软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3470−3484 [doi: 10.13328/j.cnki.jos.006290]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于生成对抗网络的空域彩色图像隐写失真函数设计方法*

廖 鑫[1,2], 唐志强[1], 曹 纭[2,3]

[1](湖南大学 信息科学与工程学院, 湖南 长沙 410082)

[2](信息安全国家重点实验室 (中国科学院 信息工程研究所), 北京 100093)

[3](中国科学院大学 网络空间安全学院, 北京 100093)

通信作者: 廖鑫, E-mail: xinliao@hnu.edu.cn

摘 要: 自适应隐写是图像隐写方向的研究热点, 它通过有效地设计隐写失真函数, 自适应地将秘密信息隐藏在图像复杂的纹理区域, 具有很强的隐蔽性. 近年来, 基于生成对抗网络的隐写失真函数设计研究在空域灰度图像上已经取得了突破性的进展, 但是目前还没有针对空域彩色图像的研究. 与灰度图像相比, 彩色图像隐写需要考虑保护RGB通道间相关性, 同时合理地分配RGB这3个通道的嵌密容量. 设计了一个基于生成对抗网络设计空域彩色图像隐写失真函数的框架CIS-GAN (color image steganography based on generative adversarial network), 生成器网络采用两个U-Net子网络结构, 第1个U-Net子网络生成修改概率矩阵, 第2个U-Net子网络进行正负向修改概率调节, 有效地降低对彩色图像通道相关性的破坏. 针对彩色图像载体, 修改灰度图像隐写分析器作为网络的对抗部分. 在生成器损失函数中对彩色图像3个通道总的隐写容量进行控制, 生成器能够自动学习分配3个通道嵌密容量. 实验结果表明, 与现有彩色图像隐写失真函数设计方法相比, 提出的网络结构能够更好地抵抗彩色图像隐写分析器的检测.

关键词: 图像隐写; 隐写失真函数; 生成对抗网络; RGB通道相关性

中图法分类号: TP391

中文引用格式: 廖鑫, 唐志强, 曹纭. 基于生成对抗网络的空域彩色图像隐写失真函数设计方法. 软件学报, 2022, 33(9): 3470–3484. http://www.jos.org.cn/1000-9825/6290.htm

英文引用格式: Liao X, Tang ZQ, Cao Y. Steganographic Distortion Function Design Method for Spatial Color Image Based on GAN. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3470–3484 (in Chinese). http://www.jos.org.cn/1000-9825/6290.htm

Steganographic Distortion Function Design Method for Spatial Color Image Based on GAN

LIAO Xin[1,2], TANG Zhi-Qiang[1], CAO Yun[2,3]

[1](College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China)

[2](State Key Laboratory of Information Security (Institute of Information Engineering, Chinese Academy of Sciences), Beijing 100093, China)

[3](School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China)

Abstract: Adaptive image steganography has been becoming a hot topic, as it conceals covert information within the texture region of an image by employing a defined distortion function, which guarantees remarkable security. In spatial gray-scale image steganography, the research on designing steganographic distortion functions using generative adversarial networks has achieved a significant breakthrough recently. However, related studies of spatial color image steganography have not been reported yet so far. Compared with the gray-scale image steganography, color image steganography should preserve the RGB channel correlation and reasonably assign the embedding capacity among RGB channels simultaneously. This study first proposes a framework based on a generative adversarial network to

* 基金项目: 国家自然科学基金 (61972142, 61872356); 国家重点研发计划 (2019QY(Y)0207, 2019QY2202); 湖南省自然科学基金 (2020JJ4212); 信息网络安全公安部重点实验室开放课题 (C20611)

收稿时间: 2020-08-11; 修改时间: 2020-10-12; 采用时间: 2020-12-16; jos在线出版时间: 2021-04-20

automatically learn to generate the steganographic distortion function for spatial color images, which is termed CIS-GAN (color image steganography based on generative adversarial network). The generator is composed of two U-Net subnetworks. One of them generates the modification probability matrix, while the other adjusts the positive/negative modification probability to effectively weaken the damage to the RGB channel correlation. The analyzer of gray-scale image steganography is modified as an adversarial part of the network for color images. In addition, the generator can automatically learn to allocate the embedding capacity for the three channels via controlling the total steganographic capacity in the generator's loss function. The experimental results show that the proposed framework outperforms the advanced steganographic schemes for spatial color images in resisting color image steganalysis.

Key words: image steganography; steganographic distortion function; generative adversarial network (GAN); RGB channel correlation

1 引 言

图像隐写是一门将秘密信息嵌入到载体图像中并尽量不降低原始载体质量的技术, 利用载体图像具有高度冗余性的特性, 将秘密信息嵌入后不易被察觉, 从而实现隐秘通信[1−3]. 载体图像嵌入秘密信息后会修改部分像素值, 不同位置像素值修改后与原图像差异各不相同, 这种差异称为失真. 一般来说, 在载体图像纹理复杂的地方进行修改产生的失真要小于在纹理平滑的地方进行修改. 隐写失真函数用于对载体图像修改后的差异大小进行建模量化. 在隐写技术发展过程中, 自适应隐写算法成为主流, 主要包括隐写失真函数和最小化失真嵌入编码算法两部分. 以 Syndrome-Trellis Code[4]为代表的最小化失真嵌入编码算法效果接近理论边界, 并且在实践中得到了很好地运用, 研究主要集中为设计有效的隐写失真函数来提高隐写算法的安全性. 代表性的隐写失真函数设计方法有 S-UNIWARD[5]和 HILL[6]等, 这些算法主要的思路是为纹理复杂区域的像素点分配高失真值, 平滑区域分配低失真值, 结合嵌入编码算法最小化失真后, 自适应地将秘密信息更多地隐藏在图像纹理复杂区域, 这些区域难以进行统计建模, 从而隐蔽性更好. 隐写分析是针对隐写技术的检测方法, 通过对多媒体信息进行分析, 检测其是否隐藏有秘密信息. 由于隐写算法的主要目的是进行隐蔽通信, 因此其性能采用隐写分析器进行评估, 即在嵌入相同秘密信息的前提下, 能够更好地抵抗隐写分析检测的隐写算法效果更好. 近年来, 硬件水平的不断提高和数据量的增长, 极大推动了深度学习的发展[7,8], 卷积神经网络[9], 生成对抗网络[10]等技术取得重大突破, 并在多个领域得到了广泛的应用[11,12]. 卷积神经网络在隐写分析中得到了很好的应用, 例如 Xu-Net[13]等网络性能已经超过了传统基于富模型 SRM[14]检测方法. 隐写分析与隐写技术相互对抗, 相互促进, 这和生成对抗模型非常相似, 于是研究者们想到利用生成对抗网络设计隐写失真函数, 并取得了较好的研究成果[15−17].

虽然目前基于生成对抗网络设计隐写失真函数在空域灰度图像上已经取得了重大突破, 但是目前还没有针对空域彩色图像采用生成对抗网络设计隐写失真函数的研究. 在现实生活中, 彩色图像应用更为广泛, 与灰度图像相比, 彩色图像通道之间具有相关性, 为了抵抗 SCRM[18]、CFA-aware-CRM[19]等加入了通道相关性分析的彩色图像隐写分析器的检测, 设计失真函数的时候需要考虑降低嵌密对通道相关性的破坏. 同时由于彩色图像 3 个通道纹理和嘈杂程度不尽相同, 适合嵌密的容量也各不一样, 因此需要合理地对 3 个通道进行容量分配. 考虑到实际应用的问题, 本文设计了一个基于生成对抗网络设计彩色图像隐写失真函数的框架 CIS-GAN (color image steganography based on generative adversarial network), 生成器网络采用两个 U-Net[20]子网络结构, 第 1 个 U-Net 子网络生成修改概率矩阵, 第 2 个 U-Net 子网络进行正负向修改概率调节, 针对彩色图像这一载体的特点修改灰度图像隐写分析器, 在预处理层, 首先将 3 个通道分别通过 SRM 滤波核进行卷积操作, 然后将得到的 3 部分特征图合并. 在生成器损失函数中对彩色图像 3 个通道总的隐写容量进行控制, 生成器能够自动学习分配 3 个通道容量. 本文的主要贡献如下.

(1) 首次针对空域彩色图像, 提出一种基于生成对抗网络的隐写失真函数设计方法, 与传统的基于经验手工设计彩色图像隐写失真函数的方法不同, 该方法能够在与彩色图像隐写分析器的对抗训练过程中自动地学习设计.

