Canny-Gauss 通用域图像隐写算法

李季瑀[1)] 付章杰[1),2)] 王 帆[1)]

[1)](南京信息工程大学数字取证教育部工程研究中心 南京 210044)

[2)](西安电子科技大学空天地一体化综合业务网全国重点实验室 西安 710126)

摘 要 自适应图像隐写算法是一种以图像为载体,通过手工设计嵌入失真代价,指导隐写码在图像载体中嵌入秘密消息的信息隐藏算法.长期以来,这类算法将秘密消息尽可能隐藏在图像纹理更深更复杂的位置以对抗基于富特征的隐写分析检测.然而,伴随着深度学习在隐写分析领域的快速发展,人工设计的自适应算法受到严重挑战.此外,基于加性失真的隐写编码在嵌入消息时,复杂纹理向边界聚集所产生的统计异常问题也亟待解决.因此,本文总结了各类人工失真代价的优势和不足,归纳出当前自适应算法在空域的设计范式,并结合 UNIWARD 在各嵌入域的转换规则,提出基于嵌入失真代价 ρ 的通用域隐写转换公式.然后,从隐写嵌入失真代价与图像纹理稀疏关系的角度出发,以 Canny 算子划分纹理、Gauss 模糊缩放轮廓、AutoML 搜索阈值的方式,提出了一种通用域隐写算法 Canny Gauss.实验结果表明,本文所提通用域隐写转换公式能够有效应用于现有主流算法.同时,在 UNIWARD 所有可行嵌入域中,本文所提算法表达出更高嵌入失真代价稳定性和隐写隐蔽性,在第三方权重加持下的深度隐写分析表现与 UNIWARD 相比至少提升 2.6%、最高提升 14.6%.这为自适应隐写算法的通用域设计,以及抵抗基于纹理特征的深度隐写分析检测提供了新思路.

关键词 自适应隐写术;隐写失真设计;通用域;Canny 算子;自动化机器学习

中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2024.00213

Canny-Gauss Universal Domain Image Steganography Algorithm

LI Ji-Yu[1)] FU Zhang-Jie[1),2)] WANG Fan[1)]

[1)](*Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing* 210044)

[2)](*State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an* 710126)

Abstract Adaptive image steganography algorithms have emerged as a means of concealing secret messages within image carriers, employing a manual design of distortion costs to guide the process of message embedding. The primary objective of these algorithms has been to hide secret information in regions of the image that possess intricate and complex textures, thereby thwarting feature-based steganalysis detection methods. However, the rapid advancements in deep learning within the field of steganalysis have posed significant challenges to the efficacy of manually designed adaptive algorithms. Furthermore, there is a pressing need to address the statistical anomalies that arise from the aggregation of complex textures towards the boundaries when employing additive distortion-based steganographic encoding techniques. To tackle these challenges, this paper provides a general summary of the strengths and limitations associated with various handcraft distortion cost design. It also presents an paradigm of the existing design paradigms for adaptive algorithms in the spatial domain, considering the transformation rules of UNIWARD

收稿日期:2023-02-16;在线发布日期:2023-07-17.本课题得到国家重点研发计划(2021YFB2700900)、国家自然科学基金(U22B2062,62172232)、江苏省杰出青年基金(BK20200039)、江苏省大气环境与装备技术协同创新中心基金资助.**李季瑀**,硕士研究生,主要研究方向为启发式图像隐写.E-mail:lijiyu@nuist.edu.cn.**付章杰**(通信作者),博士,教授,中国计算机学会(CCF)专业会员,主要研究领域为信息隐藏、网络与信息安全.E-mail:fzj@nuist.edu.cn.**王 帆**,博士研究生,主要研究方向为信息隐藏、深度学习.

across different embedding domains. In order to improve upon the existing techniques, the paper proposes a universal domain steganographic transformation formula based on the embedding distortion cost ρ. This formula provides a flexible framework that can be applied to a wide range of mainstream algorithms, enhancing their performance and adaptability. Moreover, this paper introduces a groundbreaking universal domain steganographic algorithm known as Canny Gauss, which capitalizes on multiple techniques to achieve highly effective message embedding. Firstly, the algorithm employs the Canny operator to perform texture segmentation, enabling the identification and selection of regions within the image that possess rich texture information suitable for embedding secret messages. By leveraging this approach, Canny Gauss ensures that the embedded messages are strategically placed within areas that can effectively camouflage the hidden information. In addition, the algorithm utilizes Gaussian blur to scale the contours of the image. This step is crucial in guaranteeing a seamless integration of the embedded messages with the surrounding textures, making them inconspicuous to visual inspection and steganalysis techniques. To further optimize the performance of the algorithm, an AutoML framework is employed to automatically search for suitable threshold values. This technique enhances the overall robustness and effectiveness of the steganographic process by dynamically adjusting the thresholds based on the characteristics of the input image. By adapting the thresholds to each specific image, Canny Gauss maximizes the concealment of secret messages while minimizing any adverse effects on image quality or detectability. Experimental results demonstrate the efficacy of the proposed universal domain steganographic transformation formula when applied to existing algorithms. In comparison to UNIWARD, the algorithm presented in this paper exhibits improved stability in embedding distortion costs and enhanced steganographic security. Moreover, when coupled with third-party weights, the algorithm showcases notable improvements in deep steganalysis performance, with a minimum enhancement of 2.6% and a maximum enhancement of 14.6% compared to UNIWARD. This paper not only provides valuable insights for the design of adaptive steganography algorithms in universal domains but also offers a new strategie to counter deep steganalysis detection techniques that rely on texture features.

Keywords adaptive steganography; steganographic cost design; universal domain; Canny operator; auto machine learning

1 引 言

图像信息隐藏是提高秘密消息与图像冗余信息耦合的信息嵌入技术，使经过可逆变换改造的秘密消息与载体冗余信息汉明距离高度接近，用于降低消息嵌入后的视觉影响. 目前，主流启发式图像隐写算法将秘密消息尽可能隐藏在图像纹理更深更复杂的位置，以对抗更先进的隐写检测技术. 自伴随式隐写码问世以来[1]，由于通信编码的对偶隐写码能够逼近嵌入极限的理论值上界，研究者们因而从寻找图像冗余信息中的合适嵌入特征[2]转向解释并计算图像嵌入失真函数的研究方向，由此诞生了一系列经典的启发式图像隐写算法[3-10].

当前，基于加性嵌入失真的自适应隐写算法仍是主流，Filler 等人提出的 HUGO[4] 是以减少 SPAM[11] 特征的统计相关性为出发点设计的；UNIWARD[5] 为了能实现对通用域的嵌入失真代价计算，将 2 维小波滤波器与像素方向残差结合寻找适合嵌入的复杂纹理区域；同时 Holub 等人在其提出的 UNIWARD 的基础上设计 WOW[6]，该算法可以看作是 UNIWARD 在空域的一种 P 范数变形；类似地，Li 等人提出的 HILL[7] 仅通过几层 KB 滤波核过滤嵌入关键位置，该算法作为同类算法中效率最高的代表仍受到了广泛关注；此外，Sedighi 等人的 MIPOD[8] 通过维纳滤波和费雪信息确定嵌入位置并借助递归树解决嵌入问题；Guo 等人也以 UED 为原型设计了适用于 JPEG 域和边信息域的 UERD[9]，这是一种

将显著性统计特征平均化并分散至 DCT 系数的方案；此前一段时间，Su 等人回顾其团队在 UERD 的设计思路并拓展 MIPOD 设计，思考隐写分析特征对协方差等相关统计量的影响，提出非加性算法 GMRF[10]，在 SRM 特征层面取得了超越 MIPOD 和 HILL 的安全性表现.

近年来，随着深度学习在隐写分析领域的广泛应用，自适应隐写算法的安全性受到挑战. 先进的隐写分析算法大多采用深度神经网络作为特征提取和分析手段[12]，研究人员从结合滤波核特征与神经网络的方案如 XuNet[13]等，过渡到改造残差网络的方案 SRNet[14]，再到利用 EfficientNet[15]等模型迁移学习隐写分析任务，不断提高隐写分析精度，弥补了隐写分析特征一直以来依靠经验设计的局限性[16]. 为了抵抗这类新的隐写分析技术，研究人员尝试在图像隐写时结合深度学习算法，Tang 等人[17]于 2017 年率先提出 ASDL-GAN，该方案首次跳过隐写失真代价设计，根据隐写分析器 XuNet 的检测结果对抗生成嵌入概率图，并结合所提 Ternary Embedding Simulator 子网络模拟±1 嵌入，模拟验证深度设计不弱于传统自适应隐写术的安全性. 紧接着，Yang等人[18]以U-Net作为生成器提出UT-GAN，该方案通过设计 Double-tanh 函数替代 TES 网络，简化嵌入概率向像素增减变化的过程，进一步提升了隐写嵌入的安全性. 同一时间，误导深度模型分类的对抗样本技术兴起，让攻击深度隐写分析器的隐写术实现成为可能. Zhang 等人[19]提出将自适应算法的目标载体对抗样本化，以应对多种分析特征检测. 随后，基于对抗样本的非加性隐写失真设计 ADV-EMB[20]也被 Tang 等人提出，该算法首先根据非加性规则分配隐写算法嵌入对应网格，其次通过对梯度误差的反向计算，将用于攻击隐写分析器的对抗样本，代入存放原有秘密消息的其中一层网格中，牺牲部分嵌入容量以增大秘密信息对深度分析器的隐蔽性，该方案可实际应用于大部分加性自适应隐写算法，因此有效提升了大量加性失真设计的隐蔽性. 此后，该团队在 ASDL-GAN 和 UT-GAN 的基础上分析 GAN 的优化困境，采用强化学习策略并提出 SPAR-RL[21]取得了优于 HILL 的隐写安全性表现. 此前，Mo 等人[22]基于蒙特卡洛树搜索，提出了一种强化学习通用域框架 MCTSteg，最近，该团队又基于 SPAR-RL 和 ADV-EMB 的经验，提出 ReLOAD[23]以 A3C 强化学习框架，以 SRM 滤波核的期望作为评估标准改进现有加性设计，进一步提高了现有加性失真对深度隐写分析特征的统计安全性.

尽管上述研究已经取得了显著的隐写隐蔽性表现，但基于特征对抗的方法受原始失真代价作用域的制约缺乏合理的泛化手段. 同时，Zhang 等人[24]指出隐写算法为了增加对某种显著特征的隐蔽性，其设计势必要引入另一种新的特征. 而上述的新特征对深度学习而言，只是样本量和训练成本问题，这将对现有算法的安全性构成巨大威胁[25]. 现有基于深度学习的隐写术，本质上等同于针对某一类训练好的神经网络的定向投毒攻击，以干扰原始决策边界，但这类攻击方式对训练完备的隐写分析器，尤其是基于统计特征的集成分析器收效甚微. 此外，主流的隐写失真针对不同作用域（空域、频域、边信息域）的设计存在较大分歧，一类以 UERD 为例，认为在同一个正交变换基（DCT、DWT 等）下的作用域能够适用于同一种失真设计；而一类如 MIPOD，认为不同作用域应分别采取不同的设计策略；此外，以 UNIWARD 为主的方案也提出了三域通用的设计原则. 由于某一特定的隐写失真设计在不同域的嵌入成本并未发生实质性改变，让特定域的隐写失真设计向其他域扩展成为可能. 但如何扩展某一特定域的隐写失真函数以及如何保持扩展后的失真设计在其他域抵抗分析特征的鲁棒表达是目前难点所在.

