软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(7):2411−2446 [doi: 10.13328/j.cnki.jos.006590]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

人脸识别反欺诈研究进展*

张 帆[1], 赵世坤[1], 袁 操[1], 陈 伟[2], 刘小丽[3,4], 赵涵捷[5]

[1](武汉轻工大学 数学与计算机学院, 湖北 武汉 430023)

[2](南京邮电大学 计算机学院, 江苏 南京 210023)

[3](暨南大学 信息科学技术学院, 广东 广州 510632)

[4](暨南大学 网络空间安全学院, 广东 广州 510632)

[5](台湾东华大学 电机工程学系, 台湾 花莲 08153719)

通信作者: 赵涵捷, E-mail: hcchao@gmail.com

摘 要: 当前, 人脸识别理论和技术取得了巨大的成功, 被广泛应用于政府、金融和军事等关键领域. 与其他信息系统类似, 人脸识别系统也面临着各类安全问题, 其中, 人脸欺诈(face spoofing, FS)是最主要的安全问题之一. 所谓的人脸欺诈, 是指攻击者采用打印照片、视频回放和 3D 面具等攻击方式, 诱骗人脸识别系统做出错误判断, 因而是人脸识别系统所必须解决的关键问题. 对人脸反欺诈(face anti-spoofing, FAS)的最新进展进行研究: 首先, 概述了 FAS 的基本概念; 其次, 介绍了当前 FAS 所面临的主要科学问题以及主要的解决方法及其优缺点; 在此基础上, 将已有的 FAS 工作分为传统方法和深度学习方法两大类, 并分别进行详细论述; 接着, 针对基于深度学习的 FAS 域泛化和可解释性问题, 从理论和实践的角度进行说明; 然后, 介绍了 FAS 研究所使用的典型数据集及其特点, 并给出了 FAS 算法的评估标准和实验对比结果; 最后, 总结了 FAS 未来的研究方向并对发展趋势进行展望.

关键词: 人脸反欺诈; 呈现攻击检测; 人脸识别安全; 深度学习; 域泛化; 可解释性

中图法分类号: TP18

中文引用格式: 张帆, 赵世坤, 袁操, 陈伟, 刘小丽, 赵涵捷. 人脸识别反欺诈研究进展. 软件学报, 2022, 33(7): 2411–2446. http://www.jos.org.cn/1000-9825/6590.htm

英文引用格式: Zhang F, Zhao SK, Yuan C, Chen W, Liu XL, Chao HC. Research Progress of Face Recognition Anti-spoofing. Ruan Jian Xue Bao/Journal of Software, 2022, 33(7): 2411−2446 (in Chinese). http://www.jos.org.cn/1000-9825/6590.htm

Research Progress of Face Recognition Anti-spoofing

ZHANG Fan[1], ZHAO Shi-Kun[1], YUAN Cao[1], CHEN Wei[2], LIU Xiao-Li[3,4], CHAO Han-Chieh[5]

[1](School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430023, China)

[2](School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

[3](College of Information Science and Technology, Jinan University, Guangzhou 510632, China)

[4](College of Cyber Security, Jinan University, Guangzhou 510632, China)

[5](Department of Electrical Engineering, Dong Hwa University, Hualien 08153719, China)

Abstract: Currently, face recognition theory and technology have achieved great success, and face recognition systems have been widely deployed in key fields such as government, finance, military, etc. Similar to other information systems, face recognition systems also face various security issues, among which, face spoofing is one of the most important issues. The so-called face spoofing refers to the use of attack methods such as printing photos, video re-play, and 3D masks to trick the face recognition system into making false decisions, and

* 基金项目: 国家重点研发计划(2019YFB2101704); 国家自然科学基金(61906140); 湖北省自然科学基金(2020CFB761); 武汉轻工大学科研项目(2021Y38)

本文由“智能系统的分析和验证”专题特约编辑明仲教授、张立军教授和秦胜潮教授推荐.

收稿时间: 2021-09-08; 修改时间: 2021-10-14; 采用时间: 2022-01-10; jos 在线出版时间: 2022-01-28

thus it must be addressed by a face recognition system. The recent progress of face anti-spoofing (FAS) is investigated. Initially, FAS-related concepts are outlined. Then, the main scientific problems of FAS and corresponding solutions, including the advantages and disadvantages of these solutions, are introduced. Next, existing FAS approaches are divided into two folds, i.e., traditional approaches and deep learning-based approaches, and they are depicted in detail, respectively. Moreover, regarding the domain generalization and interpretability issues of deep learning-based FAS, a detailed introduction is given from the perspective of theory and practice. Then, mainstream datasets adopted by FAS are discussed, and evaluation criteria and experimental results based on these datasets are explained as well. Finally, the future research directions are discussed and concluded.

Key words: face anti-spoofing (FAS); presentation attack detection; face recognition security; deep learning; domain generalization; interpretability

1 背景介绍

人脸识别技术具有识别速度快、准确率高和无须用户接触等优点, 已广泛用于门禁、支付等领域. 人脸识别在带来便利的同时也面临着各类信息系统安全问题, 其中最典型的是人脸欺诈攻击(face spoofing attack). 人脸欺诈攻击也称为人脸呈现攻击(face presentation attack), 其利用伪造的“人脸”图像或者视频——如打印照片[1,2]、回放视频[3,4]、3D 面具[5,6]等——诱骗人脸识别系统做出错误判断, 因此是目前人脸识别系统所必须解决的最关键问题之一. 本文对人脸反欺诈(face anti-spoofing, FAS)的最新研究进展进行了调研. 在具体介绍 FAS 的最新进展之前, 先给出 FAS 的相关概念以及 FAS 算法的基本流程.

1.1 FAS相关概念

(1) 真实人脸(real face): 指相机拍摄的、与人脸识别系统所录入的用户身份信息相一致的个体人脸.

(2) 欺骗人脸(spoofing face): 指用于欺骗人脸识别系统的非真实人脸, 典型的如打印照片、回放视频和 3D 面具等.

(3) 打印攻击(print attack): 指攻击者将受害者的人脸打印照片呈现给人脸识别系统以进行欺骗攻击. 攻击时需要对照片进行一定处理, 如弯曲照片以模仿人脸的 3D 结构, 或者裁剪掉照片的眼睛/嘴巴等关键区域,并用真实人脸的关键区域代替以满足系统的眨眼/动嘴要求等.

(4) 重放攻击(replay attack): 指攻击者重放(从网上下载, 或者通过相机拍摄的)受害者的视频信息以欺骗人脸识别系统.

(5) 面具攻击(mask attack): 指攻击者穿戴依据受害者人脸所伪造的 3D 面具, 以欺骗人脸识别系统. 现代高逼真的 3D 面具通常由和皮肤相似的材料制作而来, 可逼真地还原人脸纹理信息, 如皱纹和斑点等.

(6) 时空信息(temporal-spatio information): 时间和空间信息的简称. 本文中, 时间信息指人脸视频中连续多帧图像之间与时序相关的运动信息, 空间信息指单帧图像中的深度(depth)信息.

(7) 多模态(multimodal): 本文将信息的来源称为模态. 因此, 多模态指来源于两种或两种以上不同信息源的图像, 如 RGB 图(red-green-blue)、红外图 IR(infra-red)、深度图(depth)等.

1.2 FAS算法流程

FAS 算法的一般流程图如图 1 所示. 无论是传统的 FAS 方法还是基于深度学习的 FAS 方法, 其基本流程都包含 4 个步骤: 输入、预处理、特征提取和分类. 具体来说:

- 首先, 将图像或者视频作为 FAS 算法的输入.
- 然后, 将输入的图像/视频经过预处理(如人脸裁剪、模态转换等)后进行特征提取.
- 特征提取可分为传统的方法和基于深度学习的方法, 其中, 传统方法通常需要手工设计描述算子以提取可区分特征; 基于深度学习的方法则通过给定损失函数, 在损失函数的监督下自动提取可区分特征.
- 最后, 所提取的特征被送入分类器, 由分类器给出真实人脸(live)或欺骗人脸(spoof)的最终判定结果.

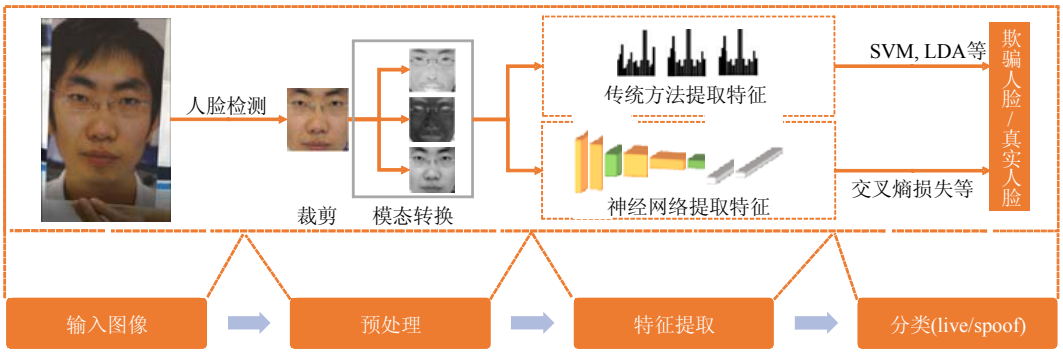


图 1 FAS 流程图, 灰色实线方框为可选步骤

接下来, 本文将对 FAS 的最新研究进展进行详细介绍. 本文第 2 节总结当前 FAS 所面临的主要科学问题以及对应解决方案的思路及其优缺点. 第 3 节和第 4 节分别对传统的 FAS 方法和基于深度学习的 FAS 方法进行详细论述. 第 5 节对基于深度学习的 FAS 方法的域泛化和可解释性两大问题从基本理论和典型方法两方面进行说明. 第 6 节展示 FAS 常用的数据集. 第 7 节介绍 FAS 算法的评估标准以及部分代表性 FAS 算法在不同数据集上的实验比对. 第 8 节对 FAS 未来的研究工作进行展望. 第 9 节进行总结.

2 FAS 面临的主要科学问题和所采用的主要方法

本节对当前 FAS 所面临的主要科学问题以及针对这些问题所发展出的主要理论和技术进行总体介绍, 更详细的方法说明请参见第 3 节–第 5 节.

2.1 FAS面临的主要科学问题

当前 FAS 主要面临如下科学问题.

(1) 域泛化(domain generalization)研究有待加强. 传统的机器学习建立在独立同分布假设(independent identically distributed, i.i.d)的基础上, 即假设训练数据集和测试数据集是相同且独立分布的[7], 但是在实践中, 训练数据集和测试数据集的概率分布往往会存在差异, 这会导致 FAS 模型的性能发生恶化(deteriorate). 域泛化研究的目标在于使得模型对于未知(unseen)的数据集仍然能够保持良好性能, 是目前深度学习的研究热点和难点[7].

(2) 模型可解释性(model interpretability)研究不足. 一方面, 深度学习的黑盒特点使得人们难以理解模型的行为; 另一方面, 深度学习又可在模型自身“高自信”的情况下做出非预期的行为[8,9], 这种冲突使得研究人员越来越关注模型的可解释性[10]. 总体上, 可解释性可以分为结论可解释性和技术可解释性. 从结论可解释性的角度, 研究人员可能并不关注于如何对技术细节进行解释, 而是更加关注于模型做出结论的依据. 例如, 当用于刑事重犯认证时, 系统需要给出认证当前对象是或者不是罪犯的具体依据以供人工审核, 否则若发生误判, 可能造成严重后果. 从技术可解释性的角度, 研究人员则更加关注于模型实现技术的可解释性. 例如, 基于深度学习的 FAS 算法[11−14]需要说明图像的哪部分对最终决策贡献最大以及为什么最大——对于防御方, 上述可解释性可为模型的改进提供思路和方向; 对于攻击方, 上述可解释性研究可以帮助定位模型的弱点, 进而提升针对模型的攻击能力. 可解释性也是目前深度学习研究的热点和难点[10].

(3) 时空(temporal-spatio)融合需要加强. 单幅图像内在的人脸空间(深度)信息以及视频流中不同图像帧之间的人脸时序(运动)信息, 均可以作为人脸识别的特征来源, 然而现有的方法对于时空特征挖掘存在两大问题: 一是许多方法基于单帧图像的高层(high-level)语义进行 FAS 检测, 而忽略了挖掘网络中低层(middle/low-level)的语义信息(如空间梯度振幅等)[15]; 二是基于单帧图像的方法遗失了重要的时序运动特征, 即真假人脸视频流中连续多帧图像之间由于运动所导致的人脸信息差异[15]. 多

数情况下, 多帧方法比单帧方法要好, 但也导致了计算复杂性的上升. 挖掘新的时空可利用信息以及寻求新的时空融合方法, 是目前 FAS 研究的重点方向之一.

(4) 多模态研究欠缺. 现有的 FAS 算法大多从 RGB 图像中提取特征信息, 当面临复杂的场景时, 如光照变化明显的雨天和雾天, 或者攻击者有意遮挡等, 此时从 RGB 图像可提取到的特征将会受限, 进而会影响 FAS 算法的性能. 利用来自不同模态图像(如 RGB 图、红外图和深度图等)之间的互补特性研究多模态 FAS 算法, 可以有效解决上述不足. 多模态研究离不开多模态数据集, 但是当前的多模态数据集非常少见[16]. 为此, 文献[16]首先发布了 CASIA-SURF; 在此的基础上, 文献[17]进一步从模态数量、种族类别、个体数量和攻击类型这 4 个方面对 CASIA-SURF 进行扩充, 将其扩充为目前最大的多模态数据集 CASIA-SURF CeFA. 上述数据集为基于多模态融合的 FAS 研究打下了基础, 但相关多模态数据集和多模态 FAS 算法研究仍然不足.

