

NFT 图像隐写

王子驰^{1,2}, 冯国瑞¹, 张新鹏¹

(1. 上海大学通信与信息工程学院, 上海 200444;

2. 深圳大学深圳媒体信息安全重点实验室, 广东 深圳 518060)

摘要: NFT (non-fungible token) 图像为元宇宙中进行创作、交易、分享和收藏的数字艺术作品。不同于自然图像, NFT 图像的内容为用户自定义, 在数据空间分布较广, 这为秘密信息的隐藏提供了极大便利, 因此借助 NFT 图像进行隐蔽通信成为图像隐写的一个新分支。提出了一种 NFT 图像的隐写方法。对一幅给定的 NFT 图像, 对高频与边缘轮廓部分进行增强, 以丰富图像中有利于掩盖隐写修改痕迹的细节部分, 从而使增强后的图像更适合隐写, 将其作为载体。根据增强前后图像像素的差异确定载体图像各像素加 1 或减 1 的倾向修改方向, 并根据此差异调整载体各像素的修改代价以满足确定的倾向修改方向, 进一步提升隐写抗检测性。利用主流隐写编码框架在载体图像中进行信息嵌入。实验结果表明, 所提方法应用于 NFT 图像时的抗检测性优于现有的数字图像隐写方案, 对于 HILL、MiPOD、DFEI 隐写方案, 所提方法可分别将隐写分析错误率 (P_E 值) 平均提升 8.7%、9.2%、6.2% (所有嵌入率与隐写分析特征情况平均值)。所提方法适用于 NFT 图像, 为除自然图像与生成图像以外的第 3 类载体 (即 NFT 图像) 提供了针对性的隐写方法。待 NFT 图像数量较为可观后, 可充分利用神经网络强大的拟合与学习能力, 设计深度学习的 NFT 图像隐写方法。

关键词: 元宇宙; NFT 图像; 隐写; 多样性

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022029

Steganography in NFT images

WANG Zichi^{1,2}, FENG Guorui¹, ZHANG Xinpeng¹

1. School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

2. Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China

Abstract: The images with non-fungible token (NFT) are employed as the digital artistic works in metaverse for creation, transaction, sharing, and collection. Being different from natural images, the content of NFT images is

收稿日期: 2022-03-04; 修回日期: 2022-05-06

通信作者: 张新鹏, xzhang@shu.edu.cn

基金项目: 国家自然科学基金 (62002214, 62072295, U1936214); 深圳媒体信息安全重点实验室开放基金 (ML-2022-01)

Foundation Items: The National Natural Science Foundation of China (62002214, 62072295, U1936214), Research Fund of Shenzhen Key Laboratory of Media Security (ML-2022-01)

引用格式: 王子驰, 冯国瑞, 张新鹏. NFT 图像隐写[J]. 网络与信息安全学报, 2022, 8(3): 18-28.

Citation Format: WANG Z C, FENG G R, ZHANG X P. Steganography in NFT images[J]. Chinese Journal of Network and Information Security, 2022, 8(3): 18-28.

defined by user and distributed in the digital space widely. It is convenient for the hidden of secret data. In this case, covert communication with NFT images is a new branch of image steganography. Then, a steganographic method for NFT images was proposed accordingly. Given a NFT image, the regions of its profile and the components with high frequency were enhanced firstly to enrich the details which were beneficial to hide the modification trace of steganography. In this way, the enhanced image was used as cover since it is more suitable for steganography. Then, the tendency modification direction of each pixel was determined by the differences between the enhanced image and the given image. The differences were also used to determine the cost value of modification amplitude. Thus, the undetectability of steganography can be increased further. Secret data was embedded into the cover image using the popular steganographic coding schemes. Experimental results showed that the proposed method had improved undetectability on NFT images compared with existing digital steganographic schemes. Compared with HILL, MiPOD, and DEFI, the proposed method can increase the detection error P_E of steganalysis by 8.7%, 9.2% and 6.2%, respectively (the average value for the cases of different payload and steganalytic features). Therefore, the proposed method is suitable for NFT images and it provides targeted steganographic method for the third kind of images, i.e., NFT images, except of natural images and generated images. For further study, the deep learning-based steganographic method can be designed for NFT images using the strong fitting and learning ability of neural networks.

Keywords: Metaverse, NFT images, steganography, diversity

0 引言

图像隐写旨在将秘密信息嵌入数字图像中以实现隐蔽通信。隐写分析则根据隐写修改引起的载体数据统计异常来判断可疑图像中是否含有秘密信息^[1]。近年来二者均得到快速发展。图像隐写采用的载体类型主要包括两种: 自然图像(对真实场景拍摄的成像)^[2]和生成图像(利用深度学习技术生成)^[3]。其中, 自然图像为真实世界的客观描述, 而生成图像则是可能场景的合理呈现。

借助元宇宙进行隐蔽通信的研究是图像隐写的一个重要分支。非同质化代币(NFT, non-fungible token)是元宇宙中用于表示数字资产的唯一加密“货币”令牌, 可用于买卖^[4]。在 NFT 中国(NFTCN)平台^[5]上, 用户可自由创作、交易、分享和收藏 NFT 数字艺术作品(目前均为数字图像)。NFTCN 平台上的数字艺术作品被称为 NFT 图像。NFT 图像的内容既非真实世界的客观描述, 也非可能场景的合理呈现, 而是虚拟空间的自定义。本文旨在利用 NFT 图像进行隐写。

NFTCN 平台与推特、微博、微信等社交网络不同, 该平台不会对上传的图像进行预处理, 即下载的图像与上传版本完全一致。因此, NFTCN 平台是理想的隐写信道, 利用其进行隐写时无须

考虑信息嵌入的鲁棒性。所有隐写方案均可保证将含密图像上传至 NFTCN 平台后, 在下载的图片中正确提取秘密信息。此外, NFT 图像的多样性为隐写提供了天然屏障。NFT 图像为数字艺术作品, 由用户自行创作, 其内容多种多样, 在数据空间分布较广。此种情形下几乎所有包含语义的图像内容可视为“正常”, 即正常 NFT 图像的范围覆盖较广, 这为秘密信息的隐藏提供了极大便利。为了保证原创, 上传的 NFT 图像不能与其他作品相同或高度相似, 从而保证了 NFT 图像的多样性。因此, 在 NFT 图像上进行隐写有较大的可行性与应用价值, 但目前还未有以 NFT 图像为载体的隐写方法。

