图像非加性隐写综述

王垚飞[1,2]，张卫明[1,2]，陈可江[1,2]，周文柏[1,2]，俞能海[1,2]
（1. 中国科学技术大学网络空间安全学院，安徽 合肥 230027；
2. 中科院电磁空间信息重点实验室，安徽 合肥 230027）

摘　要：图像非加性隐写不仅能更好地维持图像元素的分布，而且具有较高的抗检测性能。首先对图像非加性隐写方法进行了梳理，将其分为两大类：非加性失真设计和非加性隐写编码设计。进一步将非加性失真设计总结为 3 类：基于理论模型、基于修改原则和基于对抗检测的方法，对这些方法进行了对比。最后分析了非加性隐写面临的困难问题和未来的发展思路。

关键词：信息隐藏；图像隐写；非加性失真；非加性隐写编码
中图分类号：TP391
文献标识码：A
DOI: 10.11959/j.issn.2096−109x.2021102

Survey on image non-additive steganography

WANG Yaofei[1,2], ZHANG Weiming[1,2], CHEN Kejiang[1,2], ZHOU Wenbo[1,2], YU Nenghai[1,2]
1. School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, China
2. Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China

Abstract: Image non-additive steganography not only can better maintain the distribution of image elements, but also has high detection resistance. Firstly, the image non-additive image steganography methods were sorted out and divided into two major categories: non-additive distortion design and non-additive steganography coding design. The non-additive distortion was designed into three categories: theoretical models based, modification principles based and adversarial detection based, and compared these methods. Finally, the difficult problems faced by non-additive steganography and the future development ideas were analyzed.

Keywords: information hiding, image steganography, non-additive distortion, non-additive steganography code

1　引言

隐写是用于隐蔽通信的技术，它将秘密消息以不可感知的形式隐藏在内容公开的载体中，在保护秘密通信内容的同时隐藏了通信行为。自适应隐写通信系统如图 1 所示，近年来，随着图像成为流行

收稿日期：2021−07−13；**修回日期**：2021−10−28
通信作者：陈可江，chenkj@ustc.edu.cn
基金项目：国家自然科学基金（62102386, 62002334, 62072421, 62121002）；安徽省自然科学基金（2008085QF296）
Foundation Items: The National Natural Science Foundation of China (62102386, 62002334, 62072421, 62121002), The Nature Science Foundation of Anhui Province (2008085QF296)

论文引用格式：王垚飞, 张卫明, 陈可江, 等. 图像非加性隐写综述[J]. 网络与信息安全学报, 2021, 7(6): 1-10.
WANG Y F, ZHANG W M, CHEN K J, et al. Survey on image non-additive steganography[J]. Chinese Journal of Network and Information Security, 2021, 7(6): 1-10.

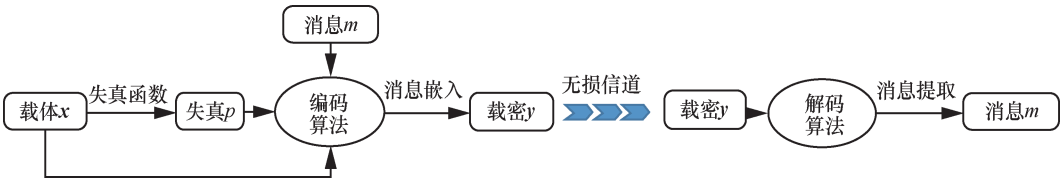

图 1　自适应隐写通信系统

Figure 1　Adaptive steganography communication system

的社交媒体，基于图像的隐写算法研究成为热点，针对图像的主流隐写都是自适应隐写，基于最小化失真隐写框架，即在负载率一定时最小化总失真。失真通常由失真函数衡量每个图像元素修改所产生的影响来获得。一般情况下，由于隐写在各个元素上的修改对失真的影响是不独立的，基于最小化失真框架很难求解，因此研究人员假设隐写在各个位置上的修改对失真的影响是独立的，即为加性隐写模型。目前的 STC（syndrome-trellis code）[1]和 SPC（steganographic polar codes）[2]在给定嵌入率下，能够在加性模型下接近失真理论界。

大量流行的图像隐写自适应失真函数设计都是加性的，根据嵌入域将它们分为空域与 JPEG 域，其中针对空域图像的失真函数有：HUGO (highly undetectable stego)[3]，WOW (wavelet obtained weights)[4]，S-UNIWARD（spatial-universal wavelet relative distortion）[5]，HILL（high-pass, low-pass and low-pass）[6]，MG（multivariate gaussian），[7]和 MiPOD（minimizing the power of optimal detector）[8]等，针对 JPEG 域的失真函数有：J-UNIWARD（jpeg-universal wavelet relative distortion）[5]，UED（uniform embedding distortion）[9]，RBV（residual block values）[10]和 BET（block entropy transformation）[11]等。为了更好地定义加性失真函数，学者们提出了 3 个基本原则：复杂度优先原则、扩散原则和争议元素优先原则。

