

彩色图像隐写分析研究进展

徐萌¹, 罗向阳¹, 王金伟², 王昊³

(1. 数学工程与先进计算国家重点实验室, 河南 郑州 450001;

2. 南京信息工程大学计算机与软件学院, 江苏 南京 210044; 3. 南京理工大学自动化学院, 江苏 南京 210014)

摘要: 传统的加密通信技术易被察觉, 很难满足安全通信的需求。数字隐写术通过对载体修改来隐藏信息, 可以实现隐蔽通信。但是, 隐写术也可能被用于非法行为, 使检测隐写术的隐写分析技术受到了越来越多的关注, 因此隐写分析技术具有重大的研究意义。深度学习在计算机视觉、模式识别和自然语言处理等领域取得了很多研究成果, 给隐写分析带来了新机遇, 提出了新挑战, 促使隐写分析产生新思想、新方法。彩色图像是互联网传输过程中的主流载体。但现有的彩色图像隐写分析特征主要依赖于人工设计, 且隐写分析特征将彩色图像当成3个独立的灰度图像, 并未充分考虑彩色图像三通道间的内部联系, 对载密图像的检测能力有待提高。深度学习在彩色图像隐写分析领域中的应用也处于初步阶段。因此在介绍隐写和隐写分析的概念、分类和研究意义, 概述其研究现状后, 重点介绍了几种针对彩色图像的隐写分析关键技术, 对其进行对比和总结, 并分析其发展趋势。

关键词: 隐写术; 隐写分析; 深度学习; 彩色图像

中图分类号: TP309.7

文献标志码: A

DOI: 10.11959/j.issn.cn2096-109x.2024052

Research progress of color image steganalysis

XU Meng¹, LUO Xiangyang¹, WANG Jinwei², WANG Hao³

1. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China;

2. Department of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China;

3. School of Automation, Nanjing University of Science and Technology, Nanjing 210014, China

Abstract: Traditional encrypted communication technologies have been easily detected and have struggled to meet the needs of secure communication. Steganography, capable of hiding information by modifying the carrier, has been utilized to realize covert communication. However, the potential for steganography to be employed in illegal acts has directed increasing attention towards steganalysis for the detection of steganography, thereby bestowing great research significance upon it. Deep learning has yielded numerous research achievements in fields such as computer vision, pattern recognition, and natural language processing, introducing new opportunities and challenges

收稿日期: 2024-01-22; 修回日期: 2024-07-28

通信作者: 罗向阳, luox_y_ieu@sina.com

基金项目: 国家自然科学基金 (U1804263; U1736214; 62172435)

Foundation Item: The National Natural Science Foundation of China (U1804263; U1736214; 62172435)

引用格式: 徐萌, 罗向阳, 王金伟, 等. 彩色图像隐写分析研究进展[J]. 网络与信息安全学报, 2024, 10(4): 49-62.

Citation Format: XU M, LUO X G, WANG J W, et al. Research progress of color image steganalysis[J]. Chinese Journal of Network and Information Security, 2024, 10(4): 49-62.

to steganalysis. These advancements have propelled the generation of new ideas and methods in steganalysis. Currently, color images constitute the mainstream carrier in the process of internet transmission. Nevertheless, existing steganalysis features for color images primarily rely on manual design and often treat the color image as three independent grayscale images, without fully considering the internal relationships between the three color channels, thus necessitating an improvement in detection capabilities for encrypted images. The application of deep learning in the field of color image steganalysis remains in its preliminary stage. The concepts, classifications, and research significance of steganography and steganalysis were introduced, along with an outline of their current research status. Several key techniques for the steganalysis of color images were introduced, compared, summarized, and their development trends were analyzed.

Keywords: steganography, steganalysis, deep learning, color image

0 引言

信息时代人们在享受信息飞快传播与分享的便利时,存在隐私被侵犯、利益受损等威胁。因此,保护隐私和传递数据的完整性已经成为信息传递过程中的重点安全问题。虽然加密能保护通信内容,但杂乱无序的加密内容却使通信行为产生突出,从而导致通信被分析和检测。因此,仅对通信内容进行加密无法有效地保证通信安全。由于隐写术(steganography)使用数字媒体(如图像、文本、视频和音频)作为载体,在载体上嵌入秘密信息的同时保证图像视觉质量,最终将其根据目的进行传递,本质上隐藏了传递秘密信息这个行为,能够克服加密行为所带来的异常,使通信更加隐蔽和安全。隐写术的发展对国家安全、军事情报以及个人隐私等都有重大意义,关于隐写术的相关研究也得到了国家和政府的大力支持与资助。隐写术的发展虽然促进了互联网的安全通信,却也给一些不法分子或非法组织提供了隐蔽途径,使个人、商业甚至国家的隐私和财产受到了严重危害。在这种情况下,隐写分析(Steganalysis)技术作为隐写术的对抗技术被提出。隐写分析技术能够监视并防止非法使用隐写术,得到了各界广泛的重视。因此,研究隐写分析技术对国家、社会以及个人都有重大的意义。

隐写术利用各种多媒体作为载体实现对秘密信息的掩护。将秘密信息嵌入载体的同时保证载体的视觉质量不被肉眼所观察到,从而实现隐蔽通信。以图像为载体的隐写技术可以根据信息嵌入域分为空间域隐写技术和变换域隐写技术。空间域隐写技术直接对原始图像的像素操作来嵌入秘密信息,而变换域隐写技术则是先将载体图像

进行变换,再根据变换得到的系数进行修改以嵌入秘密信息。在过去的研究中,隐写术在复杂性和性能上得到了较大提高,由于修改图像来嵌入秘密信息会产生代价,目前一个主要的思路就是通过降低隐写所产生的代价来优化隐写算法性能,这类隐写算法对隐写分析提出了更大的挑战。

隐写分析术利用隐写前后载体的视觉特性和统计特性会发生变化的特点,来判断载体是否被篡改。根据算法的适用范围,隐写分析技术可以分为通用隐写分析和针对特定隐写算法的专用隐写分析两种。专用隐写分析是针对某一类特定隐写技术的专用技术,这类隐写技术往往会在篡改载体时产生相同的嵌入痕迹,因此容易被捕捉。但是随着新的隐写技术不断地被提出,无法穷举所有的隐写算法,并且特定隐写算法的嵌入痕迹可能被消除,因此专用隐写分析具有一定的局限性,实用性较弱。通用隐写分析的研究成为了隐写分析领域的主要研究方向。由于人工智能的兴起,深度学习技术被运用于图像隐写分析领域。隐写分析技术可以分为传统的隐写分析和基于深度学习的隐写分析。

对比灰度图像,彩色图像在人们的日常生活、工作中的有着更加广泛的使用。在彩色图像中,不同颜色通道中的像素具有很强的相关性和相互作用。目前绝大多数隐写分析技术的实验对象是灰度图像^[1]。当分析彩色图像时,许多算法仅是简单地对彩色图像各个通道分别使用灰度图像隐写分析算法,这种做法很明显忽略了彩色图像各通道间的相关性,对载密彩色图像的检测准确率也较低。一些学者针对这个问题提出了改进的彩色图像隐写分析算法,一定程度上开发了彩

色图像内部的相关性,对载密图像检测结果的准确率得到了一定的提高。此外随着彩色图像隐写技术的发展,彩色图像隐写分析也面临着更加严峻的挑战。总体而言,现有的彩色图像隐写分析算法对彩色图像三通道间相关性的开发还不够充分,深度学习技术在彩色图像隐写分析领域的应用处在初步阶段。针对以上两个不足研究基于彩色图像的隐写分析技术具有重要意义。

