文章编号: 1003-0530(2021)12-2235-16

面向图像视频取证的机器学习综述

谭舜泉[1,2,3] 黎思力[1,2,3] 陈保营[1,2,3] 李 斌[1,3]

(1. 深圳市媒体信息内容安全重点实验室, 广东省智能信息重点实验室, 广东深圳 518060;
2. 深圳大学, 计算机与软件学院, 广东深圳 518060; 3. 深圳人工智能与机器人研究院, 广东深圳 518060)

摘 要: 近年来, 随着机器学习技术, 特别是深度学习技术的飞速发展, 使得一般人也能够生成非常逼真的高质量造假图像和视频。这给社会和个人带来了极大的风险, 也引起了世界各国相关部门以及学术界的高度重视。针对图像和视频的篡改技术和取证技术是相互对抗相互促进的矛盾双方。机器学习技术的飞速发展, 同样地也触发了图像/视频取证技术的跨越式演化。本文对近年来, 特别是过去三年面向图像/视频取证的机器学习技术的飞速发展现状进行了综述, 展示了基于传统人工构造特征以及端到端的图像视频取证机器学习方法, 并探讨了不同检测技术的优缺点, 重点对 Deepfake 换脸视频的取证技术以及基于深度学习的取证与反取证的对抗进行了介绍。对现有的科研工作进行了科学的归类。最后对其未来的发展趋势进行了展望, 旨在为后续学者的研究进一步推动图像/视频取证的机器学习技术提供指导。

关键词: 图像视频取证; 机器学习; 深度学习; 反取证; 对抗样本

中图分类号: TP391.4 **文献标识码**: A **DOI**: 10.16798/j.issn.1003-0530.2021.12.001

引用格式: 谭舜泉, 黎思力, 陈保营, 等. 面向图像视频取证的机器学习综述[J]. 信号处理, 2021, 37(12): 2235-2250. DOI: 10.16798/j.issn.1003-0530.2021.12.001.

Reference format: TAN Shunquan, LI Sili, CHEN Baoying, et al. A survey of deep learning in image and video forensics [J]. Journal of Signal Processing, 2021, 37(12): 2235-2250. DOI: 10.16798/j.issn.1003-0530.2021.12.001.

A Survey of Deep Learning in Image and Video Forensics

TAN Shunquan[1,2,3] LI Sili[1,2,3] CHEN Baoying[1,2,3] LI Bin[1,3]

(1. Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Shenzhen, Guangdong 518060, China; 2. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China; 3. Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, Guangdong 518060, China)

Abstract: In recent years, with the rapid development of machine learning technologies, especially deep learning technologies, even normal people can produce vivid and high-quality forged images and videos, which introduces great risk to our society and brings great attention of governments and scholars. Image/video forgery technologies and the corresponding forensics technologies are the two aspects in a contradiction. Also with the rapid development of machine learning technologies, the evolution of image/video forensics technologies are ongoing. In this essay, the latest development of image/video forensics oriented machine learning technologies is surveyed. The machine learning methods based on traditional handcrafted features and end to end methods are introduced. We discussed the advantages and disadvantages of different detection technologies, focusing on forensics technologies targeted at Deepfake face-transplant videos and deep learning based confrontation between forensics and counter forensics. The existing scientific research work has been scientifically classified. In

收稿日期: 2021-06-18; 修回日期: 2021-09-29

基金项目: 广东省重点领域研发计划项目(2019B010139003); 国家自然科学基金(U19B2022, 61772349, 61872244); 广东省自然科学基金(2019B151502001); 深圳市基础研究项目(JCYJ20180305124325555)

the end, this essay further outlines future research directions, aiming to provide guidance for follow-up scholars to further promote the machine learning technology of image/video forensics.

Key words: image/video forensics; machine learning; deep learning; counter forensics; adversarial examples

1 引言

近年来,随着机器学习技术的迅速发展,特别是深度学习技术,越来越多的机器学习技术应用到图像和视频篡改上,使用这些技术能够生成非常逼真,高质量的造假图片和视频,比如利用对抗生成网络(Generative Adversarial Network, GAN)合成伪造图片或者 Deepfake 换脸伪造视频。当这些造假的图片或视频被用来制作假新闻传播时,往往给社会或个人带来极大的风险。针对数字媒体,尤其是图像和视频取证技术的研究,作为涉及国家政治、军事、经济、社会和文化的重要战略性问题,近年来引起了世界各国相关部门以及学术界的高度重视。

针对图像及视频的篡改技术和取证技术是相互对抗相互促进的矛盾双方。早期的图像/视频取证技术,集中于发掘篡改导致的图像和视频场景中违反常理的失真,如阴影和光照角度等。这样的取证技术,往往严重依赖于取证专家的专业知识和经验,而无需借助高级的机器学习技术。但近十年来,随着一些高效的图像/视频修改编辑工具,如 PhotoShop、After Effects、GIMP 等的普及化,以及高级机器学习技术,使得图像/视频篡改技术的复杂程度和逼真程度都飞速发展。这时候与之相对抗的图像/视频取证技术就必须要倚仗于高维的取证特征空间,和与之相匹配的高级机器学习技术了。

在此期间,大数据及深度学习技术的出现,无论对于针对图像和视频的篡改技术和取证技术,都是武器库的一次彻底的更新换代。在大数据及深度学习赋能下,只需结合少量的取证专家知识,针对图像和视频的篡改技术和取证技术都能显著的以一种接近于端到端的形式提升自己的性能。其中,在基于深度学习的图像/视频取证框架中,按照是否需要预处理可分为两大分支:(1)手工特征与深度学习技术相结合的取证方法,该方法首先需要经过预处理提取手工特征;(2)无需预处理端到端的深度学习取证方法。按照取证任务主要分为两大任务:(1)篡改检测;(2)篡改定位。其中篡改检测属于分类任务,而篡改定位属于分割任务。在这一阶段,矛盾双方的性能优势更多的是和其中某一方所拥有的数据和模型计算能力有关。由于深度学习技术本身的可解释性缺失,某一方所取得的性能优势受多种因素制约,其鲁棒性往往存疑。但无论如何,基于深度学习的图像/视频取证技术已经在性能上占据绝对的统治地位。

本文试图为近年来,特别是过去三年的面向图像/视频取证的机器学习技术给予一个及时的综述,并基于我们对于本领域的把握和理解,给出在深度学习等新一代机器学习技术赋能下的图像/视频取证未来发展路向的展望,与各位读者探讨。本文先在第2节讨论基于传统人工构造特征的图像/视频取证机器学习方法;然后在第3节介绍近年来已经占据事实上统治地位的基于深度特征的端到端图像视频取证机器学习方法;在第4节,我们着重讨论在大数据及深度学习赋能下的图像/视频取证与反取证的对抗;最后,我们在第5节归纳面向图像/视频取证的机器学习未来重要研究问题,并在第6节对全文做出总结。

2 基于传统人工构造特征的图像视频取证机器学习方法

本章节主要介绍基于人工构造特征的传统机器学习方法和深度学习方法,其中包括有监督学习、One-class 弱监督学习和有监督 Deepfake 检测方法。其中表1总结了有监督学习和 One-class 弱监督学习的方法,表2总结了有监督 Deepfake 检测方法。

2.1 有监督学习方法

本小节主要介绍基于人工构造特征(也称手工特征)的有监督取证方法,分为基于手工特征的传统机器学习方法和基于手工特征的深度学习方法。合适的手工特征能够提高原始图像和篡改图像的区分度,然后利用大量的有标签数据集训练出一个分类器。然而,提取手工特征往往需要对篡改图片或视频有较为深入的理解和分析。

在深度学习兴起之前,大多数取证方法是基于传

统的机器学习算法,同时这些方法大多需要结合手工特征[1-2]。一些手工特征往往是针对特定的伪造痕迹而设计的,特别是由 JPEG 重压缩导致的或者与相机反应函数(Camera Response Function, CRF) 相关的。

当一个 JPEG 压缩的图像经过局部的篡改后,除篡改区域以外,整个图像都会进行重压缩而留下重压缩失真。早期的图像取证方法往往利用重压缩失真相关的手工特征对 JPEG 图像进行篡改检测和定位,并使用支持向量机(Support Vector Machine, SVM) 作为二分类器[2]。同样的,视频重压缩也会留下失真。Jiang 等人[3]通过分析重压缩视频发现: 在使用相同编码参数进行多次重压缩之后,视频质量趋于几乎不变,因此可以利用质量退化程度来区分单压缩视频和双压缩视频。根据这一线索,在视频帧内编码过程中提取舍入和截断误差的统计特征,在视频帧间编码过程中提取基于宏块模式的特征,最后将两组特征拼接输入到径向基函数(Radial Basis Function, RBF) 内核函数 SVM 分类器(RBF-SVM) 中训练和检测。