本文提出了一种用于 NFT 图像的隐写方法。由于图像纹理细节有利于掩盖隐写的修改痕迹, 首先对 NFT 图像的高频与边缘轮廓部分进行增强, 使增强后的图像更适合隐写, 将增强后的图像作为隐写的载体图像。之后将增强前图像看作增强后图像的相似版本, 以增强前图像为参考, 确定载体图像每个像素加 1 或减 1 的倾向修改方向, 据此对隐写失真代价函数进行优化, 调整载体各像素的修改代价以满足确定的倾向修改方向, 进一步提升隐写抗检测性。最后利用主流隐写编码框架在载体图像中进行信息嵌入。

有研究者通过增强载体图像提高隐写抗检测性。文献[6]利用 Gamma 变换对载体图像进行增强。与之不同, 本文对高频与边缘轮廓增强。除载体增强外, 本文还提出了针对 NFT 图像的代价优化方法。以增强图像与原始图像的差异为参考, 调整载体像素代价值, 进一步提高了 NFT 图像隐写的抗检测性。本文的主要贡献如下。

1) 提出了针对 NFT 图像的载体增强方法。通过对图像高频与边缘轮廓进行增强, 增加了像素值的随机性。上述操作使隐写的修改痕迹被掩盖在像素值的随机性中, 从而使隐写分析工具难以区分隐写修改痕迹与增强引起的随机性。

2) 提出了针对 NFT 图像的代价优化方法。以增强图像与原始图像的差异为参考, 确定载体图像各像素加 1 或减 1 的倾向修改方向, 据此调整载体像素代价值, 进一步提高了 NFT 图像隐写的抗检测性。

1 相关工作

NFT 图像隐写的相关工作包括图像隐写与图像隐写分析。

1.1 图像隐写

目前最有效的数字图像隐写框架通过最小化含密图像在预定义失真代价函数下的加性失真来保证隐写安全^[7], 其中失真代价函数为图像各像素分配一个代价值以衡量修改风险。之后通过最优隐写编码如 STC (syndrome trellis coding)^[8]或 SPC (steganographic polar code)^[9]来实现最小化加性失真嵌入。有许多适用于数字图像的失真代价函数, WOW (wavelet obtained weights)^[10]将图像各方向定向残差加和的倒数作为像素的失真代价值; UNIWARD (universal wavelet relative distortion)^[11]将一组小波方向滤波器在像素或 DCT 系数变化前后的改变量作为失真代价; HILL (high-pass, low-pass, and low pass)^[12]采用聚集修改位置策略来提高抗检测性; MiPOD (minimizing the power of optimal detector)^[13]根据似然比检验求解最佳检测器, 通过求解非线性代数方程得到像素的修改概率, 并反推各像素的失真代价。也有运用深度学习技术设计的隐写失真函数, UT-GAN (U-Net ternary embedding with generat-

ive adversarial network)^[14]利用 U-Net 网络^[15]将载体图像转换为像素修改概率图。

有研究者对已有隐写失真函数进行优化。文献[16]通过多次嵌入, 根据对抗性样本的反向梯度方向决定代价值的倾向修改方向。Denemark 等^[17]将同一场景中的多幅图像作为边信息来优化 JPEG 失真函数。Chen 等^[18]通过放大局部纹理, 在保持图像原有特性的同时突出纹理区域。Zhang 等^[19]通过将像素的联合失真拆分为单个失真来考虑像素修改后对其邻域像素的影响。Zhou 等^[20]将多种性能相近的失真函数赋予的代价值差别较大的像素称作“争议像素”, 通过对争议像素赋予较高的修改概率来融合多种失真函数。文献[21]则利用维纳滤波构造参考图像, 并据此确定代价值的倾向修改方向。Li 等^[22]改进参考图像的构造方法, 采用去块效应滤波器构造参考图像。除自然图像外, 也有针对非自然图像的失真函数, 如纹理合成图像^[23]、虹膜图像^[24]、表情图像^[25]、增强图像^[26]等。本文将针对 NFT 图像提出新的失真代价函数优化策略。

1.2 图像隐写分析

主流的图像隐写分析采用机器学习方法检测秘密信息的存在性, 主要分为人工特征隐写分析与深度学习隐写分析。

人工特征隐写分析主要研究隐写操作引起的图像统计特性变化, 首先提取图像特征, 之后训练分类器进行检测。主要捕捉隐写对载体像素修改造成的图像统计特性变化, 需要抑制图像内容、放大隐写噪声。目前已有许多隐写分析特征提取方法, 如 SPAM (subtractive pixel adjacency matrix)^[27]、SRM (spatial rich model)^[28]、PSRM (projections of spatial rich model)^[29]、maxSRMd2 (selection-channel-aware rich model)^[30]、SCRMQ1 (spatial color rich model)^[31]、TLBP (threshold local binary pattern)^[32]等。大多数隐写分析特征为高维特征, 主要采用集成分类器^[33]进行训练。集成分类器由若干个子分类器组成, 其中子分类器采用 Fisher 线性分类判决。各子分类器从高维空间中以一定步长随机抽取若干维特征进行训练, 取训练误差最小的子分类器。实际检测时各子分类器投票给出最终决策。

深度学习方法的出现在很大程度上改变了隐写分析研究方向。文献[34]用多个高通滤波器获取图像残差,将残差送入 CNN 进行训练得到 TLU-CNN。Li 等^[35]采用不同的高通滤波器分别训练 3 个子网络,再通过全连接层整合。文献[36]提出了一种深度残差网络结构来避免启发式地设计网络以提高隐写分析网络的开发效率。文献[37]考虑利用已有网络在不重新训练其参数的情况下对不同大小的图像进行隐写分析。文献[38]利用联邦学习技术实现隐写分析,多个参与者共同训练一个隐写分析网络,彼此仅共享网络模型信息,不共享训练参数。