1 国内外研究现状

1.1 图像隐写术研究概述

由Simmons提出的“囚徒问题”可以很好地阐述以图像为载体的隐写技术的流程^[2]。Alice和Bob是监狱中的两个囚徒,他们分别对应通信时的发送方和接收方,二人要避开守卫Eve的监视来实现秘密通信。因此,“囚徒问题”包含5个主要因素,分别是载体、秘密信息、嵌入/提取秘密信息的算法、密钥和通信信道。以图像为载体的隐写术所选用的对象,可以是JPEG格式图像的频域,也可以是RAW、BMP、PNG等图像的空域。待嵌入的信息通常转换为比特流来嵌入。嵌入算法的设计重点是尽量保持载体的感知特性、统计特性等,使通信更加安全。利用提取算法可以从载密载体中恢复秘密信息,可以看作是嵌入算法的逆过程。为了进一步提高通信的安全性,利用接收方和发送方共享的密钥来控制秘密信息的嵌入位置和顺序。可以看出,隐秘信息的嵌入和提取算法是核心,只要设计的嵌入算法足够有效,通信系统就能安全运行。基于图像的隐蔽通信框架如图1所示。

隐写算法的衡量指标通常可以分为以下几种。

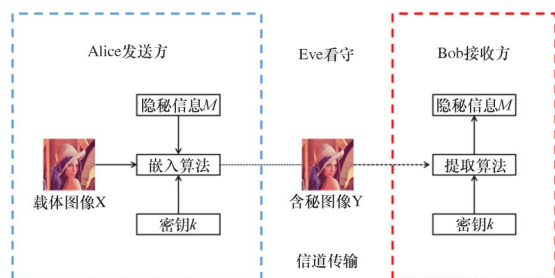


图1 基于图像的隐蔽通信框架

Figure 1 Concealed communication framework based on image

1) 不可察觉性:指嵌入算法向图像嵌入秘密信息后,图像的质量不被破坏,图像变化不易被人所感知。主要用到峰值信噪比(peak signal to noise ratio, PSNR)作为图像的质量评价标准。

2) 安全性:基于统计的隐写分析算法可以检测隐写行为,若能抵抗检测,那么算法就具有一定的安全性。通过使用各种隐写分析算法来检测隐写算法,由此衡量算法的安全性。检测率越低,表明算法越安全。

3) 容量:指隐写算法最多可以嵌入秘密信息的容量。随着嵌入容量的提高,隐写算法的前两个评价指标将会明显降低。实际嵌入过程中使用嵌入率(bits per pixel, bpp)来表示嵌入容量。bpp是单位像素所嵌的秘密信息比特数。

4) 复杂性:指嵌入和提取秘密信息时的算法复杂性。在一些实时性要求较高、数据量较大的实际应用中,该评价指标比较重要。

需要注意的是,以上几种指标往往不能同时兼顾。保证视觉质量的同时,算法的容量就不能过高。算法越复杂,安全性会提高,但是对图像的改动也有可能更大。因此需要综合考虑各个指标来评价隐写算法的优劣。

通常可以将图像隐写算法分为空域隐写算法和变换域隐写算法两种。其中,空域隐写算法直接通过修改图像的像素值来嵌入秘密信息,而变换域隐写算法先对图像进行变换,再修改变换得到的变换域系数,从而实现秘密信息的嵌入,变换域隐写算法主要是基于JPEG图像的隐写算法。

1.1.1 空域隐写算法

多媒体在获取图像时会产生随机噪声,这些随机噪声不易被人感知,也就是图像的不重要部分。图像的最低有效位(least significant bit, LSB)可以很好地隐藏秘密信息,基于最低有效位的算法具有算法简单易实现、容量大等特点,该方法受到广大学者的关注。大部分空间域隐写算法都是利用LSB来隐藏信息^[3-8],其中包括典型的扩频图像隐写算法^[9],基于像素值的分布特性的空间域隐写算法^[10-11],还有LSB替换(LSB replacement)隐写算法和LSB匹配(LSB matching)隐写算法^[12]等。

自适应隐写算法根据图像的视觉特性和统计特性,自动调节算法的参数或者嵌入方式,是自适应隐写算法和传统隐写算法的最大区别。例如收集并统计图像相邻像素间的像素值变化差异,自适应地嵌入秘密信息,从而尽可能少地破坏原始图像的统计特性。通过设计一个代价函数来控制嵌入失真最小,提高隐写算法的隐蔽性和安全性。利用人眼的视觉特性,早期自适应隐写算法根据图像像素或区域的复杂度不同,嵌入不同容量的秘密信息,以保证算法的安全性。早期的自适应隐写算法包括基于位平面复杂度分割的BPCS (bit-plane complexity segmentation) 算法^[13]及其改进算法^[14-15],以及基于像素值差分的PVD (pixel value difference) 算法^[10]及其改进算法^[16]。该类算法具有简单、高容量、不易被感知等优点,但是不能抵抗基于统计检测的隐写分析。

1.1.2 变换域隐写算法

在通信过程中往往会对加密后的图像进行滤波、有损压缩、滤波、修改格式等操作,这些操作会破坏图像的数据,使秘密信息无法被完整、准确地提取出来。因此,简单将秘密信息嵌入空间域像素值的隐写算法无法保证通信的安全。针对该不足,一些学者提出了在变换域上嵌入隐秘信息。通过对载体图像进行离散傅里叶变换 (discrete fourier transform, DFT)、离散余弦变换 (discrete cosine transform, DCT)、Mellin 傅里叶变换 (Mellin-Fourier transform) 以及离散小波变换 (discrete wavelet transform, DWT) 等,再在变换系数上嵌入秘密信息。网络中最常被使用的图像是 JPEG 图像,因此大部分变换域隐写算法是基于 JPEG 图像^[17]的。JPEG 图像的压缩是有损的,并且质量因子不同,量化表也不同,因此攻击者较难破解隐写算法。经典的基于 JPEG 图像的隐写算法包括 JSteg^[18]、JPHide&Seek^[19]、Out Guess^[1]以及基于模型的MB (model-based) 系隐写^[20]等。在此基础上,又提出了许多新的 JPEG 图像隐写算法,如 F5 算法^[21]、扰动量化隐写算法^[22]以及 YASS (yet another steganographic scheme) 隐写算法^[23]等。基于离散小波变换的隐写算法也取得了一定的成果。Cogranne 等^[24]提出了 4 种不同嵌入代价的负载分配策略,利用 DCT

变换的线性属性和多元高斯随机变量的线性变换规律,将最小化最优检测器 MiPOD (minimizing the power of optimal detector) 扩展至 JPEG 域。Taburet 等^[25]利用图像生成过程中相邻块间 DCT 系数的相关性,提出了适应于启发式代价的 JPEG 图像隐写同步修改机制。针对使用量化取整误差进行代价调制和基于邻域修改方向一致进行代价调制这两种不同的非对称代价调制方法的修改方向存在相同或冲突的前提,Boroumand 和 Fridrich^[26]设计了两种方法的组合优化策略。为了符合更完整的图像生成过程,Giboulot 等^[27]将 gamma 校正考虑到噪声模型中。Bernard 等^[28]通过多次迭代训练,并在每一轮迭代过程中选取最难检测的载密图像组成训练集,然后将该集合加入检测集并用于训练新的检测器。此外,Bernard 等^[29]用一个 Gumbel Softmax 分布的期望梯度来评估多个隐写分析检测器,再调整对抗样本的价值,提高了算法的抗检测能力。

同样地,在变换域也提出了许多自适应隐写算法。该类方法设计了一个失真度量函数来指导信息的嵌入,能够在一定程度上保持图像的统计特性,从而避开隐写分析。Fridrich 等^[30]将质量因子为 85 的图像重压缩,重压缩后图像的质量因子为 70,根据重压缩前后的不同量化步长,计算因秘密信息的嵌入而产生的失真,应用湿纸编码技术,最终完成信息的嵌入。基于文献[30]学者提出了改进算法 PQt (perturbed quantization with texture) 和 PQe (perturbed quantization with energy)^[31],改进算法利用在图像纹理区域藏信息引起的失真和能量度量引起的失真,保持图像局部统计特性和嵌入的噪声是一致的,从而使算法更加安全。BCH (bose chaudhuri hooquenghem) 算法^[32]捕捉量化 DCT 系数度量后嵌入秘密信息引起的失真,量化值越大,失真越大,设计失真函数。在 DCT 变换得到的每一个 8×8 子块中,计算量化取整系数,利用失真函数和 BCH 码来选择秘密信息的嵌入位置,使载体图像的失真最小。Kim 等^[33]改进了矩阵编码技术,在每次单独的计算中允许嵌入多于 1 bit 的信息,设计出了基于 MME (modified matrix encoding) 的算法。MME2 和 MME3 则可以在每个块中分别修

