

ISEM: 基于生成多对抗网络的图像隐写增强模型

周赛星^{1,2}, 叶苗欣^{1,2}, 骆伟祺^{1,2}

(1. 中山大学计算机学院, 广东 广州 510006; 2. 广东省信息安全技术重点实验室, 广东 广州 510006)

摘要: 图像隐写技术根据在同一嵌入单元内+1和-1的嵌入代价是否相等, 通常分为对称隐写和非对称隐写。对称隐写技术未能充分利用相邻像素之间的相互关系, 限制了隐写安全性的提升。非对称隐写技术虽能显著增强隐写的安全性, 但在生成非对称嵌入代价的过程中, 现有方法往往存在训练过程不稳定、依赖人工经验或预设参数等问题, 难以确保隐写的安全性。针对上述问题, 提出了一种基于生成对抗网络的图像隐写增强模型 (image steganography enhancement model, ISEM)。生成器网络采用相对轻量级的双分支网络结构, 判别器网络则由3个基于深度学习的隐写分析网络组成。ISEM的训练分为两个阶段。第一阶段学习载体图像与现有嵌入代价之间的映射关系; 第二阶段通过对抗训练调整初步学习的嵌入代价, 生成高安全性的非对称嵌入代价。此外, 在生成器中引入几何均值损失, 在判别器中引入均值绝对偏差损失, 以进一步优化训练效果。实验结果表明, ISEM能够有效增强对称 (以及非对称) 隐写方法的安全性, 其相较于现有的非对称隐写方法, 具有更佳的平均安全性。

关键词: 图像隐写; 生成对抗网络; 几何均值损失; 均值绝对偏差损失

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025053

ISEM: image steganography enhancement model based on generative multi-adversarial network

ZHOU Saixing^{1,2}, YE Miaoxin^{1,2}, LUO Weiqi^{1,2}

1. School of Computer Science and Engineering, SunYat-sen University, Guangzhou 510006, China

2. Guangdong Key Laboratory of Information Security Technology, Guangzhou 510006, China

Abstract: Image steganography techniques were typically categorized into symmetric and asymmetric types based on whether the embedding costs of +1 and -1 within the same embedding unit were equal. The interrelationship between adjacent pixels was not fully utilized by symmetric steganography techniques, which limited the enhancement of steganographic security. Although asymmetric steganography techniques were able to significantly im-

收稿日期: 2024-12-31; 修回日期: 2025-03-19

通信作者: 骆伟祺, luoweiqi@mail.sysu.edu.cn

基金项目: 国家自然科学基金 (62472458); 广东省自然科学基金 (2025A1515012823)

Foundation Items: The National Natural Science Foundation of China (62472458), The Natural Science Foundation of Guangdong Province (2025A1515012823)

引用格式: 周赛星, 叶苗欣, 骆伟祺. ISEM: 基于生成多对抗网络的图像隐写增强模型[J]. 网络与信息安全学报, 2025, 11(5): 126-136.

Citation Format: ZHOU S X, YE M X, LUO W Q. ISEM: image steganography enhancement model based on generative multi-adversarial network[J]. Chinese Journal of Network and Information Security, 2025, 11(5): 126-136.

prove security, existing methods were often faced with issues such as unstable training processes, reliance on manual experience, or dependence on preset parameters when generating asymmetric embedding costs, making it difficult to ensure steganographic security. To address these issues, an image steganography enhancement model (ISEM) based on a generative multi-adversarial network was proposed. A relatively lightweight dual-branch structure was adopted by the generator network, while the discriminator network was composed of three deep learning-based steganalysis networks. The training of ISEM was divided into two stages. In the first stage, the mapping relationship between the carrier image and existing embedding costs was learned. In the second stage, the initially learned embedding costs were adjusted through adversarial training to generate high-security asymmetric embedding costs. Additionally, a geometric mean loss was introduced in the generator, and a mean absolute deviation loss was incorporated in the discriminator to further optimize the training effect. Experimental results show that the security of both symmetric and asymmetric steganography methods are effectively enhanced by ISEM. Compared to existing asymmetric steganography methods, the highest average security is achieved by the proposed approach.

Keywords: image steganography, generative adversarial network, geometric mean loss, mean absolute deviation loss

0 引言

图像隐写技术利用载体图像的信息统计冗余特性和人类视觉感知特征,通过修改特定的嵌入单元(如像素值、DCT系数等),将秘密信息隐蔽地嵌入载体图像,从而实现隐蔽通信^[1-3]。目前,自适应空域图像隐写领域的研究重点主要集中在两个方面:一是根据载体图像的内容将秘密信息的嵌入操作集中在难以进行统计建模的区域;二是借助最小化失真框架完成信息嵌入^[4]。在该框架下,Fridrich等^[5]提出的校验子格编码(syndrome-trellis code, STC)算法,因效果已接近“率-失真”理论界限^[6],故而在实际应用中得到广泛采用。

设计科学有效的失真函数是自适应隐写方法面临的研究挑战。其设计目标是为纹理复杂区域分配较低的嵌入代价,而为纹理相对简单的区域(如平滑区域)分配较高的嵌入代价,以此实现将更多秘密信息隐藏于嵌入代价较低的纹理复杂区域。这种分配策略的理论依据在于:纹理复杂区域中像素值的细微变化既不易被人类视觉系统察觉,也难以通过统计模型进行有效建模,从而能够显著提高信息的隐蔽性。

根据在同一嵌入单元内进行+1和-1修改的嵌入代价是否相等,隐写技术可分为对称隐写和非对称隐写。传统的隐写方法大多数采用对称隐写模式,代表性算法包括HUGO (highly undetectable stego)^[7]、WOW (wavelet obtained weight)^[8]和

