文章编号:1671-6833(2023)04-0010-06

基于对抗机制的彩色图像隐写分析算法

张 涛[1],葛育伟[2],韩 旭[2],张 昊[1],汪 然[1]

(1.战略支援部队信息工程大学 信息系统工程学院,河南 郑州 450001;2.苏州大学 计算机科学与技术学院,江苏 苏州 215006)

摘 要:针对彩色图像的隐写分析问题,引入逐通道卷积、多激活模块以及对抗机制,提出了一种应用于彩色图像隐写分析的深度卷积网络。逐通道卷积能够避免削弱不相关噪声信号,保留更多的隐写嵌入特征;多激活模块利用多种激活函数对卷积结果进行非线性映射,针对嵌入痕迹做出不同反馈,丰富嵌入特征的多样表达;对抗机制能够将内容信息特征和隐写嵌入特征从域类别上严格划分,从而分离出更多的隐写存在性特征。在 PPG-LIRMM-COLOR 数据集上针对多种隐写算法进行了检测实验。结果显示:所提算法比对照方法中性能最好的还要高 1.83%~4.99%。实验结果验证了该彩色图像隐写分析方法的有效性。

关键词:信息隐藏;隐写分析;深度学习;多激活模块;对抗机制

中图分类号:TP391;TN915.08;O235 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2023.04.013

近年来,各种安全高效的空域图像隐写算法和对应的精准通用的隐写分析算法层出不穷,大多是以灰度图像作为载体。但现实情况是彩色图像占据网络上图像的主流部分,因此以彩色图像为载体的隐写与隐写分析算法成为研究人员的关注重点。

在针对图像的隐写算法方面,基于失真函数构造[1]和校验子格编码[2]的自适应隐写成为当前隐写方法的主流。这类算法隐写位置集中在图像纹理复杂区域或边缘区域等难以建模的地方,提高了防检测性。典型的空域灰度图像隐写算法有:WOW(wavelet obtained weights)算法[3]、S-UNIWARD(spatial-universal wavelet relative distortion)算法[4]以及 HILL(high-pass,low-pass,low-pass)算法[5]。

彩色图像隐写可通过各个通道进行的独立灰度图像隐写嵌入,将隐秘消息嵌入到彩色图像中,但该方法忽略了 3 个通道之间的相关性,因此 Tang 等[6]提出一种非加性隐写算法,在保留通道内像素相关性的同时,充分利用通道间的相关性进行隐写嵌入。由于大多数彩色图像算法在生成过程中采用彩色滤光阵列解马赛克算法,该做法会在一定程度上削弱图像邻域像素相关性,因此最早的彩色图像隐写分析便利用这一特性进行隐写检测,典型的彩色图像检测算法有 CRM[7]、SGRM[8]和 GCRM[9]等。

近年来,研究人员将深度学习与图像隐写分析相结合,取得了优异的性能。Tan 等[10]利用栈式卷积自编码器进行隐写分析;Qian 等[11]将高斯激活函数引入卷积神经网络结构开发出 GNCNN(Gaussian-neuron convolutional neural network);Xu 等[12]提出包含 5 层卷积层的 Xu-Net,检测性能首次超过空域富模型方法;Ye 等[13]设计了全新的激活函数 TLU,在此基础上提出了 TLU-CNN;Zhang 等[14]提出适用于多尺度图像的隐写分析网络 Zhu-Net;Zeng 等[15]提出一种先分离后聚合的宽卷积网络 WISERNet,并取得了较好的效果;Yedroudj 等[16]结合 Xu-Net 和 Ye-Net 中的优秀成果,提出了 Yedroudj-Net。

总体来看,现有方法主要沿用基于 CNN 的图像分类框架,是一个二分类问题。但是通过对比不难发现,图像分类网络提取更多的是内容信息,难以提取能够刻画隐藏信息存在性的特征。此外,为加快网络收敛,多数方法引入高通滤波器处理输入图像,但人工设计的滤波器并非最优,会对部分载密信号起到抑制效果。另一方面,现有算法多针对灰度图

收稿日期:2023-03-01;**修订日期**:2023-04-02

基金项目:国家自然科学基金资助项目(62072057)

