多色彩通道特征融合的GAN合成图像检测方法

乔 通[1,2,3],陈彧星[1],谢世闯[1],姚 恒[4],罗向阳[3*]

(1. 杭州电子科技大学网络空间安全学院,浙江杭州 310016;2. 中国科学院信息工程研究所信息安全国家重点实验室,北京 100093;3. 河南省网络空间态势感知重点实验室,河南郑州 450001;4. 上海理工大学光电信息与计算机工程学院,上海 200093)

摘 要: 当前,生成对抗网络(Generative Adversarial Networks, GAN)合成的逼真图像难以识别,严重危害国家网络安全及社会稳定. 与此同时,多数基于深度神经网络模型设计的检测器需要大规模训练样本,且存在模型可解释度不高、泛化性能差等问题. 为了克服上述亟待解决的关键性难题,本文提出一种多色彩通道特征融合的GAN合成图像检测方法. 首先,探索分析真实自然图像和GAN合成图像在不同色彩空间相邻像素之间的差异,并设计差异度量算法,完成色彩通道选择. 其次,利用图像像素间的高度相关性,在八个方向上通过二阶马尔可夫链对相邻像素之间的差分数组进行建模,提取差分像素邻接矩阵特征. 最后,利用上述特征,设计一种简单且高效的集成分类器完成GAN合成图像的检测任务. 在基于StyleGAN模型合成的伪造人脸数据集中,所提出方法的检测准确率高达100.00%;在小样本训练约束条件下,正负样本对数仅仅为2时,检测准确率高达99.65%;在单类样本训练约束条件下,正样本数仅仅为50时,检测准确率高达92.84%. 在基于更先进的StyleGAN2和PGGAN模型合成的伪造场景数据集中,所提出方法的检测准确率达到99.96%以上. 以上大量实验表明,本文所提出的方法明显优于比较的GAN合成图像检测方法. 本文方法已经开源:https://github.com/cyxcyx559/ccss.

关键词: 图像取证;色彩通道;特征融合;生成对抗网络;马尔可夫链;集成分类器

基金项目: 浙江省属高校基本科研业务费专项资金(No.GK219909299001-007);模式识别国家重点实验室开放课题基金(No.202200026);河南省网络空间态势感知重点实验室开放课题基金(No.HNTS2022016);国家重点研发计划(No.2022YFB3102900);国家自然科学基金(No.U1804263,No.62172435,No.62172281);中原科技创新领军人才项目(No.214200510019)

中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2024)03-0924-13

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220711

GAN Synthetic Image Detection Using Fused Features in the Multi-Color Channels

QIAO Tong[1,2,3], CHEN Yu-xing[1], XIE Shi-chuang[1], YAO Heng[4], LUO Xiang-yang[3*]

(1. *School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou, Zhejiang* 310016, *China*;
2. *State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093, *China*;
3. *Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou, Henan* 450001, *China*;
4. *School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)

Abstract: Currently, it is very difficult to identify the images synthesized by generative adversarial networks (GAN), which severely poses the threat on national cyber security and social stability. Meanwhile, most classifiers based on deep neural networks require large-scale samples for training, where the problems such as low model interpretability and poor generalization performance are less addressed. To overcome the limitations, we propose to design the ensemble classifier using fused features in the multi-color channels. First of all, by studying the discrimination of adjacent pixels in the multi-color channels between natural and GAN synthetic images, the difference metric is designed based on the correlation of adjacent pixels, in order to select the optimal color channels. Secondly, by utilizing the highly-correlated relationship among pixels, the difference array between adjacent pixels are modeled through a second-order Markov chain along eight directions, and meanwhile the subtractive pixel adjacency matrix features are successfully extracted. Finally, based on

收稿日期:2022-06-20;修回日期:2023-02-23;责任编辑:梅志强
*通讯作者:罗向阳

the extracted features, a simple but efficient detector for identifying GAN synthetic images is constructed. In the image dataset synthesized by the StyleGAN model, the results show that the accuracy of the proposed detector can reach 100.00%. It can also identify GAN synthetic images very well when the pair number of positive and negative training samples is 2 (99.65% accuracy) or only 50 positive training samples are provided (92.84% accuracy). The accuracy can also reach more than 99.96% in the image dataset synthesized by StyleGAN2 and PGGAN models. Numerous experiments show that the proposed method in this paper is better than the compared forensic methods. Our code is available at https://github.com/cyxcyx559/ccss.

Key words: image forensics; color channels; features fusion; generative adversarial networks; Markov chain; ensemble classifier

Foundation Item(s): Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK219909299001-007); Open Projects Program of National Laboratory of Pattern Recognition (No. 202200026); Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (No. HNTS2022016); National Key R&D Program of China (No. 2022YFB3102900); National Natural Science Foundation of China (No. U1804263, No. 62172435, No.62172281); Zhongyuan Science and Technology Innovation Leading Talent Project of China (No.214200510019)

1　引言

通信技术和计算机技术的飞速发展,加速了数字媒体技术和电子设备的快速普及,促使数字图像被广泛应用于日常生活和工作中,使其成为互联网信息交流的重要媒介,人类日常生活中交换信息的主要载体.数据显示,2014年,Whatsapp的日均图片发送量达到了惊人的八亿张.2015年,Instagram用户每日上传图片数量超过了8 000万张.与此同时,随着深度神经网络技术的飞速发展,图像合成技术同样取得了巨大的成功.利用一些合成软件或者开源网络平台,人们可以轻松合成高度逼真的伪造图像,达到以假乱真、混淆是非的目的.这些高度逼真的合成图像若被用在新闻报道、身份认证、司法鉴定等领域,会导致严重的真实性问题[1].例如,计算机合成图像伪造假新闻,影响社会舆论导向及社会稳定;通过计算机合成的人脸图像伪造个人信息,不法分子可以轻而易举的实施网络诈骗等犯罪活动;若在司法鉴定中错误地把计算机合成的伪造图像认定为真实的自然图像,可能对当事人造成无法弥补的损失.总而言之,数字图像的真实性及原始性遭受前所未有的挑战.

早期的计算机合成图像利用合成软件如3D MAX, Softimage-XSI, Maya, Terragend等实现.此类方法建立在复杂的数学模型之上,对技术要求较高,需要专业人士参与实现,导致其推广度不高.近些年,生成对抗网络[2](Generative Adversarial Networks, GAN)技术的提出,进一步促进了计算机视觉领域的发展,同时也把计算机图像合成技术带入到一个崭新的时代.真实自然图像如图1所示,GAN合成伪造图像如图2所示,仅凭肉眼很难分辨图1和图2的真实性.利用GAN技术,合成的伪造图像不再需要专业软件进行复杂繁冗的手工操作,经过训练得到的生成对抗网络可以自动地合成伪造图像,并且不断优化伪造图像的视觉质量.大量的GAN合成图像在互联网上传播扩散,严重危害国家网络安全及社会稳定.因此,不论学术界还是工业界,加强自然图像和GAN合成图像真伪鉴别的研究迫在眉睫.

