综述

基于可视数据的可信身份识别和认证方法

彭春蕾[1]，高新波[2]，王楠楠[1]，李洁[1]
（1. 西安电子科技大学综合业务网理论及关键技术国家重点实验室，陕西 西安 710071；
2. 重庆邮电大学重庆市图像认知重点实验室，重庆 400065）

摘　要：近年来，深度学习技术在基于视频和图像等可视数据的身份识别和认证任务（如人脸、行人识别等）中得到了广泛应用。然而，机器学习（特别是深度学习模型）容易受到特定的对抗攻击干扰，从而误导身份识别系统做出错误的判断。因此，针对身份识别系统的可信认证技术研究逐渐成为当前的研究热点。分别从基于信息空间和物理空间的可视数据身份识别和认证攻击方法展开介绍，分析了针对人脸检测与识别系统、行人重识别系统的攻击技术及进展，以及基于人脸活体伪造和可打印对抗图案的物理空间攻击方法，进而讨论了可视数据身份匿名化和隐私保护技术。最后，在简要介绍现有研究中采用的数据库、实验设置与性能分析的基础上，探讨了可能的未来研究方向。

关键词：身份识别；可视数据；可信认证；身份匿名化
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020293

Dependable identity recognition and authorization based on visual information

PENG Chunlei[1], GAO Xinbo[2], WANG Nannan[1], LI Jie[1]
1. State Key Laboratory of Integrated Services Networks, Xidian University, Xi’an 710071, China
2. Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: Recently, deep learning has been widely applied to video and image based identity recognition and authorization tasks, including face recognition and person identification. However, machine learning models, especially deep learning models, can be easily fooled by adversarial attacks, which may cause the identity recognition systems to make a wrong decision. Therefore, dependable identity recognition and authorization has become one of the hot topics currently. Recent advances on dependable identity recognition and authorization from both information space

收稿日期：2020−04−26；**修回日期**：2020−11−10
通信作者：高新波，gaoxb@cqupt.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2016QY01W0200，No.2018AAA0103202）；国家自然科学基金资助项目（No.61806152）；陕西省重点研发项目（No.2020ZDLGY08-08）；陕西省自然科学基金资助项目（No.2019JM-289）
Foundation Items: The National Key Research and Development Program of China (No.2016QY01W0200, No.2018AAA0103202), The National Natural Science Foundation of China (No.61806152), The Key Research and Development Program of Shaanxi (No.2020ZDLGY08-08), The Natural Science Basic Research Plan in Shaanxi Province of China (No.2019JM-289)

and physical space were presented, where the development of the attack models on face detection, face recognition, person re-identification, and face anti-spoofing as well as printable adversarial patches were introduced. The algorithms of visual identity anonymization and privacy protection were further discussed. Finally, the datasets, experimental protocols and performance of dependable identity recognition methods were summarized, and the possible directions in the future research were presented.

Key words: identity recognition, visual information, dependable recognition, identity anonymization

1 引言

身份识别和认证是判断目标身份的过程，其中识别身份的依据有用户口令验证、指纹识别、人脸识别、行人重识别等多种形式[1]。近些年来，随着深度学习等人工智能技术在视频和图像处理分析方面的快速发展，基于视频或图像等可视数据的身份识别和认证研究，即通过提取视频或图像中目标的生物特征信息来判断其身份，逐渐受到广泛的关注和应用。根据视频或图像等可视数据中目标与数据采集设备的距离不同，可有效利用的生物特征信息有所区别，基于可视数据的身份识别方法包括远距离场景下的行人重识别[2]、步态识别[3]以及近距离场景下的人脸检测与识别[4-5]等。

然而，机器学习特别是深度学习模型易受到对抗样本的攻击，即通过对输入数据的微小修改来影响甚至误导模型产生错误的输出。Szegedy等[6]在2013年首次发现神经网络容易受到对抗样本攻击的问题，通过最大化神经网络的预测误差，可以对特定图像增加人眼不易察觉的微小扰动使神经网络分类结果出错，并提出了首个对抗样本攻击方法。然而，该方法存在运行速度慢的不足，因而 Goodfellow 等[7]提出一种快速的对抗样本生成方法。该方法利用深度网络的线性假设，即在使用深度网络过程中往往将其在高维也看作近似的线性模型，导致容易受到对抗样本的攻击。为了在对抗样本生成过程中尽可能少地对图像内容进行修改，Moosavi-Dezfoolis 等[8]提出一种基于线性近似的迭代攻击方法，通过寻找输入图像样本和模型判别边界的最短距离，可以实现比参考文献[7]更少的图像修改。更进一步地，Su 等[9]提出一种极端的对抗样本生成方法，采用差分计算的思想，只需要修改图像中的单个像素即可欺骗深度网络模型，此类少量像素修改的方法使得对抗样本更加难以被人眼察觉。考虑到对抗样本攻击对深度神经网络的干扰，如何有效地检测对抗样本也是重要的研究课题之一。清华大学朱军等提出一种鲁棒的对抗样本检测方法[10]，利用深度神经网络学习输入样本的低维表示来区分对抗样本与正常图像。中国科学院信息工程研究所操晓春团队提出一种端到端的图像压缩模型来检测对抗样本[11]，通过一个图像压缩网络在保持图像内容结构的同时消除其中的对抗扰动，进而利用图像重建网络恢复高质量的原始图像。该方法可以作为应对对抗样本攻击的预处理方法，将对抗图像转化为不包含对抗扰动的重建图像，进而完成图像分类或识别任务。关于对抗样本攻击与防御的最新进展见参考文献[12-13]。中国科学院信息工程研究所陈恺团队近期发表了一个更全面的关于深度学习系统安全性的综述文章[14]，探讨了模型提取攻击、模型逆向攻击、模型注入攻击和对抗攻击等攻击类型对深度学习模型的隐私和安全影响。本文和上述参考文献的不同在于，目前尚未有针对基于可视数据的身份识别和认证系统安全性和隐私保护等问题进行总结分析的工作，因而本文侧重于讨论基于可视数据的可信身份识别研究思想和模型方法之间的异同，并对当前研究方法的局限性和发展趋势进行总结展望。

在2019年3月《Science》杂志发表的论文[15]

讨论了在医疗领域的机器学习模型面临对抗攻击的问题，包括针对医学图像内容识别算法和针对医疗病例文本的对抗攻击。随后，同年 10 月的《Nature》杂志上发表的论文[16]分析了为什么深度学习模型容易被攻击，指出尽管深度学习模型有很深的网络层数，但也意味着这些层在分析输入数据的不同特征时有可能会学习到局部纹理或背景等线索与输入目标之间的联系，从而导致输入内容的很小变化就可以使深度神经网络的输出结果发生明显变化。上述隐患也同样存在于基于可视数据的身份识别系统中，从而误导身份识别系统做出错误判断，产生严重的安全隐患。因此，针对身份识别和认证系统的可信研究尤为重要。可信身份识别的主要内涵就是确保身份识别系统的安全性与可靠性[17]，如图 1 所示，本文从信息空间的可视数据身份识别攻击、物理空间的可视数据身份识别攻击以及身份匿名化与隐私保护 3 个方面对可信身份识别研究进行介绍。其中，信息空间的可视数据身份识别攻击根据身份识别依据不同可以分为针对人脸识别系统的攻击方法和针对行人重识别系统的攻击方法；而物理空间的可视数据身份识别攻击根据攻击方式的不同可以分为基于人脸活体伪造的攻击方法和基于可打印对抗图案的攻击方法。与此同时，本文还从可视数据身份匿名化和身份属性隐私保护两个方面对身份识别系统的隐私泄露与保护研究进行讨论。最后，介绍了相关研究中采取的数据库与实验设置，并对代表性方法的性能进行分析，进而探讨了潜在的研究方向与发展趋势。

2　信息空间的可视数据身份识别攻击

根据身份识别过程中目标与可视数据采集设备的距离不同，从近距离地针对人脸检测和识别系统的攻击方法开始讨论，进而介绍远距离场景下针对行人重识别系统与步态识别系统的攻击方法。最后，针对其他目标识别系统的代表性攻击方法进行简单介绍，以提供信息空间的可视数据身份识别攻击研究的借鉴和算法改进思路。