(2) 构建了一个可以自适应调整正负向修改概率的双 U-Net 生成器, 有效地减少对彩色图像通道相关性的破坏. 在生成器的损失函数中对彩色图像 3 个通道总嵌密容量进行控制, 使网络通过学习自适应地为彩色图像各个通道分配嵌密容量. 实验结果表明, 与现有的彩色图像隐写失真函数设计方法相比, 本文提出的网络结构能够更好

地抵抗彩色图像隐写分析器的检测.

本文第 2 节介绍相关工作; 第 3 节主要介绍本文设计的生成对抗网络结构, 具体包括网络的生成器、隐写分析器、损失函数和模型使用, 并对模型特点进行了分析; 第 4 节对实验设置、参数设置和实验结果进行介绍; 第 5 节对全文进行总结.

2 相关工作

自适应隐写算法主要由失真函数和最小化隐写失真嵌入编码算法两部分组成. 目前已有 Syndrome-Trellis Code[4]等成熟的嵌入编码算法, 研究主要集中为设计有效的隐写失真函数来提高隐写算法的安全性. 本文根据手工设计和网络设计分别对现有方法进行介绍.

2.1 手工设计隐写失真函数

Pevny 等人在文献 [21] 中提出了 HUGO (highly undetectable stego) 算法, 根据隐写前后从 SPAM (subtractive pixel adjacency matrix) [22]特征空间提取出的特征向量的变化量设计失真函数, 使得嵌密修改的像素主要集中在不易检测的纹理和边缘区域. Holub 等人在文献 [23] 中提出了 WOW (wavelet obtained weights) 算法, 该算法使用一组定向的高通滤波器计算方向残差, 通过聚合嵌密修改对每个方向残差的影响量, 促使容易建模的平滑和边缘区域失真值高, 而纹理和噪声区域失真值低, 可以更好得抵抗富模型隐写分析[14]检测. Holub 等人随后又在文献 [5] 中提出了 UNIWARD (universal wavelet relative distortion) 算法, 该算法可以应用在多种域中进行失真函数设计, 其中应用到空域被称为 S-UNIWARD (spatial universal wavelet relative distortion) 算法. S-UNIWARD 算法与 WOW 算法实现思路相近, 因此具有相似的性能. Li 等人在文献 [6] 中提出了 HILL (high-pass, low-pass, low-pass) 算法, 该算法使用一个高通滤波器和两个低通滤波器, 使嵌密修改集中在纹理区域. 该算法避免了纹理区域出现高失真值, 低失真值聚集在较大区域内的情况, 嵌密修改后更加难以检测. Fridrich 等人在文献 [24] 中采用模型驱动的方法来获取失真值, 将图像建模成一个服从高斯分布的独立变量序列, 载体图像进行嵌密操作后, 通过最小化载体图像和载密图像模型分布的 KL 散度来计算嵌密修改概率, 进而转化得到失真值. 随后, Sedighi 等人在文献 [25] 中对该方法进行拓展, 采用多元广义高斯模型代替多元高斯模型对像素点进行建模. Sedighi 等人在文献 [26] 中对图像噪声残差进行建模, 在模型中加入了与内容相关的建模误差来提高隐写算法的安全性. Zhou 等人在文献 [27] 中提出了失真分配规则 CPP (controversial pixels prior). 首先采用多种隐写方法设计失真函数, 这些方法在某些点分配的失真值差异较大, 在这些点嵌入秘密信息不容易被检测, 因此 CPP 规则为其分配低失真值, 使嵌密修改向这些位置集中. Hu 等人为了充分利用图像自身的纹理信息设计失真函数, 在文献 [28] 中采用非负矩阵因子分解的方法来预测图像像素值, 并利用像素之间相互依赖来计算失真值. Qin 等人在文献 [29] 中提出了 MRG (multivariate Gaussian for residuals) 算法, 该算法将经过高通滤波核处理得到的图像残差建模成服从多元高斯分布的独立变量, 安全性相比已有的模型驱动方法得到了进一步提高.

上述工作针对的是灰度图像, 在实际应用中, 彩色图像应用更加广泛, 同时彩色图像 3 个通道都可以嵌入秘密信息, 与灰度图像相比具有更大的容量, 因此有研究人员专门针对彩色图像载体设计失真函数. 早期的隐写方法将彩色图像 3 个通道看做 3 个独立的灰度图像, 这些方法没有考虑彩色图像通道之间的相关性. Tang 等人在文献 [30] 中提出了 CMDC (clustering modification directions for color components) 隐写策略, 根据嵌密后周围 6 个邻居的修改情况更新失真值, 使得同一像素位置上 3 个通道尽可能向相同方向修改, 从而降低对彩色图像通道相关性的破坏. Liao 等人在文献 [31] 中提出了 ACMP (amplifying channel modification probabilities) 隐写策略来增强彩色图像通道间相关性, 该方法同时考虑了彩色图像通道内相邻像素点之间的相关性与彩色图像通道间的相关性, 能够有效抵抗现有彩色图像隐写分析算法的检测.

2.2 网络设计隐写失真函数

Tang 等人在文献 [16] 中首次提出了基于生成对抗网络设计空域灰度图像隐写失真函数框架, 生成器网络输入灰度载体图像, 输出修改概率矩阵, 修改概率矩阵通过一个训练好的子网络进行模拟嵌入后得到修改矩阵, 灰度载体图像与修改矩阵相加得到灰度载密图像, 载体图像与载密图像一起作为隐写分析器的输入进行训练. Yang 等

人在文献 [17] 中设计了一个 Double-tanh 函数进行模拟嵌入, 并且将 U-Net 网络[20]作为生成器设计空域灰度图像的隐写失真函数, 性能超过了现有手工设计隐写失真函数的方法. Wu 等人在文献 [32] 中提出在 U-Net 收缩路径中结合多个特征图构建生成器网络, 该方案隐写安全性得到提高.

在现实生活场景下, 彩色图像较灰度图像应用更广泛, 使用频率更高, 因此更加适合作为隐写载体传递秘密信息. 然而, 现有的基于生成对抗网络的隐写失真函数设计方案都是针对灰度图像载体, 如果简单地将彩色图像 3 个通道视为灰度图像嵌入秘密信息, 会破坏彩色图像通道相关性, 从而很容易被加入通道相关性分析的彩色图像隐写分析器检测出来, 同时这种方法无法合理地对彩色图像 3 个通道进行嵌密容量分配.

3 本文提出的生成对抗网络结构

为了方便理解以及后续的使用, 首先对论文中使用的符号进行说明. 符号 i, j 指代彩色图像像素点下标, k 指代通道下标. 符号 $C=(c_{i,j,k})^{H\times W\times 3}$ 和 $S=(s_{i,j,k})^{H\times W\times 3}$ 分别表示彩色载体图像和对应的彩色载密图像, 其中 H 和 W 分别表示彩色图像的高和宽. 符号 $M=(m_{i,j,k})^{H\times W\times 3}=S-C$ 表示修改矩阵, 其中 $m_{i,j,k}\in\{+1,-1,0\}$. $P^{+}=(p_{i,j,k}^{+1})^{H\times W\times 3}$ 表示正向修改概率矩阵, 其中 $p_{i,j,k}^{+1}$ 表示 $c_{i,j,k}$ 进行+1 修改的概率. 同理 $P^{-}=(p_{i,j,k}^{-1})^{H\times W\times 3}$ 和 $P^{0}=(p_{i,j,k}^{0})^{H\times W\times 3}$ 分别代表负向修改概率矩阵和不进行修改的概率矩阵. 定义符号 $P=(p_{i,j,k})^{H\times W\times 3}$ 表示修改概率矩阵, 包含正向修改概率和负向修改概率, 即 $p_{i,j,k}=p_{i,j,k}^{+1}+p_{i,j,k}^{-1}$.

结合彩色图像载体的特点, 本文提出一个基于生成对抗网络学习框架 CIS-GAN (color image steganography based on generative adversarial network) 来设计彩色图像隐写失真函数. CIS-GAN 总体框架如图 1 所示, 主要包括双 U-Net 生成器和彩色图像隐写分析器两大模块. 彩色载体图片 C 首先作为双 U-Net 生成器的输入, 输出正向修改概率矩阵 P^{+} 和负向修改概率矩阵 P^{-}. 双 U-Net 生成器在单个 U-Net 网络的基础上增加一个 U-Net 网络来学习正负向修改概率的分配, 从而使彩色图像 3 个通道上的修改方向具有一定规律, 减少对彩色图像通道间相关性的破坏. 正向修改概率矩阵 P^{+} 和负向修改概率矩阵 P^{-} 经过 Double-tanh 模拟嵌入模块[17]得到修改矩阵 M, 彩色载体图像 C 与修改矩阵 M 相加得到彩色载密图像 S, 载体图像 C 与载密图像 S 一起作为彩色隐写分析器的输入进行训练. 彩色隐写分析器针对彩色图像进行了特殊的设计, 从而能够更好地学习彩色图像通道相关性这一特征, 在与生成器的不断对抗训练中, 促使生成器不断学习设计彩色图像隐写失真函数. 下面将对 CIS-GAN 网络结构的各个部分进行详细阐述.