本文采用上述的部分隐写分析方法检验隐写抗检测性,并与现有隐写方案进行对比。

2 NFT 图像隐写方法

本文提出了一种 NFT 图像的隐写方法,其框架如图 1 所示,对一幅给定的原始 NFT 图像 X_0 ,首先将其高频与边缘轮廓部分进行增强,以丰富图像中有利于掩盖隐写修改痕迹的细节部分,从而使增强后的图像 X_c 更适合隐写。将 X_c 作为载体,利用现有的失真代价函数为 X_c 的各像素分配失真代价值 ρ_u 。之后对 ρ_u 进行优化,根据 X_0 与 X_c 像素的差异确定 X_c 各像素加 1 或减 1 的倾向修改方向,并根据此差异调整 X_c 各像素的修改代价以满足确定的倾向修改方向,得到可进一步提升隐写抗检测性的失真代价值 ρ_v 。最后利用主流隐写编码框架在 X_c 中嵌入秘密信息 m ,得到含密图像 X_s 。下面分别对本文所提方法的 4 个主要过程进行介绍。

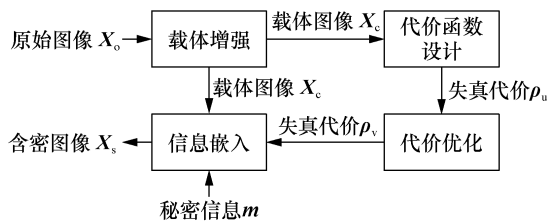


图 1 本文所提方法框架

Figure 1 Architecture of the proposed method

2.1 载体增强

由于图像纹理细节有利于掩盖隐写的修改痕迹,本文首先对原始 NFT 图像 X_0 的高频与边缘

轮廓部分进行增强,使增强后的图像 X_c 更适合隐写,将 X_c 作为隐写的载体图像。

首先在 X_0 中添加轻微的高斯噪声 $z \sim N(0, 0.000\ 1)$,以在不影响 NFT 图像正常的创作、交易、分享和收藏等功能的条件下丰富图像的高频成分,得到图像 X_t 。可证明在图像中添加参数分别为 μ 与 σ^2 ($\sigma > 0$) 的高斯噪声 $N(\mu, \sigma^2)$ 后,图像的峰值信噪比 (PSNR, peak signal to noise ratio) 为

$$\text{PSNR} = 10 \lg \frac{\sqrt{2\pi}}{\sqrt{2\pi\sigma^2 + 2\mu\sigma + \sqrt{2\pi}\mu^2}} \quad (1)$$

证明过程见附录。加入噪声 $z \sim N(0, 0.000\ 1)$ 后,所得图像 X_t 相对于图像 X_0 的峰值信噪比为 40 dB,即图像保持了较高质量。因此添加轻微高斯噪声的操作不影响图像内容及其在元宇宙中的正常使用。由于 NFT 图像不同于自然图像,其内容为用户自定义,因此只要不影响正常使用,添加噪声等操作在 NFT 图像中可被允许。

之后对图像 X_t 的边缘轮廓部分进行增强,进一步丰富有利于掩盖隐写修改痕迹的细节部分。采用 Sobel 算子计算图像像素值的梯度,将梯度方向变化较大的区域作为边缘,即图像 X_t 第 (i, j) 个像素处的梯度值 $g(i, j)$ 为

$$g(i, j) = \sqrt{g_h(i, j)^2 + g_v(i, j)^2} \quad (2)$$

其中, $g_h(i, j)$ 与 $g_v(i, j)$ 分别为水平与垂直方向的梯度值,分别由式(3)与式(4)计算得到, G_h 与 G_v 分别为 $g_h(i, j)$ 与 $g_v(i, j)$ 对应的梯度矩阵。

$$G_h = X_t \otimes F_h \quad (3)$$

$$G_v = X_t \otimes F_v \quad (4)$$

算子 F_h 与 F_v 分别为

$$F_h = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (5)$$

$$F_v = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (6)$$

由于隐写算法的修改幅度通常为 ± 1 且集中于纹理较为复杂的边缘轮廓区域,本文对图像 X_t 中定位好的边缘轮廓部分进行随机的加减 1 操作,最终得到增强后的载体图像 X_c 。即隐写信息

嵌入的对象为增强后的载体图像，而不是原始的 NFT 图像。图像高频与边缘轮廓部分由于随机性强、难以预测，历来为隐写与隐写分析双方关注的重点。尽管部分隐写分析特征是从纹理区域来构造特征模型的，增强载体图像边缘轮廓仍有利于提高隐写抗检测性，这是由于边缘轮廓增强后，其像素值的随机性增强。上述过程使隐写过程中对边缘轮廓像素的修改痕迹被掩盖在边缘轮廓的随机性中。隐写分析工具难以区分隐写修改痕迹与增强引起的随机性，即增强载体图像边缘轮廓可使隐写分析难以从纹理区域构造准确的特征模型。

2.2 代价函数设计

本文将增强后的图像 $X_c = [x_c(1), x_c(2), \dots, x_c(n)]^T \in \{0, 1, \dots, 255\}^{n \times 1}$ 作为隐写载体，采用 STC 隐写编码框架嵌入秘密信息 $m = [m(1), m(2), \dots, m(k)]^T \in \{0, 1\}^{k \times 1}$ 。由最小化加性失真框架^[8-9]可知，在含有 n 个像素的 X_c 中嵌入 k 比特信息的理论最小失真 D 为

$$D = \sum_{i=1}^n [p^+(i)\rho^+(i) + p^-(i)\rho^-(i)] \quad (7)$$

其中， $\rho^+(i)$ 与 $\rho^-(i)$ 分别为对 $x_c(i)$ 进行加 1 与减 1 修改的代价。而 $p^+(i)$ 与 $p^-(i)$ 则分别为对 $x_c(i)$ 进行加 1 与减 1 修改的概率，如式(8)与式(9)所示。

$$p^+(i) = \frac{e^{-\lambda\rho^+(i)}}{1 + e^{-\lambda\rho^+(i)} + e^{-\lambda\rho^-(i)}} \quad (8)$$

$$p^-(i) = \frac{e^{-\lambda\rho^-(i)}}{1 + e^{-\lambda\rho^+(i)} + e^{-\lambda\rho^-(i)}} \quad (9)$$