相机反应函数(CRF) 是一种将输入辐照度(Input Irradiance) 映射到输出图像强度(Output Image Intensity) 的能表征相机基本属性的隐函数。Hsu 等人[4]首先将测试图像自动分割为不同的任意形状的区域,然后使用来自局部平面辐照点(Locally Planar Irradiance Points, LPIP) 的几何不变量,从每个区域估计一个 CRF。为了将两个区域之间的边界段分类为真实的或拼接的,首先计算出基于 CRF 的交叉拟合(CRF-based cross fitting) 特征和局部图像特征,然后将两种特征输入到 SVM 分类器得到分割级别结果,最后将分割级别结果进一步推断出图像级别结果。

除此以外,移除图片语义内容引入的噪声残差(Noise Residual) 特征也被广泛应用于媒体取证。Li 等人[5]使用隐写分析富集模型(Steganalysis Rich Model, SRM) [6]提取空域噪声残差手工特征对 11 种典型的图像处理操作和四种反取证操作进行检测。他们使用由多个线性判别分析(Linear Discriminant Analysis) 分类器组合而成的集成分类器来检测图像处理操作和反取证操作,同时使用带有高斯核的 SVM 作为分类器对论文提取方法进行鲁棒性分析实验,该分类器的参数 C 和 γ 是通过网格搜索获得。

特别地,将图像匹配技术应用于图像取证也有一定成效。李等人[7]结合并 SIFT 图像匹配算法和 LPP 降维算法,提出了一种高效的复制-移动篡改检测算法。该算法在 SIFT 算法的基础上,使用 LPP 算法对特征点以及特征向量进行降维,最后通过凝聚型层次聚类算法对相似的特征点进行聚类,从而完成对图像复制-移动的篡改区域的检测。相似的,于等人[8]提出一种自适应提取特征点、快速匹配和鲁棒性篡改区域定位的复制-移动篡改检测算法。该算法第一步将图像分割成不重叠的块, 构造波动函数, 根据波动函数在图像内均匀提取 SURF 特征点; 第二步使用 SURF 特征进行特征表示; 第三步引入双比特量化局部敏感哈希(DBQ-LSH) 快速匹配特征; 第四步去除孤立匹配, 利用 K-means 聚类方法对匹配点对进行聚类, 将灰度图像转化为不变矩 LBP 图像, 从而定位篡改区域。

随着深度学习技术的发展,越来越多方法利用了深度学习技术进行端到端的图像视频取证,但是结合手工特征非端到端的深度学习技术在图片视频取证上发挥重要的作用,其中图像重压缩的手工特征或者图像的噪声残差特征被广泛应用到基于深度学习取证方法中。Wang 等人[9]提出一种基于卷积神经网络(Convolutional Neural Network, CNN) 的图像双重 JPEG 压缩检测算法,该方法设计了一个简单的 CNN(包含两个卷积层,两个池化层和三个全连接层) 作为分类器,然后提取 DCT 系数直方图手工特征作为 CNN 的输入。为了检测篡改区域,该方法使用重叠滑动窗口遍历得到整个图片篡改区域的预测结果。相似的,Park 等人[10]也提出一种基于 CNN 的图像双重 JPEG 压缩检测算法,该算法提取每个 JPEG 块的 DCT 系数直方图特征和量化表(Quantization table) 这两种手工特征作为 CNN 的输入,最后也通过滑动窗口方式得到篡改区域。该算法使用的 CNN 由四个卷积层,三个最大池化层和三个完全连接的层组成,其中量化表向量与最后一个最大池化层和两个完全连接的层组合在一起。为了检测对齐或者非对齐的图像双重压缩,Barni 等人[11]使用 8×8 的 DCT 卷积核初始化 CNN 的第一个卷积层,用来提取类似 DCT 系数直方图手工特征,这不同于前面通过预处理方式提取手工特征。

同时他们也提出一种基于噪声残差特征和CNN相结合的图像双重压缩检测算法,该算法提取图像噪声残差是通过计算原图和原图的去噪图像之差得到,而去噪图像是通过最小均方误差估计在小波域(Wavelet domain)中计算得到。其中第一种方法在检测对齐图像双重压缩上表现较优,而第二种方法在检测非对齐图像双重压缩上表现较优。相似的,Zhou等人[12]提出一种基于双流Faster R-CNN网络(RGB-N)实现端到端的篡改定位方法,该网络的双流之一是RGB流,其目的是从RGB图像中提取含有篡改失真的特征,另外一流是噪声流(Noise Steam),该流使用SRM滤波卷积层(SRM filter layer)提取的噪声残差手工特征来发现真实区域和篡改区域在噪声域中不一致性,其中他们只使用了三个SRM滤波内核初始化SRM滤波卷积层。

针对视频取证,Nam等人[13]把噪声残差手工特征与CNN相结合的方法扩展到视频取证上,提出一种双流神经网络,该网络包含两个分流CNN分别输入预测帧(Predictive Frames, PF)和帧内编码帧(Intra-coded Frames,IF)。对于PF分流网络,先把输入的RGB图片转换为YUV图片,然后使用高斯滤波器对YUV图片的Y通道进行低通滤波预处理提取手工特征后再输入在网络中;而对于IF分流网络,同样使用由三个SRM滤波内核初始化的SRM滤波卷积层作为预处理层来提取噪声残差特征后再输入到网络中。最后把两个分流网络的学习特征组合起来再一起输入到全连接层,得到判断视频是否经过重压缩的二分类预测结果。为了对视频中移除目标进行检测,白等人[14]提出一种基于双通道CNN的视频目标移除取证算法,该算法利用双通道CNN结构(主干网络均为Inception-v3),分别提取视频绝对帧差图像的RGB特征和噪声特征,并利用双线性池化对二者进行特征融合,最后通过全连接层分类输出视频帧的分类结果,从而有效地识别经过篡改的视频帧。

2.2 One-class弱监督图片视频取证方法

本小节主要介绍基于人工构造特征的单类(one-class)弱监督取证方法。基于人工构造特征的图像视频取证机器学习方法往往需要大量的篡改数据和未篡改原始数据,但是现实中的未篡改的原始数据是比较容易获取的,样本相对多,而篡改数据的相对比较难获取,样本少。因此,在不使用篡改数据的条件下,通过收集大量未篡改原始数据进行one-class弱监督取证也是一种选择。

当前与手工特征相结合的单类弱监督取证方法主要是依赖于相机指纹(Camera Fingerprint),而相机指纹往往是与相机传感器相关的。由于相机制造的缺陷,传感器元件与其预期性能存在微小的偏差,这种偏差形成了一种非常稳定的类噪声模式,称为光响应非均匀(Photo-response Non-uniformity, PRNU)噪声。在特定相机采集的所有图像上都有独特共有的PRNU模式痕迹。利用PRNU进行取证主要分为两步:首先需要通过计算大量相机拍摄图像的平均噪声残差进行离线估计得到相机的PRNU指纹;在测试时,通过去噪滤波器估计目标图像PRNU指纹,然后与相机的PRNU指纹进行相关性评估。基于该原理,Chakraborty等人[15]提出一种基于PRNU的判别随机场(Discriminative Random Fields, DRF)图像篡改定位算法。DRF是一种条件随机场,在马尔可夫随机场(Markov random field, MRF)标记先验条件下,DRF大大降低了检测的假阳率。PRNU是一种设备相关(Device-related)的相机指纹,相似的,Verdoliva等人[16]利用一种模型相关(Model-related)的相机指纹进行图像取证。利用该模型相关的相机指纹进行取证也分为两步:首先使用多维高斯模型(Multidimensional Gaussian Model)从原始图像块的残差噪声中训练得到相机模型相关的稠密局部描述符(Dense Local Descriptor),即模型相关的相机指纹;在测试时,以滑动窗口形式从目标图像噪声残差中提取相同的描述符,并与先前提取的相关描述符进行比较得到目标图像的篡改平滑决策图,最后通过简单的阈值控制得到二分类的篡改掩码。

不同于前面利用相机指纹的取证方法,Li等人[17]通过提取隐写分析领域的邻域像素差分矩阵(Subtractive Pixel Adjacency Matrix, SPAM)手工特征进行视频重压缩检测,其使用的分类器是基于高斯密度的单类分类器(Gaussian distribution-based One-Class Classifier, Gaussian-OCC)。该分类器只需输入未经过双重压缩的视频帧进行训练,并结合集成策略,大大提升了分类器的鲁棒性,同时在算法复杂度和检测性能上能够优于大部分的基于完全有监督的检测算法。

表 1　基于手工特征的图像视频取证方法总结

Tab. 1　Summary of image and video forensics methods based on handcrafted features

