基于深度学习的图像隐写分析综述 阅读

原创

躲躲世道 于 2021-03-03 16:23:53 发布 2357 收藏 34

分类专栏： 隐写

本文链接：https://blog.csdn.net/ByYastal/article/details/114299858

隐写 专栏收录该内容

1 篇文章 0 订阅

订阅专栏

背景

隐写术英文为Steganography。

现有的通信安全保障主要分为加密和信息隐藏:加密主要对秘密信息本身进行操作,但经过特殊处理后的明文更加容易受到第三方的怀疑;而信息隐藏则隐藏秘密数据的存在性,使秘密数据在不引起第三方的怀疑下进行隐蔽通信。

囚徒模型中,可以很好地阐述隐写术中各方的角色:Alice和Bob是监狱中不同牢房的犯人,他们之间的通信需要在狱警Eve的监视下完成;同时,Eve能够看见他们的通信内容.为了降低狱警Eve防范心的同时完成通信,隐写术孕育而生.Alice将想要传达的秘密信息进行隐写操作隐藏在载体当中,Bob则需要将含密载体中的秘密信息进行提取,狱警Eve时刻监视Alice和Bob的通信,一旦发现任何可疑信息就断绝双方通信。

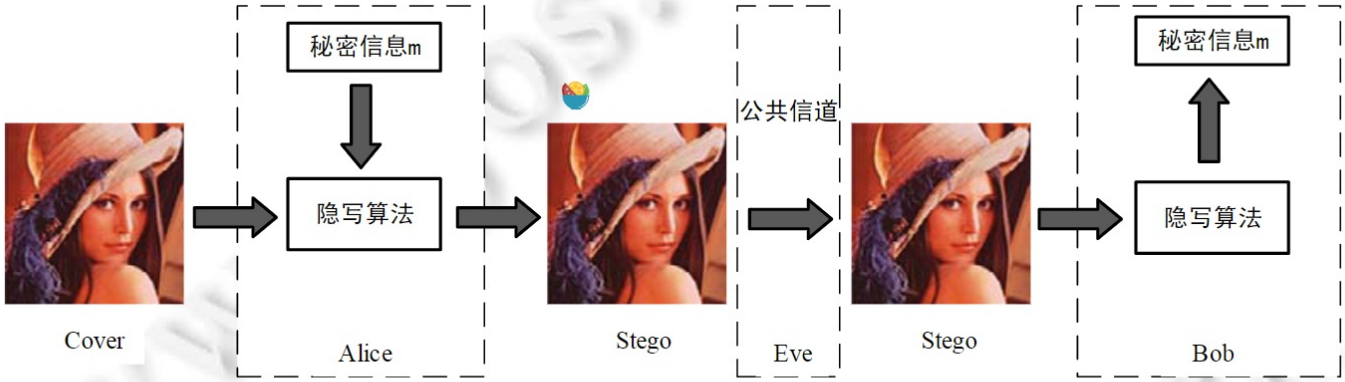
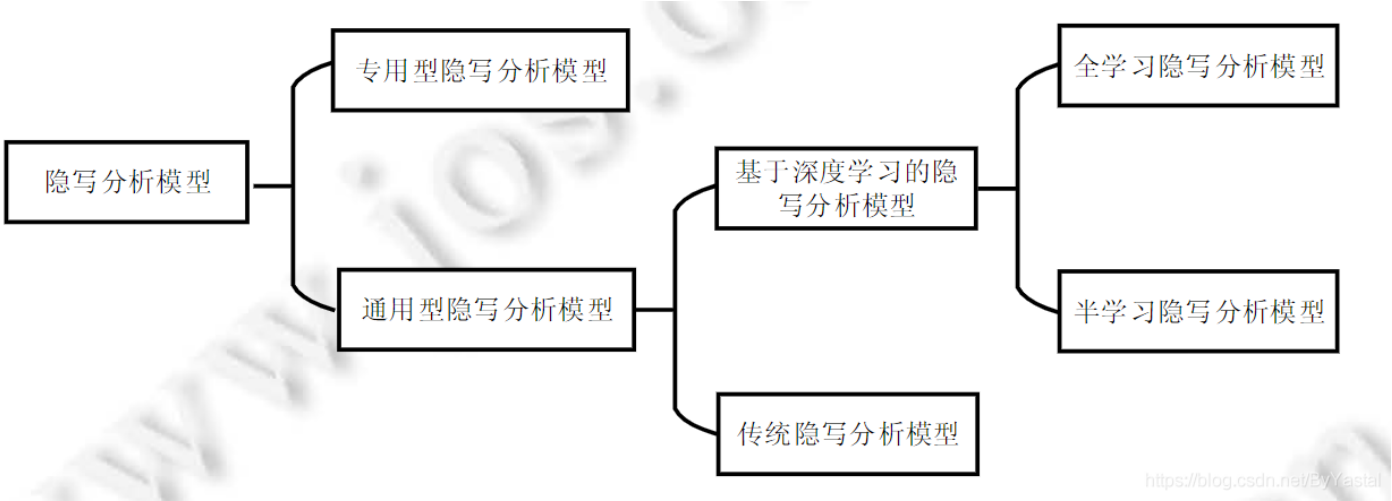