参数 λ 用于使修改概率的信息熵等于需要嵌入的信息量 k ，如式(10)所示。

$$-\sum_{i=1}^n \{p^+(i)\lg p^+(i) + p^-(i)\lg p^-(i) + [1 - p^+(i) - p^-(i)]\lg[1 - p^+(i) - p^-(i)]\} = k \quad (10)$$

即在最小化加性失真框架下需要设计失真代价函数来为载体各像素赋予修改代价。目前已有许多适用于数字图像的失真代价函数，本文利用现有隐写方法得到各像素的代价值 ρ_u ，之后对 ρ_u 进行优化。

2.3 代价优化

得到初始代价值 $\rho_u = [\rho_u^+(1), \rho_u^+(2), \dots, \rho_u^+(n), \rho_u^-(1), \rho_u^-(2), \dots, \rho_u^-(n)]^T$ 后，本文根据 X_o 与 X_c

像素的差异确定 X_c 各像素加 1 或减 1 的倾向修改方向，并根据此差异调整 X_c 各像素的修改代价以满足确定的倾向修改方向，得到可进一步提升隐写抗检测性的代价值 $\rho_v = [\rho_v^+(1), \rho_v^+(2), \dots, \rho_v^+(n), \rho_v^-(1), \rho_v^-(2), \dots, \rho_v^-(n)]^T$ 。

设图像 X_o 与 X_c 像素的差异为 $\Delta = [\delta(1), \delta(2), \dots, \delta(n)]^T$ ，即 $\delta(i) = x_c(i) - x_o(i)$ 。首先利用改进的 Sigmoid($\cdot$)函数 $\Psi(\cdot)$ 调整像素差异取值，如式(11)所示。

$$\Psi(x) = \frac{1}{1 + e^{-\alpha x}} - 0.5 \quad (11)$$

其中，参数 $\alpha > 0$ 用于调节函数值域的变化速度。式(11)用于将像素残差的取值限定在 -0.5 到 0.5 之间，并将调整后残差的绝对值作为调整隐写代价值的幅度。原始 Sigmoid($\cdot$)函数值域为 0~1，并关于 (0, 0.5) 中心对称。当残差取值为 0 时，表示载体图像与原始图像像素无差异，此时不应修改隐写代价值。将 Sigmoid($\cdot$)函数值减去 0.5，使值域关于 (0, 0) 中心对称，以满足隐写代价优化所需，则初始代价值 ρ_u 可由式(12)~式(15)进行优化。

$$\rho_v^+(i) = \beta(i)^{[1-\varepsilon(i)]/2} \cdot \rho_u^+(i) \quad (12)$$

$$\rho_v^-(i) = \beta(i)^{[1+\varepsilon(i)]/2} \cdot \rho_u^-(i) \quad (13)$$

$$\beta(i) = 1 - \Psi[|\delta(i)|] \quad (14)$$

$$\varepsilon(i) = \text{sign}[\delta(i)] \quad (15)$$

即初始状态下 $x_c(i)$ 的加 1 与减 1 的代价值 $\rho_u^+(i)$ 与 $\rho_u^-(i)$ 分别等于 $\rho_u^+(i)$ 与 $\rho_u^-(i)$ ，若 $x_c(i) > x_o(i)$ ，减小 $\rho_u^+(i)$ 使 $x_c(i)$ 进行加 1 修改的概率大于减 1 修改，从而使修改后的 $x_c(i)$ 远离 $x_o(i)$ 。这是由于增强后的图像 X_c 比原始图像 X_o 更适合隐写，因此当隐写过程中需要修改图像像素值时，应朝着远离 X_o 的方向修改。反之，若 $x_c(i) < x_o(i)$ ，则通过减小 $\rho_u^-(i)$ 使 $x_c(i)$ 进行减 1 修改的概率大于加 1 修改，同样使修改后的 $x_c(i)$ 远离 $x_o(i)$ 。此外若 $x_c(i) = x_o(i)$ ，则不对代价值进行调整。

2.4 信息嵌入

最小化加性失真理论给出了固定嵌入量下的理论最小失真界限，STC 则提供了逼近上述理论界限的实际编码方法。信息嵌入过程如式(16)所示。其中 $D(X_c, X_s)$ 为含密图像 X_s 相对于载体图像 X_c 的 Y 失真，失真度量方式由失真函数决定。即在 m 的陪集 $C(m)$ 中寻找与 X_c 距离最近的码字作

为含密对象,即寻找加性失真最小的嵌入方案。

$$\text{Emb}(X_c, m) = \arg \min_{X_s \in C(m)} D(X_c, X_s) \quad (16)$$

陪集 C 定义如式(17)所示, $H \in \{0,1\}^{k \times n}$ 为校验矩阵。STC 将秘密信息看作线性码的伴随式,而线性码由 H 确定。

$$C(m) = \{r \in \{0,1\}^{n \times 1} \mid Hr = m\} \quad (17)$$

陪集中所有码字 r 均满足 $Hr = m$, 即陪集中的所有码字左乘校验矩阵均可得到秘密信息。因此提取过程很简单,如式(18)所示,将含密图像 X_s 左乘 C 的校验矩阵 H 即可得到秘密信息。

$$\text{Ext}(X_s) = HX_s \quad (18)$$

上述过程中的载体增强与代价优化方法均针对 NFT 图像设计,不适用于一般图像。对于载体增强,本文通过添加轻微高斯噪声以及随机修改图像边缘部分,对图像高频与边缘轮廓进行增强。此操作不属于正常的图像处理,在自然图像中属于异常行为,可被隐写分析工具判为含密。而 NFT 图像内容本身就具有自定义性,几乎所有包含语义的图像内容可视为“正常”,只要不影响使用,添加噪声等操作在 NFT 图像中可被允许,其不能成为秘密信息存在的证据。因此,本文提出的载体增强方法为针对 NFT 图像设计。对于代价优化,本文以增强图像与原始图像的差异为参考,确定载体图像各像素加 1 或减 1 的倾向修改方向,据此调整载体像素代价值。一般图像隐写虽也有代价优化的方法,但其思想为使载体图像像素值朝着靠近参考图像像素值的方向调整像素代价值。而本文朝着远离参考图像像素值的方向调整载体像素代价值。这是由于增强后的载体图像比参考图像更适合隐写。因此,本文提出的代价优化方法也是针对 NFT 图像设计的。