目前,一系列数字图像取证技术已经被用于检测自然图像和GAN合成图像.目前,多数GAN合成图像检测方法基于深度神经网络[3](Deep Neural Networks, DNN)架构,往往需要大规模的训练样本,且存在模型可解释度不高、泛化性能差、过拟合程度较高等问题,同时存在遭受对抗样本攻击的风险.此外,伪造图像样本相较于真实图像样本数量较少,如何在小样本或者单类正样本约束条件下,训练高效可靠的取证检测器,是目前学术界亟待解决的难题.针对以上问题,本文没有采用目前广泛流行的DNN网络模型框架,转而继续沿用基于传统机器学习框架的检测方法.一方面,在多

(*a*)　数据集LSUN图例

(*b*)　数据集CelebA-HQ图例

图1　真实自然图像

(*a*) 基于LSUN数据集训练,使用StyleGAN2模型合成的伪造场景

(*b*) 基于CelebA-HQ数据集训练,使用PGGAN模型合成的伪造人脸

图2 GAN合成伪造图像

色彩通道选择后,具有可解释性的手工提取特征足以很好表征自然图像与GAN合成图像的差异性;另一方面,精心设计的集成分类器在处理小样本及单类样本的问题时,可以达到优于目前同类算法的检测准确率.具体而言,本文主要的贡献如下:

(1)本文分析了不同色彩通道上,自然图像和GAN合成图像之间的差异性,并利用差别度量算法,选择最优的色彩通道,用于提取两类图像特征.

(2)本文提出利用图像像素间的高度相关性,在八个方向上提取差分像素邻接矩阵(Subtractive Pixel Adjacency Matrix, SPAM)特征[4]. 通过二阶马尔可夫链对相邻像素之间的差分数组进行建模,基于随机森林的原理构建了高效可靠的集成分类器,实现了对自然图像和GAN合成图像的有效鉴别.

(3)大量实验结果表明,在基于StyleGAN模型合成的伪造人脸数据集中,所提出方法的检测准确率高达100.00%;在小样本训练约束条件下,正负样本对数仅仅为2时,检测准确率高达99.65%;在单类样本训练约束条件下,正样本数仅仅为50时,检测准确率高达92.84%. 在基于更先进的StyleGAN2和PGGAN模型合成的伪造场景数据集中,所提出方法的检测准确率达到99.96%以上.

2 国内外研究现状

2.1 GAN伪造图像的合成

生成对抗网络GAN是一种深度学习模型,2014年由Goodfellow等人提出,主要包括生成器和判别器两部分. 两者通过对抗形式训练,从而实现对高维数据分布的隐式估计. GAN模型通过计算合成图像和真实图像分布的相似性来优化生成器和判别器,利用互相博弈学习产生逼真的合成图像. 具体的,判别器的任务就是判断一个实例是真实的还是由模型生成的;生成器的任务就是生成一个实例来欺骗判别器. 两个模型相互对抗,最终达到一个平衡,即生成器生成的实例与真实的没有显著差异,判别器无法区分输入数据是生成的还是原始真实的数据. 将GAN应用于图像合成领域,生成器学习训练样本在色彩空间的数据分布,从而模拟真实图像的概率分布,不断将参数进行优化更新,最终使得判别器无法识别自然图像和GAN合成图像(如图1和图2所示).

GAN一经提出,广泛应用于图像合成等领域,并取得了巨大的成功. Denton等人提出了LAPGAN[5],Zhang等人提出了stackGAN[6,7],训练多分辨率金字塔组织的GAN堆线,达到合成高分辨率图像的目的. Karras等人提出了PGGAN[8],通过逐渐增长的生成器和鉴别器来合成高分辨率的图像. Gulrajani等人提出了WGAN-GP[9],对梯度范数进行惩罚,是一种削减权重的替代方法,从而进一步提高学习的稳定性,摆脱模式崩溃的风险. Zhang等人提出了将分层嵌套的对抗目标合并到网络分层结构中的HDGAN[10]. 在BIGGAN[11]中,应用更好的正则化技术来合成高分辨率的图像. 在Lin等人提出的COCO-GAN[12]中,生成器基于空间坐标,按部分合成比训练样本更大的图像. 受PGGAN启发,Karras等人提出了面向风格迁移的StyleGAN[13],实现了自动地无监督地将合成图像的高级属性和随机变化分离. 此外,该团队随后提出了升级版StyleGAN2[14],重点修复人造痕迹问题,并进一步提高了合成图像的质量. 然而,GAN合成图像可能会被不法分子恶意乱用,在互联网上传播扩散,严重危害国家网络安全及社会稳定. 在如今的互联网时代,如何有效地鉴别真实的自然图像与伪造的GAN合成图像变得至关重要.

2.2 GAN伪造图像的检测

目前,如何有效鉴别自然图像与GAN合成图像在数字图像取证研究中变得越来越重要[15]. GAN伪造图像的检测方法大致可以分为两类:一类是基于DNN提取特征的方法,一类是基于手工提取特征的方法.

基于DNN提取特征的方法[16]主要依靠数据驱动,特征提取器针对每一个特定的类别会自动学习相关特征. 依据大量样本的学习能够得到深层的数据集的特征表示,同时可以探索图像类别中的底层模式. Mo等人使用高通滤波器将输入图像转换为残差后送到神经网络中,修改了CNN的架构,以监督学习的方式实现了对深度伪造图像的检测[17]. Dang等人提出了一种定制的卷积神经网络CGFace[18],通过自定义卷积层数来实

现对GAN合成人脸图像的检测. Zhang等人根据GAN的指纹特征,提出了一种基于频谱输入的分类器模型,实现了对CycleGAN等模型所合成的伪造图像的准确检测[19]. Yu等人通过学习GAN合成图像的指纹属性,分类自然图像和GAN合成图像[20]. Zhuang等人提出了一种两步成对的学习方法,在使用不同生成模型生成的训练图像上捕捉伪像或真像的局部和全局特征,从而进一步提高检测性能[21]. Fu等人提出了一种基于双通道卷积神经网络的GAN图像检测算法,采用高通滤波器计算高频残差,低通滤波器计算低频分量,分别输入浅层神经网络. 该方法虽然能够提升对部分后处理操作的鲁棒性,但仍然难以抵抗JPEG压缩[22]. 但是,基于DNN设计的分类器普遍存在训练模型时间较长,计算开销较大,模型可解释性不高等问题[23].

基于手工提取特征的方法主要依靠人工设计的特征和传统机器学习技术相结合. McCloskey等人提出红绿双变量直方图和异常曝光像素比例的特征,结合支持向量机进行GAN图像检测,但实验结果表明该算法在缺乏训练样本的情况下性能不理想[24]. Nataraj等人提出了一种提取三个色彩通道上的共生矩阵,并用深度卷积神经网络框架相结合的方法来检测GAN合成图像的方法[25]. Barni等人除了使用空间共生矩阵之外,还使用跨频带共生矩阵作为CNN模型的输入,主要关注GAN合成的面部图像和自然图像的检测[26]. Li等人观察发现在残差域上伪造图像与真实图像在色度分量上更容易区分,提出了一种获取彩色图像统计信息的特征集的方法识别伪造图像[27]. 另外,研究人员发现,使用已广泛应用的或改进的隐写分析领域统计特征,结合分类器对自然图像和GAN合成图像进行检测,也能达到很好的效果. Marra等人提出将空域富模型(Spatial Rich Model, SRM)隐写分析特征[28]用于真伪图像检测[29]. Gan等人受隐写分析的启发,将笛卡尔标准化JPEG富模型(cartesian calibrated JPEG Rich Model, ccJRM)隐写分析特征[30]用于伪造取证算法的设计,可以自动检测视频伪造帧[31].