改2 bit和3 bit的数据。Huang等^[34]对MME算法进行了改进,设计出了N-MME(normalized modified matrix encoding)算法。由于对载体图像量化后的DCT系数进行取整会引起一定的误差,载体图像量化取整前的系数和经过隐写修改的系数也有一定的差异,根据这两个差异设计嵌入失真函数,使算法能够抵御多种检测,安全性较高。Luo等^[35]通过在具有丰富纹理的载体图像中自适应地添加扰动来增强图像隐写术的安全性,提出了内容自适应对抗性隐写术(content-adaptive adversarial steganography, CAAS)。Sabyasachi结合小波变换和遗传算法设计了自适应隐写算法^[36]。Yao等^[37]设计了“自适应STC-ECC(syndrome-trellis codes with error-correction code)”算法,通过在不影响鲁棒性的前提下最小化校验码的数量来增强安全性。

目前隐写术的研究目标是基于最小代价框架提出更有效的隐写算法,并能降低载体被隐写分析检测的概率,也就是对抗式隐写^[38]。同时结合深度学习技术设计更加安全的隐写算法,提高隐写效率以及鲁棒隐写也是目前隐写术的主要研究热点。

1.2 图像隐写分析术研究概述

隐写分析包含检测秘密信息存在性、确定秘密信息长度、判断隐写算法、破解秘密信息内容。目前后面3种算法的实现具有一定的困难,因此隐写分析的主要研究重点是判断载体中有无隐秘信息。基于图像的通用隐写分析框架如图2所示。对样本中的载体图像进行处理,提取图像特征用于训练,实现对载密图像的识别。

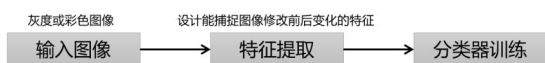


图2 基于图像的通用隐写分析框架
Figure 2 Universal steganalysis framework based on image

隐写分析算法的评价指标主要有以下几点。

1) 准确性:指算法能否正确检测出被篡改的图像,是评价隐写分析算法最重要的标准。通常使用检测错误率来评价算法的优劣,错误率越低,表明检测算法越有效。

2) 适用范围:指可以准确检测出的隐写算法范围,也就是算法的通用程度。

3) 复杂度:指算法进行隐写分析的计算复

杂度。与隐写术相对应,该指标在一些实时的场合下优先级比较高。

4) 实用性:指隐写分析算法在现实应用中的实用程度,例如一些检测的算法可能需要提供嵌入算法的类型或者嵌入率,该类算法实用性就比不需要先验信息的算法差。

图像隐写分析的发展早期提出了一些基于LSB和人工设计特征的非深度学习隐写分析算法,随着人工智能尤其是深度学习的快速发展,深度学习也被成功运用到隐写分析领域并取得了较好的成果。根据适用范围,可以将图像隐写分析技术分为通用隐写分析技术和专用隐写分析技术。其中通用隐写分析技术因其具有较好的通用性和实用性,因此成为隐写分析的主要研究方向。还可以根据载体的不同将其分为空域图像隐写分析技术和变换域图像隐写分析技术,其中变换域图像隐写分析以JPEG隐写分析为主。本文按照图像隐写分析的发展过程来介绍隐写分析算法,如图3所示。



图3 图像隐写分析的发展过程
Figure 3 The development process of image steganalysis

1.2.1 早期的隐写分析算法

早期的空域图像隐写分析算法,大多是针对特定的空域隐写算法而设计的。最早的LSB替换会导致嵌入后图像的直方图平滑化,因此Westfeld等针对这一特性,对图像像素对进行统计分析,获得了对LSB替换算法较好的检测效果^[39]。随后又提出了许多检测LSB随机嵌入的隐写分析算法,Fridrich等^[40]提出了对空域LSB隐写进行估计秘密信息长度的方法;Ker等^[41]提出了加权隐写分析算法,也能较好地估计LSB隐写的秘密信息长度。这些方法能够对早期的LSB替换隐写

算法起到不错的检测效果,同时还能在一定程度上对嵌入信息长度进行估计。另外Fridrich等^[42]提出通过基于最大似然估计的隐写分析算法,也能对LSB Matching隐写的信息长度进行估计。

1.2.2 基于低维特征的隐写分析算法

后来研究人员发现可以通过结合统计特征和机器学习分类器构建隐写分析检测器。例如Avcibas等^[43]使用图像质量指标的直方图特征进行隐写分析,这是最早使用特征提取结合分类器的方法。Harmsen等^[44]构建的直方图特性函数(histogram characteristic function, HCF)对可加性噪声建模,能够捕捉LSB Matching嵌入时对直方图的改变痕迹。除了直方图特征,研究人员也提出了其他可用于隐写分析的高阶统计特征。随后Goljan等^[45]在残差图上计算小波绝对值特征,提出了WAM(wavelet absolute moment)隐写分析算法。Pevný等^[46]通过对图像滤波后得到残差图像,构建残差图像相邻像素的概率转移矩阵得到SPAM(subtractive pixel adjacency matrix)隐写分析特征,并使用支持向量机(support vector machine, SVM)进行训练和测试。SPAM特征能比较准确地检测基于LSB Matching的隐写算法。

同样地, JPEG域隐写分析算法也是利用JPEG图像隐写算法的嵌入特点来设计隐写分析特征,根据特征对图像进行分类。例如Shi等^[47]提出SHI特征,通过马尔可夫链构建DCT系数二阶统计隐写分析特征,并通过阈值处理方法降低了特征的维数。该方法在检测Out Guess、F5和MB等隐写算法时具有较好的性能。Chen等^[48]提出可以通过块间和块内DCT系数概率转移矩阵构造隐写分析特征,提升了隐写分析性能。Fridrich^[49]首先通过计算DCT系数的边缘分布和联合分布的L1范数作为隐写分析特征,随后Pevný等^[50]将该特征与SHI特征合并构成PEV(pairwise error variance)隐写分析特征,并通过笛卡尔校准(cartesian calibration)进一步提升了SHI特征和PEV特征的隐写分析性能。在这一时期,研究人员也尝试探索可以同时检测空域和JPEG域的隐写分析算法。Taburet等^[51]对彩色图像的JPEG隐写和隐写分析进行了研究,提出调整每个通道的J-UNIWARD和UERD隐写方案的嵌入

过程,使用YCbCr色彩空间的亮度和色度通道来分散负载。Cogranne等^[52]提出了ALASKA挑战,反映了法医隐写分析所面临的限制。Yousfi等^[53]描述了为应对ALASKA隐写挑战而建立的设计结构和检测器训练。许多用于SRNets的多级检测器在3个输入通道(亮度和两个色度通道)的许多排列上进行训练。SRNet是使用跳跃连接的CNN(convolutional neural network),被广泛用于图像分类。Butora等^[54]提出了新的通用JPEG图像隐写分析方法,可以在少量载荷的情况下检测多种隐写算法。在该方案中, JPEG图像解压缩后的空间域中的摄入误差作为检测统计量,在压缩过程中对整数值信号进行离散余弦变换,就会产生攻击。然而,该技术仅限于质量因子为99和100。Wu等^[55]提出了基于去噪网络和Attention模块的JPEG隐写分析网络,该网络可以有效提取隐藏在复杂纹理区域中的隐写信号。