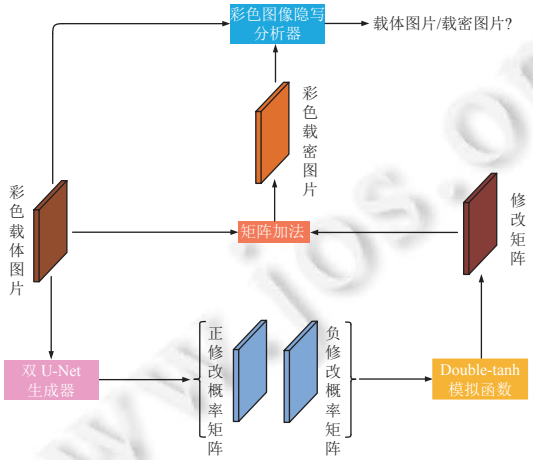


图 1 CIS-GAN 生成对抗网络框架

3.1 生成器结构

载体图像不同位置像素值修改后与原图像差异各不相同, 这种差异称为失真. 记 $\rho_{i,j,k}^{+}$ 和 $\rho_{i,j,k}^{-}$ 分别表示彩色图

像点 (i, j) 上第 k 个通道位置正向+1 和负向−1 修改的失真值. 自适应隐写遵循最小化失真原则, 最小化失真是在满足嵌入 L 比特秘密信息的约束条件下, 期望嵌入失真最小, 该模型采用如下公式进行表达.

$$\min \sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{3}(p_{i,j,k}^{+}\rho_{i,j,k}^{+}+p_{i,j,k}^{-}\rho_{i,j,k}^{-}) \tag{1}$$

$$\text{subject to } \sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{3}(-p_{i,j,k}^{+1}\log_2 p_{i,j,k}^{+1}-p_{i,j,k}^{-1}\log_2 p_{i,j,k}^{-1}-p_{i,j,k}^{0}\log_2 p_{i,j,k}^{0})=L \tag{2}$$

该最优化模型的最优解[4]为:

$$\begin{cases} p_{i,j,k}^{+}=\dfrac{\mathrm{e}^{-\lambda\rho_{i,j,k}^{+}}}{1+\mathrm{e}^{-\lambda\rho_{i,j,k}^{+}}+\mathrm{e}^{-\lambda\rho_{i,j,k}^{-}}} \\ p_{i,j,k}^{-}=\dfrac{\mathrm{e}^{-\lambda\rho_{i,j,k}^{-}}}{1+\mathrm{e}^{-\lambda\rho_{i,j,k}^{+}}+\mathrm{e}^{-\lambda\rho_{i,j,k}^{-}}} \end{cases} \tag{3}$$

其中, 参数 λ 可以通过将解代入公式 (2) 进行求解. 基于最小化失真设计隐写算法的框架包括两个步骤, 首先设计有效的失真函数, 然后根据公式 (3) 计算得到的概率分布进行最小化失真嵌入. 目前, 以 Syndrome-Trellis Code 为代表的最小化失真编码算法效果接近最小化失真理论上界, 并且在实际应用中得到了很好地运用, 因此研究的难点为设计 $\rho_{i,j,k}^{+}$ 和 $\rho_{i,j,k}^{-}$. 失真值和修改概率通过等式 (3) 进行相互转化, 这表示确定了彩色图像修改概率后, 可以得到隐写失真值. 在生成对抗网络 CIS-GAN 中通过生成器来生成 $p_{i,j,k}^{+}$ 和 $p_{i,j,k}^{-}$, 再通过公式 (3) 求解得到 $\rho_{i,j,k}^{+}$ 和 $\rho_{i,j,k}^{-}$ 后完成隐写失真函数的设计.

彩色图像包含红绿蓝 3 个通道, 通道之间具有相关性, 基于内容自适应算法会更多地倾向在纹理复杂度高的位置嵌入信息, 由于彩色图像 3 个通道具有相似性, 如果简单地将 3 个通道视为 3 张灰度图像进行秘密信息嵌入, 倾向于集中在同一像素 3 个通道上的位置进行修改. 现有的灰度图像生成隐写失真函数框架, 设置每个位置 $p_{i,j,k}^{+}=p_{i,j,k}^{-}$, 同一像素在 3 个通道对应位置上的修改方向完全随机, 容易破坏彩色图像通道间的相关性. 传统手工设计彩色图像隐写失真函数方案 CMDC[30]基于经验制定了一种方向一致性策略来调整正向修改失真 $\rho_{i,j,k}^{+}$ 和负向修改失真 $\rho_{i,j,k}^{-}$, 进而达到调整 $p_{i,j,k}^{+}$ 和 $p_{i,j,k}^{-}$ 的目的, 抵抗彩色图像隐写分析器检测的能力得到了很大的提高, 受 CMDC 的启发, 本文的生成器网络结构中, 利用一个子网络来学习正负向修改概率的调整. 在已知修改发生的情况下, 对应位置元素进行+1 和−1 修改的概率分别采用正向修改占比和负向修改占比表示. 定义符号 $W^{+}=(w_{i,j,k}^{+})^{H\times W\times 3}$ 和 $W^{-}=(w_{i,j,k}^{-})^{H\times W\times 3}$ 分别表示正向修改概率占比矩阵和负向修改概率占比矩阵, 其中 $w_{i,j,k}^{+1}=1-w_{i,j,k}^{-1}$. 生成器第 1 个子网络用来生成修改概率矩阵 P, 第 2 个子网络生成正向修改概率占比矩阵 W^{+} 来学习正负向修改概率的比例分配, 这两个任务都可以看作是图像生成图像任务. U-Net[20]是轻量级网络, 并且在图像生成图像任务应用非常成功, 因此两个子网络都使用 U-Net 网络.

双 U-Net 生成器网络结构如图 2 所示, 生成器采用两个 U-Net 子网络结构, 分别用 U1 和 U2 指代. U1 子网络输入彩色载体图片 C, 生成修改概率矩阵 P, U2 子网络根据输入的载体图片 C 和 U1 子网络的输出 P, 生成正向修改概率占比矩阵 W^{+}, 负向修改概率占比矩阵 W^{-} 通过数值 1 与正向修改概率占比矩阵 W^{+} 相减得到, 然后与修改概率矩阵 P 进行矩阵点乘运算分别得到正向修改概率矩阵 P^{+} 和负向修改概率矩阵 P^{-}.

$$p_{i,j,k}^{+}=p_{i,j,k}\times w_{i,j,k}^{+} \tag{4}$$

$$p_{i,j,k}^{-}=p_{i,j,k}\times w_{i,j,k}^{-} \tag{5}$$

图 2 中使用了一个简单的例子来展现这个过程. 假设 U1 子网络生成的修改概率矩阵 P 某一 2×2 大小的正方形小块位置上 3 个通道取值为 ((0.35, 0.39, 0.40, 0.43), (0.12, 0.20, 0.20, 0.26), (0.15, 0.22, 0.23, 0.22)), U2 子网络生成的正向修改概率占比矩阵 W^{+} 对应位置上 3 个通道取值为 ((0.69, 0.61, 0.40, 0.02), (0.19, 0.68, 0.25, 0.97), (0.54, 0.47, 0.30, 0.30)). 数值 1 与正向修改概率占比矩阵 W^{+} 相减得到负向修改概率占比 W^{-}, 计算得到对应位置值为 ((0.31, 0.39, 0.60, 0.98), (0.81, 0.32, 0.75, 0.03), (0.46, 0.53, 0.70, 0.70)). 正向修改概率占比矩阵 W^{+} 与修改概率矩阵 P 点乘得到正向修改概率矩阵 P^{+}, 对应位置值为 ((0.24, 0.24, 0.16, 0.01), (0.02, 0.14, 0.05, 0.25), (0.08, 0.10, 0.07,

0.07)), 同理可得负向修改概率矩阵 P^- 对应位置值为 ((0.11, 0.15, 0.24, 0.42), (0.10, 0.06, 0.15, 0.01), (0.07, 0.12, 0.16, 0.15)). 通过矩阵 W^+ 完成了对正负向修改概率的调整.