Fig.1 General process of steganography

图 1 隐写的一般过程

隐写分析模型分为专用型隐写分析模型与通用型隐写分析模型，由于专用型隐写分析模型仅针对特定的隐写算法且对于不匹配的或者未知的隐写算法检测效果较差,随着各式各类的自适应隐写算法的不断涌现,专用型隐写分析模型显得力不从心,也逐渐退出历史舞台,通用型隐写分析模型也逐渐成为主流隐写分析模型。



相关技术

LSB（least significant bit)

作为早期的隐写方法,是一种基于图片最低有效位修改并储存信息的隐写方法.利用人眼对于色彩差异的不敏感性,将秘密信息通过一定的嵌入方法放入图片的最低有效位,从而将我们所需要隐藏的信息通过一定方法放入图片的最低有效位上.。
LSB还有一种变化形式LSB匹配(LSB matching,简称LSBM)[20].二者之间的差距在于:LSB对于最低有效位进行的是替换操作;LSBM采用的则是随机±1原则,采用三元伴随式矩阵编码(syndrome-trellis codes,简称STC)[21]嵌入秘密信息.用LSB算法的图像格式需为位图形式,即图像不能经过压缩,所以LSB算法多应用于png,bmp等空域图像中.图3是LSB类隐写流程图,可以看到,载体图像Lena(戴帽子的女人)在隐写前后并不存在明显的差距.