HILL^[9]等。这类基于启发式设计的对称隐写方法由于未充分考虑图像内部的像素相关性,且忽略了不同方向修改对图像的差异化影响,容易被隐写分析算法提取有效特征。

近年来,基于卷积神经网络^[10] (convolutional neural network, CNN)的对称隐写方法逐渐兴起。Tang等^[11]首次提出基于生成对抗网络(generative adversarial network, GAN)的隐写失真函数模型ASDL-GAN,通过对抗学习机制逐步优化隐写性能,但其安全性仍不及传统HILL算法。Yang等^[12]在ASDL-GAN基础上改进生成器结构,采用UNet^[13]架构,提出了UT-GAN隐写方法,其安全性能已超越传统隐写算法。Tang等^[14]提出基于GAN的增强策略SPAR-RL (steganographic pixel-wise actions and rewards with reinforcement learning),将GAN模块引入强化学习框架,进一步提高了隐写安全性。Huang等^[15]提出的Steg-GMAN (steganographic method based on generative multi-adversarial network)通过构建一组CNN隐写分析网络作为GAN的判别器,有效增强了隐写方法的抗分析能力。

虽然上述对称隐写方法取得了一定的进展,但仅依赖对称嵌入代价的方式在安全性提升上存在固有局限性。现有研究表明^[16-18, 20-23],设计非对称嵌入代价能够更充分地利用像素间的相关性,进一步提高隐写信息的隐蔽性,从而提高隐写安全性。在基于预设规则修改对称嵌入代价的隐写方法中, Li等^[16]和Denemark等^[17]基于非加

性失真框架,考虑了嵌入单元之间的相互作用,分别提出了CMD和SMD隐写策略。这些策略通过子格交替嵌入和动态调整嵌入代价,促使相邻像素尽可能保持相同的修改方向,有效降低了被检测风险。Wang等^[18]提出了一种基于图像纹理和像素波动的失真函数调整方法,该方法通过分析对像素进行+1或-1修改的嵌入代价来优化隐写过程,使修改后的像素波动特性与邻域的波动特性更加一致,进而提高隐蔽性。受快速梯度符号方法^[19]启发,一些方法通过对抗本来设计非对称嵌入代价。ADV-EMB^[20]、JS-IAE^[21]和MAE^[22]等方法利用预训练CNN隐写分析器的梯度信息调整原始嵌入代价,从而有效对抗目标CNN隐写分析器的检测。这类非对称隐写方法通常先通过HILL等对称隐写方法初始化代价,再调整为非对称嵌入代价,这可被视为一种隐写增强策略。与之不同,基于生成对抗网络的非对称方法可以直接生成非对称嵌入代价。Huang等^[23]提出的方法(本文称为Huang-GAN),将Steg-GMAN扩展到非对称隐写领域。该方法通过双分支网络生成与输入图像大小相同的正向和负向嵌入概率图,并设计了一种结合图像残差、梯度、非对称概率图和修改符号的对抗损失函数。通过这种方式,生成器在与判别器对抗过程中,从零开始自动学习生成非对称嵌入代价。然而,该方法的训练过程往往不稳定,难以保证生成的嵌入代价的安全性。相比之下,隐写增强策略在已有嵌入代价的基础上进行优化,能够实现当前最优的安全性能。因此,本文主要聚焦于隐写增强策略。

受Huang-GAN模型在生成对抗网络中设计非对称嵌入代价的启发,本文提出了基于GAN的ISEM(improved steganography enhancement model),旨在解决对称隐写方法未能充分利用像素间相关性导致的安全性提升受限问题。与现有隐写增强策略不同,ISEM采用先学习后优化调整的训练方式,无须依赖手工设计的调整规则或固定参数。具体而言,该模型首先学习隐写方法中载体图像与嵌入代价之间的映射关系,然后通过迭代对抗训练自动捕捉像素间的复杂相关性,从而优化生成更安全的非对称

嵌入代价,以降低被检测的概率。实验结果表明,ISEM模型不仅能够显著提升对称隐写方法的安全性,还能够增强非对称隐写方法的安全性。

本文的主要贡献如下。

1) 提出一种新的基于GAN的隐写增强策略。该方法采用一个相对轻量级的双分支生成器和一个由3个隐写分析网络组成的判别器。生成器首先学习现有隐写方法中载体图像与嵌入代价之间的映射关系,然后通过对抗训练动态调整生成非对称嵌入代价。该设计使ISEM能够在不依赖人工规则的情况下,自适应优化嵌入代价,提高隐写安全性。

2) 设计新型损失函数以优化训练过程。为生成器设计了几何均值损失函数,并为判别器设计了均值绝对偏差损失函数,通过加权计算判别器内所有隐写分析网络的图像判别损失,这一设计有效平衡了生成器与判别器的训练过程,增强了模型的收敛性,并逐步提升了生成器的隐写性能。

3) 实验验证与消融分析。为验证所提增强模型的有效性,本文在多个对称和非对称隐写方法上进行测试,实验结果表明,ISEM能够有效提升安全性能。此外,通过消融实验验证了各模块的有效性,确保模型的合理性和可解释性。

1 隐写增强模型

1.1 模型框架

ISEM的整体框架如图1所示,其训练过程分为两个阶段。①第一阶段:初步学习原始载体图像与已有嵌入代价的映射关系。②第二阶段:利用GAN对第一阶段学习到的映射关系进行优化。下面详细介绍这两个阶段的实现方法。

第一阶段的主要目标是让生成器学习现有隐写方法在载体图像与嵌入代价之间的映射关系,为后续对抗训练奠定基础。具体实现步骤如下:① 给定一个隐写方法,计算其在载体图像 C 上的正向嵌入概率 P_{ori}^{+1} 和负向嵌入概率 P_{ori}^{-1} 。其中, P_{ori}^{+1} 和 P_{ori}^{-1} 分别表示在像素点 (i, j) 上进行+1修改和-1修改的概率。②向模型中的生成器 G 输入载体图像 C ,进行训练,使得生成的嵌入概率