针对上述问题，本文结合图像分割视角提出了一种更加直观高效的隐写失真计算方案，改善对完备分类器的隐写隐蔽性. 同时为了简化隐写失真设计流程，提高嵌入失真函数的复用和鲁棒性表达，本文研究了 UNIWARD 的相关嵌入设计，在此基础上提出了一种通用域隐写失真代价转换公式. 下面将阐明本文的主要贡献：

（1）为改善加性失真引发的嵌入聚集降低统计安全性的问题，本文从图像分割和熵优化角度出发，通过 Canny 算子过滤主要纹理、Gauss 模糊缩放边界轮廓、AutoML 推断隐写分析特征计算搜索超参的方式，提出一种同时满足空域、频域和边信息域嵌入要求的通用域图像隐写失真设计 Canny Gauss.

（2）本文通过对比分析多个主流自适应隐写设计，总结出当前启发式图像隐写设计的重要共识，即明确纹理复杂区域发生变化时的异常信号扩散边界，并提出基于嵌入失真代价 ρ 的通用域隐写转换公式.

（3）通过 Canny Gauss 与其他主流算法在任意

域下的安全性实验分析，本文验证了所提出的隐写设计共识和通用域转换公式的有效性，进一步表明基于该共识的隐写失真设计在抵抗作用域变换时的鲁棒性，实验结果表明 Canny Gauss 在不同作用域下表现出 2.6%～14.6%的安全性提升.

本文将按照如下顺序展开叙述，在第 2 节本文简述 Canny 算子的原理并从共性层面归纳多个自适应隐写算法，进而从图像分割角度总结基于率失真界的隐写优化目标；接着在第 3 节量化任意域失真代价转换公式，并针对总结出的优化目标给出求解方案 Canny Gauss；然后在第 4 节实验部分对比所提出算法和其他同类算法在不同作用域下的安全表现；最后，在第 5 节对全文进行总结和展望，其中，全文多次出现的符号如表 1 定义.

表 1 全文符号定义

数学符号	符号描述	数学符号	符号描述
K	Sobel 卷积因子	F	多贝西小波滤波核
$Gauss$	高斯模糊滤波	$Median$	中位数滤波
$\rho_{x_i\pm1}$	空域失真代价	rot_{180°	水平翻转，原位置顺时针旋转 180°
$\lvert\cdot\rvert$	取绝对值	$\lVert\cdot\rVert_p$	p-范数
$\boldsymbol{I}'$	非零元素为 1 的上三角矩阵	$Wiener$	维纳滤波
X^*	8 个方向上包含该像素的同尺寸像素块	$pad(\cdot)/unpad(\cdot)$	对称补白/取消补白
$\boldsymbol{A}_{(i,j)}$	第(i,j)个元素为 1 其余为 0 的矩阵	$B(k,l)$	以第(k,l)位置为中心的 8×8 像素块
$D(\cdot)$	汉明距离	$H(\cdot)$	信息熵
$Z(\cdot)$	配平系数	$Hist(\cdot)$	取直方图统计量
$\varepsilon(\cdot)$	平滑纹理范围上界	$\gamma(\cdot)$	复杂纹理范围上界

2 技术背景

2.1 Canny 边缘检测

Canny 作为图像处理领域最常用的边缘检测器被学者悉知，具体介绍参见文献[26]. 其计算过程由 6 部分构成：(1) 图像灰度化；(2) 高斯滤波去噪点；(3) 根据 Sobel 边缘算子计算像素点水平、垂直、主对角线、斜对角线四个方向的梯度；(4) 非极大值抑制；(5) 双阈值选择；(6) 双阈值连通及独立弱边缘剔除.

$$K=\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}\begin{cases}G_x=K\odot X\\ G_y=(-K)^{\mathrm{T}}\odot X\end{cases}\tag{1}$$

本文重点关注式(1)，$\odot$表示卷积操作不再对协相关的符号单独定义，将卷积核按水平竖直方向翻转即为协相关. Sobel 算子梯度计算和双阈值选择的设计，其滤波核沿图像 X 横轴 x 与纵轴 y 两个方向做卷积处理，重点评估图像边缘的特征即灰度值剧烈变化程度，表现为像素梯度变化. 同时，双阈值决定了 Canny 边缘检测所计算的图像纹理范围，对应在图像中表现为边缘线条的厚度. 这对于隐写算法是有意义的，为了保证嵌入的低检测性，选择纹理复杂度高的区域(图像纹理边缘区域)嵌入消息是隐写设计的常用手段. 而在边缘的信息的改变势必会引起统计特性的剧烈波动，因此需要容纳这种波动的具体范围. 类比向平静的水面扔石头会激起波纹时的冲激现象，波纹会随着波的传递逐渐减小直到消失. 在后文中将统一用 Canny 表示使用 Sobel 算子的 Canny 边缘检测器处理载体图像的过程.

2.2 自适应隐写设计共性分析

结合前面提到的，本节首先回顾几种主流的基于伴随式隐写码设计的自适应隐写算法，其次根据空域隐写的失真代价公式解释自适应隐写算法的设计逻辑，为后面导出变换域隐写转换公式实现自适应隐写算法的任意域转换提供技术背景.

如图 1 所示，目前主流的启发式算法可按空域、变换域(频域一般指 JPEG 域)、边信息域划分. 一般认为，最初实现模拟嵌入辅助伴随式隐写码设计的空域自适应隐写算法是 HUGO，但其设计初衷仅用来抵抗 SPAM 特征，导致其安全性过低. 此外，NS 与现有自适应隐写设计理念不同. GMRF 在 MIPOD 的基础上增加统计量，但实际表现略低于 MIPOD，因此本文不对这几种算法做过多介绍. 下面本文将从空域角度解释其中主要算法的隐写失真设计.

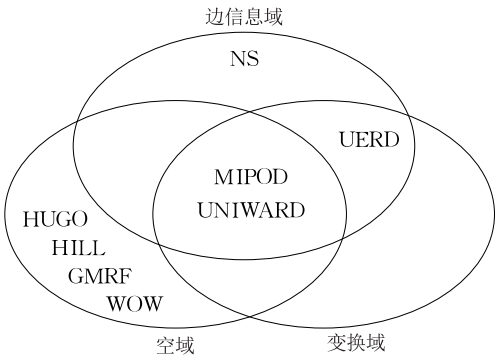


图 1 自适应隐写算法分类

$$\rho_{x_i\pm1}^{\text{S-UNIWARD}}=\sum_{t\in F}\frac{1}{|X\odot F^{(t)}|}\odot rot_{180^\circ}(|F^{(t)}|) \quad (2)$$

UNIWARD[5]. 图 2(a)首先通过多贝西小波滤波核 F 在 LH、HL、HH 三个方向进行卷积过滤，选择三个方向的过滤合并结果确定纹理平滑区域. 如式(2)所示，对经滤波的图像 X 取倒数，用于计算纹理复杂区块的位置信息，而对 F 过滤的结果取绝对值，被认为用于平衡三元嵌入中的加减变化；对 F 翻转后的滤波，作者同样解释为保护嵌入边界. 综上，UNIWARD 表明像素梯度变化剧烈(纹理复杂度高)的区域中相对平滑的位置更适合嵌入，这验证了为抵消嵌入时产生的信号冲激反映，需要向嵌入点周围的平缓区域传播一些额外的变化以平缓波动带来的剧烈影响.

(a)　(b)　(c)　(d)　(e)

图 2　从左至右分别为 UNIWARD、WOW、HILL、MIPOD、UERD 在空域下的嵌入失真代价图

$$\rho_{x_i\pm1}^{\text{WOW}}=\left\|\,|X\odot F|\odot rot_{180^\circ}(|F|)\,\right\|_p \quad (3)$$

WOW[6]. 图 2(b)与 UNIWARD 类似，对比式(2)、(3)，设计上主要发生变化的只有度量表达. WOW 由 p-范数的性质决定，隐写失真代价图表现出更加分明的层次. 这里给出一个不严谨的证明，范数的作用是度量赋范线性空间中向量的长度，对于三个方向上的滤波信息可以构建一个空间坐标系表示. 因此式(3)是在求 Cover 到小波滤波得到的平滑区域的距离. 显然，在去除平滑区域后剩下的部分即为复杂纹理，WOW 以此来计算适合嵌入的相对区域.

$$\rho_{x_i\pm1}^{\text{HILL}}=\frac{1}{|X\odot KB|\odot M_3}\odot M_{15} \quad (4)$$

HILL[7]. 图 2(c)是目前最简单高效的空域隐写设计，其构成如式(4)所示，对比式(2)，两个设计在形式上是相似的，其中，KB 表示 Ker-Böhme 滤波核[7]，M 表示均值滤波，其下标代表滤波核的大小. 强调一点，这里的 KB 核中心元素为 1 是为保留对应的中心待嵌像素，此时的 KB 核满足能量由中心向四周扩散的形式，这和像素点嵌入消息时引起的统计变化趋势是一致的. 同样地，KB 滤波用来获取纹理复杂区域，两次均值滤波用于获取纹理周围区域的边界. 这样结合前面的方法则可得出如下结论：UNIWARD 和 WOW 将获取纹理复杂区域和约束信号扩散边界的工作交给 F 同时处理，而 HILL 分离了该过程.

$$\begin{cases} r_n=\boldsymbol{G}a_n+\xi_n\approx X-Wiener_3(X) \\ \boldsymbol{G}\approx\left[DCT^{-1}(\boldsymbol{I}'_{(1,1)}\cdots\boldsymbol{I}'_{(1,p)}\cdots\boldsymbol{I}'_{(p,1)})\right]_{p^2\times q} \\ \hat{r}_n=\boldsymbol{G}\hat{a}_n=\boldsymbol{G}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}r_n \end{cases} \quad (5)$$

MIPOD[8]. 图 2(d)主要借助费雪信息获取嵌入位置. 费雪信息是一种评估极大似然估计值方差的统计量. 具体来讲，作者通过求载体像素区块残差的极大似然值与对应的线性模型最小二乘估计值的样本方差 $S^2(\xi)$ 确定费雪信息. 式(5)中 $r_n=\boldsymbol{G}a_n+\xi_n$ 为局部像素残差的线性模型[8]可由维纳滤波间求出，参数 $\hat{a}_n=(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}r_n$ 由该模型的最小二乘估计求得. 矩阵 $\boldsymbol{G}$ 是 $q=p(p+1)/2$ 个形如 $\boldsymbol{I}'$ 的矩阵做 DCT 逆变换构成，其中 $\boldsymbol{I}'$ 是非零元素为 1 的上三角矩阵，$\boldsymbol{I}'_{(k,l)}$ 是 $\boldsymbol{I}'$ 的第 (k,l) 个位置为 1 其余元素为 0 的 p 维方阵，用于评估某一像素区块下，每个元素对整体区块的信号能量变化的影响.

$$\begin{cases} S^2(\xi)=\dfrac{\sum(r_n-\hat{r}_n)^2}{p^2-q}=\dfrac{\sum\left[(I_n-\boldsymbol{G}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}})r_n\right]^2}{p^2-q} \\ \rho_{x_i\pm1}^{\text{MIPOD}}=\left(\dfrac{1}{S^2(\xi)}\right)^2\odot M_7 \end{cases} \quad (6)$$

该算法目标是求 ξ_n，即图像自然噪声误差的统计量，以此获得自然噪声的似然比检验区间. 而方差是判断数据波动的统计量，在隐写设计中能衡量局部像素块对额外噪声(秘密消息)的承载力. MIPOD 筛选出嵌入改变后仍满足自然噪声特性区域，而式(6)中的均值滤波正是平滑这个区域，这与围绕复杂纹理的区域筛选是一致的.

$$\begin{cases} X'=\sum_{t\in F}pad(X\odot F^{(t)}) \\ \rho_{x_i\pm1}^{\text{S-UERD}}=\dfrac{1}{unpad(X')+0.125\cdot\sum X^*} \end{cases} \quad (7)$$

UERD[9]. 图 2(e)原始设计仅围绕变换域，式(7)

是将其作用在空域的表达，该算法的原始思路是把嵌入位置的信号分散到周围的填充区域中以平衡统计波动，这里本文参考 UNIWARD 中的多贝西小波核 F 给出初步嵌入位置，并扩大局部扰动范围以分解统计信号的显著变化. 具体地，空域 UERD 将原始失真代价分配给待嵌像素块 X 按比例扩展后，与其相邻的 8 个方向的同尺寸邻接像素块 X^* 中，并按一定比例与原始失真累加，实现局部失真代价的均匀化.