3 实验结果

本节首先介绍实验环境,之后确定方法参数取值,再进行消融实验,最后用隐写分析工具验证本方法的有效性。

3.1 实验环境

(1) 图像库

由于本文方法是针对 NFT 图像设计的,因此未采用隐写领域常用的自然图像库。实验采用

NFTCN 平台^[5]下载的 1 000 幅不同尺寸的彩色 NFT 图像(大多数尺寸为 400×400,此外主要包括 400×500、400×533,以及 400×560 等),作为原始图像 X_0 ;采用 2.1 节所述方法进行增强后得到对应的 1 000 幅载体图像即 X_c 。本文方法隐写信息嵌入与隐写分析的对象均为增强后的载体图像,而不是 NFTCN 平台直接下载的图像。

(2) 隐写方法

用于获取初始代价 ρ_u 的隐写方法为 HILL、MiPOD,以及针对视觉增强图像设计的隐写失真函数 DFEI^[26],所有嵌入过程均采用 STC 编码的模拟嵌入器^[39]完成。嵌入率分别设定为 0.1 bit/pixel、0.2 bit/pixel、0.3 bit/pixel、0.4 bit/pixel、0.5 bit/pixel,嵌入完成后即可得到含密图像 X_s 。NFTCN 平台上的图像数量还较少,未搜集到大量的 NFT 图像,因此本实验未采用基于深度学习的 UT-GAN 等隐写方法。

(3) 隐写分析方法

用于检验隐写抗检测性的隐写分析特征采用 SPAM^[27]、maxSRMd2^[30]、SCRMQ1^[31]、TLBP^[32]。并采用集成分类器训练及测试特征。即采用上述隐写分析方法分别提取 1 000 幅增强的载体图像 X_c 以及对应的含密图像 X_s 的特征。其中,将特征集的一半作为训练集,另一半作为测试集。评价分类性能的指标为测试误差 P_E ,如式(19)所示。

$$P_E = \min_{P_{FA}} \left(\frac{P_{FA} + P_{MD}}{2} \right) \quad (19)$$

其中, P_{FA} 与 P_{MD} 分别为虚警率(将正常图像 X_c 判断为含密图像 X_s 的概率)与漏检率(将含密图像 X_s 判断为正常图像 X_c 的概率)。 P_E 越小表示隐写分析性能越好,隐写抗检测性越差。通常 $0 \leq P_E \leq 0.5$, $P_E = 0.5$ 代表随机猜测,即分类器完全无法区分正常图像特征与含密图像特征。为了消除随机性,实验所有的测试误差数据均为独立测试 10 次的平均值。

3.2 参数寻优

式(11)中参数 α 用于调节函数值域的变化速度,对代价优化的效果起到了重要作用,本节讨论其取值。将本文方法用于 1 000 幅增强后的载体图像 X_c 中嵌入秘密信息。嵌入率为 0.3 bit/pixel 与 0.5 bit/pixel 的条件下,不同 α 值下的测试误差 P_E 如图 2 所示。

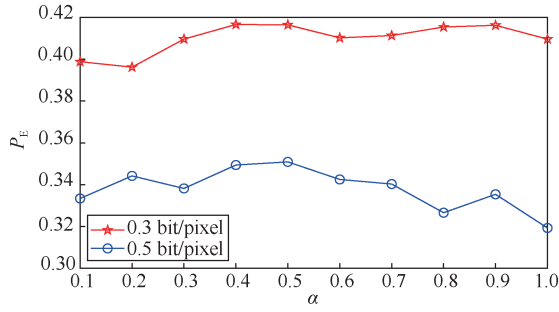


图2 不同 α 值下的测试误差 P_E
Figure 2 Values of P_E with different α

参数 α 取值越大表示对初始代价的调整幅度越大, 调整幅度过大反而会降低隐写抗检测性。由图 3 可看出嵌入率 0.3 bit/pixel 与 0.5 bit/pixel 的情况下, 测试误差 P_E 均在 $\alpha = 0.5$ 时达到最大, 即隐写抗检测性最好。因此参数 α 的取值定为 0.5。

3.3 消融实验

本文方法在现有最小化加性失真隐写框架的基础上加入了载体增强与代价优化两个环节。为验证二者均能提高隐写抗检测性, 本文首先采用 HILL 在 1 000 幅原始图像 X_0 中嵌入信息, 记为“HILL-o”;

之后采用 HILL 在 1 000 幅增强后的载体图像 X_c 中嵌入信息, 记为“HILL-c”, 即仅采用本文方法中的载体增强策略; 最后采用经过代价优化的 HILL 在 1 000 幅增强载体 X_c 中嵌入信息, 记为“HILL-p”, 即同时采用本文方法中的载体增强与代价优化策略。三者在隐写分析特征 SPAM 与 maxSRMd2 检测下的抗检测性对比如图 3 所示。

由图 3 可看出 HILL-c 与 HILL-p 方案对应的 P_E 值均高于 HILL-o 方案, 大多数情况下 HILL-p 方案对应的 P_E 值高于 HILL-c 方案。即在现有隐写方法 HILL 的基础上, 本文方法中的载体增强与代价优化策略均能进一步提升隐写抗检测性。因此相比于直接采用现有隐写方法在原始 NFT 图像中嵌入信息, 本文方法更适用于多样化的 NFT 图像。本文方法为除自然图像与生成图像以外的第 3 类载体 (即 NFT 图像) 提供了针对性的隐写方法。