作者简介:张涛(1977—),男,湖北天门人,战略支援部队信息工程大学教授,博士,博士生导师,主要从事图像处理、多媒体信息安全等方向的研究,E-mail:brunda@163.com。

引用本文:张涛,葛育伟,韩旭,等. 基于对抗机制的彩色图像隐写分析算法[J]. 郑州大学学报(工学版),2023,44(4):10-15.(ZHANG T, GE Y W, HAN X, et al. Color image steganalysis algorithm based on adversarial mechanisms [J]. Journal of Zhengzhou University (Engineering Science),2023,44(4):10-15.)

像，少有能够应对多通道彩色图像隐写的分析算法。鉴于此，本文提出一种新的网络结构，通过引入逐通道卷积、多激活模块与对抗机制实现针对彩色图像的隐写分析，提高了隐写检测的精度。

1　本文方法

1.1　网络架构

提出的基于 CNN 的隐写分析网络结构如图 1 所示，命名为 MAAMNet（multiple activation modules and adversarial mechanisms）。该网络整体结构是由 1 个逐通道卷积模块、2 个多激活模块、1 个梯度反置层、多个基础卷积模块以及全连接层构成的端到端网络。输入为 256×256 像素的真彩图像，输出类别为载体图像或隐写图像。

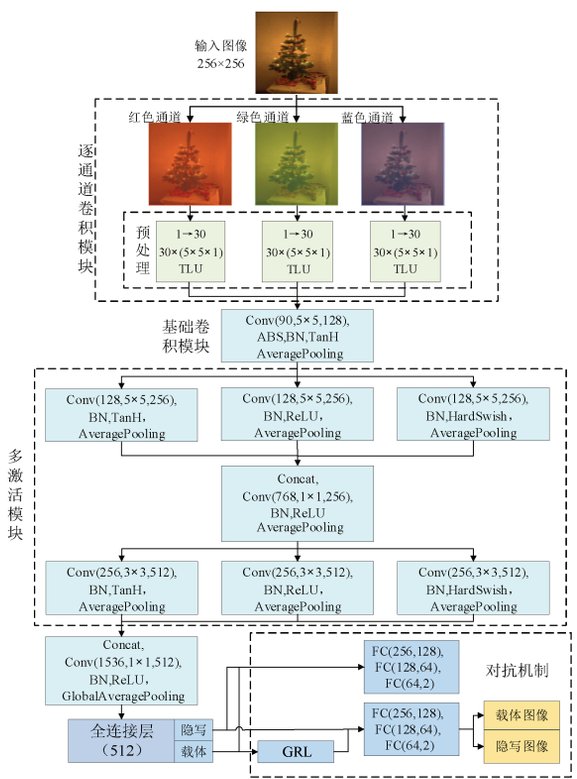

图 1　MAAMNet 的网络结构图

Figure 1　Network structure of MAAMNet

如图 1 所示，逐通道卷积模块和多激活模块的卷积层中，$C_{out}\times(K\times K\times C_{in})$ 和（C_{in}，$K\times K$，C_{out}）代表输入输出通道数分别为 C_{in} 和 C_{out}、卷积核大小为 $K\times K$ 的卷积层，BN 表示批归一化，FC 表示全连接层。整个网络的基本结构按功能划分为 3 部分：预处理、特征提取以及分类。预处理部分将彩色图像拆分成 3 个独立通道，分别运用 SRM 的 30 个基础高通滤波器（非学习权重）对红色通道、绿色通道以及蓝色通道进行逐通道卷积，计算噪声残差，并将得到的结果拼接后送入后续的层次。特征提取部分主要由多激活模块和基础卷积模块构成，多激活模块利用 TanH、ReLU 以及 Hardswish[17] 3 种激活函数同时对上层卷积输出的特征图进行非线性激活，丰富隐写嵌入特征。分类由全连接层组成。另外，对抗模块由 GRL[18] 和多层全连接层组成。下面分别对网络结构中引入的逐通道卷积、多激活模块以及对抗机制进行详细的介绍。