论文	分类器	手工特征特点	任务	方法类型
Jiang 等人[3]	带 RBF 核 SVM	帧内编码舍入和截断误差的统计特征+帧间编码基于宏块模式的特征	视频,重压缩检测	有监督传统机器学习方法
Hsu 等人[4]	SVM	基于相机反应函数(CRF) 的统计特征	图像,检测+定位	
Li 等人[5]	集成分类器+带高斯核 SVM	图像噪声残差域的富模型(SRM) 特征	图像,篡改检测	
李等人[7]	SIFT+LPP	LPP 对 SIFT 特征降维,图像匹配,聚类	图像,篡改检测	
于等人[8]	DBQ-LSH+K-means	提取 SURF 特征点,图像匹配,聚类	图像,篡改定位	
Wang 等人[9]	CNN	DCT 系数直方图特征	图像,重压缩检测	有监督深度学习方法
Park 等人[10]	CNN	DCT 系数直方图特征和量化表	图像,检测+定位	
Barni 等人[11]	带 DCT 卷积层的 CNN+CNN	DCT 卷积层提取类 DCT 系数直方图特征+图像噪声残差特征	图像,重压缩检测	
Zhou 等人[12]	双流 Faster R-CNN 网络(RGB-N)	图像噪声残差域的富模型(SRM) 特征+RGB	图像,篡改定位	
Nam 等人[13]	双流 CNN	Y 通道低通滤波特征+图像噪声残差域的富模型(SRM) 特征	视频,重压缩检测	
白等人[14]	双通道 FCN	绝对帧差图像的 RGB 特征和噪声特征	视频,移除帧检测	
Chakraborty 等人[15]	判别随机场(DRF)	设备相关的 PRNU 相机指纹	图像,检测+定位	One-class 弱监督机器学习方法
Verdoliva[16]	高斯模型	残差噪声中提取模型相关的失真特征	图像,篡改定位	
Li 等人[17]	高斯模型单分类器	隐写分析领域引入的 SPAM 特征	视频,重压缩检测	

2.3　基于传统人工构造特征的 Deepfake 检测方法

基于传统人工构造特征的方法主要是通过对比 Deepfake 生成的视频和原视频之间的区别,或者寻找前者违背生理逻辑的地方,然后设计相应的方法提取特征后,输入机器学习分类器进行分类。

针对 Deepfake 假视频和原视频的差异,Matern 等人[18] 发现 GAN 生成的人脸具有一定的局限性,这种局限性会产生特征性的篡改伪影,比如眼睛会产生异瞳现象(两只眼睛的瞳孔颜色不一样) 、鼻子两侧的光照不一样等,通过选择眼睛、牙齿和面部轮廓上的视觉特征,最后使用逻辑回归模型和多层感知机(Multilayer Perceptron, MLP) 进行分类。Li 等人[19] 认为在计算资源和制作时间的限制下,Deepfake 算法只能生成有限分辨率的人脸图像,而且假脸贴合到原脸时,必须使用仿射变换以匹配原脸的位置,会使生成的假脸视频造成一定的扭曲,所以直接模仿仿射变换的步骤来简化生成负样本的过程,从而大量生成负样本,并提取负样本的感兴趣区域(region of interest, ROI) 区域作为特征,训练 VGG16,ResNet50 等分类器进行分类。

针对部分 Deepfake 生成方法,利用违反人体生理的生物特征进行检测,是一种有效的手段。Li 等人[20] 发现 Deepfake 产生成的视频中几乎没有眨眼的动作,而正常人的眨眼具有特定的频率和时间。通过基于关键点(landmark) 的算法定位提取出人眼区域,再通过 VGG16 提取人眼特征,输入 LSTM 和全连接层进行判断。Ciftci 等人[21] 认为 GAN 生成的人脸虽然很逼真,但是人物的生物信号比如脉搏还是难以伪造。通过使用医学生物信号特征提取方法如光电容积脉搏波标记法(PPG) 和基于该方法的一些改进方案,来提取视频中人脸的脉搏信号,将这些信号转化为特征,再使用 SVM 进行分类。同理,Fernandes 等人[22] 则是通过测量由血流引起的面部皮肤颜色变化、前额的平均光强度的方式来提取心率特征,再使用神经常微分方程模型训练和分类。Yang 等人[23] 发现基于深度学习生成的假脸需要保持源人脸的表情,但是这个过程没法保证生成人脸和源人脸面部的 Landmark 相一致,虽然 Landmark 位置的误差不能用肉眼观察到,但可以通过面部真实部分和伪造部分的二维 Landmark 估计的头部姿势,然后训练一个 SVM 来区分真实视频和 Deepfake 视频。类似的,Agarwal 等人[24] 发现在 Deepfake 视频中,对人脸区域进行篡改,导致说话人的面部表情和头部运动模式与人物不相符,即面部表情和头部运动存在着明显的模式差异。他们通过使用开源的

工具包 OpenFace 对人物说话时面部表情和头部运动模式(总共20种面部或头部运动特征)进行建模,然后检测目标视频中的人物说话时的面部和头部动作一致性是否符合所建立的模型来判断真假。实验表明该方法能有效对抗图像压缩、大小缩放和传输噪声等带来的特征丢失,具有一定的鲁棒性。

传统人工构造的特征大多数是基于篡改视频的瑕疵或不合逻辑的地方构造的。然而这些特征很容易受到视频压缩、尺寸缩放和传输噪声等的影响,而且,Deepfake 的制作手段也会相应对抗升级,针对前人发现的瑕疵进行改进,日益完善。所以具有一定实用意义的 Deepfake 检测技术,基本上不采用人工构造特征的方式。表2总结了基于传统人工构造特征的检测模型,特点和使用的数据集。使用到的数据集分别是 UADFV[23],DeepfakeTIMIT[25],FaceForensics 即 FF[26],FaceForensics++即 FF++[27],Celeb-DF[28],Deepfake 即 DF[21] 和 MFC[29],除此之外,还包括一些未公开的自建数据集。

表2 基于手工特征的 Deepfake 检测方法总结

Tab. 2 Summary of Deepfake detection methods based on handcrafted features

论文	机器学习模型	手工特征特点	数据集
Matern 等人[18]	LR+MLP	提取篡改伪影特征	使用 CelebA 自己生成 Deepfake 数据集
Li 等人[19]	CNN	使用仿射变换制造大量合成人脸的特征	UADFV,DeepfakeTIMIT
Li 等人[20]	CNN+LSTM-RNN	设计提取视频中人眼特征,检测眨眼是否符合自然频率	使用互联网公开视频自建数据集
Ciftci 等人[21]	SVM	设计提取脉搏信号特征	FF,FF++,Celeb-DF,DF
Fernandes 等人[22]	Neural-ODE model	设计提取皮肤颜色变化特征和心率特征	DeepfakeTIMI,自建 Deepfake 数据库
Yang 等人[23]	SVM	提取面部真实部分和伪造部分的二维 Landmark 估计的头部姿势	UADFV, MFC
Agarwal 等人[24]	SVM	设计提取面部和头部的动作单元作为特征	使用 YouTube 公开的视频自建 Deepfake 数据集

3 基于深度特征的端到端图像视频取证机器学习方法

本章节主要介绍基于深度特征的端到端图像视频取证的机器学习方法,其中包括基于深度特征的有监督学习方法、one-class 弱监督方法和端到端的 Deepfake 检测方法。其中表3总结了有监督学习和 One-class 弱监督学习的方法,表4总结了端到端的 Deepfake 检测方法。

3.1 有监督学习方法

随着深度学习技术的发展,越来越多的图像视频取证方法是基于有监督的深度学习方法,特别是基于可学习 CNN 端到端地提取深度特征的检测方法,这些方法的检测性能往往优于基于人工构造特征的传统机器学习方法。其中部分方法是针对检测特定的篡改方法设计的,比如拼接(Splicing)、复制-移动(Copy-move)、修复(Inpainting)和 Photoshop(PS)脚本处理等篡改方法;另外一部分方法是在不考虑特定的篡改方法的条件进行端到端的图像视频取证。

针对拼接图像取证,Salloum 等人[30] 提出了一种利用全卷积网络(Fully Convolutional Network, FCN)来进行拼接图像的篡改定位的算法,该算法包含一种基于双分支(Two-branch)的多任务全卷积网络(Mutil-task FCN, MFCN),该网络除了使用一个分支进行常规的定位篡改区域任务外,还额外增加一个分支进行篡改轮廓的检测。同样的,为了对拼接图像进行篡改定位,Bi 等人[31] 提出一种环状残差 U-Net(Ringed Residual U-Net, RRU-Net),RRU-Net 的核心思想是加强 CNN 的学习方式,该网络的设计是受人脑的回忆和巩固机制的启发,通过 CNN 中残差的传播和反馈过程来实现。

针对复制-移动图像取证,Wu 等人[32] 提出一种基于 CNN 端到端的复制-移动篡改区域检测方法,该方法通过 CNN 实现传统复制-移动检测方法的三个主要步骤:特征提取、特征匹配和后处理,然后再经过由反卷积网络(Deconvolutional Network)组成的伪造掩码解码器(Fogery Mask Decoder)得到篡改区域的预测结果。相比于多步骤的传统检测方法,该方法能够实现端到端的训练,同时取得更优的检测性能和更强的鲁棒性。相似的,Zhong 等人[33] 提出一种基于 Dense-InceptionNet 的端到端复制-移动伪造检测方法,该网络由金字塔特征提取器(Pyramid Feature Extractor, PFE)、特征相关匹配(Feature Correlation Matching, FCM)和分层后处理(Hierarchical Post-Processing, HPP)模块组成。其中,PFE 模块用来提取多维和多尺度的密集特征;FCM 模块用来学习深度特征的高相关性并获得三个候选匹配映射;