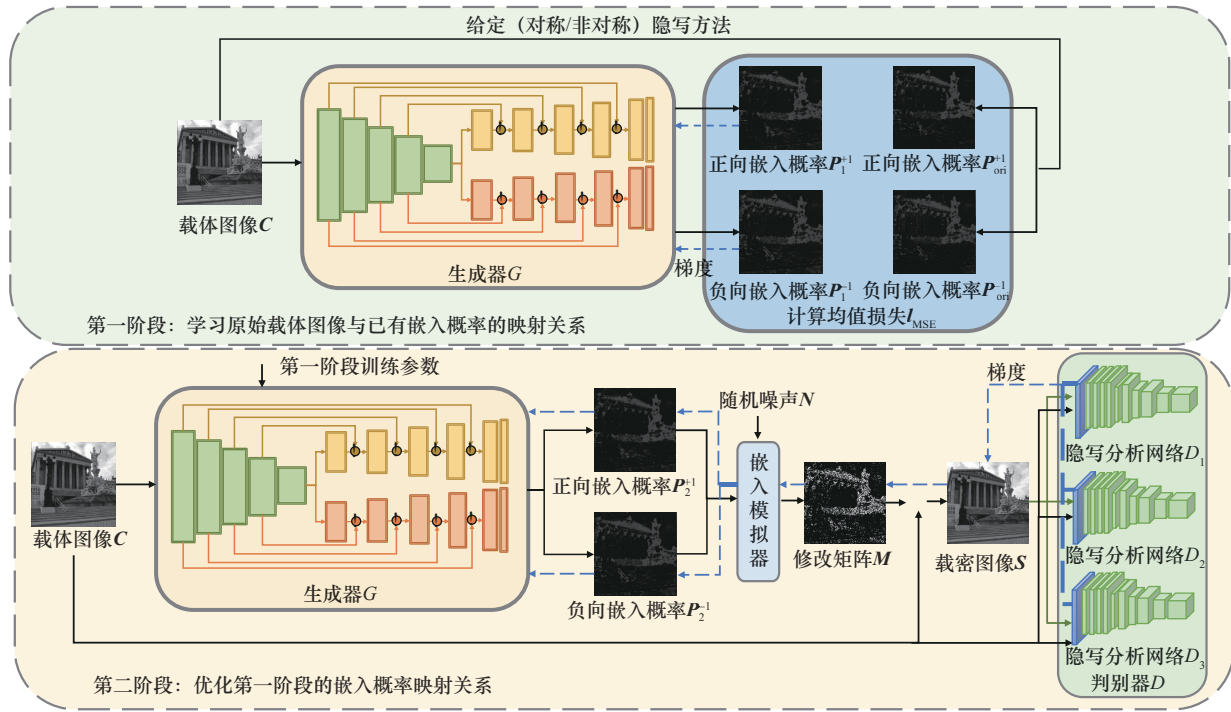


图1 ISEM的整体框架

Figure 1 Overall framework of ISEM

P_1^{+1} 和 P_1^{-1} 尽可能接近 P_{ori}^{+1} 和 P_{ori}^{-1} 。③为量化生成器的学习效果, 定义的损失函数由均方误差 (mean squared error, MSE) 计算得到, 如式(1)所示。

$$l_{MSE} = \frac{1}{H \times W} \left(\sum_{i=1}^H \sum_{j=1}^W \left(P_1^{+1}(i,j) - P_{ori}^{+1}(i,j) \right)^2 + \sum_{i=1}^H \sum_{j=1}^W \left(P_1^{-1}(i,j) - P_{ori}^{-1}(i,j) \right)^2 \right) \quad (1)$$

其中, H 、 W 分别表示载体图像 C 的高和宽, i 、 j 表示元素位置下标。这一训练过程有助于使生成器生成的嵌入概率与已有嵌入概率具备相似特性。

第二阶段的主要目标是对第一阶段学习到的嵌入概率映射关系进行优化, 通过生成非对称嵌入代价来显著提升隐写方法的安全性能。具体实现过程如下。

首先, 以第一阶段训练得到的生成器参数作为初始值, 此时生成器 G 已初步建立载体图像与已有嵌入概率之间的映射关系。在模型训练阶段, 生成器 G 以载体图像 C 作为输入, 输出正向嵌入概率 P_2^{+1} 和负向嵌入概率 P_2^{-1} 。考虑到需结合 STC 算法进行信息嵌入, 先将生成的嵌入概率转

换为嵌入代价, 才能应用 STC 算法完成信息嵌入操作。然而, 这一转换过程会显著降低训练效率。为提升训练效率, 本文采用阶梯函数模拟 STC 嵌入这一相对复杂的过程, 实现秘密信息向载体图像的嵌入, 从而获得载密图像 S 。因此, 需先生成一个随机噪声 N , 根据式(2)将 N 与 P_2^{+1} 和 P_2^{-1} 进行比较, 从而得到 C 的修改矩阵 M ; 再利用式(3)计算出载密图像 S 。

$$M(i,j) = \begin{cases} +1, N(i,j) < P_2^{+1}(i,j) \\ -1, N(i,j) > 1 - P_2^{-1}(i,j) \\ 0, \text{其他情况} \end{cases} \quad (2)$$

$$S(i,j) = C(i,j) + M(i,j) \quad (3)$$

随后, 判别器 D 以载体图像 C 和载密图像 S 作为输入。 D 内部集成的所有隐写分析网络对 C 和 S 两类图像进行判别, 并将判别梯度反向传播至生成器 G 。生成器 G 通过与判别器 D 的迭代对抗训练, 持续调整优化映射关系, 从而生成非对称嵌入概率。在实际信息嵌入阶段, 通过式(4)实现非对称嵌入概率 P_2^{+1} 、 P_2^{-1} 向非对称嵌入代价 ρ^{+1} 、 ρ^{-1} 的转化。最终结合 STC 算法完成秘密信息嵌入, 生成载密图像。

$$\begin{cases} \rho^{+1}(ij) = \ln\left(\frac{1}{P_2^{+1}(ij)} - 2\right) \\ \rho^{-1}(ij) = \ln\left(\frac{1}{P_2^{-1}(ij)} - 2\right) \\ \rho^0(ij) = 0 \end{cases} \quad (4)$$

1.1.1 生成器网络设计

ISEM 与 Huang-GAN 方法在生成器结构设计上存在显著差异。ISEM 减少了 3 层下采样和 3 层上采样操作，同时将上采样过程中原有的跳跃连接拼接方式改为特征相加方式。这一改进不仅简化了特征融合过程，还将网络参数量缩减至原始模型的 1/4。值得注意的是，ISEM 保留了生成器的双分支输出结构，以支持非对称嵌入概率的生成。ISEM 生成器的整体架构如图 2(a) 所示，其采用对称的 5 层下采样层（卷积块）和 5 层上采样层（反卷积块）架构，分别对应生成器的输入和输出。其中，每个卷积块包含 2 个 3×3 卷积核的卷积层，结构如图 2(b) 所示；每个反卷积块包含 1 个 5×5 卷积核的卷积层和 1 个 5×5 卷积核的反卷积层，结构如图 2(c) 所示。