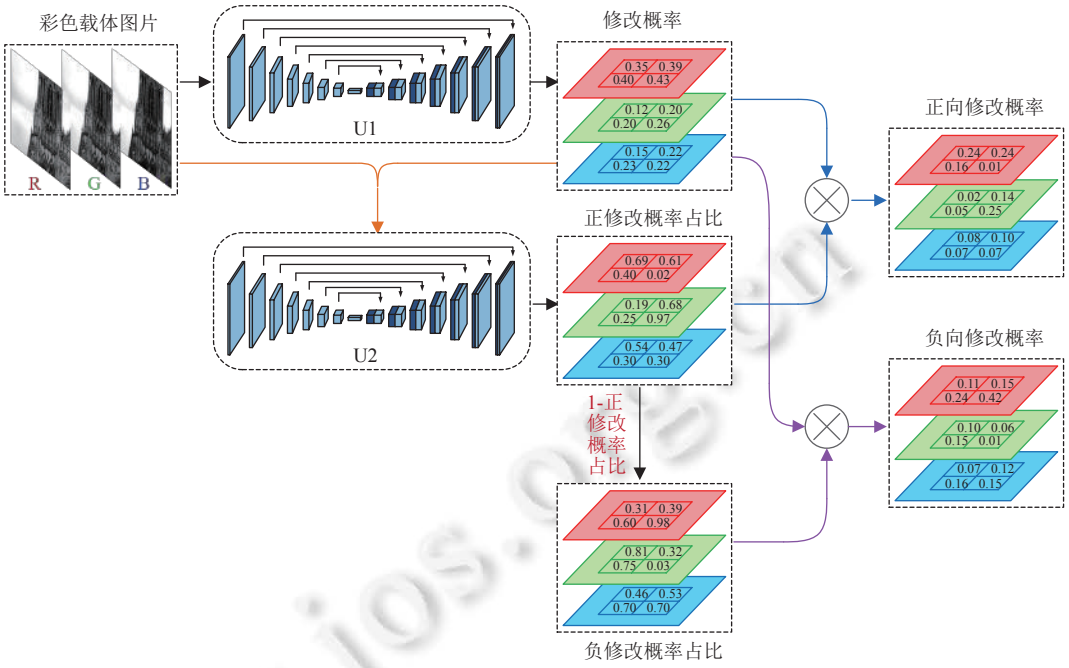


图 2 双 U-Net 生成器网络结构图

与 Yang 等人[17]直接应用 U-Net 网络作为生成器模型不同的是, 本文在考虑彩色图像通道相关性的基础上, 提出双 U-Net 网络结构, 该结构可以学习对正负向修改概率进行调节. 相比传统手工基于领域知识和经验调整正负向修改概率的彩色图像隐写失真函数设计方案 CMDC, 提出的双 U-Net 生成器网络结构可以自动地在与生成器对抗训练中进行学习.

3.2 彩色隐写分析器

在生成对抗网络模型中, 生成器网络依靠对抗网络的反馈来进行学习, 对抗网络是决定最终训练效果的关键因素之一. Xu-Net[13]是经典的灰度图像隐写分析网络, 它的网络模型较小并且性能优异. 为了让隐写分析器有效地检测彩色图片是否嵌密, 从而更好地指导生成器设计隐写失真函数, 本文对该灰度图像隐写分析器进行相应的改进. 改进后彩色图像隐写分析器的网络结构如图 3 所示, 在预处理阶段, 为了更好地保持信噪比, 采用先分后合[33]的方案, 首先将彩色图像 3 通道分别与 8 个固定的 SRM 滤波核[14]进行卷积, 然后将得到的 3 部分特征图合并成一个通道数为 24 的特征图.

3.3 损失函数

彩色隐写分析器的损失函数:

$$l_D = -\sum_{i=1}^{2} y_i' \log(y_i) \tag{6}$$

其中, y_1 和 y_2 是判别器 D 经过 Softmax 激活函数后的输出, y_1' 和 y_2' 对应真实的标签值.

生成器的损失函数包括两个部分, 第 1 部分用于与隐写分析器进行对抗, 将其设为隐写分析器损失函数的相反数:

$$l_G^1 = -l_D \tag{7}$$

第 2 部分用于控制嵌密容量, 彩色图像 3 个通道纹理各不相同, 适合嵌密的容量也不一样, 简单地将彩色图像

当做 3 张灰度图片嵌入相同容量的秘密信息并不合适, 因此对彩色图像 3 个通道嵌密容量求和来控制总的嵌密容量固定, 生成器在对抗学习的过程中可以学习为不同的通道分配合适的嵌密容量. 彩色图像嵌密容量 capacity (单位是 bit) 计算如下:

$$cap_k = \sum_{i=1}^{H}\sum_{j=1}^{W}(-p_{i,j,k}^{+1}\log_2 p_{i,j,k}^{+1} - p_{i,j,k}^{-1}\log_2 p_{i,j,k}^{-1} - p_{i,j,k}^{0}\log_2 p_{i,j,k}^{0}) \tag{8}$$

$$capacity = \sum_{k=1}^{3} cap_k \tag{9}$$

生成器第 2 部分损失函数为:

$$l_G^2 = (capacity - 3\times H\times W\times q)^2 \tag{10}$$

其中, q 表示设置的嵌密率, 单位为 bpc (bits per channel pixel), cap_k 表示通道 k 的嵌密容量. 生成器在训练的过程中会让 l_G^2 减小, 从而让嵌密容量不断接近期望值 $3\times H\times W\times q$. 由于控制的是 3 个通道嵌密容量的总和, 具体每个通道嵌密容量 cap_k 并没有约束, 因此生成器可以在与隐写分析器对抗学习的过程中自动学习如何进行通道间容量分配.

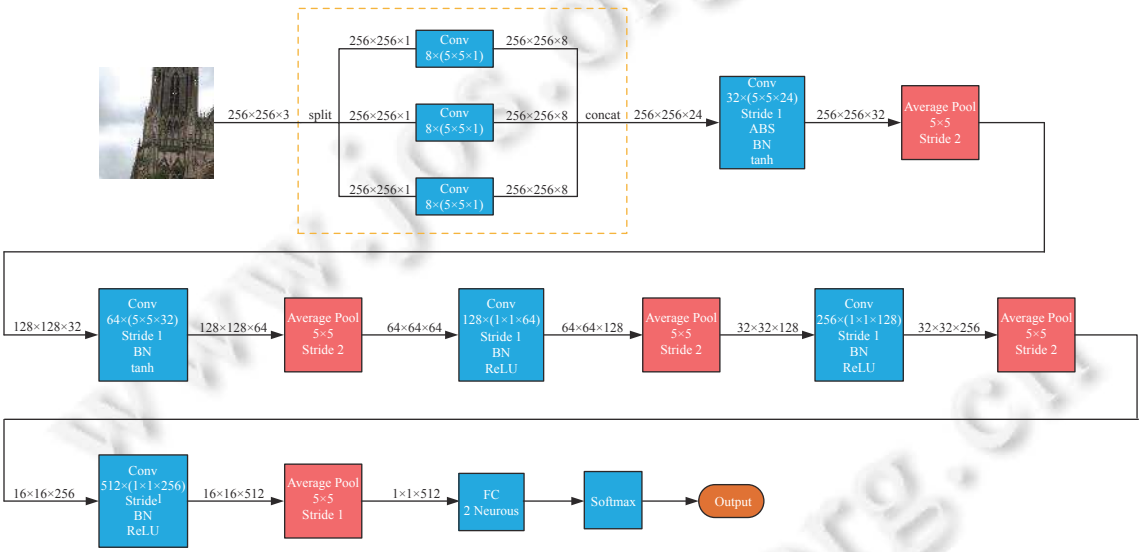


图 3 彩色图像隐写分析器网络结构图

生成器需要兼顾对抗隐写分析器和控制嵌密容量的任务, 通过一定的比例将两部分损失函数进行相加:

$$l_G = l_G^1 + \alpha\times l_G^2 \tag{11}$$

其中, α 是超参数, 用来平衡安全性和嵌密容量, α 值太小无法有效控制容量, α 值太大不能很好地对抗隐写分析器, 具体的取值需要在实验中进行尝试.