1.2.3 基于富模型的隐写分析算法

基于最小化失真框架的隐写算法被提出之后,早期的低维隐写分析特征已经无法对使用最小化失真框架隐写的图像达到较高的检测准确率。Fridrich等^[56]首次探讨了使用较高维数特征结合集成分类器进行隐写分析,随后又在文献[57]中指出要尽可能地增加高维特征多样性以捕捉更多的隐写痕迹,富模型的命名也来源于Fridrich等^[58]之后提出的空域富模型隐写分析算法。大多数隐写分析富模型算法使用具有数万维的特征,而支持向量机用于高维特征的有监督学习时计算复杂度高,因此大多数富模型隐写分析算法都使用计算成本更小的集成分类器进行有监督学习。集成分类器显著降低了训练复杂度,因此隐写分析研究人员可以设计更高维数的统计特征,并且在较大的数据集上进行训练,从而提高了隐写分析的检测性能。大多数空域隐写分析富模型和JPEG域隐写分析富模型具有相似的特征提取步骤。首先使用特定的滤波内核进行卷积以获得各种噪声残差图,以抑制图像内容保留隐写噪声,提高隐写分析所需隐写噪声的信噪比。然后对这些噪声残差图进行多组量化和截断操作以提高模型的多样性。最后对量化截断后的值进行高阶统计量的提取,例如通过灰度共生矩阵或局部二值模式等

方式提取特征, 并利用图像旋转不变性和对称性等特性对特征进一步的降维合并。Fridrich等^[58]提出的空域富模型 (spatial rich model, SRM) 隐写分析算法, 所有子模型特征合并得到 34 671 维 SRM 特征。Holub等^[59]代替共生矩阵, 选取根据 SRM 隐写分析算法中噪声残差图的投影直方图构建特征 PSRM (projection spatial rich model)。PSRM 实现了卓越的隐写分析性能, 并且以增加计算复杂度为代价降低了维数, 最终特征维数为 12 870。

多数 JPEG 域隐写分析富模型将 JPEG 图像解压到空域, 并通过大量滤波器提取残差图, 最后在量化截断后的残差图上提取统计特征。少数 JPEG 域隐写分析富模型还可以利用不同的子模型分别从图像的频率域和空间域进行特征提取。Kodovsky等^[60]提出的 JPEG 域富模型 (JPEG rich model, JRM) 特征, 与 SRM 特征类似, 包含多个子模型以捕捉 DCT 系数块内和块间相关性。同时利用笛卡尔校准, 提出了 CC-JRM (cartesian cCalibration-JPEG rich model) 特征。并根据提取 CDF 特征的思路, 将 CC-JRM 特征和 SRM 特征合并成 JSRM (JPEG spatial rich model) 特征, 获得了检测性能的提升。离散余弦变换残差 (discrete cosine transform residual, DCTR) 特征是由 Holub等^[59]提出的 JPEG 隐写分析 DCTR 特征, 该特征具有较低的复杂度, 以相对较小的维度获得了较好的性能。通过 64 个 8×8 的 DCT 基内核对 JPEG 图像进行卷积运算, 并统计卷积后量化截断残差图的一阶直方图组成 8 000 维的特征。Holub等^[61]又提出了相位映射模型 (phase aware projection model, PHARM) 特征, 通过相位将残差投影到向量上计算统计特征, 获得比 DCTR 特征更高的检测准确率。Song等^[62]提出了基于贾柏滤波残差的 GFR (gabor filters residual) 特征, 通过使用不同方向的贾柏滤波器卷积 JPEG 图像得到对应的残差图组, 在残差图组上统计直方图构成隐写分析 GFR 特征。通过多个方向的 Gabor 滤波进行卷积增加模型多样性, GFR 特征获得了比 DCTR 和 PHARM 等特征更高的检测准确率。Jonathan等^[60]使用富模型提取特征, 主成分分析 (principal component analysis,

PCA) 法来降维, 提出了神经空域富模型 (neural spatial rich model, NSRM)。Ma等^[64]提出了基于多维评价和动态阈值分配的隐写分析特征选择 MEDTA (multiple event-driven threshold adjustment) 方法, 该方法不仅在检测精度和特征降维方面具有很好的效果, 并且摆脱了对分类器的依赖。

1.2.4 基于深度学习的隐写分析算法

深度学习方法能够很好地将特征提取和特征分类结合在一起, 通过训练构建端到端的隐写分析器, 借助图形处理单元大量的训练数据以及优越的并行计算能力, 相比传统算法, 深度学习框架已经在诸多模式识别和机器学习问题中已取得了很大的优势, 并被成功引入隐写分析领域。

大多数基于深度学习的图像隐写分析算法采用卷积神经网络结构, 对网络结构的设计经历了层数越来越深、结构越来越复杂的过程。最早的空域图像深度学习隐写分析器是 Tan等^[65]构建的堆叠式卷积自动编码器网络, 其性能尚未达到 SRM 富模型特征的隐写分析性能。直到 Xu等^[66-67]提出的卷积神经网络 Xu-Net, 结合了高通滤波器和批量归一化结构, 首次获得了超越 SRM 富模型特征的性能。后继研究人员也开始在深度学习中结合隐写分析富模型的一些特性, 例如使用高通滤波核初始化底层卷积核和引入截断激活函数等, 引入这些结构在一定程度上提高了深度学习隐写分析算法的检测性能。Ye等^[68]将截断激活函数引入隐写分析 CNN 模型的构造中, 设计出更深的包含 10 层卷积层的网络, 使用 30 个 5×5 的卷积内核初始化第一层网络 (Ye-Net)。

对于 JPEG 隐写分析, 可以利用 JPEG 图像的 DCT 系数的相位信息来构建检测特征集, 许多研究人员也利用了这一领域知识构建 JPEG 隐写分析网络。同一时期的 ResNet (residual network) 在计算机视觉领域取得了优异的性能, 受 ResNet 中快捷连接结构 (shortcut connection) 和瓶颈结构 (bottleneck architecture) 的启发, 后续许多 JPEG 域深度学习隐写分析算法也采用类似结构取得了优异的性能。Chen等^[69]受 JPEG 压缩过程启发, 通过将网络底层输入特征图划分为 64 个并行通道, 提出了两种将相位信息与网络架构相

结合的方法,即 P-Net 和 V-Net。Xu 等^[70]受深层残差网络 ResNet 的启发,提出了具有快捷连接结构的 20 层卷积神经网络 Xu-Resnet (Xu deep residual network),并在底层使用截断操作,获得了优异的 JPEG 域隐写分析性能。Wu 等^[71]受深层残差网络 ResNet 的启发提出了另一个残差深度隐写分析网络 (Wu-Net)。Chen 等^[72]提出了多尺度深度神经网络模型,该模型基于深度残差网络,采用端到端设计。针对一些隐写分析网络内部对先验信息的利用和处理不充分的问题,Zhang 等^[73]提出了基于空间和通道多重注意力机制的策略,以提高网络的准确性。

早期的隐写分析算法,大多是针对灰度图像。随后一些基于低维特征的隐写分析算法被用于彩色图像隐写分析中,但这类算法分开看待彩色图像的 3 个通道,在每个颜色通道内利用相邻像素之间的相关性提取对应的特征,再将其合并最终提取彩色图像隐写分析特征,忽略了彩色通道间的相关性。彩色富模型设计出更全面、对隐写信号更敏感的彩色图像高维统计特征。随后一些基于深度学习的算法被应用于彩色图像隐写分析之中。

1.3 彩色图像隐写分析

目前,对于彩色图像隐写实施检测,大部分算法采用“检测特征+分类器”的框架,而这类算法设计的重点与难点主要在于如何设计并提取更全面、对隐写信号更敏感的图像统计特征。早期彩色图像隐写分析算法首先将一定比例的数据随机嵌入待检测的彩色载体图像中,然后根据秘密信息嵌入前后彩色图像相近颜色对数量发生变化的特点来检测隐写行为。还有一类方法是分开看待彩色图像的 3 个通道,在每个颜色通道内利用相邻像素之间的相关性提取对应的特征,再将其合并最终提取彩色图像隐写分析特征。Goljan 等^[74]针对彩色图像多通道的特点,对灰度图像具有良好隐写分析效果的富模型特征进行改进,首次提出了彩色图像富模型特征。该算法利用单通道图像 SRM 特征的残差计算模型,在每个颜色通道计算多尺度、多方向的残差,从中分别提取通道间的共生矩阵特征 5 404 维和各个通道的富模型特征 12 573 维,将其合并为 18 157 维的彩色图像富模型特征。这种隐写分析方法兼顾了彩色