虽然目前基于DNN提取特征的取证方法层出不穷,但是在解决某些问题上,例如鉴别自然图像与GAN合成图像,未必基于手工提取特征的传统方法比新兴的基于DNN的方法差. 基于DNN的模型虽然能够达到较高的准确率,但由于存在模型复杂、可解释性不高,对设备的算力要求高等问题,可能在实际应用中难以部署. 通过低维度的手工提取特征结合传统机器学习算法[32],构建"轻量级"的机器学习模型势在必行,该类方法计算效率较高,模型复杂度低,可解释强,易于部署且能够达到较高的准确率. 因此,在训练样本数量有限或者仅存在单一样本集的条件下,我们认为采用基于手工提取特征的方法就能够对自然图像和GAN合成图像达到准确的识别.

本文提出一种轻量级的手工提取特征方案,结合强大的集成分类器,达到精准地、高效地识别GAN合成图像的目标.

3　检测框架设计

本节按照所提出系统的一般框架展开论述. 该检测方法主要由三个模块组成:(1)色彩通道转换;(2)特征提取;(3)集成分类器搭建. 本文所提出框架的流程图如图3所示.

3.1　色彩通道转换

GAN合成的伪造图像和自然图像的生成机理存在明显差异. GAN合成图像是由计算机通过不断优化神经网络参数等形成的,而自然图像通常是经历了透镜,光学滤波,传感器光电转化、CFA插值等线性及非线性变换后,再经过一些后处理形成的. 自然图像和GAN合成图像成像渠道的不同,造成其色彩通道的差异. 因此,我们可以利用这些色彩通道的差异,表征为区分度明显的特征向量,从而鉴别两类图像. RGB色彩空间将色调、亮度、饱和度三个分量统一表征,缺乏对颜色分量和亮度分量的细粒度刻画. 相反,在其他色彩空间(例如HSV、YCbCr等)中,我们可以更好地区分颜色分量和亮度分量. HSV、YCbCr色彩空间在图像处理这一领域应用广泛,一般采用某些合适的通道就能完成大多数图像预处理工作. 我们在多个色彩通道上对自然图像和GAN合成图像进行全面分析,选择最优的若干通道完成特征提取任务.

我们首先在RGB、HSV、YCbCr这三个色彩空间定义了一个度量标准[33]来识别哪个色彩通道在识别GAN合成图像时更为行之有效. 对于每一张图像,在三个色彩空间的色彩通道 $c(c\in\{\mathrm{R,G,B,H,S,V,Y,Cb,Cr}\})$ 上计算相邻像素之间的相关系数,若图像尺寸为 $a\times b$,定义的度量标准如下:

$$L^c=\frac{\sum_{m=1}^{a-1}\sum_{n=1}^{b}(\boldsymbol{F}^c_{m+1,n}-\overline{\boldsymbol{F}^c})(\boldsymbol{F}^c_{m,n}-\overline{\boldsymbol{F}^c})}{\sqrt{\sum_{m=1}^{a-1}\sum_{n=1}^{b}(\boldsymbol{F}^c_{m+1,n}-\overline{\boldsymbol{F}^c})^2\sum_{m=1}^{a-1}\sum_{n=1}^{b}(\boldsymbol{F}^c_{m,n}-\overline{\boldsymbol{F}^c})^2}}\tag{1}$$

其中,a 为图像的高度,b 为图像的宽度,a 和 b 在本文中的大小均为512,$\boldsymbol{F}^c$ 是图像在该通道上的像素矩阵,$\overline{\boldsymbol{F}^c}$ 是 $\boldsymbol{F}^c$ 的平均值. 由此可以得到,L^c 的值越大,相邻像素的相关性越高. 本文中仅考虑垂直方向相邻像素的相关性,在其他方向也能得到类似结果.

然后,在每个色彩通道上计算每一张图像的 L^c,并构造 L^c 的直方图,其中横坐标代表 L^c 的数值,纵坐标代表相应的归一化频次,自然图像的直方图表示为 H^c_{R},

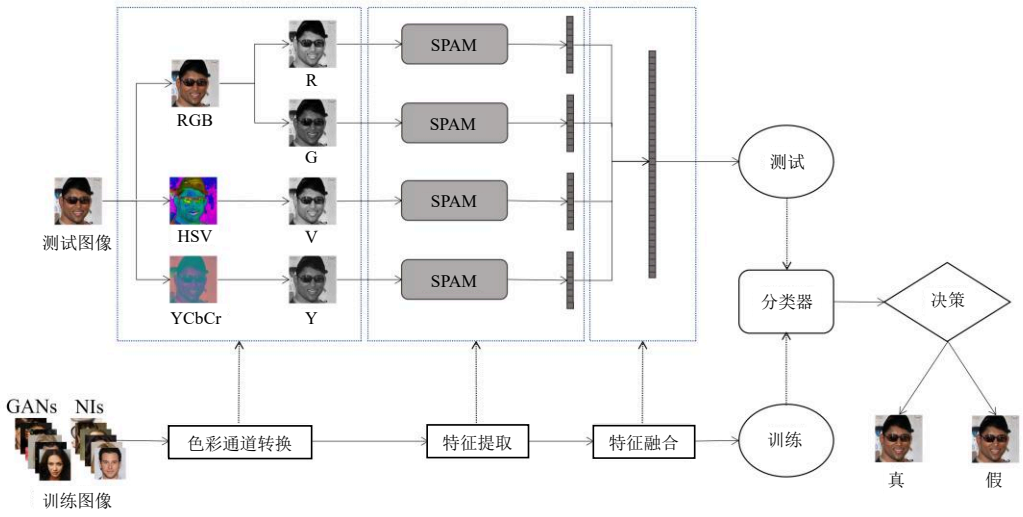


图3 本文所提出框架的流程图 (GANs: GAN合成图像,NIs: 自然图像)

GAN合成图像的直方图表示为H_{G}^{c}.

两个直方图之间的差异性可以用卡方距离来度量,卡方距离的计算方法如下:

$$d_{x^2}(H_{\mathrm{G}}^{c},H_{\mathrm{R}}^{c})=\frac{1}{2}\sum_{k}\frac{(H_{\mathrm{G}}^{c}(k)-H_{\mathrm{R}}^{c}(k))^2}{H_{\mathrm{G}}^{c}(k)+H_{\mathrm{R}}^{c}(k)} \tag{2}$$

其中,k是直方图的取值索引. 由式(2)可知,d_{x^2}的值越大,GAN合成图像和真实图像在这一通道的差别就越大,因此在这一色彩通道可以更好的识别自然图像和GAN合成图像.

为了全面评估自然图像和GAN合成图像的有效色彩通道,我们使用由CelebA-HQ[8]数据集训练的StyleGAN模型作为图像生成模型,生成了大量的GAN合成图像. 同时,随机选择了大小为512×512的10 000张自然图像和10 000张GAN合成图像计算R, G, B, H, S, V, Y, Cb, Cr这九个通道的相邻像素之间的相关系数和每个通道直方图之间的卡方距离,绘制的每个通道的直方图如图4所示. 可以观察到在R, G, V, Y四个色彩通道上,非重叠区域更大,而B, H, S, Cb, Cr这五个色彩通道的直方图重合面积较大甚至几乎重叠. R, G, V, Y四个色彩通道的$d_{x^2}(H_{\mathrm{G}}^{c},H_{\mathrm{R}}^{c})$的值都超过了0.3,均大于B, H, S, Cb, Cr这五个色彩通道. 这说明自然图像和GAN合成图像在R, G, V, Y这四个色彩通道上差异更大,更容易区分.

一方面,根据自然图像的生成机理,G通道通常携带更多的信息;另一方面,GAN合成图像一般需要色域映射以便在不同的输出设备上保持颜色感知属性,导致生成图像相较于原始图像纹理发生改变[34]. 此外,由于光照不同造成人眼视觉的感知不同,采用V, Y亮度通道进行计算可以更好的提取物体的边缘信息[35]. 我们得到的实验结果和以上理论分析基本一致,因此选择R, G, V, Y这四个色彩通道的特征,对自然图像和GAN合成图像进行检测.