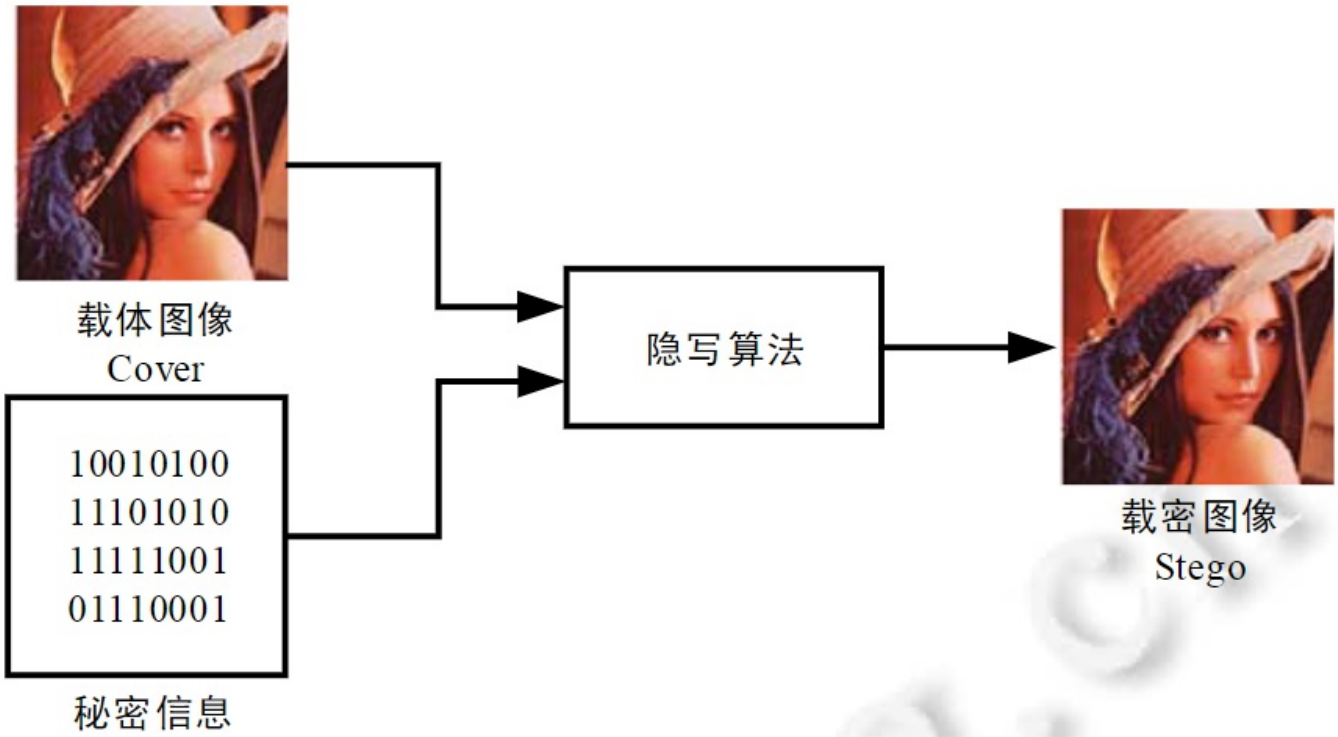


Fig.3 LSB steganography process

图 3 LSB 隐写过程

参考论文：

- Chan CK, Cheng LM. Hiding data in images by simple LSB substitution. Pattern Recognition, 2004,37(3):469−474.
- Mielikainen J. LSB matching revisited. IEEE Signal Processing Letters, 2006,13(5):285−287.
- 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.

LSB还是LSBM,都是一种非自适应的隐写算法.非自适应隐写术的思想是:对载体图像中像素内容修改地越少,隐写算法抗隐写分析能力就越强.非自适应隐写术通常与纠错编码(隐写码)相结合来实现具体的嵌入过程。

自适应隐写术

自适应隐写术则考虑载体图像的自身属性,例如图片内容的纹理信息、边缘信息,根据图像纹理复杂区域难于检测的特点,有选择地将秘密信息嵌入到载体纹理复杂或者边缘丰富的区域,提高了载密图像的抗隐写分析检测能力。常见的自适应隐写算法有HUGO[25]、WOW[26]、UNIWARD[27]、HILL[28]等,各类自适应隐写算法都与STC[21]编码方法结合,差异在于失真函数的不同。