3.4 实际隐写嵌入

生成器直接生成的不是失真值, 而是修改概率, 网络训练完成后实际使用时需要采用多层 STC 构造的翻转引理[4]将修改概率转换为失真值. 假设等式 (2) 成立, 即生成器网络生成的修改概率矩阵对应的实际嵌密容量 capacity 等于目标嵌密容量, 不失一般性, 令参数 λ=1, 通过近似处理, 由最优修改概率公式 (3) 推导出隐写失真与对应修改概率之间的转化公式.

$$\begin{cases}\rho_{i,j,k}^{+1} = -\ln(p_{i,j,k}^{+1}/(1-p_{i,j,k}^{+1}-p_{i,j,k}^{-1}))\\ \rho_{i,j,k}^{-1} = -\ln(p_{i,j,k}^{-1}/(1-p_{i,j,k}^{+1}-p_{i,j,k}^{-1}))\end{cases} \tag{12}$$

其中, 在计算得到失真值后, 真实的λ取值可以通过采用二分查找的方法使等式 (2) 成立进行求解.

实际应用中采用 Syndrome-Trellis Code[4]等最小化嵌入编码算法将秘密信息嵌入彩色载体图像中. 为了进一步保障信息的安全性, 将要发送的信息在使用 Syndrome-Trellis Code 编码嵌入之前先进行加密, 然后再采用 Syndrome-Trellis Code 编码, 结合 3 个通道各自的正负失真值矩阵和嵌密容量, 将加密后的秘密信息嵌入到彩色载体图片的 3 个通道中得到载密图片. 发送方将载密图片通过网络媒介传输给接收方, 接收方收到图片后先用 Syndrome-Trellis Code 解码得到加密后的秘密信息, 然后对加密信息进行解密最终完成一次通信.

3.5 模型分析

本文提出的基于生成对抗网络的彩色图像隐写失真函数设计方法, 生成器由生成修改概率矩阵与调整正负向修改概率占比两部分组成. 生成器通过与隐写分析器进行对抗训练, 学习使彩色图像 3 个通道位置上的修改方向符合某种特殊的规律, 从而减少对彩色图像通道间相关性的破坏. 与现有的传统彩色图像隐写失真函数设计方案 CMDC[30]相比, CMDC 调整正负失真的方案基于经验进行设计, 需要结合相应的领域知识, 而在本方案中, 调整正负失真值的任务由双 U-Net 生成器网络中的 U2 子网络完成, 通过不断地与隐写分析器进行对抗训练, 自动地从海量数据中学习到更好的设计. 实验第 4.3.2 节采用 RCCI (relative channel correlation index) 指标[30]进行通道相关性分析, 实验结果证明了 CIS-GAN 网络设计隐写失真函数方案与其他彩色图像隐写失真函数设计方案相比, 对彩色图像通道相关性影响最小.

彩色图像 3 个通道纹理和嘈杂程度不尽相同, 适合嵌密的容量也各不一样, 特别是针对一些特殊的彩色图片, 3 个通道纹理区域差别很大, 如果简单地设置平分嵌密容量, 纹理相对复杂的通道还可以嵌入更多的秘密信息, 然而纹理相对简单的通道嵌入的信息超过其最大安全隐写容量, 很容易被隐写分析器检测出来. 提出的 CIS-GAN 网络在生成器的损失函数中对彩色图像 3 个通道总的嵌密容量进行控制, 首次通过网络自动学习对彩色图像 3 个通道进行容量分配, 与彩色图像 3 个通道均等分配方案相比能够更好地抵抗彩色图像隐写分析方法的检测.

计算量用来衡量模型的复杂度. 设一个卷积核的大小为 $K_w \times K_h$, 输入特征图通道数为 C_{in}, 大小为 $H_{\text{in}} \times W_{\text{in}}$, 输出特征图通道数为 C_{out}, 大小为 $H_{\text{out}} \times W_{\text{out}}$. 卷积层的计算量为 $C_{\text{out}} \times K_w \times K_h \times C_{\text{in}} \times H_{\text{out}} \times W_{\text{out}}$. 反卷积层计算量为 $C_{\text{out}} \times K_w \times K_h \times C_{\text{in}} \times H_{\text{in}} \times W_{\text{in}}$. 根据上述计算公式, 提出的模型中所有卷积层和反卷积层总的计算量为 32.36 亿次浮点运算.

4 实验结果与分析

4.1 数据集和实验设置

本文实验使用 BOSSBase[34]图像库中的 10 000 张全分辨率原始彩色图像, 首先采用 Photoshop CS6 进行去马赛克处理, 然后采用双线性插值算法缩放得到大小为 512×512 的彩色图像. 随机从上述 10 000 张图像选出 8 000 张图像, 上下、左右分别对半裁剪得到 32 000 张大小为 256×256 的彩色图像作为 CIS-GAN 训练图像集, 记为图像集 A. 剩余的 2 000 张图像上下、左右分别对半裁剪得到 8 000 张大小为 256×256 的彩色图像作为测试图像集, 记为图像集 B. 安全性性能评估采用 SCRM[18]和 CFA-aware-CRM[19]这两个经典的彩色图像隐写分析算法提取隐写分析特征, 并结合集成分类器[35]进行分类. 具体的做法是, 从图像集 B 中选取一半的图像和相对应的载密图像一起采用隐写分析算法提取特征后训练集成分类器, 剩余的一半图像与对应的载密图像采用隐写分析算法提取特征后, 使用训练好的集成分类器进行分类. 隐写算法检测性能采用分类测试错误率来衡量, 测试错误率是由误检率和漏检率组成, 测试错误率越高, 隐写算法抵抗隐写分析的性能越好. 提出的方案通过 TensorFlow 框架实现, 实验采用 NVIDIA RTX2080ti 显卡.

4.2 训练流程和参数设置

图 4 展示了 CIS-GAN 整体训练流程. 生成器和隐写分析器的网络参数在训练前随机进行初始化, 在训练阶段每一批次 (batch size) 设为 24, 每迭代一次交替更新一次隐写分析器和生成器的参数. 实验中采用 Adam 优化器来训练网络, 生成器和隐写分析器学习率均设置为 0.000 1. 训练的时候先固定生成器的参数, 通过 Adam 优化器优化

公式 (6) 更新隐写分析器的参数, 再固定隐写分析器的参数, 通过 Adam 优化器优化公式 (11) 对生成器进行参数更新, 迭代 120000 次时结束训练. 训练完成后, 使用训练好的 CIS-GAN 模型生成图像集 B 对应的载密图像, 作为隐写分析器数据集, 进行后续的安全性分析.

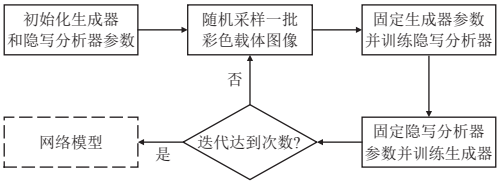


图 4 CIS-GAN 整体训练流程框图

为了让生成器兼顾与隐写分析器对抗和控制嵌密容量的任务, 在公式 (11) 中通过参数 α 来平衡安全性和嵌密容量. 本实验中针对 0.5 bpc 嵌密率进行 α 参数选取实验, 隐写分析检测算法采用 SCRM, 模型训练好后对数据集 B 进行测试, 实际嵌入率取平均值用于衡量控制嵌密容量的效果. α 值太小无法有效控制嵌密容量, 尝试取值从 10^{-9} 开始进行实验, 增加一个数量级取值为 10^{-8}, 以此类推, 每次增加一个数量级进行实验. 表 1 展示了在 α 取不同值时, CIS-GAN 在抵抗隐写分析器检测以及控制嵌密容量两个方面的效果, 其中测试错误率越高, 隐写算法抵抗隐写分析检测的性能越好, 平均嵌密率越接近目标值 0.5 bpc, 控制嵌密容量的效果越好. 通过分析发现, 随着 α 值不断增大, CIS-GAN 抵抗隐写分析器的效果逐渐变差, 特别是 α 取值为 10^{-5} 时, 效果下降非常明显, 因此大于 10^{-5} 的取值就不予以考虑. 而当 α 取值为 10^{-9} 和 10^{-8} 时, 抵抗隐写分析器检测的效果相比 α 取值为 10^{-7} 时效果提升不大, 且不能很好地控制嵌密容量, 经过权衡, 后续实验 α 取值设置为 10^{-7}.