1.1.2 判别器网络设计

在 ISEM 训练过程的第二阶段，判别器的作用是区分载体图像和载密图像。为此，本文借鉴现有的隐写分析网络架构来设计判别器。在训练过程中，判别器通过对比生成图像与真实载体图像，计算交叉熵损失来评估生成器的性能。生成器则基于判别器的反馈信息，不断优化其参数并

调整嵌入概率分布，从而生成更优的非对称嵌入概率。判别器中隐写分析网络的选择对维持生成器与判别器之间的动态平衡至关重要。研究表明^[15, 23]，若判别器性能过弱，生成器将因接收的反馈信息有限而难以有效优化；反之，若判别器性能过于强大，生成器可能因梯度信息过于严苛而训练不稳定。因此，为确保判别器具备适度的判别能力，本文选用 SRNet^[24]、CovNet^[25]和 Zhu-Net^[26] 3 种代表性的隐写分析网络来构建判别器。

1.2 损失函数

在 GAN 框架下，ISEM 的第二阶段训练通过生成器和判别器的对抗博弈实现动态优化。在这一过程中，生成器致力于生成逼近真实分布的载密图像，判别器则不断提高其鉴别能力，二者通过相互竞争推动模型向纳什均衡收敛。损失函数作为训练过程的核心优化目标，不仅为网络参数更新提供明确方向，同时还要确保生成器与判别器在对抗训练中的动态平衡。

在 ISEM 的第二阶段训练中，损失函数体系由判别器损失和生成器损失两部分组成。这两个损失函数的具体数学表达和优化机制将在后续章节进行详细阐述。

1.2.1 判别器损失

在模型迭代训练过程中，将交叉熵损失函数作为判别器中隐写分析网络 D_k 的优化目标，其数学表达式如式(5)所示。

$$l_{D_k}(C, S) = -z_0 \log(D_k(C)) - z_1 \log(D_k(S)) \quad (5)$$

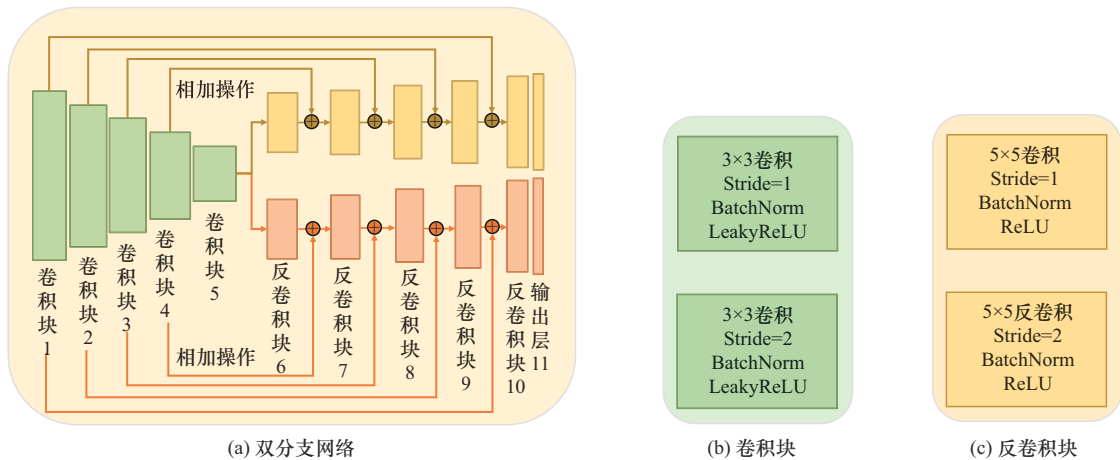


图2 生成器网络结构

Figure 2 Generator network structure

其中, $D_k(\mathbf{C})$ 和 $D_k(\mathbf{S})$ 分别表示判别器 D 内第 k 个隐写分析网络对载体图像 $\mathbf{C}$ 和载密图像 $\mathbf{S}$ 产生的分类结果向量, $k \in \{1, 2, 3\}$ 。 $\mathbf{z}_0$ 和 $\mathbf{z}_1$ 表示图像的真实标签。交叉熵 l_{D_k} 的数值越小表明隐写分析网络 D_k 的判别性能越强; 反之, 则说明其判别性能越弱。

此外, 本文设计了一种名为均值绝对偏差(mean absolute deviation, MAD)的损失函数, 其数学表达式如式(6)所示。

$$l_{\text{MAD}} = \frac{1}{3} \sum_{i=1}^3 \left| l_{D_i}(\mathbf{C}, \mathbf{S}) - \frac{1}{3} \sum_{j=1}^3 l_{D_j}(\mathbf{C}, \mathbf{S}) \right| \quad (6)$$

该损失函数用于评估判别器内部多个隐写分析网络之间性能的一致性。 l_{MAD} 数值越大, 表明各隐写分析网络的判别结果差异越显著。判别器 D 以最小化该损失值为目标, 以推动内部网络形成一致的判断结论。基于此, 判别器 D 的最终损失函数设计如下。

$$l_D = \max_{k \in \{1, 2, 3\}} l_{D_k}(\mathbf{C}, \mathbf{S}) + \omega_k \cdot l_{\text{MAD}} \quad (7)$$

$$\omega_k = \frac{e^{\left(1 - \max_{k \in \{1, 2, 3\}} l_{D_k}(\mathbf{C}, \mathbf{S})\right)}}{\sum_{i=1}^3 e^{\left(1 - l_{D_i}(\mathbf{C}, \mathbf{S})\right)}} \quad (8)$$

其中, l_D 表示在每次迭代训练中, 判别器 D 会依据当前性能表现最差(即 l_{D_k} 数值最大)的隐写分析网络, 并结合均值绝对偏差 l_{MAD} 来更新判别器整体的网络参数, 从而提高其判别性能和稳定性。 ω_k 是 l_{MAD} 的权值, 反映了 l_{MAD} 在损失计算中的重要性。 ω_k 数值越大, 表明 l_{MAD} 在损失计算中的权重越高, 对判别器整体损失的影响越显著。