(5) 模型效率有待提升. 研究表明, 早期通过眨眼、张嘴等交互式行为应对打印攻击的方法[18−20]平均检测时长在 4 s 左右, 无法应用在实时领域. 此外, 许多模型需要从多帧图像中提取时空特征, 导致具有较高的时间复杂度或者空间复杂度, 从而难以在低配设备和手持设备上运行, 这也限制了部分方法的实际应用.

2.2 FAS当前主要的解决方案

FAS 方法种类繁多, 总体上可以划为传统方法(第 3 节)和基于深度学习(第 4 节、第 5 节)的方法两大类. 限于篇幅, 本节选取近几年的代表性工作.

(1) 将传统 FAS 方法进一步细分为: 基于纹理的方法、基于运动的方法和基于远程光电容积脉搏波描记 rPPG (remote photoplethysmography)的方法等, 并在第 3.1 节−第 3.3 节分别对这 3 大方法进行详细说明.

(2) 将深度学习的 FAS 方法进一步细分为二元监督方法、深度图方法、rPPG 信号方法、时空信息融合方法、多模态方法和分离欺骗痕迹方法等, 并在第 4.1 节−第 4.6 节分别对这 6 大方法进行详细说明.

(3) 特别地, 针对基于深度学习的 FAS 的域泛化和可解释性问题, 我们单独在第 5 节中, 从理论和实践两方面进行了详细讨论.

表 1 在上述分类的基础上给出了方法概览, 介绍了现有方法的代表性文献、方法的思想以及方法的优点和不足. 由于方法是为了解决科学问题, 因此表 2 对现有 FAS 方法“所解决的科学问题以及可抵御的攻击等”两方面进行了总结.

表 1 FAS 方法概览

方法分类	方法细化	代表性文献	思想简介	优点	不足
传统人脸欺骗检测方法(第 3 节)	基于纹理的方法(第 3.1 节)	[2−4,21−41]	使用纹理描述算子如 LBP 等提取欺骗人脸呈现的质量下降、颜色失真等欺骗伪影	计算量小, 环境相对稳定的情况下, 可有效抵御打印攻击和重放攻击	对外界环境的变化较为敏感, 鲁棒性较差
	基于运动的方法(第 3.2 节)	[18−20,26,28,31,32,38,39,42−45]	通过检测眨眼等特定动作, 或者通过光流、动态模式分解等算法捕获真假人脸动态信息的差异	考虑了运动特征, 可有效抵御打印攻击	难以应对重放视频攻击; 且检测时间较长, 实时决策较为困难
	基于 rPPG 的方法(第 3.3 节)	[6,46−49]	捕捉人脸 rPPG 信号随脉搏发生的变化, 进而区分真假人脸	稳健的 rPPG 信号可有效抵御打印攻击和面具攻击	rPPG 信号易受噪声干扰, 鲁棒性较差
基于深度学习的人脸欺骗检测方法(第 4 节)	二元监督方法(第 4.1 节)	[11−14,50−53]	使用 1 和 0 作为真假人脸标签监督网络学习	相对传统方法可自动学习可区分特征; 可用于时间复杂模型预处理阶段的“粗”分类, 从而降低时间复杂模型的计算量	难以学习欺骗人脸本质特征, 决策结果较为粗糙, 缺乏可解释性

表 1 FAS 方法概览(续)

方法分类	方法细化	代表性文献	思想简介	优点	不足
基于深度学习的人脸欺骗检测方法(第 4 节)	基于深度图的方法(第 4.2 节)	[15,54−64]	通过深度图监督网络学习	深度特征是真假人脸的本质区别之一, 结合深度特征可有效应对深度信息欠缺的攻击, 如打印攻击和重放攻击等	难以抵御超具有深度信息的攻击, 如超真实 3D 面具攻击等; 且深度图自身质量会影响算法性能
	基于 rPPG 的方法(第 4.3 节)	[64−66]	通过结合 rPPG 信号和神经网络提取的图像/视频高层特征以提升 FAS 的区分效果	rPPG 信号对目标遮挡比较敏感, 可有效抵御对目标 rPPG 信号有影响的攻击, 如超真实 3D 面具攻击等	rPPG 信号本身易受干扰, 如运动、光照、角度等均可对 rPPG 信号造成全局或者局部干扰, 对环境要求较高
	基于时空信息的方法(第 4.4 节)	[57,67−72]	融合时间和空间维度的可区分特征来区分真假人脸	综合利用时间维度的运动信息和空间维度的深度信息, 可有效提升对打印攻击和重放攻击的抵御能力	对于参数量大的模型, 检测时间相对较长, 难以实时做出决策
	基于多模态的方法(第 4.5 节)	[16,73−84]	利用多模态数据优势互补的特点, 融合同一目标的多模态图像特征进行分类	可利用同一目标的不同模态图像之间优势互补的特点, 解决单模态图像信息不够丰富的问题	需要额外相机设备(深度相机、红外相机). 数据集收集成本相对较高
	分离欺骗痕迹的方法(第 4.6 节)	[61,85−88]	通过从欺骗人脸图像中分离出对 FAS 检测贡献最大的欺骗痕迹来区分真假人脸	泛化能力较强而且模型具有一定的可解释性	有一定难度, 理论和方法均有待加强
域泛化和可解释性(第 5 节)	域泛化的方法(第 5.1 节、第 5.2 节)	[58−60, 89−100]	通过缩小源域和目标域的分布差异, 使得分类器在不同目标域仍能保持较好性能	有效提升在未经训练的不可见(unseen)域上的检测能力	寻找源域和不可见域之间的共性关系是关键和难点; 当源域数据集分布差异较大时, 泛化可能会降低算法在单个数据集上的检测性能
	可解释性方法(第 5.3 节、第 5.4 节)	[61,71,88, 101−104]	通过可视化或对黑盒模型施加以可解释性为目标的约束, 使得人类能够理解模型做出某种决策的原因	提升可解释性有助于理解模型和改进模型, 增强模型的可信性(trust)、安全性(security)和可靠性(safety)	缺少一般化的可解释性理论模型和评估标准, 目前可解释性程度受限

注: 更详细的方法说明可以参阅“方法分类”和“方法细化”栏中小节编号所对应的小节

表 2 FAS 方法、所解决的科学问题和可抵御的攻击类型概览

方法分类	文献	发表年份	解决的科学问题	打印攻击	重放攻击	面具攻击
传统人脸欺骗检测方法	[27]	2018	(1), (5)	√	√	×
	[29]	2018	(1)	√	√	×
	[39]	2018	(1)	×	×	√
	[48]	2018	(1)	×	×	√
	[49]	2018	(5)	√	×	√
	[46]	2017	(1)	√	√	√
	[4]	2016	(1)	√	√	×
	[6]	2016	(1)	√	√	√
	[24]	2016	(1)	√	√	×
	[30]	2016	(1)	√	√	×
	[42]	2016	(3)	√	√	√
	[26]	2015	(3)	√	√	×
	[28]	2015	(1), (3)	√	√	×
	[35]	2014	(5)	√	√	×
	[36]	2014	(5)	√	√	×
	[37]	2015	(1)	√	√	×
	[38]	2014	(3), (5)	√	√	×
	[22]	2013	(1)	√	√	×
	[25]	2013	(1)	√	√	×
	[45]	2013	(3)	√	×	×

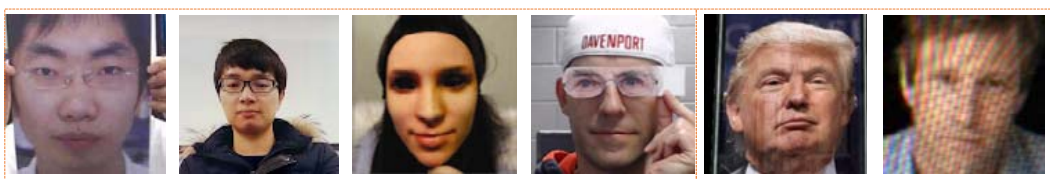
表 2 FAS 方法、所解决的科学问题和可抵御的攻击类型概览(续)

方法分类	文献	发表年份	解决的科学问题	打印攻击	重放攻击	面具攻击
传统人脸欺骗检测方法	[31]	2012	(3)	√	√	×
	[32]	2012	(3)	√	√	×
	[21]	2011	(1), (5)	√	×	×
	[33]	2011	(5)	√	×	×
基于深度学习的人脸欺骗检测方法	[84]	2021	(1), (4)	√	√	√
	[98]	2021	(1)	√	√	×
	[99]	2021	(1)	√	√	×
	[100]	2021	(1)	√	√	×
	[103]	2021	(1), (2)	√	√	×
	[105]	2021	(1)	√	√	×
	[15]	2020	(3)	√	√	×
	[55]	2020	(1), (5)	√	√	√
	[56]	2020	(1)	√	√	√
	[59]	2020	(1)	√	√	√
	[60]	2020	(1)	√	√	×
	[72]	2020	(3)	√	√	×
	[79]	2020	(4)	√	√	√
	[81]	2020	(1), (4)	√	√	×
	[82]	2020	(1), (4)	√	√	×
	[86]	2020	(1), (2)	√	√	√
	[91]	2020	(1)	√	√	×
	[93]	2020	(1)	√	√	×
	[95]	2020	(1)	√	√	×
	[104]	2020	(1), (2), (5)	√	√	√
	[106]	2020	(1)	√	√	×
	[58]	2019	(1)	√	√	×
	[71]	2019	(1), (2), (3)	√	√	×
	[96]	2019	(1)	√	√	√
	[57]	2018	(3)	√	√	×
	[64]	2018	(1), (2)	√	√	×
	[69]	2018	(1), (3)	√	√	×
	[85]	2018	(2)	√	√	×
	[89]	2018	(1)	√	√	×

注: 表中√表示可以防御, ×表示不能防御(以文中实验使用的数据集所包含的欺骗攻击类型为依据)

3 传统 FAS 方法

传统 FAS 方法建立在如下假设之上: 攻击者呈现给相机的欺骗人脸与真实人脸之间一定存在可区分线索. 以图 2 为例, 图 2(a1)–图 2(a4)给出了常见的欺骗攻击类型: 当攻击者采用打印照片攻击时, 所打印的照片会出现质量下降、颜色失真等情况(如图 2(a1)所示); 当攻击者采用回放攻击时(如图 2(a2)所示), 回放视频的欺骗电子载体会在相机前出现屏幕反光或者莫尔图案(如图 2(b1)和图 2(b2)所示)等欺骗伪影; 当攻击者采用 3D 面具攻击时, 3D 面具难以产生表情等细微的脸部变化(如图 2(a3)所示). 因此, 基于前述假设, 研究人员通过手工定义特征提取算子, 并利用所定义的特征提取算子检测(预学习到的)真实人脸特征与当前所呈现的人脸特征之间是否存在差异, 从而发现人脸欺诈攻击.



(a) 常见欺骗攻击类型 (b) 常见欺骗伪影

图 2 常见欺骗攻击类型和欺骗伪影

根据抓取欺骗线索的不同, 传统的 FAS 方法可以分为 3 大类: 基于纹理的方法、基于运动的方法和基于 rPPG 的方法, 以下进行详细说明.

3.1 基于纹理的FAS方法

基于纹理的 FAS 方法通过捕捉欺骗人脸再次呈现在摄像头前时, 与真实人脸相比所呈现的质量下降、颜色失真和图像伪影等纹理差异来发现攻击. 具体地说, 上述纹理差异可以通过特征提取算子如高斯差分 DOG (difference of Gaussian)[2,21]、方向梯度直方图 HOG (histogram of oriented gradient)[22,23]、局部二值模式 LBP (local binary patterns)[3−6,24−27]、局部相位量化 LPQ (local phase quantization)[28,29]、加速稳健特征 SURF (speeded up robust features)[29,30]以及 LBP 的变种如旋转不变均匀局部二值模式 RI-LBP (rotation invariant uniform local binary patterns)[29]、三正交平面局部二值模式 LBP-TOP (local binary patterns from three orthogonal planes)[31,32]、多尺度局部二值模式 MSLBP (multi-scale local binary patterns)[33,34,42]等进行捕获. 表 3 对常用的特征提取算子进行了说明.

表 3 特征提取常用描述算子描述的比较

名称	描述	优点	缺点
DOG	将图像在不同参数下的高斯滤波结果相减, 得到差分图, 对图像边缘进行检测	计算复杂度小, 有效检测边缘信息, 可处理具有高频噪声的图像	提取的特征容易受到图像对比度变化的影响
HOG	通过计算图像的梯度信息来捕捉图像的轮廓信息	减弱光照对图像颜色的影响	生成特征速度慢, 容易受噪点影响
LBP	通过比较周围像素和中心像素值的大小来描述图像的纹理信息	计算复杂度小, 运算速度快	对方向信息和光照变化较为敏感
LPQ	利用短时傅里叶变换从图像中提取局部相位信息	能从模糊和低分辨率的图像中提取有效的纹理信息	如果图像过于模糊, LPQ 从图像提取的纹理信息有限
SURF	对尺度不变特征变换 SIFT (scale-invariant feature transform)进行改进, 主要提取图像所包含的角点、边缘点等信息	具有尺度不变性和旋转不变性, 对光照不敏感	难以从边缘不明显的图像中准确提取特征

3.1.1 基于质量失真的方法

当欺骗人脸载体(纸张、照片、电子屏幕等)再次呈现在摄像头面前时, 会由于各种原因而出现图像质量失真, 如高频信息丢失[2]、光学特性差异(吸收、反射和折射等)[3,33]、莫尔图案[4]、固有属性改变(清晰度、亮度、色度、对比度等)[35−37]以及结构扭曲[36]等现象. 除此以外, 寻求新的图像和视频质量失真线索, 是这类 FAS 方法的要点.