表 1 CIS-GAN 在嵌密率 0.5 bpc 下 α 不同取值对比

α	10^{-9}	10^{-8}	10^{-7}	10^{-6}	10^{-5}
测试错误率	0.1910	0.1893	0.1876	0.1763	0.0424
平均嵌密率(bpc)	0.4842	0.4933	0.4981	0.4983	0.4991

4.3 实验结果

本文对比实验采用传统的彩色图像隐写方案 CMDC[30]和 ACMP[31], 这两种方案均需要与灰度图像隐写方法相结合, 实验中选取了两种经典的灰度图像隐写方法 S-UNIWARD[5]和 HILL[6], S-UNIWARD 与 CMDC 组合命名为 CMDC-SUNIWARD, HILL 与 CMDC 组合命名为 CMDC-HILL. 同理 S-UNIWARD 和 HILL 与 ACMP 组合分别得到 ACMP-SUNIWARD 和 ACMP-HILL 两组实验. 由于目前还没有采用生成对抗网络生成彩色图像隐写失真函数方法, 因此本文采用经典的灰度图像隐写失真函数设计网络 UT-6HPF-GAN[17]并结合均分方案进行对比实验. 将彩色图片训练集 A 拆成 32000×3 张灰度图片作为 UT-6HPF-GAN 网络训练数据集, 嵌密的时候对测试数据集 B 中彩色图片的 3 个通道分别进行嵌密.

4.3.1 CIS-GAN 网络对抗学习有效性

图 5 展示了 CIS-GAN 网络在 0.5 bpc 嵌密率 (嵌入容量为 98304 bit) 下, 针对图 5(a) 进行嵌密的结果, 其中图 5(b)–(d) 分别对应彩色图像 RGB 这 3 个通道图. 图 5(e)–(g) 为采用 CIS-GAN 生成的修改矩阵图, 其中灰色的点表示该位置不进行嵌密修改, 白色的点表示该位置进行−1 修改, 黑色的点表示该位置进行+1 修改. CIS-GAN 方法为彩色图像 3 个通道自动分配的容量分别为 45066 bit、25068 bit、28170 bit. 图 5(h)–(j) 为采用 CMDC-HILL 生成的修改矩阵图, 图 5(k)–(m) 为采用 ACMP-HILL 生成的修改矩阵图, 图 5(n)–(p) 为采用 UT-6HPF-GAN 生成的修改矩阵图. 通过观察发现 CIS-GAN 方法不仅能够像传统方法 CMDC-HILL 和 ACMP-HILL 一样将秘密信息嵌入到图像纹理复杂区域, 并且能够在与隐写分析器对抗学习的过程中自动学习如何进行通道间容量分配.

(a) 彩色载体图像

(b) 红色通道

(c) 绿色通道

(d) 蓝色通道

(e) 红色通道嵌密 (CIS-GAN)

(f) 绿色通道嵌密 (CIS-GAN)

(g) 蓝色通道嵌密 (CIS-GAN)

(h) 红色通道嵌密 (CMDC-HILL)

(i) 绿色通道嵌密 (CMDC-HILL)

(j) 蓝色通道嵌密 (CMDC-HILL)

(k) 红色通道嵌密 (ACMP-HILL)

(l) 绿色通道嵌密 (ACMP-HILL)

(m) 蓝色通道嵌密 (ACMP-HILL)

(n) 红色通道嵌密 (UT-6HPF-GAN)

(o) 绿色通道嵌密 (UT-6HPF-GAN)

(p) 蓝色通道嵌密 (UT-6HPF-GAN)

图 5　在 0.5 bpc 下嵌密修改矩阵图

4.3.2　通道相关性分析

为了验证 CIS-GAN 网络能够有效降低对彩色图像通道相关性的破坏, 采用相对通道相关性系数 RCCI[30]作为衡量指标, RCCI 值越小表明嵌密对通道相关性的破坏程度越小. 隐写分析器首先采用滤波器得到残差图像, 再

从残差图像提取特征, 本文采用 KV 核[36]提取残差图像, KV 滤波核 $F^{(0)}$ 的定义如下:

$$F^{(0)}=\frac{1}{12}\times\begin{bmatrix}-1 & 2 & -2 & 2 & -1\\ 2 & -6 & 8 & -6 & 2\\ -2 & 8 & -12 & 8 & -2\\ 2 & -6 & 8 & -6 & 2\\ -1 & 2 & -2 & 2 & -1\end{bmatrix} \tag{13}$$

给定彩色载体图像 C 和对应的载密图像 S, 经过 KV 核处理后得到的残差图像分别记为 C_r 和 S_r.

$$CCI(I)=CC(I_1,I_2)+CC(I_2,I_3)+CC(I_1,I_3) \tag{14}$$

$$RCCI=\frac{CCI(C_r)-CCI(S_r)}{CCI(C_r)} \tag{15}$$

其中, I_1, I_2, I_3 分别表示彩色图像 3 个通道, $CC(I_1,I_2)$ 用于计算 I_1 和 I_2 两个通道之间皮尔逊相关系数[37].

实验中针对图像集 B 中的 8 000 张图像, 在嵌密率为 0.5 bpc 条件下计算 RCCI 并求平均值, 实验结果如表 2 所示. 本文提出的 CIS-GAN 在对比方法中 RCCI 值最小, 在嵌入相同容量秘密信息的前提下, CIS-GAN 方法能够更好地保护彩色图像 3 个通道之间的相关性.

表 2 不同隐写方法在嵌密率 0.5 bpc 下 RCCI ($\times10^{-2}$) 指标

隐写方法	RCCI
ACMP-SUNIWARD[31]	13.77
ACMP-HILL[31]	12.99
CMDC-SUNIWARD[30]	8.00
CMDC-HILL[30]	7.77
UT-6HPF-GAN[17]	12.21
CIS-GAN	4.88

4.3.3 抵抗经典彩色图像隐写分析器

为了有效利用已有训练模型, 实验采用迁移学习的方法, 首先训练 0.5 bpc 嵌密率下的模型, 然后将训练好的 0.5 bpc 嵌密率下的模型权重初始化 0.4 bpc 嵌密率下的模型, 以此类推到用训练好的 0.2 bpc 嵌密率下的模型权重初始化 0.1 bpc 嵌密率下的模型. 在做迁移学习时, 为了避免学习率过大而丢失之前嵌密率下学习到的信息, 将学习率设置为 0.000 01. 为了保证公平对比, UT-6HPF-GAN 采用相同的迁移学习方式得到 0.4 bpc、0.3 bpc、0.2 bpc 和 0.1 bpc 嵌密率下的模型.

表 3 展示了采用 SCRM 隐写分析器进行检测时, 现有隐写算法 ACMP-SUNIWARD、ACMP-HILL、CMDC-SUNIWARD、CMDC-HILL、UT-6HPF-GAN, 以及本文提出的 CIS-GAN 在 0.5 bpc、0.4 bpc、0.3 bpc、0.2 bpc 和 0.1 bpc 这 5 种嵌密率下测试错误率. 表 4 展示了在 5 种嵌密率下, CFA-aware-CRM 隐写分析器分别针对不同彩色图像隐写方法的检测效果. 图 6 是直观展示表 3 和表 4 数据的折线图.

表 3 不同隐写方法抵抗 SCRM 隐写分析测试错误率

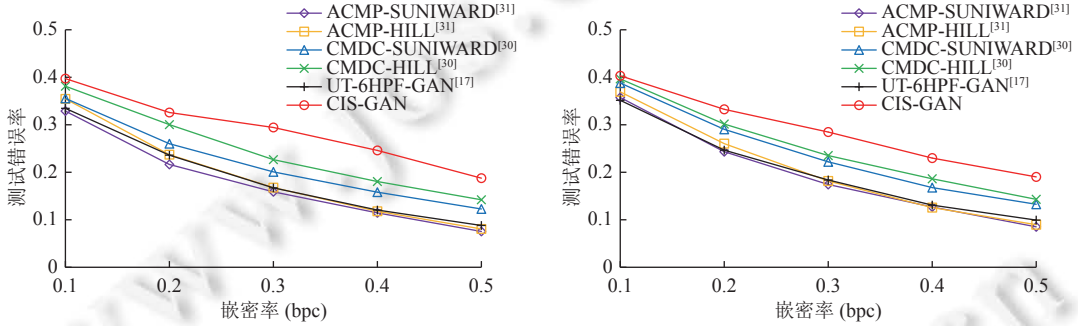
隐写方法	0.5 bpc	0.4 bpc	0.3 bpc	0.2 bpc	0.1 bpc
ACMP-SUNIWARD[31]	0.075 4	0.114 5	0.158 8	0.216 4	0.328 9
ACMP-HILL[31]	0.080 7	0.117 9	0.167 7	0.237 1	0.354 5
CMDC-SUNIWARD[30]	0.122 7	0.158 1	0.200 8	0.260 1	0.355 5
CMDC-HILL[30]	0.142 1	0.180 5	0.226 6	0.300 3	0.381 2
UT-6HPF-GAN[17]	0.088 0	0.120 5	0.167 2	0.236 0	0.334 6
CIS-GAN	0.187 6	0.246 0	0.294 3	0.325 8	0.397 2

