成图像的分布。pix2pix 模型的最终训练目标如式(6)所示，由 cGAN 损失和 L1 距离损失两部分组成，λ 为权重参数。

$$G^{*}=\arg\min_{G}\max_{D}\mathcal{L}_{\mathrm{GAN}}(G,D)+\lambda\mathcal{L}_{\mathrm{L1}}(G) \tag{6}$$

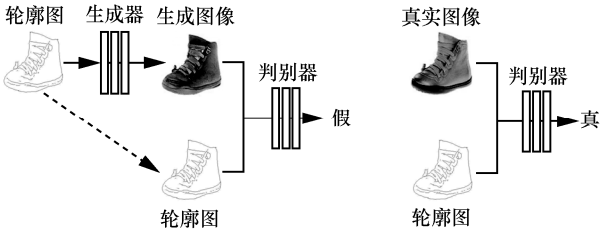


图 2　pix2pix 模型框架

pix2pix 模型中生成器采用了 U-Net 网络[28]结构，U-Net 网络在传统的编解码器网络中加入了跳跃连接，连接了降采样与升采样的相应卷积层，从而保留了不同分辨率的图像细节信息，对完善生成图像的细节起到重要的作用。pix2pix 采用 PatchGAN[27]作为判别器。PatchGAN 并非将整张图像输入判别器中，而是将图像分割成 $N\times N$ 大小的图像块，这些图像块是可以被认为相互独立的。判别器对每个图像块进行真假判别，然后取所有图像块结果的平均值作为判别器的判别结果。这些设计被证明不仅能保证较快的训练速度，还能保证较高的图像生成质量[27]。

3　算法设计

3.1　整体框架

为了解决现有构造式图像隐写方法的隐藏容量低以及秘密信息提取困难的问题，本文提出一种基于轮廓自动生成的构造式图像隐写方法，总体框架如图 3 所示。在隐写阶段，以待隐写的秘密信息为驱动，利用基于 LSTM 的轮廓自动生成模型生成相应的轮廓线；然后，将生成的轮廓线输入基于 pix2pix 的轮廓−图像可逆变换模型，从而将生成的轮廓线变换为含密图像。接收方利用该轮廓−图像可逆变换模型，可以将含密图像逆变换为轮廓线，然后从轮廓线中恢复出秘密信息。

3.2　秘密信息驱动的轮廓自动生成

为了保证最终生成的含密图像在视觉上不容易引起攻击者怀疑，用于生成图像的轮廓线应符合真实物体轮廓线的统计规律。基于以上考虑，本文首先建立基于 LSTM 的轮廓自动生成模型，然后以秘密信息为驱动生成相应轮廓线。

本文以生成简单的山脉图像为例，设计轮廓自动生成算法。对于一幅宽为 w、高为 h 的山脉轮廓图像而言，其山脉轮廓线可以视为在二维图像平面的 w 个连续的轮廓点组成的曲线 $C=\{c_i\mid i=1,2,\cdots,n\}$，其中 c_i 表示该曲线的第 i 个轮廓点，其取值应不超过图像的高度值，即 $c_i\in[1,h]$。本文通过从大量自然图像中提取轮廓线数据，输入 LSTM 中进行训练以获得轮廓自动生成模型，其中 LSTM 的输出函数采用 softmax 激活函数。

根据式(1)和式(2)，LSTM 的第 t 时刻的输出与前 $t-1$ 个时刻的输入都有关系。同时，为了简化表达，f_{LSTM} 表示 LSTM 的输出函数，第 t 时刻的网络

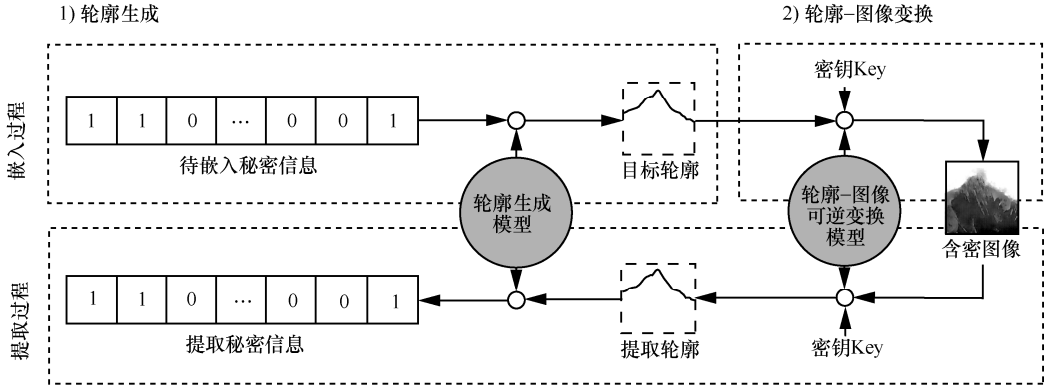


图 3　本文方法总体框架

输出可以表示为

$$O_t = f_{\mathrm{LSTM}}\left(x_t \mid x_1, x_2, \cdots, x_{t-1}\right) \tag{7}$$