图像失真线索可以采用单一线索[2,3,33], 也可以采用多重线索[4,35−37].

- 首先, 单一线索方面, Zhang 等人[2]发现, 欺骗人脸与真实人脸相比会发生高频信息丢失. 为此, 他们通过使用多个不同的 DOG 滤波器提取人脸图像中的高频信息, 并将滤波后的图像输入到支持向量机 SVM (support vector machine)中以分类出欺骗人脸. Chingovska 等人[3]通过 LBP 捕捉真实人脸和欺骗人脸在光学特性上的差异来防止打印攻击和重放攻击. Määttä 等人[33]发现, 打印照片由于反光特性比真实人脸更强而导致图像质量失真, 提出可使用 MSLBP 分别从人脸图像的整体和局部区域捕捉因反光导致的质量失真来区分真假人脸.
- 其次, 多重线索方面, Patel 等人[4]发现, 打印照片或者重放视频会出现表面反光、莫尔图案等质量失真现象, 因此可通过结合 LBP 和颜色矩 CM (color moments)分析图像质量来发现欺骗人脸. Galbally 等人[35]采用 14 种图像质量度量方式, 从清晰度、颜色和亮度等方面对人脸质量进行评估, 然后将所获得的质量特征利用线性判别分析 LDA (linear discriminant analysis)进行分类, 以发现欺骗人脸攻击. 为了应对不同的场景变化以增强鲁棒性, Galbally 等人[36]进一步选取了 25 种图像质量评估算法捕捉图像的多重质量失真现象, 以区分真假人脸. Di 等人[37]利用 4 种不同的算法分别捕获欺骗人脸所呈现的 4 种质量失真, 即镜面反射、图像模糊、色度变化和对比度变化, 然后将得到的特征向量送入 SVM 分类出真假人脸, 提升了算法的泛化能力.

3.1.2 基于颜色失真的方法

无论是打印机还是显示器, 其颜色感知能力都是有限的. 具体地说, 打印机和显示器的色域都不能完全覆盖可见光的色域(如图 3 所示), 这使得当欺骗人脸载体(如纸张、照片、电子屏幕等)再次呈现在摄像机前面时, 会不可避免地会丢失颜色信息. 基于这种颜色差异, 可从颜色空间中通过捕获颜色失真纹理信息实现 FAS 检测[24,29,30]. 在具体实现时, 常用的颜色空间包括红绿蓝空间 RGB (red, green, blue), 色调饱和度亮度空间 HSV (hue, saturation, value), 亮度蓝色红色空间 YCbCr 等.

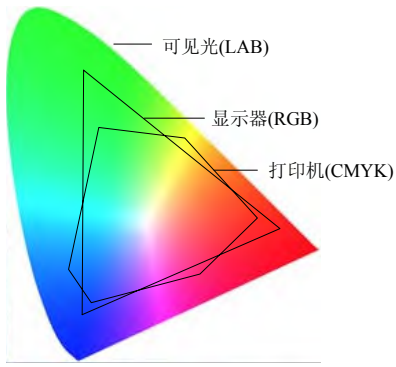


图 3 显示器和打印机的色域范围比较

颜色失真的代表性工作由 Boulkenafet 等人[24,29,30]发起. 考虑到仅从图像的亮度(即灰度)分析纹理差异[33]难以发现色域的差异, Boulkenafet 等人[24]利用共生相邻局部二值模式 CoALBP (co-occurrence of adjacent local binary patterns)算子和 LPQ 算子分别从 HSV 和 YCbCr 颜色空间中捕捉色度和亮度差异进行 FAS 检测. 但从颜色空间捕捉颜色失真时, 需要解决颜色纹理特征的稳定性问题. Boulkenafet 等人[30]观察到, 从色度通道提取的颜色特征比从 RGB 通道中提取的颜色特征更加稳定[24], 为此提出可使用 SURF 算子(参见表 3)从人脸图像的 HSV 和 YCbCr 颜色空间的色度通道提取纹理特征, 经过主成分分析(principal component analysis, PCA)降维处理后再送入 Softmax 分类器中以检测欺骗人脸. 进一步地, 为提升抵御打印照片攻击和视频重放攻击的泛化能力, Boulkenafet 等人[29]选取了 7 种不同的颜色纹理描述算子进行实验, 结果表明, SURF 算子和 RI-LBP 算子分别在抵御打印攻击和重放攻击上表现最佳. 他们提出, 可以结合这两种算子来提升 FAS 的泛化能力.

3.1.3 基于纹理失真的方法

基于质量差异和颜色失真的 FAS 方法仅仅从单帧图像(即图像 *XY* 平面)中捕获静态纹理差异, 但是它们忽略了多帧图像之间随着时间推移(即图像的 *XT* 和 *YT* 平面)而呈现的动态纹理信息. 对于上述位于 *XY*, *XT* 和 *YT* 这 3 个正交平面的静态和动态纹理信息, 可以利用与三正交平面相关的算子, 如三正交平面局部二值模式 LBP-TOP 算子、三正交平面多尺度二值化统计图像特征 MBSIF-TOP (multiscale binarized statistical image features on three orthogonal planes)算子、三正交平面多尺度局部相位量化 MLPQ-TOP (multiscale local phase quantization on three orthogonal planes)算子以及低层 CNN (convolutional neural networks)进行捕捉.

LBP-TOP 除了能够描述图像平面的空间纹理信息(静态纹理信息)外, 还可描述图像沿水平和垂直方向的时间纹理信息(动态纹理信息), 因而可利用 LBP-TOP 同时捕获静态和动态两种纹理特征, 以增强对人脸欺骗攻击的抵御能力[31,32,38]. Arashloo 等人[28]考虑到 MBSIF-TOP 和 MLPQ-TOP 具有优势互补的特点——即 MBSIF-TOP 对人脸的动态纹理差异较为敏感从而具有较好的真假人脸区分能力, 但不足之处在于鲁棒性较差; 而 MLPQ-TOP 是模糊可容忍的(blur-tolerant), 其长处在于鲁棒性较好——提出可结合上述两大算子, 利用计算高效的谱回归核判别分析 SR-KDA (spectral regression kernel discriminant analysis)对 MBSIF-TOP 和 MLPQ-TOP 生成的核信息进行融合之后实现真假人脸区分, 取得了较好的效果. 针对 3D 面具攻击问题, Shao 等人[39]利用低层 CNN 生成的特征图捕获面部的动态纹理信息, 并通过加权的方式从特征图中筛选出可用于

区分的通道信息和空间信息, 达到识别真假人脸的目的.

除了上述从动态信息中捕获纹理失真以外, 将不易直接应用的纹理特征设法增强或变换后寻求纹理差异也是一种思路.

Chan 等人[27]发现, 闪光灯环境可放大真实人脸和欺骗人脸之间的差异. 为此, 他们将闪光灯和非闪光灯环境下拍摄的人脸图像同时作为模型输入, 然后使用 4 种不同的算子分别从有无闪光的图像中提取纹理信息和人脸结构信息来检测人脸欺骗攻击. Agarwal 等人[41]观察到, 图像经过冗余离散小波变换后提取的 Haralick 纹理特征可明显增强欺骗人脸和真实人脸的差异. 为此, 他们首先利用离散小波变换将输入图像变换到小波域, 然后从分解后的冗余小波变换子带中提取出 Haralick 特征, 最后将 Haralick 特征利用 PCA 降维后送入 SVM 中进行分类以识别真假人脸.

3.2 基于运动的方法

3.2.1 交互式运动方法

交互式运动方法要求被检测对象按照人脸识别系统的要求进行交互式响应, 如眨眼、张嘴等, 无法完成交互式响应的对象则被判定为欺骗人脸.

Pan 等人[18,19]将眨眼视为一个从睁开到闭合到再睁开的连续过程, 为了从视频图像序列中捕获到这一连续过程, 他们在条件随机场 CRF (conditional random fields)[107]中对眨眼行为进行建模, 并使用自适应增强 Adaboost (adaptive boosting)算法[108]测量眼睛的闭合程度以区分真假人脸. Kollreider 等人[20]提出可基于嘴部运动状态进行 FAS 检测, 其基本思想是: 首先生成一串随机数字序列并要求被检测对象读出; 然后将被检测对象的嘴部划分为 4 个不同区域, 使用光流法[109]对每个区域的连续 5 帧图像进行运动估计, 判断对应区域嘴部运动状态和所给随机数字序列是否相匹配, 当匹配时为真实人脸, 否则为虚假人脸.

3.2.2 非交互式运动方法

非交互式方法无需被检测对象与系统进行交互, 其通过直接捕获人脸运动(如表情变化、头部转动等)达到真假人脸识别的目的.

光流(optical flow)具有描述运动信息的能力, 被广泛应用于非交互式 FAS 检测领域[20,42,44,45]. 所谓的光流是指由于目标对观察者的相对运动所形成的目标、目标表面和目标边缘的运动模式. 大多数计算光流的方法都假定像素的颜色/强度在从一个视频帧到下一个视频帧时是不变的[110]. 当光流应用于打印照片攻击检测时, 由于二维平面的相对运动只有平移、旋转、前后运动和摆动这 4 种方式, 故而只会对应产生 4 种不同的光流场, 因此, 二维的打印照片所产生的运动光流场只可能是上述 4 种光流场的线性组合. 相反, 真实人脸是不规则的三维形态, 其面部表情或者头部运动的变化所产生的光流远比上述 4 种光流场的线性组合丰富, 特别是两者在摆动时生成的光流场, 其差异尤为明显. 基于上述观察, Bao 等人[44]提出: 可计算二维平面理想状态下的光流场以及受试区域的真实光流场, 然后度量“理想-真实”两个光流场之间的差异, 当该差异超过某一设定的阈值时即为真实人脸, 否则为虚假人脸. Anjos 等人[45]考虑到 2D 静态照片与 3D 真实人脸相比, 其每一部分的运动轨迹一定存在差异, 因而可以首先利用光流计算水平和垂直方向的速度分量以对人脸的运动状态进行描述, 然后通过二分类器识别出打印照片攻击.

除了光流以外, 将运动信息与 LBP 纹理信息结合, 也是一种常见的非交互式检测方法. Tirunagari 等人[26]指出, 动态模式分解 DMD (dynamic mode decomposition)可同时捕捉真实人脸呈现的面部动态信息(如眨眼、张嘴、表情变化等)以及欺骗人脸所呈现的伪影信息(如莫尔图案、平面信息等), 故而可首先利用 DMD 捕获输入图像的动态信息建立动态模式图像, 然后利用 LBP 从动态模式图像中提取纹理特征, 最后送入 SVM 以分类出真假人脸. Siddiqui 等人[42]发现, 未经裁剪的视频序列会呈现多种欺骗线索, 为此提出了一种多特征聚合的 FAS 检测方法. 具体地说, 首先, 对于视频片段的每一帧, 使用 MSLBP 从该帧的整张图像(整体)以及该帧整张图像中裁剪出来的单纯人脸图像(局部)两者中同时提取纹理特征; 然后, 使用定向光流直方图 HOOF (histogram of oriented optical flow)从视频片段的整体和局部同时提取运动特征; 最后, 将所提取的纹理和运动两种特征聚合后送入 SVM 进行真假人脸判定.

3.3 基于rPPG的方法

光电容积脉搏波描记 PPG (photoplethysmography)的原理如图 4 左侧所示[47]: 当光线照射到皮肤时, 光线穿过皮肤被人体的组织和血液所吸收、反射, 再次反射出的光线将发生衰减并被传感器所接收. 当没有大幅度运动时, 人体组织(如肌肉、骨骼等)对光线的吸收基本恒定不变; 相反, 血液对光线的吸收却会发生周期性变化, 这是因为随着心跳, 动脉会发生扩张和收缩, 动脉中的血液也会随之流动, 因而血液所吸收的光线会随着心跳发生周期性变化. 当攻击者戴上面具或者使用照片遮挡人脸时(图 4 右侧), 由于面具和照片会阻挡大部分光线, 导致穿透皮肤组织的光线大大减少, 进而光线和血液的相互作用被削弱, 表现为检测到的 PPG 信号变化十分微弱. 获取 PPG 信号可以采用接触式方法, 也可以采用非接触式方法, 如利用普通相机即可以获取 PPG 信号[111]. 为区别起见, 非接触式的 PPG 称为 rPPG. 在 FAS 中, 一般采用 rPPG 来检测人脸欺骗攻击.