虽然基于加性失真的模型设计在技术上更容易处理，但显然没有反映实际的失真变化情况。早在自适应隐写设计之初，研究人员就已经开始考虑该问题，Filler 和 Fridrich 使用 Gibbs 结构[12]进行了首次尝试，通过将宏观特征表达成局部特征的总和，该方法在 HUGO-BD（HUGO- bounding distortion）[3]中实现并显示了其有效性，但其安全性仍弱于最近提出的加性失真方案。如何建立有效的非加性隐写方案，一直困扰隐写领域的研究者。在 2013 年的国际信息隐藏大会上[13]：设计非加性隐写编码和非加性隐写失真函数被列入隐写领域的两个公开问题。2014 年，Holub 等[5]提出了基于小波函数定义的通用的失真函数 UNIWARD，设计之初它的失真定义方式是非加性的，即考虑了其他元素的修改对当前元素失真所造成的影响，但没有很好的非加性失真隐写编码方案，最终将该失真近似为加性失真用于嵌入。2015 年，非加性失真函数的设计有了新的突破，Li 等[14]和 Denemark 等[15]同时提出了方向一致性原则并将该原则用于非加性失真函数的定义，有效地提升了原有加性失真函数的安全性。随后 Zhang 等[16]提出了联合失真分解编码方法 DeJoin，将非加性编码问题等效分解成几个加性编码问题，从而实现了快速嵌入。随后，在 JPEG 域也出现了更多的原则和方案来优化基于非加性失真的嵌入，如 BBC[17]、BBC++[18]和 BBM[19]等原则。

针对现有的非加性隐写的研究，参考 2013 年的公开问题，将其分为两大类：非加性隐写编码研究和非加性失真函数研究。非加性失真函数研究又分为 3 类：① 基于理论模型的非加性失真函数，即以理论模型为指导优化失真函数的设计；② 基于修改原则的非加性失真函数，即以某种原则为指导来更新失真函数；③ 基于对抗检测的非加性失真函数设计，即通过直接考虑抵抗检测器的嵌入来设计非加性失真函数。

下面首先介绍最小化失真隐写框架，然后概述加性失真函数设计原则，在此基础上详细介绍非加性隐写模型。

2　最小化失真隐写框架

不失一般性，设载体为 $\boldsymbol{x}=(x_1,\cdots,x_n)\in X\triangleq I^n$，对于空域图像，典型地，$I=\{0,1,\cdots,255\}$，嵌入消息后的载密样本为 $\boldsymbol{y}=(y_1,\cdots y_n)\in Y\subset X$，

$\pi(\boldsymbol{y}) \triangleq P(\boldsymbol{y}|\boldsymbol{x})$ 表示修改转移分布（亦称为修改分布），不同的分布对应不同的嵌入处理操作，记 $Y = I_1 \times I_2 \times \cdots \times I_n, I_i \subset I$。本文只考虑三元嵌入，有 $I_i = \{x_i - 1, x_i, x_i + 1\}, x_i \neq \{0, 255\}$。对应特定的 $\boldsymbol{x}$，不同的 $\boldsymbol{y}$ 代表不同的嵌入方法，该修改方法下的总体失真定义为 $D(\boldsymbol{y}) \triangleq D(\boldsymbol{x}, \boldsymbol{y})$，可称为总失真函数。由于存在元素嵌入之间的相互影响，$D(\boldsymbol{y})$ 是每个元素失真上的加性总和。一般情况下，根据信息论理论，隐写的信息传输量（消息量）为分布函数 $\pi(\boldsymbol{y})$ 的熵：

$$H(\pi) = -\sum_{y \in Y} \pi(\boldsymbol{y}) \mathrm{lb} \pi(\boldsymbol{y}) \tag{1}$$

平均总失真为

$$E_\pi(D) = \sum_{y \in Y} D(\boldsymbol{y}) \pi(\boldsymbol{y}) \tag{2}$$

在隐写中，通常假设需要发送的隐秘消息量 m 是确定的，此时最优嵌入问题要最小化平均总失真，可以归结为求解限负载发送问题

$$\min_{\pi(y)} E_\pi(D) = \sum_{\boldsymbol{y} \in Y} D(\boldsymbol{y}) \pi(\boldsymbol{y}) \tag{3}$$

$$\text{s.t.}\ H(\pi) = -\sum_{\boldsymbol{y} \in Y} \pi(\boldsymbol{y}) \mathrm{lb} \pi(\boldsymbol{y}) = m \tag{4}$$

其中，$\sum_{\boldsymbol{y} \in Y} \pi(\boldsymbol{y}) = 1$，即在保证传输消息量 m 的能力下，最小化失真。遵循最大熵原理，最优解应当具有 Gibbs 分布的形式

$$\pi_\lambda(\boldsymbol{y}) = \frac{\exp(-\lambda D(\boldsymbol{y}))}{\sum \exp(-\lambda D(\boldsymbol{y}))} \tag{5}$$

在计算中，往往需要给出 $D(y)$ 的具体形式。但是由于每个元素嵌入之间存在相互影响，很难精确地给出 $D(y)$ 的具体形式。在此通常考虑加性失真的情况，设 ρ_i 表示仅 x_i 被修改为 y_i 所引起的失真，在加性模型下，假设每个元素的修改不相互影响，总体失真是每个元素上修改所造成失真的和，即

$$D(y) = \sum_{i=1}^{n} \rho_i(y_i) \tag{6}$$

假设 $\pi(y_i)$ 是 x_i 被修改为 y_i 的修改概率，那么最小化失真的优化问题转化为

$$\min_{\pi} E_\pi(D) = \sum_{i=1}^{n} \sum_{t_i \in I_i} \pi(t_i) \rho(t_i) \tag{7}$$

$$\text{s.t.}\ H(\pi) = -\sum_{i=1}^{n} \sum_{t_i \in I_i} \pi(t_i) \mathrm{lb} \pi(t_i) = m \tag{8}$$