2.1　针对人脸识别系统的攻击方法

由于对抗样本攻击技术对基于人脸数据的身份认证系统安全性具有严重的威胁，近年来人脸检测和识别系统的可信研究逐渐成为热点问题之

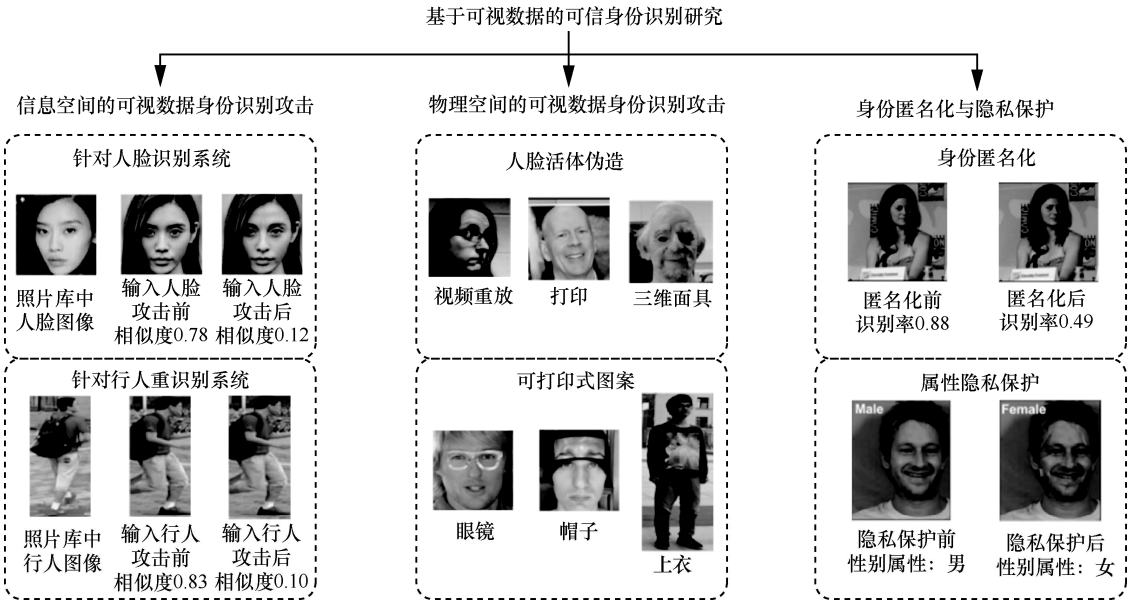


图 1　基于可视数据的可信身份识别研究分类

一。例如，通过对抗攻击可以误导人脸识别系统将输入人脸识别为特定的身份，从而对目标身份进行伪装假冒；也可以欺骗人脸识别系统做出非特定的错误判断，从而扰乱基于人脸识别的身份识别系统正常工作。

卡内基梅隆大学研究团队首先考虑了对抗样本攻击对人脸识别算法的影响[18]，并设计了一种眼镜图案来攻击基于深度神经网络的人脸识别模型，在白盒攻击和黑盒攻击两种场景下进行了对抗样本干扰的实验。其中，白盒攻击指的是可以获取到被攻击网络模型的所有信息和参数，而黑盒攻击无法获得被攻击模型的参数信息，只能根据模型的输入和输出信息来生成对抗样本。该方法进一步拓展到针对人脸检测算法的攻击，由于现有人脸识别系统中往往都包含人脸检测和识别两个环节，针对人脸检测算法的攻击使识别系统无法检测到人脸图像将产生更严重的隐患。

Goswami 等[19]设计了包括图像失真处理、人脸局部区域失真处理和生成对抗样本在内的多种攻击方案来测试基于深度学习的人脸识别算法在面对对抗攻击时的鲁棒性问题，并提出一种面对对抗攻击的检测方法，通过检测真实人脸与经过微小修改后的对抗人脸在深度神经网络模型中间层特征的区别来防御对抗样本的攻击。Song 等[20]提出一种基于视觉注意模块的对抗样本生成网络，通过引入视觉注意模块来学习输入人脸和特定目标人脸图像之间的语义联系，并将人脸识别网络作为判别模型添加到生成框架中，实现误导识别网络将输入人脸识别为特定错误身份的目的。在针对人脸识别系统的攻击过程中，保障图像在添加对抗干扰后的视觉质量也十分重要，因此密西根州立大学 Jain 团队提出一种高质量对抗人脸图像生成方法[21]。该方法同样通过在人脸图像中添加对抗干扰的方法完成特定目标或非特定目标的对抗攻击，并且可以通过用户交互控制所添加对抗干扰的程度。最近，Wang 等[22]提出一种新的基于图像形变的对抗攻击方式，可以在不添加额外对抗噪声的前提下，仅借助语义信息对人脸图像进行形变如面部表情变化实现对抗攻击，从而更加不容易被人眼所察觉。

考虑到人脸检测是人脸识别系统中不可或缺的预处理环节，针对人脸检测模型的攻击将具有更大的威胁性。多伦多大学 Bose 等[23]提出一种针对人脸检测模型 Faster R-CNN[24]的对抗攻击方法，通过对抗生成网络算法在人脸图像中添加微小扰动后，可以生成针对人脸检测模型的对抗样本，从而使人脸图像不被人脸检测模型所发现，其对应身份信息也就无法被识别。Yang 等[25]提出可以在人脸图像中添加一种通用的对抗图像块来欺骗人脸检测系统，从而无须对每张输入人脸图像进行单独学习。

此外，人脸属性信息在基于人脸数据的身份识别过程中是重要的辅助信息，通过欺骗人脸属性检测模型做出错误的判断，也可能误导身份识别的结果。Rozsa 等[26]对人脸属性检测模型的对抗攻击问题进行研究，提出一种快速属性翻转算法，可以误导人脸属性检测模型做出相反的属性判断。欺骗人脸属性检测模型的对抗攻击方式一方面将对人脸识别系统的可靠性产生干扰；另一方面也可以应用到人脸的隐私保护任务中，如保护人脸图像所对应身份的性别、年龄等属性隐私，本文将在后面进行详细讨论。

2.2 针对行人重识别系统的攻击方法

行人重识别系统在公共安全领域特别是基于监控视频的目标身份重识别与跨监控区域运动轨迹追踪等具有重要的意义和应用价值，然而目前关于行人重识别系统的可信研究仍处于起步阶段。针对行人重识别系统进行对抗攻击，将欺骗行人重识别算法对输入目标身份产生错误的重识别结果，从而可能误导和干扰对目标轨迹的有效追踪。

牛津大学研究团队首次提出面向行人重识别

算法的对抗攻击与防御问题[27]。考虑到行人重识别大多通过相似度距离矩阵来判断两幅行人图像是否属于同一身份，发现现有的距离矩阵容易受到对抗攻击，并提出了一种对抗矩阵攻击的方法来干扰行人重识别算法的性能。与此同时，该团队提出一种矩阵保持网络来应对和防御针对距离矩阵的对抗攻击。西安交通大学与厦门大学研究团队提出一种针对行人重识别算法的通用对抗干扰方法[28]。考虑到行人重识别任务中往往更关注重识别后结果的排序情况，而现有的对抗样本攻击大多针对图像分类模型进行干扰而难以直接应用到对排序结果的干扰当中。因此，参考文献[28]提出基于与排序结果相关的攻击目标来实现对抗干扰，并通过一种全局变量最小化方案提高对抗攻击的泛化能力。

在远距离的身份识别方法中，除了行人重识别以外，步态识别也是一种重要的身份识别方式，因此针对步态识别系统的攻击所产生的干扰和隐患也是值得关注的研究课题。北京航空航天大学王蕴红团队提出一种基于伪造视频生成的步态识别系统攻击方法[29]，借助生成对抗网络模型并根据特定目标的步态视频和特定背景即可生成该目标的伪造视频，并设计了多尺度生成模型以提高伪造视频的逼真度，来欺骗步态识别系统。中国科学院自动化所研究团队近期提出一种时域稀疏对抗攻击的方法对步态识别系统进行攻击[30]。由于步态识别研究中往往采用行走过程中的黑白轮廓图作为步态特征，使得其与传统图像分类任务的数据类型有很大差异，且不同目标身份的步态差异在黑白轮廓图序列中往往难以有效区分，导致在图像分类任务中采取的对抗样本生成方法难以直接应用到步态识别的对抗攻击中。因此，参考文献[30]提出利用不受限对抗样本生成策略，在高斯分布初始化基础上生成时域稀疏的步态黑白轮廓图作为对抗样本数据，并将生成的对抗图像替换或插入原始的步态序列中，干扰步态识别算法的结果。