1.2.2 生成器损失

生成器的损失函数 l_G 由有效载荷嵌入损失 l_G^1 和对抗损失 l_G^2 构成, 其数学表达式如式(9)所示。

$$l_G = \alpha \cdot l_G^1 + \beta \cdot l_G^2 \quad (9)$$

其中, α 、 β 分别表示 l_G^1 和 l_G^2 的权重。在本文实验参数设置中, $\alpha = 10^{-7}$, $\beta = -1$ 。

有效载荷嵌入损失 l_G^1 用于控制载体图像的嵌入容量, 其数学表达式如式(10)所示。

$$l_G^1 = \left(-\sum_{m \in \{-1, 0, +1\}} \mathbf{P}_2^m(ij) \cdot \log_2(\mathbf{P}_2^m(ij)) - H \cdot W \cdot q \right)^2 \quad (10)$$

其中, m 表示在像素上的修改值, $m \in \{-1, 0, +1\}$ 。

q 表示设置的嵌入率, 单位为bpp (bits per pixel)。

对抗损失 l_G^2 用于与判别器 D 进行对抗, 其数学表达式如式(11)所示。

$$l_G^2 = \frac{1}{H \cdot W} \sum_{\forall(ij)} \mathbf{r}(ij) \cdot (\log(\mathbf{P}_2^{+1}(ij)) \cdot I(\mathbf{M}(ij) = +1) + \log(\mathbf{P}_2^{-1}(ij)) \cdot I(\mathbf{M}(ij) = -1)) \quad (11)$$

其中, $I(\cdot)$ 是一个指标函数。权重 $\mathbf{r}(ij)$ 的数学表达式如下。

$$\mathbf{r}(ij) = \epsilon \cdot \mathbf{M}(ij) \cdot \mathbf{g}(ij) \cdot \mathbf{T}(ij) \quad (12)$$

$$\mathbf{g}(ij) = \frac{\partial L_{\omega}}{\partial \mathbf{M}(ij)} \quad (13)$$

其中, $\epsilon = 10^7$ 。 $\mathbf{g}(ij)$ 表示判别器 D 关于修改矩阵 $\mathbf{M}(ij)$ 的梯度, 定义为损失函数 L_{ω} 对 $\mathbf{M}(ij)$ 的偏导数, 具体计算如式(13)~式(15)所示。 $\mathbf{T}(ij)$ 表示载体图像 $\mathbf{C}$ 应用拉普拉斯8邻域算子取绝对值后获得的残差矩阵, 其作用是将嵌入修改集中在难以进行统计建模的区域。当修改矩阵 $\mathbf{M}(ij)$ 的像素修改方向与梯度 $\mathbf{g}(ij)$ 的方向一致时, $\mathbf{r}(ij)$ 作为激励(旨在鼓励生成器在生成载密图像时保持像素修改方向与梯度方向的一致性)。这种设计有助于生成更安全的非对称嵌入概率分布, 从而有效提高隐写系统的安全性能。

本文还设计了一种几何均值(geometric mean, GM)损失, 其数学表达式如式(14)所示。

$$l_{\text{GM}} = e^{\left(\frac{\sum_{k=1}^3 e^{\left(1 - l_{D_k}(\mathbf{C}, \mathbf{S})\right)} \cdot \log(l_{D_k}(\mathbf{C}, \mathbf{S}))}{\sum_{i=1}^3 e^{\left(1 - l_{D_i}(\mathbf{C}, \mathbf{S})\right)}} \right)} \quad (14)$$

该损失函数用于对判别器中隐写分析网络的损失进行加权几何平均计算, 为生成器调整嵌入概率提供综合反馈和指导。 l_{GM} 数值越大, 表明生成器优化后的非对称嵌入概率对判别器检测的规避能力越强。生成器 G 以最大化 l_{GM} 为目标, 以增强对抗能力并实现更隐蔽的信息嵌入。基于此, 生成器的最终损失函数 L_{ω} 设计如式(15)所示。

$$L_{\omega} = \min_{k \in \{1, 2, 3\}} l_{D_k}(\mathbf{C}, \mathbf{S}) + l_{\text{GM}} \quad (15)$$

L_{ω} 代表在每次迭代训练中, 生成器 G 会依

据当前性能最优（即 I_{D_k} 数值最小）的隐写分析网络，并结合加权对数 I_{GM} 的反馈信息来更新生成器整体的网络参数。这种优化策略使生成器能够动态调整嵌入概率，从而生成更安全的非对称嵌入代价，进一步提升隐写系统的抗检测能力。

2 隐写实验与结果分析

2.1 实验设置

在训练阶段，从 SZUBase^[14] 数据集中随机选取 40 000 张图像用于模型训练；在测试阶段，从 BOSSBase^[27] 数据集中选取 10 000 张图像用于模型测试。为评估隐写安全性，本文选取了 6 种具有代表性的隐写分析器（包括 1 种传统隐写分析器 SRM^[28] 和 5 种基于 CNN 的隐写分析器 Xu-Net^[29]、Yedroudj-Net^[30]、SRNet^[24]、CovNet^[25] 和 Zhu-Net^[26]）。具体配置如下。

1) 传统隐写分析器 SRM：随机选取 5 000 组图像对用于训练，剩余 5 000 组图像对用于测试。

2) 基于 CNN 的隐写分析器：使用 5 000 组载体一载密图像对进行模型训练，1 000 组图像对用于验证，4 000 组图像对用于性能评估。

本文选取 1 种传统的对称隐写方法 HILL^[9]、4 种基于 GAN 的对称隐写方法（ASDL-GAN^[10]、UT-GAN^[12]、SPAR-RL^[14]、Steg-GMAN^[15]），以及 1 种非对称隐写方法 Huang-GAN^[23]，分别与 ISEM 结合（组合后分别命名为 ISEM-HILL、ISEM-ASDL、ISEM-UT、ISEM-SPAR、ISEM-Steg 和 ISEM-Huang），进行安全性分析。此外，还将非对称隐写策略 CMD^[16] 和 SMD^[17] 分别与 HILL 结合（分别命名为 CMD-HILL 和 SMD-HILL）作为对比组。

ISEM 基于 PyTorch 深度学习框架实现，实验的硬件环境配置为 Intel (R) Core (TM) i7-6900K @3.20 GHz CPU 及 4 个 NVIDIA TITAN X GPU。实验中网络模型的关键参数设置如下。