3.2 差分像素邻接矩阵特征提取

像素直方图中的数据量随像素数量呈指数增长,因此直接在像素上寻找差异是困难的. 由于自然图像中的像素值存在高度空间相关性[4],相邻像素的取值高度相似,两个相邻像素$(\boldsymbol{F}_{m,n},\boldsymbol{F}_{m,n+1})$出现的概率$P(\boldsymbol{F}_{m,n},\boldsymbol{F}_{m,n+1})$分布成形为中间对称的形状,因此像素差$\boldsymbol{F}_{m,n}-\boldsymbol{F}_{m,n+1}$近似独立于像素$\boldsymbol{F}_{m,n}$. 使用差分像素邻接矩阵SPAM特征,能够有效刻画自然图像和GAN合成图像的差异. 具体实现过程如下:

首先计算八个方向差值数组:

$$\boldsymbol{T}^{\rightarrow}_{m,n}=\boldsymbol{F}_{m,n}-\boldsymbol{F}_{m,n+1} \tag{3}$$

其中$m\in\{i|1\leqslant i\leqslant a\}$,$n\in\{i|1\leqslant i\leqslant b-1\}$,上标$\{\leftarrow,\uparrow,\rightarrow,\downarrow,\nearrow,\swarrow,\searrow,\nwarrow\}$分别表示计算的方向,$\boldsymbol{F}$表示图像的像素矩阵,$\boldsymbol{T}$表示差值数组. 由于其他方向的计算和水平方向类似,本文仅定义了水平方向上的计算方法.

通过二阶马尔可夫链对相邻像素之间的差分数组的转移概率矩阵进行建模,得到二阶的SPAM特征. 在水平方向上一阶特征可以表示为:

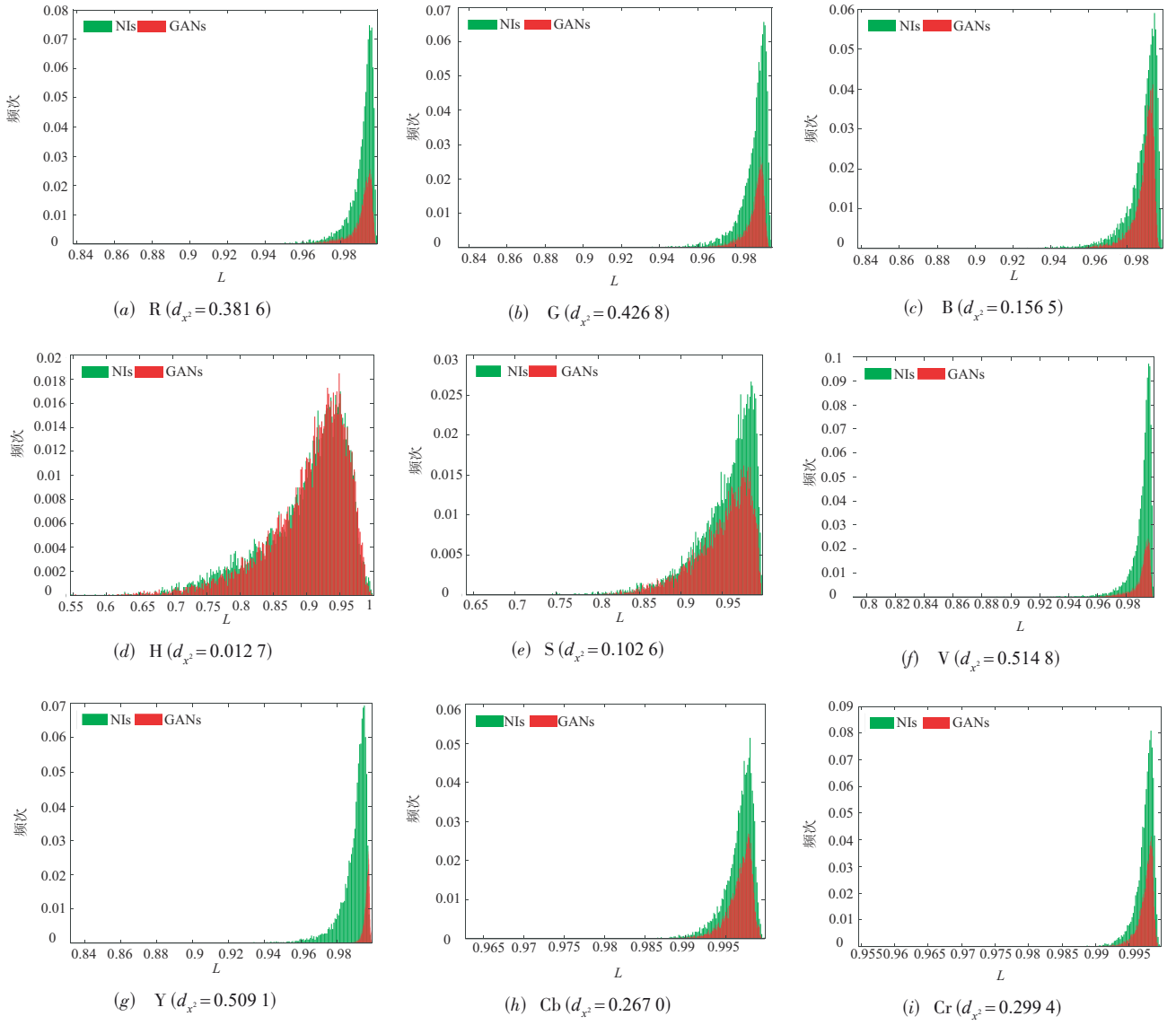
$$\boldsymbol{K}^{\rightarrow}_{x,y}=P(\boldsymbol{T}^{\rightarrow}_{m,n+1}=x|\boldsymbol{T}^{\rightarrow}_{m,n}=y) \tag{4}$$

其中,$x,y\in\{i|-L\leqslant i\leqslant L\}$,$P(\boldsymbol{T}^{\rightarrow}_{m,n+1}=x|\boldsymbol{T}^{\rightarrow}_{m,n}=y)$表示在$\boldsymbol{T}^{\rightarrow}_{m,n}=y$的情况下,$\boldsymbol{T}^{\rightarrow}_{m,n+1}=x$的概率. 若$P(\boldsymbol{T}^{\rightarrow}_{m,n+1}=x)=0$,则$\boldsymbol{K}^{\rightarrow}_{x,y}=P(\boldsymbol{T}^{\rightarrow}_{m,n+1}=x|\boldsymbol{T}^{\rightarrow}_{m,n}=y)=0$.

在水平方向上二阶特征表示为:

$$\boldsymbol{K}^{\rightarrow}_{x,y,z}=P(\boldsymbol{T}^{\rightarrow}_{m,n+2}=x|\boldsymbol{T}^{\rightarrow}_{m,n+1}=y,\boldsymbol{T}^{\rightarrow}_{m,n}=z) \tag{5}$$

其中,$x,y,z\in\{i|-L\leqslant i\leqslant L\}$,$P(\boldsymbol{T}^{\rightarrow}_{m,n+2}=x|\boldsymbol{T}^{\rightarrow}_{m,n+1}=y,\boldsymbol{T}^{\rightarrow}_{m,n}=z)$表示在$\boldsymbol{T}^{\rightarrow}_{m,n+1}=y$且$\boldsymbol{T}^{\rightarrow}_{m,n}=z$的情况下,$\boldsymbol{T}^{\rightarrow}_{m,n+2}=x$的概率. 若$P(\boldsymbol{T}^{\rightarrow}_{m,n+1}=y,\boldsymbol{T}^{\rightarrow}_{m,n}=z)=0$,则$\boldsymbol{K}^{\rightarrow}_{x,y,z}=P(\boldsymbol{T}^{\rightarrow}_{m,n+2}=x|\boldsymbol{T}^{\rightarrow}_{m,n+1}=y,\boldsymbol{T}^{\rightarrow}_{m,n}=z)=0$.