最后 HPP 模块充分利用了三个候选匹配映射来获得交叉熵的组合,从而便于通过反向传播进行更好地训练网络。该检测方法能够抵御大多数已知攻击方法。为了进一步区分复制-移动图像的源和目标区域,Wu 等人[34] 提出一种对复制-移动图像源和目标区域进行定位的网络 BusterNet,该网络包含一个双分支模块和一个融合模块,这两个分支分别通过视觉失真定位潜在的操纵区域,并通过视觉相似性定位复制-移动区域,最后再融合模块合并两个分支的特征来预测区分原始、源拷贝和目标拷贝类的复制-移动像素级掩码。相似的,李等人[35] 提出一种基于条件生成对抗网络(conditional Generative Adversarial Networks,cGANs) 的可区分源和目标区域的复制-移动伪造检测方法,该方法通过优化 cGANd 的损失函数和使用弱监督样本提高算法的性能。

针对修复图像视频取证,Li 等人[36] 提出基于图像残差的高通全卷积网络(high-pass Fully Convolutional Network, HP-FCN) 的检测方法,他们设计了一个可学习的高通预滤波模块来获取图像残差,以增强修复痕迹,然后使用四个串联的 ResNet 块构建特征提取模块,该模块从图像残差中学习判别特征,最后通过上采样模块将学习到的特征图放大,从而获得逐像素级别的图像修复定位掩码。为了解决篡改区域过小而导致的样本不平衡问题,他们使用 FocalLoss 损失函数训练模型。图像修复主要分为传统修复和深度修复两种类型,为了提高不同深度修复图像检测的泛化性,Li 等人[37] 提出了一种双分支的噪声图像交叉融合网络 (Noise-Image Cross-fusion Network, NIX-Net) 。该网络通过交叉融合 RGB 图像分支和图像噪声残差分支,有效地利用了图像及其噪声模式中包含的因深度修复而导致的失真判别信息。相似的,为了对不同类型的修复图像(包括传统修复和深度修复) 进行检测,Wu 等人[38] 结合神经架构搜索(Neural Architecture Search, NAS) 算法和注意力机制提出了一种端到端的修复图像检测算法 IID-Net。

针对 PS 工具处理的图像取证,Wang 等人[39] 提出一种基于 CNN 的 PS 脚本处理人脸图像的检测方法。为了检测图像是否经过 PS 脚本篡改,他们使用改进的空洞残差网络(Dilated Residual Network variant, DRN-C-26) [40] 作为二分类器; 为了检测 PS 变形处理后图像的篡改区域,他们设计了一个基于空洞残差网络结构的模型去预测从原始图像到变形图像的光流场(Optical Flow Field) 。为了提高模型的鲁棒性,他们使用了丰富的数据增强进行训练模型,其中包括缩放图像的大小、JPEG 压缩和各种类型的直方图编辑。

在不考虑特定的篡改方法的条件下,由于现实中的数据存在被各种不同的篡改方法修改的可能,再加这些修改的历史不可追溯性,给基于端到端有监督学习的取证方法带来巨大的挑战。Wu 等人[41] 提出一种基于全卷积网络的端到端图像篡改定位方法,可以检测被许多已知的伪造类型方法篡改过的任意大小的图像。他们设计的全卷积网络叫 ManTra-Net,该网络包含两个子模块: 创建统一特征表示的图像操作痕迹特征提取器(Image Manipulation-trace Feature Extractor, IMFE) 和对伪造区域进行定位的局部异常检测网络(Local Anomaly Detection Network, LADN) 。为了在 IMFE 模块提取更好的统一特征表示,他们通过对 385 种图像操作类型进行分类来学习更加鲁棒的图像操作痕迹; 为了在 LADN 模块捕捉更好的局部异常,他们将伪造定位问题描述为一个局部异常检测问题,设计了一个 Z-score 特征来捕获局部异常,并提出了一种新的长短期记忆(Long Short Term Memory, LSTM) 方法来评估局部异常。Hu 等人[42] 在 ManTra-Net 的基础上,提出了一种基于空间金字塔注意网络(Spatial Pyramid Attention Network, SPAN) 的图像篡改定位方法,SPAN 是由特征提取模块, 金字塔空间注意传播(Pyramid Spatial Attention Propagation) 模块和篡改区域预测模块这三个模块组成。在特征提取模块,他们使用预训练 ManTra-Net 的特征提取器作为特征提取器,在金字塔空间注意传播模块通过局部自我注意和金字塔传播进一步模拟空间相关性,最后在篡改区域预测模块计算得到篡改掩码。此外,Bappy 等人[43] 提出一种基于混合 LSTM 和编码器-解码器(Encoder-Decoder) 网络(H-LSTM) 的图像篡改定位方法,该方法框架由三部分组成: LSTM 网络、卷积编码器和卷积解码器。其中,提取图像的重采样特征(Resampling Features) 作为 LSTM 网络的输入,以捕捉伪造图像中,块与块之间在频域上的不一致性; 而卷积编码解码器则是直接输入空域的 RGB 图像以提取图像的空域特征; 最后把 LSTM 网络和卷积编码解码器的输出特征拼接在一起作为卷积解码器的输入以预测出图像篡改区域的掩码。Zhou 等人[44] 提出一种新颖的基于生成、分割和细化(Generate, Segment, and Refine) 的图像伪造检测框架,该框架能够在训练期间使用 GAN 在线生成囊括各种篡改方法的难检测伪造样本,然后使用基于 CNN 的分割和细化网络对图像的篡改区域和篡改边界进行检测。

为了对视频进行帧重复(Frame Duplication) 检

测和定位，Long 等人[45]提出一种新颖的基于深度 CNN 的从粗到细(Coarse-to-fine)的视频取证框架。他们首先使用 I3D 网络[46]在候选重复帧序列和相应的选定原始帧序列之间找到粗略匹配，然后使用基于 ResNet 结构的孪生网络(Siamese Network)进一步识别单个重复帧与相应选定帧之间的精细对应关系，最后使用基于 I3D 的不一致性检测器实现对视频重复帧的检测和定位。

3.2 One-class 弱监督方法

使用基于深度 CNN 算法进行有监督图像视频取证，往往需要大量有标签的训练数据，而要想得到一个能够检测任何可能篡改操作的取证算法，则必须需要收集囊括所有可能篡改操作的数据集，显然这是不现实的。因此，使用原始数据的单分类弱监督取证方法是一种可能的替代方案。该方法通过充分学习原始数据的共同固有的特征，间接得到能够区分任何可能篡改操作的检测能力。

Cozzolino 等人[47]提出一种基于自编码器(Autoencoder)的单分类图像拼接篡改检测方法。该方法从图像噪声残差中提取富有代表性的局部特征，然后输入到生成数据的隐式模型(Implicit Model)的自编码器中。最后通过迭代判别性特征的标注和自动编码，隐式模型最终拟合原始数据，从而识别出与原始数据存在较大异常的拼接区域。D'Avino 等人[48]把文献[47]方法扩展到视频拼接检测上，提出了一种基于自动编码器和递归神经网络(Recurrent Neural Networks, RNN)的视频拼接检测方法。两种方法都是提取图像块(Image Patches)的特征输入自编码器中。所不同的是，文献[48]的方法利用了 LSTM 模型学习视频前后帧之间在时序上的相关性。

不同的相机型号在图像上存在独特的相机模型特征(Camera-model Features)，即原始图像的所有像素只含有单一相同的相机生成痕迹，而篡改合成图像则存在多个不同的相机生成痕迹。根据这个原理，Bondi 等人[49]使用 CNN 从图像块中提取相机模型特征，然后结合置信得分(Confidence Score)使用 K-means 算法进行聚类评估，从而对伪造图像进行篡改检测和定位。相似的，Mayer 等人[50]在提出了一种比较两个图像块是否源自同一相机模型的算法，该算法首先使用 CNN 提取反映相机模型信息的图像块高维特征(High-level Features)，然后把两个图像块的特征输入到一个基于全连接层的相似性度量网络中，从而预测出两个图像块是否源自相同的相机模型，即实现源区分(Source differentiation)。同时，在图像块源区分基础上，可以进一步对拼接图像进行检测。另外，Huh 等人[51]利用原始图像 EXIF 头部的元数据(Meta-data)作为有监督标签训练一个孪生网络，该网络能够区分两个图像块是否具有相同的元数据，从而进一步把该网络应用于图像拼接检测和定位的任务。

不同于前面的相机模型特征的方法，Cozzolino 等人[52]提出一种基于 CNN 孪生网络提取相机噪声指纹(Camera Noiseprint)的图像篡改定位方法。该方法使用成对的图像块训练孪生网络，其中来自相同相机拍摄的一对图像在相同位置上的图像块具有相同的标签，否则具有不同的标签。更进一步，Cozzolino 等人[53]把文献[52]相机噪声指纹方法扩展到视频取证上，通过使用孪生网络学习到视频的相机噪声指纹，从而能够对视频帧进行篡改检测和定位，甚至能够对 Deepfake 视频进行检测。