对应的最优修改概率 π_λ 为

$$\pi_\lambda(y_i) = \frac{\exp(-\lambda \rho(y_i))}{\sum_{t_i \in I_i} \exp(-\lambda \rho(t_i))} \tag{9}$$

以上即为最小化失真隐写框架，由于加性失真容易刻画和求解，常见的自适应隐写一般基于失真的加性模型设计。以最小化加性隐写失真为目的的隐写嵌入，其中两个关键技术就是隐写编码和隐写失真的设计。鉴于采用隐写编码 STC[1] 和 SPC[2]可以在接近加性失真的理论界的条件下完成消息的嵌入和提取，留下的另一个问题就是对加性失真函数的设计和非加性隐写设计。

3　加性失真函数设计原则

目前的非加性失真模型都是以加性模型为基础的，其失真也是基于加性失真改进设计的，因此本文总结了加性失真函数设计的 3 项基本原则。

（1）复杂度优先原则

该原则是最重要的原则，它的理念是对高复杂度和噪声区域的元素分配较低的修改失真。因为这些难以预测的复杂区域和噪声区域难以建模，在这些区域的修改对应于隐写分析特征的变化较小，从而难以被检测到。目前许多隐写失真函数都基于这一基本原则[4-6]。

（2）扩散原则

它的理念是要求相邻元素的修改失真差异不应过大[20]。换句话说，当一个元素具有较高或者较低的修改失真时，相邻元素也应当具有相当的修改失真，该原则已经被成功地应用于空域图像[6,20]和 JPEG 图像中[21]。

（3）争议元素优先原则[22]

该原则认为，当性能相当的基础失真函数对同一元素的失真定义有很大差别时，应当对这种争议元素赋予更低的修改失真以鼓励其修改。

4　非加性隐写模型

本节探讨非加性隐写方法，根据这些方法的特点，本文将现有的非加性隐写研究分为两大类：

非加性隐写失真设计和非加性隐写编码设计。

4.1　非加性隐写失真设计

对于非加性隐写失真设计，根据设计的目标和过程将其分为 3 类（如图 2 所示）。第一类是基于理论模型的非加性失真函数设计，即以理论模型为指导优化失真函数的设计，最早期的非加性失真设计采用该方法，如 Gibbs 构造[12]；第二类是基于修改原则的非加性失真函数设计，即以某种原则为指导来更新失真函数，如 CMD[14]，目前大多数方法属于这一类，简单有效；第三类是基于对抗检测的非加性失真函数设计，即通过直接考虑抵抗检测器的嵌入来设计非加性失真函数，如早期对抗手工特征检测器的 HUGO-BD[3]和现在对抗深度学习的检测器 ADV-EMB[23]等，利用深度学习捕捉元素相关性的优势来指导非加性失真的调整。研究人员在设计失真函数的同时，为了更多地考虑相关性，往往会结合多种方法，如 ITE-SYN[24]既采用了修改原则的方式又采用了对抗检测的方式。以下将详细介绍这 3 类非加性失真函数设计。

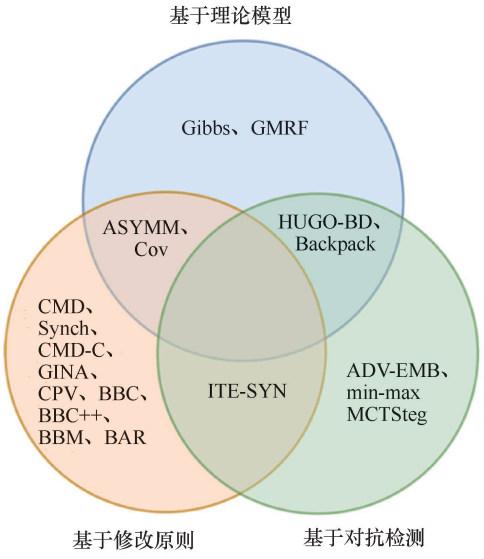

图 2　图像非加性隐写失真分类

Figure 2　Classification of image non-additive steganographic distortion

（1）基于理论模型的非加性失真函数设计

最早的非加性失真方案是采用 Gibbs 构造[12]模拟嵌入变化之间的相互作用，通过将宏观特征写成局部特征函数的总和，为加性失真函数创建了一个上限。为了进一步简化该方案，Filler 等将载体输入区域相互隔离，采用迭代的方式优化最优嵌入，该方法中的载体划分方案为后来的基于修改原则的非加性方法提供了参考，HUGO-BD[3]就是采用 Gibbs 构造的非加性隐写。类似地，为了进一步考虑不同嵌入单元的相互作用，Su 等提出了一种基于高斯马尔可夫随机场（GMRF）[30]的图像隐写算法，将图像隐写任务表示成在给定嵌入容量下最小化载体和载密分布 KL-散度的优化问题。Hu 等[29]通过考虑相邻元素的影响，提出了基于模型的非对称隐写框架 ASYMM，通过在嵌入时采用方向一致性原则，将相邻修改通过高斯混合模型进行优化。以上方法的性能和特点如表 1 所示。Taburet 等[31]从另一个角度构建模型，通过统计 DCT 系数之间的相关性设计了一个非加性的同步嵌入策略，具体而言是先统计系数之间的相关性，然后将其转化为高斯分布，其方差直接从嵌入失真中计算而来，最后按照分块嵌入的思想，将已嵌入部分的修改作为先验更新未嵌入部分的失真。