1.2　逐通道卷积

逐通道卷积是常规卷积的变体，能够对输入图像的每个通道都单独进行卷积。针对彩色图像，本文引入 WISERNet[17] 中的逐通道卷积，如图 2 所示。网络将引入的逐通道卷积和 SRM 滤波器组相结合，利用 SRM 中的 30 个高通滤波器对输入彩色图像的 3 个通道进行单独卷积，分别得到 30 个特征图，用作后续特征提取层次的输入。

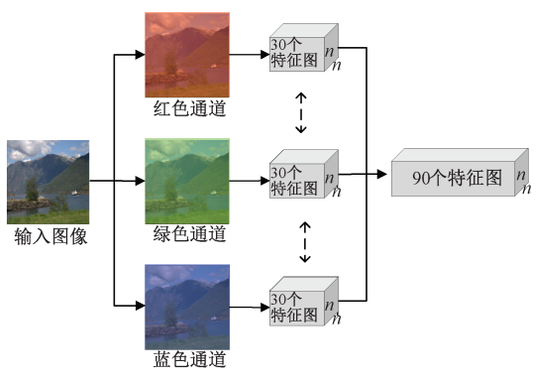

图 2　逐通道卷积结构示意图

Figure 2　Diagram of the channel-wise convolution structure

彩色图像多通道间微弱的噪声相关性为通道拆分提供了可能。CNN 中常规卷积会保留强相关的图像内容信息，削弱噪声信号，不利于增强信噪比，所以在底层网络中引入逐通道卷积取代常规卷积。对于利用加性失真框架自适应隐写算法进行隐写嵌入的彩色图像来说，其嵌入的加性隐写噪声在 3 个通道的同一位置像素表现出极其微弱的相关性，即便是考虑了彩色图像多通道之间相关性的 CMD-C 隐写算法，也仅仅只是起到了略微提升的作用，使得 3 个通道之间的隐写噪声呈现弱相关性。为尽可能降低隐写嵌入带来的失真、保障隐写安全，不可避免地导致彩色图像多通道之间的隐写噪声相关性变得很微弱，这种特性为彩色图像能够进行通道拆分、并对其进行逐一卷积提供了可能。此外，多尺度特征融合也会进一步增强特征表征能力[19]。

1.3　多激活模块

多激活模块利用多种激活函数对输入进行非线性映射，得到同一输入的不同响应。而引入激活函

数能够给神经元带来非线性因素,增加了网络的非线性拟合能力,多激活模块的结构如图 3 所示。

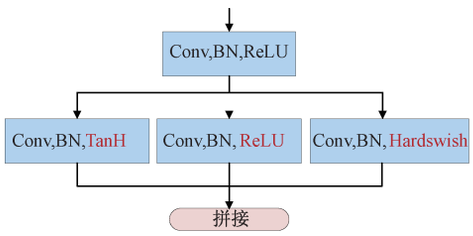

图 3 多激活模块结构示意图

Figure 3 Schematic of the structure of a multiple activation module

上层卷积层的输出被同时传递给 3 个分支层,每层的卷积、批归一化操作保持一致,区别在于使用分别为 TanH、ReLU 和 Hardswish 的不同激活函数进行激活。其中 Hardswish 激活函数即硬编码的 swish 函数,其计算方法如式(1)所示,其集合分布如图 4 所示。

$$\text{Hardswish}(x)=\begin{cases}0, x\leqslant -3;\\ x, x\geqslant 3;\\ x(x+3)/6, -3<x<3。\end{cases} \tag{1}$$

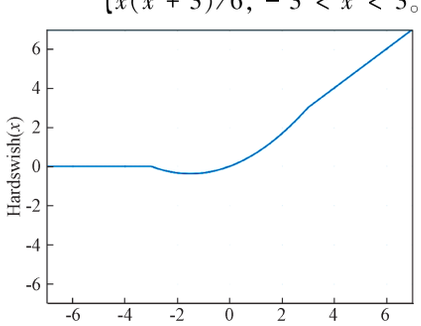

图 4 Hardswish 激活函数

Figure 4 Hardswish activation function

一般认为,更宽的 CNN 结构能够携带更多重信息[20]。而引入的多激活模块在拓宽网络的同时,增加了多种非线性映射来捕获更多隐写嵌入痕迹。

1.4 对抗机制

相较于图像内容信息,隐写嵌入噪声信号难以提取。本文借鉴迁移学习的思想,在特征提取层和标签分类层之间引入对抗训练,尽可能地抑制图像内容信息,凸显隐写信息。对抗机制主要分为 3 部分,如图 5 所示。

