中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-17

论文引用格式:Li Rong, Zhang Yi, Zhang Mingliang, Liu Yanmei, Luo Xiangyang. Robust Image Steganography with Multi-Adversarial Guidance and Hash-Based Error Correction[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260049.(李荣, 张祎, 张明亮, 刘燕美, 罗向阳. 融合多对抗引导和哈希纠错的鲁棒图像隐写[J/OL]. 中国图象图形学报,XXXX:1-17. DOI: 10.11834/jig.260049.)[DOI:10.11834/jig.260049]

融合多对抗引导和哈希纠错的鲁棒图像隐写

李荣[1,2],张祎[1,2*],张明亮[1,2],刘燕美[1,2],罗向阳[1,2]

1. 信息工程大学,河南郑州　450001; 2. 河南省网络空间态势感知重点实验室,河南郑州　450001

摘　要:**目的**　图像隐写是信息隐藏技术的重要分支,其核心目标是将秘密信息嵌入载体图像中实现隐蔽通信,在军事通信与隐私保护等领域具有重要意义。近年来,基于编码-解码网络的鲁棒图像隐写方法通过端到端训练,在兼顾高容量、良好视觉质量与强鲁棒性方面取得显著进展。然而,现有方法多侧重于全局视觉质量优化,缺乏对局部空间特征的自适应能力,导致平滑区域易产生伪影、纹理区域细节丢失;同时,受限于纠错机制的能力,在复杂信道干扰下信息恢复性能仍显不足。为此,本文提出一种融合多对抗引导和哈希纠错的鲁棒图像隐写方法。**方法**　首先,将隐写分析网络引入对抗训练,利用其对平滑区域嵌入修改的高敏感性,引导秘密信息嵌入低敏感的纹理区域,从而降低嵌入扰动的可感知性。其次,设计双级纠错机制:在发送端,先通过BCH码对信息进行初步校正;针对BCH无法纠正的错误,则根据纠错位置与自然图像哈希的映射,从共享图像库中检索对应图像,并与载密图像同时传输。接收端则利用哈希生成模型计算哈希值以定位错误位置,并与BCH纠错协同恢复信息。同时,为确保哈希映射的唯一性,提出一种无歧义的哈希映射方案,从而系统性地保障信息在干扰信道中恢复的准确性。**结果**　实验结果显示,与当前先进方法相比,所提方法在图像质量上显著提升,PSNR和MS-SSIM分别提高5.88%和3.13%,LPIPS降低31.25%,表明视觉不可感知性更高。在鲁棒性方面,对多数常见非几何攻击,秘密信息提取正确率达100%;特别地,在"旋转60°-缩放-裁剪"复合攻击下,提取正确率提升30.88%。**结论**　该方法兼顾高视觉质量与强鲁棒性,为构建实用化隐写系统提供了有效新思路。

关键词:隐蔽通信;鲁棒隐写;图像隐写;纠错;哈希映射

Robust Image Steganography with Multi-Adversarial Guidance and Hash-Based Error Correction

Li Rong[1,2], Zhang Yi[1,2*], Zhang Mingliang[1,2], Liu Yanmei[1,2], Luo Xiangyang[1,2]

1. Information Engineering University, Zhengzhou 450001, China; 2. Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou 450001, China

Abstract: **Objective** Image steganography is a vital branch of information hiding technology, whose core objective is to embed secret information into cover images for covert communication, and it holds great significance in fields such as military communication and privacy protection. In recent years, robust image steganography methods based on encoder-

收稿日期:2026-01-22;**修回日期**:2026-03-16

* **通信作者**:张祎 tzyy4001@sina.com

基金项目:河南省优秀青年科学基金项目(252300421233);国家自然科学基金项目(62172435, 62202495, U23A20305);中原学者项目(254000510007)

Supported by: Natural Science Foundation for Excellent Young Scholars of Henan Province (252300421233); National Natural Science Foundation of China (62172435, 62202495, U23A20305); Zhongyuan Scholars Program (254000510007)

decoder networks have achieved remarkable progress in balancing high capacity, strong robustness and favorable visual quality through end-to-end training. However, existing methods mostly focus on global visual quality optimization while lacking adaptive capacity for local spatial features, which leads to artifacts in smooth regions and detail loss in textured regions. Meanwhile, constrained by the capability of error correction mechanisms, the performance of information recovery under complex channel interference remains inadequate. To address these issues, this paper proposes a robust image steganography with multi-adversarial guidance and hash-based error correction. **Method** First, on the basis of the TransStego architecture, this scheme introduces a steganalyzer to jointly construct a multi-adversarial guidance mechanism together with a discriminator. Through adversarial training, the encoder leverages the steganalyzer's high sensitivity to smooth regions, thereby adaptively embedding secret information into texture regions with low perceptual sensitivity. Meanwhile, the discriminator ensures global visual fidelity, whereas the steganalyzer suppresses local statistical anomalies. The collaborative optimization of the two modules significantly reduces embedding artifacts and improves both the visual quality and statistical invisibility of stego images. Second, in terms of error correction, a dual-level error correction mechanism based on BCH codes and hash mapping is designed. At the sender side, the secret information is first preliminarily corrected using BCH codes. If there remain uncorrectable errors beyond the capability of BCH codes, the error location information is encoded into a robust hash sequence by means of a pre-synchronized hash generation model and a shared image database. This hash sequence is transmitted together with the stego image after matching the corresponding natural image in the database. In addition, to avoid hash collisions, an unambiguous hash mapping scheme is designed in this paper: the hash generation model is employed to generate initial hash sequences for all images in the database and perform global ascending sorting; for subsets of images with hash collisions, secondary ascending sorting is conducted based on average pixel values to eliminate collisions, while the remaining images maintain their original order. All images are then integrated into a unified sequence, assigned consecutive unique IDs, and these IDs are mapped to binary sequences as the final hash codes. At the receiver side, accurate positioning is achieved through hash values and average pixel values, and the secret information is recovered by combining with BCH codes, thus realizing accurate information recovery in interference channels. **Results** We evaluate the method in terms of imperceptibility, robustness, security, and ablation. For imperceptibility, it outperforms TransStego with a PSNR increase of 5.88%, an MS-SSIM improvement of 3.13%, and an LPIPS reduction of 31.25%. Under common non-geometric attacks including Gaussian blur, median filtering, JPEG compression, and image scaling, the method achieves near-perfect secret extraction accuracy. Under challenging geometric attacks such as affine transformation, rotation, rotation cropping, and rotation scaling cropping, it maintains strong recovery despite pixel loss and attains an average extraction accuracy of 79.2%. This exceeds the performance of four representative baselines: TransStego at 52.97%, StegaStamp at 54.43%, RoSteALS at 47.32%, and FNNS-R at 49.97%, yielding an average gain of 27.78 percentage points. Notably, under the severe composite attack of 60-degree rotation followed by scaling and cropping, the method improves extraction accuracy by 30.88% over TransStego. In security evaluation against the advanced steganalyzer UCNet, it achieves the lowest and most stable detection rate, outperforming PRIS, FNNS-R, and TransStego. Ablation studies show that the two-stage error correction mechanism boosts robustness, while steganalytic guidance enhances both imperceptibility and security. Their combination yields the highest extraction accuracy across diverse attacks, with performance gains exceeding 30% in geometric scenarios. **Conclusion** This method achieves the synergistic optimization of high visual fidelity and strong robustness under complex interferences through steganalysis-guided local adaptive embedding and the BCH-hash dual-level error correction mechanism. It takes the steganalyzer as the embedding guidance source to direct secret information to be preferentially embedded into anti-disturbance texture regions, and avoids anomalies and reduces distortions via gradient feedback; the dual-level error correction mechanism improves information extraction accuracy. Experiments demonstrate that this design significantly optimizes the visual quality and robustness of stego images, providing a new paradigm for constructing practical and highly reliable robust image steganography systems.

Key words: covert communication; robust steganography; image steganography; error correction; hash mapping

0 引 言

图像隐写术是指将秘密信息嵌入到数字图像中，以实现隐蔽通信的技术。其核心在于设计安全可靠的隐写算法，避免秘密信息被第三方察觉或分析破坏，从而保障安全传输（Li等，2011）。传统隐写方法通常假设载密图像在无损信道中传输，然而在实际网络环境中，尤其是通过社交平台传输时，图像常会遭受JPEG压缩、尺寸缩放等处理，导致嵌入信息被破坏，信息提取准确率显著下降（Fridrich等，2007）。为此，研究者提出鲁棒隐写技术，旨在保证隐写图像视觉质量的基础上，提升嵌入信息抵抗图像处理操作的能力，如图1所示。因此，载密图像的鲁棒性已成为衡量其在真实网络环境中可靠性和实用性的关键指标。如何平衡鲁棒性与隐蔽性仍是当前研究的核心挑战。

图1 社交网络平台有损处理下的鲁棒图像隐写示意图

Fig. 1 Schematic diagram of robust image steganography under lossy processing on social network platforms

为应对社交网络中有损信道的挑战，鲁棒图像隐写方法需要同时保证隐蔽性与鲁棒性，现有技术主要分为基于人工设计和基于深度学习两类。基于人工设计的方法通常遵循“鲁棒区域选择+纠错编码（如RS码）（Gore，1969）+最小失真编码（Syndrome-Trellis Codes，STC）（Filler等，2011）”框架（Zhang等，2025），通过优化嵌入代价函数，优先选择低代价区域嵌入信息，并融合纠错码与最小失真编码，确保在压缩、加噪、滤波等图像处理攻击下仍能可靠恢复信息（Zhang等，2022）。该类方法按抗攻击能力可分为两类：一类旨在对抗特定类型的攻击，以JPEG压缩为代表。Zhang等人（2015）提出首个抗JPEG重压缩的鲁棒隐写方法，选择中频系数作为鲁棒载体，显著提升了安全性，但嵌入容量有限且仅支持二元STC编码。随后，Yu等人（2020）扩展了嵌入域并引入了广义抖动调制，能与三元STC编码结合使用，从而进一步提升了安全性。然而，此类方法的鲁棒性适用范围有限，多数方案仅适用于低质量图像，难以应对高质量因子重压缩。为此，Zeng等人（2023）采用精确抖动调制与补偿机制以增强鲁棒性，但仍存在迭代过程耗时，且在极端质量因子下性能不足的问题。后续的研究通过扩展鲁棒嵌入域（Duan等，2023；Yao等，2024）与优化纠错编码（Cheng等，2025）等途径，进一步提升了整体性能。另一类方法旨在提升对多种图像处理操作的鲁棒性。Zhang等（2019）提出基于系数差值构建多重鲁棒载体，兼具鲁棒性与抗检测能力；其后续工作（Zhang等，2021）利用量化取整原理，实现对压缩与缩放攻击的鲁棒嵌入，并在多嵌入率下实现100%信息提取。除上述框架外，也有研究采用不同思路，如Rajan等人（2025）提出一种通用的鲁棒图像隐写框架，通过猎豹优化器动态选择像素洗牌参数，结合RC4加密与Hash-LSB嵌入，有效提升不可感知性与抗攻击能力。然而，随着深度学习隐写分析技术的发展，基于人工设计的方法因依赖专家知识且泛化能力有限，面临较高的被检测风险。