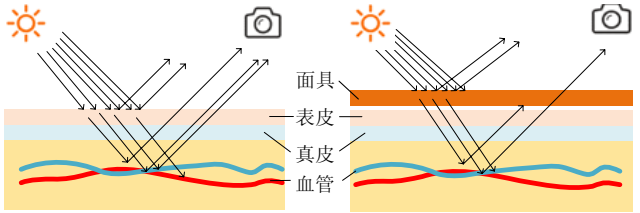


图 4 皮肤(左)和面具(右)对 PPG 的影响

当前, 基于最新 3D 打印技术所生成的超真实面具可以高保真地还原皱纹等精细纹理信息以及脸部的深度信息, 这对基于纹理信息[32]和深度信息的方法[39,50]提出了挑战. 针对超真实 3D 面具攻击问题, Li 等人[6]首先提出可从 RGB 图像的红绿蓝 3 个通道分别提取 rPPG 信号, 然后利用 3 种时间滤波器降低噪声对 rPPG 信号的干扰, 最后将降噪后的 rPPG 信号转化为向量送入 SVM 分类出真假人脸.

rPPG 的思想并不复杂, 但是在实践中, rPPG 很容易受到噪声干扰. 例如, 当相机发生运动、头部姿态变化、光照发生变化等, 都会引入噪声. 因此, 必须提升 rPPG 方法的鲁棒性[46−48]. Nowara 等人[46]考虑到: 当光照强度发生变化时, 对于欺骗人脸, 其前景的欺骗人脸 rPPG 信号和背景的 rPPG 信号变化是相同的; 对于真实人脸, 其前景的真实人脸 rPPG 信号与背景的 rPPG 信号变化则是不同的. 因此, 利用背景区域的 rPPG 信号可以克服光照变化对 FAS 的干扰. 另一方面, 由前述 rPPG 的原理可知, rPPG 信号和个体心跳是紧密相关的. 基于这种相关性, Liu 等人[47]认为, 对于真实人脸, 同一个体的面部不同区域的 rPPG 信号必定具有相似的波形, 其差异很小; 相反, 3D 面具的不同“人脸”区域由于光线被面具遮挡从而导致所接收到的 rPPG 信号主要是环境噪声, 在频率和周期上存在较大差异. 为此, 对于任意两个分属脸部不同区域的 rPPG 信号, 可计算其互相关(cross-correlation)频谱的最大值来度量两者在频率和周期上的相关性. 最终, 将所有可能的 rPPG 信号的两两组合分别计算其相关性, 然后取并集作为所提取的 rPPG 模式, 即可用于发现 3D 面具攻击. 采用互相关计算的优势在于可以放大(真实人脸不同区域的)相似的心跳频率, 并抑制(3D 面具虚假“人脸”不同区域的)随机噪声干扰, 起到正反馈的作用. 但是, 在噪声强于心跳信号的情况下——如摄像机运动产生的全局噪声、光线昏暗等——互相关计算将对噪声进行正反馈, 这将导致错误的检测结果. 为了解决噪声处于主导地位时的 rPPG 应用问题, Liu 等人[48]提出了名为一致特征 rPPG 的方法, 即 CFrPPG (correspondence feature rPPG). 其基本思想是: 避免从 rPPG 信号中直接提取心跳信号, 而是利用相关性(correlation)计算反映真实人脸不同区域 rPPG 信号之间共性的频谱模板, 此即共性的心跳信息. 之后, 通过计算所学习到的频谱模板和待检测 rPPG 信号之间的关联关系, 实现噪声主导情况下的 3D 面具攻击检测.

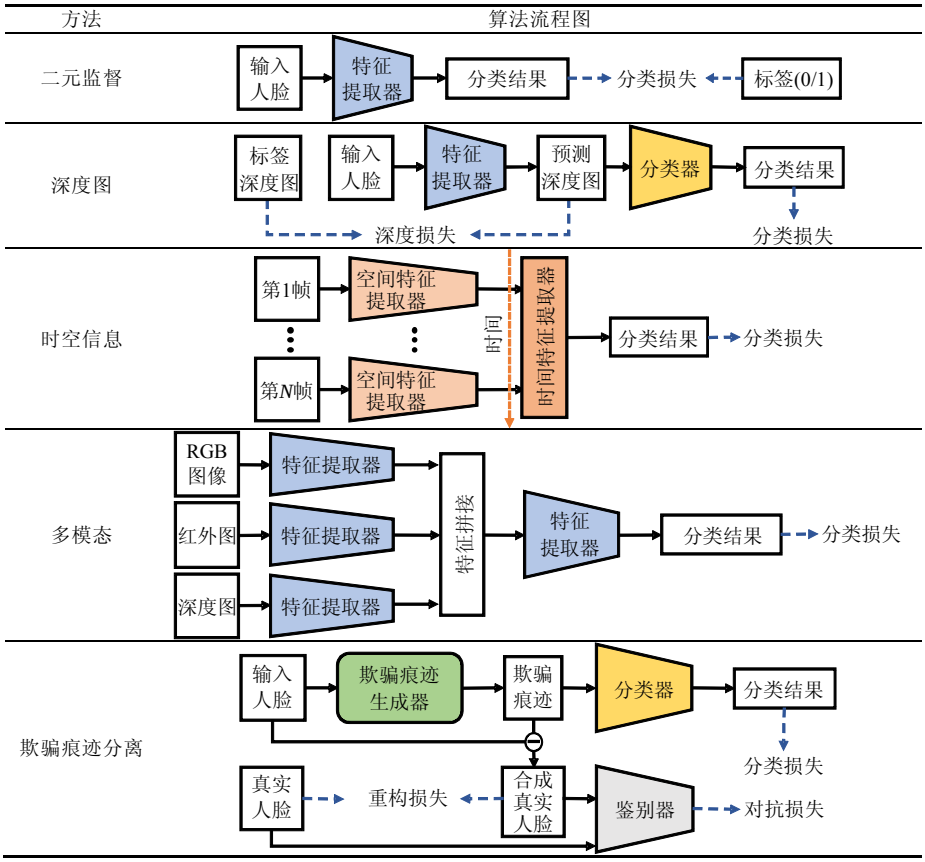
4 基于深度学习的 FAS 方法

传统的 FAS 方法需要手工设计算子以提取特征, 但手工算子依赖于专家经验, 工作量大, 且只能提取预设的特征. 基于深度学习的 FAS 方法则克服了传统方法的不足: 其无须手工算子, 可在给定目标函数的情况下利用梯度下降算法自动更新网络参数以优化网络模型, 进而自动学习到人脸特征. 自 Yang 等人[13]将 CNN

引入到 FAS 领域以来, 基于深度学习的 FAS 方法已在各方面超越传统 FAS 方法而成为研究人员的首选.

根据欺骗线索标记方式或者欺骗线索来源的不同, 本文将基于深度学习的 FAS 方法分为 6 类, 分别是二元监督方法、深度图方法、rPPG 方法、时空信息融合方法、多模态方法和欺骗痕迹分离方法, 表 4 给出了上述 FAS 方法的算法流程示例图.

表 4 基于深度学习的 FAS 方法流程图



4.1 二元监督

在深度学习应用于 FAS 的初期, 研究人员将真假人脸区分看作简单的二分类问题[11−14,39,50−53]: 在训练神经网络时, 将训练数据用 0 和 1 这两个标签分别表示虚假人脸和真实人脸, 然后通过分类损失计算网络预测结果与 0/1 标签之间的差异完成分类(参见表 4). 由于采用 0/1 二元标签, 因而这类方法也称为二元监督方法. 二元监督多以二元交叉熵作为损失函数, 其优化目标如公式(1)所示.

$$F^{*}=\arg\max_{\theta}\frac{1}{N}\sum_{i=1}^{N}\mathcal{Y} \tag{1}$$

其中, $\mathcal{Y}=(y_i)\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)$, y_i 表示输入人脸对应的标签, $\hat{y}_i$ 表示网络预测的结果, θ表示网络参数.

既然二元监督将 FAS 检测看作二分类问题, 一个自然的想法是, 利用图像分类 CNN 进行 FAS 设计. 常用的图像分类 CNN 包括视觉几何组 VGG (visual geometry group)[112]、Inception[113]和残差网络 ResNet (residual network)[114]等等. Li 等人[12]针对传统 CNN 仅由全连接层进行分类决策的方式, 忽略了卷积层自身内在包含的丰富特征问题, 提出了深层 CNN, 即 DPCNN (deep part CNN). DPCNN 将每一个卷积核均看作一个单独的滤

波器, 并用预训练的 VGG[112]提取人脸的特征信息, 最后经过 PCA 降维后送入 SVM 分类出真假人脸. Lucena 等人[51]提出了一种基于 VGG16[112]的迁移学习方法: 通过对已经在 ImageNet 数据集上完成预训练的 VGG16 模型进行微调, 再利用 sigmoid 函数完成真假人脸的二分类, 在 3DMAD[5]和 REPLAY-ATTACK[3]数据集上取得了当时最优的效果. Nagpa 等人[52]将 Inception-v3[113], ResNet50[114]和 ResNet152[114]调整为二分类网络, 在手机人脸欺骗数据集 MFSD (MSU mobile face spoofing database)[37]上对三者进行性能评估, 给出了评估结果以及对三者的应用建议. 针对深度学习时难以获取欺骗人脸数据的问题, Guo 等人[53]提出了一种将平面的 2D 打印照片合成为 3D 虚拟欺骗照片的方法, 其基本思想是: 首先将 2D 打印照片网格化为 3D 对象, 然后在 3D 空间中模拟打印照片的弯曲和旋转等操作以合成大量 3D 虚拟欺骗照片, 最后将合成的样本用于训练修改后的 ResNet50[114]并给出二分类结果.

二元监督方法有一定的效果, 但其只是进行笼统的二分类而并没有考虑欺骗信息的来源. 研究表明, 从欺骗信息的源头挖掘欺骗特征来设计 FAS 算法, 可以取得比二元监督好得多的分类效果. 目前, 常用的欺骗特征来源包括深度图、rPPG 信号、时空信息、多模态信息和欺骗痕迹等, 第 4.2–4.6 节将分别加以介绍.

4.2 深度图

在计算机视觉领域, 深度图是一幅图像或者图像通道, 其中包含有从目标对象表面到视点的距离信息. 当用于 FAS 检测时, 深度图可呈现出真假人脸表面与相机之间的距离差异, 原理如下: 对于真实人脸, 其为立体结构, 因而真实人脸的不同区域(如鼻子和耳朵等)与相机之间存在较为明显的距离差异; 相反, 对于虚假人脸, 其为平面结构, 因而虚假“人脸”的不同区域与相机之间的距离几乎没有变化. 由此, 可以利用深度图作为真假人脸的可区分特征来源. 在具体操作时, 基于深度图的 FAS 方法其流程如表 4 所示. 首先, 对输入的人脸图像通过网络模型预测该输入人脸对应的深度图; 然后, 基于深度损失计算所预测的深度图与给定的(标签深度图)正样本(ground truth)之间的距离; 最后, 根据预测的深度图做出分类决策. 形式地, 人脸深度信息的像素级损失函数如公式(2)所示, 当 L_{depth} 最小化时, 模型估计人脸面部深度的能力达到最优.

$$L_{depth}=\frac{1}{H\times W}\sum_{i\in H,\,j\in W} f(D_{pre(i,j)},D_{gt(i,j)}) \tag{2}$$

其中, H, W 分别表示网络生成的深度图的高和宽; D_{pre} 和 D_{gt} 分别表示网络预测的深度图和深度图的正样本; $f(\cdot)$为衡量对应像素值距离的函数, 通常选择 $L2$ 范数的平方[55,56].

Atoum 等人[54]提出一种结合人脸全局和局部特征的双重 CNN 方法, 其基本思想是: 采用一个 CNN, 从全局获取图像的深度图, 并依据深度图给出是否为真实人脸的“活性(live)”评分; 采用另一个 CNN, 从局部随机选取图像的小块区域, 并对所选取的小块区域从 HSV 和 YCbCr 颜色空间以及 LBP 图学习可区分特征, 并给出另一个评分; 最后, 融合上述两个 CNN 评分得到最终判定结果. 其中, 全局的深度图和局部的随机小块区域分别从空间和纹理的角度提供了判定线索, 增强了判定的准确性. Yu 等人[55]针对目前 FAS 方法所面临的两大难题——即第一, 基于多帧图像序列抽取动态特征导致模型时间复杂度升高; 第二, 需要专家设计网络结构, 难以捕获图像中的细粒度信息以及难以快速适应不同环境——提出了名为中心差分卷积网络 CDCN++ (central difference convolutional networks)的深度图方法. CDCN 首先对传统 CNN 的聚合操作进行扩展, 引入了新的中心差分卷积 CDC (central difference convolutional)用以捕获 CNN 采样值的中心梯度; 然后, 在聚合时对传统 CNN 的采样值以及 CDC 所捕获的中心梯度同时进行聚合, 从而从单帧图像中得到粒度更细、鲁棒性更强的深度图. 为了提升 CDCN 的自动设计和细粒深度特征获取能力, 他们首次将基于梯度的神经架构搜索 NAS (neural architecture search)引入到 FAS, 以自动寻找在捕获 CDC 时性能更优的网络主干架构; 同时, 设计了一个多尺度注意力融合模块 MAFM (multiscale attention fusion module), 通过空间注意力来求精, 并融合低、中、高 3 级(low-mid-high levels) CDC 特征; 在此基础上, 将 CDCN 扩展为更具自动化和性能更优的 CDCN++. Yu 等人[56]认为, 欺骗伪影是由于虚假人脸载体(如打印照片、电子屏幕、3D 面具)和人体皮肤的材质差异造成的, 因此可将材质感知技术[115]用于 FAS. 为了从不同的角度学习材质差异, 他们提出了双边卷积网络 BCN (bilateral convolutional networks). BCN 分别使用深度图、反光图(reflection map)[116]和补丁图(真假人

脸分别对应全 1 和全 0 的图)作为正样本来监督网络, 以学习不同材料之间的深度信息、反光信息和纹理信息差异, 进而达到基于材质区分真假人脸的目的.

4.3 rPPG信号

rPPG 信号方法能够较好地防御超真实 3D 面具等攻击手段, 但是 rPPG 信号对噪声干扰很敏感[46−48]: 光照强度变化、人脸姿态变化、相机运动等都会对 rPPG 信号产生干扰. 为此, 在深度学习中 rPPG 信号常常与其他方法结合应用.

Liu 等人[64]基于人脸深度图和 rPPG 信号提出了新颖的 CNN-RNN 结构, 其中, 卷积神经网络 CNN 在深度图的监督下学习人脸的深度信息, 循环神经网络 RNN (recurrent neural network)在 rPPG 信号的监督下学习心率信息, 最后根据深度信息和心率信息计算出最终分类结果. 需要指出的是, 这里的心率信息是通过 RNN 学习得到的, 而并非直接从视频中提取, 这是深度学习 rPPG 方法与传统 rPPG 方法的重要区别.