(1)特征提取器。该部分对输入的图像进行特征提取,并根据图像内容和隐写信息两个域将展平的特征划分为内容特征和隐写特征。正常情况下隐写特征对隐写分析是最有用的,图像内容特征对隐写检测来说是干扰信号,应当尽可能抑制。

(2)标签分类器。经特征分解得到的隐写特征

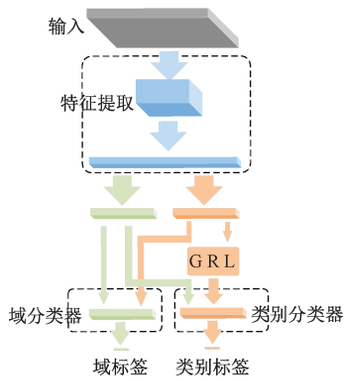

图 5 对抗机制结构示意图

Figure 5 Diagram of the adversarial mechanism structure

通过梯度反置层(gradient reversal layser, GRL),送入后续的标签分类器,完成隐写检测的分类任务。梯度反置层能够向负梯度方向优化即最大化目标函数,提取出更多的图像内容特征来误导判别器的分类,从而更有利于隐写信息存在性的检测。通过对抗训练,能分离出更多的隐写嵌入特征,提升网络的准确率。

(3)域分类器。该部分可以视作一个二分类器。将特征提取层输出的结果按照图像内容信息和隐写信息两个域进行特征分解,通过域分类器的训练,尽可能让图像内容特征和隐写特征从两个域的类别空间上区分开。

网络中引入了对抗学习利用生成器和判别器之间的对抗,抑制图像内容信息,凸现隐写嵌入信息,所以训练过程中设置的损失函数同样由隐写嵌入特征向量输入标签分类器的损失交叉熵 L_s、图像内容特征向量输入标签分类器的损失 L_c、域分类器的损失 L_d 组成,网络的训练就是在最小化该损失函数组,当取得最优值时,网络的性能达到最佳。

对于给定的训练集 $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^{N}$,$\boldsymbol{x}_i$ 表示输入量,$\boldsymbol{y}_i$ 表示输出类别向量,θ_E、θ_C、θ_D 分别表示特征提取器 $\boldsymbol{E}$,标签分类器 $\boldsymbol{C}$ 和域分类器 $\boldsymbol{D}$ 的参数。特征提取器得到的特征向量先被分解为内容向量 $\boldsymbol{c}_i$ 和隐写向量 $\boldsymbol{s}_i$,所属的域标签分别用向量 $\boldsymbol{b}_i$ 和 $\boldsymbol{q}_i$ 表示,然后 $\boldsymbol{c}_i$ 和 $\boldsymbol{s}_i$ 被输入到后续的标签分类器和域分类器中,标签分类器得到输出 $\hat{\boldsymbol{h}}_i$ 和 $\hat{\boldsymbol{y}}_i$,域分类器得到输出 $\hat{\boldsymbol{b}}_i$ 和 $\hat{\boldsymbol{q}}_i$。损失值分别如下。

(1)隐写嵌入特征向量输入标签分类器的交叉熵损失 L_s:

$$L_s(\theta_E,\theta_C)=\sum_{i=1}^{N}CE(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)。 \tag{2}$$

(2)利用 GRL 简化对抗训练,图像内容特征向量输入标签分类器的损失 L_c:

$$L_c(\theta_E,\theta_C)=\sum_{i=1}^{N}CE(\boldsymbol{y}_i,\hat{\boldsymbol{h}}_i)。\tag{3}$$