基于理论模型的非加性隐写方法，以理论为指导优化目标函数，取得了一定的成效，但求解的过程比较复杂，因此研究者开始研究启发式的基于修改原则的非加性失真函数设计。

（2）基于修改原则的非加性失真函数设计

虽然 Gibbs 构造[12]具有较好的理论解释，且在 HUGO-BD[3]上产生了一定效果，但其安全性依然不能超过近年来所提出的自适应加性隐写失真函数，然而该工作给了研究者一个很好的启发，即可以将载体分成隔离的部分，然后采用更新失真的方法动态更新未嵌入部分的失真。

2015 年 Li 等[14]以及 Denemark 与 Fridrich[15]针对空域灰度图像分别提出了 CMD（clustering modification directions）非加性隐写与 Synch（synchronization）非加性隐写，有效地提升了抗检测性能，其主要思想都是将载体图像划分没有交集的子图，然后依次嵌入，根据已修改像素的修改方向按照方向一致性原则来更新未嵌入区域的失真，从而达到聚集修改方向的目的。CMD 和 Synch 的主要不同点是图像划分方案和失真更新方法不同，两者对比及特点如表 1 所示。

受启发于方向一致性原则，Tang 等[32]直接将该原则应用到彩色空域图像上设计了 CMD-C 策略，但由于未考虑通道间的相关性，安全性并没

表 1　针对灰度空域图像的不同非加性失真定义方法对比

Table 1　Comparison of different methods of defining non-additive distortion for grayscale spatial domain images

方法	性能	性能比较	特点	缺点
Gibbs[12]	相比于加性失真略有提升，如在 BOWS2[25]中采用二元嵌入，嵌入率为 0.4 bpp，抵抗 SPAM[26]特征的性能可以提升约 0.02	ASYMM>CMD>Synch>GMRF>HUGO-BD>Gibbs	基于 Gibbs 构造具有一定的理论性，为非加性方法定义提供了参考	安全性提升较少，实行比较复杂
HUGO-BD[3]	相比于加性失真有较大提升，如在 BOWS2 中采用二元嵌入，嵌入率为 0.4 bpp，抵抗二阶 SPAM 特征的性能可以提升约 0.16		采用模型矫正，通过动态更新元素 ±1 的失真，在一定限度上克服了加性模型的局限	仅适用于二元嵌入，且难以抵抗基于富模型特征的隐写分析
CMD[14]	相比于加性失真有较大提升，如在 BOSSBase[27]中采用三元嵌入，嵌入率为 0.4 bpp，以 HILL[6]为初始失真抵抗 SRM[28]特征的性能可以提升约 0.05		基于方向一致性原则，通过更新失真的方式鼓励相邻像素的修改方向同向修改，方法简单有效，可以和任意初始失真进行组合	基于启发式，缺乏一定的理论性
Synch[15]	相比于加性失真有部分提升，如在 BOSSBase 中采用三元嵌入，嵌入率为 0.4 bpp，以 HILL 为初始失真抵抗 SRM 特征的性能可以提升约 0.03		基于方向一致性原则，简单有一定效果，可以和任意初始失真进行组合	基于启发式，缺乏理论性，弱于 CMD
ASYMM[29]	抗监测性能相比于 CMD 有微弱的提升		基于非对称模型，分块嵌入，将相邻嵌入作为先验对待嵌入像素的修改率进行优化	抗检测性能提升微弱，计算复杂度高
GMRF[30]	略好于 MiPOD[8]		基于四元素交叉邻域的高斯马尔可夫随机场模型来描述载体图像的局部元素之间的相互作用，将隐写问题表述为与模型相关的载体载秘 KL 最小化问题	计算复杂度高，考虑相关性带来的提升较小

注：嵌入率（bpp，bits per pixel）

有得到显著提升。随后 Wang 等[33]通过考虑通道间相关性和差异性后，参考 CMD 策略提出了基于 G 通道的通道间非加性策略 GINA，该方法考虑到了通道间的相关性，即鼓励 R 和 B 通道与 G 通道同步修改，并考虑到通道间的差异性，即在复杂度优先原则的基础上有选择地更新失真，此外 GINA 采用了载荷自适应分配的策略来提升隐写的安全性。相同的追求也在 Qin 等[34]提出的 CPV[34]方法中体现，即通过将同一位置不同通道的像素考虑为一个超像素，根据复杂度优先原则和通道间的相关性定义了 27 种失真，采用联合失真分解 DeJoin 方案[16]将消息自适应地分配到 3 个通道。

表 2　针对彩色空域图像的不同非加性失真定义方法对比

Table 2　Comparison of different methods of defining non-additive distortion for color spatialdomain images

方法	性能	性能比较	特点	缺点
CMD-C[32]	相比加性失真有所提升，如在 BOSSbasePPGBIC 中采用三元嵌入，嵌入率为 0.4bpcp，以 HILL 为初始失真抵抗 SCRMQ1[36]特征的性能可以提升约 0.07	GINA>CPV-CMD>CMD-C>CPV	基于方向一致性原则，通过更新失真的方式不加区分地鼓励通道内和通道间的修改方向一致，简单有一定效果，可以和任意初始失真进行组合	没有挖掘彩色通道间的相关性和差异性，缺乏理论性，部分情况弱于基于 CMD 的嵌入
CPV[34]	相比加性失真有所提升，如在 BOSSbasePPGBIC 中采用三元嵌入，嵌入率为 0.4bpcp，相比于 HILL 抵抗 SCRMQ1 特征的性能可以提升约 0.007，和 CMD 结合后 CPV-CMD 可进一步提升 0.072		一种基于颜色像素向量的失真定义，通过定义 3 个通道的相同位置处的共同失真，并采用失真分解的方式获得单个像素的失真，可以在 3 个通道中自适应分配消息	没有完全考虑到通道间的相关性，失真计算复杂度较高
GINA[33]	相比 CMD-C 和 CMD 有所提升，如在 BOSSbasePPGBIC 中采用三元嵌入，嵌入率为 0.4bpcp，以 HILL 为初始失真抵抗 SCRMQ1 特征的性能可以提升约 0.10		挖掘了彩色通道间的相关性和差异性，权衡方向一致性原则和复杂度优先原则，简单有效，可以和任意初始失真进行组合	缺乏理论性，参数设计仅基于启发式获得