本文将曲线 C 的轮廓点依次输入，那么 LSTM 的第 i 个输出可以作为轮廓点 c_i 的条件概率分布，表示为 $\Pr\left(c_i \mid c_1, c_2, \cdots, c_{i-1}\right)$，$c_i$ 的位置值可以是[1,h]中的任意值。为了使训练的轮廓自动生成模型与真实的轮廓线样本分布特性一致，本文使用每个轮廓线的概率值定义损失函数为

$$\begin{aligned}&\mathrm{Loss} = -\log\left(\Pr(C)\right) = -\log\left(\Pr\left(c_1, c_2, \cdots, c_n\right)\right) = \\ &-\log\left(\Pr\left(c_1\right)\Pr\left(c_2 \mid c_1\right)\cdots\Pr\left(c_n \mid c_1, c_2, \cdots, c_{i-1}\right)\right) = \\ &-\sum_{i=1}^{n}\log\left(\Pr\left(c_i \mid c_1, c_2, \cdots, c_{i-1}\right)\right)\end{aligned} \tag{8}$$

在以上训练过程中，本文采用梯度下降算法更新网络参数。通过最小化损失函数，可以得到与真实轮廓线的统计特性基本一致的轮廓线自动生成模型。然后利用该生成模型，将以秘密信息驱动生成山脉轮廓线。轮廓线生成过程包含以下几个步骤。

1) 秘密信息序列切分。假设需要嵌入的秘密信息是一个长度为 L 的二值序列 S，首先将其均等切分成一系列长度为 l 的二进制片段，$l \in \left[1, \mathrm{lb}h\right]$ 为整数，具体取值将在实验部分详细讨论。然后转换为十进制表达形式，即 $S = \left\{s_i \mid i = 1, 2, \cdots, m\right\}$，$s_i$ 表示第 i 个序列片段对应的十进制形式。

2) 起始位置初始化。为了使生成的山脉轮廓符合自然山脉轮廓的统计规律，应尽可能避免出现山体陡峭等极端情况。因此，山脉轮廓的起始位置不宜过高或过低。利用随机函数 Rand 和密钥 Key_1 将 s_1 映射到[h/4, 3h/4]，得到起始点 c_1 的位置为

$$c_1 = \mathrm{Rand}_{\mathrm{Key}_1}\left(s_1\right) \tag{9}$$

3) 候选池生成。在轮廓生成过程中，输入已确定的轮廓点 $c_1 \sim c_{i-1}$ 到预训练的轮廓自动生成模型，可以得到当前点 c_i 所有可能位置的概率分布。然而，用于生成轮廓线不仅可以用概率最高的位置作为轮廓点位置，其他概率较高的位置也同样可以作为轮廓点位置。因此，本文可以把秘密信息编码为 c_i 的位置选择，从而在轮廓线生成过程中嵌入秘密信息。为了便于每个轮廓点 c_i 的选择，本文首先建立相应的轮廓点位置候选池。

假设前期确定的山脉轮廓点为 $[c_1, c_2, \cdots, c_{i-1}]$，输入轮廓线自动生成模型中进行计算，模型的输出为下一个轮廓点 c_i 所有可能位置的概率为 $\Pr\left(c_i \mid c_1, c_2, \cdots, c_{i-1}\right)$。将 c_i 所有可能位置的概率按降序排列，选取概率较高的前 $K = 2^l$ 个位置。为了加强信息隐写的安全性，利用 Key_i 对这 K 个位置进行置乱，得到所有候选项的最终位置候选池 Pool_i，如式(10)所示。

$$\mathrm{Pool}_i = \left\{p_1, p_2, \cdots, p_K\right\} \tag{10}$$

4) 轮廓点选择。对于轮廓点 c_i，以秘密信息为驱动，从位置候选池 Pool_i 中选择相应的位置作为该轮廓点 c_i 的位置。由于步骤 1)中将秘密信息二进制序列以长度 l 为单位进行切分，那么对于秘密信息片段的十进制 s_i，其取值范围为 $\left[0, 2^l - 1\right]$，而候选池总共有 $K = 2^l$ 个位置。那么本文可以用 Pool_i 中第 $s_i + 1$ 位置作为 c_i 的位置，表示为

$$c_i = \mathrm{Pool}_i\left[s_i + 1\right] \tag{11}$$

根据式(11)，通过对 c_i 的位置选择可以隐藏 lb$K$ bit 信息。假设总共生成含有 n 个轮廓点的轮廓线，那么在轮廓线生成时可以总共隐藏 $n$lb$K$=nl bit 秘密信息。为了进一步解释信息嵌入过程，以图 4 为例进行说明。首先将秘密信息切分成一系列长度为 $l = 3$ bit 的序列段；对于秘密信息 $s_1 = 6$，通过式(9)，得到 c_1 的位置；对于秘密信息 $s_2 = 2$，那么可以选择 $\mathrm{Pool}_1[3]$ 作为轮廓点 c_2 的位置；按照以上方式确定之后所有轮廓点 s_i 的位置($i \geqslant 2$)，最终生成的轮廓线总共可以隐藏 3n bit 信息。具体步骤如算法 1 所示。