式中：$\hat{\boldsymbol{h}}_i=C(R(\boldsymbol{s}_i))$；$R$ 表示实现 GRL 功能。

(3)域分类器的损失 L_d 表示图像内容特征和隐写特征的分类损失，利用交叉熵损失计算两类特征通过域分类器的输出结果与实际类别的误差和：

$$L_d(\theta_E,\theta_D)=\frac{1}{2}\left[\sum_{i=1}^{N}CE(\boldsymbol{b}_i,\hat{\boldsymbol{b}}_i)+\sum_{i=1}^{N}CE(\boldsymbol{q}_i,\hat{\boldsymbol{q}}_i)\right]。\tag{4}$$

式中：$\hat{\boldsymbol{b}}_i=D(\boldsymbol{c}_i)$；$\hat{\boldsymbol{q}}_i=D(\boldsymbol{s}_i)$。

通过最小化该目标损失函数，能够有效地将两类特征从空间上区分开，分离出更多的隐写特征，进而提升网络的识别性能。

2　实验结果与分析

2.1　数据集与软件平台

PPG-LIRMM-COLOR[8]为实验所用彩色图像数据集。该数据集包含有 10 000 张尺寸为 512×512 像素、格式为 ppm 的彩色图像，图像分为人物、建筑、风景等多个实际生活场景下的常见类别。实验中通过下采样将该数据集中的原始图像处理为 256×256 像素，并利用 S-UNIWARD、HILL、CMD-C-SUNIWARD 以及 CMD-C-HILL 等算法进行嵌入率为 0.2 bpc 和 0.4 bpc 的隐写嵌入，后续的实验主要对这两种嵌入率进行性能对比。为了使 HILL 和 S-UNIWARD 能够完成彩色图像的隐写嵌入，实验对彩色图像的每个通道都采用相应的嵌入率进行隐写嵌入。

本文所有的实验均使用拥有 Tesla P100 显卡的 Ubuntu16.06 服务器。在实验过程中利用 MATLAB 工具进行隐写数据集的构建，网络的训练、验证和测试使用的是 PyTorch 深度学习框架。

2.2　训练、验证与测试

(1)第 1 部分实验主要考察逐通道卷积、多激活模块和对抗机制的有效性。实验使用基于 PPG-LIRMM-COLOR 数据集的 10 000 张彩色修改图像，隐写后共计 20 000 张图像，包含训练集 6 000 对图像，验证集 2 000 对图像，剩下的 2 000 对图像作为测试集，图像集合中无覆盖。

(2)第 2 部分实验是与其他彩色图像隐写分析方法进行对比实验，具体包括利用彩色图像富模型特征进行隐写分析的方法以及调整后的 Ye-Net 和 WISERNet。

2.3　实验结果

(1)为了验证网络结构中引入的逐通道卷积、多激活模块以及对抗机制的有效性，在底层网络中用常规卷积代替了逐通道卷积，表 1 展示了在底层网络中应用常规卷积和逐通道卷积模型时的准确率。

表 1　底层网络应用常规卷积和逐通道卷积的检测准确率

Table 1　Detection accuracy of conventional convolution and channel-wise convolution applied to the underlying network

嵌入率/bpc	检测准确率/%			
	常规卷积		逐通道卷积	
	S-UNIWARD	CMD-C-SUNIWARD	S-UNIWARD	CMD-C-SUNIWARD
0.2	74.36	72.58	81.43	78.94
0.4	83.66	81.07	91.35	87.76

从表 1 可以看出，在底层网络中用逐通道卷积代替常规卷积能够在不同的嵌入率下显著提升隐写检测的准确率。底层引入的常规卷积操作会对彩色图像的 3 个通道求取加权和，这被认为是一种线性共谋攻击。在求取多通道像素之间的线性组合过程中，常规卷积更多保留的是强相关性的图像内容信息，削弱了相关性较弱的隐写嵌入噪声信号，降低了信噪比，不利于隐写检测。由表 1 数据可得，在嵌入率为 0.2 bpc 和 0.4 bpc 的 S-UNIWARD 和 CMD-C-SUNIWARD 隐写算法进行检测时，准确率提升了 6.36%～7.69%，如此大的性能提升归功于引入的逐通道卷积，这表明逐通道卷积相比于常规卷积能够显著增强信噪比，充分提取隐写嵌入特征、提高检测性能。