Lin 等人[65]提出了一种融合 rPPG 信号和纹理特征的 FAS 方法, 其基本思想是:

- 在提取 rPPG 信号时, 将长期统计频谱 LTSS (long-term statistical spectral)[117]改进为多尺度长期统计频谱 MS-LTSS (multi-scale LTSS). 不同于 LTSS 仅从恒定时长的 rPPG 信号中提取频谱特征, MS-LTSS 可结合不同长度和不同重叠尺寸的滑动窗口来统计频谱特征, 从而可以获得更为精细的 rPPG 信息.
- 在提取纹理特征时, 使用基于上下文补丁的 CP-CNN (contextual patch-based CNN)从全局和局部区域同时提取纹理特征, 从而可以更好地发现真实人脸与重放视频之间的纹理差异.
- 最后, 将 MS-LTSS 和 CP-CNN 的分类结果通过权重求和进行融合即得到最终的判定结果.

Yu 等人[66]提出了 rPPG 转换器方法, 以充分利用图像的全局和局部 rPPG 信号, 其基本思想是: 首先从输入人脸视频的背景区域和人脸区域分别提取 rPPG 信号, 并对应转化为各自的多尺度时空图[118]; 然后采用自动化的视觉转换器(vision transformer)[119]代替手工特征提取的方式, 从上述两个多尺度时空图中分别提取出 rPPG 特征, 以充分捕捉时空图中的活体信息并给出二分类预测结果.

4.4 时空信息

单帧图像含有人脸的空间信息, 典型的如真实人脸的深度信息, 或者欺骗人脸的平面信息等. 视频流中连续多帧图像之间也含有人脸由于运动而产生的信息, 典型的如真实人脸的非刚性运动信息, 或者欺骗人脸的刚性运动信息等. 注意, 这种信息是时序相关的(蕴含在连续的多帧图像中), 因而我们将其简称为时间信息. 传统的方法大多基于单帧图像从空间信息的角度进行 FAS 检测, 但这忽略了多帧图像之间蕴含的时间信息. 研究表明, 多帧方法在检测效果上可能要优于单帧方法. 直觉上, 这是因为多帧方法相对于单帧方法可学习到额外的时间域特征, 但是多帧方法也不可避免地增加了算法的时间复杂度. 近年来, 主流的方法趋向于基于时空融合进行 FAS 模型设计[15], 其基本流程如表 4 所示: 首先, 空间特征提取器(如 CNN)从输入的多帧图像中探索纹理或者深度等空间特征; 然后, 时间特征提取器(如 RNN)融合各帧的空间特征, 并从时间维度发现可用于区分的运动信息; 最后, 基于上述时空特征完成真假人脸分类. 目前, 时空信息融合方法研究的重难点在于: 对于空间信息, 如何挖掘有别于传统方法(如深度图)的新型空间信息源; 对于时间信息, 如何设计可捕获长期(long-term)运动行为、并可与空间信息完美融合的网络结构; 最后, 如何减少时空融合算法的时间复杂度也是需要考虑的问题.

Xu 等人[67]将 FAS 检测看作视频分类问题, 并提出了长短期记忆 CNN 网络, 即 LSTM-CNN (long short term memory-CNN). LSTM-CNN 是一种堆叠式架构, 其中, CNN 位于底层, 用于对视频流的每一帧挖掘空间信息; LSTM 位于中层, 用于接收 CNN 所挖掘的空间信息, 并从中继续挖掘时间信息以及完成时空融合; 最上层是 Softmax, 其接收 LSTM 输出的时空融合信息, 并根据条件概率判定人脸的真假. Li 等人[69]针对传统 2D CNN 只考虑空间信息而忽略时间信息的问题, 提出了时空融合的 3D CNN 人脸检测方法. 在 3D CNN 中, 每一个卷积核都是一个形如 $W\times H\times T$ 的 3 维时空结构, 其中, $W\times H$ 是传统卷积核的空间大小, T 是卷积帧数(即时间深度). 与已有的方法相比, 3D CNN 可学习到额外的时间可判别特征. Wang 等人[57]发现, 真假人脸在运动

时会出现深度差异, 由此提出了“时序深度 TD (temporal depth)”的方法. TD 架构分为单帧和多帧两部分, 其中, 单帧部分采用传统的 CNN, 用以生成人脸深度图; 多帧部分将光流引导特征块 OFFB (optical flow guided feature block)[120]和卷积门控循环单元 ConvGRU (convolution gated recurrent units)级联, 用以捕获时空梯度和形态等人脸运动特征, 并生成多帧深度图(multi-frame depth map). 最后, 通过对多帧深度图进行深度监督和二元监督完成真假人脸分类. 针对寻求新的可区分空间信息问题, Wang 等人[15]发现: 除了传统的“高层”(high-level)语义特征外, “低层”(low-level)语义特征——如在卷积堆叠时可能被丢弃的空间梯度振幅 SGM (spatial gradient magnitude)——也可以作为重要的空间特征来源, 该发现拓宽了空间信息的挖掘思路. 文献[15]在文献[57]的基础上改进实现, 其架构也分为单帧和多帧两部分, 其中, 单帧部分设计了新的残差空间梯度块 RSGB (residual spatial gradient block), 用以捕获 SGM; 多帧部分利用性能更好的短期空时块 STSTB (short-term spatio-temporal block)取代 OFFB, 并仍然与 ConvGRU 级联后生成多帧深度图. 最后, 基于多帧深度图进行深度监督学习, 实现真假人脸判定.

时空融合可以基于整张图像学习全局时空信息, 也可以对图像裁剪之后学习局部时空信息. 比较而言, 后者虽然增加了计算量, 但是由于粒度更细, 因而可能获得一些独特的性能. Lin 等人[70]发现: 回放视频的相邻帧之间存在着真实人脸所不具备的关联运动模式, 而这种关联运动模式可利用计算机视觉领域的平面单应性(planar homography)进行描述. 为此, 他们提出可将视频流中相邻两帧图像均裁剪成大小相同的 9 块局部区域, 并计算两帧图像中对应的局部区域是否存在平面单应性关系: 如果存在, 即可判定为虚假人脸. Yang 等人[71]提出了一种基于“全局时间-局部空间”的 FAS 方法. 其基本思想是: 首先, 利用 CNN-LSTM 从视频流中融合全局时空信息, 并经过二元监督后进行初步的真假人脸识别; 然后, 将上述过程生成的 CNN 信息送入区域注意力模块, 由区域注意力模块根据 CNN 提供的信息识别出图像重要的局部区域; 最后, 重点对图像的重要局部区域进行特征学习完成模型训练. 这种“全局时间-局部空间”的方法不仅可以增强对欺骗攻击的抵御能力, 而且具备一定的模型可解释能力. Cai 等人[72]受人类区分真假人脸方式——即首先从全局照片定位可能的局部欺骗区域, 然后从可能的局部欺骗区域寻找欺骗线索——的启发, 提出了一种“全局空间-局部时间”的 CNN-RNN 架构, 其基本思想是: 首先, 利用 ResNet18 网络对视频流中的每帧图像学习全局空间特征进行初步的真假人脸判断; 然后, 利用门控循环单元 GRU (gated recurrent unit), 采用强化学习的方式, 从初判的真假人脸图像中递归学习局部时间特征, 得到需要重点关注的局部区域; 最后, 将全局空间特征和局部时间特征相融合完成真假人脸分类.

4.5 多模态

所谓的多模态(modality), 是指来自不同来源并以不同形式展示的信息, 如视频、语音和文字等, 其每种都是一种模态. 特定到 FAS 领域, 有 3 种常用的模态图像, 即 RGB 图、深度图和红外图. 三者优势互补, 各有所长, 其中, RGB 图具有高保真的特点, 可以清晰地呈现纹理信息, 但是仅能覆盖光谱中的可见光区域, 且对光照的鲁棒性较差; 深度图可以详细地提供空间结构信息, 且对光照的鲁棒性较强; 红外图可以全天候工作, 并能够穿透视觉上的遮挡和障碍, 但是对比度和分辨率较差, 在图像上具有较大的灰度同质区域. 针对同一个对象, 多模态可以提供语义相关、内容互补的异构信息, 从而基于多模态可以发掘出从单模态所无法发掘的特征, 进而增强 FAS 的检测能力. 基于多模态的 FAS 算法流程图见表 4: 首先, 特征提取器(通常为 CNN)分别从 RGB 图、红外图和深度图中提取各自模态的可区分特征; 然后, 将上述不同模态的可区分特征进行融合; 最后, 基于融合后的特征完成真假人脸判定.

针对当前 FAS 数据集绝大多数为 RGB 数据集[73]的问题, 文献[16]发布了首个面向多模态的大规模数据集 CASIA-SURF, 其中包含 RGB 图、深度图和红外图等 3 种不同的模态图像. 在与 CVPR2019 协办的 Chalearn LAP 多模态 FAS 攻击挑战赛[121]中, Parkin 等人[73]在 CASIA-SURF 数据集上取得了第一的优异成绩, 其基本思想是, 将网络分为两部分.

- 一部分是主干部分, 由 ResNet-34 和 ResNet-50 网络以及挤压和激励 SE (squeeze and excitation)模块构成. 主干部分用以学习 3 种模态各自的特征, 其方法是为每种模态建立 1 个独立的通道, 每个通道均

利用 res_1, res_2 和 res_3 残差块学习不同层次的特征. 3 个通道的 res_3 输出经过挤压和激励后送入统一的 res_4.

- 另一部分是聚合部分, 用以学习 3 种不同模态之间的关联特征. 聚合部分对主干部分的 3 个残差块 res_1, res_2 和 res_3 分别设计了对应的聚合块 agg_1, agg_2 和 agg_3, 其中, 每一个聚合块 agg_i 都从其对应的残差块 res_i 获取特征, 并与前一个聚合块 $agg_j(j=i-1)$的输出结果进行聚合(agg_1 不执行这一步, 只获取 res_1 的特征). 最后, agg_3 的结果也送入 res_4.

显然, 该方案不仅考虑各个模态自己的独立特征, 同时也考虑了多个模态之间的关联特征, 取得了很好的效果. Shen 等人[74]为解决过拟合问题提出了 FaceBagNet 网络, 该网络首先从不同模态人脸图像的局部区域提取特征, 然后将上述不同模态的人脸特征进行融合, 最后再对融合后的特征随机擦除某一模态特征之后进行分类. Zhang 等人[75]指出: 多模态方法虽然比单模态方法性能更优, 但是增加了模型的复杂度. 为此, 他们将深度卷积 DWConv (DepthWise convolution)[122]用于网络中以降低网络参数量, 同时使用逐步判定方法进一步降低时间复杂度. 具体来说, 其首先将深度图作为网络的输入, 若无法给出输入样本的具体类别, 再使用红外图做进一步判断.

种族的差异会影响人脸欺骗检测算法的泛化能力, 为此, CASIA-SURF CeFA[17,76]通过对 CASIA-SURF 进行东亚、中亚和非洲三地的种族图像扩展, 成为目前最大的跨种族多模态数据集. Liu 等人[76]基于 CASIA-SURF CeFA 提出了部分共享分支多模态网络 PSMM-Net (partially shared branch multi-modal network), PSMM-Net 首先根据 CASIA-SURF CeFA 中 3 种模态的静态图像生成其各自的动态图像; 然后利用 3 个分支分别从不同模态的静态图像和动态图像中提取各自的模态特征; 最后将 3 个分支所提取的特征相互融合, 以充分利用不同模态图像之间的互补信息来提升真假人脸的识别率. Yu 等人[77]使用 CDC[55]代替传统卷积方式得到 CDCN[55], 并将 CDCN 扩展成多模态网络结构, 以从 CASIA-SURF CeFA 中学习多模态跨种族特征, 提升了模型在跨种族情形下的 FAS 能力. 针对多模态方法难以抵御高质量面具的问题, Yang 等人[78]提出了 PipeNet, 该网络的优势在于, 针对 CASIA-SURF CeFA 提供的不同模态图像, 选择最合适的分支网络结构, 以最大化利用多模态信息. 具体来说, 他们为 RGB 图像和红外图选择了 ResNeXt[123], 为深度图选择了 SE-ResNet[124], 将网络从多个模态学习的特征相拼接后, 送入融合模块得到最终判定结果.

4.6 欺骗痕迹分离

许多基于深度学习的 FAS 方法将网络所挖掘出的人脸欺骗特征视为一个不可分割的整体, 事实上, 这种欺骗特征是多种“强相关-弱相关-不相关”特征互相纠缠在一起的“纠缠体”. 这种表现为“纠缠体”的欺骗特征一方面制约了 FAS 识别性能的进一步提升; 另一方面, 也使得深度学习方法缺乏可解释性, 进而可能导致算法的可信性和安全性问题. 为此, 一些 FAS 算法[61,85−87]考虑从欺骗人脸中相对精确地分离出本质的欺骗特征, 我们将其称为欺骗痕迹分离. 这类 FAS 算法可分为 3 个基本步骤(参见表 4): 首先, 将欺骗人脸 $\hat{I}$ 输入到生成器 G 中, 生成欺骗痕迹图像 $G(\hat{I})$; 然后, 利用欺骗人脸 $\hat{I}$ 减去欺骗痕迹图像 $G(\hat{I})$, 生成“合成的真实图像”, 即 $\hat{I}-G(\hat{I})$; 最后, 根据分离的欺骗痕迹计算出最终的分类结果. 为分离出可靠的欺骗痕迹 $G(\hat{I})$. 需要达到两个目标: 其一是使得“合成的真实人脸” $\hat{I}-G(\hat{I})$ 和“原有的真实人脸”I 服从同一分布; 其二是使得“合成的真实人脸 $\hat{I}-G(\hat{I})$ ”和“原有的真实人脸”I 之间的差异最小.