通过以上几种设计可以得出结论：主流的隐写失真设计目标是明确纹理复杂区域发生变化时的异常信号扩散边界，寻找全局改动最小的局部熵最大的失真设计.

3 Canny Gauss 算法设计

在本节中，结合 2.2 节总结归纳的任意域隐写设计共识，首先归纳出根据实现通用域隐写失真设计的具体方法，然后根据纹理复杂位置的统计传播范围，构造 Canny Gauss 隐写算法. 下面，将分别展开叙述.

3.1 变换域隐写转换公式

从前面例子可以看出，空域下嵌入失真函数与图像纹理密切相关. 以 UNIWARD 为例，不同作用域的嵌入失真函数中纹理滤波输出是共用的，而发生变化的只有滤波出的纹理结果和各域间的失真表达. 为此，将现有的嵌入失真函数中有关图像纹理部分和失真代价表达部分分离，将独立出来的部分分别称为嵌入失真代价图和嵌入途径. 而任意的启发式算法都是在设计不同的嵌入失真，嵌入途径相对独立于失真代价，途径的改变势必会导致嵌入改变，失真代价会随着正交基(如 DCT、DWT、DFT 等)出现异常表达. 因此，通用域算法的设计关键是保证嵌入失真代价在嵌入途径变换后的稳定性.

$$\boldsymbol{A}'_{(i,j)}=DCT^{-1}(\boldsymbol{A}_{(i,j)})\cdot Q^{75}_{i,j} \tag{8}$$

以量化因子 75 时的 JPEG 图像为例，上式(8)是根据量化表 Q^{75} 计算的每个像素占整个像素块的信号能量比. $\boldsymbol{A}'_{(i,j)}$ 也被称为空域因子，在 UNIWARD 的任意域设计规范中被首先提出，用于实现空域失真代价向变换域的转换. 在 JPEG 格式的编解码过程中，8×8 图像块中的每个像素值都由对应的 DCT 系数的逆变换和量化表的乘积计算得出. 因此，原空域嵌入改变映射到对应的 JPEG 域时，失真代价的分量应执行同样操作，以保证变换域系数与嵌入代价匹配. 同样地，由于 DCT 系数将像素块的信号能量按量化表集中，导致原始卷积失效，所以变换域中卷积的表达应回归其统计学本质 $f(n)\odot g(n)=\sum_{x=-\infty}^{\infty}f(x)g(n-x)$，即卷积是某个随机分布函数的翻转与另一个随机分布函数在给定范围内的乘积再求和的结果. 此时 UNIWARD 在 JPEG 域第(k,l)位置上的嵌入代价的计算如式(9)所示：

$$\rho^{\text{J-UNIWARD}}_{k,l}=\sum_{i=1}^{2r}\sum_{j=1}^{2r}\sum_{t\in F}\frac{|\boldsymbol{A}'_{(i,j)}\odot F^{(t)}|}{|X\odot F(t)|_{B(k,l)}} \tag{9}$$

对比式(2)、(9)不难验证，式(9)的分母是其中一个随机分布函数 $g(x)$，其倒数可以类比函数翻转 $g(n-x)$，而 $B(k,l)$表示以第(k,l)个位置为圆心，r 为半径的方形像素块，取该像素块的空域失真用于 JPEG 失真设计是考虑到 JPEG 图像的 DCT 变换围绕 8×8 的块展开. 此外，将式(9)分母单独分离出来，等价于式(2)中缺少翻转卷积的 UNIWARD 空域失真代价，所以在 J-UNIWARD 设计中，分子的 $\boldsymbol{A}'_{(i,j)}$(对应另一个随机分布函数 $f(x)$)结合了 F 以实现翻转卷积过程. 现在给出通用 JPEG 域嵌入代价转换公式：

$$\rho^{\text{JPEG}}_{k,l}=\sum_{i=1}^{2r}\sum_{j=1}^{2r}|\boldsymbol{A}'_{i,j}|\cdot\rho^{\text{空域}}_{B(k,l)} \tag{10}$$

与 UNIWARD 不同，隐写失真 ρ 与空域因子 $\boldsymbol{A}'$ 以乘积形式连接，这是根据纹理滤波核筛选位置和失真代价的特性决定的，当最简形式排除了滤波核 F 的干扰后，乘积的形式满足 JPEG 格式 DCT 变换的正向过程. 同时，为了保护像素块信号的完整性，像素块半径 r 常常大于 4，通过 pad 显著提升变换域嵌入失真代价的表现.

$$\begin{cases}DCT^{\text{real}}=DCT(X-128)/Q^{75}\\ \boldsymbol{e}=\lfloor DCT^{\text{real}}\rceil-DCT^{\text{real}}\end{cases} \tag{11}$$

而在边信息隐写中，最为常见的嵌入方法被称为 Perturbed Quantization[27-28]，与变换域隐写不同，边信息隐写在已知量化表和原始图像 PreCover 的前提下，通过量化取整前后产生的系数变化嵌入秘密信息. 因此，取整前后的像素差值矩阵 $\boldsymbol{e}$(如式(11)所示)成为了主要修改目标. 由于取整操作的变换范围在[−0.5,0.5]区间内，对于靠近边界值的点，其像素值极易在量化取整过程中发生改变，这为 ±1 嵌入提供了冗余空间. 反观那些靠近区间中心的像素点在量化过程中并没有丢失信息，由于对低频信号的改动会造成图像的明显失真，因此优先不考虑

作为嵌入位置.具体地,当像素点差值近似±0.5 时将不执行嵌入操作,此时嵌入消息会向下一个像素区间延拓,破坏已经确定好的舍入边界;反之,当靠近±1 等整数边界时则应优先嵌入;同理当 $e_{(i,j)}=0$ 时关系到视觉影响不作为嵌入考量.为了简化边信息算法的计算过程,本文先给出通用边信息域 SI 方法[27]的嵌入代价转换公式:

$$\rho_{k,l}^{\mathrm{SI}}=\left|e_{k,l}-sgn(e_{k,l})\right|\cdot\rho_{B(k,l)}^{\mathrm{JPEG}} \tag{12}$$

其中,$sgn(e_{k,l})$表示量化边界,而量化误差 $e_{k,l}$与其边界差值的绝对值,即为 JPEG 域失真向 SI 边信息域转换的权重系数.当误差 $e_{k,l}$靠近边界时,二者差值接近 0.5,权重系数及其失真代价更小.同理,当 $e_{k,l}$值接近 0 时,二者差值接近 1,失真代价更大.这使得量化取整像素值损失较大的点优先于损失较少的点嵌入,以最简形式实现了量化误差掩盖像素修改产生的统计波动的具体过程.

但从高质量 PreCover 向低质量 Cover 压缩过程中,由于量化表引起的溢出问题一直以来未能妥善解决,这在一定程度上影响了 SI 边信息域隐写的安全性.因此,Butora 等人[28]提出了联合乘子 c_{ij}提升 SI 方法的隐蔽性潜力,具体如下:

$$\begin{cases} g_{i,j}=Q_{i,j}^{75}/gcd(Q_{i,j}^{75},Q_{i,j}^{100}) \\ c_{i,j}=\begin{cases}+1, & g_{i,j}od\,2=0\ \text{或}\ g_{i,j}=1\\ +\infty, & \text{其他}\end{cases} \\ \rho_{k,l}^{\mathrm{PQ}}=(1-2\left|e_{k,l}\right|)\cdot c_{i,j}\cdot\rho_{B(k,l)}^{\mathrm{JPEG}}\end{cases} \tag{13}$$

如式(13)所示,取 PreCover 的量化表为 Q^{100},其中$(1-2\left|e_{k,l}\right|)$是 JPEG 域向 SI 域转化的权重系数的另一种表达形式.因此,PQ 算法与 SI 算法的差异主要集中在 $c_{i,j}$,而 Butora 等人通过实验验证了高质量因子向低质量因子转化时,低质量因子对应的量化表 Q^{75}只有在被 Q^{100}完全约分或约分后的剩余因数为偶数时,截断产生的影响才更容易被吸收,反之则可能产生溢出(定理 4.1[28]).所以对于溢出部分需要赋予湿码以降低溢出对统计特征的影响,即 $c_{i,j}$所示.

至此,当已知空域的嵌入失真代价时,根据需求向对应嵌入域转换并计算对应的失真代价湿码范围和嵌入边界,即完成对应变换域的设计.同理,在获得了非空间域的嵌入失真函数时,也能够一一对应地反推对应的空域设计.实际上,各个域中得到 ρ 沿用 UNIWARD 的边界设定已经足够了,当然真正的湿码应该是能够首先排除的不适合嵌入的位置,这在 Canny Gauss 的设计中能充分体现.

3.2　Canny Gauss 隐写失真代价设计

现在,本文将简述 Canny Gauss 与主流自适应隐写算法在设计思路和关键性创新上的主要区别.首先需要指出,本文设计主要遵循 Fridrich 等人[4]提出的有关 MCMC 采样下等效嵌入思想,后续实验是通过 Gibbs 采样模拟固定有效载荷下的最大嵌入状态得出的.其中,Gibbs 族分布满足熵最大性质,但隐写代价要求全局改动最小,在嵌入失真函数设计阶段,这意味着全局修改有极大的概率使嵌入残差围绕在复杂纹理周围,此时嵌入需要的额外信息量更小、信息熵也更小;同理,需要修改的局部图像随机聚集形成不规则的簇,反映在嵌入过程中表示更大的不确定性,对应信息熵也更大.因此,本文将其归纳为全局熵最小和局部熵最大的双层优化关系.这与以往 UNIWARD 或 WOW 与之多贝西小波滤波、HILL 与之 Ker-Böhme 滤波、MIPOD 或 GMRF 与之自然噪声误差的统计量、UERD 与之 DCT 系数误差、UT-GAN/ADV-EMB 与之深度隐写分析特征等,基于隐写分析特征设计的嵌入失真函数设计原则存在本质区别.其次与主流方案类似的是 Canny Gauss 在选择熵最大区域范围时,是通过明确纹理复杂区域发生变化时的异常信号扩散边界获得的.本文认为在启发式图像隐写中二者是高度重合的,这一观点也是在对比了其他主流方案的嵌入失真代价图和安全性表现后得出的.将之作为一种简化运算的技巧,用来加速 Canny Gauss 在 AutoML 搜索阶段的收敛速度,就目的而言,这也与 HILL 等方案直接根据相关滤波核调整嵌入区域的方式存在较大区别.此外,Canny Gauss 的提出是用来验证前文提出的自适应隐写失真设计共识、通用域嵌入转换公式以及手工设计在抵抗深度检测器的潜力等目标.详细介绍其设计细节和具体实现,具体框架如图 3 所示.其整体实现流程可分为三个步骤:首先根据像素平滑区间IPSR(Inter Pixel Smoothing Ranges)计算干码区获得浅层嵌入范围,接着通过 Canny 算子与统计相关的纹理密度来计算集中嵌入区块边界 CGPM(Canny Gaussian cost Probability density Maps),最后借助 AutoML 技术全面提高算法安全性表现.