注：嵌入率（bpcp，bits per color pixel）

CMD-C、CPV 和 GINA 的特点和比较如表 2 所示。

方向一致性原则虽然在空域图像展示了其有效性，但并不能直接应用到 JPEG 图像，为此 Li 等[17]针对 JPEG 图像提出了一个新的原则：块边界连续性（BBC，block boundary continuity）原则，即在嵌入修改过程中鼓励空域块边界的连续性，基于 BBC 原则的第一个方法只考虑相邻 DCT 块中同一模式下的相关性来鼓励其同向或反向修改，并采用联合失真分解 DeJoin 进行嵌入。然而 DCT 块中一个系数的修改会影响到空域中 64 个像素，且在修改过程中只考虑一对 DCT 系数的相关性来维持空域块边界连续性是不够的，基于此 Wang 等[18]提出了一种块边界连续性增强策略 BBC++，即通过考虑全局 DCT 系数的修改来更新 DCT 系数和失真来维持空域块边界连续性，实验证明该策略能够充分利用块边界连续性原则来提升隐写的安全性。基于 BBC 原则的方法虽然取得了一定的效果，但它只考虑相邻块间系数的相关性，没有考虑到块内系数的相关性，为此 Wang 等[19]通过探索 DCT 系数嵌入后在空域块中不同区域的修改分布和抗检测性能，发现空域块边界的修改数量和幅度明显大于空域块内部，且抗检测性能弱于空域块内部。基于此，通过探寻空域块内部的相关性，理论推导证明：当块内一对同行或同列的 DCT 系数坐标差为偶数时，反向修改会导致更少的空域块边界修改量，由此提出了块边界维持（BBM，block boundary maintenance）原则[19]，即通过考虑块内 DCT 系数的相关性减少空域块边界的修改，基于此原则的方法有效地提升了 JPEG 图像的隐写安全性，且该原则可以和 BBC 原则进一步结合来提升隐写的安全性。这也启发研究人员不同修改原则并不是冲突的，非加性失真设计可以参考多种原则进行设计。此外，Lu 等[35]针对 JPEG 图像提出了块效应去除原则 BAR，即通过衡量相邻块的修改对块效应的影响来调节嵌入失真，它同时考虑了相邻块和块内部系数的相关性，这类似于 BBC 和 BBM 的组合。BBC、BBC++、BBM 和 BAR 的特点和比较如表 3 所示。

基于修改原则的非加性设计需要研究者根据不同类型图像的特点，挖掘其内在的相关性和差异性来总结出一种修改原则指导非加性失真的设计。目前针对空域图像和 JPEG 图像，虽然产生了多种修改原则，但基于这些修改原则的失真更新都是启发式的，缺乏一定的理论指导，且人工寻找新的原则往往耗时耗力，如何利用机器学习设计非加性失真开始成为人们关注的热点。

表 3　针对灰度 JPEG 图像的不同非加性失真定义方法对比

Table 3　Comparison of different methods of defining non-additive distortion for grayscale JPEG images

方法	性能	性能比较	特点	缺点
BBC[17]	相比于加性失真有所提升，如在质量因子为 75 的 BOSSBase 中采用三元嵌入，嵌入率为 0.5bpnzac，以 UERD[37]为基础失真抵抗 GFR[38]特征的性能可以提升 0.01	BAR>BBC++>Cov>BBM>BBC	首个针对 JPEG 图像的非加性失真定义原则，通过定义相邻块相同位置的一对 DCT 系数的联合失真来鼓励修改后的空域块边界连续，考虑相邻块 DCT 系数的相关性	DCT 系数之间的相关性考虑不完备，效果提升较少
BBC++[18]	相比于 BBC 有所提升，如在质量因子为 75 的 BOSSBase 中采用三元嵌入，嵌入率为 0.5bpnzac，以 UERD 为基础失真抵抗 GFR 特征的性能可以提升 0.040。		BBC 方案的增强版，通过分块嵌入，嵌入过程中更新载体和失真来维持空域块边界连续性，相较于 BBC 具有更明显的提升性能	仅考虑了块间系数的相关性
BBM[19]	相比于加性失真有所提升，如在质量因子为 75 的 BOSSBase 中采用三元嵌入，嵌入率为 0.5bpnzac，以 UERD 为基础失真抵抗 GFR 特征的性能可以提升 0.030		首个考虑 DCT 块间相关性的非加性失真定义原则，通过失真更新的方式，减少空域块边界修改扰动，有一定的效果	仅考虑了块内系数的相关性
BAR[35]	相比于加性失真有所提升，如在质量因子为 75 的 BOSSBase 中采用三元嵌入，嵌入率为 0.5bpnzac，以 UERD 为基础失真抵抗 GFR 特征的性能可以提升 0.040		同时考虑了 DCT 块内和块间系数的相关性，通过失真更新的方式，维持空域块边界的分布，效果提升显著	基于启发式设计，参数基于实验调节
Cov[31]	相比于加性失真性能有一定的提升，如在质量因子为 75 的 BOSSBase 中采用三元嵌入，嵌入率为 0.28bpnzac，以 UERD 为基础失真抵抗 DCTR[39]特征的性能可以提升 0.030		在图像处理流程中利用 DCT 系数间的统计分析设计同步策略，对相邻块构建协方差矩阵，将每个系数相关的经验失真转化为高斯分布	计算复杂度高，考虑到相关性带来的提升较小