2.3　针对其他目标识别系统的攻击方法

在信息空间中除了上述人脸识别、行人重识别和步态识别等基于可视数据的可信身份识别研究外，还有面向非身份数据的目标检测、语义分割、视频分类等任务中的对抗攻击与防御研究。尽管这些研究方法未直接应用到可视身份识别任务中，但这些相关领域的研究技术可以为可信身份识别研究提供丰富的借鉴和算法提升思路。

例如，在目标检测任务上，Lu 等[31]首次提出针对目标检测模型的对抗攻击方法。在语义分割任务上，参考文献[32]首次展示了对抗攻击可以欺骗基于深度神经网络的语义分割模型，从而在保持大部分目标分割结果不变的情况下，移除某一特定类别的分割，如在室外监控图像中移除行人目标的分割结果，给场景内容理解造成隐患。在显著性区域检测任务上，图像显著性检测模型的准确性和鲁棒性对于图像内容解析和目标检测等具有重要的作用。Che 等[33]首次对图像显著性检测模型的对抗样本攻击问题开展研究，提出一种基于稀疏特征空间的对抗攻击方法，通过对图像添加稀疏的对抗干扰，误导和改变显著性区域的检测结果。图像缩放操作也可能受到对抗攻击的影响，参考文献[34]研究通过对抗攻击算法可以使图像在缩放操作后的结果误导图像分类模型产生错误的判断。这一隐患也会对基于可视数据的身份识别产生威胁，如在安防监控场景中采集到的人脸或行人图像分辨率往往存在差异，需要先通过图像缩放操作将分辨率对齐后进行身份识别，因此图像缩放产生的对抗干扰也会误导身份识别系统的结果。

除了在静态图像上添加对抗扰动外，清华大学朱军等首次提出针对视频分类的稀疏对抗干扰方法[35]。在面向视频数据的对抗攻击过程中，不仅要考虑视频帧的空域信息，还需要考虑到视频序列的时域关系。为了提高对抗攻击的效率和隐

蔽程度，应当在尽可能少的视频帧上进行对抗干扰。基于上述考虑，参考文献[35]假设针对视频的对抗噪声添加在时域上是稀疏分布的，并且所添加的干扰可以传播影响到相邻的视频帧，并提出一种以视频行为识别模型为攻击目标的白盒攻击方法。Jiang 等[36]提出一种针对视频行为识别模型的黑盒攻击方法。现有黑盒攻击方法往往需要对目标模型进行大量次数的查询操作，以根据模型输入和输出关系作为黑盒攻击的线索。然而，当输入数据为包含大量图像帧的视频时，在查询过程中所需要的运算成本将成倍增强，使得难以将面向图像的黑盒攻击方法直接应用到视频数据中。因此，参考文献[36]提出两阶段的黑盒攻击方法，首先对视频帧添加通用的试探性对抗干扰，进而通过选择策略对局部对抗干扰进行调整，从而在只需要少量查询次数的情况下实现对视频识别模型的黑盒攻击。最近，Naeh 等[37]提出在视频帧中添加一个不包含空域信息的固定偏差值作为对抗干扰来攻击视频行为识别模型的方法。由于视频在拍摄过程中随着拍摄场景和拍摄手法的改变出现光照变化，而参考文献[37]中所提出的添加固定偏差值的对抗干扰在视觉上可以模拟光照改变的效果，更加不易被察觉。与此同时，该方法中由于并未在视频帧内添加常规的空域噪声，因而也难以被现有的对抗样本检测方法所发现。

除了通过添加对抗干扰来误导深度神经网络模型做出错误判断外，上海交通大学杨杰团队提出一种新的对抗攻击方式[38]，对输入图像做出巨大的改动而欺骗神经网络模型输出结果保持不变。将这类攻击方式定义为统计学中的第一类错误，如可以将人脸图像编辑为视觉上完全不同的样貌，但人脸识别系统仍然误识为同一身份。该团队提出一种有监督的变分自编码器模型来实现第一类错误的对抗攻击，并以手写数字识别和人脸识别任务为例验证了这一新型对抗攻击的性能。上述针对其他任务中的对抗攻击研究可以为本文所关注的基于可视数据的可信身份识别提供一定的借鉴和启发。

3 物理空间的可视数据身份识别攻击

参考文献[39]首次指出对抗样本攻击可以应用到现实物理空间中，该方法通过将对抗图案打印出来，利用手机拍照后可以误导图像分类软件产生错误的判断。物理空间中的可视数据身份识别攻击需要面临更多的挑战，如打印过程中的颜色失真对攻击效果的影响、物理空间中复杂的光照环境、视频或图像的拍摄角度变换以及打印后对抗图案的移动、旋转及变形等复杂情况下对攻击鲁棒性的干扰等。根据物理空间中攻击方式的不同，本文将从基于人脸活体伪造的攻击方法和基于可打印对抗图案的攻击方法两个方面展开讨论。其他类型的可打印对抗攻击场景，如对交通标识检测和识别系统的攻击[40]、车辆目标检测系统攻击[41]等由于不涉及基于可视数据的身份识别任务，在本文中暂不做讨论。

3.1 基于人脸活体伪造的攻击方法

人脸活体伪造攻击指的是在物理空间中利用可以冒充特定身份的虚假人脸欺骗人脸识别系统，如利用打印照片、重放视频、三维硅胶面具等形式。人脸活体伪造与检测方法主要根据真实的活体人脸与非活体虚假人脸之间的差异实现。例如，传统的人脸活体伪造与检测方法可以利用眨眼动作检测[42]、心率信号检测[43]等线索实现。

中国科学院自动化研究所李子青团队首次将神经网络引入人脸活体伪造与检测任务中[44]。密西根州立大学刘小明团队在人脸活体伪造检测中分别利用卷积神经网络提取深度信息和利用循环神经网络提取心率信息[45]，在活体检测性能上首次超过了传统方法。电子科技大学马争团队提出利用卷积神经网络模型级联长短期记忆网络模型来学习视频帧之间的眨眼、嘴部运动和头部摆动等运动特征[46]，从而利用视频中人脸的时域动态

作为检测依据。腾讯优图研究团队近期提出一种基于反射光的实时人脸活体伪造检测方法[47]，首先对真实或伪造人脸的反射光线进行分析，进而利用多任务卷积神经网络提取深度图进行活体检测。Yu 等[48]首次利用神经网络架构搜索（neural architecture search，NAS）解决活体伪造检测问题，以克服现有方法所采用的人工设计网络结构的不足。

由于人脸活体伪造的类型多样，且实施攻击的物理环境如光照、背景、采集设备分辨率等具有差异化，因而提高人脸活体伪造检测方法的泛化能力来应对未知场景和未知类型的攻击是当前最新的研究热点。Li 等[49]最先考虑到变化的光照、相机质量等条件下活体检测的泛化能力问题，并提出一种无监督的领域自适应方法来应对实际应用场景中的人脸活体检测。腾讯人工智能实验室与上海交通大学研究团队考虑到现有方法大多只局限于在单个数据集上进行模型训练[50]，导致在实际场景下模型泛化能力不足，提出一种基于数据搜索和数据合成的解决方案，并构建了基于时空信息的活体检测网络，分别提取空域信息、视觉注意区域信息和时域信息作为检测的特征。密西根州立大学刘小明团队将未知类型的活体伪造攻击检测定义为零次学习问题[51]，提出一种基于深度树状结构网络的方法将未知类型的攻击样本划分到最相似的数据类别中进行鉴别，并发布了一个包含 13 种活体伪造类型的大规模数据库。为了研究在跨模态、跨人群种族场景下的活体检测问题，Liu 等[52]发布了一个包含彩色图像、深度图像和近红外图像 3 种模态以及 3 种不同人群种族的人脸活体伪造检测数据库，并提出一种基于静态和动态特征融合的检测方法。