为应对这一挑战，基于深度学习的鲁棒图像隐写方法日益受到关注。该类方法以数据驱动为核心，通过大量数据训练神经网络，使其端到端地学习秘密信息的嵌入与恢复，无需依赖人工先验知识（Fu等，2020）。根据载密图像的生成机制，现有方法可分为选择式、生成式和修改式三类。其中，为提升抗检测能力，研究者提出不修改载体的选择式鲁棒隐写方法，该类方法利用深度神经网络提取图像的高维语义特征（如深度哈希、目标属性），构建与秘密信息的映射关系（Tang等，2024），并从图像库中

筛选匹配载密图像直接传输。接收端基于相同特征恢复信息。例如，Meng 等人(2022)利用端到端深度卷积网络从自然图像生成固定长度的鲁棒哈希序列，并以该哈希值为索引映射秘密信息；Li 等人(2023)通过 DWT 提取图像低频特征并构建特征矩阵完成秘密信息映射。此类方法因未修改载体图像而具有较强的抗隐写分析能力，但嵌入容量较低。为提升嵌入容量，生成式隐写方法应运而生，该类方法以秘密信息为驱动生成载密图像。例如，Yang 等人(2023)通过逆变换采样将信息映射至生成对抗网络(Generative Adversarial Network，GAN)(Goodfellow 等，2014)的潜在空间，在提高容量的同时结合可微分噪声层与梯度下降优化鲁棒性，但受限于 GAN 的生成能力，视觉质量仍有不足。相比之下，基于扩散模型(Ho 等，2020)的方法能生成更自然的图像，如 Peng 等人(2024)基于潜在空间扩散模型(Rombach 等，2022)，通过将隐空间重建误差划分为多个区间，并建立其与秘密信息的映射关系来实现隐写；Hu 等人(2024)则采用了双密钥策略，将秘密信息编码为符合高斯分布的噪声，直接用于预训练扩散模型生成载密图像。虽然这类方法生成图像质量较高，但容易受到干扰，鲁棒性和生成效率有待进一步提升。修改式方法通过在原始载体上施加不可见扰动嵌入秘密信息，在嵌入容量、视觉质量与鲁棒性之间实现良好平衡，通常采用编码器-解码器架构，并结合可微分信道模拟实现端到端训练。根据其网络设计侧重点，现有研究主要分为两类：一是聚焦于噪声层设计，旨在通过模拟复杂信道失真来增强抗干扰能力。Zhu 等人(2018)首次引入可微分的 JPEG 压缩模拟，但其对其他多种失真的适应能力仍然有限；Tancik 等人(2020)进一步模拟多种现实失真，并结合 BCH(Bose - Chaudhuri - Hocquenghem)码(Forney，1965)与空间变换网络(Spatial Transformer Network，STN)(Jaderberg 等，2015)，虽提升了鲁棒性，却也导致视觉质量下降。二是优化嵌入与提取过程的网络机制，旨在通过策略与结构改进提升整体性能。Kishore 等人(2022)优化载体扰动以改善视觉质量，但其随机解码器导致提取稳定性下降；Bui 等人(2023)改为在冻结自编码器的潜在空间中嵌入，虽提升了训练效率，但图像质量受限于解码器重建能力。此外，Yang 等人(2024)采用可逆网络实现高保真恢复，却带来较高计算开销；Xiao 等人(2025)将自注意力机制引入特征建模以优化载密图像视觉质量，融合对抗训练、空间几何校正策略，结合随机遮挡与多类型图像扰动的数据增强手段提升鲁棒性，但其在高强度与复合攻击下的鲁棒性仍显不足。

以上分析表明，修改式方法通过端到端训练可自适应学习针对不同信道干扰的鲁棒嵌入策略，在嵌入容量、视觉质量与鲁棒性之间取得了较好平衡。然而，现有方法仍存在两方面局限：一是现有多数方案的对抗训练聚焦全局视觉质量，忽视嵌入扰动与局部纹理的协调，导致载密图像在细节区域不自然；二是在多种信道攻击下，信息提取准确率显著下降，而现有纠错机制较为单一，难以保障高强度干扰下的可靠恢复。针对上述问题，本文提出一种融合多对抗引导和哈希纠错的鲁棒图像隐写方法。具体而言，在修改式鲁棒图像隐写方法框架下，引入隐写分析器参与对抗训练，以引导编码器将秘密信息优先嵌入图像对修改不敏感的区域，进一步提升载密图像的视觉自然性；同时，受选择式鲁棒图像隐写方法的启发，利用自然图像传递少量纠错信息，设计融合 BCH 纠错码与哈希生成模型的双级纠错策略，并针对现有方法中多图生成同一哈希序列、影响信息恢复可靠性的歧义问题，设计无歧义哈希映射方案。

本文工作的主要贡献如下：1)提出一种隐写分析网络引导的嵌入机制，利用其对嵌入修改的敏感性构建多元对抗训练，有效抑制载密图像在局部区域的视觉失真。2)提出一种融合 BCH 纠错码与哈希生成模型的双级纠错机制，能够同时处理随机误码与结构性失真，显著提升秘密信息的提取准确率。3)设计了一种无歧义哈希映射方案，通过初始哈希排序与冲突图像平均像素值的两级排序生成唯一哈希表示，解决了哈希冲突问题，确保秘密信息恢复准确。

2 提出的方法

为提升载密图像的局部视觉质量并增强其在社交网络有损信道下的鲁棒性，本文提出一种融合多对抗引导和哈希纠错的鲁棒图像隐写方法。本节首先介绍方法的整体流程，接着阐述模型结构、损失函数与训练策略，并重点探讨基于哈希映射的双级纠错机制的设计与实现。

2.1 概述

本文提出的端到端的鲁棒隐写框架如图2所示，由训练端的三元对抗模块和应用端的双级纠错模块组成。三元对抗模块学习嵌入提取策略，而双级纠错模块修复信道攻击导致的比特错误。二者协同工作，预先约定图像传输与接收的机制，保障发送端到接收端的可靠传输。判别器与隐写分析器仅

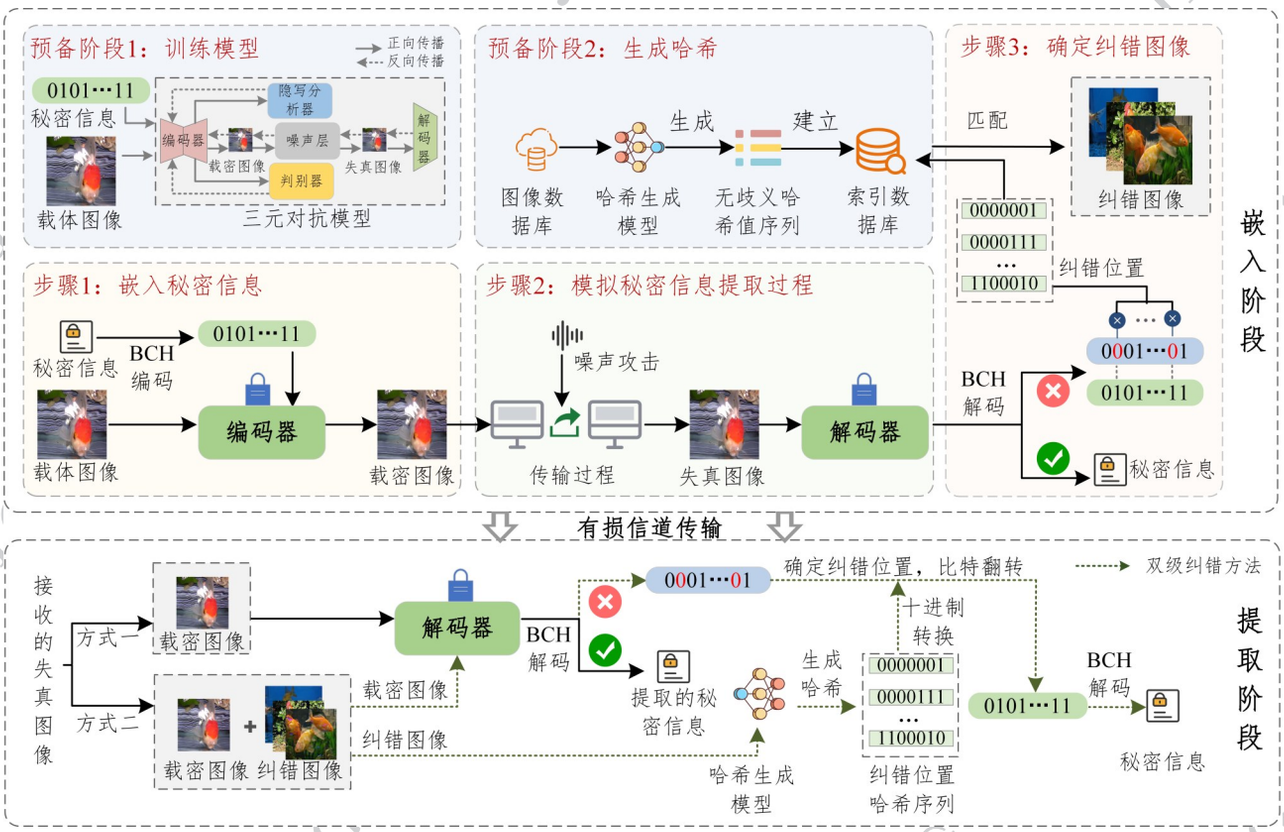


图2 方法框架图

Fig. 2 Overview of the framework

参与训练过程，不介入实际的隐写操作。训练完成后，纠错模块可进一步提升传输可靠性。

2.2 模型结构

本框架基于编码器-解码器-判别器对抗架构，引入了隐写分器模块以增强安全性。整体网络包含四个核心组件：编码器（作为生成器）$\boldsymbol{G}$、判别器$\boldsymbol{D}$、隐写分析器$\boldsymbol{S}$、解码器（作为提取器）$\boldsymbol{E}$。