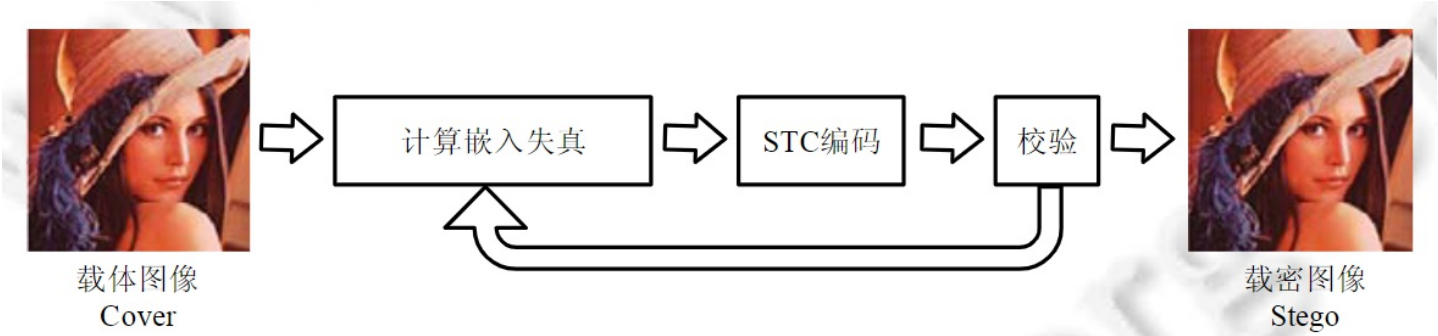


Fig.4 Adaptive steganography process

图 4 自适应隐写过程

参考论文：

- Pevný T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography. In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 2010. 161−177.
- Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: Proc. of the 2012 IEEE Int'l Workshop on Information Forensics and Security (WIFS). IEEE, 2012. 234−239.
- Holub V, Fridrich J. Digital image steganography using universal distortion. In: Proc. of the 1st ACM Workshop on Information Hiding and Multimedia Security. 2013. 59−68.
- Li B, Tan S, Wang M, Huang J. Investigation on cost assignment in spatial image steganography. IEEE Trans. on Information Forensics and Security, 2014,9(8):1264−1277.

不仅在空域上存在隐写算法,在频域即JPEG域上的隐写方法也很多,其中较早且具有代表性的是Jsteg[29]、F5[22]、J-UNIWARD[30]、UED[31]、UERD[32].根据图像经过离散余弦变换之后得到63个AC系数和1个DC系数,由于DCT分块后得到的结构信息存储在DC系数中,所以对于中频的AC系数的修改并不会引起结构上可见的变化,这样修改AC系数既可以达到隐写的目的,也不会对图像结构造成太大的破坏,保证了隐写的隐蔽性.

参考论文：

- [29] Zhang T, Ping X. A fast and effective steganalytic technique against JSteg-like algorithms. In: Proc. of the 2003 ACM Symp. on Applied Computing. 2003. 307−311.
- [22] Westfeld A. F5—A steganographic algorithm. In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 2001. 289−302.
- [30] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security, 2014,2014:Article No.1.
- [31] Guo L, Ni J, Shi YQ. Uniform embedding for efficient JPEG steganography. IEEE Trans. on Information Forensics and Security, 2014,9(5):814−825.
- [32] Guo L, Ni J, Su W, Tang C, Shi Y. Using statistical image model for JPEG steganography: Uniform embedding revisited. IEEE Trans. on Information Forensics and Security, 2015,10(12):2669−2680.

GAN

2016年,Volkhonskiy[33]提出了SGAN的隐写模型,结合已有的DCGAN网络生成更加符合隐写规则的载体图像.2018年,ASDL-GAN[34]和UT-6HPF-GAN[35]相继被提出,将对抗网络应用在修改概率图的生成上.由于生成对抗网络需要一个‘对手’共同进步,通常将基于深度学习的隐写分析模型作为生成对抗网络中的对立方,这样两个模型可以在对抗学习中共同进步.这种新型的隐写方法不仅减少了人为参与,还可以有效提升隐写安全性.但是仍然存在一些问题,例如网络架构不稳定、GAN网络的不可逆性导致隐写内容无法准确提取等.

隐写分析

隐写分析是检测隐写术的一种手段，可分为三个阶段。

- 第1个阶段:判断载密图像(stego)中是否隐藏秘密信息,即判断数字图像是载体图像(cover)还是载密图像.这是现在大多数隐写分析模型最重要的步骤,也被称为盲隐写分析.
- 第2个阶段:判断载密图像中秘密信息的容量和秘密信息隐藏的位置等(多为纹理复杂处或者图像边缘处).
- 第3个阶段:从载密图像中提取秘密信息,这个阶段需要具体了解隐写方法、隐写位置、隐写容量等各种信息[36].