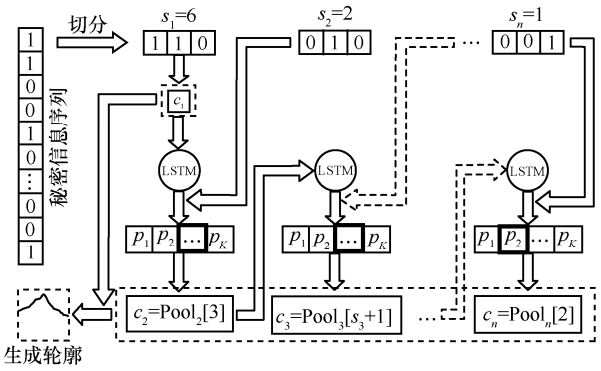


图 4 以秘密信息为驱动的轮廓自动生成过程

算法 1 以秘密信息为驱动的轮廓生成算法

输入 待嵌入秘密信息的十进制序列 $S = \left\{s_i \mid i = 1, 2, \cdots, n\right\}$

输出 所有轮廓点位置集合 $[c_1, c_2, \cdots, c_n]$ 组成的轮廓图像

1) 初始化起始位置 $c_1 = \mathrm{Rand}_{\mathrm{Key}_1}(s_1)$

2) 循环

3) for $i = 2:1:n$

4) 循环

5) 根据基于 LSTM 的轮廓自动生成模型，计算 c_i 所有可能位置的概率

6) 取概率最高的前 K 个点，利用 Key_i 进行置乱，生成位置候选池 Pool_i

7) 根据 s_i 确定 c_i 的位置 c_i=$\mathrm{Pool}_i[s_i+1]$

8) end for

注意到，此时的生成轮廓只是比较粗糙的曲线，并非“真实自然”图像，不能直接用于隐蔽通信任务。因此，下一步将轮廓线转换为包含该轮廓线信息的“真实自然”图像。

3.3 含密图像构造

通过 3.2 节以秘密信息为驱动的轮廓线生成，秘密信息已经被嵌入生成的目标轮廓中。接下来，需要将生成轮廓变换成尽量逼真的图像，并保证秘密信息能从生成图像中准确提取。为了达到以上目标，本文对 pix2pix 模型进行改进，在原有的基础上添加了提取器，建立了基于 pix2pix 的轮廓−图像可逆变换模型。

轮廓−图像可逆变换模型结构如图 5 所示，包括生成器 G、判别器 D、提取器 E 共 3 个部分。生成器和判别器沿用了 pix2pix 模型的设计，与生成器和提取器网络均采用 U-Net 网络结构。需要注意的是，虽然轮廓−图像可逆变换中生成器与提取器的结构是一致的，但其中的网络参数并不一致，这是因为需要分别对生成器和提取器进行训练。另外，为了获得多样性的图像输出，pix2pix 在训练和测试中使用 dropout 机制时加入了随机噪声。而在图像隐写任务中，为了保证秘密信息提取的准确性，一个秘密信息序列有且仅有一幅含密图像与之对应，因此本文在网络设计中没有引入随机噪声。

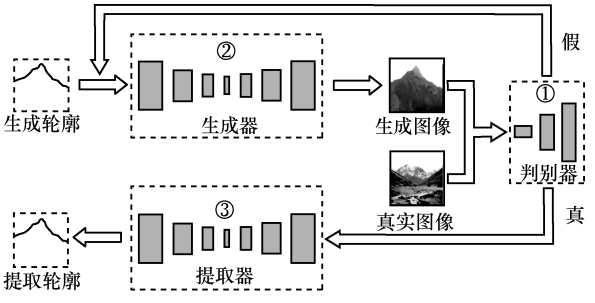


图 5　轮廓−图像可逆变换模型结构

在训练阶段，与 pix2pix 模型相似，该模型的生成器、判别器、提取器的训练分别独立进行。模型训练时首先固定生成器和提取器的网络参数，将生成的含密图像和真实图像输入判别器中，通过梯度下降更新参数判别器参数 θ_D，使判别器尽量判别出含密图像的真假；接着固定判别器和提取器，更新生成器中的网络参数 θ_G，以尽量欺骗判别器，即让判别器难以判别图像的真伪；然后固定生成器和判别器，更新提取器网络参数 θ_E，以尽量准确提取出轮廓线。当生成器、判别器、提取器均独立训练一次时，完成一次迭代训练。随着迭代训练次数的增加，生成的含密图像越来越逼真，提取的轮廓更接近原始目标轮廓。训练目标函数如式(12)所示。该目标函数对 pix2pix 的目标函数进行了改进，在原有基础上添加了 $\mathcal{L}_{\mathrm{L1}}(E)$，即提取轮廓与生成轮廓的 L1 距离，用于提取器的训练和优化，λ 和 μ 为权重参数。

$$G^*, E^* = \arg\min_G \max_D \min_E \mathcal{L}_{\mathrm{GAN}}(G,E) + \lambda\mathcal{L}_{\mathrm{L1}}(G) + \mu\mathcal{L}_{\mathrm{L1}}(E) \tag{12}$$