2.2.1 编码器

编码器是框架的核心嵌入模块，负责将二进制秘密信息$\boldsymbol{M}$嵌入到载体图像$\boldsymbol{I}_{\mathrm{c}}$，生成载密图像$\boldsymbol{I}_{\mathrm{s}}$。该模块采用Transformer架构，在保持视觉质量的同时实现隐蔽嵌入。从整体架构上看，编码器基于ViT（Vision Transformer）（Dosovitskiy 等，2020）的类U-Net（Ronneberger 等，2015）设计，通过跳跃连接融合多尺度特征，并利用ViT模块捕获全局上下文，引导信息自适应嵌入纹理复杂区域，其网络结构如图3所示。

为实现上述嵌入过程，该编码器的前向传播处理流程可分为三个阶段：首先，在信息融合阶段，将长度为k的$\boldsymbol{M}\in\{0,1\}^k$经全连接层投影为特征张量$\boldsymbol{T}_{\mathrm{info}}\in\boldsymbol{R}^{H\times W\times C}$与载体图像张量$\boldsymbol{I}_{\mathrm{c}}\in\boldsymbol{R}^{H\times W\times 3}$沿通道维度进行拼接为混合张量$\boldsymbol{X}$，实现秘密信息与图像内容的初步融合；其次，在特征提取阶段，对$\boldsymbol{X}$进行卷积层下采样得到深层局部特征$\boldsymbol{F}_{\mathrm{conv}}=f_{\mathrm{conv}}(\boldsymbol{X})$，

再输入ViT模块，再经多头自注意力机制（Vaswani 等，2017）提取全局上下文特征$\boldsymbol{F}_{\mathrm{attn}}$；最后，在图像重建阶段，将$\boldsymbol{F}_{\mathrm{attn}}$经转置卷积层上采样，结合跳跃连接融合的多尺度特征，逐步恢复图像空间分辨率，输出残差图像$\boldsymbol{I}_{\mathrm{r}}$，并通过$\boldsymbol{I}_{\mathrm{s}}=\boldsymbol{I}_{\mathrm{c}}+\boldsymbol{I}_{\mathrm{r}}$生成载密图像。该过程整体可视作$\boldsymbol{G}$对输入$(\boldsymbol{I}_{\mathrm{c}},\boldsymbol{M})$的映射，即

$$\boldsymbol{I}_{\mathrm{s}}=G(\boldsymbol{I}_{\mathrm{c}},\boldsymbol{M}) \tag{1}$$

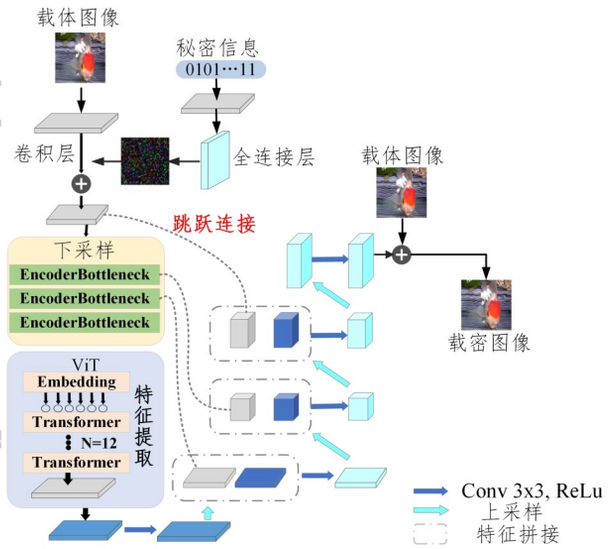


图3　编码器网络模型图

Fig. 3　The network structure diagram of the encoder network

2.2.2　判别器

判别器采用WGAN(Wasserstein GAN)(Arjovsky等,2017)框架,其输出为无界的实值分数,用于估计I_s与I_c所服从分布之间Wasserstein距离。在对抗训练中,编码器通过最小化判别器输出的该距离估计值,促使I_s的分布逼近I_c的分布,从而提升载密图像的视觉保真度。该距离反映了将一个分布转化为另一个分布所需的最小"代价",定义如下:

$$W(P_{I_c},P_{I_s})=\inf_{\gamma\in\Pi(P_{I_c},P_{I_s})}\mathbb{E}_{(x,y)\sim\gamma}\left[\|x-y\|\right] \tag{2}$$

式中,P_{I_c}与P_{I_s}分别表示I_c和I_s的概率分布;inf表示在所有可能的联合分布γ中,期望传输代价的下确界;x与y为从对应分布采样的图像样本;$\Pi(P_{I_c},P_{I_s})$是所有以P_{I_c}和P_{I_s}为边缘分布的联合分布　的集合;$\|\cdot\|$表示像素空间中的距离度量。

2.2.3　隐写分析器

隐写分析器对图像纹理平滑区域的嵌入修改敏感,能够识别嵌入扰动与自然图像的分布偏差,进而分辨载体图像I_c和与载密图像I_s。为此,本文将其引入对抗训练,与判别器共同构成多元对抗约束,以提升载密图像的视觉质量、降低嵌入扰动的可感知性。二者均以I_c和I_s为输入,但优化目标各有侧重,前者聚焦局部统计一致性,后者侧重全局视觉保真度。训练过程中,编码器通过最小化被隐写分析器识别的概率,借助其敏感性引导秘密信息嵌入到纹理区域,最终生成局部纹理协调、统计特性贴近自然图像的载密图像,有效减少细节伪影,实现视觉质量的提升。

2.2.4　解码器

解码器从可能失真的载密图像中恢复秘密信息。该网络首先通过STN对输入图像进行自适应几何对齐,学习仿射变换以缓解旋转、缩放、平移等几何失真;随后通过多层卷积神经网络提取特征;最终经Sigmoid激活函数输出长度为k的向量$\boldsymbol{M}_{\text{out}}\in(0,1)^k$,其中每个分量表示对应秘密比特为1的概率估计。

2.3　纠错模块

本方法提出一种结合BCH纠错码与哈希生成模型的双级纠错机制,通过将纠错位置编码为图像的鲁棒哈希序列,并以对应自然图像为载体传递纠错位置信息,构建主动纠错通道。

2.3.1　双级纠错机制

本研究设计的双级纠错框架,通过"基础纠错+精准修正"的协同机制消除信道失真比特错误,解决超出BCH纠错能力的信道失真时而导致的秘密信息提取失败的问题。双级纠错框架的完整流程如图2所示。其核心逻辑可定义如下:

在双级纠错框架的核心流程中,发送方首先对遭受攻击后的载密图像进行解码,然后使用BCH进行预纠错。若超出BCH纠错能力,则触发基于哈希生成模型的纠错机制。发送方与接收方需预先同步哈希生成模型、共享图像库、随机种子及索引规则。其中,一级纠错为BCH基础修正,将解码器输出的二进制串输入BCH解码器校正,校正与UTF-8解码均成功则直接输出秘密信息,失败则触发二级纠错;二级纠错为哈希映射精准修正,对接收的纠错图像生成唯一哈希值以定位错误位置,随后翻转错误位置的比特,将修正后的二进制串再次输入BCH解码,恢复正确的秘密信息。

2.3.2　无歧义哈希映射

在基于哈希映射的纠错机制中,哈希序列通过固定长度的二进制串建立纠错位置与自然图像的映射,以确保接收端能准确提取纠错信息。为解决哈希映射中的冲突并确保每个纠错位置与自然图像的唯一映射,本研究提出了一种无歧义哈希映射方案,示意图如图4所示。

首先,通过哈希生成模型为图像数据库中的所有图像生成初始哈希序列,并依据哈希序列值进行

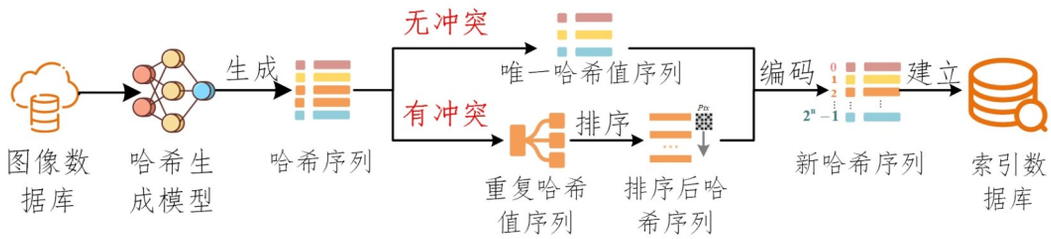


图4　无歧义哈希映射方案

Fig. 4　Unambiguous hash mapping scheme

首次全局升序排序。接着，对哈希值冲突的图像子集，以图像平均像素值(Pix)为依据进行二次升序排列，以消除哈希冲突；其余图像保留其初次排序结果。最后，将无冲突图像与二次排序后的冲突图像整合为统一序列，从0开始为每个图像分配唯一的连续整数ID，并将该ID映射为二进制序列作为最终哈希编码。该方法通过二次排序确保了每个图像均可获得唯一的哈希表示，有效解决了哈希值冲突问题。为便于理解，现给出哈希二次编码的一个示例，如图5所示。