(a) R (d_{x^2}=0.381 6)　(b) G (d_{x^2}=0.426 8)　(c) B (d_{x^2}=0.156 5)

(d) H (d_{x^2}=0.012 7)　(e) S (d_{x^2}=0.102 6)　(f) V (d_{x^2}=0.514 8)

(g) Y (d_{x^2}=0.509 1)　(h) Cb (d_{x^2}=0.267 0)　(i) Cr (d_{x^2}=0.299 4)

图4　自然图像和GAN合成图像在不同色彩通道的直方图

由于差值数组值大量分布在[−3, +3]之间,我们使用范围[−3, +3]来截断差值数组的值. 因此,差值数组的值共有[−3, −2, −1, 0, +1, +2, +3]七种可能. 对于二阶特征,有 $i=7^3=343$ 种可能,因此在每个方向上的特征为343维.

假设自然图像中的统计信息是对称的,对水平和垂直矩阵、对角矩阵进行平均可以得到:

$$\boldsymbol{W}_j=\frac{1}{4}[\boldsymbol{K}^{\rightarrow}+\boldsymbol{K}^{\leftarrow}+\boldsymbol{K}^{\uparrow}+\boldsymbol{K}^{\downarrow}],\quad 0\leqslant j\leqslant i \tag{6}$$

$$\boldsymbol{W}_j=\frac{1}{4}[\boldsymbol{K}^{\rightarrow}+\boldsymbol{K}^{\leftarrow}+\boldsymbol{K}^{\uparrow}+\boldsymbol{K}^{\downarrow}],\quad i<j\leqslant 2i \tag{7}$$

联合式(6)和式(7),得到最后的686维的SPAM特征.

虽然GAN技术已经能够实现非常逼真的图像输出,但究其根本,两类图像的产生过程存在巨大差异,尤其是在颜色通道和图像纹理上具有不同的统计特性. SPAM特征利用自然图像的对称性来降低提取特征的维度,并使用二阶马尔科夫链模拟自然图像中相邻像素之间的差异,从而对差异进行建模,该方法能够识别出偏离该模型的偏差. 其建模的是自然图像相邻像素间既有的关系,而不是非自然图像所拥有的一些特性,因此无论是隐写图像还是GAN生成图像,它们在相邻像素之间的差异是偏离自然图像相邻像素差异模型的,因此我们有理由认为可以通过SPAM特征对两类图像进行分类. 我们在多个色彩通道上提取SPAM特征,从而可以更好地表征图像中相邻像素之间的差异化特征. 根据3.1小节中对色彩通道的分析,在R, G, V, Y这四个色彩通道上,自然图像和GAN合成图像更容易

区分,我们提取R, G, V, Y四个色彩通道的SPAM特征,每个通道的特征向量均为686维,并把它们融合成2 744维特征向量进行下一步的验证.

3.3 集成分类器搭建

支持向量机(Support Vector Machine, SVM)是一个二元分类器,它根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷,以获得最优的泛化能力,即找到一个超平面,使两类数据远离超平面,同时设置阈值,从而判断输入数据的所属类别. LIBSVM[36]是林智仁等人开发的一套支持向量机的库,它采用SVM中支持高斯核的LIBSVM,建立一类分类器. 本文利用该分类器,通过提取自然图像及GAN合成图像的特征向量来构造分类模型,训练SVM分类器.

同时,受到文献[37]的启发,本文进一步提出基于随机森林的集成分类器. 集成分类器最终的决策是通过聚合各个基础学习器的决策而形成的. 对于训练过程,每个学习器在比全维度小的多的特征空间上从训练集提取的引导样本上进行训练,而不是在整个训练集上进行训练. 即使单个基础学习器的性能可能较弱,融合后的准确率也会迅速提高,最终趋于平稳. 其中每个基础学习器为Fisher线性判别器(Fisher Linear Discriminant, FLD)[38],使用63%的样本训练,剩余的37%可用于验证. 对于测试特征,基础学习器通过计算投影并将其与阈值进行比较来做出决定. 在收集了所有的判决之后,通过使用未加权多数投票策略将它们组合起来形成最终的分类器输出.

集成分类器可以进行快速的机器学习,相比深度学习而言,其计算复杂度低,并且能够有效地处理高维特征和大型训练数据集,具有比单一的SVM分类器更好的准确率. 我们根据实际需求选择使用的分类器,把这些提取到的特征向量随机分为训练集和测试集,并利用这些特征向量来训练分类器. 其中自然图像的特征向量作为正样本,GAN合成图像的特征向量作为负样本.

在测试实验中,把测试集中的特征向量输入搭建好的集成分类器中. 通过分类器输出预测的标签,测试集被判断为自然图像或者GAN合成图像. 最后根据输出标签的准确率来验证提出的方法是否具有良好的性能.

4 实验结果

我们构建用于实验仿真的大规模图像库,其中包含真实的自然图像(部分示例图像如图5所示)和伪造的GAN合成图像(部分示例图像如图6所示). 首先,从CelebA-HQ标准数据集中随机选择10 000幅分辨率为1 024×1 024的人脸自然图像. 接着,使用了最近广泛流行的StyleGAN模型来构建GAN合成图像数据集. StyleGAN模型是一种基于风格的生成对抗网络,能够自动地无监督地将合成图像的高级属性(姿势)和随机变化属性(雀斑、头发)分离,是目前最先进的高分辨率图像合成方法之一,被证明可以应用在多种原始数据集上,并合成逼真的伪造图像. 具体的,我们将CelebA-HQ标准图像库中的10 000幅自然图像作为正样本,并输入StyleGAN中训练得到模型,用这个模型随机生成了10 000幅GAN合成图像作为图像库中的负样本. 其中StyleGAN的参数设置为:截断参数truncation_psi = 0.7, truncation_cutoff = 8,重新随机化噪声输入randomize_noise = True等一系列默认参数. 同时,为了提高模型的训练效率,我们将来自于CelebA-HQ图像库的真实样本裁剪为分辨率为512×512的原始图像. 利用StyleGAN模型,我们生成了相同尺寸,即分辨率为512×512的GAN合成图像. 此外,为了进一步丰富实验样本,我们使用FFHQ和LSUN标准数据集作为原始真实图像,利用StyleGAN2和PGGAN作为生成伪造图像的模型. 若无特别描述,在本文实验中使用6 000对自然图像和GAN合成图像作为训练集,其余4 000对自然图像和GAN合成图像构成测试集,实验数据配置如表1所示. 另外,本文采用准确率ACC作为评估指标.