1) 训练参数设置：两个阶段训练批次皆为 24，周期数量设置为 72，并采用 Adam 优化器进行网络训练。

2) 学习率设置：第一阶段生成器学习率为 1×10^{-4} ，第二阶段生成器学习率为 1×10^{-5} 、判别器为 1×10^{-4} 。学习率每 20 个训练周期衰减 50%。

2.2 安全性分析

本节对各类隐写方法在 0.4 bpp 嵌入率下的安全性进行评估，实验结果见表 1。基于表 1 数据可得出以下结论。

1) ISEM 能够显著提升对称隐写方法的安全性。ISEM-HILL、ISEM-ASDL、ISEM-UT、ISEM-SPAR 和 ISEM-Steg 的平均安全性能分别提升 12.38%、12.68%、8.36%、6.79% 和 5.40%；与非对称隐写方法 SMD-HILL 和 CMD-HILL 相比，ISEM-HILL 平均安全性能分别提升了 7.57%（从 27.23% 提升至 34.80%）和 6.85%（从 27.95% 提升至 34.80%）。

2) ISEM 能够优化非对称隐写方法的嵌入代价。在抵抗 CovNet 隐写分析器检测时，ISEM-Huang 的安全性能从 31.63% 提升至 34.94%，增幅达 3.31%。

综上所述，实验结果充分证明了 ISEM 的有效性。在抵抗传统隐写分析器和基于 CNN 的隐写分析器检测时，现有隐写方法与 ISEM 结合后，安全性能均得到显著提升。与基于预设规则的隐写增强策略 SMD、CMD 相比，ISEM 通过引入生成对抗网络架构，借助迭代对抗训练机制，能够更精准地捕捉图像嵌入单元间的相关性，动态生成安全性更高的非对称嵌入代价分布，从而在隐写安全性能上实现显著突破。

此外，本文对基于深度学习的隐写方法在训练阶段的耗时进行了对比分析，实验结果见表 2。为保证对比的公平性，所有实验均在相同的服务器环境下开展。从表 2 的实验数据可以看出，ISEM 的训练时间最长。这是因为该模型采用分阶段的训练方式，第二阶段网络复杂度的增加是为了进一步优化模型性能。虽然训练时间有所延长，但这一设计显著提升了模型的隐写性能。

2.3 消融实验

2.3.1 判别器内隐写分析网络组合对模型安全性能的影响

为探究 ISEM 第二阶段训练中判别器内不同隐写分析网络组合在嵌入率为 0.4 bpp 条件下对模型安全性能的影响，本文设计 3 组对比实验，结果见表 3。除非特殊说明，后续实验中 ISEM 第一阶段训练均以 Huang-GAN 隐写方法为基础。

表1 不同隐写方法及与ISEM组合的抵抗隐写分析器检测的错误率
Table 1 Detection error rates of different steganographic methods and their combinations with ISEM against steganalysis detectors

隐写方法	SRM ^[28]	Xu-Net ^[29]	Yedroudj-Net ^[30]	SRNet ^[24]	CovNet ^[25]	Zhu-Net ^[26]	平均值	提升值
HILL ^[9]	26.92%	27.97%	19.54%	17.33%	16.10%	26.66%	22.42%	
SMD-HILL	30.91%	35.35%	25.02%	19.55%	19.63%	32.91%	27.23%	4.81% ↑
CMD-HILL	32.61%	35.17%	25.77%	20.45%	20.26%	33.43%	27.95%	5.53% ↑
ISEM-HILL	34.05%	40.35%	33.58%	32.45%	28.75%	39.62%	34.80%	12.38% ↑
ASDL-GAN ^[10]	24.50%	28.75%	23.12%	16.23%	13.74%	29.71%	22.68%	
ISEM-ASDL	34.64%	39.71%	35.43%	33.05%	30.36%	38.98%	35.36%	12.68% ↑
UT-GAN ^[12]	28.64%	33.61%	27.14%	21.92%	21.97%	34.20%	27.91%	
ISEM-UT	34.73%	40.93%	36.24%	35.45%	30.92%	39.36%	36.27%	8.36% ↑
SPAR-RL ^[14]	30.34%	35.53%	28.59%	22.81%	23.21%	33.00%	28.91%	
ISEM-SPAR	35.31%	39.95%	35.30%	33.37%	30.34%	39.93%	35.70%	6.79% ↑
Steg-GMAN ^[15]	30.62%	35.29%	30.85%	26.15%	25.38%	34.10%	30.40%	
ISEM-Steg	36.31%	39.63%	35.85%	32.70%	31.19%	39.09%	35.80%	5.40% ↑
Huang-GAN ^[23]	35.78%	40.96%	39.91%	36.62%	31.63%	38.52%	37.24%	
ISEM-Huang	35.94%	41.63%	37.56	37.45%	34.94%	41.44%	38.16%	0.92% ↑

注: 黑色加粗数字表示模型增强前后的最高安全性能, 蓝色加粗并带有符号↑数字表示隐写方法经过模型(隐写策略)增强后的平均提升值

表2 不同深度学习隐写方法的训练时间对比
Table 2 Comparison of training times for different deep learning steganography methods

隐写方法	ASDL-GAN ^[10]	UT-GAN ^[12]	SPAR-RL ^[14]	Steg-GMAN ^[15]	Huang-GAN ^[23]	本文方法
训练时间/h	13.1	6.5	6.7	9.8	11.3	31.2

表3 判别器内不同网络组合的抵抗隐写分析检测的错误率对比
Table 3 Comparison of the error rates of resisting steganalysis detection for different network combinations within the discriminator

判别器内隐写分析网络组合	SRNet ^[24]	CovNet ^[25]	Zhu-Net ^[26]	SRM ^[28]	Xu-Net ^[29]	Yedroudj-Net ^[30]	平均值
Yedroudj-Net&SRNet&Zhu-Net	33.80%	30.91%	39.53%	35.95%	39.68%	31.30%	35.20%
Yedroudj-Net&SRNet&CovNet	36.75%	34.22%	39.12%	36.46%	40.82%	37.25%	37.44%
SRNet&CovNet&Zhu-Net (本文方法)	37.45%	34.94%	41.44%	35.94%	41.63%	37.56%	38.16%