3.4 抗检测性对比

本文方法的两个策略均提升了隐写抗检测性。图 4 与图 5 展示了本文方法与现有隐写方

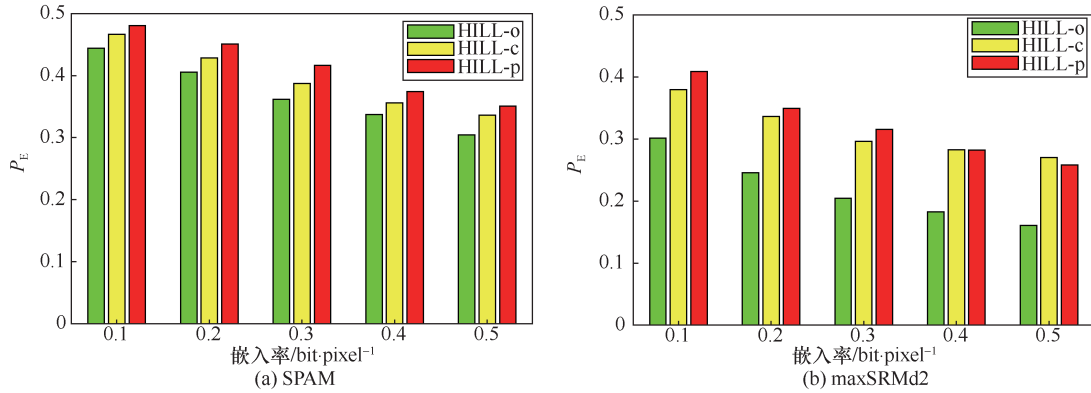


图3 不同嵌入策略抗检测性对比
Figure 3 Undetectability comparison of different embedding strategies against

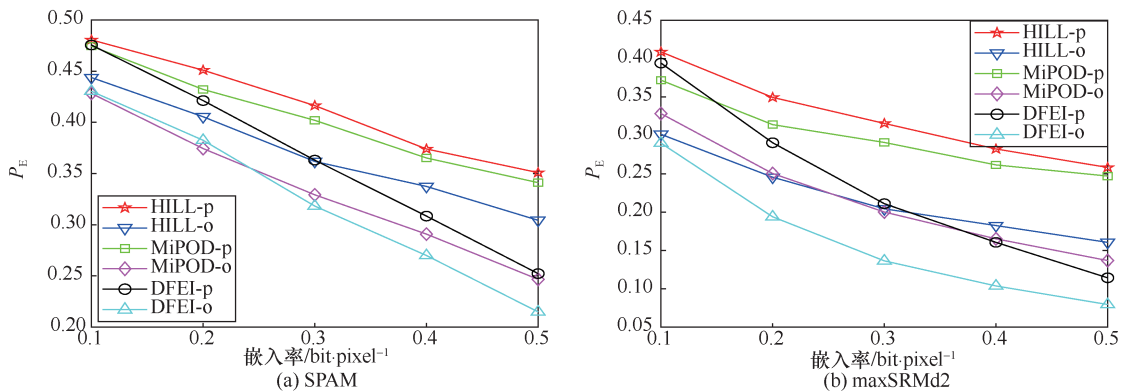


图4 本文方法与现有隐写方案抗检测性对比 1
Figure 4 Undetectability comparison between proposed method and existing steganographic schemes against 1

案的抗检测性对比, 对应的具体数值如表 1 所示。MiPOD-o 与 DFEI-o 表示分别采用 MiPOD 与 DFEI 在 1 000 幅原始图像 X_o 中嵌入信息, MiPOD-p 与 DFEI-p 则分别表示采用经过代价优化的 MiPOD 与 DFEI 在 1 000 幅增强后的载体图像 X_c 中嵌入信息。

由图 4 与图 5 可看出, 所有情况下本文方法对应的 P_E 值均明显高于直接在 NFT 图像上使用 HILL、MiPOD、DFEI 隐写。例如, 本文方法优化后的 HILL, 对嵌入率为 0.3 bit/pixel 的情况, 在 SPAM 与 maxWSRMd2 特征下 P_E 值分别提高了 5.5%与 11.1%; 在 TLBP 与 SCRMQ1 特征下,

分别提高了 9.0%与 14.0%。对 MiPOD 与 0.3 bit/pixel 情况, 则在 4 个特征下分别提高了 7.2%、9.1%、13.7%, 以及 19.4%。而对于针对增强图像设计的隐写方法 DFEI, 嵌入率为 0.3 bit/pixel 的情况在 4 个特征下分别提高了 4.5%、7.5%、2.8%, 以及 12.3%。此外, 可看出抗检测性的提升量与嵌入率呈正相关趋势。这是因为嵌入率低时 P_E 值距离理论界限 50% 较近, 提升空间小。

整体上看, 对于 HILL、MiPOD、DFEI 隐写方案, 本文方法可分别将 P_E 值平均提升 8.7%、9.2%, 以及 6.2% (所有嵌入率与隐写分析特征情况平均值)。因此本文方法应用于 NFT 图像时的

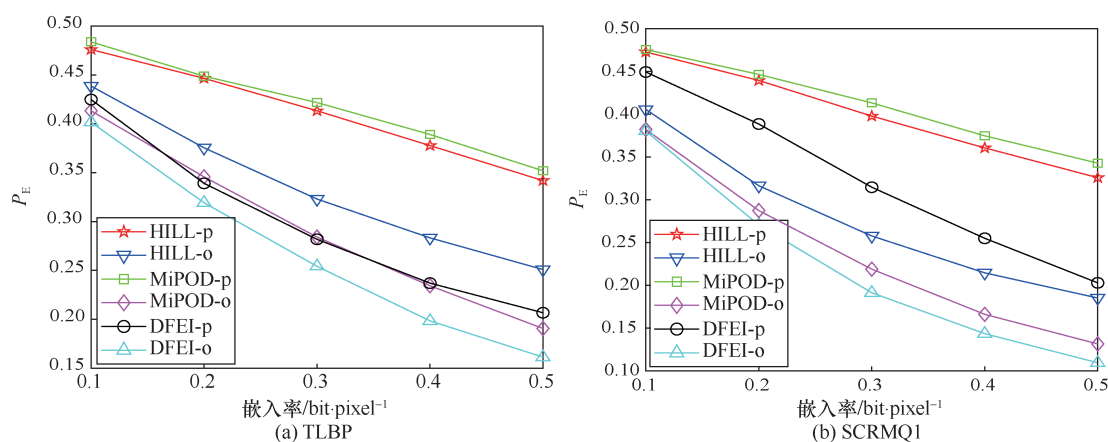


图5 本文方法与现有隐写方案抗检测性对比 2

Figure 5 Undetectability comparison between proposed method and existing steganographic schemes against 2