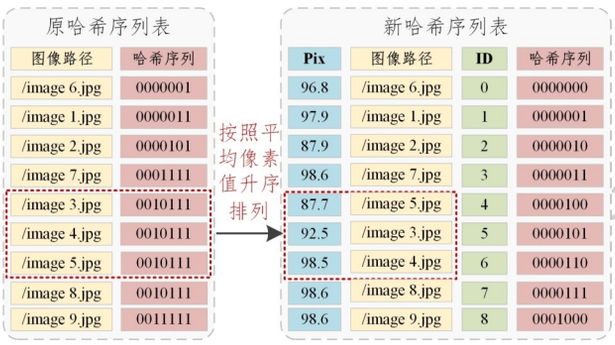


图5　哈希二次编码示意图

Fig. 5　Example of repeated hashing with double encoding

总之，发送方与接收方通过预同步的图像数据库、哈希生成模型及无歧义哈希映射规则，结合哈希值与图像像素均值，实现对错误位置的精准定位。

2.4　损失函数设计

训练过程中，编码器$\boldsymbol{G}$、判别器$\boldsymbol{D}$、隐写分析器$\boldsymbol{S}$与解码器$\boldsymbol{E}$协同优化。各模块均有独立的损失函数：编码器和解码器联合最小化综合损失以提升嵌入与提取性能；判别器最大化其判别能力；隐写分析器最小化分类误差以增强检测能力。整体目标由图像失真损失、秘密信息损失与对抗损失三部分构成，具体如下：

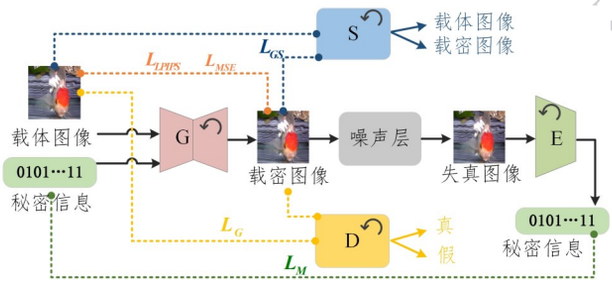


图6　基于三元对抗网络的鲁棒图像隐写框架

Fig. 6　Robust steganography framework based on Ternary adversarial network

2.4.1　图像失真损失

从像素级和视觉感知两方面来保障图像质量，像素级损失采用L2损失(MSE，载体图像与载密图像之间差的平方的平均值)，公式如下：

$$L_{\mathrm{MSE}}=\frac{\|\boldsymbol{I}_c-\boldsymbol{I}_s\|_2^2}{C\times H\times W} \tag{3}$$

式中，$\|\cdot\|_2^2$表示L_2范数的平方，用于计算差异的平方和，C、H和W分别表示图像通道数、长、宽。

感知损失采用LPIPS(Learned Perceptual Image Patch Similarity)(Zhang等，2018)，将$\boldsymbol{I}_s$与$\boldsymbol{I}_c$输入到预训练深度网络，提取多层级特征，并计算加权的特征空间距离：

$$L_{\mathrm{LPIPS}}(\boldsymbol{I}_s,\boldsymbol{I}_c)=\sum_{f\in F}w_f\left\|\phi_f(\boldsymbol{I}_s)-\phi_f(\boldsymbol{I}_c)\right\|_2^2 \tag{4}$$

式中，F表示从预训练网络中选取的特征层数量，$\phi_f(\cdot)$表示第f层的特征图，w_f为对应层的通道权重(本文采用官方预训练权重，固定不更新)。

2.4.2　秘密信息损失

由于秘密信息是二进制比特流，因此使用二分类交叉熵损失函数来衡量原始秘密信息与提取秘密信息的一致性，定义如下：

$$L_{\mathrm{M}}=-\frac{1}{k}\sum_{i=1}^{k}[\boldsymbol{M}_{\mathrm{in},i}\log\boldsymbol{M}_{\mathrm{out},i}+(1-\boldsymbol{M}_{\mathrm{in},i})\log(1-\boldsymbol{M}_{\mathrm{out},i})] \tag{5}$$

式中，$\boldsymbol{M}_{\text{in}} \in \{0,1\}^k$ 和 $\boldsymbol{M}_{\text{out}} \in \{0,1\}^k$ 分别是原始与解码消息，k 为秘密信息长度。

2.4.3 对抗损失

为生成视觉质量更高的载密图像，本方法通过 $\boldsymbol{G}$、$\boldsymbol{D}$ 和 $\boldsymbol{S}$ 之间的多重对抗训练实现模型稳定收敛。对抗损失由判别器对抗损失和隐写分析对抗损失两部分组成。

（1） 判别器损失

在WGAN框架中，$\boldsymbol{D}$ 被训练来估计 $\boldsymbol{I}_{\text{c}}$ 分布 P_{I_c} 与 $\boldsymbol{I}_{\text{s}}$ 分布 P_{I_s} 之间的Wasserstein距离。$\boldsymbol{D}$ 为输入图像输出一个实值分数，期望对 $\boldsymbol{I}_{\text{c}}$ 输出较高分数，对 $\boldsymbol{I}_{\text{s}}$ 输出较低分数。因此，$\boldsymbol{D}$ 的损失函数定义为：

$$L_{\text{D}} = -(\boldsymbol{D}(\boldsymbol{I}_{\text{c}}) - \boldsymbol{D}(\boldsymbol{I}_{\text{s}})) \tag{6}$$

通过最小化 L_D，$\boldsymbol{D}$ 被优化以增大 $\boldsymbol{D}(\boldsymbol{I}_{\text{c}})$ 与降低 $\boldsymbol{D}(\boldsymbol{I}_{\text{s}})$，从而拉大两者的分数差 $\boldsymbol{D}(\boldsymbol{I}_{\text{c}}) - \boldsymbol{D}(\boldsymbol{I}_{\text{s}})$，以增强对分布 P_{I_c} 和 P_{I_s} 的区分能力。编码器 $\boldsymbol{G}$ 则通过最小化以下对抗损失来提升 $\boldsymbol{I}_{\text{s}}$ 的 $\boldsymbol{D}$ 评分，以欺骗 $\boldsymbol{D}$：

$$L_{\text{G}} = -\boldsymbol{D}(\boldsymbol{I}_{\text{s}}) \tag{7}$$

最小化 L_G 即最大化 $\boldsymbol{D}(\boldsymbol{I}_{\text{s}})$，从而驱动 P_{I_s} 向分布 P_{I_c} 逼近，实现视觉不可感知性。

（2） 隐写分析器损失

隐写分析器的损失包括自身分类损失与编码器的对抗损失。其中，$\boldsymbol{I}_{\text{c}}$ 的标签为 $y_{\text{c}} = 0$，$\boldsymbol{I}_{\text{s}}$ 的标签为 $y_{\text{s}} = 1$。则分类损失分别为：

$$L_{\text{steg-c}} = \text{CrossEntropy}(\boldsymbol{S}(\boldsymbol{I}_{\text{c}}), y_c) \tag{8}$$

$$L_{\text{steg-s}} = \text{CrossEntropy}(\boldsymbol{S}(\boldsymbol{I}_{\text{s}}), y_s) \tag{9}$$

总分类损失为二者均值：

$$L_{\text{steg}} = \frac{1}{2}\left(L_{\text{steg\_c}} + L_{\text{steg\_s}}\right) \tag{10}$$

$\boldsymbol{S}$ 的训练目标是最大化分类准确率，等价于最小化 L_{steg}，期望形式可表示为：

$$\min_{S} L_{\text{steg}} = \mathbb{E}_{I_c \sim P_{I_c}}[-\log S_0(\boldsymbol{I}_{\text{c}})] + \mathbb{E}_{I_s \sim P_{I_s}}[-\log S_1(\boldsymbol{I}_{\text{s}})] \tag{11}$$

式中，$S_0(\cdot)$、$S_1(\cdot)$ 分别为 $\boldsymbol{S}$ 输出“载体/载密图像”的概率。

编码器针对 $\boldsymbol{S}$ 的对抗损失 L_{GS}，目标为使 $\boldsymbol{I}_{\text{s}}$ 被误判为原始图像：

$$\min_{G} L_{\text{GS}} = -\mathbb{E}_{I_c \sim P_{I_c}}[\log S_0(\boldsymbol{I}_{\text{s}})] \tag{12}$$

该损失与 $\text{CrossEntropy}(\boldsymbol{S}(\boldsymbol{I}_{\text{s}}), y_{\text{c}})$ 等价，通过推动 $S_0(\boldsymbol{I}_{\text{s}}) \to 1$，迫使编码器将秘密信息隐藏到统计不可检测区域。

（3） 编码器的对抗目标

编码器最小化对抗损失 $L_{\text{adv-total}}$，使视觉分布趋近一致，且保证统计特性无法被识别。

$$\begin{aligned}\min_{G} L_{\text{adv-total}} &= \lambda_{\text{G}} \cdot L_{\text{G}} + \lambda_{\text{GS}} \cdot L_{\text{GS}} \\ &= -\lambda_{\text{G}} \cdot \mathbb{E}_{I_c \sim P_{I_c}}[\boldsymbol{D}(\boldsymbol{I}_{\text{s}})] \\ &\quad -\lambda_{\text{GS}} \cdot \mathbb{E}_{I_c \sim P_{I_c}}[\log S_0(\boldsymbol{I}_{\text{s}})]\end{aligned} \tag{13}$$

式中，λ_{G}、λ_{GS} 分别为判别器、隐写分析器对抗损失权重，控制优化优先级。

2.4.4 总损失函数

编码器总损失由图像像素级损失、感知损失、两类对抗损失、秘密信息损失构成：

$$\begin{aligned}Loss &= \lambda_{\text{MSE}} L_{\text{MSE}} + \lambda_{\text{LPIPS}} L_{\text{LPIPS}} \\ &\quad + \lambda_{\text{G}} L_{\text{G}} + \lambda_{GS} L_{\text{GS}} + \lambda_{\text{M}} L_{\text{M}}\end{aligned} \tag{14}$$

式中，λ_{MSE}、λ_{LPIPS}、λ_{G}、λ_{GS}、λ_{M} 为各损失权重，分别控制不同损失项的重要性。

2.5 训练策略

为避免多任务训练初期的优化冲突，本文采用交替优化策略协调 $\boldsymbol{G}$、$\boldsymbol{D}$、$\boldsymbol{S}$ 与 $\boldsymbol{E}$ 的联合训练：首先聚焦秘密信息的嵌入与提取，待其初步收敛后，再引入图像保真与对抗约束进行协同优化。整体训练流程见算法1。