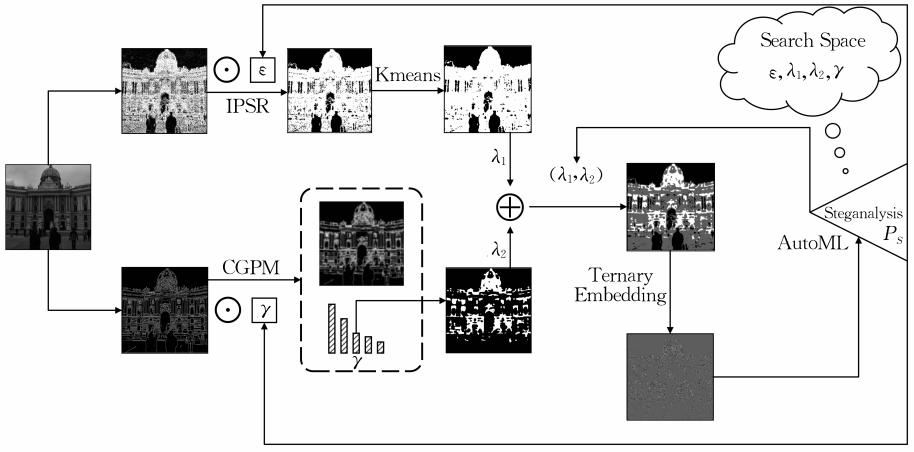


图 3 Canny Gauss 完整设计

3.2.1 像素平滑区间

当前隐写湿码往往被设计人员用于修饰图像的嵌入边界,以防止图像严重失真,作为嵌入失真代价图的收尾设计. 干湿码的设计起初是用于舍弃那些没有嵌入价值的位置,这类像素在整张载体图像的占比很高,因此干湿码分离极大的简化了嵌入计算量.

但在隐写码体系下,模拟和实际嵌入都逼近率失真界,在给定有效载荷时嵌入失真代价图的合适嵌入位置往往不够分配,导致嵌入信息向次优位置转移,严重制约最小化嵌入失真. 如图 4 所示,图 4(b)表示 bpnzAC(bit per none zero AC coefficient)为 0.4 时 SI-UNIWARD 根据图 4(a)计算的嵌入失真代价,而图 4(c)是三元模拟嵌入根据中间图像生成的空域残差. 白色椭圆标记的区域是该载体图像的两块湿码范围,且嵌入失真代价对该区域的标识表示这些区域相对不适合嵌入,但最终模拟嵌入后的实际结果表明,嵌入算法出现了秘密消息向次优位置转移的现象. 这是由于隐写术目标 $\min\limits_{H(\pi_\lambda)=m} D(X,Y)$ 的优先级高于嵌入失真代价 ρ,下游编码算法为了保证嵌入失真代价最小化,冒险选择了不适合嵌入的湿码区. 因此针对上述问题,考虑到直接在计算失真代价前期对显著位置约束会造成模拟嵌入逼近率失真界时消息的丢失,本文提出像素平滑区间 IPSR 以提高嵌入的优先级. IPSR 顾名思义是像素点之间相对平滑的区域,与纹理复杂的位置相比这些区域的局部差异较小. 其核心思想如下:

$$E_{diff\leq\alpha}\left(\sum(X^*_{i,j}-X_{i,j})\right)\leqslant\varepsilon(\alpha) \tag{14}$$

如式(14)所示,本文以阈值 α 分割纹理平滑和复杂区域,二者的边界可由局部残差期望与 α 的映射关系得出,且平滑区域残差存在上界 $\varepsilon(\alpha)$,由隐写分析特征决定. 这里借助 Gauss 模糊+K-Means 聚类求解 α 的映射,对 K-Means 聚类算法而言,前一步 Gauss 模糊在图像分割中充当放大作用,通过隐写分析器的反馈修正,解决嵌入失真代价中有关 α 的映射计算问题. 具体地讲,在给出一定的约束条件后,初步满足约束范围的区域通过 Gauss 模糊放宽,再由 K-Means 根据现有隐写分析特征给出相对应的聚类边界. 每个非边界像素都存在与其直接相连的 8 个邻居,而当该像素位于平滑区域时,有极大的概率满足该点与其邻接点的差值存在较小的数量关系乃至直接相等,而如果相邻像素的差值超过了阈值则认为该点处于纹理复杂区域. 现在进一步严格约束该假设,对于任意的平滑区域像素的相邻像素满足围绕该点的中心对称性质.

此时如图 5 所示,取该像素块←、↑、↖、↗四个方向的像素残差来表示该点与其周围邻居的数量关系,实心圆表示该点所在,空心圆对应向各方向平移后像素块中心位置,平移步长是 α 个像素单位,重叠的阴影部分即为残差. 那么,当图 5(a)、(b)、(c)、(d)

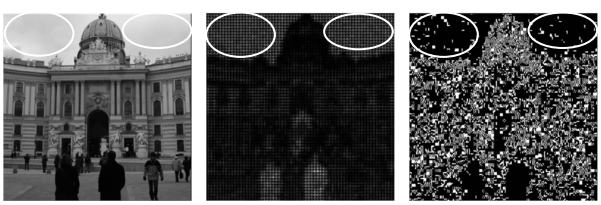


图 4 嵌入点溢出示意图

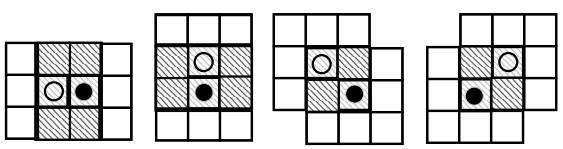
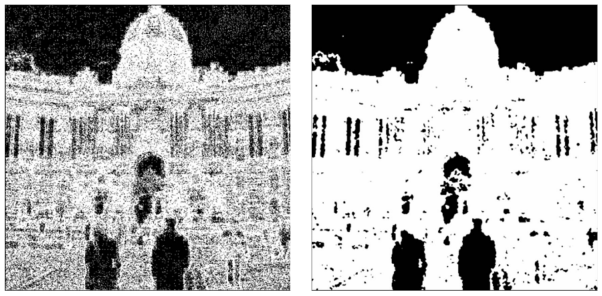


图 5 左侧、上方、左上和右上四个方向的像素残差示意图

4 种情况的残差 $r=a\bigcap b\bigcap c\bigcap d=0\neq\varnothing$ 时，则能严格确定该像素块在光滑区域中，此时该中心像素为湿点.

$$\mathrm{IPSR}=K\text{-Means}\left(\prod(X-X^{*})\odot Gauss\right) \quad (15)$$

但这种方式的约束过于严格且筛选出的区域离散严重(如图 6(a))，不利于后续设计. 因此如式(15)所示，IPSR 先通过 Gauss 模糊平滑残差，再借助聚类激活(如图 6(b)). 其中，Gauss 模糊的滑动窗口边长应设置为相邻像素数量与 α 减嵌入载荷对应信息量差值的和，但实际上该统计量应由隐写分析特征给出，此时称黑色的区域为 IPSR 即湿码区，对于该区域的 ρ 通常赋予一个较大的正值.



(a)　　(b)

图 6　IPSR 残差即聚类结果示意图

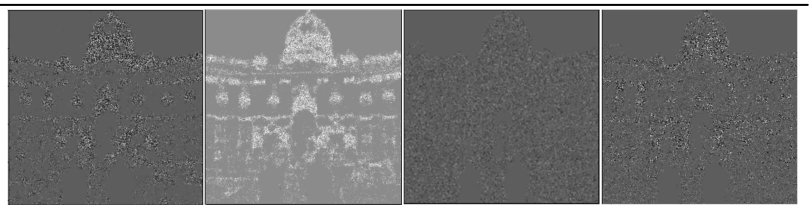
仔细观察图 6，(a)为直接平移计算的图像残差，(b)为经过 Gauss 模糊和聚类后的残差. 对于那些穿插在密集干码区的湿点，当 $\alpha=1$ 时，K-Means 将它们划分进了白色区域，实现在明确湿码的同时扩大干码范围，保障嵌入位置的充足. 同时干码区统计量随嵌入不断变化，因此在图 6(a)中，白色区域包围的湿点远比黑色区域中的干点安全. 而图 6(b) K-Means 的聚类标签为后文图 7(a)，因此 Canny Gauss 的干码区才能尽可能多的包含纹理，为纹理区成块提供基础. 其中，图 7(a)为 Canny 算子阈值为中位数时筛选出的纹理图，而图 7(b)为经过 Gauss 模糊和直方图筛选处理后的重点嵌入区域.

(a)　　(b)

图 7　CGPM 纹理及中位数筛选结果示意图

3.2.2　Canny 纹理密度筛选

若仅根据 IPSR 筛选出的干码区作为嵌入失真代价图模拟嵌入秘密消息，则将导致得到的像素分布更加混乱即更趋于熵增的异常结果，具体如图 8 所示. 尽管这种嵌入失真函数设计更符合率失真界约束，但对于隐写分析方而言，检测时增加的特异统计量严重削弱算法安全性. 换句话说，目前的隐写框架在给定载荷时，一味追求更低的嵌入改变量不能弥补由此带来的暴露风险. 因此应考虑从上游嵌入代价角度给出约束限制随机的、混乱的加性失真.



Schemes	HILL	MIPOD	Ours(IPSR)	Ours(Canny Causs)
Change Rate	0.0906	0.0948	0.0706	0.0965
P_S(MaxSRM d2q2)	0.5920	0.6547	0.6623	0.6196

图 8　Canny Gauss 与 HILL 和 MIPOD 在空域城堡灰度图像素改动及统计差异对比

如图 8 中所展现的，嵌入改变率(Change rate)刻画了嵌入秘密消息前后城堡灰度图的像素改变幅度，而 maxSRMd2q2 表示富特征共生矩阵扫描方向为“d2”且量化步长为 2 的 maxSRM 特征，用来验证各方案的空域统计隐蔽性[29]，同时 P_S 表示基于 MIPOD 训练的 maxSRMd2q2 特征在各图像上的检测正确率. MIPOD 和 HILL 的嵌入残差均向着纹理更复杂的区域聚集，这种聚集一方面加剧了与聚类相关的统计特异性，另一方面也显著降低了加性失真带来的其他负面影响. 尽管二者的稀疏部分根据呈现出嵌入差异，但残差的聚集趋向基本保持一致，这与图 7(a)的纹理稠密区域相对应. 因此有充分证据表明自适应隐写的实质就是根据有效载荷分离出上述纹理稠密区域. 此外需要补充说明的是，像素改动的幅度，以及残差向复杂纹理聚集的程度在适当范围内是正相关的，若算法仅关注降低像素

改变率,将出现图 8 中 IPSR 的隐写分析统计异常结果(像素改变率低但被检测率高);同理,若失真设计仅关注复杂纹理的聚集,也会导致嵌入范围受限,迫使残差向湿码点转移影响隐写隐蔽性.因此,尽管 Canny Gauss 的空域像素改动数量更大,但通过构造 IPSR 与 MIPOD 之间的纹理残差聚集状态,可以避免上述两种极端状况的发生.