4.1 实验结果

首先,本实验对R, G, V, Y四个色彩通道进行了特征提取,并将四个通道的特征联合起来得到共计2 744维特征向量,放入集成分类器中进行训练和验证. 本文所提方法和其他方法比较结果如表2所示,其中Co-occurrence[26]、Li[27]代表基于手工提取特征的方法,VE[39]、Capsule[40]、Two-step[21]代表基于DNN提取特征的方法,训练集和测试集的比例都是6 000:4 000. 实验结果表明,本文提出的方法、Capsule、Li和Co-occurrence的准确率能够接近或达到100%,VE和Two-step这两种方法的准确率在90%左右,表明使用这些方法均可以在一定程度上准确地识别自然图像和GAN合成图像. 从准确率来说,基于手工提取特征的方法堪比基于DNN提取特征的方法. 另外,如表2所示,我们还比较了提取特征、训练及测试所需消耗的时间. 由于我们的方法和Li的方法都是轻量级的方法,完成所有过程所需的时间较短,进行测试和训练均只需几秒,而Co-occurrence方法由于训练测试过程是在神经网络中进行的,耗时较长. VE方法虽然是基于DNN提取特征的方法,但由于使用的是浅层网络,所需时间较短,但其准确率相对不高,Capsule和Two-step这两种基于DNN的特征提取是融合在训练过程中的,因此在表格中我们把该过程所需时间划分为训练时间,可见这两种方法的消耗时间均较长. 此外,为了验证本文所提出

图5　正样本示例图(来自CelebA-HQ的真实自然图像)

图6　负样本示例图(来自GAN合成伪造图像)

表1　实验数据配置

正样本来源	CelebA-HQ[8], FFHQ[13], LSUN[41]
负样本来源	StyleGAN[13], StyleGAN2[14], PGGAN[8]
色彩通道	R, G, V, Y
图像尺寸	512×512
图像格式	JPEG
正负样本对数量	10 000
比较方法	Capsule[40], VE[39], Two-step[21], Co-occurrence[26], Li[27]

表2　本文的方法和其他方法比较

方法	本文	Co-occurrence[26]	Li[27]	Capsule[40]	VE[39]	Two-step[21]
ACC/%	100.00	97.20	99.89	99.98	90.30	87.07
特征提取时间/s	5472	1653	2733	—	4458	—
训练时间/s	6	24 633	2	17 523	20	82 080
测试时间/s	1	10 649	1	916	1 895	861

框架的可扩展性,我们提出采用另外两种经典的隐写分析特征SRM[28]和ccJRM[30],搭载本文提出的集成分类器,完成自然图像和GAN合成图像的鉴别任务. 在相同的实验环境配置下,实验结果表明,检测准确率同样可以达到100.00%. 值得注意的是,采用SPAM特征、SRM特征、ccJRM特征均可以达到近乎完美的检测准确率. 但是,针对单一通道,SRM特征维度是34 671,ccJRM特征维度是22 510,而本文采用的SPAM特征仅为686. 即使联合四个色彩通道,总共特征维度为2 744,远远小于对比的两类特征向量. 因此,采用本文所提出的检测方法可以显著地降低计算复杂度.

4.2　TSNE可视化

t分布随机邻域嵌入(t-distributed Stochastic Neighbor Embedding,TSNE)是目前效果最好的数据降维与可视化方法之一,其基本思想是:如果把高维空间相似的数据点映射到低维空间距离,它们也是相似的. TSNE把这种距离关系转换为一种条件概率来表示相似性. 当我们想要对高维数据进行分类,又不清楚这个数据集有没有很好的可分性时,可以通过TSNE投影到2维或者3维的空间中进行观察. 如果在低维空间中具有可分性,则数据就是可分的;如果在高维空间中不具有可分性,可能数据不可分,也可能数据无法投影到低维空间. 在这组实验中,我们对本文使用的差分像素邻接矩阵特征进行了TSNE可视化. 结果如图7所示,我们总共选取了10 000对正负样本进行实验,红色和蓝色分别表示GAN合成图像和自然图像,从图中我们可以清楚的分辨出两类样本特征的分布情况. 该实验证明

了本文使用的特征对自然图像和GAN合成图像的分类是非常有效的.

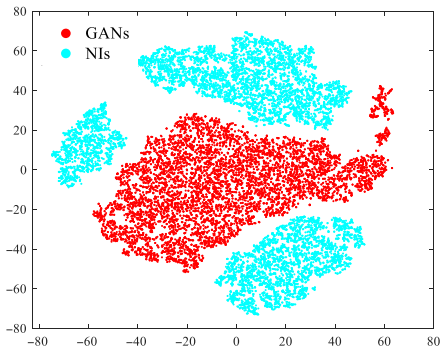


图7 本文提出特征的TSNE可视化结果

4.3 小样本训练及检测结果

利用少量标注数据学习具有一定泛化能力的模型尤为重要[42,43]. 因此,在减少训练样本数的情况下对本文方法进行评估. 由于神经网络需要大量数据进行训练,本组实验仅和SRM、ccJRM、Li的特征放入集成分类器的情况进行对比. 图8和表3展示了不同数量、不同特征训练图像样本对的验证结果. 值得一提的是,表3中"本文 + SRM" 表示采用本文所提出的检测框架,同时搭载SRM隐写分析特征向量;"本文+ ccJRM" 表示采用本文所提出的检测框架,同时搭载ccJRM隐写分析特征向量. 本组实验图像库包含10 000对正负样本,训练集和测试集比例6:4. 在小样本训练和测试中,训练样本在6 000对正负样本中随机选取,每种情况随机选取10次,得到10个检测模型. 测试样本采用除训练样本之外的4 000对正负样本. 其中每种情况下所得结果是10次测试结果的平均值. 我们采用σ对检测器的稳定性和可靠性进行综合评估. σ代表测试结果的标准差,σ越小,说明结果越稳定越可靠. 实验结果表明,本文所提出的方法在训练样本对为10以上时,准确率基本能达到99.96%以上,而训练样本对为5或2时,准确率也没有明显下降,维持在99.65%以上. 以上结果表明,本文提出的方法在训练数据很少的情况下,仍能保持非常高的检测性能. 在本文提出的检测框架下,SRM特征和本文特征的性能相当,而Li的准确率在样本只有10对的情况下下降到87.56%,在只有2对样本的情况下准确率仅为61.19%,ccJRM的准确率在样本只有2对的情况下下降到90.28%. 另外,ccJRM和Li方法具有较大的σ值,分类器的稳定性相对较差. 然而,不论采用本文特征或是SRM特征,检测结果的稳定性接近,且都维持较小的标准差. 尤其是本文所提出方法,在只有2对训练样本时σ仅为0.34%,这在一定程度上证明了我们的方法具有较好的稳定性. 需要特别强调的是,SRM的特征维度远大于本文所采用SPAM的特征维度,证明本文所提出的方法具有一定的优势.