表3 端到端图像视频取证方法总结

Tab. 3 Summary of end-to-end image and video forensics methods

有监督学习方法			
论文	机器学习模型	方法特点	任务
Salloum 等人[30]	多任务 FCN(MFCN)	多任务，关注篡改区域和篡改轮廓	拼接图像，篡改定位
Bi 等人[31]	环状残差 U-Net (RRU-Net)	环状残差思想，U-Net 结构网络	拼接图像，篡改定位
Wu 等人[32]	FCN	端到端，反卷积网络解码	复制-移动图像，篡改定位
Zhong 等人[33]	Dense-InceptionNet	包含金字塔特征提取器(PFE)，特征相关匹配(FCM)和分层后处理(HPP)三个模块	复制-移动图像，篡改定位
Wu 等人[34]	BusterNet	包含一个双分支模块和一个融合模块，区分源区域和目标区域	复制-移动图像，篡改定位
李等人[35]	FCN	基于 cGAN 的 FCN，弱监督样本，区分源区域和目标区域	复制-移动图像，篡改定位
Li 等人[36]	高通 FCN	可学习的高通预滤波卷积层来获取图像残差，ResNet 作为主干网络	修复图像，篡改定位
Li 等人[37]	NIX-Net	RGB 图像分支和图像残差噪声分支交叉融合	修复图像，篡改定位

续表 3

有监督学习方法			
论文	机器学习模型	方法特点	任务
Wu 等人[38]	IID-Net	神经架构搜索(NAS) ,注意力机制	修复图像,篡改定位
Wang 等人[39]	DRN-C-26+FCN	分类用改进版本的空洞卷积网络(DRN-C-26) ,定位用主干网络为 DRN-C-26 的 FCN 预测光流场	PS 处理图像,检测+定位
Wu 等人[41]	ManTra-Net	图像操作痕迹特征提取器(IMFE) 和局部异常检测网络(LADN) ,SRM 卷积,LSTM	任意篡改图像,篡改定位
Hu 等人[42]	空间金字塔注意网络(SPAN)	使用 ManTra-Net 提取特征,使用金字塔空间注意传播模块模拟空间相关性	任意篡改图像,篡改定位
Bappy 等人[43]	混合 LSTM 和编码器-解码器网络(H-LSTM)	LSTM 提取频域上的重采样特征,编码器提取 RGB 特征	任意篡改图像,篡改定位
Zhou 等人[44]	GAN+CNN	把篡改图像生成、分割和细化的步骤整合成一个框架,利用了自监督思想	任意篡改图像,篡改定位
Long 等人[45]	I3D + CNN-based 孪生网络	粗略匹配用 I3D, 精细对应关系识别用基于 ResNet 的孪生网络,最后使用 I3D 实现视频重复帧的检测和定位	视频,重复帧检测+定位
One-class 弱监督方法			
论文	机器学习模型	方法特点	任务
Cozzolino 等人[47]	Autoencoder	图像噪声残差局部特征,Autoencoder 多次迭代	拼接图像,篡改定位
D' Avino 等人[48]	Autoencoder+ LSTM	图像噪声残差局部特征,Autoencoder 多次迭代,LSTM 学习时序上的相关性	拼接视频,篡改检测
Bondi 等人[49]	CNN+K-means	CNN 提取相机模型特征,K-means 聚类	任意篡改图像,篡改定位
Mayer 等人[50]	CNN	CNN 提取反映相机模型信息的图像块高维特征,全连接层的相似性度量	图像块,源区分; 拼接图像,篡改检测
Huh 等人[51]	CNN-based 孪生网络	原始图像 EXIF 头部的元数据(Meta-data) 作为有监督标签训练一个孪生网络	拼接图像,检测+定位
Cozzolino 等人[52]	CNN-based 孪生网络	相同相机拍摄的一对图像在相同位置上的图像块具有相同的标签+1,否则为-1	任意篡改图像,篡改定位
Cozzolino 等人[53]	CNN-based 孪生网络	相同相机拍摄的一对图像在相同位置上的图像块具有相同的标签+1,否则为-1	常规篡改视频+Deepfake 视频,检测+定位

3.3　端到端 Deepfake 检测方法

端到端的 Deepfake 检测方法中大多数使用深度学习的方法进行检测,因为输入直接就是图片,所以会在模型结构或者其他辅助模块比如注意力机制上进行深入探究,尽量让模型学习到 Deepfake 篡改特征。

Afchar 等人[54]将检测篡改视频的任务置于中层语义水平来分析,提出了由少量卷积模块组成的浅层卷积网络来检测。Rössler 等人[27]使用 Xception 来检测 Deepfake 视频,通过实验发现基于关键点提取出来的人脸图像作为输入要比一整帧作为输入效果好; 深度较深的通用网络会比浅的效果好; 对于压缩率比较高的 Deepfake 视频则难以检测。Dang 等人[55]则认为检测模型要更多关注篡改的重点区域,提出了多任务学习的方式,即在检测 Deepfake 图像的同时,通过在主干网络上使用注意力机制生成伪造区域进行定位,并通过伪造区域的强化来增强检测结果。龚等人[56]认为 Dang 采用的是参数不可学习的注意力机制模块,所以通过设置可学习参数的双层注意力机制来完成异常特征及相关区域的自适应捕获,双层注意力模块分别是通道注意力模块和空间注意力模块,两种模块分别与主干网络拼接,主干网络可以是 Xception 和 ResNet。

Kumar 等人[57]提出了使用度量学习的方式来检测高压缩率的 Deepfake 视频。将人脸图像输入网络获得人脸特征并输入三元组网络,将真帧和假帧进行分离,从而进行 Deepfake 检测。这个方法在高压缩率的情况下,能够比基于卷积神经网络的方法好很多。在使用度量学习的基础上,Qian 等人[58]进一步

将目光投向了频域信息去解决低质量视频上的检测难题,因为传统的频域处理如离散余弦变换(DCT) 或者傅里叶变换(FFT) 不满足平移不变性和局部一致性,不适合直接输入 CNN,所以作者设计了两种频域特征,第一种将自适应的滤波器放在 DCT 变换后,再做反 DCT,重组回原来的图片; 第二种是对输入的图片进行滑窗 DCT,提取并统计局部的频率信息。两种频域特征虽然不同,但都是用 DCT 提取,具有一致性,然后再设计一个模块融合这两种特征,使用双路网络来完成预测,主干网络使用的是 Xception,实验结果显示对低质量的视频具有很好的检测效果。

除了常规的深度学习分类网络,Nguyen 等人[59]也尝试引入胶囊网络(Capsule Network) 。将图像输入 VGG16 提取特征,再使用胶囊网络进行检测。Güera 等人[60]认为不同帧场景下光源引起的脸部闪烁现象,这种基于时间的信息可以用 LSTM 捕捉。通过将图片输入 Inception v3 网络提取特征,再输入到 LSTM 来进行检测。Sabir 等人[61]借鉴了行为识别领域利用时间信息处理视频的方法,使用 ResNet 或 DenseNet 作为主干网络,后接 RNN 网络进行端到端的训练,并且发现双向 RNN 的效果最好。Amerini 等人[62]认为 Deepfake 视频是由计算机合成的,和由摄像机拍摄的原视频存在显著差异,这种差异可以使用光流来捕捉。他们使用 PWC-Net 来提取帧间的光流后再使用分类网络比如 VGG16 或 ResNet50 进行分类。Li 等人[63]关注 Deepfake 视频合成过程中的融合(blending) 阶段,提出结合了分类和分割的方法来输出 Deepfake 视频的篡改轮廓的方法——Face X-ray,作者使用了一种融合(blending) 的方式来构造伪造人脸,在数据集中选择真实人脸并计算 Landmark,然后通过 Landmark 的欧氏距离在数据集中寻找最相似的人脸进行融合(blending) ,另一方面通过 Landmark 得到融合的 Mask。使用 HRNet 全卷积网络来预测 Mask。实验结果表明,该方法在跨数据集的测试上表现非常好。韩等人[64]认为像 Face X-ray 等只用到了视觉特征,而使用记忆能力的 LSTM 和 CNN 结合的方法不完全适用于伪造检测,没有充分利用视频的时间信息。所以提出了仅包含两个 Inception 模块的高效 I3D(Inception3D) 网络,整体是双流结构,使用 dlib 开源库提取眼部和口部的视觉信息作为双流网络的输入,并且使用 3D 网络提取时间特征。

随着更多的模型发表出来,模型的泛化性能也逐渐被大家关注,因为在单一数据集训练和测试,准确率往往很高,一旦跨域测试,性能就下降很快,Wang 等人[65]设计了一种辅助监督的方法去引导网络学习伪造视频中显著但又更通用的篡改特征。作者设计了两种解码器,一种是纹理解码器,用来估计真实图片的局部二值图和伪造图片的零值图。一种是合成边界解码器,用来估计造假图片的伪造边界。两种辅助特征通过双路网络结合在一起用于最终的检测任务。实验结果显示,在单一数据集上训练和测试的准确保持非常高的水平下,跨数据集的测试也能达到当时最好的水平。