通过以上迭代训练，直至目标函数收敛，模型达到纳什均衡。经过训练后的生成器和提取器可以实现轮廓与图像的双向变换。本文利用轮廓−图像可逆变换模型中的生成器，实现轮廓到图像的变换，将得到的图像作为含密图像，同时也可以实现含密图像到轮廓的逆变换，实现秘密信息的提取。为了方便通信双方发送和接收秘密数据，本文中的轮廓自动生成模型与轮廓−图像可逆变换模型均是公开的，秘密通信双方共享一组密钥 $\mathrm{Key}=\{\mathrm{Key}_1,\mathrm{Key}_2,\cdots,\mathrm{Key}_n\}$。

3.4 秘密信息提取

在接收端，对于一幅接收到的含密图像，接收方进行秘密信息恢复。该秘密信息恢复是秘密信息嵌入的逆过程。其具体步骤如下。

1) 含密图像到轮廓线转换。接收方利用轮廓−图像可逆变换模型中的提取器从含密图像提取出轮廓线 $C=\{c_i \mid i=1,2,\cdots,n\}$。

2) 轮廓线到秘密信息恢复。该步骤分为以下 3 个小步骤。

①利用 Key_1 对 c_1 进行解密得到 s_1。

②将前期确定的山脉轮廓点 $\{c_1,c_2,\cdots,c_{i-1}\}$ 输入 LSTM 的隐藏层中进行计算，从而输出下一个轮廓点 c_i 所有可能位置的概率 $\Pr(c_i \mid c_1,c_2,\cdots,c_{i-1})$。将

c_i 所有可能位置的概率按降序排列，选取概率较高的前 $K=2^l$ 个位置；然后利用 Key_i 对前 K 个点的概率进行置乱，得到所有候选位置的候选池 Pool_i。观察 c_i 的位置在 Pool_i 的序号 ind_i。根据式(13)，可以得到

$$s_i = \text{ind}_i - 1 \tag{13}$$

③根据以上步骤，可以得到所有的十进制形式的秘密信息片段 $S=\{s_i \mid i=1,2,\cdots,m\}$，将其转换为二进制序列，从而恢复出最终的秘密信息。

4 实验结果与分析

本节主要介绍实验结果并进行分析，首先对参数设置进行讨论，然后分别从提取率、抗隐写分析能力和生成图像质量方面对所提方法进行性能评价，并与现有的经典方法进行比较。

以上实验均是在 Ubuntu 20.04 环境下采用 Tensorflow 2.1.0 实现的，采用的显卡型号是 GTX 1080ti。需要隐写的秘密信息是由计算机随机生成的二值序列，共 10 000 份。本文建立了包含 500 幅分辨率为 256 像素×256 像素的彩色真实山脉图像库作为训练集；采用本文方法和训练好的模型，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，总共得到 10 000 幅含密图像。

4.1 参数设置讨论

本文方法包含 3 个重要的参数：训练轮廓−图像可逆变换模型的权重参数 λ 和 μ，以及秘密片段切分长度 l。轮廓−图像可逆变换模型训练的目的是得到一个生成器和一个提取器，λ 取值较大时优先训练生成模型，μ 取值较大时优先训练提取模型。因此，λ 和 μ 取不同值将对信息提取准确率和图像质量有影响；参数 l 的取值直接影响隐藏容量、信息提取准确率以及生成的图像质量。由于本文实验采用图像的大小为 256 像素×256 像素，l 的最大取值为 $\text{lb}256=8$。当 l 取值较大时，表示每个轮廓点可以嵌入更多的比特数，隐藏容量上升但可能引起信息提取准确率和生成图像质量的下降。

表 1 列出了 λ 和 μ 分别取不同值时对信息提取率和生成图像质量的影响，此时 l 固定为 5，提取率和图像质量分别用 BER（bit error rate）和 EMD（earth move distance）来衡量。BER 值越小表示提取准确率越高；EMD 值表示生成图像集和真实图像集之间的统计距离，越接近 0 则表示生成图像与真实图像的距离越小，生成图像质量越高。从表 1 可以看出，当 λ 和 μ 取值相同时，BER 和 EMD 的波动较小；当固定 λ 时，随着 μ 的增长，BER 呈现不明显的下降趋势而 EMD 逐渐上升；当固定 μ 时，随着 λ 的增长，BER 呈现上升趋势而 EMD 下降。为了在信息提取准确率和图像质量之间获得较好的平衡，在本文实验中令 λ=50，μ=1。