对于多激活模块有效性的验证包含多个方面，首先将多激活模块从网络结构中移除，以此来验证多激活模块的有效性。表 2 展示了未引入多激活模块(标记为 MAAMNet/wodam)和引入多激活模块(标记为 MAAMNet/widam)的两种网络在检测使用 S-UNIWARD 和 CMD-C-SUNIWARD 隐写算法对 PPG-LIRMM-COLOR 数据集进行隐写时的准确率。

表 2　未引入和引入多激活模块的检测准确率

Table 2　Detection accuracy rates of non-introduction and introduction multiple activation modules

嵌入率/bpc	检测准确率/%			
	MAAMNet/wodam		MAAMNet/widam	
	S-UNIWARD	CMD-C-SUNIWARD	S-UNIWARD	CMD-C-SUNIWARD
0.2	77.32	74.55	81.43	78.94
0.4	86.21	83.94	91.35	87.76

从表 2 中可以看出，同样在对使用嵌入率为 0.2 bpc 和 0.4 bpc 的 S-UNIWARD 和 CMD-C-SUNIWARD 隐写算法嵌入图像进行检测时，准确率提升了 3.82%～5.14%。此部分性能的提升归功于多激活模块的引入，该模块利用多种激活函数对上层卷

积结果进行非线性映射,相较于普通卷积模块使用单一激活函数,该模块能够获取嵌入痕迹的不同信息,丰富隐写嵌入特征,从而提升网络的分类性能。上述实验结果充分验证了在网络结构中引入多激活模块的有效性。

(2)该部分实验选择对照算法时,综合考虑了传统使用富模型特征的隐写分析方法和基于深度学习的彩色图像隐写分析器,最终选择CRM、YeNet以及WISERNet作为实验中的比较对象,其中CRM方法使用常用的FLD集成分类器进行分类,而对于YeNet和WISERNet采用验证集上性能最好的模型对测试集进行评估。为了确保实验的公平,YeNet并没有引入选择信道感知的先验信息,仅仅使用普通版本的网络,而且为了使针对灰度图像的YeNet能够更好地进行彩色图像隐写分析,将YeNet的底层网络进行了相应的修改,用逐通道卷积代替了原始的常规卷积。表3展示了这部分实验的结果。

如表3所示,在针对多种隐写算法和不同嵌入率的情况时,本文方法取得了比其他方法更高的检测准确率,比另外3种方法中效果最好的WISERNet还要高1.83%~4.99%。在检测嵌入率为0.4 bpc的S-UNIWARD隐写嵌入图像时,所提方法更是取得了91.35%的检测准确率。

表3 多种彩色图像隐写分析算法的检测准确率

Table 3 Detection accuracy of various color image steganalysis algorithms

隐写算法	检测准确率/%							
	CRM		YeNet(channel-wise)		WISERNet		MAAMNet	
	0.2 bpc	0.4 bpc	0.2 bpc	0.4 bpc	0.2 bpc	0.4 bpc	0.2 bpc	0.4 bpc
S-UNIWARD	70.09	83.26	76.03	86.82	78.52	87.34	81.43	91.35
CMD-C-SUNIWARD	66.83	78.66	68.49	82.53	73.95	85.12	78.94	87.76
HILL	67.15	80.74	73.84	84.91	77.58	86.96	79.54	88.82
CMD-C-HILL	63.41	75.92	65.07	78.86	71.41	83.75	73.24	85.63

3 结论

以彩色图像为载体,提出一种新的彩色图像隐写分析方法,该方法比目前已有的彩色图像隐写分析器具有更高的检测精度。其中引入的逐通道卷积将多个通道拆分后进行逐一卷积,提高了信噪比;多激活模块通过获取卷积特征图的不同映射,捕获更多的隐写信息,丰富了嵌入特征;对抗机制能够迫使特征提取器提取更多的内容信息特征,从而隔离出更多有用的隐写嵌入特征。大量的实验表明,与现有彩色图像隐写分析算法相比,所提出的方法明显提高了检测的准确率。

未来工作将聚焦于引入强表征能力的网络模型,提取出具有强分类能力的隐写嵌入特征,进一步提高隐写检测的精度。此外,本文网络输入的是固定尺寸的图像,如何应对真实生活场景下任意尺寸的彩色图像,仍需进一步的研究。

参考文献:

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

[2] 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.