总的来说,现有端到端的 Deepfake 检测方法基本可以分为两类。如图 1 所示,第一类是通过提取脸部区域后,直接使用分类器进行分类,同时可以添加一些注意力图(Attention Map) 让分类器关注更加感兴趣的特征; 第二类是提取脸部区域后使用特征提取模型提取特征(一般是基于 CNN 的模型比如 Inception) ,然后再输入时序网络如 LSTM 等,再输出分类结果。端到端的 Deepfake 检测会比手工构造特征的方法性能好,但是对数据集的依赖程度较高,通常需要大量的数据进行训练。而且在模型鲁棒性方面也面临着很多挑战,比如在跨数据集的表现上性能往往会下降很多。表 4 总结了 Deepfake 检测方法的模型构造、方法特点和原论文所用到的数据集,数据集的全称和他们的简称分别如下: Faceforensics 即 FF[26] ,Meso 数据集[54] ,FaceForensics++即 FF++[27] ,Diverse Fake Face Dataset 即 DFFD[55] ,DeepFake Detection Challenge 即 DFDC[66] ,Celeb-DF [28] 以及从互联网和 HOHA 数据集[67] 中选出 600 个视频自建的 Deepfake 数据集(HOHA-DF) 。

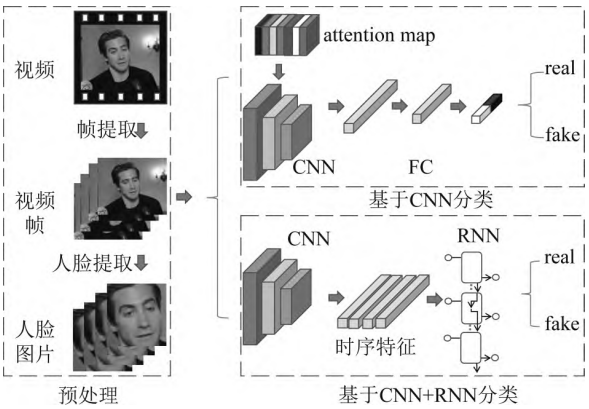

图 1 端到端的 Deepfake 检测模型的两种类别

Fig. 1 Two categories of end-to-end Deepfake detection models

表 4　端到端的 Deepfake 检测方法总结

Tab. 4　Summary of end-to-end Deepfake detection methods

论文	机器学习模型	方法特点	数据集
Afchar 等人[54]	CNN	网络深度浅,直接输入整张图片	FF,Meso 数据集
Rössler 等人[27]	Xception	提取脸部区域后再输入网络	FF++
Dang 等人[55]	CNN+attention map	加入注意力机制	DFFD,FF++
龚等人[56]	CNN+attention	使用可学习参数的双层注意力机制	FF++
Kumar 等人[57]	Triplet Network	使用度量学习检测高压缩率的视频	FF++,Celeb-DF
Qian 等人[58]	CNN	设计并使用了两种频域特征	FF++
Nguyen 等人[59]	CNN+胶囊网络	首次使用胶囊网络来检测 Deepfake	FF,Meso 数据集
Güera 等人[60]	CNN+LSTM	捕捉脸部光源的时序信息	HOHA-DF
Sabir 等人[61]	CNN+RNN	双向 RNN	FF++
Amerini 等人[62]	PWC-Net+CNN	利用光流学习合成视频和自然视频的差异	FF++
Wang 等人[65]	CNN	使用了辅助监督	FF,FF++
Li 等人[63]	HRNet	利用自监督合成伪造人脸来增广数据	FF++,DF,DFDC
韩等人[64]	Inception3D	利用双流特征和时间信息	FF++,Celeb-DF,DF,DFDC

4　取证与反取证的对抗

本章节主要介绍取证与反取证的对抗,其中包括应用于反取证的机器学习方法和生成对抗网络在生成伪造样本中的应用。

4.1　应用于反取证的机器学习方法

数字媒体在当今社会发挥的作用越来越大,而随之兴起的媒体取证技术也日益壮大。但与之对抗的反取证技术也层出不穷。早期的反取证技术通常会根据特定的取证方法进行攻击,这些方法大多数是基于特定的手工特征的机器学习方法,比如抹除傅里叶域中的重采样痕迹、抹除传感器痕迹或者 PRNU 指纹等等。

近年来很多取证技术都依赖于深度学习,而随着神经网络研究的深入,神经网络的脆弱性也逐渐暴露出来。Szegedy 等人[68] 发现神经网络模型中可用通过添加一些难以感知的扰动使模型分类错误,这个扰动可以通过最大化模型预测误差来得到,而且,这些扰动不是特定学习的产物,将同样的扰动添加到其他数据集也会产生分类错误的效果。Szegedy 的研究一般被认为是对抗样本的开山之作,之后,基于 Szegedy 理论的生成对抗样本方法不断被提出,Goodfellow 等人[69] 提出了 FGSM(Fast Gradient Sign Method) 算法,通过梯度生成对抗扰动添加到输入图片从而生成对抗样本,Kurakin 等人[70] 改进 FGSM 为 I-FGSM(Iterative-FGSM) 算法,在 FGSM 的基础上,进行梯度的多次迭代来生成攻击性更强的对抗样本。Dong 等人[71] 提出了 MI-FGSM(Momentum Iterative FGSM) ,在梯度更新的基础上,加入动量,通过累积损失函数方向上的矢量,有效避免局部最优。作者还扩展了有目标攻击,即指定目标生成对抗样本,在白盒和黑盒迁移攻击都具有优秀的效果。除此之外,Carlini[72] 等人提出一种攻击方法——CW(由作者名字首字母构成) ,他们认为 I-FGSM 等算法生成的对抗样本干扰噪声比较明显,肉眼可观察到,所以在 loss 增加了对抗样本和原样本之间的距离,使得生成的对抗样本在具有较高攻击称功率的同时,保持着与原样本极高的相似度。在现实世界中,对抗样本一般发挥不出其作用,比如摄像头拍摄产生的噪声会掩盖掉对抗样本的扰动,导致失效。Brown 等人[73] 抛弃了迭代更新整幅图像的方法,而是替换掉图像的一部分来发起攻击,这个部分称为对抗补丁(Adversarial Patch) ,补丁的形状可以是任意的,然后添加到图像中,并随机进行平移、缩放和旋转等,只使用梯度更新补丁。生成对抗补丁后,可以在网络上进行传播,攻击者可以直接打印使用去攻击任何适用的模型。

许多反取证的方法都是借鉴攻击神经网络的方法而实现的。Chen 等人[74] 对使用 SPAM 特征的 SVM 分类器进行攻击,使用基于梯度的攻击方法,使分类器的检测结果降低,并能有效应用在该类型的 SVM 分类器中。图片拍摄相机源检测是取证的基本任务之一,Marra 等人[75] 对基于 CNN 的几种模型使用 FGSM 算法进行攻击,结果显示在图像正常的情况下攻击非常有效,但是如果图像经过压缩之后,攻击效果则会下降。Carlini 等人[76] 在检测伪造人脸的 CNN 模型上展开攻击实验,他们假设了白盒的场景下,使用了多种攻击方法包括 I-FGSM 和对抗补丁[73] 。两种攻击方法均能让检测模型准确率大幅度下降。

在计算机视觉领域,对抗样本有着比较强的迁移性。即针对某个模型产生的对抗样本,用于攻击另一个模型也同样可能导致分类错误。在图像视频取证领域,Barni 等人[77] 研究了在图像取证方法上,对抗样本攻击的迁移能力。他们用同一个数据集训练两个基于 CNN 的识别图片相机来源模型。使用 I-FGSM 算法对其中一个模型生成对抗样本,

并用这些对抗样本去攻击另一个模型。他们报道的实验结果却显示迁移性很差。

4.2 生成对抗网络在生成伪造样本中的应用

生成对抗网络(GAN) 自提出以来不断的发展,在生成类任务具有很好的效果,许多反取证方法是基于 GAN 来设计的。对于深度神经网络模型,使用基于梯度或者优化的方法生成对抗样本时,因为需要多次迭代,耗时较慢。Xiao 等人[78] 针对这些缺点提出了基于 GAN 生成对抗样本的半白盒方法——advGAN,他们将原始图片作为生成器的输入,生成扰动,然后将扰动添加到原始图片形成对抗样本。advGAN 在训练完成后获得的生成器,其生成对抗样本的时间非常短。

Chen 等人[79-80] 设计了一个生成器来伪造能够欺骗相机源识别模型的图片。他们根据对攻击方相机模型的掌握程度,把训练场景分为白盒训练和黑盒训练。白盒模型是在假设获取了相机源分类模型的全部信息下,将分类模型直接拿来使用。训练完成后,图片会被分割成很多小块分别进行攻击,然后再组合成大图片; 在黑盒场景下,则使用迁移攻击的方式,用替代模型进行对抗训练。实验结果表明,在白盒场景和黑盒场景下都能达到较好的攻击效果,而且伪造图片的视觉质量不低。

在图像的操作痕迹反取证中,一般通过加入中值滤波来消除操作痕迹,但是中值滤波很容易被现有的多媒体取证方法检测出来。Kim 等人[81] 提出了一种基于 GAN 的结构的中值滤波反取证方法,能够有效消除经过中值滤波处理的图像,显著增强了中值滤波图像的反取证性能。不同的 GAN 生成的伪造图像会留下特定的指纹特征,很多针对这种指纹特征的检测器经过数据处理很容易将这些伪造图片检测出来。Neves 等人[82] 提出了一种自动编码器,能够将合成的伪造图片指纹信息移除,让大多数现有的检测器失效。