目前关于物理空间中人脸活体伪造的攻击和检测研究方法很多，上述讨论中只介绍了代表性的研究思路，更多和人脸识别活体检测相关的研究介绍见参考文献[53-54]。

3.2　基于可打印对抗图案的攻击方法

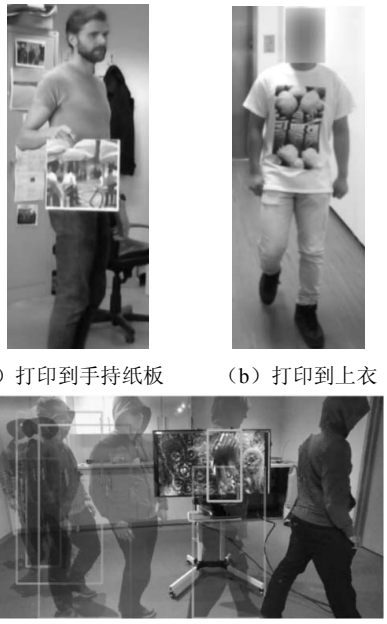
将信息空间中生成的对抗样本图案打印出来可以在物理空间中对可视身份识别系统进行攻击。卡内基梅隆大学 Sharif 等[55]在信息空间中基于眼镜图案的对抗攻击方法[18]，将其打印出来验证在物理空间中欺骗人脸识别系统的效果。将对抗图案打印出来以在物理空间中进行攻击面临更多的挑战性问题，如需要考虑打印后对抗图案在光照环境变化以及佩戴者姿态角度变动时的对抗攻击鲁棒性问题，以及对抗图案中的颜色能否与现有的普通彩色打印机相匹配以避免打印后颜色失真等。此外，考虑到打印的对抗眼镜图案需要用户进行佩戴，因此还需要考虑对抗图案尽可能贴合真实的眼镜框款式以避免用户佩戴过程中引起他人视觉注意。加州大学伯克利分校研究团队提出通过污染训练数据的方法攻击人脸识别系统[56]。该方法在人脸识别系统的训练数据集中人为添加少量的对抗样本，如佩戴特定款式眼镜的人脸图像。随后，在人脸识别系统投入使用后，当测试用户佩戴该特定款式眼镜时将被识别为特定的假冒身份，而佩戴其他类型眼镜时则不会触发该后门攻击，使得该方法在测试阶段具有很好的隐蔽性。

莫斯科国立大学和华为莫斯科研究中心的 Komkov 等[57]提出一种在帽沿上粘贴对抗图案的方法攻击人脸识别系统。该方法将对抗样本图案用彩色打印机打印出来，然后粘贴到用户佩戴的帽子和额头之间的帽沿区域，即可误导人脸识别系统做出错误的判断。除此之外，打印后的对抗图案还可以粘贴人脸面部其他区域，如鼻梁、腮部区域等，均会影响人脸检测和识别系统的结果可信性。

除针对人脸检测和识别系统的可打印对抗图案攻击外，还可以将可打印对抗图案用于攻击行人目标检测和行人重识别系统，以实现在监控场景内的隐身效果。例如，在安防监控场景中采集

到的行人视频规模通常很大，难以通过人工进行逐帧辨识，因而当嫌疑目标穿戴或手持可打印对抗图案经过监控区域时将欺骗和误导行人检测与行人重识别算法，从而避免被发现。Thys 等[58]提出通过将对抗图案打印到纸板上并在胸前手持后即可欺骗人体目标检测模型。为了评估对抗样本图案在经过彩色打印机制作后是否会出现颜色偏差，该方法中提出一种可打印分数损失函数来衡量对抗图案中的像素值与常见打印机中的可选颜色之间的差异。然而，参考文献[58]将对抗图案打印到手持纸板上在实际应用场景中存在较大的使用局限性，因此美国东北大学与麻省理工学院提出一种可以直接打印到上衣 T 恤上的对抗图案生成方法[59]。与打印到手持纸板上相比，将对抗图案打印到衣服上不仅要考虑打印过程中的颜色偏差，还要考虑衣服在穿戴过程中的褶皱等不规则形变对攻击效果的干扰。武汉大学王骞团队提出一种针对行人重识别系统的物理空间对抗攻击方法[60]。该方法分别考虑了无特定目标的逃逸攻击与有特定目标的伪装攻击两种攻击场景，均通过将对抗图案打印到衣服上达到欺骗行人重识别系统的目的。无特定目标攻击过程中，可以欺骗行人重识别系统将穿戴有对抗图案衣服的用户识别为任意其他身份。而在有特定目标攻击场景下，可以误导行人重识别将用户错识为特定的身份。两种攻击场景均会对基于视频监控数据的身份识别系统产生严重的威胁。图 2 中展示了将对抗图案打印到手持纸板[58]、打印到上衣[59]以及打印到背景板[61]等物理空间的攻击类型对行人目标检测系统的干扰。

和上述攻击场景下需要用户手持或穿戴打印后的对抗图案不同，Wiyatno 等[61]提出可以将对抗图案打印成海报或背景板等形式放置到监控画面中，而无须用户亲自穿戴的攻击方法。当行人目标经过该监控场景时，行人目标跟踪算法将聚焦到对抗图案上，而使得跟踪结果出现漂移或跟踪丢失。



（a）打印到手持纸板　　（b）打印到上衣

（c）打印到背景板

图 2　针对行人目标检测系统的可打印对抗图案攻击示例

4　可视数据身份匿名化与隐私保护

随着基于人脸、行人等可视数据的身份识别技术快速发展与应用，关于可视数据中身份的隐私泄露问题也逐渐引起人们的担忧。可视数据身份匿名化与隐私保护，指的是在面对身份识别系统时将可视视频或图像数据转化为匿名化后的结果，从而在保持原有图像空域结构信息的基础上移除其中的身份信息，在不影响人眼视觉观看与辨识的前提下，避免被身份识别技术所发现以实现隐私保护的目的。考虑到越来越多的个人照片在互联网上进行分享，对这些照片中的身份信息进行隐私保护具有重要的意义。值得注意的是，身份匿名化与攻击身份识别系统的区别在于[62]，针对身份识别系统的攻击通常具有特定的攻击目标，即利用对抗样本生成等技术欺骗特定的身份识别系统，而身份匿名化往往没有特定的针对性，旨在通过改变可视数据本身包含的身份信息或特定属性来实现身份的“去识别（de-identification）”[63]。

针对人脸图像的身份信息移除与匿名化研究最早开始于 2015 年，参考文献[64]尝试将人脸图

像进行模糊处理或像素化处理来实现身份匿名化与视觉观看质量之间的平衡。Newton 等[65]提出一种名为 *K*-Same 的人脸图像去识别算法，将每个需要匿名化的人脸图像用其 *K* 个最相似的人脸图像平均值进行表示。然而，上述对人脸图像模糊化、像素化或平均化的操作均会在视觉上引入严重的噪声，影响匿名化后人脸图像的视觉质量。基于传统机器学习方法的可视数据中身份匿名化与隐私保护研究见参考文献[66]，近年来随着深度学习技术的发展和在可视数据匿名化与隐私保护方面的应用，匿名化后的可视数据在视觉质量和鲁棒性方面均取得一定进展。下面将对近年来基于深度学习的最新研究进行讨论。