[3] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//2012 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE, 2013: 234-239.

[4] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, 2014(1): 1-13.

[5] LI B, WANG M, HUANG J W, et al. A new cost function for spatial image steganography[C]//2014 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2015: 4206-4210.

[6] 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.

[7] GOLJAN M, FRIDRICH J, COGRANNE R. Rich model for steganalysis of color images[C]//2014 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE, 2015: 185-190.

[8] ABDULRAHMAN H, CHAUMONT M, MONTESINOS P, et al. Color image steganalysis based on steerable Gaussian filters bank[C]//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Securi-

ty. New York: ACM, 2016: 109-114.

[9] ABDULRAHMAN H, CHAUMONT M, MONTESINOS P, et al. Color image steganalysis using correlations between RGB channels[C]//2015 10th International Conference on Availability, Reliability and Security. Piscataway: IEEE, 2015: 448-454.

[10] TAN S Q, LI B. Stacked convolutional auto-encoders for steganalysis of digital images[C]//Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific. Piscataway: IEEE, 2015: 1-4.

[11] QIAN Y L, DONG J, WANG W, et al. Deep learning for steganalysis via convolutional neural networks[C]//Conference on Media Watermarking, Security, and Forensics. Palos Verdes:SPIE, 2015:171-180.

[12] XU G S, WU H Z, SHI Y Q. Structural design of convolutional neural networks for steganalysis[J]. IEEE Signal Processing Letters, 2016, 23(5): 708-712.

[13] YE J, NI J Q, YI Y. Deep learning hierarchical representations for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.

[14] ZHANG R, ZHU F, LIU J Y, et al. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1138-1150.

[15] ZENG J S, TAN S Q, LIU G Q, et al. WISERNet: wider separate-then-reunion network for steganalysis of color images[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(10): 2735-2748.

[16] YEDROUDJ M, COMBY F, CHAUMONT M. Yedroudj-Net: an efficient CNN for spatial steganalysis[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 2092-2096.

[17] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 1314-1324.

[18] SU K, KUNDUR D, HATZINAKOS D. Statistical invisibility for collusion-resistant digital video watermarking[J]. IEEE Transactions on Multimedia, 2005, 7(1): 43-51.

[19] 张坚鑫, 郭四稳, 张国兰, 等. 基于多尺度特征融合的火灾检测模型[J]. 郑州大学学报(工学版), 2021, 42(5): 13-18.
ZHANG J X, GUO S W, ZHANG G L, et al. Fire detection model based on multi-scale feature fusion[J]. Journal of Zhengzhou University (Engineering Science), 2021, 42(5): 13-18.

[20] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1-9.

Color Image Steganalysis Algorithm Based on Adversarial Mechanisms

ZHANG Tao[1], GE Yuwei[2], HAN Xu[2], ZHANG Hao[1], WANG Ran[1]

(1. School of Information System Engineering, Strategic Support Force Information Engineering University, Zhengzhou 450001, China; 2. School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Aiming at the steganalysis of color images, a deep convolutional network applied to the steganalysis of color images was proposed by introducing channel-wise convolution, multiple activation module and adversarial mechanism. Channel-wise convolution could avoid weakening irrelevant noise signals and retain additional steganographic embedded features; and multiple activation modules could use various activation functions to nonlinearly map convolution results and make different feedback for embedded traces to enrich the diverse expressions of embedded features; adversarial mechanisms could divide content information features and steganographic embedding features from domain categories, thereby separating additional steganographic existence features. Experiments were carried out on the PPG-LIRMM-COLOR dataset for various steganographic algorithms. The proposed algorithm was 1.83%-4.99% higher performance than the best performance in the control methods. Results verified the effectiveness of the proposed color image steganalysis method.

Keywords: information hidding; steganalysis; deep learning; multiple activation modules; adversarial mechanisms