训练过程分为两个阶段：

（1）嵌入与提取预训练阶段（前 X 步）：通过 L_M

算法 1 分阶段交替训练策略

输入：载体图像 $\boldsymbol{I}_{\text{c}}$，秘密信息 $\boldsymbol{M}$

输出：训练完成的模型 $\{\boldsymbol{G},\boldsymbol{D},\boldsymbol{S},\boldsymbol{E}\}$

```
FOR t = 1 to X DO
    IF t ≤ X THEN
        计算 L_M = L_M(E(G(I_c, M)), M);
        通过 ∇_{G,E} L_M 更新 G 和 E;
    ELSE
        // 阶段一：更新 D 与 S
        计算 L_D 和 L_steg;
        通过 ∇_D L_D 更新 D;
        通过 ∇_S L_S 更新 S;
        // 阶段二：更新 G 与 E
        计算总损失 Loss;
        通过 ∇_{G,E} Loss 更新 G 和 E;
    END IF
END FOR
```

更新$\boldsymbol{G}$与$\boldsymbol{E}$参数，不使用图像保真或对抗损失，以快速建立基本的嵌入与提取能力。

(2)联合对抗训练阶段(第$X+1$步起)：引入$\boldsymbol{D}$与$\boldsymbol{S}$，通过对抗训练机制引导$\boldsymbol{G}$将秘密信息嵌入到视觉不敏感区域，从而在保持图像质量的同时提升隐蔽性与鲁棒性。

具体地，每步迭代先固定$\boldsymbol{G}$，分别用$\nabla_{\mathrm{D}}L_{\mathrm{D}}$和$\nabla_{\mathrm{S}}L_{\mathrm{S}}$更新$\boldsymbol{D}$与$\boldsymbol{S}$；再固定$\boldsymbol{D}$和$\boldsymbol{S}$，用$\nabla_{\mathrm{G,E}}Loss$更新$\boldsymbol{G}$与$\boldsymbol{E}$。该交替更新策略可缓解梯度冲突，促进各模块稳定收敛。

3 实验结果与分析

本节将通过实验对所提方法的综合性能进行验证与评估。实验部分主要包括以下五个方面：实验设置、不可感知性测试、鲁棒性测试、安全性验证以及消融实验。

3.1 实验设置

3.1.1 数据集和模型

本文模型基于Python 3.8.18与PyTorch深度学习框架实现，采用CUDA 11.4，在NVIDIA RTX A6000(48GB显存)显卡上训练。实验采用ImageNet-1K(ILSVRC2012)数据集，该数据集包含超120万张图像、涵盖1000个类别，本研究从中随机采样30000张图像作为训练集，1000张图像作为测试集，且所有图像在训练前统一处理为400×400的尺寸。实验采用UCNet隐写分析器(Wei等，2022)；模型训练的核心参数配置如下：编码器、解码器、判别器的学习率均设置为1e-5，隐写分析器的学习率设置为1e-3；批处理大小设为8；损失权重λ_{MSE}、λ_{LPIPS}、λ_{G}、λ_{GS}、λ_{M}分别默认设为：2、1.5、0.5、0.01、1.5；训练步数X设置为2000。

3.1.2 评价指标

本文从视觉质量、鲁棒性和安全性三个维度对所提方法进行综合评估。具体采用的量化指标如下。

(1) 视觉质量

采用峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)、多尺度结构相似性指数(Multi-scale Structural Similarity Index，MS-SSIM)(Wang等，2003)和学习型感知图像相似度(LPIPS)三种指标评估视觉质量。

1) MS-SSIM

MS-SSIM是对结构相似性指数(SSIM)的多尺度扩展，通过高斯金字塔在多个分辨率下分解图像，在最低分辨率层保留亮度相似性，在其余层融合对比度与结构相似性，并加权组合得到整体质量评估。取值范围为[0, 1]，越接近1表示图像质量越高。计算公式为：

$$\mathrm{MS-SSIM}(x,y)=\left[\mathcal{B}_{\Lambda}(\boldsymbol{x}_{\Lambda},\boldsymbol{y}_{\Lambda})\right]^{\alpha}\times\prod_{\lambda=1}^{\Lambda-1}\left[\mathcal{C}_{\lambda}(\boldsymbol{x}_{\lambda},\boldsymbol{y}_{\lambda})\right]^{\beta_{\lambda}} \tag{15}$$

式中，Λ为尺度总数，$\boldsymbol{x}_{\lambda}$和$\boldsymbol{y}_{\lambda}$为通过高斯金字塔生成的第λ尺度的下采样图像；$\mathcal{B}_{\Lambda}$为最低分辨率层($\lambda=\Lambda$)的亮度相似项；$\mathcal{C}_{\lambda}$为第λ尺度的对比-结构相似项；α通常取1，$\boldsymbol{\beta}=[\beta_1,\cdots,\beta_{\Lambda-1}]$为预设权重向量。

2) LPIPS

LPIPS是一个基于深度学习的图像质量评价指标。它利用神经网络进行特征提取，并计算这些特征之间的距离以评估图像之间的感知相似性。由于其基于人类识别模式提取特征，因此能更准确地反映人眼对图像质量的感知。LPIPS的取值范围为[0, 1]，较小的值表示图像具有较高的感知质量。

(2) 鲁棒性

评估算法的鲁棒性，通常采用恢复准确率(Recovery Accuracy，Acc_{rec})作为衡量标准，即攻击后正确提取的秘密信息比特数占总嵌入比特数的比例。计算公式如下：

$$Acc_{\mathrm{rec}}=1-\frac{1}{N}\sum_{i=1}^{N}\frac{n_{\mathrm{err}}(i)}{n} \tag{16}$$

式中，n为单张图像的嵌入容量，N为测试图像总数，$n_{\mathrm{err}}(i)$表示第i张图像提取出的错误比特数。

(3) 安全性

在隐写分析中，模型性能常通过误检率或检测准确率(Detection Accuracy，Acc_{dec})进行评价。本文统一采用检测准确率作为评估指标，其反映了模型正确区分载体图像(阳性类)与载密图像(阴性类)的能力。准确率计算如式(17)所示：

$$Acc_{\mathrm{dec}}=\frac{TP+TN}{TP+TN+FP+FN} \tag{17}$$

式中，TP和TN分别表示正确分类的载体图像与载密图像数量，FP和FN对应将载密图像误判为载体图像、将载体图像误判为载密图像的数量。该指标从整体上衡量了模型在二分类任务中的判别性能。

表 1　不同模型的视觉质量对比结果

Table 1　Results of visual quality comparison for different models

方法	Image Size	PSNR ↑	MS-SSIM ↑	LPIPS ↓
StegaStamp	400×400	27.183	0.934	0.131
FNNS-R	400×400	31.434	0.895	0.223
PRIS	224×224	30.681	<u>0.952</u>	0.114
RoSteALS	256×256	32.429	0.931	<u>0.087</u>
TransStego	400×400	<u>33.742</u>	0.932	0.096
本文	400×400	**35.728**	**0.961**	**0.066**

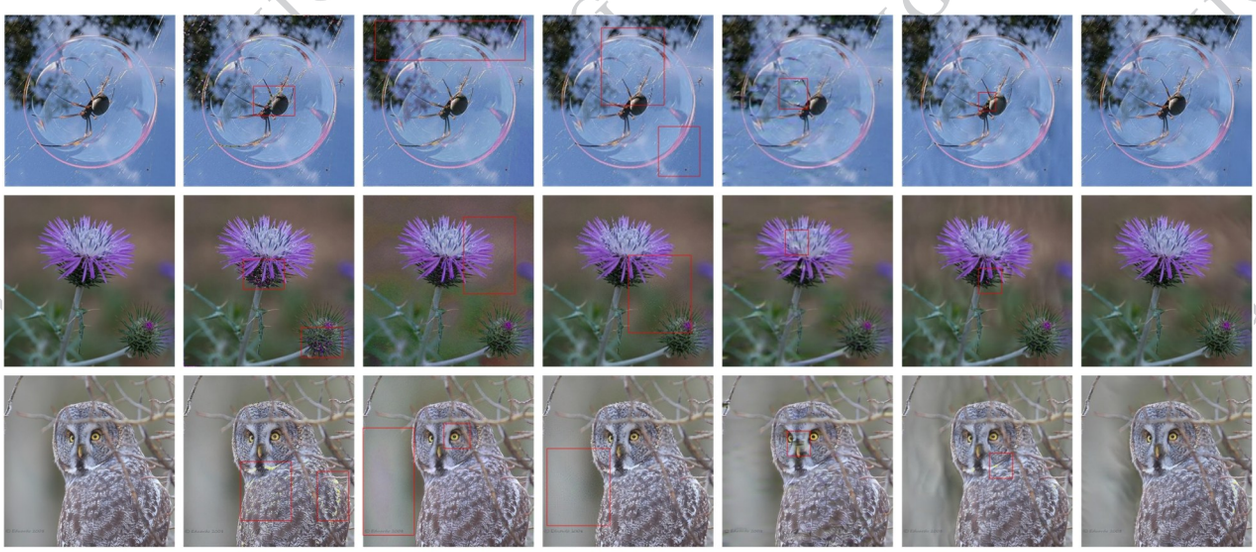
注：加粗、下划线字体分别表示各列最优、次优结果，“↑”表示值越大越好，“↓”表示值越小越好。

3.2　不可感知性

为了验证不可感知性，从 PSNR、MS-SSIM、LPIPS 等角度对所提方法进行评估。表 1 展示了在 ImageNet 数据集上，本文方法与 StegaStamp（Tancik 等，2020）、FNNS-R（Kishore 等，2022）、RoSteALS（Bui 等，2023）、PRIS（Yang 等，2024）、TransStego（Xiao 等，2025）五种主流的隐写模型在统一嵌入容量 100 bit 下的视觉不可感知性能对比结果。

如表 1 结果显示，所提方法在多项图像质量评估指标上均优于对比模型，展现出优异的视觉不可感知性。具体而言，其 PSNR 达到 35.728dB，较 TransStego 的 33.742 dB 提升 1.986 dB，相对提高约 5.88%；MS-SSIM 为 0.961，高于 TransStego 的 0.932，提升 3.13%；LPIPS 值降至 0.066，较 TransStego 的 0.096 降低 31.25%，表明生成图像在人眼感知层面与原始图像高度一致。上述结果表明，所提方法在嵌入秘密信息时，能有效减少视觉失真，使生成的载密图像在人眼感知上与原始图像高度一致，实现了优异的不可感知性。