由于模拟嵌入的随机性质,在密集和稀疏区域内的嵌入残差呈现均匀分布的特点,此时每个集中区域较离散区块表现出更密集的混乱状态.若此时列出一个优化关系,则可表述为求尽可能降低残差点向纹理密集区聚集引发的像素改变时,满足局部区块内最大熵的决策,如式(16)所示:

$$L=\sum_{i,j\in X^{\text{rough}}}\min_{D(y_{i,j})}\max_{H(\pi_\lambda)}$$
$$\text{s.t. } H(\pi_\lambda)=-\sum_{t\in I}\frac{1}{Z(\lambda)}e^{-\lambda D(y)}\log\frac{1}{Z(\lambda)}e^{-\lambda D(y)} \tag{16}$$

根据图 8 所示,结合复杂纹理对应的各算法局部残差,以及各残差的隐写分析结果可知,呈现出无序状态的局部残差在增大熵值 H 的同时,起到了减小像素改动 D 的作用,与干码区完全无序的随机嵌入相比,这显著降低了统计异常风险.综上,Canny Gauss 的核心思是从图像分割的角度对这一特殊现象进行再现,如图 7(b)所示,本文通过对纹理密集区域进行特定的失真设计,使其残差结果向 HILL 和 MIPOD 的集中嵌入区域靠近.

$$\begin{cases}X^{\text{Blur}}=Canny(X)\odot Gauss_{15}\\ \gamma=Median[Hist(X^{\text{Blur}})]\end{cases} \tag{17}$$

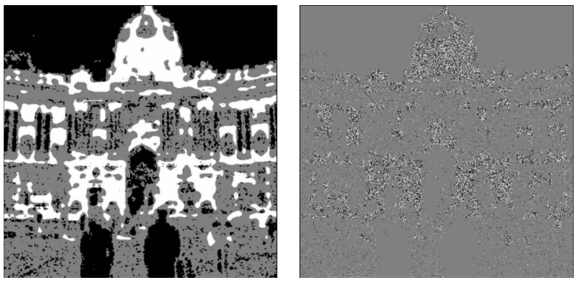
具体而言,根据 Canny 算子获得纹理进而定义干码区的重点嵌入密度,此时纹理聚集程度与密度呈正相关.如式(17)所示,首先通过大范围的 Gauss 模糊柔化纹理图像分离稀疏和稠密纹理区块;接着使用直方图处理柔化后的纹理图并取其中位数作为密度分界阈值;然后以该阈值划分柔化纹理图得出对应嵌入密度,表示为式(18):

$$X^{\text{Density}}=\begin{cases}X_{i,j}^{\text{Blur}}\left(1+\dfrac{255-X_{i,j}^{\text{Blur}}}{255}\right), & X_{i,j}^{\text{Blur}}\geqslant\gamma\\ 0, & \text{其他}\end{cases} \tag{18}$$

最后接着将 IPSR 和 X^{Density} 归一化后的线性组合再取倒数,即为 Canny Gauss 空域嵌失真代价 ρ:

$$\rho_{\text{空域}x_i\pm1}=\frac{1}{\dfrac{IPSR(\varepsilon)}{\lambda_1}+\lambda_2X^{\text{Density}}(\gamma)} \tag{19}$$

此时由嵌入失真代价引导的像素改变概率如图 8 所示为 0.0965,这一结果也证实了前文提到的嵌入失真会由算法设计因素而增大的结论.在获得基础干码失真图与重点嵌入密度图后,按照式(19)实现前文总结的隐写优化目标,结果如图 9 所示,(a)为改造后的嵌入失真代价率图,(b)为模拟嵌入后的空域残差.Canny Gauss 通过筛选干码区,极大地提高了嵌入的可选择范围以降低溢出影响;同时将纹理密度较高的区域,根据嵌入比重再次划分以增大局部复杂纹理区块占比.



(a) (b)

图 9 失真代价及空域残差示意图

Canny Gauss 计算流程如算法 1 所示,当确定好待嵌载体和秘密消息后,对各项超参数按照式(15)和(17)进行初始化.载体图像经过算法的 4～6 行的结果如图 9(a)所示,在经过 IPSR 粗筛和 X^{Density} 精筛后,重叠的结果即为所求.对于任意聚类算法而言,其结果只标定了与聚类中心相关的簇,而具体标签需要人为定义.对于当前问题为 2 类的簇,其中一组与粗筛结果相背的区块显然错误,7～9 行是当精筛与粗筛相交重合部分较少时,对此结果取反的操作.算法的 12、13 行分别对应了不同作用域下的嵌入失真函数的具体计算规则,Canny Gauss 默认构造空间域失真函数并通过前文的通用域转换公式,转换到频域和边信息域,最后,根据近优模拟嵌入或几种主流的隐写编码方案[1,30-31]完成嵌入流程.在实验部分,具体比较了基于嵌入失真代价图分割的 Canny Gauss 与其他原生与转换方案的安全性表现,这进一步验证了图像分割在隐写失真设计中的独到之处.另外,阈值 γ 初始由隐写分析中最常见的特征直方图统计量给出,换句话说 γ 的取值不唯一且一般应由隐写分析相关特征决定.因此,为进一步提高隐写安全性,密度分割标准 γ 不妨由隐写分析特征集给出.

算法 1. Canny Gauss.

输入:载体图像 *Cover*;超参数集 *hy_set*;随机种子 $rs=1$;$\varepsilon=$式(15);$\gamma=$式(17);$\lambda_1=\lambda_2=10$

输出:含密载体 *Stego*

1. IF *hy_set* is None:

2.　Initialize；
3. END IF
4. $dry\_blur=Gauss(IPSR(Cover),\varepsilon)$；
5. $canny=Canny(Cover)$；
6. $dry=K\text{-}Means(dry\_blur)$；
7. IF $dry\cap canny>(\sim dry)\cap canny$：
8.　$dry=\sim dry$；
9. END IF
10. $canny\_blur=Gauss(canny,\gamma)$；
11. $canny_density=$式(18)；
12. $\rho_{空域}=$式(19)；
13. $\rho_{频域/边信息域}=$式(10)/式(12)、(13)；
14. STC,SPC,SPSC,模拟嵌入；
15. return *Stego*

3.2.3　AutoML 超参数搜索

由于 Canny Gauss 从划分干湿码和分离局部嵌入密度 2 个角度刻画了信息嵌入过程中引起的统计波动，因此隐写失真代价式(18)围绕着隐写分析决策与阈值 ε,γ 密切相关. 由于从两个不同方面刻画隐写嵌入设计，因此需要额外的配平系数以保证设计的准确性. 当前由隐写分析特征求出的未知数个数为四，对于求解一个四元高次方程，在可行域下的蒙特卡洛方法自然而然就成为了首选的求解方案. 伴随着 AutoML 技术的推广普及，蒙特卡洛方法下的优化和计算方式被广泛集成和工具化，使目标求解更加简单易行. 目前 AutoML 被广泛应用于阈值选择、特征工程、模型构造等各式机器学习任务，而基于 AutoML 的阈值筛选策略日趋完善，催生出包括进化计算、模拟退火、随机下降、神经网络等大批可行有效的具体实现. 本文借助 AutoML 框架 NNI 实现对优化目标的高效求解，搜索过程如算法 2 所示.

算法 2.　超参数搜索.

输入：载体图像数据集(如 Bossbase1.01、BOWS2 等)；超参数变量集 $hy\_set(\varepsilon;\gamma;\lambda_1;\lambda_2)$；此时 $\varepsilon,\gamma\in 2k+1,k\in N^+<10;\lambda_1;\lambda_2\sim U_{(1,10)}$

输出：单位时间内最优参数集 *best_hy_set*

1. $stack=0;score=$INF$;m=size($datasets$)$
2. WHILE $current\_step<max\_steps$：
3.　$current\_set=hyparameter\_search$(Random Walk, Evolution, Anneal)
4.　FOR i in range(m)：
5.　$stegos[i]=$算法 1($datasets[i],current\_set,rs$)
6.　$current\_score+=steganalysis(stegos[i])/m$
7.　END FOR
8.　IF $curren\_score<the\_lowest\_score$：
9.　PUSH *current_hy_set* to *stack*
10.　broadcast(Random Walk,GA,Anneal)
11.　resample($hyparameter\_interval$)
12.　ELSE：
13.　random resample($hyparameter\_interval$)
14.　IF $size(stack)>10$：
15.　POP *worst_set* from *stack*
16.　END IF
17.　END IF
18. END WHILE
19. return *best_hy_set*

由算法 1 可知，算法 2 的目标是借助 AutoML 技巧求解影响嵌入失真的四个关键超参数，而这组解的取值与隐写分析器的特征直接相关，本文选择多种 AutoML 计算模式，包括随机算法、进化计算与模拟退火，而不同方案间取值的变换范围也存在较大区别，因此算法 2 通过筛选在各算法当前轮更新的取值中安全性表现最好的解作为下一轮搜索的公共起始值，并将上述方案得出的优势结果依次进栈，保留安全性表现最好的 10 项解集在栈内，当搜索时长达到规定上限或安全性指标收敛时按入栈顺序抛出栈顶解集即为所求.

综上，此时算法求出阈值明确失真设计函数的分割边界，再交给隐写分析器做简单判别，通过搜索筛选出满足最大安全性的超参数集作为最终嵌入方案. 与以往工作不同，Canny Gauss 用于对抗的隐写分析器可以是任意分类器的线性组合，若指定某个训练好的深度隐写分析器，此时搜索出的阈值能大幅度降低隐写分析表现原理；而当使用 SRM 等富特征时[32]，新的阈值能规避灰度共生矩阵等相关统计量特征以提高安全性. 为了提高单次搜索的泛化性，本文在几种主流的隐写分析方案中进行初步筛选，最终选择训练完备的 SRNet 的软输出作为超参数取值的安全性指标. 上述安全性表现与结论在实验部分得到了验证.

4　实验分析

4.1　实验设置

目前隐写分析常用 $P_E=\dfrac{P_{FA}+P_{MD}}{2}$ 及其改进公式 $P_{E^*}=\min\limits_{P_{FA}}(P_{FA},P_E)$ 作为隐写安全性评价指标[33]，其中 F_{FA} 为 False-Alarm，P_{MD} 为 Missed-Detection，二者分别表示常见的样本假阳和假阴概率. 对比目前深度学习环境下二分类问题的误差评价方法 $P_S=P_{TP}+P_{TN}=1-P_E$，在实际编程阶段更为科研人员所熟知，在分类器完备状态下给定载体图像时 $P_{FP}\approx$

P_{FN}，因此本文使用更常见的 P_S 即分类正确率作为误差评价标准①. 同理，*AUC*(Area Under Curve)表示 ROC(Receiver Operating Characteristic)曲线与 x 轴的面积，用来评价集成分类器下袋外数据的分类正确率，也就是在载体已知的条件下集成分类器测试集的正确分类概率.

本文实验部分所用数据集灰度图为 Bossbase1.01 数据集，彩图为 Alaska2 比赛数据集，各取前 1000 张图像计算给出如下实验数据. 其中针对集成学习类隐写分析判别采用 Scikit-learn 框架计算，配置 FLD(Fisher Linear Discrimination)子分类器 50 个，Bootstrap 袋内采样数据占比 50%、随机子空间特征占比 80%，训练和测试样本按给定随机种子从 1000 张图片中对半划分；深度隐写分析器根据第三方权重分别采用 Pytorch 和 Tensorflow 两种框架，其中自训练分类器均为 Pytorch 复现，训练集和验证集由各数据集除 1000 张图片外的剩余图片按 9∶1 比例划分. 所有训练验证测试过程和 *Stego* 样本生成过程随机种子设置为 1，实验对比的所有嵌入域参考 UNIWARD 设计，对比算法均来自原作者公开代码. 其中，部分非通用域算法根据本文提出的通用域隐写转换公式转换到对应域参与对比分析，名称前加"*"用以区分.