图像通道内和通道间的特征,大大提高了彩色图像自适应隐写的检测效果。Goljan 等^[75]考虑照相机在生成彩色图像时的传感器原理,根据 Bayer CFA (bayer color filter array) 的特点对图像像素分块,再借鉴中通道间共生矩阵的提取方法,计算不同通道内的每个块中的像素残差,由这些残差得到共生矩阵,将这些矩阵合并,得到隐写分析的特征集。根据自适应隐写算法将嵌入更改集中在纹理复杂区域的特点,为隐写分析特征提取构造新的特征提取源,从而对现有特征进行改进。首先根据自适应隐写算法计算各颜色通道的纹理复杂区域,通过并集操作得到任一通道纹理较为复杂的区域,再通过交集操作得到各个通道纹理都较复杂的区域,然后在这两个区域内分别计算各通道间和各通道内残差的共生矩阵,合并两个共生矩阵得到最终的隐写分析特征,能有效检测 WOW (weighted optimization watermarking) 和 S-UNIWARD (spatial-UNiversal wAvelet relative distortion) 等自适应隐写。Yedroudj 等^[76]结合 Xu-Net 和 Ye-Net 中的优点,提出了 Yedroudj-Net。Kang 等^[77]根据彩色图像隐写将减弱不同颜色通道梯度幅值间的相关性的特点,提出了基于通道梯度相关性的彩色图像隐写分析算法。Zeng 等^[78]提出了针对彩色图像隐写分析的先分离后聚合,并且具有宽卷积层结构的深度卷积网络 (wider separate-then-reunion network, WISERNet)。Wei 等^[79]使用 62 个高通滤波器提取图像残差,设计了空域和 JPEG 域都适用的彩色图像隐写分析网络。目前彩色图像隐写分析的主要难点是如何充分开发彩色图像三通道间的相关性,结合深度学习等技术,设计出更加全面、有效的隐写分析特征。

2 经典彩色图像隐写分析算法介绍

2.1 彩色图像富模型

灰度图像空域富模型用 SPAM 来表示高通滤波后的残差图像,SRM 本质上是基于多残差的 SPAM 特征。图像的残差值可以通过高通滤波器处理后计算像素与其预测值的差值。获得残差的公式如下:

$$R_{ij} = \hat{X}_{ij}(N_{ij}) - cX_{ij} \quad (1)$$

其中, N_{ij} 为 X_{ij} 的若干领域像素, $\hat{X}_{ij}(N_{ij})$ 通过该领域预测 cX_{ij} , 一般 N_{ij} 中像素的数量和 c 相等, c 为残差阶。然后将实数残差 R_{ij} 进行量化与截断:

$$R_{ij} \leftarrow \text{trunc}_T \left(\text{round} \left(\frac{R_{ij}}{q} \right) \right) \quad (2)$$

其中, q 为量化步长, 截断函数 trunc_T 的定义是

$$\text{trunc}_T(x) = \begin{cases} x, & x \in [-T, T] \\ \text{Tsign}(x), & x \notin [-T, T] \end{cases} \quad (3)$$

对残差进行截断操作有两个原因: 一是可以减少特征的维度, 降低计算复杂度; 二是截断操作可以保留相关性较强、残差较小的区域, 该区域能更有效地提高检测准确度。

在 SRM 特征中, 计算残差图像的共生矩阵, 提取得到的共生矩阵来构造最终的隐写分析特征。特征的设计是算法的关键, 为了降低失真, 秘密信息主要集中嵌入在图像的边缘和纹理区域, 如 HUGO (highly undetectable steganography using GOffset) 算法等。针对该隐写特点, 算法设计了 39 (106) 个子模型来构造残差, 以此丰富特征的多样性, 从而提高隐写分析算法的准确度。同时为了降低特征维度防止过拟合, 算法采取符号对称和空间对称两种对称方法来获取残差, 并增加隐写分析特征的鲁棒性。降维后的子模型数目依然较多, 且每个模型都有较高的特征维数, 根据量化步长, 最终两个主流的版本的 SRM 特征维数分别是 12 753 维和 34 671 维。

彩色图像富模型 (spatial color rich model, SCRM) 在 RGB 每个色彩通道分别计算 SRM 特征, 得到各个通道的 12 573 维富模型特征, 再提取 3 个通道间的 5 404 维共生矩阵特征, 将两个特征合并最终得到 18 157 维的彩色图像富模型特征 SCRMQ1。该方法既考虑了彩色图像各通道内的相关性, 又考虑了 3 个通道间的特征, 极大地提高了彩色图像自适应隐写的检测准确率。

2.2 JPEG 图像富模型 JRM

JPEG 隐写主要是在 JPEG 系数上进行修改。JRM 特征也包含多个子模型, 通过分析每一块 JPEG 系数的相关性, 设计每个子模型的结构。因此, JRM 特征兼顾了 JPEG 系数块间和块内的相关性, 增加了特征的多样性, 能从多个方面检

测隐写对隐写系数相关性的破坏, 从而提高了 JPEG 图像隐写的检测精确度。JRM 特征主要由分块 JPEG 系数及其之间差值的 2 阶共生矩阵组成。定义如下矩阵:

$$\begin{aligned} A_{ij}^x &= |D_{ij}|, i = 1, \dots, M, j = 1, \dots, N \\ A_{ij}^{\rightarrow} &= |D_{ij}| - |D_{i+1,j}|, i = 1, \dots, M, j = 1, \dots, N-1 \\ A_{ij}^{\downarrow} &= |D_{ij}| - |D_{i,j+1}|, i = 1, \dots, M-1, j = 1, \dots, N \\ A_{ij}^{\nearrow} &= |D_{ij}| - |D_{i+1,j+1}|, i = 1, \dots, M-1, j = 1, \dots, N-1 \\ A_{ij}^{\rightarrow\rightarrow} &= |D_{ij}| - |D_{i+8,j}|, i = 1, \dots, M, j = 1, \dots, N-8 \\ A_{ij}^{\downarrow\downarrow} &= |D_{ij}| - |D_{i,j+8}|, i = 1, \dots, M-8, j = 1, \dots, N \end{aligned}$$

其中, D 为原始 JPEG 图像的系数矩阵。

在 JPEG 系数矩阵 A^* 中, $*$ $\in \{x, \rightarrow, \downarrow, \nearrow, \rightarrow\rightarrow, \downarrow\downarrow\}$ 由 (x, y) 和 $(x + \Delta x, y + \Delta y)$ 模式上的元素提取 2 阶共生矩阵 $C_T^*(x, y, \Delta x, \Delta y)$, 形成 JRM 中单个子模型, 其中, (x, y) 也是分块内坐标, 但 Δx 与 Δy 是固定的位置偏移, 其偏移范围可以至其他分块。 C_T^* 中的元素由下式计算:

$$\begin{aligned} C_{kl}^*(x, y, \Delta x, \Delta y) &= \\ \frac{1}{Z} &|\{T_{xy}^{(ij)} | T = \text{trunc}_T(A^*); \\ T_{xy}^{(ij)} &= k, T_{x+\Delta x, y+\Delta y}^{(ij)} = l\}| \end{aligned} \quad (4)$$

其中, 总情况数 Z 是归一化常数, 用于确保 $\sum_{k,l} C_{kl}^* = 1$; $\text{trunc}_T(\cdot)$ 表示截断操作:

$$\text{trunc}_T(x) = \begin{cases} T \cdot \text{sign}(x), & |x| > T \\ x, & \text{其他} \end{cases} \quad (5)$$

JRM 特征的总维数是 $8\ 635 + 2\ 620 = 11\ 255$, 通过类似笛卡尔校准等改造可以进一步提升 JRM 特征的分析效果。

2.3 基于通道梯度相关性的彩色图像隐写分析算法

基于通道梯度相关性的彩色图像隐写分析算法利用 RGB 三通道的梯度值来表现各个通道内的纹理变化程度, 根据信息嵌入前后纹理发生变化的特点, 设计隐写特征来检测载密图像。具体的特征提取框架如图 4 所示。