为从视觉角度进一步验证所提方法的不可感知性，将其与基线方法生成的载密图像进行直观对比，如图 7 所示。实验结果显示，StegaStamp 和 FNNS-R 生成的载密图像存在局部异常高亮问题；PRIS 的输出图像出现明显纹理细节丢失；RoSteALS 与 TransStego 则在部分边缘位置伴有点状高亮或轻微结构扭曲。相比之下，所提方法生成的载密图像未引入明显视觉伪影，在整体清晰度、纹理细节及色彩一致性上均与原始图像高度接近，尤其在猫头鹰羽毛、花朵花瓣等复杂纹理区域，展现出更强的细节保留能力。上述表象表明，所提方法生成的载密图像在测试样本中均保持了良好的视觉连续性与自然性，有效提升了隐写图像的视觉质量，验证了方法优异的视觉不可感知性。



(e) RoSteALS; (f) TransStego; (g) Ours)

图 7　视觉效果对比图

Fig. 7　Comparison of visual effects chart ((a) original image; (b) StegaStamp; (c) FNNS-R; (d) PRIS;

3.3 鲁棒性

为全面评估所提模型的鲁棒性，实验依次从载密图像受攻击后的效果可视化分析、不同模型在常见攻击下的鲁棒性能对比以及不同模型在典型几何攻击场景下的鲁棒性能对比三个维度展开，并测试其在真实社交网络信道中的鲁棒性能。

从载密图像受攻击后效果可视化分析的角度，对载密图像进行了几种常见的噪声攻击，包括JPEG压缩、高斯噪声、中值滤波、高斯模糊、裁剪、缩放、旋转、旋转裁剪、旋转-缩放-裁剪、仿射变换和小范围随机失真，图8展示了不同噪声攻击下载密图像的失真效果。从图8可见，不同攻击对载密图像的视觉影响差异显著，且所提模型在面对多种攻击时均表现出优异的鲁棒性。所提模型在面对滤波（如高斯模糊、中值滤波）、压缩和缩放等攻击时均表现出优异的鲁棒性。所提模型在面对滤波（如高斯模糊、中值滤波）、压缩和缩放等攻击时，整体结构保持完整，表现出较强的鲁棒性。特别地，在强空间扰动的几何攻击中，“旋转裁剪”、“旋转-缩放-裁剪”因像素丢失而视觉失真最显著的强破坏性攻击，信息提取也保持了较高的准确率。尽管“旋转裁剪”因像素丢失成为最具破坏性的攻击类型，该模型的信息提取正确率仍能维持在65%及以上，体现了其在处理复合攻击时的优越性能。

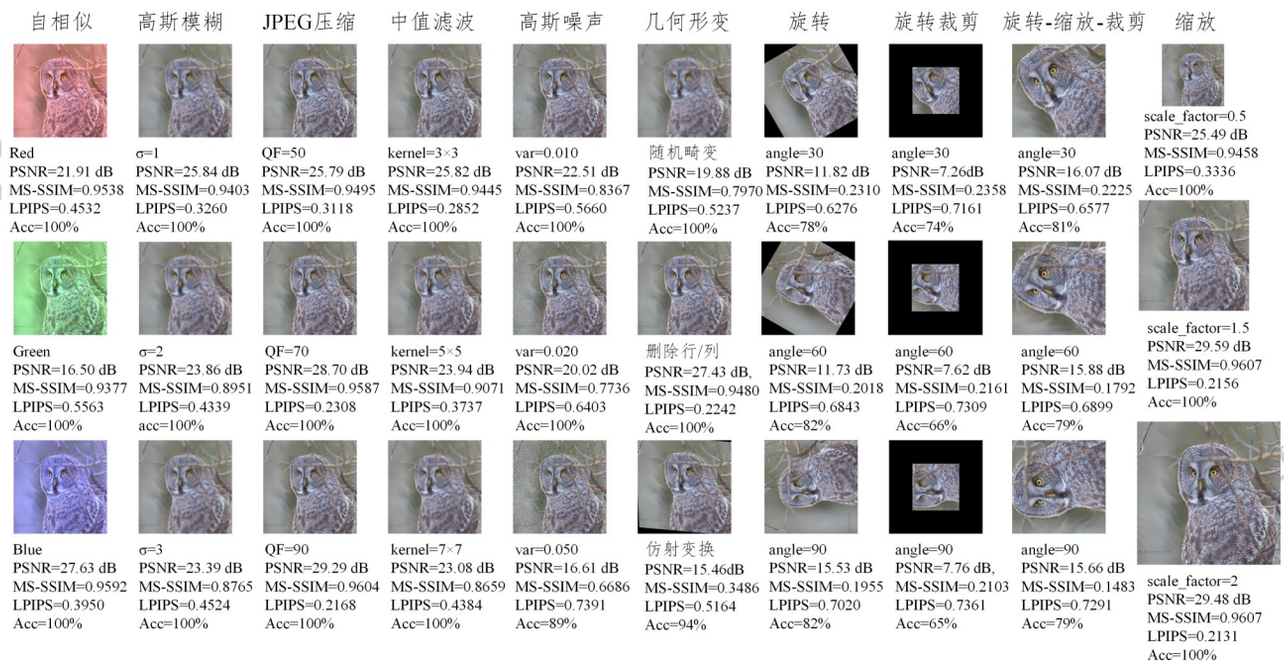
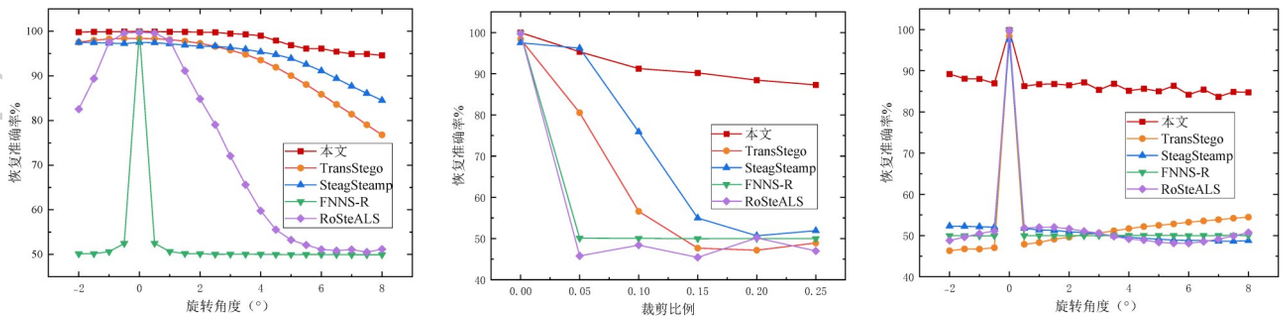


图8　不同噪声攻击下的载密图像失真效果

Fig. 8　Distortion effects of stego images under various noise attacks

从不同模型在典型几何攻击场景下的鲁棒性能对比视角，为评估模型在几何扰动下的鲁棒性，实验聚焦“旋转”、“裁剪”及“旋转-缩放-裁剪”三种典型的几何攻击。实验对比结果如图11所示。在旋转攻击下，本实验测试了[-2°, 8°]角度范围的提取性能。其中，在[-2°, -1°]区间，所提方法保持最优提取精度，TransStego、StegaStamp紧随其后维持稳定性能，RoSteALS的性能呈现逐步上升的趋势；在[-1°, 1°]区间，所提方法性能依旧最优，TransStego、StegaStamp及RoSteALS均表现出较好性能。FNNS-R对旋转攻击的鲁棒性较弱，几乎无法抵抗此类扰动。当角度超过1°时，RoSteALS的精度急剧下降；超过4°后，所提方法、TransStego与StegaStamp虽均出现明显性能下降，但所提方法的下降趋势更为平缓，并能保持更高的精度。在裁剪攻击下，本实验测试了[0.00, 0.25]裁剪比例的提取性能。其中，裁剪比例在[0.00, 0.05]区间时，所提方法与StegaStamp仍保持良好精度，而其他模型性能急剧下降，尤其是RoSteALS与FNNS-R的结果已接近随机猜测水平。相比之下，所提方法始终表现稳定在“旋转-裁剪-缩放”复合攻击下，本实验测试了[-2°, 8°]角度范围的提取性能。所提方法仍表现出较好性能与较



（a） 旋转　　（b） 裁剪　　（c） 旋转-裁剪-缩放

图9 各模型的抗几何扰动性能对比实验

Fig. 9 Performance comparison of different models against geometric perturbations ((a) rotation; (b) cropping; (c) rotation-cropping-scaling)

强的抗扰动能力，进一步验证了其在此类复合几何攻击下的鲁棒性。

为验证所提方法在真实社交网络信道中的鲁棒性，本文基于微博平台开展实际图像传输测试。实验随机选取100张载密图像，以原图形式分别上传至微博动态、相册及私信三类典型场景，并通过不同下载方式获取经平台处理后的图像。在此基础上，计算载密图像的比特准确率与秘密信息提取效果，实验结果如表3所示。实验结果显示，比特准确率平均为99.99%，秘密信息提取的成功率为100%。这表明，本方法能够有效抵抗基于微博平台对载密图像的处理攻击，进一步验证了其在实际社交网络环境中的鲁棒性与可靠性。

3.4 安全性

为了验证提出方法的安全性，本文使用Weng等人(2019)方法中的训练策略，即观察隐写分析网络需要多少对载体/载密样本学习训练，才能在测试阶段正确区分出载体图像和载密图像。在嵌入容量为100 bit的条件下，测试不同训练样本量下隐写分析模型(UCNet)对各方法载密图像的检测准确率，以此评估各方法的抗检测性能。同时，将共享图像纳入训练与测试流程，进一步验证其安全性，结果如图10所示。

实验结果表明，共享图像的检测准确率接近50%的随机判别水平，说明其不存在可被隐写分析模型利用的统计特征。本文方法生成的载密图像检测准确率始终保持最低且增长缓慢，表现出优异的统计安全性与抗分析稳定性。相比之下，FNNS-R、RosteALS、TransStego和StegaStamp方法虽具备一定隐蔽性，但在训练样本充足时，易被隐写分析器有效检测。综上，所提方法在抵御先进隐写分析器方面具有显著优势，可有效保障共享图像与载密图像同步传输场景下的整体安全性。