针对第 1 个目标, 可引入鉴别器 D, 用以尽可能区分出 $\hat{I}-G(\hat{I})$ 和 I 之间的分布差异, 其对抗损失如公式(3)所示[85,86].

$$L_{GAN}(G,D)=\mathbb{E}_{I\in R}[\log D(I)]+\mathbb{E}_{\hat{I}\in S}[\log(1-D(\hat{I}-G(\hat{I})))] \tag{3}$$

其中, R 和 S 分别为原数据集中的真实人脸集合和欺骗人脸集合.

针对第 2 个目标, 需要测量 $\hat{I}-G(\hat{I})$ 和 I 之间的差异, 对应损失函数如公式(4)所示[85,86].

$$L_{rec}=\|(\hat{I}-G(\hat{I}))-I\|_1 \tag{4}$$

最终分离欺骗痕迹的优化目标为

$$G^* = \arg\min_G \max_D (L_{GAN} + L_{rec}) \tag{5}$$

其中, 对于鉴别器 D, 总损失 L_{GAN}+L_{rec} 越大, 说明其辨别原始人脸和合成人脸的能力越强; 对于生成器 G, 总损失 L_{GAN}+L_{rec} 越小, 说明 D 越难以区分原始人脸还是合成人脸, 即分离的欺骗痕迹越可靠.

Jourabloo 等人[85]设计了新型的 CNN 结构, 其包含欺骗痕迹分离网络 DS Net (de-spoof network)、鉴别质量网络 DQ Net (discriminative quality network)和视觉质量网络 VQ Net visual quality net)这 3 个部分, 其中, DS Net 用于从欺骗人脸中分离出欺骗痕迹图, 并与欺骗人脸相减后得到重构的真实人脸; DQ Net 用于估计所重构的真实人脸的深度信息; VQ Net 用于尝试区分所重构的真实人脸和原有的真实人脸. 这 3 个网络结构协同反馈迭代工作, 当最终由 DQ Net 得到的所重构人脸的深度信息与真实人脸的深度信息相似并且 VQ Net 无法区分所重构的真实人脸和原有的真实人脸时, 说明 DS Net 分离出的欺骗痕迹质量非常高. 类似地, Liu 等人[86]提出了欺骗跟踪解纠缠网络 STDN (spoof trace disentanglement network). STDN 除了重构真实人脸图像以外, 还通过从欺骗人脸分离出的欺骗痕迹图合成新的欺骗人脸图像. 这样, 相比于文献[85]就存在两个待区分元组: (原有的真实人脸图像,重构的真实人脸图像)和(原有的欺骗人脸图像,合成的欺骗人脸图像), 当判别器对两个元组均不能区分, 并且合成的欺骗人脸再次经过欺骗痕迹分离, 得到的新欺骗痕迹和原有欺骗人脸分离出的欺骗痕迹差异达到最小时, 证明生成了可靠的欺骗痕迹图. 相比于文献[85], 文献[86]具有更强的可解释性和抗多样攻击能力. Feng 等人[87]提出, 由欺骗痕迹生成器和辅助分类器组成的 FAS 检测网络. 具体地说, 欺骗痕迹生成器以 U-Net[125]为主干从输入图像中学习欺骗痕迹, 辅助分类器基于欺骗痕迹生成器的输出进一步放大欺骗信息以区分出真假人脸, 所提出的方案能够较好地解决由于过拟合所导致的泛化能力差的问题.

本节对基于深度学习的 6 种代表性 FAS 方法进行了介绍. 其中, 第 4.1 节的二元监督方法实现相对简单且速度相对较快, 但由于分类准确率不高, 一般只用于预处理阶段的“粗”分类, 其得到的结果再输入到其他方法进行下一步的精确分类; 第 4.2 节–第 4.5 节的深度图、rPPG 信号、时空信息、多模态信息本质上是人脸欺骗信息的不同来源, 基于这些欺骗信息来源设计新的 FAS 检测方法是一项重要的工作, 但更有挑战性的是能否发掘出新的欺骗信息来源. 目前, 这方面的进展暂时不大. 第 4.6 节的欺骗痕迹分离本质上是一种“求精”(表征学习/特征解耦)方法, 它建立在如下观察之上: 许多 FAS 方法所挖掘出的特征是与挖掘目标“强相关-弱相关-不相关”的多类特征相互纠缠在一起的“纠缠体”, 若能从这种“纠缠体”中解耦出强相关的特征, 显然可以提升方法的识别性能和可解释性. 上述“求精”(表征学习/特征解耦)方法是目前研究的热点和难点; 另一方面, 在实际实现中也并不局限于挖掘欺骗特征, 亦可以挖掘包括活体特征在内的其他感兴趣特征. 最后需要指出的是, 在深度学习领域, 域泛化[7,126,127]和可解释性[10,127,128]是目前较为前沿的研究领域, 基于深度学习的 FAS 也不例外. 接下来, 我们将对基于深度学习的 FAS 的域泛化和可解释性问题进行探讨.

5 基于深度学习的 FAS 域泛化与可解释性

在深度学习领域, 域泛化指在某个训练数据集(也称为源域)上训练模型, 然后在除了训练数据集之外的测试数据集(称为不可见域 unseen domain, 或者目标域)上测试模型的通用性. 由于深度学习所基于的 i.i.d 假设——源域与不可见域独立同分布——在实践中往往并不成立, 这导致基于源域所训练的模型在面对不可见域时会出现性能恶化. 因此, 所有基于深度学习的方法都必须面对域泛化 DG (domain generalization)问题. 第 5.1 节和第 5.2 节将分别对域泛化的基本理论以及 FAS 的域泛化代表性方法进行说明.

另一方面, 深度学习海量的参数与复杂的处理机制使得人类很难追溯与理解其推理过程, 导致对这类黑箱学习很难进行解释, 进而引发了人们对深度学习的可信性、安全性和公平性的担忧与质疑. 例如, 对输入数据施加人类无法察觉的微小对抗扰动[129,130], 即可能完全改变深度学习系统的行为: 当加入人眼无法察觉的扰动后, 一张猫的图片可以被系统以高置信识别为狗; 毫无意义的白噪声也可被误认为是某个特定的对象[10]. 当上述技术应用于人脸识别系统时, 攻击者可以在人类无法感知的情况下恶意控制人脸识别的结果. 解决上述的问题的途径是研究 FAS 的可解释性, 第 5.3 节和第 5.4 节将分别对可解释性基本理论以及 FAS 的可解释性代表性方法进行说明.

5.1 域泛化及其基本理论

2021 年, Wang 等人[7]针对域泛化问题从理论、方法、应用和数据集方面给出了最新的综述. 根据文献[7], 理论上, 解决域泛化 DG 问题有两种主要思路.

(1) 一是借鉴域适应 DA (domain adaption)的思想. DA 和 DG 的区别在于, 目标域对于 DA 是已知的, 对于 DG 是未知的, 因而 DG 更具一般性. 借鉴 DA 的思想, 可通过最小化所学习到的模型与目标域之间的分类错误风险(risk of classification error)来实现 DG[131]. 形式地, 上述分类错误风险可表示为 $\varepsilon^t(h,h^{*t})$[7]. 这里, h 是模型所学习到的分类器分类, h^{*t} 是目标域的真实分类, ε是两者之间的差异. 前已说明, 目标域对于 DG 是未知的, 因而 h^{*t} 未知, 故而直接最小化$\varepsilon^t(h,h^{*t})$是不可行的. 但是, 注意到源域是已知的, 即 h^{*s} 已知, 人们转而寻求基于已知的源域分类错误风险$\varepsilon^s(h,h^{*s})$来间接最小化未知的目标域分类错误风险$\varepsilon^t(h,h^{*t})$, 其基本方法是: 首先, 建立$\varepsilon^s(h,h^{*s})$和$\varepsilon^t(h,h^{*t})$之间的不等式约束关系; 然后, 通过调整$\varepsilon^s(h,h^{*s})$来达到间接最小化$\varepsilon^t(h,h^{*t})$的目的[7].

(2) 二是基于域不变表达 DA-DIR (DA based on domain-invariant representation)实现 DG[132]. 其基本思想是: 对于给定的源域$\mathcal{X}$和目标域$\mathcal{Y}$, 由于两者之间的分布差异是固定不可改变的, 因而可寻找一个映射函数 $g:\mathcal{X}\to\mathcal{Z}$, 将源域$\mathcal{X}$映射到一个中间表示空间(representation space)域$\mathcal{Z}$, 并通过缩小$\mathcal{Z}$和$\mathcal{Y}$之间的分布差异来实现 DG[132].

方法上, 解决 DG 问题有 3 种主要思路[7]: 一是数据操作, 二是表征学习, 三是学习策略.

(1) 数据操作

数据操作从数据多样性的角度提升模型的泛化能力. 除了通过传统方法如反转、旋转、缩放、裁剪、添加噪音等对数据施加扰动以减少模型的过拟合实现泛化之外, 当前主流的方法是生成更多不属于源域$\mathcal{X}$的多样性数据来增强模型的泛化能力. 例如, 随机数据生成[133]基于有限的训练样本通过模拟复杂的环境生成新的随机数据. 但是, 引入随机性不可避免地会破坏数据的语义空间, 进而导致模型识别准确率的降低. 为此, 关于数据操作, 人们更加关注于如何在生成多样性数据的同时仍能保证模型分类的准确率. 解决这个问题的关键在于"在语义空间约束的前提"下生成数据: 由于数据的语义空间不变, 因而所学习到的分类器的判定准确率自然也不会降低. 为在语义空间约束下生成数据, 当前常用的方法包括对抗数据生成[134,135], 以及利用相关生成模型如变分自编码器 VAE (variational autoencoder), 生成对抗网络 GAN (generative adversarial networks)进行数据生成等.

(2) 表征学习

表征学习在机器学习中占有重要的地位[136], 由于其解除了对人类专家知识和经验的依赖, 可自动学习数据的表征, 因而获得了广泛关注, 也是当前 DG 领域最流行的方法.

所谓的表征是指: 对所观察到的(关于输入数据的)所有潜在解释因素的后验分布, 深度学习的目标是通过多种非线性变换组合的方式, 为预测器生成更为抽象和有用的表征[136]: 形式地, 若将深度学习的预测器表示为复合函数 $f(g(x))$, 其中, f 是分类函数, g 是表征函数, x 是输入, 则表征学习需要学习函数 g, 使得预测器 $f(g(x))$所预测的分类与真实分类之间的数学期望最小[7].

根据对不同域之间共同表征获取方法的不同, 表征学习可以分为域不变表征学习 DIRL (domain-invariant representation learning)和特征解耦 FD (feature disentanglement)两大类.

① 域不变表征学习 DIRL

域不变表征学习寻求不依赖于域的不变表征, 其建立在如下理论的基础上: 如果一个特征表示(feature representations)对于所有的域都是不变的, 那么该特征就是泛化的, 并且可以迁移到不同的域[132]. 将上述理论应用到 DG, 如果我们能在一个特定的特征空间(feature space)中, 尽可能减少不同源域之间的表征差异(representation discrepancy), 就得到域不变表征学习的方法. 在具体实现上, 核函数、特征对齐、对抗学习等均可用来学习域不变表征.

② 特征解耦 FD

特征解耦基于如下事实: 任何一个特征都是由多重表征因素构成的, 这些多重表征因素可以表示为一个向量. 特征解耦尝试将上述表征分量从特征向量中解耦出来, 当每个域的特征均被解耦为多重表征之后, 再寻找不同域之间的共同表征作为域不变表征. 在实现时, 特征解耦可以从模型和数据两方面进行.

1) 第一, 当从模型方面进行特征解耦时, 其将每个域的特征向量的表征分量对应于模型的参数. 不失一般性, 形式地, 对于任意第 i 个域, 假设其模型参数集(特征向量)为 p_i, 则 p_i 可以表示为 $p_i=p+\Delta_i$, 其中, p 是所有域共享的域不变参数集(域不变表征), Δ_i 是特定于第 i 个域的域特定参数集(域特定表征). 由此, p 即为所求的域不变表征.

2) 第二, 当从数据方面进行特征解耦时, 其从领域级别(domain-level)、样本级别(sample-level)和标签级别(label-level)这 3 个方面制定样本的生成机制, 在保持空间分布语义约束的前提下, 生成更多的数据以寻求域不变表征. 在生成新的数据时, 会借助一些数据生成模型, 最常见的是变分自编码器 VAE. 自编码器 Autoencoder 是一种人工神经网络, 以无监督的方式训练网络忽略信号“噪声”来学习一组数据的表征并表达为编码的形式; 变分 Variational 则约束模型以避免过拟合以及在感兴趣的潜在空间分布上编码.

(3) 学习策略

一些工作尝试从新的学习范式的角度入手进行泛化, 将其统称为学习策略的方法.

① 集成学习(ensemble learning)

集成学习建立在如下假设之上: 任何一个样本都看作是多个源域的一个综合样本, 因此整体预测结果可以看作是多个域网络预测结果的叠加. 在实现时, 集成学习结合多个模型, 通过特定的网络架构设计和训练策略使得不同模型之间互相协作, 最后对不同模型预测结果的权重进行聚合以给出最终预测结果.

② 元学习(meta-learning)[137]

元学习尚难以给出统一的定义[137], 但是元学习与传统深度学习相比, 其关键差异在于“学习如何学习(learn-to-learn)”, 因而元学习基于少量的基础学习即可以有很好的应对新任务的能力. 从数学上, 可以认为元学习比传统深度学习多了一维参数, 不失一般性, 假设数据集为$\mathcal{D}=\{(x_0,y_0),\ldots,(x_N,y_N)\}$; 深度学习所得到的模型为 $\hat{y}=f_\theta(x)$, 其中, θ为模型 f 的参数集, 则元学习需要求解的问题为

$$\theta^*=\arg\min_\theta \mathcal{L}(\mathcal{D},\theta,\omega) \tag{6}$$

其中, $\mathcal{L}$是损失函数, θ是模型 f 的参数集, ω表示“如何学习(how to learn)”的目标, f 和θ两者均是ω的函数. 对于传统的深度学习, ω是一个预定义的定值, 因而求解公式(6)只与$\mathcal{D}$和θ相关; 对于元学习, ω根据目标(如模型性能、准确率、泛化能力等)会发生动态变化, 因而求解公式(6)与$\mathcal{D}$, θ和ω均相关, 进而函数 f 及其参数集θ均会随着目标ω的变化而变化. 显然, 元学习良好的适应新任务能力可以增加泛化能力[137].