4.2 通用域实验分析

在本节中，给出各算法在空域、JPEG 域、边信息 SI 域和 PQ 域在不同测度下的安全性对比. 为简化实验流程嵌入算法的有效载荷固定为 0.4bpp/bpnzAC，并根据各域的性质绘制载荷或质量因子增加时 P_S 变化情况的点线图.

4.2.1 空域实验分析

本小节将展示空域模拟嵌入时主流的三种隐写分析检测器对各算法的检测情况，实验结果如表 2 和表 3 所示. 结果数值越小安全性越高，其中 SRNet 为第三方用户贡献权重，后两种深度检测器模型来自 Alaska2 隐写分析比赛由 Tesla K80 训练. "″"表示当前分类器使用第三方权重测试，此时 SRNet 使用 brijeshiitg 开源权重②，此外 SE-ResNet 及 EfficientNet-b0 为 Alaska2 隐写分析比赛金奖和银奖模型，基于 UNIWARD 训练权重；而 P_S(maxSRM)是根据 MIPOD 训练得到的，区别于在已知载体和隐写失真条件下的分类器 *AUC* 结果，这里使用 P_S 作为检测标准.

表 2 空域灰度图 bpp=0.4 安全性比较

Metrics	Schemes						
	S-UNIWARD[5]	WOW[6]	MIPOD[8]	HILL[7]	*UERD[9]	GMRF[10]	Canny Gauss
AUC(PSRM q1)	0.6922	0.7322	**0.6250**	0.6511	0.8926	0.6894	0.7878
AUC(maxSRM q2d2)	0.7228	0.7503	**0.6497**	0.6616	0.7874	0.7027	0.7835
P_S(PSRM q1)	0.6500	0.6600	**0.6200**	0.6500	0.7700	0.6600	0.7000
P_S(maxSRM q2d2)	0.7210	0.7315	0.7035	0.7000	0.6990	0.7300	**0.6925**
P_S(SRNet″)	0.7260	0.7660	**0.6195**	0.6410	0.7280	0.6940	0.6200
P_S(SE-ResNet)	0.7180	0.6665	0.5955	0.5975	0.7525	0.6110	**0.5125**
P_S(EfficientNet-b0)	0.7640	0.7450	0.5985	0.6240	0.7895	0.6615	**0.5145**

表 3 基于 ρ 修改算法的空域灰度图 bpp=0.4 安全性比较

Metrics	Schemes					
	ADVEMB-SUNIWARD[20]	ADVEMB-HILL[20]	ADVEMB-MIPOD[20]	ReLOAD-HILL[23]	ReLOAD-MIPOD[23]	Canny Gauss
AUC(PSRM q1)	0.7070	0.6356	0.5884	0.5508	**0.5324**	0.7878
AUC(maxSRM q2d2)	0.9006	0.8874	0.8709	**0.6399**	0.6429	0.7835
P_S(PSRM q1)	0.8700	0.8600	0.8800	**0.6200**	0.6400	0.7000
P_S(maxSRM q2d2)	0.8310	0.8095	0.8175	**0.6770**	0.6920	0.6925
P_S(SRNet″)	0.5530	**0.5505**	0.6280	0.6260	0.5905	0.6200
P_S(SE-ResNet)	0.8015	0.7670	0.7265	0.5975	0.5305	**0.5125**
P_S(EfficientNet-b0)	0.8500	0.7970	0.7580	0.6145	0.5490	**0.5145**

如图 10 通过算法 2 的搜索，其中 c_k_size，d_k_size，c_multiplier，d_multilplier 分别代表超参数 ε，γ，λ_1，λ_2，default 表示 SRNet 在第三方权重下的隐写分析判别精度. Canny Gauss 根据采样函数和判别条件在参数空间选择出合适的数值解，此时搜索轮数仅为 1500 轮，一轮的平均搜索时间小于 1 min. 结合图 11，当前的搜索结果验证了 AutoML 对 Canny Gauss 的安全性提升效果. 此时，针对第三方

① 当未知载体数据时，此时在样本空间中待测样本的统计代价与载体代价不平衡，须将 P_{FA} 与 P_{MD} 分开处理，即采用 P_{E^*} 防止出现第一类(拒真)错误.

② https://github.com/brijeshiitg/Pytorch-implementation-of-SRNet

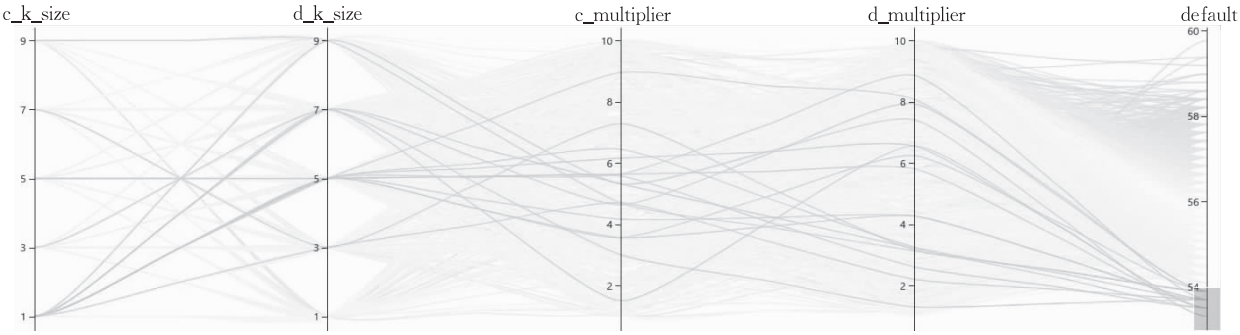


图 10　NNI 框架超参数收敛曲线

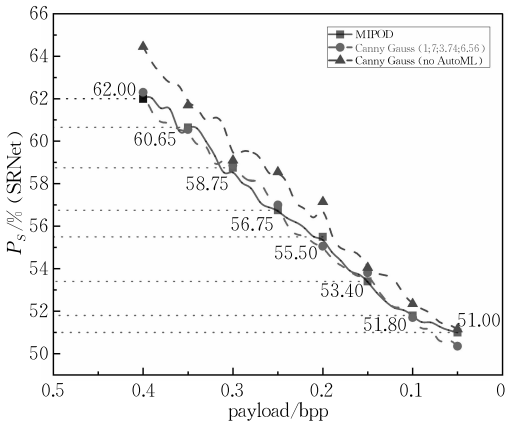


图 11　Canny Gauss 经过 AutoML 空域安全性的提升效果图

权重 SRNet 的检测，本文所提算法，对比载荷为 0.4bpp 的S-UNIWARD隐蔽性提升了 14.6%，接近 MIPOD 的安全性表现. 同时，在低于 0.4 bpp 的嵌入状态下，Canny Gauss 表现出了略优于 MIPOD 的深度隐写分析特征隐蔽性.

本文根据大量实验得出如下结论，类似于集成分类器的方式，由袋外数据产生的分类精度作为判断隐写安全性的方案也是有偏的. 尽管这类方法可以通过极大似然比检验，验证其在一定范围内的安全性. 然而，对于 Canny Gauss 而言，借助 AutoML 工具将袋外数据分类正确率作为奖惩反馈机制则能轻易绕过对应的集成式隐写分析特征. 同时目前集成式隐写分析结果由集成 FLD 训练得出，FLD 由特征集费雪信息按类内与类间的比值训练分类. 以空域下 MIPOD 和 UNIWARD 为例，结合表 2，MIPOD 训练的特征分类器对 UNIWARD 的安全性缺乏有效判断，而使用 *AUC* 值作为判断标准则很容易认为 MIPOD 比 UNIWARD 更加安全，但实际上这是已知另一种算法的设计思路和原始样本为前提的. 当 FLD 目标是训练以局部像素块方差关系的 MIPOD 时，自然会对以滤波为主的UNIWARD 和 HILL失效，反之也是如此；至于在集成训练时 HILL 和 MIPOD 表现出更高的袋外数据分类精度，与 FLD 集成分类器的分类依据和统计特征的费雪信息存在关联，刻意降低某些显著性特征而引入的新的特征对深度分类器而言未必更安全. 因此不应盲目地认为 HILL 和 MIPOD 在整体安全性上高于 UNIWARD，这从后几组变换域实验的图像质量和深度分类器的判别结果上可以看出.

同理，在表 3 中，Canny Gauss 仅在深度分类器下表现出优势，对比基于 SRNet 对抗设计的 ADV-EMB 算法，在各自 ρ 下均表现出对 SRNet 特征明显的隐蔽性，但对应的 maxSRM 特征则高度敏感，而基于 SRM 特征对抗训练的 ReLOAD 在 HILL 和 MIPOD 的集成分类器特征上表达出更强的安全性表现，但对基于 SE-Block 结构的两种深度隐写分析器优势不明显.

4.2.2　JPEG 域实验分析

由于在当前域变换域体系下UNIWARD 和WOW 的变换结果高度相似，因此接下来的实验将省略 WOW 的相关结果. 以表 4 为例，各算法在变换域彩色图像下的安全性表现，数值越小误检率越高. 本文

表 4　JPEG 域彩图 bpnzAC＝0.4、QF＝75 安全性比较

Metrics	Schemes					
	J-UNIWARD[5]	J-MIPOD[34]	*GMRF[10]	*HILL[7]	UERD[9]	Canny Gauss
AVG-SSIM	**0.9988**	0.9954	0.9986	0.9987	0.9987	0.9982
AVG-PSNR	**53.0542**	48.6025	52.0566	52.6855	52.9680	50.8615
AUC(DCTR)	0.5700	0.6700	0.6800	0.6700	**0.5500**	0.6900
AUC(GFR)	0.6100	0.6800	0.6700	0.6700	**0.6000**	0.6200
P_S(DCTR)	0.9453	0.7168	0.7811	0.8016	0.7416	**0.6900**
P_S(GFR)	0.8726	0.7079	0.7989	0.8074	0.7305	**0.6974**
P_S(Efficient-b0″)	0.7045	0.6310	0.7340	0.7325	0.6295	**0.6015**

额外对变换域的图像质量做出评估，这是因为变换域信号在经历有损变换后有效载荷接近 0.4bpnzAC 会产生肉眼可见的溢出伪影，同时三通道彩色图像主要以 Y 通道嵌入为主，将导致其与 Cr、Cb 通道匹配性不足出现额外色差，因此变换域下图像质量对算法的实际表现有重要意义. 此时 EfficientNet 使用的权重来自第三方隐写检测工具 aletheia①，基于 J-UNIWARD 训练，由于 JPEG 变换域的特殊性，该权重也被用于测试相关边信息域. 结合表 4 与图 12(a)，此时 Canny Gauss 在 JPEG 域使用的超参数并未发生改变，在空域根据隐写分析特征搜索的数值解在变换域中依然成立，由此说明该算法的嵌入失真设计在经过 DCT 变换后仍能正确表达，这是自适应隐写算法抗正交基变换的典型例子.