图像的纹理变化特征可以用梯度向量来表示。图像相邻像素间的差分可以放大隐写痕迹, 忽略图像本身内容, 利用图像水平和垂直方向的纹理变化作为隐写分析特征可以有效检测隐写。

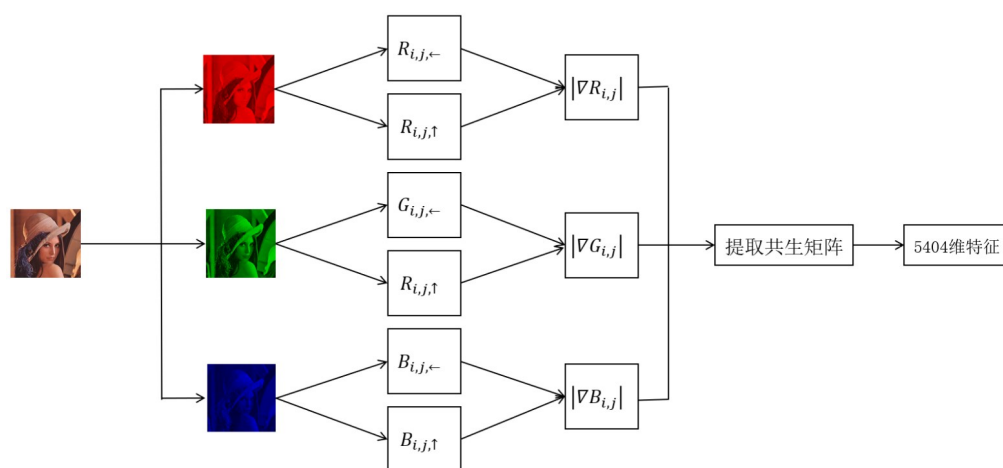


图4 CGC特征提取框架

Figure 4 Architecture of CGC feature

彩色图像各通道间的纹理特征及其变化程度的相似性，向彩色图像进行隐写时，秘密信息通常根据自适应算法被分为3段分别嵌入3个通道中。这种嵌入操作具有随机性，因此RGB这3个通道的梯度幅值的相关性极易被改变。计算各个通道的梯度幅值的残差并进行取整、量化、截断操作，由得到的残差图像计算共生矩阵，最终得到5404维彩色图像隐写分析特征。

基于通道梯度相关性的彩色图像隐写分析算法一定程度上考虑了RGB这3个通道间的梯度幅值的相关性，提高了对载密图像的检测准确率。但该算法只使用了梯度向量这一个指标，具有一定的局限性，检测隐写行为的性能只有小幅度提升。

2.4 WISERNet

先分离后聚合且具有宽结构的深度卷积网络 WISERNet (wider separate-then-reunion network) 的网络输入是真彩图像，将真彩图像拆分为RGB这3个通道，并分别对其进行逐通道卷积。在底层逐通道卷积结构中使用富模型的卷积核，每个通道中的30个卷积内核都是SRM中使用的。RGB这3个通道都分别与30个 5×5 卷积内核进行卷积，得到30个对应的特征图并输出到下一层。在具体训练过程中，根据梯度下降调整卷积内核的权重，以对网络进行优化。WISERNet网络的分离阶段即底层逐通道卷积阶段。经过卷积操作后结合3个独立的输出特征图，组合得到90个特征图组，并输入网络的第二个卷积层中。

WISERNet网络的聚合阶段从第二个卷积层往上，实质上是具有大量内核卷积层且层数相对较浅、较宽的卷积神经网络。控制第二阶段的深度不变，上层结构的级联卷积层数设为3。为了增加网络的宽度，利用一个模型放大系数 n 来放大每个卷积层中的内核数目。 n 越多，则WISERNet网络越宽。卷积层的顶部是4层全连接层，通过这些连接层最终得到隐写分析预测结果。这4层全连接层分别包含800、400、200和2个神经元。WISERNet的整体框架如图5所示。

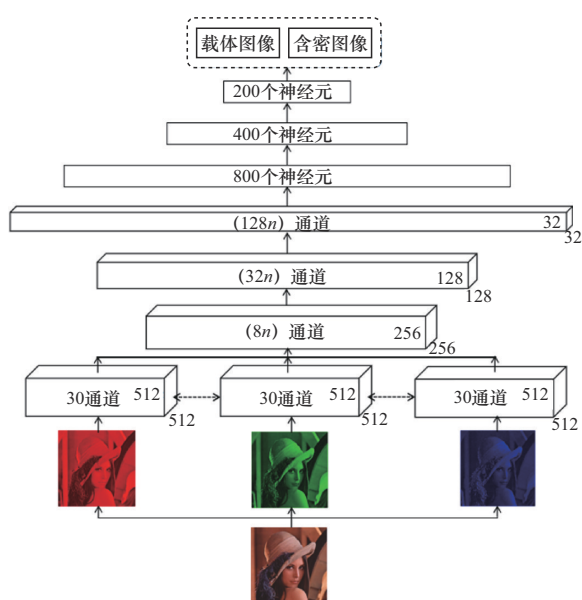


图5 WISERNet的整体框架

Figure 5 Architecture of WISERNet

常规卷积结构中的加权和操作能够保留图像较强的相关模式并削弱不相关的噪声,而隐写分析需要放大由隐写引发的噪声,因此WISERNet提出逐通道卷积结构,提高隐写分析的性能。通过理论分析和实验证明网络的有效性。

3 彩色图像隐写分析算法的比较

选取BOSSBase图像库中的10 000张彩色图像进行对比实验。选用两种彩色图像隐写算法:HILL和S-UNIWARD。验中隐写算法的嵌入率分别设置为0.1 bpc (bit per channel)、0.2 bpc、0.3 bpc、0.4 bpc和0.5 bpc。对比算法选取彩色图像隐写分析富模型CRM^[58]、Ye-Net^[68]和WISERNet^[78] 3种算法。实验中对基于富模型的隐写分析器,使用具有默认设置的FLD集成分类器。对基于深度学习的隐写分析器,采用验证集来验证训练好的模型,并采用在验证集上获得最好性能的训练参数作为测试时使用的模型。HILL和S-UNIWARD隐写分析算法的性能比较结果如图6和图7所示。

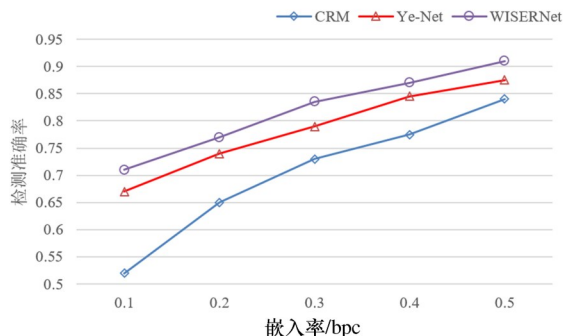


图6 HILL隐写分析算法的性能比较

Figure 6 Performance comparison of HILL steganalysis algorithms

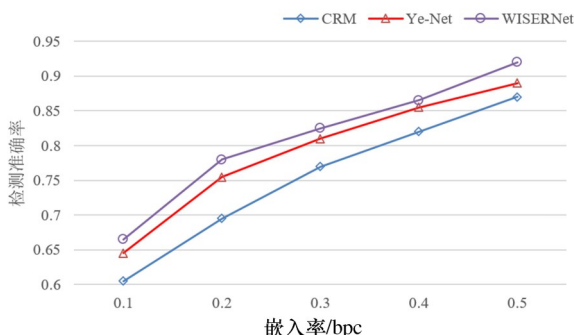


图7 S-UNIWARD分析隐写算法的性能比较

Figure 7 Performance comparison for S-UNIWARD steganalysis algorithms

与基于富模型的隐写分析算法CRM相比,另外两种基于深度学习的隐写分析模型都可以显著提高对载密图像的检测能力。WISERNet模型相对Ye-Net深度学习模型层数较浅,但其彩色图像隐写分析性能更优。同时WISERNet的参数数量比Ye-Net更少,计算复杂度更低。需要注意的是,虽然基于深度学习的隐写分析算法比基于人工设计特征的隐写分析算法效果更好,但是基于深度学习的隐写分析很容易被对抗性示例欺骗^[80],因此将人工设计的特征与深度学习相结合在理论上能提高载密彩色图像的检测效率。