以上介绍了域泛化的基本理论和思想, 具体到 FAS 域泛化的实际实现上, 目前表征学习(域不变表征学习和特征解耦)、元学习和生成对抗网络是最常用的技术手段. 限于篇幅, 我们拟选取域不变表征学习[58,89−95]、零样本或小样本学习[96,97]、元学习[59,60]等为代表, 对上述技术进行说明.

5.2 FAS域泛化方法

5.2.1 基于域不变表征学习的方法

域不变表征学习[58,89−95]将源域和目标域映射到同一特征空间中, 然后从该同一特征空间中学习可用于区分真假人脸的特征, 由此可确保从源域训练得到的模型在目标域中也能取得满意的结果. 基于域不变表征学习的 FAS 泛化方法主要包括两类: 特征对齐方法[89,90]和对抗[58,91−95]方法. 特征对齐方法通过对齐源域和目标域之间的特征分布来实现泛化; 对抗方法则利用生成对抗网络 GAN[138]和梯度反转层 GRL (gradient reversal layer)[139]等对抗思想, 让域分类器区分样本是来自源域还是目标域, 当域分类器无法区分时, 说明源域和目标域的分布相似, 从而实现了泛化.

(1) 特征对齐方法

Li 等人[89]最早针对域泛化问题展开研究, 其使用最大均值差异 MMD (maximum mean discrepancy)[140]来衡量源域和目标域之间的差异, 并通过最小化 MMD 实现源域和目标域两者在特征空间的对齐. 在对齐特征空间之后, 再从中训练区分真假人脸的分类器, 从而达到泛化的目的. Tu 等人[90]针对不同数据集的特征分布不一致的问题, 提出可通过 MMD 最小化不同域之间的距离, 并利用 CNN 最小化真假人脸的分类错误, 由此实现了不同域之间真实人脸的相互靠拢以及虚假人脸的相互靠拢.

(2) 基于对抗的方法

Shao 等人[58]认为, 不可见域和源域之间一定存在某些共同特征, 进而可通过从源域中学习这些共同特征来实现对不可见域的泛化. 以打印攻击为例, 虽然源域和不可见域在采集环境、欺骗载体材质等多种因素上可能存在差异, 但两者均要将欺骗人脸打印在纸张上才能发起攻击, 由此, 纸张即可以作为源域和不可见域的共同特征. 基于上述观察, 他们提出, 可利用对抗学习从多个源域中学习共同特征. 具体地说, 他们让特征生成器(用于生成多个源域的共有特征)和域鉴别器(用于区分所生成的共有特征是否来自于某个特定的源域)相互竞争, 最终当特征生成器能够欺骗所有的域鉴别器时, 即学到了所有源域的共同特征. 但是上述泛化操作可能会降低 FAS 分类的准确率[141,142], 这是因为不同数据域之间真实人脸的距离可能要远于同一数据域内部真实人脸和欺骗人脸的距离. 为此, 他们进一步提出了双力三元挖掘约束 DFTC (dual-force triplet-mining constraint)的方法. DFTC 利用机器学习中的三重态损失函数(triplet loss), 使得三元组(基准人脸, 真实人脸和欺骗人脸)满足: 基准人脸与真实人脸之间的距离最小化, 且基准人脸与欺骗人脸之间的距离最大化. 最终, DFTC 可达成如下两个目标.

(1) 给定一个特定的数据域 D, 任意选取 D 内两个真实人脸 R_D 和 R'_D, 则 R_D 和 R'_D 之间的距离一定小于“R_D 与 D 内任意一个欺骗人脸的距离以及 R'_D 与 D 内任意一个欺骗人脸的距离”(该目标保证了模型具有良好的分类能力).

(2) 给定任何两个跨域数据集 D_1 和 D_2, 任意选取 D_1 内一个真实人脸 R_{D_1} 以及 D_2 内一个真实人脸 R_{D_2}, 则 R_{D_1} 和 R_{D_2} 的距离一定小于“R_{D_1} 与 D_2 内任意一个欺骗人脸的距离以及 R_{D_2} 与 D_1 内任意一个欺骗人脸的距离”(该目标保证了模型具有良好的泛化能力).

文献[58]从理论上属于域不变表达 DA-DIR, 具体而言, 其映射函数 g 采用的是对抗学习模型; 中间表示域$\mathcal{Z}$选取的是所有源域$\mathcal{X}$的共有特征; 由于作者认为所寻找的共有特征$\mathcal{Z}$是不可见域$\mathcal{Y}$也具备的, 因而$\mathcal{Z}$和$\mathcal{Y}$之间的分布差异趋近于 0. 但是研究人员发现, 对于中间表示空间域$\mathcal{Z}$, 从真实人脸构造$\mathcal{Z}$是容易的, 从欺骗人脸构造$\mathcal{Z}$却比较困难. 这是因为, 对于来自不同源域$\mathcal{X}$的真实人脸(即跨域真实人脸), 其相互之间的分布差异不大; 相反, 对于跨域欺骗人脸, 其相互之间的分布差异却较大. 上述问题会使得在泛化时聚拢跨域真实人脸相对容易、但聚拢跨域虚假人脸却比较困难. 由此可能出现跨域真实人脸间距离大于跨域真假人脸间距离的情况, 进而降低泛化后 FAS 的识别准确率. 为了解决这个问题, Jia 等人[91]提出了单边域泛化框架, 该框架仅仅对跨域真实人脸进行(以泛化为目标的)“全局”聚拢, 而放弃了对跨域虚假人脸的“全局”聚拢. 具体地说, 对于虚假人脸, 文献[91]利用非对称三元损失函数 ATL (asymmetric triplet loss)[143]将跨域虚假人脸在其各自所属的域中进行“局部”聚拢, 由此得到跨域“全局”聚拢的真实人脸以及域内“局部”聚拢的虚假人脸. 最终仍然保证跨域真实人脸之间的距离小于跨域真假人脸之间的距离, 从而实现泛化.

5.2.2 基于零样本小样本学习的方法

零样本或小样本学习[59,60,96]从理论上来自于域适应 DA, 从方法上属于表征学习或者元学习. 许多方法在训练和测试时所使用的欺骗样本都是同类型的[48,54,64,85], 如同为打印攻击或者同为视频攻击等. 然而这类方法在面对未经学习过的新型攻击时可能表现不佳, 因此, 需要寻求一种可基于已学习的知识面对未知攻击的网络模型. 针对这个问题, 一种比较重要的方法是零样本和小样本学习(zero- and few-shot learning)[59,60,96]. 两者的区别在于, 零样本学习仅仅对现有的欺骗攻击类型进行学习而不包含任何新型攻击的样本[96], 而小样本学习则引入少量的新型攻击样本[59,60].

为使得 FAS 算法在无法获得新型欺骗攻击样本的情况下依然能够应对新型欺骗攻击类型, Liu 等人[96]提

出了一种零样本学习方案, 其将所有已知欺骗类型的欺骗样本分为多个组, 当有未知的欺骗攻击类型出现时, 将它划分到与欺骗类型最相似的组再给出预测结果. 基于这一思想, 他们提出了深度树状网络 DTN (deep tree network). DTN 由卷积残差单元 CRU (convolutional residual unit)、树路由单元 TRU (tree routing unit)和监督表征学习 SFL (supervised feature learning)这 3 个模块组成, 其中, CRU 从输入的图像中提取特征, 是具有残差结构的卷积模块; TRU 负责从树状网络的根节点出发, 通过递归计算所输入图像与当前节点的左右儿子节点之间的相似性而选择不同的路由路径进入子树, 直到最终到达叶子节点; SFL 在 TRU 到达叶子节点之后, 结合二元监督和像素级监督完成欺骗表征的学习.

5.2.3 基于元学习的方法

元学习的关键是学习如何学习, 其基于少量的基础学习即可以有很好的应对新任务的能力. 在学习阶段, 元学习会有多个训练任务, 若将第 i 个任务损失记为 l_i, 则在所有训练任务上的总损失为 $L(F)=\sum_{i=1}^{N} l_i$, 当 $L(F)$ 达到最小时, 元学习模型 F 达到了学习目标. 在 FAS 领域, 欺骗攻击类型众多, 元学习快速适应新任务的能力使得它在 FAS 泛化领域有着很好的潜力.

针对 FAS 泛化问题, Qin 等人[59]提出了自适应内部更新 AIU (adaptive inner-update)方法. AIU 可以让元学习器从已有的攻击类型中归纳出通用的可区分特征, 并基于目标域中的少量样本迅速更新优化元学习器的参数以适应新的欺骗攻击类型, 从而提升泛化能力. Shao 等人[60]针对元学习算法仅能模拟单个域转换的情况, 提出可利用深度图作为辅助信息来规范化特征空间, 同时, 将源域划分为多个元训练集和元测试集, 通过多个域转换训练/测试场景引导元学习模型参数向更具泛化能力的方向更新, 避免模型优化过程中过度适应某一个域.

5.3 可解释性基本理论

根据文献[10], 可解释性有 4 种呈现形式, 即基于规则(rule)、基于隐藏语义(hidden semantics)、基于属性(attribution)和基于案例(example). 这里介绍对 FAS 相对更有借鉴意义的前 3 种方法[10].

(1) 基于规则的方法将可解释性形式化为逻辑规则或者决策树.

在实现上, 又可以分为分解(decomposition)方法和教学(pedagogical)方法.

- 分解方法将模型视为白盒, 通过“分解”网络的内部连接得到逻辑规则集. 但是分解方法的时间复杂度和网络规模成指数关系, 因而这种方法只能面对小型网络.
- 教学方法将模型视为不可知的黑盒, 其直接从网络的“输入-输出”关系中学习规则, 从而将时间复杂度降低为多项式, 更具实用性. 教学方法的学习过程本质上可归约为传统的规则学习或者决策树学习, 因此可应用典型的算法, 如: 对于规则学习, 可采用序贯覆盖(sequential covering)算法来生成规则集; 对于决策树, 可采用 CART (classification and regression tree)或者 C4.5 算法来生成决策树.

在可解释性的所有 4 种呈现形式中, 基于规则的可解释性具有最为坚实的数学基础, 且能够同时对模型给予全局和局部的解释, 但是在实践中需要控制规则的复杂度, 以避免过于复杂而难以应用.

(2) 基于隐藏语义的方法多用于机器视觉领域.

根据是否需要对已学习到的模型做可解释性相关的改变, 基于隐藏语义的可解释性方法可以进一步划分为“被动式隐藏语义方法”和“主动式隐藏语义方法”. “被动式的隐藏语义”方法不对已学习到的模型做任何改变, 其尝试从“可视化(visualization)”的角度寻求模型中神经元(neuron)、通道(channel)、卷积层(layer)或者卷积核(kernel)与图像概念(concept)之间的关系. “可视化”的思想来源于神经科学的“祖母细胞”假设, 所谓的“祖母细胞”假设是指: 在对特定图像(这里假设为祖母的图像)进行记忆和识别时, 是大脑中的某个特定细胞(称为祖母细胞)——而不是整个神经网络——所完成的. 具体地说, 当特定图像(祖母图像)出现时, 这个特定细胞(祖母细胞)就被激活. 因此, “可视化”可以采用如下激活最大化(activation maximum)公式得到在图像概念识别中起决定作用的神经元/通道/卷积层/卷积核.

$$x^* = \arg\max_x (act(x;\theta) - \lambda\Omega(x)) \tag{7}$$

其中, $act(\cdot)$是感兴趣的神经元的激活函数, θ是模型训练的参数集, Ω是可选的正则化器. 在研究早期, 研究人员通过 $act(\cdot)$寻找激活的神经元, 后来人们发现高频噪声是影响识别的主要因素, 故而转向寻求具有更好先验知识或者更好正则化的Ω, 并通过压制高频振幅和高频噪声、裁剪不重要像素、基于 GAN 生成高分辨率的逼真图像学习先验知识等方法, 提升隐藏语义的解释效果. 除了可视化以外, 挖掘卷积核和图像概念之间的语义关联关系也是一种思路[10], 研究人员发现: 为了编码图像中的一个概念, 往往需要多个卷积核, 通过卷积核嵌入, 可以更好地表征图像概念. 这个发现也从侧面证明了: 虽然 CNN 的高层过滤器已经学习到了一些图像的对象级(object)概念, 如图像中的头和脚等, 但是这些概念是互相纠缠在一起的, 换句话说, 高层过滤器学习到的某个“概念”实质是多个不同概念对应模式的混合体. 针对这个问题, 主动式隐藏语义学习的方法尝试对模型进行修改, 通过增加损失项使得高层过滤器尽可能只表达唯一的概念.

(3) 基于属性的方法根据模型是白盒还是黑盒有两种不同的处理方式.