表 1 λ 和 μ 取值不同时 BER 和 EMD 的变化

λ	μ=1		μ=50		μ=100	
	BER	EMD	BER	EMD	BER	EMD
1	0.015 8	0.027 8	0.015 4	0.029 1	0.015 5	0.029 6
50	0.016 1	0.026 1	0.015 7	0.028 1	0.015 8	0.028 7
100	0.017 2	0.026 2	0.016 2	0.027 9	0.015 8	0.026 9

表 2 中列出了 l 取不同值时 Bit、BER 和 EMD 的变化趋势，其中 Bit 表示每幅图像中可以嵌入的比特数量。通过对比不难发现，当 l 取值逐渐增大时，生成图像的隐藏容量随之增大；BER 也轻微升高，即信息提取的错误率升高；同时 EMD 值也迅速增长，表示生成图像与真实图像相似度降低，生成图像的质量变差。在本文实验中，通过 l 取不同值，观察本文提出的隐写方法在不同隐藏容量下提取率、抗隐写分析能力和图像生成质量方面的性能。

表 2 l 取值不同时 Bit、BER、EMD 的变化

l	Bit	BER	EMD
2	2×256	0.012 7	0.020 5
4	4×256	0.014 5	0.027 0
6	6×256	0.017 6	0.026 6
8	8×256	0.019 8	0.055 5

4.2 提取率

表 3 表示在不同隐藏容量下本文方法和 SWE 方法[23]提取率的对比。从表 3 中可以看出，随着隐藏容量的增加，本文方法的 BER 逐渐升高，提取准确率下降，但依然优于 SWE 方法，主要原因归纳如下。SWE 方法将秘密信息隐射为噪声信号，然后输入 DCGAN 中生成含密图像，其本质实现噪声信号与图像隐式特征之间的映射。由于 DCGAN 的图像生成不是马尔可夫过程，该映射过程难以通过

训练 DCGAN 来实现，从而导致信息隐藏容量仍然非常有限，并且秘密信息提取网络同样难以训练，无法准确提取秘密信息。不同于 SWE 方法，本文方法首先将秘密信息映射为图像的显式特征即轮廓信息，然后将轮廓信息为 GAN 的输入构造出相应的含密图像。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，与现有构造式信息隐写方法相比，本文方法将很容易训练出相应的图像生成网络和秘密信息提取网络，可以同时实现较高的隐藏容量和准确的秘密信息提取。

表 3　不同隐藏容量下的本文方法和 SWE 方法提取率对比

Bit	本文方法	SWE 方法
2×256	**0.012 7**	0.071 9
4×256	**0.014 5**	0.073 2
6×256	**0.017 6**	0.079 4
8×256	**0.019 8**	0.080 3

4.3　抗隐写分析能力

本文采用常用的 SRM[13]和 XuNet[14]隐写分析器进行攻击测试，以测试各种不同隐写方法抗隐写分析检测的能力。SRM 通过滤波器手动提取隐写图像的一系列一阶和二阶统计特征，并用这些统计特征训练支持向量机（SVM, support vector machine）进行隐藏图像检测。XuNet 则采用 CNN 自动提取隐写图像的特征并学习，从而使网络具备判别隐写图像的能力。在抗隐写分析测试中，将用到三类图像：真实图像、未嵌入信息的生成图像和嵌入信息的生成图像。实际上，嵌入信息的生成图像是未嵌入信息的生成图像的一种特殊形式，二者是包含与被包含的关系。因此，本文将真实图像和未嵌入信息的生成图像合计 10 000 幅作为不含密的图像数据集，而将 10 000 幅嵌入秘密信息的生成图像作为含密图像数据集。用于对比的隐写方法包括 S-UNIWARD[5]、UT-6HPF-GAN[9]和 SWE[23]。本实验采用 P_E 对隐写算法抗隐写分析检测能力进行评价，如式(14)所示。

$$P_E = \min_{P_{\mathrm{FA}}} \frac{1}{2}\left(P_{\mathrm{FA}} + P_{\mathrm{MD}}\right) \tag{14}$$

其中，P_{FA} 和 P_{MD} 分别表示错检率和漏检率，P_E 值越接近 0.5，表示抵抗隐写分析性能越好。

表 4 表示针对 SRM 和 XuNet 隐写分析器，不同隐藏容量下各类隐写方法的抗隐写分析能力的对比。从表 4 可以发现，不论隐写分析器采取 SRM 还是 XuNet，在隐藏容量相同的情况下，S-UNIWARD 和 UT-6HPF-GAN 的 P_E 值比 SWE 方法和本文方法的 P_E 值要低，也就是说传统修改式隐写方法的抗隐写分析能力弱于构造式隐写方法。这是因为载体修改式隐写方法按照一定规则修改载体图像以实现秘密信息的嵌入，那么含密图像中必然留有篡改痕迹，从而为隐写器提供检测依据。而 SWE 方法和本文方法并不需要修改载体图像，而是在秘密信息的驱动下构造出新图像作为含密图像。在隐写过程没有留下修改痕迹，因此不容易引被现有的隐写分析器成功地检测。