4 结束语

本文介绍了目前隐写术和隐写分析术(包括彩色图像隐写分析)的研究进展,结合深度学习技术设计更加安全的隐写算法,提高隐写效率。今后隐写分析技术的主要研究重点应该是在更通用的图像数据库上设计通用隐写分析算法。

在隐写分析中所做的工作主要集中在灰度图像上,本文列出了多年来彩色图像隐写分析的突破。目前提出的大多数彩色图像隐写分析特征主要依赖于人工设计,且隐写特征并未充分考虑彩色图像三通道间的内部联系。深度学习技术在图像隐写分析领域的应用还处于初步阶段,需要对其进一步研究。近些年来,针对彩色图像的先进隐写算法的不断提出也给彩色图像隐写分析带来了更加严峻的挑战。针对以上问题,彩色图像隐写分析的一个思路是设计出更多样且更有效的隐写分析特征,另一个思路是将深度学习应用于彩色图像隐写分析。而基于深度学习的隐写分析很容易被对抗性示例欺骗,因此将人工设计的特征与深度学习技术相结合,理论上能进一步提高检测效率,这也是一种有价值的研究方向。

参考文献:

- [1] PROVOS N. Defending against statistical steganalysis[C]//Proceedings of 10th USENIX Security Symposium. Washington: ACM, 2001: 323-336.
- [2] SIMMONS G J. The prisoners' problem and the subliminal channel[M]//Advances in Cryptology. Boston, MA: Springer, 1984: 51-67.
- [3] KAWAGUCHI E, EASON R O. Principles and applications of BPCS steganography[C]//Proceedings of the SPIE, Multimedia

- Systems and Applications. SPIE, 1999: 464-73.
- [4] NGUYEN B C, YOON S M, LEE H K. Multi bit plane image steganography[M]//Digital Watermarking. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 61-70.
 - [5] LUO W Q, HUANG F J, HUANG J W. Edge adaptive image steganography based on LSB matching revisited[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 201-214.
 - [6] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[C]//Proceedings of International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2010: 161-177.
 - [7] OMOOMI M, SAMAVI S, DUMITRESCU S. An efficient high payload ± 1 data embedding scheme[J]. Multimedia Tools and Applications, 2011, 54(2): 201-218.
 - [8] 师夏阳, 马赛兰, 胡永健, 等. 一种基于像素块的纹理优先自适应隐写算法[J]. 电子学报, 2015, 43(6): 1094-1100.
SHI X Y, MA S L, HU Y J, et al. A pixel block-based adaptive steganographic algorithm with embedding priority given to image textures[J]. Acta Electronica Sinica, 2015, 43(6): 1094-1100.
 - [9] MARVEL L M, BONCELET C R, RETTER C T. Spread spectrum image steganography[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 1999, 8(8): 1075-1083.
 - [10] WU D C, TSAI W H. A steganographic method for images by pixel-value differencing[J]. Pattern Recognition Letters, 2003, 24(9/10): 1613-1626.
 - [11] WANG C M, WU N N, TSAI C S, et al. A high quality steganographic method with pixel-value differencing and modulus function[J]. Journal of Systems and Software, 2008, 81(1): 150-158.
 - [12] KER A D. Improved detection of LSB steganography in gray-scale images[C]//Proceedings of International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2004: 97-115.
 - [13] KAWAGUCHI E, EASON R O, TESCHER A G, et al. Principle and applications of BPCS steganography[J]. Multimedia Systems and Applications, 1999, 3528: 464-473.
 - [14] HIOKI H. A data embedding method using BPCS principle with new complexity measures[C]//Proceedings of Pacific Rim Workshop on Digital Steganography. Kitakyushu, Japan: Pacific Rim Workshop Digital Steganography 2002.
 - [15] NODA H, SPAULDING J, SHIRAZI M N, et al. Application of bit-plane decomposition steganography to JPEG2000 encoded images[J]. IEEE Signal Processing Letters, 2002, 9(12): 410-413.
 - [16] PARK Y R, KANG H H, SHIN S U, et al. An image steganography using pixel characteristics[M]//Computational Intelligence and Security. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 581-588.
 - [17] 李星. JPEG 图像及解压图像中的隐写分析技术研究[D]. 郑州: 解放军信息工程大学, 2012.
LI X. Research on steganalysis for JPEG images and decompressed images[D]. Zhengzhou: PLA Information Engineering University, 2012.
 - [18] KODOVSKÝ J, FRIDRICH J. Quantitative structural steganalysis of Jsteg[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(4): 681-693.
 - [19] PETITCOLAS F, ANDERSON R J, KUHN M. Information hiding-a survey[J]. Proc IEEE, 1999, 87: 1062-1078.
 - [20] PHIL S. Model-based steganography[C]//Proceedings of International Workshop on Digital Watermarking. Berlin, Heidelberg: Springer, 2004: 154-167.
 - [21] FRIDRICH J, PEVNÝ T, KODOVSKÝ J. Statistically undetectable jpeg steganography: dead ends challenges, and opportunities[C]// Proceedings of the 9th workshop on Multimedia & Ssecurity - MM&Sec '07. New York: ACM, 2007: 3-14.
 - [22] FRIDRICH J, GOLJAN M, SOUKAL D. Perturbed quantization steganography[J]. Multimedia Systems, 2005, 11(2): 98-107.
 - [23] SOLANKI K, SARKAR A, MANJUNATH B S. YASS: yet another steganographic scheme that resists blind steganalysis[C]//Proceedings of International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2007: 16-31.
 - [24] COGRANNE R, GIBOULOT Q, BAS P. Steganography by minimizing statistical detectability: the cases of JPEG and color images[C]// Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2020: 161-167.
 - [25] 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: ACM, 2020: 139-149.
 - [26] BOROUMAND M, FRIDRICH J. Synchronizing embedding changes in side-informed steganography[J]. Electronic Imaging, 2020, 32(4): 290-1-290-12.
 - [27] GIBOULOT Q, COGRANNE R, BAS P. Detectability-based JPEG steganography modeling the processing pipeline: the noise-content trade-off[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 2202-2217.
 - [28] BERNARD S, BAS P, KLEIN J, et al. Explicit optimization of min max steganographic game[J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 812-823.
 - [29] BERNARD S, BAS P, PEVNÝ T, et al. Optimizing additive approximations of non-additive distortion functions[C]//Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2021: 105-112.
 - [30] FRIDRICH J, GOLJAN M, SOUKAL D. Perturbed quantization steganography[J]. Multimedia Systems, 2005, 11(2): 98-107.
 - [31] FRIDRICH J, PEVNÝ T, KODOVSKÝ J. Statistically undetectable JPEG steganography: dead ends, challenges, and opportunities[C]// Proceedings of ACM Workshop on Multimedia and Security. Dallas, Texas, USA: ACM, 2007:3-14.
 - [32] SACHNEV V, KIM H J, ZHANG R Y. Less detectable JPEG steganography method based on heuristic optimization and