针对LSB和LSBM这两种空域隐写算法,修改最低有效位会在一定程度上破坏相邻像素之间的关联性.根据这一特性,存在相应的专用型隐写分析模型[37,38].专用型隐写分析是指隐写分析一方在已知隐写具体算法的情况下所设计的特用的隐写分析模型,数字图像在嵌入秘密信息后,载体图像的某种统计特性特征会发生相应的改变.通用隐写分析在未知载体图像和隐写算法的基础上,检测图像是否含有秘密信息.相对于通用型隐写分析,专用型隐写分析的准确率更高但具局限性.

空域隐写分析通过分析数字图像的统计特性,来检测图像是否嵌入秘密信息;而频域隐写分析由于不同的DCT与量化矩阵,则需要分析DCT系数关系而进行判别.

自适应隐写算法根据图片最小失真函数,结合STC[21]使用进行隐写.这使得隐藏的秘密信息越来越难以发现,所以图像中秘密信息的有效特征越来越难以获取,原有的隐写分析特征一般是由专业的研究人员依赖自己的先验经验和不断启发式尝试计算得出.隐写分析的特征提取和机器学习二分类训练是分开的,前者通过手工设计,后者通过机器学习方法完成,两步操作无法同时进行优化,很难达到一个异构平衡状态.利用深度学习端到端的学习过程，使得特征提取和判别器可以同时训练.依赖深度学习可以模拟人脑学习复杂的结构信息,从而提取出数字图像中的特征信息。

数据集与评价指标

数据集

隐写术和隐写分析所采用的数据集多为BOWS2(https://photogallery.sc.egov.usda.gov/)和BOSSbase-v1.01(ttp://agents.fel.cvut.cz/stegodata/),两款数据集都是512×512的一万张灰度图,数据集中包含生活、景点、建筑等多种类型图片。

Table 1 Source of BOSSbase image datasets

表 1 BOSSbase 图像数据拍摄来源

图片序号	相机型号
1–1354	CanonEOS 400D
1355–1415	CanonEOS 40D
1416–2769	CanonEOS 7D
2770–4811	CanonEOSDIGITALREBELXsi
4812–6209	PENTAXK20D
6210–7242	NIKON D70
7243–10212	M9 digital camera

Table 2 Comparison of different datasets

表 2 各类数据库对比

数据集	数据量(张)	位深度	图像类型	图片大小	格式
BOSSbase	10 000	8	灰度图	512×512	PGM
BOWS2	10 000	8	灰度图	512×512	PGM
UCID	1 338	24	彩图	512×384/384×512	TIF
NRCS	N/A	24	彩图	1500×2100	TIF/JPEG
SIPI	N/A	8/24	灰/彩图	256×256/512×512/1024×1024	TIFF

同的数据集之间存在一定的相似性,较为常用的数据集是BOSSbase[57]和BOWS2[58],这两类数据集不仅属性相似,图片的内容也存在一定的相似性,所以在隐写分析模型需要对数据集进行增强操作时,通常混用两个数据集进行网络训练.

通常,比较隐写分析网络的检测效果,采用误检率Err或准确率Acc作为模型效果的衡量标准.隐写分析的目标是从数字图像中检测载密图像,因此将载密图像作为阳性类,载体图像作为阴性类.假设载体图像和载密图像的数量分别为C和S,其中被正确分类的载体图像与载密图像的样本数为N和P,在评价隐写分析模型时,通常会用到如下几种指标:

$$Acc = \frac{P+N}{C+S}$$

$$Err = 1 - \frac{P+N}{C+S}$$

P_{FA} 代表虚警率(false alarm ratio),即代表载体图像被误判成载密图像的比率.

$$P_{FA} = \frac{C-N}{C} \tag{3}$$

P_{MD} 代表漏检率(missed detection ratio),即代表载密图像被误判成载体图像的比率.