Meden 等[67]将卷积神经网络与 *K*-Same 算法[65]相结合，利用卷积神经网络和训练集中的相似人脸图像生成高质量的匿名化人脸，并且可以控制匿名化后人脸的表情、年龄、性别等属性特征。挪威科技大学研究团队提出一种基于生成对抗网络的人脸匿名化方法[68]，可以利用条件生成对抗网络制作与输入人脸具有相同姿态和背景的匿名图像。参考文献[69]同样提出利用生成对抗网络，并在网络模型基础上增加了验证模块来移除身份特征信息，以及归一化模块来保持匿名化后人脸与原图像之间的结构相似度，从而生成去身份后的高质量匿名人脸图像。德国萨尔大学 Sun 等[70]提出基于人脸图像修复的面部隐私保护方法。该方法首先通过人脸关键点检测获取面部姿态等信息，然后将原图像中的人脸区域简单裁剪去除或模糊处理，最后以图像修复的思想将裁剪或模糊处理后的面部区域恢复出来，得到身份信息隐藏后的结果。除针对图像数据的匿名化外，Facebook 人工智能实验室 Gafni 等[71]近期首次提出面向视频的实时人脸匿名化方法。视频数据中的身份匿名化不仅仅需要去除其中可用于识别的身份信息，还要保持视频帧之间人脸的姿态、光照、表情等视觉感知一致性和视频质量。参考文献[71]在对抗自编码器模型基础上，添加了人脸分类模型提取目标人脸特征嵌入对抗自编码器的隐层特征上，实现在修改视频中人脸的眉毛、眼睛的五官细节特征的同时保持姿态、表示和肤色等与身份无关的其他视觉特征不变，以移除视频中人脸的身份信息。上述方法只考虑了对可视数据中的身份匿名化过程，而未考虑如何去除匿名化效果的逆过程问题。Gu 等[72]提出一种基于密码验证的人脸图像匿名化与去匿名化方法。该方法一方面可以对人脸图像移除其中的身份信息进行匿名化操作；另一方面在给定正确密码的情况下可以进行去匿名化操作恢复原始人脸图像，而当所给密码错误时将生成包含错误身份的人脸图像结果，具有很强的身份隐私保护功能。

除了对可视数据中的身份信息进行匿名化外，针对其他软生物特征如性别、年龄等属性的隐私保护也具有重要的意义。密西根州立大学 Othman 等[73]首次提出针对人脸软生物特征的隐私保护概念，即通过对人脸图像进行处理使其无法被检测到性别、年龄、种族等属性信息的同时仍然可用于身份识别，从而可以在提交给身份识别系统可视数据时使数据中仅包含身份信息而不会泄露其他属性隐私。参考文献[73]采取了人脸图像融合的思路，选择一张相反性别属性的人脸图像与原图像进行面部区域的形变和融合，以实现对性别属性的隐藏。密西根州立大学 Mirjalili 等[74]利用对抗样本生成的思路对人脸图像中添加有关性别属性的对抗干扰，从而在欺骗属性检测模型对人脸性别属性产生错误判断的同时，保持其他用于人脸身份识别的特征不变，以保护人脸数据中的性别隐私。在此基础上，Chhabra 等[75]进一步利用对抗干扰实现了多个人脸属性隐私的同步匿名化和保护算法。在实际应用中，身份匿名化算法的泛化能力对于应对未知的属性隐私检测模型具有重要意义，因此，Mirjalili 等[76]提出一种基于半对抗网络模型的多种人脸属性隐私保护方法，

通过将人脸图像投影到性别、年龄和种族 3 个互相独立的属性子空间，实现在保持人脸图像的身份识别功能基础上同步完成多个属性的隐私保护，并可以推广到未知属性分类场景当中。

5 数据库介绍及性能分析

本文接下来主要对信息空间和物理空间中代表性的可视数据身份识别攻击方法进行数据库和实验性能分析。表 1 中列出了在信息空间中的可视数据身份识别系统攻击方法实验概况，包括攻击任务类型、实验所采取的数据库和被攻击的身份识别模型。

在攻击人脸识别系统的代表性方法中，通常采取大规模的人脸数据集作为实验数据库，如LFW[79]、CelebA[84]和 CASIA-WebFace[77]等，以主要用于攻击模型当中深度网络部分的训练。现有研究所针对的攻击目标模型大多是目前代表性的基于深度学习的人脸识别算法，如VGG-Face[86]、SphereFace[80]和 ArcFace[81]等。所取得的攻击效果和性能大多可以成功地降低基于人脸数据的身份识别准确率等指标。例如，参考文献[20]利用基于视觉注意模块的对抗生成网络对人脸识别系统进行黑盒攻击和白盒攻击，并以对抗人脸图像攻击成功的平均准确率（mean average precision，mAP）、对抗人脸与真实人脸的余弦距离相似度以及客观图像质量评价作为评估指标。以攻击 ArcFace[95]人脸识别系统为例，在白盒攻击场景下平均准确率 mAP 为 56.26%，而黑盒攻击场景下 mAP 为 9.07%。实验结果可见由于黑盒攻击时无法获取身份识别系统的网络结构和参数信息，因而攻击难度更大。参考文献[21]提出基于对抗干扰的人脸合成方法，并分别验证了在攻击人脸识别系统过程中无特定目标身份的混淆攻击（obfuscation attack）和有特定目标身份的伪装攻击（impersonation attack）两种攻击场景。该方法采用攻击成功率和客观图像质量评价作为评估指标，在 FaceNet[83]、SphereFace[80]和 ArcFace[81]等人脸识别系统上进行攻击实验。同样以攻击 ArcFace[95]人脸识别系统为例，在混淆攻击场景下该方法可以取得 64.53%的攻击成功率而在伪装攻击时可以取得 24.30%的攻击成功率。

在攻击行人重识别系统的方法中，大多采用现有的行人重识别数据集进行模型训练，如Market-1501[87]和 DukeMTMC-reID[88]等。参考文献[27]提出对行人重识别过程中的距离矩阵进行对抗攻击的方式，并且展示了在白盒攻击和黑盒攻击场景下针对两种行人重识别算法 HACNN[89]和 Mancs[90]的实验性能。HACNN 和 Mancs 在 Market-1501 数据集上分别可以取得 75.28%和 82.50%的平均准确率 mAP，而经过黑盒攻击后平均准确率均下降为 40%左右，验证了对抗攻击对行人重识别系统的欺骗和干扰能力。参考文献[28]

表 1 信息空间的可视数据身份识别攻击实验概况

任务描述	代表性方法	实验数据库	被攻击模型
攻击人脸识别系统	参考文献[20]	CASIA-WebFace[77]、MS-Celeb-1M[78]、LFW[79]	SphereFace[80]、ArcFace[81]、MobileFaceNet[82]
	参考文献[21]	CASIA-WebFace[77]、LFW[79]	FaceNet[83]、SphereFace[80]、ArcFace[81]
	参考文献[22]	CelebA[84]、CelebA-HQ[85]	VGG-Face[86]
攻击行人重识别系统	参考文献[27]	Market-1501[87]、DukeMTMC-reID[88]	HACNN[89]、Mancs[90]
	参考文献[28]	Market-1501[87]、DukeMTMC-reID[88]、MSMT17[92]	BOT[91]
攻击步态识别系统	参考文献[29]	CASIA-A[93]、CASIA-B[94]	CNN-Gait[95]、GaitSet[96]
	参考文献[30]	CASIA-B[94]	GaitSet[96]

提出一种针对行人重识别排序结果的通用对抗干扰方法，并在多个行人重识别数据集上进行实验。该方法在攻击 BOT[91]行人重识别系统时可以将平均准确率从攻击前的 65%以上干扰为攻击后平均准确率小于 2%，取得了明显的攻击效果。为了更好地评估对抗攻击的性能，该方法中进一步提出一种平均下降率指标（mean droop rate，mDR），并分别在 Market-1501[87]、DukeMTMC-reID[88]数据集上验证了攻击 BOT[91]行人重识别系统的结果。

在攻击步态识别系统的方法中，采用的实验数据库为目前常用的步态数据集 CASIA-A[93]和 CASIA-B[94]，并以攻击基于深度学习的步态识别系统 CNN-Gait[95]和 GaitSet[96]为主。参考文献[29]提出一种基于视频伪造的步态识别攻击方法，并分别从伪造视频视觉质量、用户调查以及步态识别系统准确率等多个方面进行评估。参考文献[30]提出利用对抗样本生成的策略，直接在步态识别中的黑白轮廓图上添加对抗扰动来攻击步态识别系统。在实验中通过向步态视频序列中添加对抗干扰图像，可以使识别准确率下降 60%以上，并且随着视频中添加对抗图像帧数的增加，攻击效果也更加明显。