5 研究展望

从上述综述可以看出,过去三年在大数据和深度学习技术赋能下,无论是图像/视频取证技术还是与之相对抗的图像/视频篡改及相对应的反取证技术,均取得了长足的发展,近年来图像/视频取证,以及深度学习图像/视频取证的学术论文数量统计如图 2 (Google scholar 搜索结果图) 所示。但毫无疑问,即使经历了十多年的演化,数字多媒体上,尤其是图像/视频上的取证研究还远未成熟和完备。面向图像/视频取证的机器学习技术依然存在许多未解决的科学问题。在此我们试图基于对本领域的把握和理解,展望在大数据和深度学习技术赋能下,图像/视频取证技术以及面向取证的机器学习技术的未来发展路向,归纳出以下几个未来的有待重点研究的问题:

图 2 Google Scholar 搜索结果图

Fig. 2 The search results for Google Scholar

1) 机器学习模型在面向图像/视频取证时的广谱适用性。面对真实取证场景中千变万化的图像/视频压缩格式、媒体质量和尺寸、篡改手段和涉及区域,如何实现一种只需一次高强度训练,其后只需少量样本做微调便可自适应地应用于各种真实取证场景的机器学习模型,是一个有待重点研究的问题。

2) 机器学习模型在面向图像/视频取证时的可解释性。现有的机器学习图像/视频取证模型,特别是基于深度学习的取证模型,其最为人诟病的地方是模型本身无法解释其作出某一决策背后的理据。如何提升机器学习模型,特别是深度学习模型在应用于图像/视频取证时,其基于严谨的数学理论分析的可解释性,是另一个有待重点研究的问题。

3) 机器学习模型在面向图像/视频取证时的鲁棒性。现有的机器学习图像/视频取证模型,往往依赖于高度非线性的深度学习及集成学习框架。这些高度非线性的机器学习框架极易受到针对性的对抗样本攻击。如何提升机器学习模型在面向图像/视频取证时的鲁棒性,从根本上杜绝针对性对抗样本的构造,是第三个有待重点研究的问题。

4) 以模型为中心,而非以数据为中心的图像/视频取证机器学习技术。现有的图像/视频取证机器学习技术均以数据为中心,因此其性能表现与训练数据紧密相关,其在训练数据所无法预见的真实测试样本上的表现存疑。如何使图像/视频取证机器学习技术转为以模型为中心,使其能够有效地度量真实及篡改图像/视频之间的统计差异,是我们认为最重要的一个有待研究的问题。

6 结语

近三年来,随着基于深度学习技术的 Deepfake 视频换脸软件的流行,面向图像和视频篡改的取证技术从高高在上的象牙塔迅速走进了普罗大众的视野,也引起了包括学术界在内的各国相关部门的高度重

视。这一波媒体取证领域的技术迭代,其基础就是以深度学习为代表的高度复杂的机器学习技术的出现和兴起。本文对近年来,特别是过去三年面向图像/视频取证的机器学习技术的飞速发展现状进行了综述,并对其未来的发展趋势进行了展望,对重要的研究问题进行了归纳。希望这篇综述对投身于多媒体取证领域的研究学者们的未来研究能有所启发。

参考文献

[1] PIVA A. An overview on image forensics[J]. ISRN Signal Processing, 2013, 2013: 1-22.

[2] KORUS P. Digital image integrity-a survey of protection and verification techniques[J]. Digital Signal Processing, 2017, 71: 1-26.

[3] JIANG Xinghao, HE Peisong, SUN Tanfeng, et al. Detection of double compression with the same coding parameters based on quality degradation mechanism analysis[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(1): 170-185.

[4] HSU Y F, CHANG S F. Camera response functions for image forensics: An automatic algorithm for splicing detection[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(4): 816-825.

[5] LI Haodong, LUO Weiqi, QIU Xiaoqing, et al. Identification of various image operations using residual-based features[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(1): 31-45.

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

[7] 李子健,阮秋琦. 基于 LPP 和改进 SIFT 的 copy-move 篡改检测[J]. 信号处理, 2017, 33(4):589-594.
LI Zijian, RUAN Qiuqi. Copy-move forgery detection based on LPP and improved SIFT algorithm[J]. Journal of Signal Processing, 2017, 33(4): 589-594. (in Chinese)

[8] 于亮, 杨红颖. 基于自适应提点鲁棒定位的图像复制粘贴篡改检测[J]. 计算机系统应用, 2020, 29(12): 117-125.
YU Liang, YANG Hongying. Copy-move forgery detection based on adaptive keypoints extraction and robust localization[J]. Computer Systems & Applications, 2020, 29(12): 117-125. (in Chinese)

[9] WANG Qing, ZHANG Rong. Double JPEG compression forensics based on a convolutional neural network[J]. EURASIP Journal on Information Security, 2016, 2016(1): 1-12.

[10] PARK J, CHO D, AHN W, et al. Double JPEG detection in mixed JPEG quality factors using deep convolutional neural network[C]// Computer Vision-ECCV 2018. Munich, Germany. Springer International Publishing, 2018: 656-672.

[11] BARNI M, BONDI L, BONETTINI N, et al. Aligned and non-aligned double JPEG detection using convolutional neural networks[J]. Journal of Visual Communication and Image Representation, 2017, 49: 153-163.

[12] ZHOU Peng, HAN Xintong, MORARIU V I, et al. Learning rich features for image manipulation detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 1053-1061.

[13] NAM S H, PARK J, KIM D, et al. Two-stream network for detecting double compression of H.264 videos[C]// 2019 IEEE International Conference on Image Processing (ICIP). Taipei, Taiwan, China. IEEE, 2019: 111-115.

[14] 白珊山, 倪蓉蓉, 赵耀. 基于双通道卷积神经网络的视频目标移除取证算法[J]. 信号处理, 2020, 36(9): 1415-1421.
BAI Shanshan, NI Rongrong, ZHAO Yao. Video forensics for object removal based on two-channel convolutional neural network[J]. Journal of Signal Processing, 2020, 36(9): 1415-1421. (in Chinese)

[15] CHAKRABORTY S, KIRCHNER M. PRNU-based image manipulation localization with discriminative random fields[J]. Electronic Imaging, 2017, 2017(7): 113-120.

[16] VERDOLIVA L, COZZOLINO D, POGGI G. A feature-based approach for image tampering detection and localization[J]. 2014 IEEE International Workshop on Information Forensics and Security (WIFS), 2014: 149-154.

[17] LI Qiushi, CHEN Shengda, TAN Shunquan, et al. One-class double compression detection of advanced videos based on simple Gaussian distribution model[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021: 1.

[18] 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). Waikoloa, HI, USA. IEEE, 2019: 83-92.

[19] LI Yuezun, LYU Siwei. Exposing deepfake videos by detecting face warping artifacts[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA. IEEE, 2019: 46-52.

[20] LI Yuezun, CHANG M C, LYU Siwei. In ictu oculi: Exposing AI created fake videos by detecting eye blinking[J]. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018: 1-7.

[21] CIFTCI U A, DEMIR I, YIN Lijun. FakeCatcher: detection of synthetic portrait videos using biological signals[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020: 1.

[22] FERNANDES S, RAJ S, ORTIZ E, et al. Predicting heart rate variations of deepfake videos using neural ODE

[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South). IEEE, 2019: 1721-1729.

[23] YANG Xin, LI Yuezun, LYU Siwei. Exposing deep fakes using inconsistent head poses [C]// ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 8261-8265.

[24] AGARWAL S, FARID H, GU Yuming, et al. Protecting world leaders against deep fakes [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA. IEEE, 2019: 38-45.

[25] KORSHUNOV P, MARCEL S. DeepFakes: a new threat to face recognition? assessment and detection [EB/OL]. https://arxiv.org/abs/1812.08685,2018.

[26] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics: A large-scale video dataset for forgery detection in human faces [EB/OL]. https://arxiv.org/abs/1803.09179,2018.

[27] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: learning to detect manipulated facial images [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 1-11.

[28] LI Yuezun, YANG Xin, SUN Pu, et al. Celeb-DF: A large-scale challenging dataset for DeepFake forensics [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 3204-3213.

[29] GUAN Haiying, KOZAK M, ROBERTSON E, et al. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation [C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Waikoloa, HI, USA. IEEE, 2019: 63-72.

[30] SALLOUM R, REN Yuzhuo, JAY KUO C C. Image splicing localization using a multi-task fully convolutional network (MFCN) [J]. Journal of Visual Communication and Image Representation, 2018, 51: 201-209.

[31] BI Xiuli, WEI Yang, XIAO Bin, et al. RRU-net: The ringed residual U-net for image splicing forgery detection [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA. IEEE, 2019: 30-39.

[32] WU Yue, ABD-ALMAGEED W, NATARAJAN P. Image copy-move forgery detection via an end-to-end deep neural network [C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, NV, USA. IEEE, 2018: 1907-1915.

[33] ZHONG Junliu, PUN C M. An end-to-end dense-InceptionNet for image copy-move forgery detection [J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 2134-2146.