$$P_{MD} = \frac{S-P}{S} \tag{4}$$

P_E 代表最小平均错误率(minimum average decision error ratio),即在虚警率发生变化时,两类错误平均值的最小值.

$$P_E = \frac{1}{2}\min_{P_{FA}}(P_{FA} + P_{MD}(P_{FA})) \tag{5}$$

$MD5$ 代表当 P_{FA} 为 5%情况下的误检率.

$$MD5 = P_{MD}(P_{FA}=0.05) \tag{6}$$

$FA50$ 代表当 P_{MD} 为 50%情况下的虚警率.

$$FA50 = P_{FA}(P_{MD}=0.5) \tag{7}$$

公式(6)、公式(7)为在 ALASKA 隐写分析挑战赛[59]中的评判标准.

半学习隐写分析

半学习是指在隐写分析网络利用固定滤波核作为独立的一个预处理层,并且内部的权重参数不参与反向传播,其他的网络层则是依赖深度学习方法去优化.在本节中,按照网络的架构分为深度网络模型与宽度网络模型.

基于深度网络的半学习隐写分析模型

2015年,Qian等人[63]提出了一种新的网络,称为GNCNN(Gaussian-Neuron CNN),图6是GNCNN网络与传统隐写分析之间的对比图.

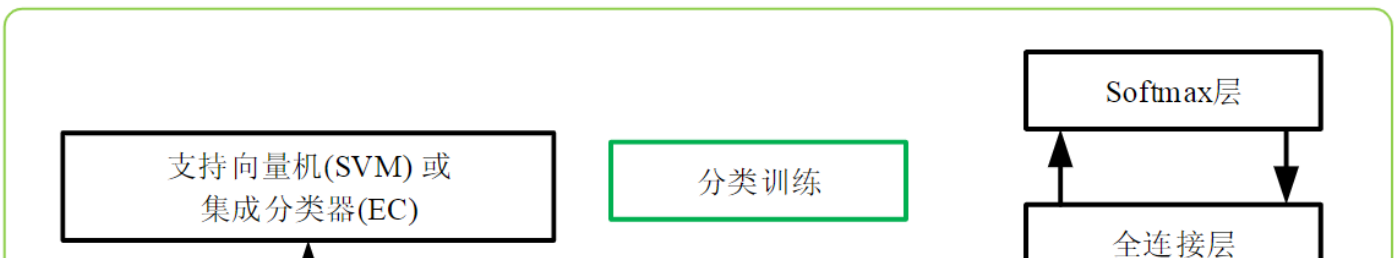


Fig.6 Traditional steganalysis and GNCNN structure

图 6 GNCNN 与传统隐写分析结构

该网络结构包括一个预处理层、5个卷积层和3个全连接层,预处理层将卷积层中的卷积核替换成固定的高通滤波核,获取数字图像的高维残差信息,辅助网络进行学习.这样不仅仅加快了隐写分析网络的训练,而且将不必要的图像内容信息移除,减少了图像低维信息干扰.在实验过程中,加入了固定的高通滤波核的GNCNN网络在训练速度和训练结果上都会优于使用在预处理层中随机初始化卷积核的网络.由于经过高通滤波器后得到的信息多为高频残差信息,最大池化容易丢失高频残差图像信息,导致网络难以拟合,所以在GNCNN中,使用平均池化操作来减少残差信息的丢失.Qian根据隐写噪声的特点提出了高斯激活函数,替代卷积层中的ReLU激活函数.如下是GNCNN中所采用高斯激活函数.

$$F(x)=1-\mathrm{e}^{-\frac{x^2}{\sigma^2}}$$

其中,σ是用来衡量函数曲线宽度的参数.该公式可以将数值较小的输入转换成一个正数。对比试验结果如下：

Table 3 Comparison of experimental results under different steganography algorithms of traditional steganalysis and GNCNN