从表 3、表 4 和图 6 可以看出本文提出的 CIS-GAN 在抵抗经典彩色图像隐写分析方法 SCRM 和 CFA-aware-CRM 上性能比现有的彩色图像隐写方法更好. 与 CMDC-HILL 方法相比, CIS-GAN 在嵌密率为 0.5 bpc 至

0.1 bpc 时, 抵抗 SCRM 隐写分析算法检测的性能提高了 0.016 至 0.067, 抵抗 CFA-aware-CRM 隐写分析算法检测的性能提高了 0.006 至 0.049. 双 U-Net 生成器网络中 U2 子网络学习调整正负失真值, 通过不断地与隐写分析器进行对抗训练, 自动地从海量数据中学习到更好的设计, 与 CMDC 方案手工调整正负失真相比更具优势. 与 UT-6HPF-GAN 方法相比, CIS-GAN 在嵌密率为 0.5 bpc 至 0.1 bpc 时, 抵抗 SCRM 隐写分析算法检测的性能提高了 0.062 至 0.125, 抵抗 CFA-aware-CRM 隐写分析算法检测的性能提高了 0.051 至 0.1. 与直接应用灰度图像隐写分析方法 UT-6HPF-GAN 相比, CIS-GAN 在网络中加入了通道相关性的设计, 同时生成器能够自动学习分配 3 个通道嵌密容量, 因此安全性能有了较大提升.

表 4　不同隐写方法抵抗 CFA-aware-CRM 隐写分析测试错误率

隐写方法	0.5 bpc	0.4 bpc	0.3 bpc	0.2 bpc	0.1 bpc
ACMP-SUNIWARD[31]	0.0845	0.1263	0.1736	0.2426	0.3567
ACMP-HILL[31]	0.0890	0.1250	0.1810	0.2596	0.3685
CMDC-SUNIWARD[30]	0.1320	0.1671	0.2213	0.2897	0.3871
CMDC-HILL[30]	0.1424	0.1856	0.2345	0.3008	0.3959
UT-6HPF-GAN[17]	0.0987	0.1302	0.1831	0.2459	0.3508
CIS-GAN	0.1896	0.2292	0.2840	0.3318	0.4024



(a) SCRM 隐写分析检测　　(b) CFA-aware-CRM 隐写分析

图 6　不同隐写方法抵抗隐写分析测试错误率

4.3.4　消融实验分析

为了进一步验证双 U-Net 生成器和网络自动学习容量分配策略的有效性, 增加两组对比实验.

1) 将 CIS-GAN 的生成器部分改成单个 U-Net 网络, 生成修改概率矩阵, 正向修改概率矩阵和负向修改概率矩阵设置成各占 1/2, 该方案记为 Variant 1.

2) 将 CIS-GAN 方案中容量控制部分采用均分策略, 即 3 个通道嵌入相同的容量, 该方案记为 Variant 2. 生成器损失函数中容量控制部分改为如下公式:

$$l_G'^2=\sum_{k=1}^{3}\left(\left(\sum_{i=1}^{H}\sum_{j=1}^{W}\left(-p_{i,j,k}^{+1}\log_2 p_{i,j,k}^{+1}-p_{i,j,k}^{-1}\log_2 p_{i,j,k}^{-1}-p_{i,j,k}^{0}\log_2 p_{i,j,k}^{0}\right)\right)-H\times W\times q\right)^2 \tag{16}$$

表 5 展示了在嵌密率为 0.5 bpc 下, CIS-GAN 与两组对比方案分别抵抗 SCRM 和 CFA-aware-CRM 隐写分析方法检测的实验结果.

表 5　不同方案抵抗隐写分析方法测试错误率

隐写分析方法	CIS-GAN	Variant 1	Variant 2
SCRM	0.1876	0.0876	0.1568
CFA-aware-CRM	0.1896	0.0945	0.1570

从表 5 实验结果看, CIS-GAN 相比 Variant 1 方案在抵抗 SCRM 和 CFA-aware-CRM 检测上性能分别提高 0.1 和 0.0951. Variant 1 方案生成修改概率矩阵后设置每个位置正向修改概率和负向修改概率相等, 同一像素在 3 个通道位置上的修改方向完全随机, 容易破坏彩色图像通道间的相关性. 双 U-Net 生成器增加了一个子网络来学习正负向修改概率的分配, 实验效果证实了这种方案能够更好地抵抗彩色隐写分析器的检测. CIS-GAN 相比 Variant 2 方案在抵抗 SCRM 和 CFA-aware-CRM 检测上性能分别提高 0.0308 和 0.0326. 由于彩色图像 3 个通道纹理各不相同, 适合嵌密的容量也不一样, CIS-GAN 在生成器的损失函数中对彩色图像 3 个通道总的隐写容量进行控制, 通过网络自动学习对彩色图像 3 个通道进行容量分配, 与彩色图像 3 个通道均等分配方案相比能够更好地抵抗彩色图像隐写分析方法的检测.

4.3.5 运行效率

CMDC 与 ACMP 是传统的失真函数设计算法, 运行在 CPU 上, 而 CIS-GAN 运行在 GPU 上, 硬件环境不一致, 无法进行公平的对比. 因此, 表 6 仅展示了本文提出的 CIS-GAN 与 UT-6HPF-GAN 在训练阶段迭代 12 万次耗时以及测试阶段平均生成单张彩色图像隐写失真函数所消耗的时间. 由于 CIS-GAN 生成器采用双 U-Net 网络, 相比 UT-6HPF-GAN 网络复杂度更高, 在训练的时候消耗的时间更长. 网络训练是算法预处理的过程, 影响有限, 因此网络更加关注测试时间消耗. 模型训练好后, 二者在实际使用中处理单张彩色图像平均耗时相差不大, 但 CIS-GAN 相比 UT-6HPF-GAN 抵抗彩色图像隐写分析器检测的安全性有了较大提升, 因此总体性能更优.

表 6 训练与测试时间

方法	训练 (迭代12万次)(h)	测试 (一张彩色图像)(s)
CIS-GAN	16.3	0.022
UT-6HPF-GAN[17]	5.3	0.019

5 结束语

本文针对空域彩色图像这一载体的特殊性, 提出了一个基于生成对抗网络设计隐写失真函数的框架 CIS-GAN. 生成器采用双 U-Net 网络, 生成的嵌密失真函数能够有效地降低对彩色图像 3 个通道间相关性的破坏. 在生成器的损失函数中对彩色图像 3 个通道总的隐写容量进行控制, 生成器能够自动学习分配 3 个通道的嵌密容量. 实验结果证明, 该网络结构在抵抗 SCRM 和 CFA-aware-CRM 两种经典空域彩色图像隐写分析器时, 安全性相比现有的彩色图像隐写失真函数设计方法有了较大的提高. 在现实应用场景中, 由于 JPEG 格式图像相比空域图像占用更小的存储空间, 因此应用更为广泛. 下一步我们将基于生成对抗网络, 深入研究 JPEG 格式的彩色图像的隐写方法.

References:

[1] Li B, He JH, Huang JW, Shi YQ. A survey on image steganography and steganalysis. Journal of Information Hiding and Multimedia Signal Processing, 2011, 2(2): 142–172.

[2] Sun X, Zhang WM, Yu NH, Wei Y. Steganography based on parameters' disturbance of spatial image transform. Journal on Communications, 2017, 38(10): 166–174 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x.2017155]

[3] Shen J, Liao X, Qin Z, Liu XC. Spatial steganalysis of low embedding rate based on convolutional neural network. Ruan Jian Xue Bao/Journal of Software, 2021, 32(9): 2901–2915 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5980.htm [doi: 10.13328/j.cnki.jos.005980]

[4] Filler T, Judas J, Fridrich J. Minimizing additive distortion in steganography using Syndrome-Trellis Codes. IEEE Trans. on Information Forensics and Security, 2011, 6(3): 920–935. [doi: 10.1109/TIFS.2011.2134094]

[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. [doi: 10.1186/1687-417X-2014-1]

[6] Li B, Wang M, Huang JW, Li XL. A new cost function for spatial image steganography. In: Proc. of the 2014 IEEE Int'l Conf. on Image Processing. Paris: IEEE, 2014. 4206–4210. [doi: 10.1109/ICIP.2014.7025854]

[7] Tian X, Wang L, Ding Q. Review of image semantic segmentation based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 440–468 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5659.htm [doi: 10.13328/j.cnki.jos.005659]

[8] Luo HL, Tong K, Kong FS. The progress of human action recognition in videos based on deep learning: A review. Acta Electronica Sinica, 2019, 47(5): 1162–1173 (in Chinese with English abstract). [doi: 10.3969/j.issn.0372-2112.2019.05.025]

[9] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: ACM, 2012. 1097–1105.