[34] WU Yue, ABD-ALMAGEED W, NATARAJAN P. BusterNet: detecting copy-move image forgery with source/target localization [C]// Computer Vision-ECCV 2018. Munich, Germany. Springer International Publishing, 2018: 170-186.

[35] 李应灿, 杨建权, 丁峰, 等. 区分来源和目标区域的图像 copy-move 伪造检测方法 [J]. 信号处理, 2020, 36(9): 1533-1543.
LI Yingcan, YANG Jianquan, DING Feng, et al. Copy-move detection method for distinguishing between source and target regions [J]. Journal of Signal Processing, 2020, 36(9): 1533-1543. (in Chinese)

[36] LI Haodong, HUANG Jiwu. Localization of deep inpainting using high-pass fully convolutional network [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 8300-8309.

[37] LI Ang, KE Qiuhong, MA Xingjun, et al. Noise doesn't lie: towards universal detection of deep inpainting [C]// 2021 International Joint Conferences on Artifical Intelligence (IJCAI). Montreal, 2021: 786-792.

[38] WU Haiwei, ZHOU Jiantao. IID-Net: Image inpainting detection network via neural architecture search and attention [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021: 1.

[39] WANG Shengyu, WANG O, ZHANG R, et al. Detecting photoshopped faces by scripting photoshop [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 10071-10080.

[40] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 636-644.

[41] WU Yue, ABDALMAGEED W, NATARAJAN P. ManTra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 9535-9544.

[42] HU Xuefeng, ZHANG Zhihan, JIANG Zhenye, et al. SPAN: spatial pyramid attention network for image manipulation localization [M]// Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 312-328.

[43] BAPPY J H, SIMONS C, NATARAJ L, et al. Hybrid LSTM and encoder-decoder architecture for detection of image forgeries [J]. IEEE Transactions on Image Processing, 2019, 28(7): 3286-3300.

[44] ZHOU Peng, CHEN B C, HAN Xintong, et al. Gener-

ate, segment, and refine: Towards generic manipulation segmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13058-13065.

[45] LONG Chengjiang, BASHARAT A, HOOGS A. A coarse-to-fine deep convolutional neural network framework for frame duplication detection and localization in video forgery [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA. IEEE, 2019: 1-10.

[46] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 4724-4733.

[47] COZZOLINO D, VERDOLIVA L. Single-image splicing localization through autoencoder-based anomaly detection [J]. 2016 IEEE International Workshop on Information Forensics and Security (WIFS), 2016: 1-6.

[48] D'AVINO D, COZZOLINO D, POGGI G, et al. Autoencoder with recurrent neural networks for video forgery detection[J]. Electronic Imaging, 2017, 2017(7): 92-99.

[49] BONDI L, LAMERI S, GÜERA D, et al. Tampering detection and localization through clustering of camera-based CNN features[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA. IEEE, 2017: 1855-1864.

[50] MAYER O, STAMM M C. Learned forensic source similarity for unknown camera models[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 2012-2016.

[51] HUH M, LIU A, OWENS A, et al. Fighting fake news: Image splice detection via learned self-consistency[M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 106-124.

[52] COZZOLINO D, VERDOLIVA L. Noiseprint: A CNN-based camera model fingerprint[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 144-159.

[53] COZZOLINO D, POGGI G, VERDOLIVA L. Extracting camera-based fingerprints for video forensics[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA. IEEE, 2019: 130-137.

[54] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network [J]. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018: 1-7.

[55] DANG Hao, LIU Feng, STEHOUWER J, et al. On the detection of digital face manipulation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 5780-5789.

[56] 龚晓娟, 黄添强, 翁彬, 等. 基于双层注意力的 Deepfake 换脸检测[J]. 网络与信息安全学报, 2021, 7(2): 151-160.
GONG Xiaojuan, HUANG Tianqiang, WENG Bin, et al. Deepfake swapped face detection based on double attention[J]. Chinese Journal of Network and Information Security, 2021, 7(2): 151-160. (in Chinese)

[57] KUMAR A, BHAVSAR A, VERMA R. Detecting deepfakes with metric learning[C]//2020 8th International Workshop on Biometrics and Forensics (IWBF). Porto, Portugal. IEEE, 2020: 1-6.

[58] QIAN Yuyang, YIN Guojun, SHENG Lu, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 86-103.

[59] 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). Brighton, UK. IEEE, 2019: 2307-2311.

[60] GÜERA D, DELP E J. Deepfake video detection using recurrent neural networks[C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Auckland, New Zealand. IEEE, 2018: 1-6.

[61] SABIR E, CHENG Jiaxin, JAISWAL A, et al. Recurrent convolutional strategies for face manipulation detection in videos[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA. IEEE, 2019: 80-87.

[62] AMERINI I, GALTERI L, CALDELLI R, et al. Deepfake video detection through optical flow based CNN[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South). IEEE, 2019: 1205-1207.

[63] LI Lingzhi, BAO Jianmin, ZHANG Ting, et al. Face X-ray for more general face forgery detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 5000-5009.

[64] 韩语晨, 华光, 张海剑. 基于 Inception3D 网络的眼部与口部区域协同视频换脸伪造检测[J]. 信号处理, 2021, 37(4): 567-577.
HAN Yuchen, HUA Guang, ZHANG Haijian. Inception3D net based video face swapping forgery detection jointly exploiting eye and mouth areas[J]. Journal of Signal Processing, 2021, 37(4): 567-577. (in Chinese)

[65] WANG Xinyao, YAO Taiping, DING Shouhong, et al. Face manipulation detection via auxiliary supervision[M]//Neural Information Processing. Cham: Springer International Publishing, 2020: 313-324.

[66] DOLHANSKY B, BITTON J, PFLAUM B, et al. The

DeepFake detection challenge dataset [EB/OL]. https://arxiv.org/abs/2006.07397,2020.
[67] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies [C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA. IEEE, 2008: 1-8.
[68] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. https://arxiv.org/abs/1312.6199,2013.
[69] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. https://arxiv.org/abs/1412.6572,2014.
[70] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world [EB/OL]. https://arxiv.org/abs/1607.02533,2016.
[71] DONG Yinpeng, LIAO Fangzhou, PANG Tianyu, et al. Boosting adversarial attacks with momentum [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, UT, USA. IEEE, 2018: 9185-9193.
[72] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//2017 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA. 2017: 39-57.
[73] BROWN T B, MANE D, ROY A, et al. Adversarial patch [EB/OL]. https://arxiv.org/abs/1712.09665,2018.
[74] CHEN Zhipeng, TONDI B, LI Xiaolong, et al. A gradient-based pixel-domain attack against SVM detection of global image manipulations [C]//2017 IEEE Workshop on Information Forensics and Security (WIFS). Rennes, France. IEEE, 2017: 1-6.
[75] MARRA F, GRAGNANIELLO D, VERDOLIVA L. On the vulnerability of deep learning to adversarial attacks for camera model identification [J]. Signal Processing: Image Communication, 2018, 65: 240-248.
[76] CARLINI N, FARID H. Evading deepfake-image detectors with white-and black-box attacks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA. IEEE, 2020: 2804-2813.
[77] BARNI M, KALLAS K, NOWROOZI E, et al. On the transferability of adversarial examples against CNN-based image forensics [C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 8286-8290.
[78] MA Yun, MAO Xudong, CHEN Yangbin, et al. Generating adversarial examples by adversarial networks for semi-supervised learning [EB/OL]. https://arxiv.org/abs/1801.02610,2019.
[79] CHEN Chen, ZHAO Xinwei, STAMM M C. Generative adversarial attacks against deep-learning-based camera model identification [J]. IEEE Transactions on Information Forensics and Security, 2019: 1.
[80] CHEN Chen, ZHAO Xinwei, STAMM M C. Mislgan: an anti-forensic camera model falsification framework using a generative adversarial network [C]//2018 25th IEEE International Conference on Image Processing (ICIP). Athens, Greece. IEEE. 2018: 535-539.
[81] KIM D, JANG H U, MUN S M, et al. Median filtered image restoration and anti-forensics using adversarial networks [J]. IEEE Signal Processing Letters, 2018, 25(2): 278-282.
[82] NEVES J C, TOLOSANA R, VERA-RODRIGUEZ R, et al. GANprintR: improved fakes and evaluation of the state of the art in face manipulation detection [J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 1038-1048.

作者简介

谭舜泉 男,1980年生,广东台山人。博士,副教授,主要研究方向为多媒体取证,隐写及隐写分析、深度学习。
E-mail: tansq@szu.edu.cn

黎思力 男,1997年生,广东汕尾人。深圳大学硕士研究生,主要研究方向为Deepfake视频检测取证、模型量化。
E-mail: 2070276026@szu.edu.cn

陈保营 男,1997年生,广东湛江人。深圳大学硕士研究生,主要研究图像篡改检测与定位和Deepfake视频检测取证。
E-mail: 1900271059@szu.edu.cn

李 斌 男,1982年生,广东五华人。博士,教授,主要研究方向为多媒体信息安全、智能信息处理。
E-mail: libin@szu.edu.cn