注: 黑色加粗数字表示在相应情况下模型的最佳安全性能

根据表3的实验结果, 对比不同判别器隐写分析网络组合的性能表现发现: SRNet&CovNet&Zhu-Net组合, 相较于Yedroudj-Net&SRNet&Zhu-Net组合, 在CovNet检测下的安全性能提升显著(从30.91%提升至34.94%, 增幅达4.03%), 平均安全性能提升2.96%(从35.20%提升至38.16%); SRNet&CovNet&Zhu-Net组合相较于Yedroudj-Net&SRNet&CovNet组合, 平均安全性

能提升了0.72%(从37.44%提升至38.16%)。基于上述对比分析, ISEM最终选择SRNet&CovNet&Zhu-Net作为判别器的默认隐写分析网络组合。

2.3.2 损失函数对模型安全性能的影响

为探究ISEM第二阶段训练中几何均值损失 I_{GM} 和均值绝对偏差损失 I_{MAD} 的使用策略在嵌入率为0.4 bpp条件下对模型安全性能的影响, 本文设计3组对比实验, 结果见表4。

表4 不同损失函数引入策略的抵抗隐写分析检测的错误率对比

Table 4 Comparison of error rates of resistance to steganalysis detection with different loss function introduction strategies

损失函数引入策略	SRM ^[28]	Xu-Net ^[29]	Yedroudj-Net ^[30]	SRNet ^[24]	CovNet ^[25]	Zhu-Net ^[26]	平均值
G 不引入 l_{GM} 且 D 不引入 l_{MAD}	34.68%	40.90%	34.57%	32.40%	30.32%	39.26%	35.36%
G 引入 l_{GM} , D 不引入 l_{MAD}	36.81%	41.73%	37.78%	36.35%	32.95%	40.93%	37.76%
G 引入 l_{GM} 且 D 引入 l_{MAD} (本文方法)	35.94%	41.63%	37.56%	37.45%	34.94%	41.44%	38.16%

注:黑色加粗数字表示在相应情况下模型的最佳安全性能

基于表4数据可得出以下结论。

1) 当在判别器损失 l_D 中不引入均值绝对偏差损失 l_{MAD} , 而仅在生成器对抗损失 l_G^2 中引入几何均值损失 l_{GM} 时, 在SRNet上的安全性能提升3.95% (从32.40%提升至36.35%), 平均安全性能提升2.40% (从35.36%提升至37.76%)。这一结果表明, l_{GM} 有效促使生成器在判别器提供的综合反馈下优化嵌入概率映射关系, 能够生成安全性更高的非对称嵌入代价, 从而显著增强模型的安全性能。

2) 当在生成器对抗损失 l_G^2 中引入几何均值损失 l_{GM} , 并在判别器损失 l_D 中引入均值绝对偏差损失 l_{MAD} 时, 在CovNet上的安全性能提升1.99% (从32.95%提升至34.94%), 平均安全性能提升

0.40% (从37.76%提升至38.16%)。这一结果表明, l_{MAD} 有效提高了判别器的性能和稳定性, 从而使生成器能够获得更全面的信息反馈, 进一步提升了ISEM的安全性能。

基于上述实验结果与分析, ISEM的最终训练方案确定为: 在生成器对抗损失 l_G^2 中引入几何均值损失 l_{GM} , 在判别器损失 l_D 中引入均值绝对偏差损失 l_{MAD} 。

2.4 扩展实验

为验证ISEM在不同数据集上的泛化能力, 本文采用BOWS2^[31]的10 000张图像作为测试集, 并在嵌入率为0.4 bpp条件下进行安全性分析, 实验结果见表5。

从表5中的数据可以看出, ISEM能够显著

表5 在BOWS2数据集上不同方法抵抗隐写分析器的检测错误率对比

Table 5 Comparison of the detection error rates of different methods in resisting steganalysis attacks on the BOWS2 dataset

隐写方法	SRM ^[28]	Xu-Net ^[29]	Yedroudj-Net ^[30]	SRNet ^[24]	CovNet ^[25]	Zhu-Net ^[26]	平均值	提升值
HILL ^[9]	30.27%	31.08%	21.61%	21.48%	19.54%	29.45%	25.57%	
SMD-HILL	35.16%	37.41%	28.05%	23.5%	22.9%	36.08%	30.52%	4.95% ↑
CMD-HILL	37.38%	38.32%	28.89%	25.05%	24.17%	36.98%	31.8%	6.23% ↑
ISEM-HILL	38.6%	41.61%	36.64%	34.13%	34.34%	42.49%	37.97%	12.40% ↑
ASDL-GAN ^[10]	25.3%	31.24%	24.02%	17.21%	16.64%	30.11%	24.09%	
ISEM-ASDL	39.24%	41.88%	37.37%	36.48%	35.81%	42.5%	38.88%	14.79% ↑
UT-GAN ^[12]	33.62%	37.04%	31.74%	30.97%	26.74%	37.09%	32.87%	
ISEM-UT	40.06%	43.02%	37.88%	34.43%	35.24%	42.56%	38.87%	6.00% ↑
SPAR-RL ^[14]	33.83%	38.22%	31.09%	26.64%	26.05%	36.19%	32%	
ISEM-SPAR	39.39%	42.14%	36.87%	38.54%	35.01%	41.4%	38.89%	6.89% ↑
Steg-GMAN ^[15]	35.55%	36.74%	34.24%	30.2%	29.03%	37.5%	33.88%	
ISEM-Steg	39.93%	43.06%	40.89%	39.62%	37.17%	43.96%	40.77%	6.89% ↑
Huang-GAN ^[23]	37.39%	42.4%	46.48%	39.55%	36.55%	38.73%	40.18%	
ISEM-Huang	39.69%	42.6%	40.31%	49.8%	39.26%	43.69%	42.56%	2.38% ↑