3.5 消融实验

为系统性地验证所提隐写分析器对抗训练机制与双级纠错机制各自的有效性及其协同作用，我们设置了四种消融配置：基线模型(A)不引入隐写分析器且不使用纠错机制；仅纠错模型(B)在基线基础上引入双级纠错机制；仅隐写分析器模型(C)引入隐写分析器参与对抗训练但不使用纠错机制；完整模型(D)同时引入隐写分析器并启用双级纠错机制。表4与表5展示了消融实验结果：表4展示了在A－D四种配置下生成的1000张载密图像的视觉质量评估结果，表5则对比了各配置在多种信道攻击下的秘密信息恢复准确率，以验证各模块对整体性能的贡献。基于实验结果，可得出以下结论：

首先，双级纠错机制显著提升了系统的鲁棒性。通过对比模型A与B，以及模型C与D的结果可以看出，使用纠错机制后，在所有攻击类型下的信息恢复准确率均有明显提高。具体而言，在非几何攻击中，模型A与B在高斯噪声(方差为0.05)场景下的正确率从79.44%提升至92.61%，提高了13.17%；而在“旋转60°”攻击下，正确率从47.94%提升至81.13%，提高了33.69%。这表明，纠错模块能够有效校正传输过程中的比特错误。

其次，引入隐写分析器显著提升了隐蔽性，但对鲁棒性产生了一定影响。通过对比模型B与D可以发现，在部分攻击(如旋转裁剪、高斯噪声)下，模型

表 2　不同攻击下的秘密信息提取准确率

Table 2　Secret message extraction accuracy under various attacks

攻击类型	参数	TransStego	StegaStamp	RoSteALS	FNNS-R	本文方法
自相似	Red	98.13%	97.26%	99.64%	52.32%	**99.88%**
	Green	98.13%	97.39%	99.46%	52.73%	**99.82%**
	Blue	97.93%	97.31%	99.53%	52.24%	**99.85%**
高斯模糊	1	98.35%	97.41%	99.76%	50.84%	**99.97%**
	2	97.67%	97.12%	99.76%	50.36%	**99.69%**
	3	96.92%	96.96%	99.72%	50.28%	**99.49%**
JPEG压缩	50	97.97%	97.25%	99.69%	52.36%	**99.80%**
	70	98.13%	97.44%	99.82%	52.47%	**99.93%**
	90	98.36%	97.47%	99.89%	52.67%	**99.94%**
中值滤波	3×3	98.47%	97.89%	99.90%	51.10%	**99.91%**
	5×5	98.35%	97.91%	**99.89%**	50.48%	99.77%
	7×7	97.24%	97.78%	**99.79%**	50.30%	99.38%
高斯噪声	0.01	93.14%	96.44%	**99.91%**	53.00%	97.45%
	0.02	88.56%	94.36%	**99.63%**	53.30%	95.06%
	0.05	79.44%	92.45%	**97.94%**	53.55%	90.41%
随机畸变	0.02	87.47%	86.76%	88.5%	49.99%	**91.70%**
删除行/列	1	98.39%	97.51%	99.76%	51.71%	**99.82%**
仿射变换	0.1	79.34%	77.97%	51.09%	49.97%	**85.70%**
旋转	30°	47.44%	51.02%	49.79%	49.99%	**92.14%**
	60°	47.94%	50.70%	46.42%	49.97%	**83.92%**
	90°	50.44%	53.28%	47.98%	49.97%	**80.30%**
旋转裁剪	30°	51.52%	53.13%	42.48%	49.99%	**73.10%**
	60°	50.49%	49.63%	43.60%	49.97%	**67.94%**
	90°	50.23%	55.21%	42.22%	49.97%	**71.17%**
旋转缩放裁剪	30°	52.96%	53.92%	50.96%	49.97%	**79.72%**
	60°	48.59%	51.09%	50.35%	49.97%	**79.47%**
	90°	50.75%	48.39%	48.30%	49.97%	**78.43%**
缩放	0.5	98.36%	97.41%	99.76%	50.89%	**99.87%**
	1.5	98.41%	97.51%	99.75%	52.01%	**99.92%**
	2	98.38%	97.50%	99.75%	52.00%	**99.97%**

注：加粗、下划线字体分别表示各行最优、次优结果。

D的提取准确率略低于模型B。这主要是因为隐写分析器引导信息嵌入视觉不敏感区域，而这些区域在高斯噪声或强几何变换下保留信息的能力有限。相比之下，模型B不受隐蔽性约束，可以更灵活地分配秘密信息的嵌入位置与嵌入强度，从而反映出隐蔽性与鲁棒性之间的权衡。

表3　实际社交网络鲁棒性实验

Table3　Robustness experiments of real social networks

测试场景	下载方式	比特准确率	信息提取成功率
微博动态	原图下载	99.99%	100%
	直接下载	99.98%	100%
微博相册	原图下载	99.99%	100%
	直接下载	99.98%	100%
微博私信	下载	100%	100%

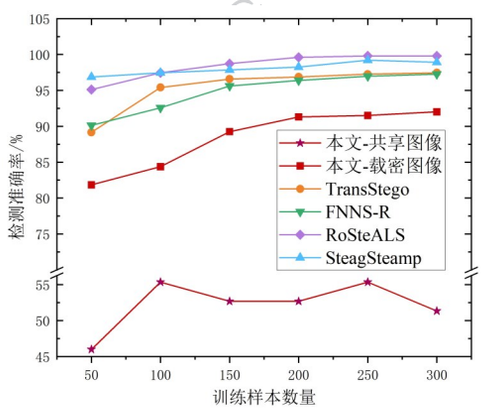


图10　不同模型在不同训练样本数量下的检测精度

Fig. 10　Detection accuracy of different models with varying training sample sizes

表4　不同消融配置下载密图像的视觉质量

Table 4　Visual quality of stego images under different ablation configurations

模型	PSNR ↑	MS-SSIM ↑	LPIPS ↓
A	33.74183	0.93158	0.09607
B	33.74183	0.93158	0.09607
C	**35.72847**	**0.96073**	**0.06687**
D	**35.72847**	**0.96073**	**0.06687**

注：加粗字体表示各列最优结果，“↑”表示值越大越好，“↓”表示值越小越好。

最后，针对不同攻击类型的表现进一步验证了方案的有效性。在常规压缩与滤波攻击（如JPEG压缩、模糊）下，所有模型均保持较高的正确率，纠错机制使正确率提升至99.6%以上。在强度较大的攻击（如噪声干扰和旋转裁剪）中，纠错机制显著减小了性能损失。特别是在单一和复合几何攻击（如旋转30°、旋转60°-缩放-裁剪）下，模型D分别比模型A提高了44.70%和30.88%的正确率，展现了良好的容错能力。

总之，得益于对图像质量与鲁棒性的协同优化，该方法在多种攻击下实现了更高的秘密信息恢复准确率，显著提升了系统的可靠性与实用性。

4　结　论

本文针对现有基于编码-解码网络的鲁棒图像隐写方法在局部视觉质量差与多种信道攻击下恢复性能有限的问题，提出一种融合多对抗引导和哈希纠错的鲁棒图像隐写方法。通过将隐写分析器引入生成训练，实现了秘密信息向统计自然且人眼不敏感区域的自适应嵌入，有效缓解了平滑区伪影与纹理细节丢失；进一步结合BCH纠错码与无歧义哈希映射构建双级纠错机制，借助预共享图像库中的自然图像哈希建立纠错映射，避免在载密图像中嵌入冗余纠错信息，显著提升了秘密信息的提取准确率。实验表明，本文方法减少了载密图像的局部区域失真，提升了整体视觉质量；同时在JPEG压缩、滤波及典型的复合攻击下均展现出较强的鲁棒性。尽管本文方法在多种攻击下展现出优于现有方法的鲁棒性，但在强几何攻击下纠错能力仍存在局限。未来将改进哈希纠错与编解码隐写网络的协作机制，进一步提升在社交平台等真实场景中的实用性与可靠性。

表 5　多种信道攻击下各消融配置的秘密信息提取准确率

Table 4　Secret message extraction accuracy of each ablation configuration under various channel attacks

攻击类型	参数	A	B	C	D	攻击类型	参数	A	B	C	D
自相似	Red	98.13%	99.66%	97.87%	**99.88%**	随机畸变	0.02	87.47%	**94.31%**	80.25%	**91.70%**
	Green	98.13%	99.61%	97.88%	**99.82%**	删除行/列	1	98.39%	99.67%	98.53%	**99.82%**
	Blue	97.93%	99.61%	98.05%	**99.85%**	仿射变换	0.1	79.34%	85.69%	70.66%	**85.70%**
高斯模糊	1	98.35%	99.71%	98.50%	**99.97%**	旋转	30°	47.44%	69.57%	58.04%	**92.14%**
	2	97.67%	99.53%	97.29%	**99.69%**		60°	47.94%	81.13%	54.25%	**83.92%**
	3	96.92%	99.40%	96.14%	**99.49%**		90°	50.44%	79.39%	50.52%	**80.30%**
JPEG压缩	50	97.97%	99.64%	97.36%	**99.80%**	旋转裁剪	30°	51.52%	**75.69%**	46.98%	73.10%
	70	98.13%	99.70%	98.17%	**99.93%**		60°	50.49%	**71.23%**	47.97%	67.94%
	90	98.36%	99.72%	98.48%	**99.94%**		90°	50.23%	**73.46%**	47.98%	71.17%
中值滤波	3×3	98.47%	99.66%	98.61%	**99.91%**	旋转–缩放–裁剪	30°	52.96%	**83.91%**	46.22%	79.72%
	5×5	98.35%	99.63%	98.22%	**99.77%**		60°	48.59%	78.31%	49.80%	**79.47%**
	7×7	97.24%	**99.44%**	96.37%	99.38%		90°	50.75%	77.22%	49.07%	**78.43%**
高斯噪声	0.01	93.14%	**98.49%**	88.54%	97.45%	缩放	0.5	98.36%	99.63%	98.48%	**99.87%**
	0.02	88.56%	**96.87%**	82.95%	95.06%		1.5	98.41%	99.70%	98.55%	**99.92%**
	0.05	79.44%	**92.61%**	73.43%	90.41%		2	98.38%	99.73%	98.51%	**99.97%**

注：加粗、下划线字体分别表示各列最优、次优结果。

参考文献（References）

Arjovsky M, Chintala S, and Bottou L. 2017. Wasserstein generative adversarial networks//Proceedings of 2017 International Conference on Machine Learning. Sydney. NSW, Australia: ACM: 214 – 223 [DOI:10.5555/3305381.3305404]

Bui T, Agarwal S, Yu N, and Collomosse J. 2023. RoSteALS: robust steganography using autoencoder latent space//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, BC, Canada: IEEE: 933 – 942 [DOI: 10.1109/CVPRW59228.2023.00100]

Cheng Y, Luo Z, and Yin Z X. 2025. Robust steganography with boundary-preserving overflow alleviation and adaptive error correction. Expert Systems with Applications, 281: 127598 [DOI: 10.1016/j.eswa.2025.127598]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2020. An image is worth 16×16 words: transformers for image recognition at scale//Proceedings of 2020 International Conference on Learning Representations. [s.l.]: OpenReview.net.