注：嵌入率（bpnzac，bits per non-zero ac coefficients）

（3）基于对抗检测的非加性失真函数设计

HUGO[3]是最早的自适应隐写失真函数，它通过考虑对抗 SPAM 特征[26]来设计，并提供了模型矫正功能，具有非加性的特点，即在嵌入中针对每个像素重新计算当前嵌入下的+1 和−1 的失真，选择其中失真较小的修改方向进行嵌入。随着隐写分析的发展，手工特征的维度越来越高，如对于空域富模型特征 SRM[28]，很难有针对性地设计非加性失真函数。因而近年来，从对抗高维度的人工隐写分析特征出发，没有产生有效的非加性失真设计方法，直到基于深度学习隐写分析器的出现，给了研究者另一条思路，即抵抗基于深度学习的隐写分析器。

基于深度学习的隐写分析器表现了较高的隐写检测性能，但深度学习神经网络分类器具有脆弱性，给样本添加微小的扰动就能误导分类器，称基于此的研究为对抗样本。借鉴对抗样本的思想，Zhang 等[44]首次提出对抗样本隐写的概念，随后 Li 等[23]提出 ADV-EMB，将图像划分成两部分区域，第一部分正常嵌入，第二部分对抗嵌入。逐渐增大对抗嵌入区域，增强隐写算法的对抗性。其中第二部分的失真调整考虑到第一部分已经发生的修改，使得修改点有更大的可能性和梯度的方向一致，从而产生对抗效果。这种基于对抗检测的方法可以和基于修改原则的方法进一步结合，如 ITE-SYN[24]将对抗修改和方向一致性原则[14]结合，设计了更有效的非加性隐写失真算法。为了增强对抗多种隐写分析器的性能和效率，Bernard 等[40-41]提出了一种 min-max 策略来进行优化对抗嵌入的过程，此外，他们又基于理论设计了一种新的失真调整策略 Backpack[43]，即将对抗性的非加性失真近似为加性，通过 Gumbel-Softmax 分布的样本来逼近离散的嵌入变化。与对抗样本的思想类似，但不使用梯度攻击，Mo 等[42]将蒙特卡洛树搜索（MCTS）和基于隐写分析器的环境模型结合，建立了自动化非加性隐写失真学习框架 MCTSteg，可以在空域和频域提高抗手工特征和深度学习隐写分析器的性能。以上方法的特点和对比如表 4 所示。

基于对抗检测的非加性失真函数设计虽然通过直接考虑抵抗检测器取得了较好的效果，但在与基于修改原则的失真函数设计进行结合后仍可以进一步提升[24]，这说明基于对抗检测的非加性失真函数设计对修改之间的相关性考虑还不完备，如何借鉴修改原则进一步提升抗检测的性能，或将对抗检测的设计理论化依然是值得研究的问题。

4.2　非加性隐写编码设计

隐写编码 STC[1]和 SPC[2]只解决了针对加性失真函数的消息嵌入问题。首个最小化非加性失真函数的次优编码方案是 Gibbs 构造[12]，该方法通过模拟嵌入变化之间的相互作用，进行不断嵌入和更新失真完成消息嵌入。这可以用来实现具有任意失真的嵌入，但是这种方案求解较为困难，因此，进一步约束将载体划分为没有交集的子图像进行交替迭代嵌入，希望嵌入模式能够收敛到最优嵌入的样本，但由于没有很好的非加性失真

表 4　基于深度学习的不同非加性失真定义方法对比

Table 4　Comparison of different non-additive distortion definition methods based on deep learning

方法	性能	性能比较	特点	缺点
ADV-EMB[23]	对于已知隐写分析网络，该方法有较高的抗检测性能，在已知隐写分析且无防御情况下，有较高的漏检率	Backpack>min-max>MCTSteg>ADV-EMB	首个利用对抗思想动态调整失真的对抗嵌入，即先嵌入一部分，另一部分根据梯度来修改失真进行嵌入，在白盒情况下有较高的安全性能	攻击成功率较低，迁移性弱
min-max 策略[40-41]	一种对抗训练策略，一定限度上增强了 ADV-EMB 的迁移性		一种通用的对抗训练策略，具有较高的迁移性	训练代价较高
MCTSteg[42]	空域和频域相比 CMD、BBC 和 BBM 都有提升		一种通用的非加性失真定义方法，采用分块嵌入，通过强化学习计算待嵌入块的最优修改方式后更新失真再嵌入	计算复杂度高
Backpack[43]	相比 ADV-EMB 性能有较大的提升		基于理论模型，通过 Gumbel-Softmax 分布的样本来逼近离散的嵌入变化，利用梯度信息来求解失真更新，具有较好的性能	