- BCH syndrome coding[C]//Proceedings of the 11th ACM workshop on Multimedia and security. New York: ACM, 2009: 131-140.
- [33] KIM Y, DURIC Z, RICHARDS D. Modified matrix encoding technique for minimal distortion steganography[C]//Proceedings of International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2007: 314-327.
- [34] HUANG F J, HUANG J W, SHI Y Q. New channel selection rule for JPEG steganography[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(4): 1181-1191.
- [35] LUO J, HE P S, LIU J Y, et al. Improving security for image steganography using content-adaptive adversarial perturbations[J]. Applied Intelligence, 2023, 53(12): 16059-16076.
- [36] PRAMANIK S. An adaptive image steganography approach depending on integer wavelet transform and genetic algorithm[J]. Multimedia Tools and Applications, 2023, 82(22): 34287-34319.
- [37] YAO Y, HUANG L C, WANG H, et al. Robust adaptive steganography based on adaptive STC-ECC[J]. IEEE Transactions on Multimedia, 2024, 26: 5477-5489.
- [38] 张卫明, 王宏霞, 李斌, 等. 多媒体隐写研究进展[J]. 中国图象图形学报, 2022, 27(6): 1918-1943.
- ZHANG W M, WANG H X, LI B, et al. Overview of steganography on multimedia[J]. Journal of Image and Graphics, 2022, 27(6): 1918-1943.
- [39] WESTFELD A, PFITZMANN A. Attacks on steganographic systems[C]//Proceedings of the Third International Workshop on Information Hiding. New York: ACM, 1999: 61-76.
- [40] FRIDRICH J, GOLJAN M. On estimation of secret message length in LSB steganography in spatial domain[C]//Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI. SPIE, 2004: 23-34.
- [41] KER A D, BÖHME R. Revisiting weighted stego-image steganalysis[C]//Proceedings of the SPIE Security, Forensics, Steganography, and Watermarking of Multimedia Contents X. SPIE, 2008: 5-17.
- [42] FRIDRICH J, SOUKAL D, GOLJAN M. Maximum likelihood estimation of length of secret message embedded using $\pm k$ steganography in spatial domain[C]//Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VII. SPIE, 2005: 595-606.
- [43] AVCIBAS I, MEMON N, SANKUR B. Steganalysis using image quality metrics[J]. IEEE Transactions on Image Processing, 2003, 12(2): 221-229.
- [44] HARMSSEN J H, PEARLMAN W. Steganalysis of additive-noise modelable information hiding[J]. SPIE Proceedings, 2003, 5020: 131-142.
- [45] GOLJAN M, FRIDRICH J, HOLOTYAK T. New blind steganalysis and its implications[C]//Proceedings of the SPIE Security, Steganography, and Watermarking of Multimedia Contents VIII. SPIE, 2006: 1-13.
- [46] 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.
- [47] SHI Y Q, CHEN C H, CHEN W. A Markov process based approach to effective attacking JPEG steganography[C]//Proceedings of International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2007: 249-264.
- [48] CHEN C H, SHI Y Q. JPEG image steganalysis utilizing both intrablock and interblock correlations[C]//Proceedings of the 2008 IEEE International Symposium on Circuits and Systems (ISCAS). Piscataway: IEEE Press, 2008: 3029-3032.
- [49] FRIDRICH J. Feature-Based steganalysis for JPEG images and its implications for future design of steganographic schemes[C]//Proceedings of the 6th International Conference on Information Hiding. New York: ACM, 2004: 67-81.
- [50] PEVNY T, FRIDRICH J. Merging Markov and DCT features for multi-class JPEG steganalysis[C]//Proceedings of the SPIE Proceedings, "Security, Steganography, and Watermarking of Multimedia Contents IX. SPIE, 2007: 67-81.
- [51] TABURET T, FILSTROFF L, BAS P, et al. An empirical study of steganography and steganalysis of color images in the JPEG domain[C]//Proceedings of International Workshop on Digital Watermarking. Cham: Springer, 2019: 290-303.
- [52] COGRANNE R, GIBOULOT Q, BAS P. The ALASKA steganalysis challenge: a first step towards steganalysis[C]//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2019: 125-137.
- [53] YOUSFI Y, BUTORA J, FRIDRICH J, et al. Breaking ALASKA: color separation for steganalysis in JPEG domain[C]//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2019: 138-149.
- [54] BUTORA J, FRIDRICH J. Reverse JPEG compatibility attack[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1444-1454.
- [55] WU T, REN W X, LI D W, et al. JPEG steganalysis based on denoising network and attention module[J]. International Journal of Intelligent Systems, 2022, 37(8): 5011-5030.
- [56] FRIDRICH J, KODOVSKÝ J, HOLUB V, et al. Breaking HUGO - the process discovery[C]//Proceedings of International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2011: 85-101.
- [57] FRIDRICH J, KODOVSKÝ J, HOLUB V, et al. Steganalysis of content-adaptive steganography in spatial domain[C]//Proceedings of International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2011: 102-117.
- [58] FRIDRICH J, KODOVSKÝ J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
- [59] HOLUB V. Content adaptive steganography - design and detection[D]. State University of New York at Binghamton, Thomas J. Watson School of Engineering and Applied Science, Department of Electrical and Computer Engineering. Binghamton: Binghamton University, 2014.
- [60] KODOVSKÝ J, FRIDRICH J. Steganalysis of JPEG images

- using rich models[C]//Proceedings of Media Watermarking, Security, and Forensics 2012. SPIE, 2012, 8303: 81-93.
- [61] HOLUB V, FRIDRICH J. Phase-aware projection model for steganalysis of JPEG images[C]//Proceedings of the SPIE Media Watermarking, Security, and Forensics 2015. SPIE, 2015: 94090T.
- [62] 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. New York: ACM, 2015: 15-23.
- [63] LWOWSKI J, CORLEY I, HOFFMAN J. Neural steganalysis with spatial rich models for image steganography detection [EB]. TechRxiv, 2020.
- [64] MA Y Y, XU L G, ZHANG Y, et al. Steganalysis feature selection with multidimensional evaluation and dynamic threshold allocation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(3): 1954-1969.
- [65] TAN S Q, LI B. Stacked convolutional auto-encoders for steganalysis of digital images[C]//Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific. Piscataway: IEEE Press, 2014: 1-4.
- [66] XU G S, WU H Z, SHI Y Q. Ensemble of CNNs for steganalysis: an empirical study[C]//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2016: 103-107.
- [67] 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.
- [68] 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.
- [69] CHEN M, SEDIGHI V, BOROUMAND M, et al. JPEG-phase-aware convolutional neural network for steganalysis of JPEG images[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2017: 75-84.
- [70] XU G S. Deep convolutional neural network to detect J-UNIWARD[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2017: 67-73.
- [71] CHEN H, HAN Q, LI Q, et al. Image steganalysis with multi-scale residual network[J]. Multimedia Tools and Applications, 2023, 82(14): 22009-22031.
- [72] ZHANG X B, ZHANG X P, FENG G R. Image steganalysis network based on dual-attention mechanism[J]. IEEE Signal Processing Letters, 2023, 30: 1287-1291.
- [73] WU S T, ZHONG S H, LIU Y. Deep residual learning for image steganalysis[J]. Multimedia Tools and Applications, 2018, 77(9): 10437-10453.
- [74] GOLJAN M, FRIDRICH J, COGRANNE R. Rich model for Steganalysis of color images[C]//Proceedings of the 2014 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE Press, 2014: 185-190.
- [75] GOLJAN M, FRIDRICH J. CFA-aware features for steganalysis of color images[C]//Proceedings of the SPIE Media Watermarking, Security, and Forensics 2015. SPIE, 2015: 01-13.
- [76] YEDROUDJ M, COMBY F, CHAUMONT M. Yedroudj-net: an efficient CNN for spatial steganalysis[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 2092-2096.
- [77] KANG Y H, LIU F L, YANG C F, et al. Color image steganalysis based on channel gradient correlation[J]. International Journal of Distributed Sensor Networks, 2019, 15(5): 155014771985203.
- [78] 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.
- [79] WEI K K, LUO W Q, TAN S Q, et al. Universal deep network for steganalysis of color image based on channel representation [J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 3022-3036.
- [80] QIN C, CHEN Y F, CHEN K J, et al. Feature fusion based adversarial example detection against second-round adversarial attacks[J]. IEEE Transactions on Artificial Intelligence, 2023, 4(5): 1029-1040.

[作者简介]



徐萌 (1993-), 女, 江苏泰州人, 数学工程与先进计算国家重点实验室博士生, 主要研究方向为信息安全。



罗向阳 (1978-), 男, 河南郑州人, 数学工程与先进计算国家重点实验室教授、博士生导师, 主要研究方向为图像隐写和图像隐写分析。



王金伟 (1978-), 男, 内蒙古呼伦贝尔人, 南京信息工程大学教授、博士生导师, 主要研究方向为多媒体版权保护、多媒体取证、多媒体加密和数据认证。



王昊 (1995-), 男, 江苏连云港人, 南京理工大学博士生, 主要研究方向为信息安全、多媒体取证。