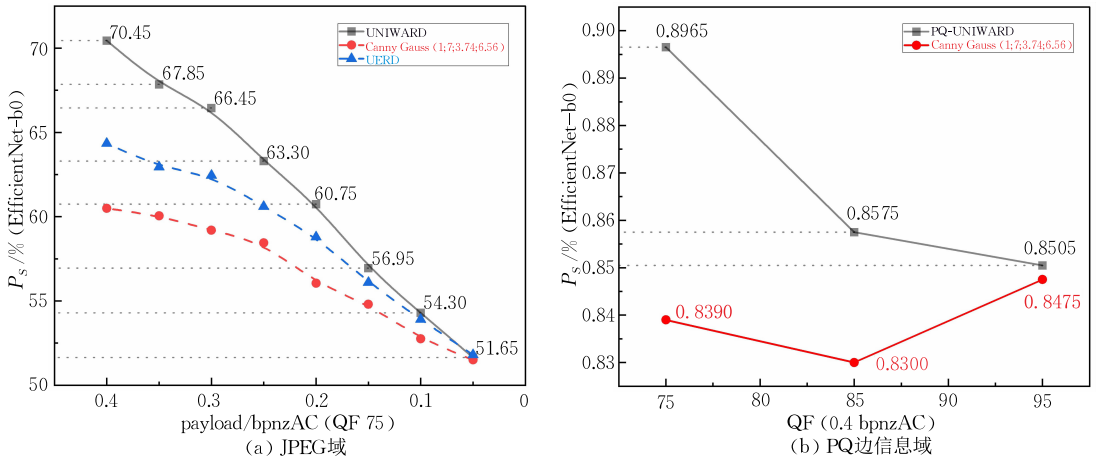


图 12 Canny Gauss 在 JPEG 域和 PQ 边信息域的实际安全性表现

4.2.3 边信息域实验分析

正如前文所提，边信息域存在的两种不同实现方式 SI 和 PQ[27-28]. 因此，本小节分别以表 5、表 6 给出两种嵌入途径下的实验结果. 由于 Alaska2 数据集包含 75、90、95 三种不同质量因子的图像，因此本文假设随机抽取的图像量化系数为 100 并作为 PreCover，图 12(b)中 PQ 边信息域图像被检测率，是按照不同的质量因子生成 Cover 后交给 Efficient-Net 得出的. 此时 MIPOD 等算法对 P_S 表达出更高的安全性，但图像质量开始出现下滑，这在一定程度上说明该算法在正交变换后信号随量化截断出现异常扩散，若此时以图像质量作为隐写图像的检测标准，那 MIPOD 的优势将完全丧失. 另外 UERD 在 SI 域下表现出较高的安全性和图像质量，但在同为边信息域的 PQ 域表现不尽如人意，这也验证了即使在相同域不同嵌入途径下，未修正的异常信号会导致出现伪影降低图像质量的问题. 但反观作为 SI 和 PQ 域的基础算法 UNIWARD 对图像质量和集成分类器特征处理上表现得更好，这与 UNIWARD 针对各域规范不同湿码防止边界溢出的设计是密不可分的. 而本文所提算法，在第三方权重的 Efficient-Net 检测下与 SI-UNIWARD 相比，隐写隐蔽性也存在不低于 2.6% 的提升. 同时由于 PQ 边信息域与 SI 边信息域的高度相似性，PQ-UNIWARD 在 0.4bpnzAC 的安全性与 SI-UNIWARD 相比并无较大改善，但 Canny Gauss 在 PQ 边信息域却提升明显. 综上，本文认为对于上述各算法总能找到某个域使某一算法在图像质量和安全性上得到平衡，表现出更高的实验精度，但尽可能降低由域变换带来的改造成本是不可忽视的因素.

表 5 SI 边信息域彩图 bpnzAC=0.4、QF=75 安全性比较(基于 PreCover)

Metrics	Schemes					
	SI-UNIWARD[5]	SI-MIPOD[35]	*GMRF[10]	*HILL[7]	UERD[9]	Canny Gauss
AVG-SSIM	**0.9030**	0.8389	0.9022	0.9027	0.9024	0.9016
AVG-PSNR	**34.9312**	32.0773	34.8153	34.9005	34.8587	34.7397
AUC(DCTR)	0.9200	0.9999	0.9000	0.9200	**0.8800**	0.9000
AUC(GFR)	0.8400	0.9900	0.8800	0.9100	**0.8200**	0.9000
P_S(DCTR)	0.9747	**0.8068**	0.9579	0.9668	0.9384	0.8842
P_S(GFR)	0.9700	**0.7153**	0.9626	0.9700	0.9347	0.9058
P_S(Efficient-b0″)	0.8965	**0.6845**	0.8945	0.8965	0.8765	0.8730

① https://github.com/daniellerch/aletheia

表 6　PQ 边信息域彩图 bpnzAC＝0.4、QF＝75 安全性比较（基于 PreCover）

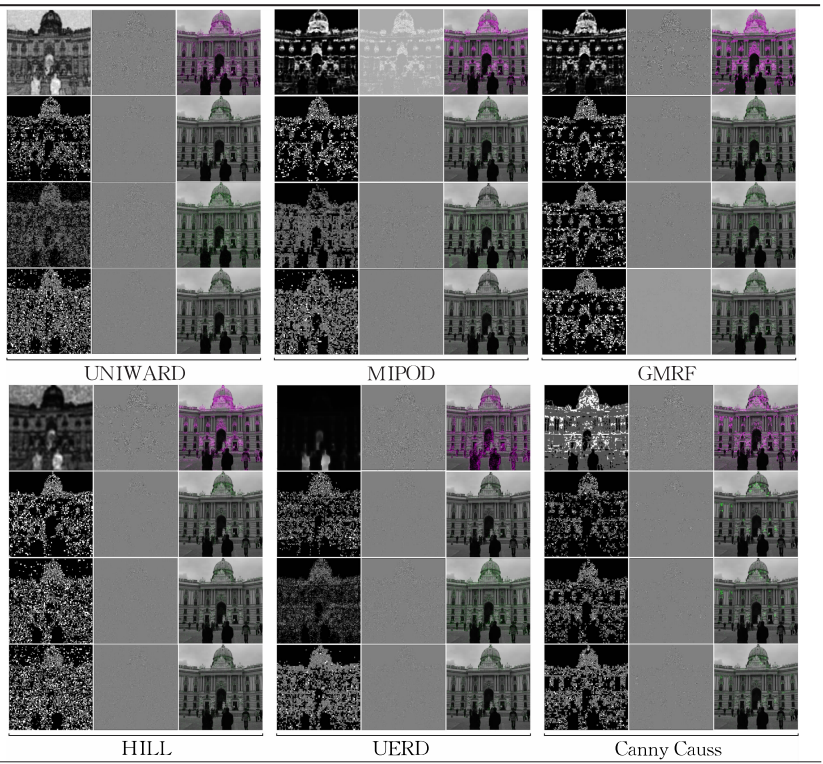
Metrics	Schemes					
	PQ-UNIWARD[28]	*MIPOD[8]	*GMRF[10]	*HILL[7]	*UERD[9]	Canny Gauss
AVG-SSIM	**0.9020**	0.8650	0.9014	0.9011	0.8813	0.8926
AVG-PSNR	**34.8339**	32.6300	34.7251	34.7911	32.9335	33.9639
AUC(DCTR)	**0.9000**	0.9999	0.9295	0.9600	0.9997	0.9900
AUC(GFR)	**0.9100**	0.9900	0.9328	0.9500	0.9999	0.9700
P_S(DCTR)	0.9879	**0.6616**	0.9832	0.9832	0.7079	0.9495
P_S(GFR)	0.9858	**0.7089**	0.9742	0.9832	0.7584	0.9047
P_S(Efficient-b0″)	0.8965	**0.6845**	0.8965	0.8945	0.6985	0.8390

结合表 5 和表 6 可知，Canny Gauss 算法在边信息域方面没有取得所有指标的最佳效果，但相较于其他算法，其表现出更强的稳健性和鲁棒性. 例如，在 SI 边信息域针对 DCTR 特征的实验中，Canny Gauss 在多数指标下优于同样经过转换的 GMRF 和 HILL 算法. 虽然 Canny Gauss 在图像质量方面的表现略逊于 SI-UNIWARD 算法，但其被检测率却明显降低，且与 SI-MIPOD 算法相比，Canny Gauss 不需要以图像质量为代价来获得更好的被检测率结果. 同时，在 PQ 边信息域中，PQ-UNIWARD 着重关注了图像质量和对其集成分类器的误导倾向，但对 EffcientNet 深度隐写分析特征的隐蔽性不足，因此其他算法根据变换域转换公式可以获得更高的抗隐写分析水平，而 Canny Gauss 算法在 PQ 边信息域中对像素质量的影响也相对更小. 因此在实际应用中，Canny Gauss 针对各嵌入域转换后的像素改变相较其他方案更具优势.

此外如图 12(b)所示，基于 JPEG 域训练的深度分类器在获得 PreCover 图像后，也表现出了较高的精确度. 但这在真实场景下是很难获得的，本文仅用于验证各算法在域变换后的安全性. 若改用对应压缩后的 Cover 图像作为参考样本，深度分类器的检测水平将大幅度下降. 由于边信息嵌入与量化压缩同向时能够掩盖住大量靠近纹理的嵌入，这相当于变换域嵌入降低载荷的结果. 进而，这种改变嵌入途径获得隐写隐蔽性的方案尚无有效的对抗手段，研究嵌入途径及其泛化方法仍至关重要.

4.3　实验总结

观察图 13 各算法在不同域间的嵌入残差变化，



Schemes	UNIWARD	MIPOD	GMRF	HILL	UERD	Canny Causs
Robust Ratio	0.0046	0.0060	0.0063	0.0053	0.0045	0.0089

图 13　各算法空域残差及其鲁棒性展示

从上到下分别表示空域、JPEG 域、SI 边信息域和 PQ 边信息域，其中各部分左上角图为空域下嵌入失真代价. 除 GMRF 和 Canny Gauss 外，其他算法在城堡灰度图的 PQ 域均存在湿码边界溢出的现象，极大地降低了算法的统计隐蔽性. 此外，各算法在 JPEG 变换域均未呈现与空域一致的嵌入分布，由于嵌入失真代价图经过 DCT 变换前后出现了表达损失，AC 系数的可嵌入位置比空域 LSB[36] 更广泛，导致基于纹理设计的算法优势降低，因此需要评估各算法在变换域失真代价的鲁棒表达. 本文引入残差鲁棒率(Robust Ratio)验证各算法在经过变换域转换后的嵌入表达水平. 残差鲁棒率计算公式为 $(diff_{空域} \cap diff_{JPEG} \cap diff_{SI边信息} \cap diff_{PQ边信息})/m$. 残差鲁棒率由 Bossbase1.01 数据集前 1000 张图像取均值计算得出，结合各算法的残差鲁棒率，本文验证了 Canny Gauss 在抵抗正交基变换时的鲁棒性，实验结果表明本文所提算法在其他各变换域的嵌入代价表达更贴合原始空域.

此外，针对 Canny Gauss 中用于实现 Gauss 模糊缩放复杂纹理轮廓的 IPSR 模块，和借助 Canny 算子和中位数滤波筛选纹理的 CGPM 模块，以及基于对抗判别的 AutoML 阈值搜索模块，本文分别进行了消融分析，结果如图 14 所示. 本文根据 SRNet、SE-ResNet 和 EfficientNet 三种深度隐写分析器在 0.1 bpp～0.4 bpp 下的嵌入被检测率，验证了各模块对算法统计隐蔽性的提升. 其中，在具有 SE-Block 的深度隐写分析器中，IPSR 模块起到了降低统计特异性的关键性作用，而 CGPM 和 AutoML 模块也有效降低了 Canny Gauss 的被检测率，在 0.4 bpp 时尤为明显，缺少各模块都会造成安全性的降低.