定义，这种方案最终没有达到很好的性能。

随着基于修改原则的非加性失真定义的出现，研究者参考 Gibbs 构造中的划分方案，采用了不同的划分区间[14-15]，只需要迭代一次便可将消息完整的嵌入，这也是目前实现非加性嵌入的最有效的方案。随后，Zhang 等[16]提出了一种联合失真分解编码 DeJoin 可以达到联合失真的理论界，将失真空间等效转换到概率空间。将多个元素视为一个整体，定义联合修改失真，根据限负载下的优化问题，可以得到联合修改概率，将联合修改概率根据链式法则分解为边缘概率和条件概率，根据概率失真翻转引理，将边缘概率和条件概率再转化成边沿失真和条件失真，由于这两种失真均为加性失真，嵌入消息时大大降低了编码复杂度，虽然这种方案可以达到联合失真的理论界，但如何定义更好的联合失真依然是值得研究的问题。其他的非加性编码方案如 variable-cost STC[45]，将动态更新融入 STC 编码过程中，但该方法的安全性没有得到有效提升，且计算复杂度较高。

对于非加性失真的设计，非加性隐写编码方案较少，大多数采用分步嵌入的方案，如何设计更有效的非加性编码依然是值得研究的课题。

5 结束语

本文对近年来图像非加性隐写的研究工作进行了总结和分析，将非加性隐写研究分为两大类：非加性隐写失真设计和非加性隐写编码设计。其中又对非加性隐写失真设计分为 3 类：基于理论模型、基于修改原则和基于对抗检测的非加性失真设计。最早的基于理论模型研究受限于无法得到精确描述的载体分布，于是过渡到启发式的基于修改原则的非加性失真函数设计；随着深度学习隐写分析的出现，通过对抗深度学习隐写分析，研究者借鉴对抗样本和强化学习的思想提出了更多基于对抗检测的非加性失真函数设计。随着非加性隐写失真设计的发展，非加性隐写编码也由原来的 Gibbs 构造到分块嵌入，再到可证明逼近联合失真理论界的高效隐写编码 DeJoin。由此可以看出，非加性隐写失真设计和非加性隐写编码互为补充，均与隐写分析的发展息息相关，面对快速发展的深度学习隐写分析算法，本文对未来的图像非加性隐写研究有如下几点展望。

1）基于理论模型的非加性隐写失真设计具有更好的解释性，但由于人工设计的理论模型和参数估计往往不能完全反映图像本身的特点，该方案下的非加性失真效果不佳，如何借鉴深度学习的强大学习能力，对图像建模设计更有效的非加性隐写失真是一个有挑战性的问题。

2）基于修改原则的非加性隐写具有简单有效的特点，但人工寻找有效的原则耗时耗力，如何利用深度学习从对抗检测的角度归纳新的原则是值得研究的问题。

3）基于对抗深度学习的非加性失真隐写设计虽然取得了较好的效果，但深度学习算法本身存在的可解释性不足的问题，依然需要更多的理论来支持非加性隐写失真设计。

4）非加性隐写编码相对于非加性隐写失真设计发展较缓，目前基于分块嵌入的策略虽然简单有效但无法达到失真理论界，联合失真分解编码 DeJoin 的提出可以达到联合失真的理论界，但针对联合失真的设计较少，如何利用 DeJoin 的特点设计联合失真依然是值得研究的问题。

综上，深度学习为图像非加性隐写设计提供了新的理念和技术，是信息隐藏领域未来发展的重要方向之一，但在理论性和效率方面仍然存在很多问题亟待解决；深度学习所具有的强大学习能力可以将失真定义和编解码过程两部分合成一个部分，目前较为成熟的是在以图藏图上的应用，即将秘密图像隐藏到载体图像，但这种方式并没有很高的抗检测性能；如何解决这些问题并进一步提升图像隐写的安全性和实用性对网络空间中信息的安全传输具有重要意义。

参考文献：

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

[2] LI W X, ZHANG W M, LI L, et al. Designing near-optimal steganographic codes in practice based on polar codes[J]. IEEE Transactions on Communications, 2020, 68(7): 3948-3962.

[3] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[M]//Information Hiding. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010: 161-177.

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

[5] HOLUB V, FRIDRICH J. Digital image steganography using universal distortion[C]//ACM Workshop on Information Hiding and Multimedia Security. 2013: 59-68.

[6] LI B, WANG M, HUANG J W, et al. A new cost function for spatial image steganography[C]//Proceedings of 2014 IEEE International Conference on Image Processing (ICIP). 2014: 4206-4210.

[7] FRIDRICH J, KODOVSKÝ J. Multivariate Gaussian model for designing additive distortion for steganography[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 2949-2953.

[8] SEDIGHI V, COGRANNE R, FRIDRICH J. Content-adaptive steganography by minimizing statistical detectability[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(2): 221-234.

[9] GUO L J, NI J Q, SHI Y Q. Uniform embedding for efficient JPEG steganography[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(5): 814-825.

[10] WEI Q D, YIN Z X, WANG Z C, et al. Distortion function based on residual blocks for JPEG steganography[J]. Multimedia Tools and Applications, 2018, 77(14): 17875-17888.

[11] HU X L, NI J Q, SHI Y Q. Efficient JPEG steganography using domain transformation of embedding entropy[J]. IEEE Signal Processing Letters, 2018, 25(6): 773-777.

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

[13] KER A D, BAS P, BÖHME R, et al. Moving steganography and steganalysis from the laboratory into the real world[C]//Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security-IH&MMSec '13. 2013: 45-58.

[14] 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.

[15] 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. 2015: 5-14.

[16] ZHANG W M, ZHANG Z, ZHANG L L, et al. Decomposing joint distortion for adaptive steganography[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(10): 2274-2280.