表 1 隐写方法抗检测性对比数值 (P_E)

Table 1 Undetectability comparison between steganographic schemes using P_E

嵌入率/ bit-pixel ⁻¹	SPAM						maxSRMd2					
	HILL-o	HILL-p	MiPOD-o	MiPOD-p	DFEI-o	DFEI-p	HILL-o	HILL-p	MiPOD-o	MiPOD-p	DFEI-o	DFEI-p
0.1	0.443 8	0.480 6	0.428 3	0.474 5	0.430 4	0.475 6	0.301 4	0.408 8	0.328 4	0.371 9	0.289 9	0.394 2
0.2	0.405 5	0.451 0	0.374 4	0.432 2	0.382 5	0.421 3	0.245 5	0.349 5	0.250 3	0.314 0	0.193 9	0.290 4
0.3	0.361 8	0.416 4	0.329 2	0.402 0	0.318 2	0.363 1	0.204 3	0.315 6	0.200 1	0.290 8	0.136 5	0.211 0
0.4	0.337 4	0.374 0	0.290 8	0.365 2	0.269 9	0.308 4	0.182 5	0.282 3	0.165 2	0.261 5	0.103 8	0.160 6
0.5	0.304 3	0.350 9	0.246 6	0.341 1	0.214 8	0.252 2	0.160 6	0.258 2	0.136 9	0.247 0	0.079 6	0.114 4

嵌入率/ bit-pixel ⁻¹	TLBP						SCRMQ1					
	HILL-o	HILL-p	MiPOD-o	MiPOD-p	DFEI-o	DFEI-p	HILL-o	HILL-p	MiPOD-o	MiPOD-p	DFEI-o	DFEI-p
0.1	0.438 6	0.476 0	0.413 3	0.483 8	0.401 6	0.424 8	0.405 6	0.472 7	0.382 1	0.475 6	0.380 2	0.449 3
0.2	0.375 4	0.446 5	0.345 7	0.448 7	0.319 2	0.339 1	0.316 3	0.439 4	0.287 3	0.446 5	0.270 4	0.388 5
0.3	0.323 0	0.413 3	0.284 1	0.421 7	0.254 2	0.281 8	0.257 6	0.397 8	0.218 8	0.413 2	0.191 2	0.314 6
0.4	0.283 3	0.377 7	0.234 3	0.389 1	0.198 4	0.237 0	0.214 5	0.360 6	0.166 0	0.374 7	0.143 4	0.254 9
0.5	0.250 9	0.341 9	0.190 6	0.351 9	0.161 3	0.206 7	0.185 1	0.325 8	0.131 4	0.342 8	0.109 4	0.202 8

抗检测性优于现有的数字图像隐写方案, 即本文方法适用于 NFT 图像。

4 结束语

针对 NFT 图像, 本文基于最小化加性失真框架提出了一种新的隐写方法。首先将原始 NFT 图像的高频与边缘轮廓部分进行增强, 以得到更适合隐写的载体。增强操作不影响图像内容及其在元宇宙中的创作、交易、分享和收藏等正常使用。之后根据增强前后图像像素的差异对隐写失真代价函数进行优化。实验结果表明本文提出的载体增强与代价优化策略均能提升隐写抗检测性。应用于 NFT 图像时, 本文方法的抗检测性优于直接使用现有的数字图像隐写方案。未来待 NFT 图像数量较为可观后, 可充分利用神经网络强大的拟合与学习能力, 设计深度学习的 NFT 图像隐写方法。

附录

定理 在图像中添加参数分别为 μ 与 σ^2 ($\sigma > 0$) 的高斯噪声 $N(\mu, \sigma^2)$ 后, 图像的峰值信噪比为

$$\text{PSNR} = 10 \lg \frac{\sqrt{2\pi}}{\sqrt{2\pi\sigma^2 + 2\mu\sigma + \sqrt{2\pi}\mu^2}}$$

证明 已知高斯噪声 $N(\mu, \sigma^2)$ 的概率密度函数 $f(x)$ 为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (20)$$

则加入该噪声后, 图像的均方误差 (MSE, mean squared error) 为

$$\begin{aligned} \text{MSE} &= \int_{-\infty}^{\infty} (255x)^2 f(x) dx = \frac{255^2}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \\ &= \frac{255^2\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{(x-\mu)^2}{\sigma^2} + \frac{2\mu x}{\sigma^2} - \frac{\mu^2}{\sigma^2} \right] e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = I_1 + I_2 - I_3 \end{aligned} \quad (21)$$

其中,

$$\begin{aligned} I_1 &= \frac{255^2\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{255^2\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} q^2 e^{-\frac{q^2}{2}} dq = \\ &= \frac{255^2\sigma^2}{\sqrt{2\pi}} \left(q e^{-\frac{q^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-\frac{q^2}{2}} dq \right) = \frac{255^2\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{q^2}{2}} dq = 255^2\sigma^2 \end{aligned} \quad (22)$$

$$I_2 = \frac{2 \cdot 255^2\mu}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx =$$

$$\frac{2 \cdot 255^2\mu\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{x-\mu}{\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx +$$

$$2 \cdot 255^2\mu^2 \int_{-\infty}^{\infty} f(x) dx = \frac{2 \cdot 255^2\mu\sigma}{\sqrt{2\pi}} \int_0^{\infty} e^{-t} dt + 2 \cdot 255^2\mu^2 = (23)$$

$$\frac{2 \cdot 255^2\mu\sigma}{\sqrt{2\pi}} + 2 \cdot 255^2\mu^2$$

$$I_3 = \frac{255^2\mu^2}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 255^2\mu^2 \int_{-\infty}^{\infty} f(x) dx = 255^2\mu^2 \quad (24)$$

则,

$$\text{MSE} = I_1 + I_2 - I_3 = 255^2\sigma^2 + \frac{2 \cdot 255^2\mu\sigma}{\sqrt{2\pi}} + 255^2\mu^2 \quad (25)$$

因此, 图像的峰值信噪比为

$$\begin{aligned} \text{PSNR} &= 10 \lg \frac{255^2}{\text{MSE}} = \\ &= 10 \lg \frac{\sqrt{2\pi}}{\sqrt{2\pi\sigma^2 + 2\mu\sigma + \sqrt{2\pi}\mu^2}} \end{aligned} \quad (26)$$