注:黑色加粗数字表示模型增强前后的最高安全性能,蓝色加粗并带有符号↑数字表示隐写方法经过模型(隐写策略)增强后的平均提升值

提升对称隐写方法的安全性能。ISEM-HILL、ISEM-ASDL、ISEM-UT、ISEM-SPAR 和 ISEM-Steg 的平均安全性能分别提升 12.40%、14.79%、6.00%、6.89%、6.89%。在优化非对称隐写方法的非对称嵌入代价方面, ISEM-Huang 在抵抗隐写分析器 SRNet 检测上, 安全性能提升 10.25% (从 39.55% 提升至 49.80%)。与隐写增强策略 SMD-HILL 和 CMD-HILL 相比, ISEM-HILL 平均安全性能分别提升 7.45% 和 6.17%。

综上所述, 实验结果充分证明: 当现有隐写方法与本文提出的 ISEM 结合后, 安全性能得到进一步增强; 与其他隐写增强策略相比, ISEM 面对各类隐写分析器的检测时始终保持最优的安全性能表现。

3 结束语

本文提出一种基于生成对抗网络的图像隐写增强模型, 旨在突破传统对称隐写技术在安全性提升方面的瓶颈。该模型基于 GAN 架构, 通过轻量级双分支生成器与 3 个隐写分析网络集成的判别器设计, 实现了隐写安全性的显著提升。生成器首先学习现有隐写方法中载体图像与嵌入代价之间的映射关系, 并通过对抗训练生成更加安全的非对称嵌入代价, 从而提升隐写的安全性。此外, 本文设计了几何均值损失函数用于生成器, 并为判别器引入了均值绝对偏差损失函数, 以进一步提升网络的稳定性和训练效果。安全性评估实验结果表明, ISEM 在隐写安全性方面优于现有的对称和非对称隐写方法。消融实验进一步验证了 ISEM 设计的合理性和有效性。总体来说, ISEM 为提升图像隐写技术的安全性提供了一个有效的解决方案, 具有良好的应用潜力。

参考文献:

- [1] LI B, HE J H, HUANG J W, et al. A survey on image steganography and steganalysis[J]. *Journal of Information Hiding and Multimedia Signal Processing*, 2011, 2(2): 142-172.
- [2] 竺乐庆, 郭钰, 莫凌强, 等. DGANS: 基于双重生式对抗网络的稳健图像隐写模型[J]. *通信学报*, 2020, 41(1): 125-133.
- ZHU L Q, GUO Y, MO L Q, et al. DGANS: robustness image steganography model based on double GAN[J]. *Journal on Communications*, 2020, 41(1): 125-133.

- [3] 翟黎明, 嘉炬, 任魏翔, 等. 深度学习在图像隐写术与隐写分析领域中的研究进展[J]. *信息安全学报*, 2018, 3(6): 2-12.
- ZHAI L M, JIA J, REN W X, et al. Recent advances in deep learning for image steganography and steganalysis[J]. *Journal of Cyber Security*, 2018, 3(6): 2-12.
- [4] FRIDRICH J, FILLER T. Practical methods for minimizing embedding impact in steganography[C]//*Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents IX*. 2007: 650502.
- [5] 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.
- [6] KER A D, PEVNÝ T, KODOVSKÝ J, et al. The square root law of steganographic capacity[C]//*Proceedings of the 10th ACM Workshop on Multimedia and Security*. New York: ACM, 2008: 107-116.
- [7] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[C]//*Information Hiding*. Berlin: Springer, 2010: 161-177.
- [8] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//*Proceedings of the 2012 IEEE International Workshop on Information Forensics and Security (WIFS)*. Piscataway: IEEE Press, 2012: 234-239.
- [9] LI B, WANG M, HUANG J W, et al. A new cost function for spatial image steganography[C]//*Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*. Piscataway: IEEE Press, 2014: 4206-4210.
- [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, 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.
- [12] 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*, 2019, 15: 839-851.
- [13] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//*Proceedings of Medical Image Computing and Computer-Assisted Intervention*. Berlin: Springer, 2015: 234-241.
- [14] 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*, 2020, 16: 952-967.
- [15] HUANG D X, LUO W Q, LIU M L, et al. Steganography embedding cost learning with generative multi-adversarial network[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 19: 15-29.
- [16] LI B, WANG M, LI X L, et al. A strategy of clustering modification directions in spatial image steganography[J]. *IEEE Transactions on Information Forensics and Security*, 2015, 10(9): 1905-1917.
- [17] DENEMARK T, FRIDRICH J. Improving steganographic security by synchronizing the selection channel[C]//*Proceedings of the 3rd*

- ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2015: 5-14.
- [18] WANG Z C, LV J P, WEI Q D, et al. Distortion function for spatial image steganography based on the polarity of embedding change [C]//Proceedings of Digital Forensics and Watermarking. Berlin: Springer, 2017: 487-493.
- [19] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples[EB]. 2014.
- [20] 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.
- [21] MO H X, SONG T T, CHEN B L, et al. Enhancing JPEG steganography using iterative adversarial examples[C]//Proceedings of the 2019 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE Press, 2019: 1-6.
- [22] LIU M L, LUO W Q, ZHENG P J, et al. A new adversarial embedding method for enhancing image steganography[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4621-4634.
- [23] HUANG D X, LUO W Q, ZHENG P J, et al. Automatic asymmetric embedding cost learning via generative adversarial networks[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 8316-8326.
- [24] BOROUHAND 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.
- [25] DENG X Q, CHEN B L, LUO W Q, et al. Fast and effective global covariance pooling network for image steganalysis[C]//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2019: 230-234.
- [26] ZHANG R, ZHU F, LIU J Y, et al. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 1138-1150.
- [27] BAS P, FILLER T, PEVNÝ T. "Break Our Steganographic System": the ins and outs of organizing BOSS[C]//Information Hiding. Berlin: Springer, 2011: 59-70.
- [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] 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.
- [30] YEDROUDJ M, COMBY F, CHAUMONT M. Yedroudj-Net: an efficient CNN for spatial steganalysis[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 2092-2096.
- [31] BAS P, FURON T. Image database of BOWS-2[J]. Accessed, 2017, 20: 2016-2017.

[作者简介]



周赛星 (2001-), 女, 广东深圳人, 中山大学硕士生, 主要研究方向为图像隐写。



叶苗欣 (1999-), 女, 广东珠海人, 中山大学硕士生, 主要研究方向为图像隐写。



骆伟祺 (1980-), 男, 广东广州人, 博士, 中山大学教授、博士生导师, 主要研究方向为数字多媒体取证、图像隐写和图像隐写分析。