[17] LI W X, ZHANG W M, CHEN K J, et al. Defining joint distortion for JPEG steganography[C]//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. 2018: 5-16.

[18] WANG Y F, LI W X, ZHANG W M, et al. BBC: enhanced block boundary continuity on defining non-additive distortion for JPEG steganography[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(5): 2082-2088.

[19] WANG Y F, ZHANG W M, LI W X, et al. Non-additive cost functions for JPEG steganography based on block boundary maintenance[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 1117-1130.

[20] LI B, TAN S Q, WANG M, et al. Investigation on cost assignment in spatial image steganography[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(8): 1264-1277.

[21] CHEN K J, ZHOU H, ZHOU W B, et al. Defining cost functions for adaptive JPEG steganography at the microscale[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(4): 1052-1066.

[22] ZHOU W B, ZHANG W M, YU N H. A new rule for cost reassignment in adaptive steganography[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2654-2667.

[23] TANG W X, LI B, TAN S Q, et al. CNN based adversarial embedding with minimum alteration for image steganography[J]. arXiv: 1803.09043, 2018.

[24] QIN X H, TAN S Q, TANG W X, et al. Image steganography based on iterative adversarial perturbations onto a synchronized-directions sub-image[C]//Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021: 2705-2709.

[25] BENNOUR J, -L DUGELAY J, MATTA F. Watermarking attack: bows contest[C]//Proc SPIE 6505, Security, Steganography, and Watermarking of Multimedia Contents IX, 2007, 6505: 443-448.

[26] PEVNY T, BAS P, FRIDRICH J. Steganalysis by subtractive pixel adjacency matrix[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 215-224.

[27] BAS P, FILLER T, PEVNÝ T. “Break our steganographic system”: the ins and outs of organizing BOSS[M]//Information Hiding. Berlin, Heidelberg: Springer Berlin Heidelberg, 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] HU X, CHEN H, NI J. A novel steganography scheme based on asymmetric embedding model[M]//Cloud Computing and Security. Cham: Springer International Publishing, 2018: 183-194.

[30] SU W K, NI J Q, HU X L, et al. Image steganography with symmetric embedding using Gaussian Markov random field model[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(3): 1001-1015.

[31] TABURET T, BAS P, SAWAYA W, et al. JPEG steganography and synchronization of DCT coefficients for a given development pipeline[C]//Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security. New York, NY, USA: ACM, 2020: 139-149.

[32] TANG W X, LI B, LUO W Q, et al. Clustering steganographic modification directions for color components[J]. IEEE Signal Processing Letters, 2016, 23(2): 197-201.

[33] WANG Y F, ZHANG W M, LI W X, et al. Non-additive cost functions for color image steganography based on inter-channel correla-

tions and differences[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 2081-2095.

[34] QIN X H, LI B, TAN S Q, et al. A novel steganography for spatial color images based on pixel vector cost[J]. IEEE Access, 2019, 7: 8834-8846.

[35] LU Y B, ZHAI L M, WANG L N. Designing non-additive distortions for JPEG steganography based on blocking artifacts reduction[M]//Digital Forensics and Watermarking. Cham: Springer International Publishing, 2020: 268-280.

[36] GOLJAN M, FRIDRICH J, COGRANNE R. Rich model for steganalysis of color images[C]//Proceedings of 2014 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE Press, 2014: 185-190.

[37] GUO L J, NI J Q, SU W K, et al. Using statistical image model for JPEG steganography: uniform embedding revisited[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(12): 2669-2680.

[38] SONG X F, LIU F L, YANG C F, et al. Steganalysis of adaptive JPEG steganography using 2D Gabor filters[C]//Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security. 2015: 15-23.

[39] HOLUB V, FRIDRICH J. Low-complexity features for JPEG steganalysis using undecimated DCT[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(2): 219-228.

[40] BERNARD S, PEVNÝ T, BAS P, et al. Exploiting adversarial embeddings for better steganography[C]//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. 2019: 6.

[41] BERNARD S, BAS P, KLEIN J, et al. Explicit optimization of min max steganographic game[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 812-823.

[42] MO X B, TAN S Q, LI B, et al. MCTSteg: a Monte Carlo tree search-based reinforcement learning framework for universal non-additive steganography[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4306-4320.

[43] BERNARD S, BAS P, KLEIN J, et al. Optimizing additive approximations of non-additive distortion functions[C]//Proceedings of 9th ACM Workshop on Information Hiding and Multimedia Security. 2021:105-112.

[44] ZHANG Y W, ZHANG W M, CHEN K J, et al. Adversarial examples against deep neural network based steganalysis[C]//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. 2018: 67-72.

[45] PEVNY T, KER A D. Exploring non-additive distortion in steganography[C]//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. 2018: 109-114.

[作者简介]

王垚飞（1996− ），男，河北邯郸人，中国科学技术大学博士生，主要研究方向为信息隐藏。

张卫明（1976− ），男，河北定州人，中国科学技术大学教授、博士生导师，主要研究方向为信息隐藏、多媒体内容安全、人工智能安全。

陈可江（1994− ），男，浙江温州人，博士，主要研究方向为信息隐藏与人工智能安全。

周文柏（1992− ），男，安徽合肥人，中国科学技术大学特任副研究员，主要研究方向为信息隐藏与人工智能安全。

俞能海（1964− ），男，安徽无为人，中国科学技术大学教授、博士生导师，主要研究方向为多媒体信息检索、图像处理与视频通信、数字媒体内容安全。