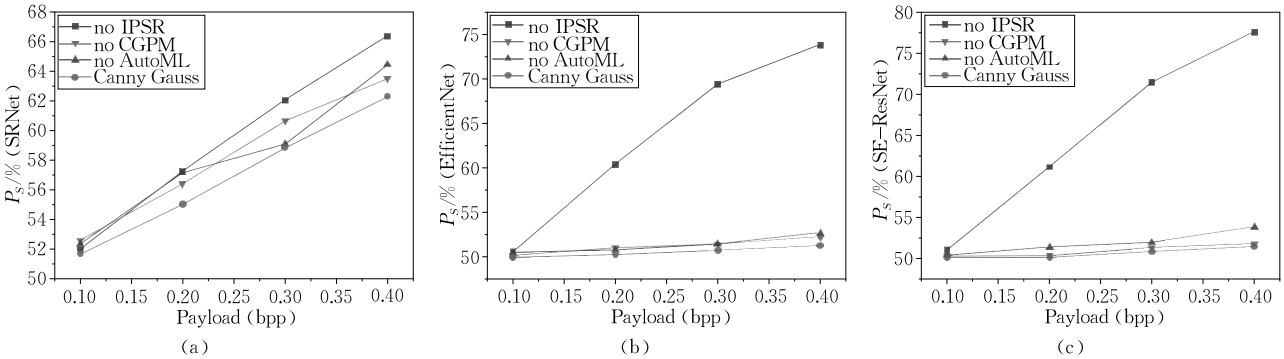


图 14 Canny Gauss 各模块消融实验

通过上述各域各算法的比较，本小节验证了各算法在变换域转换公式下的隐写隐蔽性. 实验结果表明，本文所提算法在集成分类器下的结果并不理想，同时在变换域的图像质量影响也有待提高，而集成分类器多采用灰度共生矩阵等统计量作为隐写分析的判别标准，这与 IPSR 的平滑范围是高度相关的，因此需要将 IPSR 的嵌入范围进一步收缩以规避此类富特征的检测. 同时，真实场景下有效嵌入比重大于 0.4 相对少见，现实中往往是将秘密消息按照小于 0.1 的嵌入载荷进行多次分割，嵌入不同图像载体中，以此降低泄露风险[37]，此时载荷为 0.4bpp/bpnzAC 满嵌安全性表现已由图 11 和图 12 给出，在这种情况下，基于更高载荷训练的隐写分析器易造成低载荷隐写图像的判准失真；又如图 13 所示，由于 JPEG 和 PQ 域在嵌入失真代价图上的高度相似，让跨域检测成为可能，因此信道控制方对图像进行压缩重构攻击的实际意义远大于隐写分析. 目前，自适应隐写转向了鲁棒隐写以提升载密信息的鲁棒性. 但鲁棒隐写安全性严重依赖目标信道的压缩、滤波等攻击的具体信息[38-39]，这相当于 PreCover 对 Cover 的改造过程被检测方获得，所以在通用域表现出更高嵌入代价稳定性和安全性的启发式算法将是未来一段时间的研究重点.

5 总结与展望

本文回顾了当前主流的嵌入失真函数设计原理与常见隐写作用域(空域、频域、边信息域)嵌入的具体细节. 通过对各主流设计的隐写失真代价图的共性分析和隐写作用域的关联性比较，从图像分割与优化角度直观地总结了启发式算法为抵抗隐写分析特征的通用设计原则，即满足全局扰动最小的局部熵最大策略，并设计了一套完整的图像隐写作用域转换公式实现通用域隐写. 此外，针对单一隐写分析特征设计的隐写术抵抗深度分析器暴露出大量的安全问题. 本文从嵌入失真代价和隐写嵌入域两个角度提出自适应隐写设计 Canny Gauss. 实验结果表明，本文的方案对深度隐写分析特征的隐蔽性显著，

但仍有较大改进空间.未来将继续完善失真代价,以提高算法在集成分类器下的隐写隐蔽性.

参 考 文 献

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

[2] Morkel T, Eloff J H, Olivier M S. An overview of image steganography//Proceedings of the 5th Annual Information Security South Africa Conference (ISSA2005). Sandton, South Africa, 2005, 1(2): 1-11

[3] Hussain M, Wahab A W A, Idris Y I B, et al. Image steganography in spatial domain: A survey. Signal Processing: Image Communication, 2018, 65: 46-66

[4] Filler T, Fridrich J. Gibbs construction in steganography. IEEE Transactions on Information Forensics and Security, 2010, 5(4): 705-720

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

[6] Holub V, Fridrich J. Designing steganographic distortion using directional filters//Proceedings of the 2012 IEEE International Workshop on Information Forensics and Security. Tenerife, Spain, 2012: 234-239

[7] Li B, Wang M, Huang J, et al. A new cost function for spatial image steganography//Proceedings of the 21st IEEE International Conference on Image Processing. Paris, France, 2014: 4206-4210

[8] Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability. IEEE Transactions on Information Forensics and Security, 2015, 11(2): 221-234

[9] Guo L, Ni J, Su W, et al. Using statistical image model for JPEG steganography: Uniform embedding revisited. IEEE Transactions on Information Forensics and Security, 2015, 10(12): 2669-2680

[10] Su W, Ni J, Hu X, et al. Image steganography with symmetric embedding using Gaussian Markov random field model. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(3): 1001-1015

[11] Pevny T, Bas P, Fridrich J. Steganalysis by subtractive pixel adjacency matrix. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 215-224

[12] Hussain I, Zeng J, Xinhong X, et al. A survey on deep convolutional neural networks for image steganography and steganalysis. KSII Transactions on Internet and Information Systems, 2020, 14(3): 1228-1248

[13] Xu G. Deep convolutional neural network to detect J-UNIWARD//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. Philadelphia Pennsylvania, USA, 2017: 67-73

[14] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 2018, 14(5): 1181-1193

[15] Yousfi Y, Butora J, Khvedchenya E, et al. ImageNet pre-trained CNNs for JPEG steganalysis//Proceedings of the 12th IEEE International Workshop on Information Forensics and Security. New York, USA, 2020: 1-6

[16] Fu Zhang-Jie, Wang Fan, Sun Xing-Ming, Wang Yan. Research on steganography of digital images based on deep learning. Chinese Journal of Computers, 2020, 43(9): 1656-1672(in Chinese)
(付章杰，王帆，孙星明，王彦. 基于深度学习的图像隐写方法研究. 计算机学报，2020，43(9)：1656-1672)

[17] Tang W, Tan S, Li B, et al. Automatic steganographic distortion learning using a generative adversarial network. IEEE Signal Processing Letters, 2017, 24(10): 1547-1551

[18] Yang J, Ruan D, Huang J, et al. An embedding cost learning framework using GAN. IEEE Transactions on Information Forensics and Security, 2019, 15: 839-851

[19] Zhang Y, Zhang W, Chen K, et al. Adversarial examples against deep neural network based steganalysis//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. Innsbruck, Austria, 2018: 67-72

[20] Tang W, Li B, Tan S, et al. CNN-based adversarial embedding for image steganography. IEEE Transactions on Information Forensics and Security, 2019, 14(8): 2074-2087

[21] Tang W, Li B, Barni M, et al. An automatic cost learning framework for image steganography using deep reinforcement learning. IEEE Transactions on Information Forensics and Security, 2020, 16: 952-967

[22] Mo X, Tan S, Li B, et al. MCTSteg: A Monte Carlo tree search-based reinforcement learning framework for universal non-additive steganography. IEEE Transactions on Information Forensics and Security, 2021, 16: 4306-4320

[23] Mo X, Tan S, Tang W, et al. ReLOAD: Using reinforcement learning to optimize asymmetric distortion for additive steganography. IEEE Transactions on Information Forensics and Security, 2023, 18: 1524-1538

[24] Wang Shuo-Zhong, Zhang Xin-Peng, Zhang Wei-Ming. Recent advances in image-based steganalysis research. Chinese Journal of Computers, 2009, 32(7): 1247-1263(in Chinese)
(王朔中，张新鹏，张卫明. 以数字图像为载体的隐写分析研究进展. 计算机学报，2009，32(7)：1247-1263)

[25] Liu J, Zhang W, Zhang Y, et al. Detection based defense against adversarial examples from the steganalysis point of view//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, California, 2019: 4825-4834

[26] Canny J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8: 679-698

[27] Fridrich J, Goljan M, Soukal D. Perturbed quantization steganography with wet paper codes//Proceedings of the 2004 Workshop on Multimedia and Security. Magdeburg, Germany, 2004: 4-15

[28] Butora J, Fridrich J. Revisiting perturbed quantization//Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. Virtual Event, Belgium, 2021: 125-136

[29] Denemark T, Sedighi V, Holub V, et al. Selection-channel-aware rich model for steganalysis of digital images//Proceedings of the 2014 IEEE Workshop on Information Forensic and Security. Atlanta Georgia, USA, 2014: 48-53

[30] Li W, et al. Designing near-optimal steganographic codes in practice based on polar codes. IEEE Transactions on Communications, 2020, 68(7): 3948-3962

[31] Fu H, Zhao X, He X. High-performance steganographic coding based on sub-polarized channel//Proceedings of the 21st International Workshop on Digital-forensics and Watermarking. Guilin, China, 2022: 3-19

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

[33] Fridrich J, Pevny T, Kodovsky J. Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities//Proceedings of the 9th Workshop on Multimedia and Security. Dallas Texas, USA, 2007: 3-14

[34] Cogranne R, Giboulot Q, Bas P. Steganography by minimizing statistical detectability: The cases of JPEG and color images//Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security. Denver, USA, 2020: 161-167

[35] Denemark T, Fridrich J. Model based steganography with precover. Electronic Imaging, 2017, 29: 56-66

[36] Mielikainen J. LSB matching revisited. IEEE Signal Processing Letters, 2006, 13(5): 285-287

[37] Zhang Y, Qin C, Zhang W, et al. On the fault-tolerant performance for a class of robust image steganography. Signal Processing, 2018, 146: 99-111

[38] Wang Z, Meng X. Digital image information hiding algorithm research based on LDPC code. EURASIP Journal on Image and Video Processing, 2018, 2018(1): 1-12

[39] Zeng K, Chen K, Zhang W, et al. Improving robust adaptive steganography via minimizing channel errors. Signal Processing, 2022, 195: 108498

LI Ji-Yu, M.S. candidate. His research interest is adaptive image steganography.

FU Zhang-Jie, Ph. D., professor. His research interests include information hiding, network and information security.

WANG Fan, Ph. D. candidate. Her research interests include information hiding and deep learning.

Background

As the primary branch of multimedia steganography, image steganography is a key study direction in the world of information security, and it is frequently employed in highsecurity data transmission levels, such as privacy protection and concealment communication. Information transmission is accomplished through the use of a picture as the medium for constructing a chimerical model of a secret message and Cover redundant information. In order to improve the security performance of the current framework for image steganography against advanced steganalysis, this paper presents a universal domain steganography algorithm called Canny Gauss that is based on the relationship between embedding cost map and texture sparsity. Experiments indicate that Canny Gauss has superior embedding stability in all of UNIWARD's feasible domains. The deep steganalysis performance of this algorithm under the third-party weights has improved by a minimum of 2.6% and a maximum of 14.6%. It offers a novel strategy for resisting feature-based deep steganalysis detection.

This work is supported by the National Key R&D Program of China under Grant No. 2021YFB2700900, the National Natural Science Foundation of China under Grant Nos. U22B2062, 62172232, the Jiangsu Basic Research Programs-Natural Science Foundation under Grant No. BK20200039, and the Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET) Fund. These projects aim to promote the development of information hiding technology enrich the theoretical knowledge in the field of information hiding and ensure information security in cyberspace.