表 4　针对 SRM 和 XuNet 隐写分析器，不同隐藏容量下各类隐写方法的 P_E 值对比

隐写分析器	方法	Bit			
		2×256	4×256	6×256	8×256
SRM	S-UNIWARD	0.455 8	0.452 1	0.448 9	0.441 2
	UT-6HPF-GAN	0.461 7	0.465 3	0.457 6	0.451 2
	SWE 方法	0.478 9	0.473 8	0.476 4	0.473 2
	本文方法	**0.493 4**	**0.490 2**	**0.489 6**	**0.492 1**
XuNet	S-UNIWARD	0.457 9	0.454 3	0.456 7	0.450 1
	UT-6HPF-GAN	0.463 4	0.460 9	0.468 9	0.453 5
	SWE 方法	0.481 3	0.481 8	0.486 4	0.487 6
	本文方法	**0.498 8**	**0.499 5**	**0.501 1**	**0.500 2**

在这些对比方法中，本文方法表现最好，在不同隐藏容量下 P_E 的取值均非常接近 0.5，表明本文方法在实现较大隐藏容量的同时，能成功地抵抗以上隐写分析工具的检测。

4.4　图像生成质量

由于 t-SNE 可视化方法常用于数据相似度的可视化，本文采用 t-SNE 可视化方式对比 l 取不同值时生成图像与真实图像数据分布。图 6 为 l 取不同值时的对比结果。

此外，本文采用常用的 FID（frechet inception distance）对图像生成质量进行评价，FID 越小表示生成图像的分布与真实图像越接近。本文方法在不同隐藏容量情况下，与 SWE 方法比较生成的图像质量，结果如表 5 所示。由表 5 可知，本文方法的 FID 远小于 SWE 方法，表明本文方法图像生成质量更好。这主要是因为在构建的轮廓−图像可逆变换模型中，通过对生成模型添加约束条件控制图像的生成，另外，在目标函数中加入最小化 L1 距离

为训练目标，以完善生成图像的细节。而 SWE 方法是直接利用 DCGAN 将噪声转换成含密图像，不能控制图像的生成质量，此外，以逐像素的方式生成较大图像时容易产生棋盘效应，生成图像质量较差。图 7 展示了本文方法部分真实图像与生成图像的对比，仅凭肉眼观察，生成的含密图像与真实图像几乎没有视觉差异，生成图像质量较好。

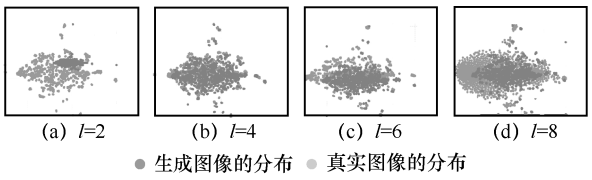


图 6　当 l 取不同值时 t-SNE 图示例

表 5　　不同隐藏比特数下生成质量对比

Bit	SWE 方法	本文方法
2×256	404.609 6	**77.449 2**
4×256	401.378 1	**84.775 5**
6×256	422.920 3	**89.736 3**
8×256	437.523 9	**91.979 6**

4.5　消融实验

为进一步验证本文方法各个部分的性能，本节设计了一系列消融实验，将本文方法与 4 种不同版本的本文方法的隐写效果进行对比。本文方法的轮廓生成模型采用 LSTM 结构，轮廓–图像可逆变换模型中的生成器、判别器和提取器分别采用 U-Net、PatchGAN 和 U-Net，方法 1 分别采用 RNN、U-Net、PatchGAN 和 U-Net，方法 2 分别采用 LSTM、Encoder-decoder、PatchGAN 和 U-Net，方法 3 分别采用 LSTM、U-Net、GAN、U-Net，方法 4 分别采用 LSTM、U-Net、PatchGAN 和 Canny 算子[29]。

表 6 展示了以上 4 种方法在秘密信息提取错误率 BER、抗隐写分析能力 P_E 以及图像生成质量 EMD 的表现。其中 P_E 针对 SRM 隐写分析器计算得出，所有数据是在 l=4 时，即嵌入比特数固定为 4×256 bit 的情况下测得的。由表 6 可得如下结论。

1) 本文方法的 EMD 值远低于方法 1 的 EMD 值，即本文方法的生成图像质量明显优于方法 1。这是因为经典的 RNN 结构没有设计门控机制，在生成轮廓点时对相邻轮廓点的依赖度过高，而对于距离较远的轮廓点，在训练阶段的反向传播过程中易产生梯度爆炸的问题，模型难以收敛，从而导致生成轮廓与真实轮廓差距较大。而具有门控机制的 LSTM 结构可以有效避免以上问题。因此，采用 LSTM 可以使最终生成的含密图像质量更高。