表 3 GNCNN 与传统隐写分析在不同隐写算法下实验结果对比

BPP	HUGO			WOWO			S-UNIWARD		
	GNCNN	SRM	SPAM	GNCNN	SRM	SPAM	GNCNN	SRM	SPAM
BOSSbase									
0.3	33.8%	29.6%	42.9%	34.3%	31.2%	42.2%	35.9%	34.3%	40.0%
0.4	28.9%	25.2%	39.1%	29.3%	25.7%	38.2%	30.9%	29.3%	35.1%
0.5	25.7%	21.4%	35.7%	24.8%	22.1%	34.9%	26.3%	24.8%	30.6%
ImageNet database									
0.4	33.6%	32.5%	–	34.1%	34.7%	–	34.7%	34.4%	–

从表3的实验结果中可以看出,GNCNN的检测效果较优于SPAM较弱于SRM.在各类的隐写算法上都满足这样一个条件:随着嵌入率(bit per pixel,简称BPP)的提升,即隐写容量的增加、载密图像中嵌入的秘密信息增加,隐写分析的准确率就会越高.BOSSbase是由10 000张经过裁减的灰度图所组成的专用数据集;表3中最下一行的ImageNet[64]数据集则是由互联网中大量彩图组成,在彩图隐写分析上,GNCNN已经与SRM的检测效果非常接近.在BOSSbase数据集上,通过大量数据测试发现:GNCNN仅仅比SRM的检测正确率低3%~5%;而对于彩图这种通道数较多的数据集而言,GNCNN与SRM的隐写检测水平相近.这是因为相对于灰度图的隐写,彩图不同通道间具有关联性且包含的信息更多,因此彩图隐写也更容易被检测,对于网络自学习的参数权重要求较低.相对于其他基于深度学习的隐写分析而言,GNCNN由于网络模型较为简单,在隐写分析的准确率上存在局限性.

xu-net在网络框架上仍然沿用了GNCNN的网络架构特点,例如依旧采用全局池化操作,减少残差图像信息的丢失.同样在网络前端添加了一个固定的高通滤波层,即KV核作为预处理层,如下所示.

$$K_{kv}=\frac{1}{12}\begin{pmatrix}-1 & 2 & -2 & 2 & -1\\ 2 & -6 & 8 & -6 & 2\\ -2 & 8 & -12 & 8 & -2\\ 2 & -6 & 8 & -6 & 2\\ -1 & 2 & -2 & 2 & -1\end{pmatrix}$$

滤波核是从SRM[48]的30个高通滤波核中挑选出来的,在区分高维特征即纹理复杂度时具有较好的效果.高通滤波器是一种中心对称的结构,这样可以有效地提取出像素点与周围像素之间的信息差距,使得隐写分析模型可以有效地获取像素之间的共生矩阵,重新排列得到信噪特征,从而帮助隐写分析模型更好地检测,各类不同的滤波核在处理相同的数字图像时会有不同的效果.其结果已足够抗衡传统隐写分析。

Table 4 Comparison of detection accuracy of Xu-Net and SRM on S-UNIWARD and HILL

表 4 Xu-Net 与 SRM 在 S-UNIWARD 与 HILL 下准确率对比

Algorithm	BPP (%)			
	0.1bit/pixel		0.4bit/pixel	
	CNN	SRM	CNN	SRM
S-UNIWARD	57.33	59.25	80.24	79.53
HILL	58.44	56.44	79.24	75.24

Xu-Net网络根据经过预处理层的残差高频噪声信号具有关于0对称且与符号无关的特性,在第1个卷积层采用添加ABS(absolute layer)层来收敛特征图的范围,从原来无意义的正负区间缩小到正向区间.添加BN层(batch normalization layer)进行批处理,使得训练数据符合正态分布.这样可以提升训练时的收敛速度,也可以避免训练时出现梯度弥散或梯度爆炸现象,导致训练结果陷入局部最小值.最后采用1×1的卷积核将特征信息集聚,并且防止模型存在过拟合的情况.