表 2 归纳了在物理空间中基于可打印图案的可视数据身份识别系统攻击方法实验概况。由于目前相关研究较少，部分方法如参考文献[55, 59]采取了自建数据集的策略，或者在攻击人脸识别系统时使用现有的大规模人脸数据集进行实验。在攻击模型的选择上，往往选择基于深度学习技术的人脸识别模型 OpenFace[97]、VGG-Face[86]或目标检测模型 Faster R-CNN[98]、YOLOv2[99]进行攻击。参考文献[55]制作对抗眼镜图案并可以打印后佩戴到脸上进行人脸识别系统攻击，并在自建数据集上进行逃逸攻击和伪装攻击实验。实验中发现在逃逸攻击时可以取得 80%以上的视频帧被错误分类的结果，以及在伪装攻击中平均 68%的视频帧中人脸成功伪装成特定目标身份。参考文献[57]提出一种将对抗攻击图案打印并粘贴到帽沿上的人脸识别攻击方法，并在大规模人脸数据集 CASIA-WebFace[77]上进行攻击人脸识别算法 ArcFace[81]的实验。该方法采用余弦距离相似度作为评估指标，在帽沿添加对抗图案后可以将相似度降低 0.5 以上，具有明显的攻击效果。参考文献[59]将对抗图案打印到上衣 T 恤上来攻击 Faster R-CNN[98]和 YOLOv2[99]等目标检测系统，并计算攻击成功的视频帧数与全部测试视频帧数的比值作为攻击成功率（attack success rate，ASR）来评估攻击方法的性能。该方法通过手机录制穿戴有对抗图案的行人移动视频，并分别在室内、室外和训练集中未出现的场景下进行实验。以攻击 YOLOv2 检测模型为例，该方法可以在室内和室外场景下分别取得 64%和 47%的攻击成功率。由于室外场景下视频背景更加复杂，导致攻击性能有所下降，但仍然可以对现有目标检测模型造成严重的威胁。参考文献[60]提出一种将对抗图案打

表 2　物理空间的可视数据身份识别攻击实验概况

任务描述	代表性方法	实验数据库	被攻击模型
打印眼镜图案，攻击人脸识别系统	参考文献[55]	自建数据库	OpenFace[97]、VGG-Face[86]
打印帽沿图案，攻击人脸识别系统	参考文献[57]	CASIA-WebFace[77]	ArcFace[81]
打印纸板图案，攻击行人检测系统	参考文献[58]	Inria[100]	YOLOv2[99]
打印上衣图案，攻击行人检测系统	参考文献[59]	自建数据库	Faster R-CNN[98]、YOLOv2[99]
打印上衣图案，攻击行人重识别系统	参考文献[60]	Market-1501[87]、自建数据集 PRCS[60]	行人重识别算法
打印背景板图案，攻击行人检测系统	参考文献[61]	自建数据库	GOTURN[101]

印到上衣中来攻击行人重识别系统的方法。该方法在行人重识别数据集如 Market-1501[87]和自建数据集 PRCS[60]上进行实验，并对基于深度学习技术的行人重识别算法进行攻击。该方法从视频监控内多个不同距离和不同视角下进行攻击实验并计算实验结果的均值，在逃逸攻击场景下可以达到平均 74.3%的 Rank-1 准确率下降和 61.1%的平均准确率 mAP 下降。与此同时，在伪装攻击场景下成功将穿戴有对抗图案的人员伪装成特定身份的 Rank-1 准确率和平均准确率分别为 47.1%和 67.9%，从而验证了可打印对抗图案攻击对基于行人重识别模型的身份识别系统造成的威胁。

从现有方法的实验评估过程中可以看到，现有基于可视数据的身份识别系统攻击研究仍处于起步阶段，缺乏可靠统一的实验设置标准和性能评估指标。由于现有研究工作中采取的攻击方式和攻击目标模型的多样性，给相关工作的性能对比与分析带来不便。因此，制定统一的身份识别系统攻击性能评估标准是未来需要重点关注的问题之一。

关于人脸活体伪造检测的公开数据集较多，在参考文献[52, 102]中对现有的人脸活体检测数据库进行了很好的总结，本文不再赘述。值得关注的是，人脸活体伪造与检测数据库搜集和实验评估设置逐渐侧重于增加人脸活体伪造的规模、类型以及应对未知攻击类型的泛化能力 3 个方面。特别是提高模型的泛化能力，可以有助于人脸活体检测技术更快更好地投入实际应用当中。例如，中国科学院计算技术研究所韩琥团队提出对抗领域自适应的方法并进行跨数据库实验[103]，即选择一个数据库进行测试时在其他数据库上进行模型训练。参考文献[51]发布了包含 13 种人脸活体伪造类型的大规模数据库 Siw-M，并对三维面具攻击、化妆攻击以及部分区域伪造攻击等类型进行细分，从而有助于设计和验证面对未知类型攻击的模型鲁棒性。参考文献[102]发布了包含多个种族、多种图像模态以及包括打印攻击、重放攻击和三维面具攻击在内的多种伪造类型，可以有助于提高人脸活体检测方法在跨模态、跨种族场景下的泛化能力[104]。

6 结束语

本文对基于可视数据的可信身份识别和认证方法进行了阐述与分析，并分别从信息空间与物理空间的身份识别攻击方法进行讨论。其中，将信息空间的身份识别攻击方法根据身份识别线索的不同划分为近距离场景下针对人脸识别系统的攻击以及远距离场景下针对行人重识别系统与步态识别系统的攻击。在物理空间的身份识别攻击方法中分别讨论了基于人脸活体伪造的攻击方法和基于可打印对抗图案的攻击方法。随后，对可视数据身份匿名化与身份属性的隐私保护进行了介绍。最后，对现有研究中所采用的数据库和实验概况进行了归纳和总结。

尽管基于可视数据的可信身份识别研究已取得一定进展，然而现有研究仍然存在一定的局限性，并且有以下问题待解决。

（1）复杂不可控环境

由于针对身份识别系统的攻击场景复杂且不可控，如光照变化、可视数据采集视角不同、采集设备距离变换引起的尺度失配等挑战，对人脸识别、行人重识别等系统的对抗攻击与防御模型的影响是一个值得研究的课题。

（2）高质量对抗图案

现有的对抗样本生成大多以攻击成功率作为衡量指标，而忽略了所生成对抗图案的视觉质量和图像美感，使得在实际应用中难以直接应用。特别是在物理空间的身份识别系统攻击中，现有的可打印对抗图案在视觉外观上仍然容易引人关注，难以在实际场景中发挥作用。未来研究中应考虑提升对抗图案的视觉质量和图像美感，从而

在保证攻击有效性的前提下设计出更加贴近日常生活和不引人注目的对抗图案。

（3）视觉线索的隐私保护

在可视数据身份匿名化与隐私保护方面，除了对人脸图像中的身份信息和人脸属性进行匿名化处理外，其他视觉信息如人物的衣着、身高、步态等线索也对身份辨识有一定的辅助作用，因而这些视觉线索的隐私保护也值得进一步研究。

（4）性能评估和公开数据集

当前研究中还缺乏统一的算法评估指标和大规模可信身份识别数据集。例如，在对抗样本生成研究中[104]发布了一个真实对抗样本数据集可用于图像分类模型的鲁棒性验证，以及提高分类模型应对对抗样本攻击的能力。然而，目前仍缺乏大规模人脸或行人对抗样本数据集，通过搜集并整理能够误导身份识别系统产生错误判断的真实视频或图像数据，对于提高身份识别系统的鲁棒性和可信研究具有重要意义。

（5）其他可视数据分析领域的可信研究

目前还有很多基于可视数据的身份分析与识别算法的可信研究仍处于空白，如针对跨模态（近红外、素描人脸等）人脸识别系统、人脸表情识别以及行人属性检测等模型的对抗攻击与防御，均是值得关注的研究课题。基于可视数据的可信身份识别与认证研究目前才刚刚起步，由于可信身份识别对于公共安全、信息安全和隐私保护等方面具有重要意义，特别是在物理空间的可信身份识别问题对当前基于监控数据的内容分析与身份识别体系产生了严峻的挑战，在未来需要进一步的关注和研究。

参考文献：

[1] 章坚武, 姚泽瑾, 吴震东. 基于生物特征和混沌映射的多服务器身份认证方案[J]. 电信科学, 2017, 33(2): 22-31.
ZHANG J W, YAO Z J, WU Z D. Multi-server identity authentication scheme based on biometric and chaotic maps[J]. Telecommunications Science, 2017, 33(2): 22-31.