2) 本文方法的 BER 和 EMD 值比方法 2 明显要低，即本文方法的信息提取正确率和生成图像质量明显高于方法 2。这是因为本文方法生成器采用 U-Net，而方法 2 生成器采用 Encoder-decoder。相较于 Encoder-decoder，U-Net 在升采样与降采样的过程中能够更好地保留图像的抽象特征。这样使本文方法的含密图像的生成质量较高，轮廓线准确提取更加容易，秘密信息提取准确率更高。

3) 本文方法的 BER 和 EMD 值稍低于方法 3，即本文方法的信息提取准确率和生成图像质量稍高于方法 3。这是因为本文方法的判别器采用 PatchGAN，而方法 3 采用传统 GAN。采用 PatchGAN 的判别器将生成图像切分成许多图像块，并对每个图像块真实度进行判别。相较采用传统 GAN 的判别器，采用 PatchGAN 的判别器对图像真实度的判别结果更加准确，从而促进生成模型生成质量更高的图像，使本文方法的轮廓线提取和

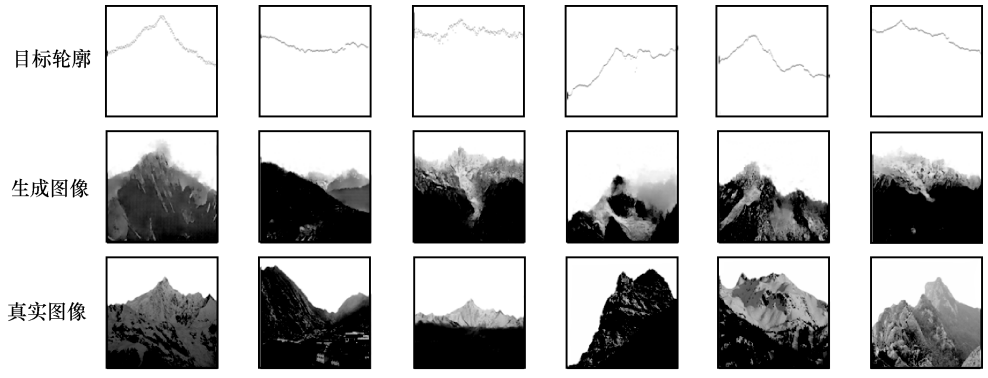


图 7　本文方法的目标轮廓、生成图像与真实图像实际效果

表 6　　本文方法与 4 种不同版本的本文方法的隐写效果对比

方法	生成模型		生成器		判别器		提取器		BER	P_E	EMD
	LSTM	RNN	U-Net	Encoder-decoder	PatchGAN	GAN	U-Net	Canny 算子			
方法 1	×	√	√	×	√	×	√	×	0.015 1	0.421 3	0.184 4
方法 2	√	×	×	√	√	×	√	×	0.048 2	0.488 1	0.049 7
方法 3	√	×	√	×	×	√	√	×	0.019 4	0.481 5	0.031 9
方法 4	√	×	√	×	√	×	×	√	0.093 2	0.489 6	0.027 7
本文方法	√	×	√	×	√	×	√	×	0.014 5	0.490 2	0.026 0

注：√表示采用此结构，×表示不采用此结构。

秘密信息提取更加准确。

4) 本文方法的 BER 值明显低于方法 4，EMD 值基本与方法 4 持平，即方法 4 的秘密信息提取错误率明显高于本文方法。这是因为方法 4 的提取器采用传统的 Canny 算子，而本文方法采用 U-Net。Canny 算子是手工设计的边缘检测方法，其参数也采用经验值，自适应性明显低于 U-Net。因此，Canny 算子提取山脉轮廓的准确度明显较低，导致秘密信息提取错误率也明显较高。

以上 4 种方法的 P_E 值与本文方法相差不大，均在 0.5 左右，即在抗隐写分析方面均表现较好。这是因为这些方法都是构造式的图像隐写方法，生成的含密图像中不含有修改痕迹，从而能较好地抵抗现有隐藏分析工具的检测。

根据以上实验结果和分析，本文方法中轮廓生成模型及轮廓−图像可逆变换模型的生成器、判别器和提取器均表现最好，本文方法在秘密信息提取准确率、抗隐写分析能力、图像生成质量方面综合性能最优。

5　结束语

本文提出了一种基于轮廓自动生成的构造式图像隐写方法。针对现有构造式信息隐写方法的隐藏容量低和秘密信息提取困难的问题，本文方法构建轮廓自动生成模型，将秘密信息映射为图像的显式特征即轮廓信息，然后构建轮廓−图像可逆变换模型，将轮廓信息作为该模型的输入，从而构造出相应的含密图像。实验结果表明，本文方法能实现较高的隐藏容量，并且在相同隐藏容量下信息提取率和生成图像质量优于现有构造式信息隐写方法，同时在抗隐写分析检测能力方面优于传统的修改式隐写方法。本文对构造式图像隐写方法进行探索，有效地推进了构造式图像隐写方法的实际应用。本文方法以秘密信息为驱动生成较简单轮廓线，从而构造出含密图像，下一步将以秘密信息为驱动构造出更复杂的多轮廓线以及生成多样性化的含密图像，以进一步提高隐写容量和生成图像质量。

参考文献：

[1] FILLER T, JUDAS J, FRIDRICH J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935.