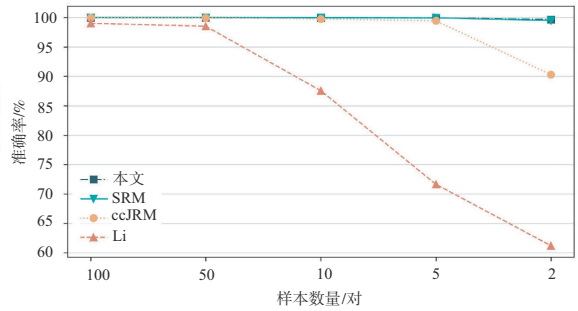


图8 不同数量小样本训练的准确率

4.4 单类样本训练及检测结果

图像分类在实际应用中可能会出现只有一种类型的标注样本,或者存在两类样本且一类样本数目远少于另一类型样本的非均衡情况[44]. 如果此时采用二元分类器,训练集正负样本不均衡,可能造成分类器过于偏向数目多的样本类别. 本小节提出在单类样本训练情况下,比较分类器的检测结果. 实际场景中,人们得到真实图像要比得到GAN合成图像容易得多,而且GAN合成图像很难确定其来源,因此我们仅使用真实图像提取的特征送入单类SVM分类器训练检测模型,通过网格搜寻最优参数. 在测试阶段,我们使用正负两类样本进行分类测试. 得到的分类结果如表4所示,本文所提出的方法在只有5 000和500个正类样本的情况下仍然表现良好,能够实现超过95.07%以上的准确率. 随着样本数量的逐渐减少,准确率逐渐降低,在仅有5个正样本的情况下,准确率下降到57.25%,说明本文特征在单类样本数据量足够的情况下能够实现自然图像和GAN合成图像的有效分类. 本文+ SRM、本文+ccJRM、Li方法在只有一种类型样本的情况下,准确率

表3 不同数量小样本训练的准确率 单位:%

训练对数	100	50	10	5	2
本文	99.98±0.02	99.98±0.03	99.96±0.03	99.94±0.06	99.65±0.34
本文+ SRM[28]	99.99±0.02	99.99±0.02	99.97±0.04	99.93±0.07	99.49±2.25
本文+ ccJRM[30]	99.97±0.03	99.96±0.04	99.75±0.17	99.43±0.72	90.28±9.35
Li[27]	99.01±0.43	98.51±0.37	87.56±4.53	71.64±5.84	61.19±7.31

只有50.00%. 因此在只有真实图像可以进行训练的情况下,利用本文提出的方法来识别GAN合成图像可达到较高的检测率. 此外,为了进一步说明我们所提出检测器的优良性能,在仅有负样本训练的情况下,比较分类结果. 如表5所示,本文所提方法同样展示了较高的检测准确率.

表4　单类训练样本的准确率(仅使用正样本训练) 单位:%

训练个数	5000	500	50	10	5
本文	95.67	95.07	92.84	75.56	57.25
本文 + SRM[28]	50.00	50.00	50.00	50.00	50.00
本文 + ccJRM[30]	50.00	50.00	50.00	50.00	50.00
Li[27]	50.00	50.00	50.00	50.00	50.00

表5　单类训练样本的准确率(仅使用负样本训练) 单位:%

训练个数	5000	500	50	10	5
本文	95.40	95.02	93.50	87.77	87.47
本文 + SRM[28]	50.69	50.00	50.00	50.00	50.00
本文 + ccJRM[30]	50.00	50.00	50.00	50.00	50.00
Li[27]	51.38	50.01	50.00	50.00	50.00

4.5　多种GAN模型检测结果

如同2.1节所述,GAN在博弈下不断提高建模能力,合成图像的质量也不断得到改进,最终实现以假乱真的图像合成[44]. 近年来,GAN模型不断优化,其种类也越来越丰富. 在这组实验中,针对多种不同的GAN模型生成的图像,在单一色彩通道和融合色彩通道上,我们对检测器性能进行了全面评估和比较. 实验结果如表6所示,其中第2行表示基于CelebA-HQ数据集,使用StyleGAN模型生成图像得到的检测结果;第3行表示基于FFHQ[13]数据集,使用StyleGAN2模型生成图像得到的检测结果;第4行表示基于LSUN[41]数据集,使用StyleGAN2模型生成图像得到的检测结果;第五行表示基于CelebA-HQ数据集,使用PGGAN模型生成图像得到的检测结果,其中正样本和负样本的数量都是10 000幅. 实验结果表明,本文提出的方法不仅可以识别StyleGAN合成的伪造图像,针对其他种类GAN模型生成的图像也能达到很高的准确率,而且只采用其中一个通道的特征进行识别,准确率达到98.81%以上. 对于现在最先进的StyleGAN2模型,本文所提出的方法对FFHQ作为原始数据集的准确率能达到99.98%,对于非人脸的LSUN数据集作为原始数据集,准确率也能达到99.96%. 整体来说,融合四个通道的特征检测性能要优于单通道,说明我们将四个通道特征融合是有必要的. 我们的方法不仅能够识别GAN生成的人脸图像,也可以识别GAN生成的其他类型图像. 因此本文的方法对GAN合成图像的识别具有一定的普适性.

表6　多种GAN模型生成图像识别的准确率　　单位:%

色彩通道	R	G	V	Y	R+G+V+Y
StyleGAN (CelebA-HQ)	99.88	99.99	99.86	99.97	100.00
StyleGAN2 (FFHQ)	99.76	99.44	99.66	98.81	99.98
StyleGAN2 (LSUN)	99.84	99.78	99.68	99.92	99.96
PGGAN (CelebA-HQ)	99.93	99.95	99.96	99.97	99.98

5　结论

本文提出了一种基于色彩通道选择特征融合的GAN合成图像识别方法. 通过对不同的色彩通道进行分析,选择最优的四个通道提取图像的差分像素邻接矩阵特征并融合. 该特征利用自然图像的对称性来降低维度,并使用二阶马尔科夫链来模拟自然图像中相邻像素之间的差异. 通过对自然图像相邻像素之间的差异建模,从而有效区分自然图像和GAN合成图像. 最后,本文使用该特征训练集成分类器,从而实现对自然图像和GAN合成图像的二分类. 我们在不同的训练条件下,对提出的检测方法进行全面且细致的评估. 实验结果表明,在训练样本充足的情况下,本文的方法能够准确地区分自然图像和GAN合成图像,对StyleGAN模型生成的图像识别的准确率高达100.00%;在训练样本只有2对的极端情况下,同样可以有效识别自然图像和StyleGAN模型生成的图像,准确率达到99.65%;在负样本未知,仅使用正样本训练且个数仅为50的情况下,提出的单类分类器在识别两类图像时,准确率也能达到92.84%以上,该实验也充分证明了我们所提出的方法能够有效识别出偏离自然图像模型的偏差. 另外,本文的方法对StyleGAN2和PGGAN模型生成的伪造图像,识别准确率也能达到99.96%以上,证明我们所提出的方法具有普适性. 需要特别强调的是,本文提出的基于传统机器学习框架的检测方法在保证准确率的同时,显著提高了计算效率.

参考文献

[1] 梁瑞刚, 吕培卓, 赵月, 等. 视听觉深度伪造检测技术研究综述[J]. 信息安全学报, 2020, 5(2): 1-17.
LIANG R G, LV P Z, ZHAO Y, et al. A survey of audiovisual deepfake detection techniques[J]. Journal of Cyber Security, 2020, 5(2): 1-17. (in Chinese)

[2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[3] GU J X, WANG Z H, KUEN J S, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[4] PEVNÝ T, BAS P, FRIDRICH J. Steganalysis by subtractive pixel adjacency matrix[C]//Proceedings of the 11th ACM workshop on Multimedia and security. New York: ACM, 2009: 75-84.

[5] DENTON E, CHINTALA S, SZLAM A, et al. Deep generative image models using a Laplacian pyramid of adversarial networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. New York: ACM, 2015: 1486-1494.

[6] ZHANG H, XU T, LI H S, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 5907-5915.

[7] ZHANG H, XU T, LI H S, et al. StackGAN++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947-1962.

[8] KARRAS T, AILA T M, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB/OL]. (2017-10-27)[2022-05-20]. https://arxiv.org/abs/1710.10196v2.

[9] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein GANs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc. 2017: 5769-5779.

[10] ZHANG Z Z, XIE Y P, YANG L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6199-6208.

[11] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis [EB/OL]. (2018-09-28) [2022-05-20]. https://arxiv. org/abs/1809.11096.