Duan X L, Li B, Yin Z X, Zhang X P, and Luo B. 2023. Robust image steganography against lossy JPEG compression based on embedding domain selection and adaptive error correction. Expert Systems with Applications, 229: 120416 [DOI: 10.1016/j.eswa.2023.120416]

Forney G D. 1965. On decoding BCH codes. IEEE Transactio-ns on Information Theory, 11 (4): 549-557 [DOI: 10.1109/TIT. 1965. 1053808]

Filler T, Judas J, and Fridrich J. 2011. Minimizing additive distortion in steganography using syndrome-trellis codes. IEEE Transactions on Information Forensics and Security, 6(3): 920 – 935 [DOI: 10.1109/TIFS.2011.2134094]

Fridrich J, Pevný T, and Kodovský J. 2007. Statistically unde-tectable JPEG steganography: dead ends challenges, and opportunities// Proceedings of 2007 Workshop on Multimedia & Security. Dallas, Texas, USA: ACM: 3 – 14 [DOI: 10.1145/1288869.1288872]

Fu Z J, Wang F, Sun X M, Wang Y. 2020. Research on steg-anography of digital images based on deep learning. Chin-ese Journal of Computers, 43(9): 1656-1672 (付章杰, 王帆, 孙星明, 王彦. 2020. 基于深度学习的图像隐写方法研究. 计算机学报, 43(9): 1656-1672) [DOI: 10.11897/SP.J.1016.2020.01656]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Sherjil Ozair, et al. 2014. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing System. Montreal, Quebec, Canada MIT Press: 2672 – 2680

Gore W C. 1969. Generalized threshold decoding and the Reed- Solomon codes. IEEE Transactions on Information Theory, 15 (1): 78-81 [DOI: 10.1109/TIT.1969.1054269]

Ho J, Jain A, and Abbeel P. 2020. Denoising diffusion probabilistic

models//Proceedings of the 34th Advances in Neural Information Processing Systems. Curran Associates, Inc.:6840 - 6851

Hu X, Li S, Ying Q, Peng W, Zhang X P, and Qian Z X. 2024. Establishing robust generative image steganography via popular stable diffusion. IEEE Transactions on Information Forensics and Security, 19: 8094-8108 [DOI: 10.1109/TIFS.2024.3444311]

Jaderberg M, Simonyan K, and Zisserman A. 2015. Spatial transformer networks//Proceedings of the 29th Annual Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc.: 28

Kishore V, Chen X Y, Wang Y, Li B Y, and Weinberger K L Q. 2022. Fixed neural network steganography: train the images, not the network//Proceedings of 2022 International Conference on Learning Representations. [s.l.]: OpenReview.net.

Li B, He J H, Huang J W, and Shi Y Q. 2011. A survey on image steganography and steganalysis. Journal of Information Hiding and Multimedia Signal Processing, 2(2): 142 - 172

Li L, Yang C, Chen J. 2023. A coverless image steganography method based on feature matrix mapping//Proceedings of the Chinese Conference on Image and Graphics Technologies. Singapore: Springer: 472 - 488 [DOI: 10.1007/978-981-99-7549-5_34]

Meng L J, Jiang X H, Zhang Z Z, Li Z H, and Sun T F. 2022. A robust coverless image steganography based on an end-to-end hash generation model. IEEE Transactions on Circuits and Systems for Video Technology, 33(7): 3542 - 3558 [DOI: 10.1109/TCSVT.2022.3232790]

Peng Y Y, Wang Y F, Hu D H, Chen K J, Rong X J, and Zhang W M. 2024. LDStega: practical and robust generative image steganography based on latent diffusion models//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 3001 - 3009 [DOI: 10.1145/3664647.3681635]

Rajan J G K and Ganesh R. 2025. Dynamic pixel shuffling and hash LSB steganography with RC4 encryption: a robust data security framework. Expert Systems with Applications, 279: 127403 [DOI: 10.1016/j.eswa.2025.127403]

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: IEEE:10684-10695 [DOI: 10.1109/CVPR52688.2022.01042]

Ronneberger O, Fischer P, and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany. Springer: 234 - 241 [DOI: 10.1007/978-3-319-24574-4_28]

Tancik M, Mildenhall B, and Ng R. 2020. StegaStamp: invisible hyperlinks in physical photographs// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Electr Network: ACM: 2114 - 2123 [DOI: 10.1109/CVPR42600.2020.00219]

Tang Y X, Di F Q, Zhang Z, and Zhang M Q. 2024. Review of coverless image steganography//Proceedings of the 4th International Conference on Electronic Information Engineering and Computer. Shenzhen, China: IEEE: 1006 - 1011 [DOI: 10.1109/EIECT64462.2024.10866313]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31th Advances in Neural Inform-ation Processing Systems. Long Beach, California, USA: Curran Associates, Inc.: 261 - 272

Wang Z, Simoncelli E P, and Bovik A C. 2003. Multiscale structural similarity for image quality assessment//Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers. Pacific Grove, CA, USA: IEEE: 1398-1402[DOI: 10.1109/ACSSC.2003.1292216]

Wei K K, Luo W Q, Tan S Q, and Huang J W. 2022. Universal deep network for steganalysis of color image based on channel representation. IEEE Transactions on Information Forensics and Security, 17: 3022 - 3036 [DOI: 10.1109/TIFS.2022.3196265]

Weng X Y, Li Y Z, Chi L and Mu Y D. 2019. High-capacity convolutional video steganography with temporal residual modeling//Proceedings of the 2019 on International Conference on Multimedia Retrieval. Ottawa, Canada: ACM: 87 - 95 [DOI: 10.1145/3323873.3325011]

Xiao C E, Peng S R, Zhang L, Wang J X, Ding D, and Zhang J Y. 2025. A transformer-based adversarial network framework for steganography. Expert Systems with Applications, 269: 126391 [DOI: 10.1016/j.eswa.2025.126391]

Yang Z J, Chen K J, Zeng K, Zhang W M, and Yu N H. 2023. Provably secure robust image steganography. IEEE Transactions on Multimedia, 26: 5040 - 5053 [DOI: 10.1109/TMM.2023.3330098]

Yang H, Xu Y T, Liu X H, and Ma X. 2024. PRIS: practical robust invertible network for image steganography. Engineering Applications of Artificial Intelligence, 133: 108419 [DOI: 10.1016 /j.engappai.2024.108419]

Yao Y, Huang L C, Wang H, Chang Q, Ren Y Z, and Xiao F J. 2024. Robust adaptive steganography based on adaptive STC-ECC. IEEE Transactions on Multimedia, 26: 5477 - 5489 [DOI: 10.1109/TMM.2023.3334487]

Yu X Z, Chen K J, Wang Y F, Li W X, Zhang W M, and Yu N H. 2020. Robust adaptive steganography based on generalized dither modulation and expanded embedding domain. Signal Processing, 168: 107343 [DOI: 10.1016/j.sigpro.2019.107343]

Zeng K, Chen K J, Zhang W M, Wang Y F, and Yu N H. 2023. Robust steganography for high quality images. IEEE Transactions on Circuits and Systems for Video Technology, 33(9): 4893 - 4906 [DOI: 10.1109/TCSVT.2023.3250750]

Zhang R, Isola P, Efros A A, Shechtman E, and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//

Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 586 - 595 [DOI: 10.1109/CVPR.2018.00068]

Zhang Y, Luo X Y, Yang C F, Ye D P, and Liu F L. 2015. A JPEG-compression resistant adaptive steganography based on relative relationship between DCT coefficients//Proceedings of the 10th International Conference on Availability, Reliability and Security. Toulouse, France: IEEE: 461 - 466 [DOI: 10.1109/ARES.2015.53]

Zhang Y, Luo X Y, Guo Y Q, Qin C, and Liu F L. 2019. Multiple robustness enhancements for image adaptive steganography in lossy channels. IEEE Transactions on Circuits and Systems for Video Technology, 30(8): 2750 - 2764 [DOI: 10.1109/TCSVT.2019.2923980]

Zhang Y, Luo X Y, Wang J W, Guo Y Q, and Liu F L. 2021. Image robust adaptive steganography adapted to lossy channels in open social networks. Information Sciences, 564: 306 - 326 [DOI: https://doi.org/10.1016/j.ins.2021.02.058]

Zhang Y, Luo X Y, Wang J W, Yang C F and Liu F L. 2022. A survey on robust image steganography. 中国图象图形学报, 27(1): 3-26)(张祎, 罗向阳, 王金伟, 卢伟, 杨春芳, 刘粉林. 2022. 数字图像鲁棒隐写综述. 中国图象图形学报, 27(1): 3-26 [DOI: 10.11834/jig.210449]

Zhang Y, Ma Y Y, Zhang Q Q, Pei Y, Luo X Y. 2025. An image robust batch steganography framework with minimum embedding signs. IEEE Transactions on Information Forensics and Security, 20: 10745-10760[DOI: 10.1109/TIFS.2025.3615446]

Zhu J, Kaplan R, Johnson J, and Fei-Fei L. 2018. HiDDeN: hiding data with deep networks//Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer: 682-697 [DOI: 10.1007/978-3-030-01267-0_40]

作者简介

李荣，女，硕士研究生，主要研究方向为图像隐写。E-mail: lr923myself@163.com

张祎，通讯作者，女，讲师，主要研究方向为图像隐写与隐写分析。E-mail: tzyy4001@sina.com

罗向阳，男，教授，主要研究方向为图像隐写与隐写分析。E-mail: luoxy_ieu@sina.com

张明亮，男，博士研究生，主要研究方向为信息隐藏和多媒体安全。E-mail: zmlmail2021@163.com

刘燕美，女，博士研究生，主要研究方向为信息隐藏和多媒体安全。E-mail: lym129@zua.edu.cn