[2] LIAO X, YU Y B, LI B, et al. A new payload partition strategy in color image steganography[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(3): 685-696.

[3] 付章杰, 王帆, 孙星明, 等. 基于深度学习的图像隐写方法研究[J]. 计算机学报, 2020, 43(9): 1656-1672.
FU Z J, WANG F, SUN X M, et al. Research on steganography of digital images based on deep learning[J]. Chinese Journal of Computers, 2020, 43(9): 1656-1672.

[4] FRIDRICH J, GOLJAN M, DU R. Detecting LSB steganography in color, and gray-scale images[J]. IEEE MultiMedia, 2001, 8(4): 22-28.

[5] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, 2014(1): 1-13.

[6] 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 (ICIP). Piscataway: IEEE Press, 2014: 4206-4210.

[7] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[C]//International Workshop on Information Hiding. Berlin: Springer, 2010: 161-177.

[8] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//2012 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE Press, 2012: 234-239.

[9] 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.

[10] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[11] 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.

[12] DUAN X T, JIA K, LI B X, et al. Reversible image steganography scheme based on a U-Net structure[J]. IEEE Access, 2019, 7: 9314-9323.

[13] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[14] 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.

[15] 张新鹏, 钱振兴, 李晟. 信息隐藏研究展望[J]. 应用科学学报, 2016, 34(5): 475-489.
ZHANG X P, QIAN Z X, LI S. Prospect of digital steganography research[J]. Journal of Applied Sciences, 2016, 34(5): 475-489.

[16] ZHOU Z L, SUN H Y, HARIT R, et al. Coverless image steganography without embedding[C]//International Conference on Cloud Computing and Security. Berlin: Springer, 2015: 123-132.

[17] 周志立, 曹燚, 孙星明. 基于图像 bag-of-words 模型的无载体信息隐藏[J]. 应用科学学报, 2016, 34(5): 527-536.
ZHOU Z L, CAO Y, SUN X M. Coverless information hiding based on bag-of-words model of image[J]. Journal of Applied Sciences, 2016, 34(5): 527-536.

[18] WU K C, WANG C M. Steganography using reversible texture synthesis[J]. IEEE Transactions on Image Processing, 2015, 24(1): 130-139.

[19] QIAN Z X, ZHOU H, ZHANG W M, et al. Robust steganography using texture synthesis[M]. Berlin: Springer, 2017.

[20] MEGÍAS D. Improved privacy-preserving P2P multimedia distribution based on recombined fingerprints[J]. IEEE Transactions on Dependable and Secure Computing, 2015, 12(2): 179-189.

[21] LI S, ZHANG X P. Toward construction-based data hiding: from secrets to fingerprint images[J]. IEEE Transactions on Image Processing, 2019, 28(3): 1482-1497.

[22] 刘明明, 张敏情, 刘佳, 等. 基于生成对抗网络的无载体信息隐藏[J]. 应用科学学报, 2018, 36(2): 371-382.
LIU M M, ZHANG M Q, LIU J, et al. Coverless information hiding based on generative adversarial networks[J]. Journal of Applied Sciences, 2018, 36(2): 371-382.

[23] HU D H, WANG L, JIANG W J, et al. A novel image steganography method via deep convolutional generative adversarial networks[J]. IEEE Access, 2018, 6: 38303-38314.

[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[25] YANG Z L, GUO X Q, CHEN Z M, et al. RNN-stega: linguistic steganography based on recurrent neural networks[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1280-1295.

[26] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv Preprint, arXiv:1411.1784, 2014.

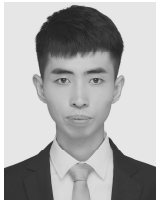
[27] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 5967-5976.

[28] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[29] CANNY J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8(6): 679-698.

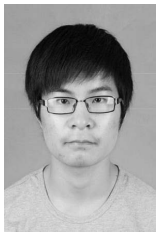
[作者简介]

周志立（1984− ），男，湖北黄冈人，博士，南京信息工程大学教授，主要研究方向为信息隐藏、数字取证、视觉密码、数字多媒体内容安全等。



王美民（1996− ），男，江苏盐城人，南京信息工程大学硕士生，主要研究方向为信息隐藏、数字取证。

杨高波（1974− ），男，湖南岳阳人，博士，湖南大学教授，主要研究方向为图像/视频信号处理、多媒体通信、数字媒体内容安全等。



朱剑宇（1996− ），男，江苏南通人，南京信息工程大学硕士生，主要研究方向为数字水印、图像处理。

孙星明（1963− ），男，湖南湘潭人，博士，南京信息工程大学教授，主要研究方向为网络与信息安全、传感器网络、自动气象观测等。