参考文献:

- [1] WANG Z C, FENG G R, REN Y L, et al. Multichannel steganography in digital images for multiple receivers[J]. IEEE Multimedia, 2021, 28(1): 65-73.
- [2] WANG Z C, FENG G R, ZHANG X P. Repeatable data hiding: towards the reusability of digital images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 135-146.
- [3] 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.
- [4] XU M T, FENG G R, REN Y L, et al. On cloud storage optimization of blockchain with a clustering-based genetic algorithm[J]. IEEE Internet of Things Journal, 2020, 7(9): 8547-8558.
- [5] Bigverse[EB].
- [6] 孙曦, 张卫明, 俞能海, 等. 基于空域图像变换参数扰动的隐写术[J]. 通信学报, 2017, 38(10): 166-174.
- [7] SUN X, ZHANG W M, YU N H, et al. Steganography based on parameters' disturbance of spatial image transform[J]. Journal on Communications, 2017, 38(10): 166-174.
- [8] WANG Z C, FENG G R, QIAN Z X, et al. JPEG steganography with content similarity evaluation[J]. IEEE Transactions on Cybernetics, 2022: 1-12.
- [9] 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.
- [10] LI W X, ZHANG W M, LI L, et al. Designing near-optimal steganographic codes in practice based on polar codes[J]. IEEE Transactions on Communications, 2020, 68(7): 3948-3962.

- [10] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//Proceedings of 2012 IEEE International Workshop on Information Forensics and Security. 2012: 234-239.
- [11] HOLUB V, FRIDRICH J. Digital image steganography using universal distortion[C]//Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security-IH&MMSec'13. 2013: 59-68.
- [12] LI B, WANG M, HUANG J W, et al. A new cost function for spatial image steganography[C]//Proceedings of 2014 IEEE International Conference on Image Processing. 2014: 4206-4210.
- [13] SEDIGHI V, COGRANNE R, FRIDRICH J. Content-adaptive steganography by minimizing statistical detectability[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(2): 221-234.
- [14] 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.
- [15] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015: 234-241.
- [16] TANG W X, LI B, TAN S Q, et al. CNN-based adversarial embedding for image steganography[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(8): 2074-2087.
- [17] DENEMARK T, FRIDRICH J. Steganography with multiple JPEG images of the same scene[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(10): 2308-2319.
- [18] CHEN K J, ZHANG W M, ZHOU H, et al. Defining cost functions for adaptive steganography at the microscale[C]//Proceedings of 2016 IEEE International Workshop on Information Forensics and Security. 2016: 1-6.
- [19] ZHANG W M, ZHANG Z, ZHANG L L, et al. Decomposing joint distortion for adaptive steganography[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(10): 2274-2280.
- [20] ZHOU W B, ZHANG W M, YU N H. A new rule for cost reassignment in adaptive steganography[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2654-2667.
- [21] WANG Z C, QIAN Z X, ZHANG X P, et al. On improving distortion functions for JPEG steganography[J]. IEEE Access, 2018, 6: 74917-74930.
- [22] LI W X, CHEN K J, ZHANG W M, et al. JPEG steganography with estimated side-information[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(7): 2288-2294.
- [23] SHI L N, WANG Z C, QIAN Z X, et al. Distortion function for steganography in texture synthesized images[M]//Recent Advances in Intelligent Information Hiding and Multimedia Signal Processing. Cham: Springer International Publishing, 2018: 313-320.
- [24] LI S, CHEN X, WANG Z C, et al. Data hiding in IRIS image for privacy protection[J]. IETE Technical Review, 2018, 35: 34-41.
- [25] SHI L N, WANG Z C, QIAN Z X, et al. Distortion function for emoji image steganography[J]. Computers, Materials & Continua, 2019, 59(3): 943-953.
- [26] WANG Z, FENG G, ZHANG X. Steganographic distortion function for enhanced images[C]//Proceedings of the 19th International Workshop on Digital-forensics and Watermarking (IWDW 2020). 2020: 31-40.
- [27] PEVNY T, BAS P, FRIDRICH J. Steganalysis by subtractive pixel adjacency matrix[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 215-224.
- [28] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
- [29] HOLUB V, FRIDRICH J. Random projections of residuals for digital image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(12): 1996-2006.
- [30] DENEMARK T, SEDIGHI V, HOLUB V, et al. Selection-channel-aware rich model for Steganalysis of digital images[C]//Proceedings of 2014 IEEE International Workshop on Information Forensics and Security. 2014: 48-53.
- [31] GOLJAN M, FRIDRICH J, COGRANNE R. Rich model for Steganalysis of color images[C]//Proceedings of 2014 IEEE International Workshop on Information Forensics and Security. 2014: 185-190.
- [32] LI B, LI Z P, ZHOU S J, et al. New steganalytic features for spatial image steganography based on derivative filters and threshold LBP operator[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1242-1257.
- [33] KODOVSKY J, FRIDRICH J, HOLUB V. Ensemble classifiers for steganalysis of digital media[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 432-444.
- [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] LI B, WEI W H, FERREIRA A, et al. ReST-net: diverse activation modules and parallel subnets-based CNN for spatial image steganalysis[J]. IEEE Signal Processing Letters, 2018, 25(5): 650-654.
- [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] 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.
- [38] YANG H W, HE H, ZHANG W Z, et al. FedSteg: a federated transfer learning framework for secure image steganalysis[J]. IEEE

Transactions on Network Science and Engineering, 2021, 8(2): 1084-1094.

- [39] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[C]// Proceedings of the 12th International Conference on Information Hiding, 2010: 161-177.



冯国瑞（1976—），男，辽宁沈阳人，博士，上海大学教授、博士生导师，主要研究方向为多媒体取证与安全、隐写分析、机器学习。

[作者简介]



王子驰（1992—），男，贵州毕节人，博士，上海大学讲师，主要研究方向为隐写、隐写分析、人工智能安全。



张新鹏（1975—），男，黑龙江牡丹江人，博士，上海大学教授、博士生导师，主要研究方向为多媒体信息安全。