[12] LIN C H, CHANG C C, CHEN Y S, et al. COCO-GAN: Generation by parts via conditional coordinating[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4512-4521.

[13] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 4401-4410.

[14] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 8110-8119.

[15] 何沛松, 李伟创, 张婧媛, 等. 面向GAN生成图像的被动取证及反取证技术综述[J]. 中国图象图形学报, 2022, 27(1): 88-110.
HE P S, LI W C, ZHANG J Y, et al. Overview of passive forensics and anti-forensics techniques for GAN-generated image[J]. Journal of Image and Graphics, 2022, 27(1): 88-110. (in Chinese)

[16] 郑远攀, 李广阳, 李晔. 深度学习在图像识别中的应用研究综述[J]. 计算机工程与应用, 2019, 55(12): 20-36.
ZHENG Y P, LI G Y, LI Y. Survey of application of deep learning in image recognition[J]. Computer Engineering and Applications, 2019, 55(12): 20-36. (in Chinese)

[17] MO H X, CHEN B L, LUO W Q. Fake faces identification via convolutional neural network[C]//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2018: 43-47.

[18] DANG L, HASSAN S, IM S, et al. Deep learning based computer generated face identification using convolutional neural network[J]. Applied Sciences, 2018, 8(12): 2610.

[19] ZHANG X, KARAMAN S, CHANG S F. Detecting and simulating artifacts in GAN fake images[C]//2019 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE, 2019: 1-6.

[20] YU N, DAVIS L, FRITZ M. Attributing fake images to GANs: Learning and analyzing GAN fingerprints[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 7556-7566.

[21] ZHUANG Y X, HSU C C. Detecting generated image based on a coupled network with two-step pairwise learning[C]//2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2019: 3212-3216.

[22] FU Y, SUN T F, JIANG X H, et al. Robust GAN-face detection based on dual-channel CNN network[C]//2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Piscataway: IEEE, 2019: 1-5.

[23] 李旭嵘, 纪守领, 吴春明, 等. 深度伪造与检测技术综述[J]. 软件学报, 2021, 32(2): 496-518.
LI X R, JI S L, WU C M, et al. Survey on deepfakes and

detection techniques[J]. Journal of Software, 2021, 32(2): 496-518. (in Chinese)

[24] MCCLOSKEY S, ALBRIGHT M. Detecting GAN-generated imagery using saturation cues[C]//2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2019: 4584-4588.

[25] NATARAJ L, MOHAMMED T M, MANJUNATH B S, et al. Detecting GAN generated fake images using co-occurrence matrices[J]. Electronic Imaging, 2019, 31(5): 532-1-532-7.

[26] BARNI M, KALLAS K, NOWROOZI E, et al. CNN detection of GAN-generated face images based on cross-band Co-occurrences analysis[C]//2020 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE, 2020: 1-6.

[27] LI H D, LI B, TAN S Q, . Identification of deep network generated images using disparities in color components [J]. Signal Processing, 2020, 174: 107616.

[28] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[29] MARRA F, GRAGNANIELLO D, COZZOLINO D, et al. Detection of GAN-generated fake images over social networks[C]//2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Piscataway: IEEE, 2018: 384-389.

[30] KODOVSKÝ J, FRIDRICH J. Steganalysis of JPEG images using rich models[C]//Proc SPIE 8303, Media Watermarking, Security, and Forensics 2012. Burlingame: SPIE, 2012: 81-93.

[31] GAN Y F, YANG J X, LAI W D. Video object forgery detection algorithm based on VGG-11 convolutional neural network[C]//2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS). Piscataway: IEEE, 2019: 575-580.

[32] 苏兆品, 吴张倩, 岳峰, 等. 自然环境背景噪声下基于低维深度特征的手机来源识别[J]. 电子学报, 2021, 49(4): 637-646.
SU Z P, WU Z Q, YUE F, et al. Source cell-phone identification under background noise based on low-dimensional deep features[J]. Acta Electronica Sinica, 2021, 49(4): 637-646. (in Chinese)

[33] SERFOZO R. Basics of Applied Stochastic Processes [M]. Berlin: Springer, 2009.

[34] BONNIER N. Contribution to Spatial Gamut Mapping Algorithms[D]. Paris: Telecom Paristech, 2008.

[35] CASTLEMAN K R. Digital Image Processing[M]. Englewood Cliffs: Prentice Hall, 1996.

[36] CHANG C C, LIN C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2(3): 1-27.

[37] KODOVSKY J, FRIDRICH J, HOLUB V. Ensemble classifiers for steganalysis of digital media[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 432-444.

[38] HART P E, STORK D G, DUDA R O. Pattern Classification[M]. Hoboken: Wiley, 2000.

[39] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE, 2019: 83-92.

[40] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 2307-2311.

[41] YU F, ZHANG Y D, SONG S R, et al. Construction of a large-scale image dataset using deep learning with humans in the loop[EB/OL]. (2015-06-10) [2022-04-20]. https://arxiv.org/abs/1506.03365.

[42] 潘兴甲, 张旭龙, 董未名, 等. 小样本目标检测的研究现状[J]. 南京信息工程大学学报(自然科学版), 2019, 11(6): 698-705.
PAN X J, ZHANG X L, DONG W M, et al. A survey of few-shot object detection[J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2019, 11(6): 698-705. (in Chinese)

[43] 葛轶洲, 刘恒, 王言, 等. 小样本困境下的深度学习图像识别综述[J]. 软件学报, 2022, 33(1): 193-210.
GE Y Z, LIU H, WANG Y, et al. Survey on deep learning image recognition in dilemma of small samples[J]. Journal of Software, 2022, 33(1): 193-210. (in Chinese)

[44] TAX D M J, DUIN R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11/12/13): 1191-1199.

[45] 乔通, 姚宏伟, 潘彬民, 等. 基于深度学习的数字图像取证技术研究进展[J]. 网络与信息安全学报, 2021, 7(5): 13-28.
QIAO T, YAO H W, PAN B M, et al. Research progress

of digital image forensic techniques based on deep learning[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 13-28. (in Chinese)

作者简介

乔 通 男,1986年10月出生于河南省新乡市. 现为杭州电子科技大学网络空间安全学院副教授、硕士生导师. 主要研究方向为数字图像取证、信息隐藏. 主持多项国家级、省部级项目,首批“浙江省高校领军人才培养计划”-青年优秀人才,在国内外发表学术论文60余篇.
E-mail: tong.qiao@hdu.edu.cn

陈彧星 女,1998年8月出生于江西省新余市. 2020年本科毕业于杭州电子科技大学网络空间安全学院,现为杭州电子科技大学网络空间安全学院硕士,主要研究方向为多媒体信息安全、人工智能安全.
E-mail: cyx299@hdu.edu.cn

谢世闯 男,1997年10月出生于河南省驻马店市. 2020年本科毕业于河南大学计算机科学与技术专业,现为杭州电子科技大学网络空间安全学院硕士,主要研究方向为多媒体信息安全、人工智能安全.
E-mail: shichuang_xie@hdu.edu.cn

姚 恒 男,1982年9月出生于安徽省芜湖市. 现为上海理工大学光电信息与计算机工程学院副教授,硕士生导师. 主要研究方向为多媒体安全、图像处理和模式识别,主持了两项国家自然科学基金项目,在国内外发表学术论文60余篇.
E-mail: hyao@usst.edu.cn

罗向阳 男,1978年2月出生于湖北省钟祥市. 现为信息工程大学教授、博导,主要研究方向为网络与信息安全,在国内外发表学术论文200余篇.
E-mail: luoxy_ieu@sina.com