[10] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 2672–2680.

[11] Zhang ZK, Pang WG, Xie WJ, Lü MS, Wang Y. Deep learning for real-time applications: A survey. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2654–2677 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

[12] Wang WL, Li ZR. Advances in generative adversarial network. Journal on Communications, 2018, 39(2): 135–148 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x.2018032]

[13] Xu GS, Wu HZ, Shi YQ. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 2016, 23(5): 708–712. [doi: 10.1109/LSP.2016.2548421]

[14] Fridrich J, Kodovský J. Rich models for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2012, 7(3): 868–882. [doi: 10.1109/TIFS.2012.2190402]

[15] Zhai LM, Jia J, Ren WX, Xu YB, Wang LN. Recent advances in deep learning for image steganography and steganalysis. Journal of Cyber Security, 2018, 3(6): 2–12 (in Chinese with English abstract). [doi: 10.19363/J.cnki.cn10-1380/tn.2018.11.01]

[16] Tang WX, Tan SQ, Li B, Huang JW. Automatic steganographic distortion learning using a generative adversarial network. IEEE Signal Processing Letters, 2017, 24(10): 1547–1551. [doi: 10.1109/LSP.2017.2745572]

[17] Yang JH, Ruan DY, Huang JW, Kang XG, Shi YQ. An embedding cost learning framework using GAN. IEEE Trans. on Information Forensics and Security, 2020, 15: 839–851. [doi: 10.1109/TIFS.2019.2922229]

[18] Goljan M, Fridrich J, Cogranne R. Rich model for steganalysis of color images. In: Proc. of the IEEE Int'l Workshop on Information Forensics and Security. Atlanta: IEEE, 2014. 185–190. [doi: 10.1109/WIFS.2014.7084325]

[19] Goljan M, Fridrich J. CFA-aware features for steganalysis of color images. In: Proc. of the SPIE 9409, Media Watermarking, Security, and Forensics 2015. San Francisco: SPIE, 2015. 94090V. [doi: 10.1117/12.2078399]

[20] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI). Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[21] Pevný T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography. In: Proc. of the 12th Int'l Conf. on Information Hiding. Berlin: Springer, 2010. 161–177. [doi: 10.1007/978-3-642-16435-4_13]

[22] Pevný T, Bas P, Fridrich J. Steganalysis by subtractive pixel adjacency matrix. IEEE Trans. on Information Forensics and Security, 2010, 5(2): 215–224. [doi: 10.1109/TIFS.2010.2045842]

[23] Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: Proc. of the IEEE Int'l Workshop on Information Forensics and Security. Costa Adeje: IEEE, 2012. 234–239. [doi: 10.1109/WIFS.2012.6412655]

[24] Fridrich J, Kodovský J. Multivariate Gaussian model for designing additive distortion for steganography. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 2949–2953. [doi: 10.1109/ICASSP.2013.6638198]

[25] Sedighi V, Fridrich J, Cogranne R. Content-adaptive pentary steganography using the multivariate generalized Gaussian cover model. In: Proc. of the SPIE 9409, Media Watermarking, Security, and Forensics 2015. San Francisco: SPIE, 2015. 94090H. [doi: 10.1117/12.2080272]

[26] Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability. IEEE Trans. on Information Forensics and Security, 2016, 11(2): 221–234. [doi: 10.1109/TIFS.2015.2486744]

[27] Zhou WB, Zhang WM, Yu NH. A new rule for cost reassignment in adaptive steganography. IEEE Trans. on Information Forensics and Security, 2017, 12(11): 2654–2667. [doi: 10.1109/TIFS.2017.2718480]

[28] Hu DH, Xu HY, Ma ZJ, Zheng SL, Li B. A spatial image steganography method based on nonnegative matrix factorization. IEEE Signal Processing Letters, 2018, 25(9): 1364–1368. [doi: 10.1109/LSP.2018.2856630]

[29] Qin XH, Li B, Huang JW. A new spatial steganographic scheme by modeling image residuals with multivariate Gaussian model. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019. 2617–2621. [doi: 10.1109/ICASSP.2019.8682688]

[30] Tang WX, Li B, Luo WQ, Huang JW. Clustering steganographic modification directions for color components. IEEE Signal Processing Letters, 2016, 23(2): 197–201. [doi: 10.1109/LSP.2015.2504583]

[31] Liao X, Yu YB, Li B, Li ZP, Qin Z. A new payload partition strategy in color image steganography. IEEE Trans. on Circuits and Systems for Video Technology, 2020, 30(3): 685–696. [doi: 10.1109/TCSVT.2019.2896270]

[32] Wu HB, Li FY, Zhang XP, Wu K. GAN-Based steganography with the concatenation of multiple feature maps. In: Proc. of the 18th Int'l Workshop on Digital Forensics and Watermarking. Cham: Springer, 2020. 3–17. [doi: 10.1007/978-3-030-43575-2_1]

[33] Zeng JS, Tan SQ, Liu GQ, Li B, Huang JW. WISERNet: Wider separate-then-reunion network for steganalysis of color images. IEEE Trans. on Information Forensics and Security, 2019, 14(10): 2735–2748. [doi: 10.1109/TIFS.2019.2904413]

[34] Bas P, Filler T, Pevný T. "Break our steganographic system": The ins and outs of organizing BOSS. In: Proc. of the 13th Int'l Conf. on Information Hiding. Berlin: Springer, 2011. 59–70. [doi: 10.1007/978-3-642-24178-9_5]

[35] Kodovský J, Fridrich J, Holub V. Ensemble classifiers for steganalysis of digital media. IEEE Trans. on Information Forensics and Security, 2012, 7(2): 432–444. [doi: 10.1109/TIFS.2011.2175919]

[36] Qian YL, Dong J, Wang W, Tan TN. Deep learning for steganalysis via convolutional neural networks. In: Proc. of the SPIE 9409, Media Watermarking, Security, and Forensics 2015. San Francisco: SPIE, 2015. 94090J. [doi: 10.1117/12.2083479]

[37] Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. The American Statistician, 1988, 42(1): 59–66. [doi: 10.1080/00031305.1988.10475524]

附中文参考文献:

[2] 孙曦, 张卫明, 俞能海, 魏尧. 基于空域图像变换参数扰动的隐写术. 通信学报, 2017, 38(10): 166–174. [doi: 10.11959/j.issn.1000-436x.2017155]

[3] 沈军, 廖鑫, 秦拯, 刘绪崇. 基于卷积神经网络的低嵌入率空域隐写分析. 软件学报, 2021, 32(9): 2901–2915. http://www.jos.org.cn/1000-9825/5980.htm [doi: 10.13328/j.cnki.jos.005980]

[7] 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述. 软件学报, 2019, 30(2): 440–468. http://www.jos.org.cn/1000-9825/5659.htm [doi: 10.13328/j.cnki.jos.005659]

[8] 罗会兰, 童康, 孔繁胜. 基于深度学习的视频中人体动作识别进展综述. 电子学报, 2019, 47(5): 1162–1173. [doi: 10.3969/j.issn.0372-2112.2019.05.025]

[11] 张政馗, 庞为光, 谢文静, 吕鸣松, 王义. 面向实时应用的深度学习研究综述. 软件学报, 2020, 31(9): 2654–2677. http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

[12] 王万良, 李卓蓉. 生成式对抗网络研究进展. 通信学报, 2018, 39(2): 135–148. [doi: 10.11959/j.issn.1000-436x.2018032]

[15] 翟黎明, 嘉炬, 任魏翔, 徐一波, 王丽娜. 深度学习在图像隐写术与隐写分析领域中的研究进展. 信息安全学报, 2018, 3(6): 2–12. [doi: 10.19363/J.cnki.cn10-1380/tn.2018.11.01]

廖鑫(1985－), 男, 博士, 副教授, 博士生导师, CCF 高级会员, 主要研究领域为多媒体安全, 数字取证, 密码学, 人工智能安全.

曹纭(1983－), 男, 博士, 副研究员, 主要研究领域为多媒体信息安全.

唐志强(1995－), 男, 硕士生, 主要研究领域为多媒体信息安全.