[2] 叶钰, 王正, 梁超, 等. 多源数据行人重识别研究综述[J]. 自动化学报, 2019, 46(9): 1869-1884.
YE Y, WANG Z, LIANG C, et al. A survey on multi-source person re-identification[J]. Acta Automatica Sinica, 2019, 46(9): 1869-1884.

[3] NAMBIAR A, BERNARDINO A, NASCIMENTO J C. Gait-based person re-identification: a survey[J]. ACM Computing Surveys, 2019, 52(2): 1-34.

[4] WANG M, DENG W. Deep face recognition: a survey[J]. arXiv preprint arXiv: 1804.06655, 2018.

[5] 唐彪, 金炜, 符冉迪, 等. 多稀疏表示分类器决策融合的人脸识别[J]. 电信科学, 2018, 34(4): 31-40.
TANG B, JIN W, FU R D, et al. Face recognition using decision fusion of multiple sparse representation-based classifiers[J]. Telecommunications Science, 2018, 34(4): 31-40.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv: 1312.6199, 2013.

[7] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412.6572, 2014.

[8] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2574-2582.

[9] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[10] PANG T, DU C, DONG Y, et al. Towards robust detection of adversarial examples[C]//Proceedings of Advances in Neural Information Processing Systems. Montréal: ACM Press, 2018: 4579-4589.

[11] JIA X, WEI X, CAO X, et al. Comdefend: an efficient image compression model to defend adversarial examples[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 6084-6092.

[12] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019(8): 15.
ZHANG S S, ZUO X, LIU J W. The problem of the adversarial examples in deep learning[J]. Chinese Journal of Computers, 2019(8): 15.

[13] YUAN X, HE P, ZHU Q, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[14] HE Y, MENG G, CHEN K, et al. Towards privacy and security of deep learning systems: a survey[J]. arXiv preprint arXiv: 1911.12562, 2019.

[15] FINLAYSON S G, BOWERS J D, ITO J, et al. Adversarial attacks on medical machine learning[J]. Science, 2019, 363(6433): 1287-1289.

[16] HEAVEN D. Why deep-learning AIs are so easy to fool[J].

Nature, 2019, 574(7777): 163.

[17] 沈昌祥, 张焕国, 王怀民, 等. 可信计算的研究与发展[J]. 中国科学: 信息科学, 2010, 40(2): 139-166.
SHEN C X, ZHANG H G, WANG H M, et al. Research and development on trusted computation[J]. Science China: Information Science, 2010, 40(2): 139-166.

[18] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2016: 1528-1540.

[19] GOSWAMI G, RATHA N, AGARWAL A, et al. Unravelling robustness of deep learning based face recognition against adversarial attacks[C]//Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018.

[20] SONG Q, WU Y, YANG L. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network[J]. arXiv preprint arXiv: 1811.12026, 2018.

[21] DEB D, ZHANG J, JAIN A K. Advfaces: adversarial face synthesis[J]. arXiv preprint arXiv: 1908.05008, 2019.

[22] WANG R, JUEFEI-XU F, XIE X, et al. Amora: black-box adversarial morphing attack[J]. arXiv preprint arXiv: 1912.03829, 2019.

[23] BOSE A J, AARABI P. Adversarial attacks on face detectors using neural net based constrained optimization[C]//Proceedings of 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP). Piscataway: IEEE Press, 2018: 1-6.

[24] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1440-1448.

[25] YANG X, WEI F, ZHANG H, et al. Design and interpretation of universal adversarial patches in face detection[J]. arXiv preprint arXiv: 1912.05021, 2019.

[26] ROZSA A, GÜNTHER M, RUDD E M, et al. Facial attributes: accuracy and adversarial robustness[J]. Pattern Recognition Letters, 2019(124): 100-108.

[27] BAI S, LI Y, ZHOU Y, et al. Metric attack and defense for person re-identification[J]. arXiv preprint arXiv: 1901.10650, 2019.

[28] DING W, WEI X, HONG X, et al. Universal adversarial perturbations against person re-identification[J]. arXiv preprint arXiv: 1910.14184, 2019.

[29] JIA M, YANG H, HUANG D, et al. Attacking gait recognition systems via silhouette guided GANs[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 638-646.

[30] HE Z, WANG W, DONG J, et al. Temporal sparse adversarial attack on gait recognition[J]. arXiv preprint arXiv: 2002.09674, 2020.

[31] LU J, SIBAI H, FABRY E. Adversarial examples that fool detectors[J]. arXiv preprint arXiv: 1712.02494, 2017.

[32] HENDRIK METZEN J, CHAITHANYA KUMAR M, BROX T, et al. Universal adversarial perturbations against semantic image segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2755-2764.

[33] CHE Z, BORJI A, ZHAI G, et al. Adversarial attacks against deep saliency models[J]. arXiv preprint arXiv: 1904.01231, 2019.

[34] XIAO Q, CHEN Y, SHEN C, et al. Seeing is not believing: camouflage attacks on image scaling algorithms[C]//Proceedings of 28th USENIX Security Symposium (USENIX Security 19). Santa Clara: USENIX Association, 2019: 443-460.

[35] WEI X, ZHU J, YUAN S, et al. Sparse adversarial perturbations for videos[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019: 8973-8980.

[36] JIANG L, MA X, CHEN S, et al. Black-box adversarial attacks on video recognition models[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 864-872.

[37] NAEH I, PONY R, MANNOR S. Patternless adversarial attacks on video recognition networks[J]. arXiv preprint arXiv: 2002.05123, 2020.

[38] TANG S, HUANG X, CHEN M, et al. Adversarial attack type I: cheat classifiers by significant changes[J]. arXiv preprint arXiv: 1809.00594, 2018.

[39] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[J]. arXiv preprint arXiv: 1607.02533, 2016.

[40] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1625-1634.

[41] ZHANG Y, FOROOSH H, DAVID P, et al. CAMOU: learning physical vehicle camouflages to adversarially attack detectors in the wild[C]//Proceedings of International Conference on Learning Representations. Vancouver: OpenReview.net, 2018: 1-20.

[42] PAN G, SUN L, WU Z, et al. Eyeblink-based anti-spoofing in face recognition from a generic Web camera[C]//Proceedings of 2007 IEEE 11th International Conference on Computer Vision. Piscataway: IEEE Press, 2007: 1-8.

[43] LI X, KOMULAINEN J, ZHAO G, et al. Generalized face anti-spoofing by detecting pulse from face videos[C]// Proceedings of 2016 23rd International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2016: 4244-4249.

[44] YANG J, LEI Z, LI S Z. Learn convolutional neural network for face anti-spoofing[J]. arXiv preprint arXiv: 1408.5601, 2014.

[45] LIU Y, JOURABLOO A, LIU X. Learning deep models for face anti-spoofing: binary or auxiliary supervision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 389-398.

[46] TU X, ZHANG H, XIE M, et al. Enhance the motion cues for face anti-spoofing using CNN-LSTM architecture[J]. arXiv preprint arXiv: 1901.05635, 2019.

[47] LIU Y, TAI Y, LI J, et al. Aurora guard: real-time face anti-spoofing via light reflection[J]. arXiv preprint arXiv: 1902.10311, 2019.

[48] YU Z, ZHAO C, WANG Z, et al. Searching central difference convolutional networks for face anti-spoofing[J]. arXiv preprint arXiv: 2003.04092, 2020.

[49] LI H, LI W, CAO H, et al. Unsupervised domain adaptation for face anti-spoofing[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(7): 1794-1809.

[50] YANG X, LUO W, BAO L, et al. Face anti-spoofing: model matters, so does data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 3507-3516.

[51] LIU Y, STEHOUWER J, JOURABLOO A, et al. Deep tree learning for zero-shot face anti-spoofing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4680-4689.

[52] LIU A, TAN Z, LI X, et al. Static and dynamic fusion for multi-modal cross-ethnicity face anti-spoofing[J]. arXiv preprint arXiv: 1912.02340, 2019.

[53] GALBALLY J, MARCEL S, FIERREZ J. Biometric antispoofing methods: a survey in face recognition[J]. IEEE Access, 2014(2): 1530-1552.

[54] 蒋方玲, 刘鹏程, 周祥东. 人脸活体检测综述[J]. 自动化学报, 2020(1): 1-24.
JIANG F L, LIU P C, ZHOU X D. A review on face anti-spoofing[J]. ACTA Automatica Sinica, 2020(1): 1-24.

[55] SHARIF M, BHAGAVATULA S, BAUER L, et al. A general framework for adversarial examples with objectives[J]. ACM Transactions on Privacy and Security, 2019, 22(3): 1-30.

[56] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv: 1712.05526, 2017.

[57] KOMKOV S, PETIUSHKO A. AdvHat: real-world adversarial attack on ArcFace face ID system[J]. arXiv preprint arXiv: 1908.08705, 2019.

[58] THYS S, VAN RANST W, GOEDEMÉ T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2019: 1-7.

[59] XU K, ZHANG G, LIU S, et al. Adversarial t-shirt! evading person detectors in a physical world[J]. arXiv preprint arXiv: 1910.11099, 2019.

[60] WANG Z, ZHENG S, SONG M, et al. advPattern: physical-world attacks on deep person re-identification via adversarially transformable patterns[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 8341-8350.

[61] WIYATNO R R, XU A. Physical adversarial textures that fool visual object tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 4822-4831.

[62] SUN Q, MA L, JOON OH S, et al. Natural and effective obfuscation by head inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5050-5059.

[63] GROSS R, SWEENEY L, DE LA TORRE F, et al. Model-based face de-identification[C]//Proceedings of 2006 Conference on Computer Vision and Pattern Recognition Workshop. Piscataway: IEEE Press, 2006: 161-161.

[64] NEUSTAEDTER C, GREENBERG S, BOYLE M. Blur filtration fails to preserve privacy for home-based video conferencing[J]. ACM Transactions on Computer-Human Interaction, 2006, 13(1): 1-36.

[65] NEWTON E M, SWEENEY L, MALIN B. Preserving privacy by de-identifying face images[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(2): 232-243.

[66] PADILLA-LÓPEZ J R, CHAARAOUI A A, FLÓREZ- REVUELTA F. Visual privacy protection methods: a survey[J]. Expert Systems with Applications, 2015, 42(9): 4177-4195.

[67] MEDEN B, EMERSIC Z, STRUC V, et al. κ-Same-Net: neural-network-based face deidentification[C]//Proceedings of 2017 International Conference and Workshop on Bioinspired Intelligence. Piscataway: IEEE Press, 2017: 1-7.

[68] HUKKELÅS H, MESTER R, LINDSETH F. DeepPrivacy: a generative adversarial network for face anonymization[C]//Proceedings of International Symposium on Visual Computing. Berlin: Springer, 2019: 565-578.

[69] WU Y, YANG F, XU Y, et al. Privacy-protective-GAN for privacy preserving face de-identification[J]. Journal of Computer Science and Technology, 2019, 34(1): 47-60.

[70] SUN Q, MA L, JOON OH S, et al. Natural and effective obfuscation by head inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5050-5059.

[71] GAFNI O, WOLF L, TAIGMAN Y. Live face de-identification in video[C]//Proceedings of the IEEE International Conference

on Computer Vision. Piscataway: IEEE Press, 2019: 9378-9387.

[72] GU X, LUO W, RYOO M S, et al. Password-conditioned anonymization and deanonymization with face identity transformers[J]. arXiv preprint arXiv: 1911.11759, 2019.

[73] OTHMAN A, ROSS A. Privacy of facial soft biometrics: suppressing gender but retaining identity[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2014: 682-696.

[74] MIRJALILI V, ROSS A. Soft biometric privacy: retaining biometric utility of face images while perturbing gender[C]// Proceedings of 2017 IEEE International Joint Conference on Biometrics (IJCB). Piscataway: IEEE Press, 2017: 564-573.

[75] CHHABRA S, SINGH R, VATSA M, et al. Anonymizing k-facial attributes via adversarial perturbations[J]. arXiv preprint arXiv: 1805.09380, 2018.

[76] MIRJALILI V, RASCHKA S, ROSS A. PrivacyNet: semi-adversarial networks for multi-attribute face privacy[J]. arXiv preprint arXiv: 2001.00561, 2020.

[77] YI D, LEI Z, LIAO S, et al. Learning face representation from scratch[J]. arXiv preprint arXiv: 1411.7923, 2014.

[78] GUO Y, ZHANG L, HU Y, et al. Ms-celeb-1m: a dataset and benchmark for large-scale face recognition[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 87-102.

[79] HUANG G B, MATTAR M, BERG T, et al. Labeled faces in the wild: a database forstudying face recognition in unconstrained environments[Z]. 2008.

[80] LIU W, WEN Y, YU Z, et al. Sphereface: deep hypersphere embedding for face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 212-220.

[81] DENG J, GUO J, XUE N, et al. Arcface: additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4690-4699.

[82] CHEN S, LIU Y, GAO X, et al. Mobilefacenets: efficient CNNs for accurate real-time face verification on mobile devices[C]//Proceedings of Chinese Conference on Biometric Recognition. Berlin: Springer, 2018: 428-438.

[83] SCHROFF F, KALENICHENKO D, PHILBIN J. FACENET: a unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE Press, 2015: 815-823.

[84] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 3730-3738.

[85] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.

[86] PARKHI O M, VEDALDI A, ZISSERMAN A. Deep face recognition[Z]. 2015.

[87] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: a benchmark[C]//Proceedings of the IEEE international Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1116-1124.

[88] ZHENG Z, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3754-3762.

[89] LI W, ZHU X, GONG S. Harmonious attention network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2285-2294.

[90] WANG C, ZHANG Q, HUANG C, et al. Mancs: a multi-task attentional network with curriculum sampling for person re-identification[C]//Proceedings of the European Conference on Computer Vision. Piscataway: IEEE Press, 2018: 365-381.

[91] LUO H, GU Y, LIAO X, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2019: 1-9.

[92] WEI L, ZHANG S, GAO W, et al. Person transfer GAN to bridge domain gap for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 79-88.

[93] WANG L, TAN T, NING H, et al. Silhouette analysis-based gait recognition for human identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(12): 1505-1518.

[94] ZHENG S, ZHANG J, HUANG K, et al. Robust view transformation model for gait recognition[C]//Proceedings of 2011 18th IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2011: 2073-2076.

[95] WU Z, HUANG Y, WANG L, et al. A comprehensive study on cross-view gait based human identification with deep CNNs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(2): 209-226.

[96] CHAO H, HE Y, ZHANG J, et al. Gaitset: regarding gait as a set for cross-view gait recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019: 8126-8133.

[97] AMOS B, LUDWICZUK B, HARKES J, et al. Openface: face recognition with deep neural networks[C]//Proceedings of IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2016: 6.

[98] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// Proceedings of Advances in Neural Information Processing Systems. Montréal: ACM Press, 2015: 91-99.

[99] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7263-7271.

[100] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2005: 886-893.

[101] HELD D, THRUN S, SAVARESE S. Learning to track at 100 fps with deep regression networks[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 749-765.

[102] LI A, TAN Z, LI X, et al. CASIA-SURF CeFA: a benchmark for multi-modal cross-ethnicity face anti-spoofing[J]. arXiv preprint arXiv: 2003.05136, 2020.

[103] WANG G, HAN H, SHAN S, et al. Improving cross-database face presentation attack detection via adversarial domain adaptation[C]//Proceedings of International Conference on Biometrics (ICB). Crete: IEEE Press, 2019: 1-8.

[104] HENDRYCKS D, ZHAO K, BASART S, et al. Natural adversarial examples[J]. arXiv preprint arXiv: 1907.07174, 2019.

[作者简介]

彭春蕾（1991− ），男，博士，西安电子科技大学综合业务网理论及关键技术国家重点实验室讲师，主要研究方向为计算机视觉、模式识别和机器学习。

高新波（1972− ），男，博士，重庆邮电大学重庆市图像认知重点实验室教授，主要研究方向为多媒体分析、计算机视觉、模式识别与机器学习等。

王楠楠（1986− ），男，博士，西安电子科技大学综合业务网理论及关键技术国家重点实验室教授，主要研究方向为计算机视觉、模式识别和机器学习。

李洁（1972− ），女，博士，西安电子科技大学综合业务网理论及关键技术国家重点实验室教授，主要研究方向为模式识别和计算机视觉。