当模型是白盒时, 主要利用显著图(saliency map)对模型进行解释. 显著图是一种图像分割模式, 显著图的目标在于将一般图像的表征简化或者变换为更易于分析的形式, 显著图获取的核心在于梯度的计算, 针对在梯度计算时可能存在的不同情况, 研究人员提出了不同的梯度定义和计算方法(参见文献[10]的表 4). 需要指出的是, 攻击者也可以对显著图进行攻击: 通过生成人类无法感知差异的对抗样本, 使得模型做出相同的分类, 但却生成截然不同的显著图, 从而做出错误的解释. 在利用显著图解释 FAS 模型时, 需要对此引起注意. 当模型是不可知的黑盒时, 显然梯度计算的方法不可行. 为此, 研究人员提出可基于合作博弈的思想, 将模型的最终输出看作是不同输入特征合作博弈的结果, 并借鉴合作博弈从合作产生的总收益中为每个参与者分配回报的做法, 计算每个特征对于最终分类的贡献. 除此以外, 通过扰动、遮挡或者修改图像观察输出结果的变化进行敏感性分析; 以及计算具有特定特征的输入图像与输出结果之间的最大互信息, 也都是提升可解释性的思路.

5.4 FAS可解释性方法

5.4.1 基于隐藏语义的方法

CNN 具有深层复杂结构及黑盒特性, 特别是在 CNN 的高层特征中, 多个图像概念相互纠缠, 这使得理解 CNN 网络内部如何运作变得十分困难, 为此, 对隐藏语义(即纠缠在一起的图像概念)进行解纠缠(disentangle)是一种常用的 CNN 可解释性方法[61,88,101].

从活体特征解纠缠的角度, Zhang 等人[61]认为人脸图像是由对人脸识别起决定性作用的活体特征 L(liveness)和对人脸识别无关的光照、背景等内容特征 C(content)共同构成的, 人脸识别的关键是从人脸图像中明确区分出活体特征 L, 并将其用于真假人脸识别. 为了将与内容特征纠缠在一起的活体特征解纠缠出来, Zhang 等人[61]提出了一种基于真假人脸活体特征交换的方法, 其基本思想是:

- 首先, 从真实人脸 R 中提取潜在的“真实活体特征”R_L 和内容特征 R_C 得到元组 $R(R_L,R_C)$.
- 然后, 对 R 对应的欺骗人脸 S 类似提取“欺骗活体特征”S_L 和内容特征 S_C 得到元组 $S(S_L,R_C)$.
- 接着, 基于 R 和 S 合成新的图像, 即利用 R 中的 R_L 替换 S 中的 S_L, 由此合成得到新图像 $S_R(R_L,S_C)$. 由于 S_R 是利用 R_L 替换了 S 中的 S_L, 因而可以认为是“真实人脸”.
- 继续从“合成的真实人脸” S_R 中提取活体特征得到 R_L' (注意: 虽然合成 S_R 时是利用 R_L 替换 S_L, 但由于“原始的真实人脸”R 和“合成的真实人脸”S_R 之间仍然存在差异, 因而提取到的活体体征为 R_L').
- 最终, 当提取得到的 R_L 和 R_L' 满足损失函数约束时, 即认为活体特征 R_L 解纠缠成功.

在具体实现时, 他们采用纹理图和深度图辅助学习完成活体特征的解纠缠.

除了活体特征解纠缠以外, 欺骗特征解纠缠也有其现实意义[88]: 其一, 可以给出具体的解释以说明为什么识别为欺骗人脸; 其二, 可以通过从欺骗人脸中“减”去欺骗特征实现真实人脸重建; 其三, 可以基于欺骗特征和真实人脸合成新的欺骗人脸, 从而提升模型的泛化能力. Liu 等人[88]提出了一种基于生成对抗的欺骗特征

解纠缠方法, Liu 等人将各类攻击形式下的人脸欺骗特征抽象为两类过程——添加过程(additive process)和修改过程(inpainting process)——的复合结果, 其中, 添加过程指欺骗材质所引入的额外模式, 如莫尔图案等; 修改过程指欺骗材质完全覆盖了真实人脸的特定区域. 为了解纠缠出上述两类欺骗特征, Liu 等人首先利用一个 CNN 主干编码器从低频(如颜色失真)、中频(如化妆攻击)和高频(如摩尔图案和面具边缘)分别抽取表征, 然后将表征送入基于 CNN 的欺骗特征解码器, 分别解纠缠出添加欺骗特征和修改欺骗特征.

5.4.2 基于属性的方法

基于属性的 FAS 方法[71,102−104]利用显著图观察输入的图像中哪些区域对模型做出决策时的贡献最大. George 等人[102]使用二元标签图(即为输入图像的每个像素提供一个二元标签. 在学习阶段, 若输入图像为真实人脸, 则对应全 1 图; 若输入图像为欺骗人脸, 则对应全 0 图)监督网络学习, 然后对输入图像的每个像素做出预测生成二元图. 二元图即清楚解释了在真假人脸判定时每个像素所起的作用. 深度图是另一种常见的解释工具. Wu 等人[103]针对双像素 DP (dual-pixel)传感器被广泛应用于智能终端的情况, 提出一种基于 DP 双图像预测深度图的方法. 由于 DP 双图之间虽然存在差异, 但是这种差异很小, 不足以作为深度图构建的基准, 为此, 他们提出了两条规则: 一是转换一致性, 即 DP 双图像之间的像素偏移必须满足特定的规则, 这条像素偏移相关的规则是由 DP 传感器的硬件布局特点所决定的; 二是相对深度标签, 即在构建深度图时依据文献[144]的思想, 采用相对深度而不是绝对深度, 由此所估计的深度图能够很好地 用于真假人脸分类和 FAS 解释. Yang 等人[71]观察到, 欺骗线索在人脸图像中并不是平均分布的, 其表现为有的区域丰富、有的区域稀疏. 为此, Yang 等人提出了一种基于热力图和区域注意力机制的人脸图像欺骗线索定位方法, 其基本思想是: 首先,利用可视化工具 Grad-CAM[145]生成热力图, 并由热力图从人脸图像中初步“粗”定位出需要关注的候选人脸区域; 然后, 再利用区域注意力模块从候选人脸区域中进一步精确定位出欺骗线索最为丰富的“学习区域”; 最后, 将“学习区域”用于真假人脸判定. 上述过程中, 热力图和“学习区域”即可用于解释图像被判定为欺骗人脸的原因. Deb 等人[104]利用得分图(score map)和二进制掩码图(binary mask map)来定位人脸图像中的欺骗区域. 具体来说:

- 他们首先使用全卷积网络在二元交叉熵损失的约束下生成得分图, 在得分图中每一个像素都有一个对应的分数值, 该分数值越大, 则代表其输入图像对应的感受野为欺骗区域的可能性越大.
- 然后, 对得分图进行最小最大归一化(min-max normalization), 以将每个像素的分数值均映射到[0,1]区间.
- 最后, 将归一化后的得分图转换为二进制掩码图, 其转换规则为: 如果得分图中某个像素的分数值不小于预设的阈值(文献中阈值取 0.5), 则该像素值被置 1; 否则置 0.

二进制掩码图中被置 1 的区域即为欺骗区域, 从视觉上, 二进制掩码图中高亮部分(像素为 1)能够清晰地显示图像的哪些部分被识别为欺骗区域.

6 数据集对比

数据集对于训练 FAS 模型以及评估模型的有效性有着极为重要的作用. 从模态的角度, 可将数据集分为单模态数据集和多模态数据集: 单模态数据集中的样本都是 RGB 模态的图像/视频; 而多模态数据集除了 RGB 模态之外, 还包含其他模态图像(如深度图、红外图等). 主流的 FAS 数据集对比结果见表 5.

表 5 人脸欺骗检测数据集对比

数据集	年份	人数	样本类型和数量 Total(real, fake)	模态类型	姿态	攻击类型	录制工具	欺骗媒介 (载体)
NUAA[1]	2010	15	I, 12 614 (5105, 7509)	RGB	Frontal	Print	Webcam (640×480)	A4 Paper, Photographic Paper
Replay-Attack[3]	2012	50	V, 1 300 (300, 1000)	RGB	Frontal	Print, Replay	MacBook (320×240)	A4 Paper iPad (1024×768), iPhone
CASIA-MFSD[2]	2012	50	V, 600 (150, 450)	RGB	Frontal	Print, Replay	USB Camera (640×480), Sony NEX-5 (1280×720)	Copper Paper, iPad
3DMAD[5]	2013	17	V, 255 (170, 85)	RGB, Depth	Frontal, Profile	3D Mask	Microsoft Kinect for Xbox 360 (640×480)	Hard Resin
MSU MFSD[37]	2015	55	V, 440 (110, 330)	RGB	Frontal	Print, Replay	MacBook Aircamera (640×480) Google Nexus 5 camera (720×480)	iPad Air, iPhone 5S, A3 Paper
OULU-NPU[146]	2017	55	V, 5 940 (1980, 3960)	RGB	Frontal	Print, Replay	Samsung, HTC, MEIZU, ASUS Zenfone Selfie, Sony, OPPO	A3 Paper, UltraSharp 1905FP (1280×1021), Macbook 13 (2560×1600)
SiW[64]	2018	165	V, 4 620 (1320, 3300)	RGB	[−90°,90°]	Print, Replay	Canon EOS T6 (1920×1080), Logitech C920 (1920×1080)	Glossy Paper, Matt Paper, iPad Pro, iPhone 7, Galaxy S8, Asus MB168B
SiW-M[96]	2019	493	V, 1 628 (660, 968)	RGB	[−90°,90°]	Print, Replay, 3D Mask, Makeup, Partial	Logitech C920, Canon EOS T6	–
CASIA-SURF[16]	2019	1 000	V, 21 000 (3000, 18000)	RGB, Depth, IR	[−30°,30°]	Print	RealSense (RGB: 1280×720, Depth and IR: 640×480)	A4 Paper
CASIA-SURF CeFA[17]	2019	1 607	V, 23 538 (4500, 19038)	RGB, Depth, IR	[−30°,30°]	Print, 3D Mask	RealSense (1280×720)	Silica Gel, Cloth
CelebA-Spoof[62]	2020	10 177	I, 625 537 (182385, 443152)	RGB	[−30°,30°]	Print, Replay, 3D Mask, Paper Cut	–	24 Devices (four types: PC, Camera, Tablet, Phone)

注: 表中第 4 栏 I, V 分别表示图像(Image)、视频(Video); 第 5 栏 RGB, Depth, IR 分别表示 RGB 图、深度图、红外图

6.1 单模态数据集

- NUAA[1]

NUAA 于 2010 年提出, 是第一个用于 FAS 的数据集. 该数据集总共包含 15 个参与者(受测试者)的 12 614 张人脸图像, 分为 5 105 张真实人脸图像和 7 509 张打印攻击人脸图像. 对于真实人脸图像, 为使其欺骗人脸图像尽可能无明显差异, 在录制时要求无眨眼、摇头等明显的面部运动及表情变化. 对于欺骗人脸图像, 分别采用相纸和 A4 纸打印出彩色人脸图像, 并对所打印的欺骗人脸图像采用平移、弯曲和旋转等方式模拟真实人脸图像.

- Idiap Replay-Attack[3]

Idiap Replay-Attack 又简称为 Replay-Attack, 该数据集由 50 个受测试者分别在不同光照条件下用 MacBook 拍摄而成, 拍摄分辨率为 320×240. 数据集共有 1 300 个视频, 分为 300 个真实人脸视频和 1 000 个

欺骗人脸视频. 对于 1 000 个欺骗人脸视频, 每个欺骗人脸视频采集时长均在 10 s 左右, 在拍摄时, 采用手持或者固定欺骗媒介(如纸张或电子屏幕等)两种方式. 当手持欺骗媒介时, 由于会发生抖动, 因而可起到欺骗眨眼检测器的作用. 相比于 NUAA, Replay-Attack 具有更多的个体数, 而且具有 NUAA 所不具备的时序(运动)特征. 该数据集可用于训练针对打印攻击和重放攻击的 FAS 模型.

- CASIA-MFSD[2]

该数据集采集了 50 名受测试者的真假人脸视频, 视频共有 600 个, 分为 150 个真实人脸视频和 450 个欺骗人脸视频, 在采集真实人脸视频时, 受测试者需要有眨眼动作. 所采集的欺骗人脸视频有两种类型.

(1) 一是由打印照片拍摄而来. 具体地说, 首先将人脸图像打印为照片, 然后对照片进行扭曲或者裁剪操作以“模拟”真实人脸, 最后对扭曲或裁剪的照片进行拍摄即得到欺骗人脸视频. 考虑到部分 FAS 算法从图像质量的角度进行 FAS 检测, CASIA-MFSD 共收集了 3 种不同质量的人脸图像, 其中的高清图像由 Sony NEX-5 相机拍摄而来.

(2) 二是由视频重放拍摄而来. 即直接在其他设备播放人脸视频, 然后对所播放的视频进行重拍摄而得到欺骗人脸视频. 和 NUAA 以及 Replay-Attack 相比, CASIA-MFSD 具有更丰富的欺骗类型和更高的图像质量.

- MSU MFSD[37]

该数据集是第一个使用手机摄像头模拟手机欺骗场景的数据集, 该数据集从 55 名受测试者中收集了 440 个视频, 其中包括 110 个真实人脸视频和 330 个欺骗人脸视频. 视频的平均时长为 12 s. 类似于 CASIA-MFSD, MSU MFSD 的欺骗人脸视频也由打印照片和重放视频拍摄而来, 其中,

(1) 对于打印照片视频, MSU MFSD 采用佳能 550D 单反相机拍摄后, 打印在 A3 纸张上, 再对打印在 A3 纸张上的照片拍摄视频得到打印照片视频.

(2) 对于重放攻击视频, MSU MFSD 又分为两种情况: 一是由佳能 550D 相机拍摄后使用 iPad Air 回放, 二是由 iPhone 5S 拍摄后使用 iPhone 5S 回放. 该数据集除了所拍摄的照片质量更高以外, 更重要的是能够用于移动设备的人脸欺骗检测.

- Oulu-NPU[146]

该数据集由 55 个受测试者在 3 种不同的光照条件下拍摄而成, 共有 5 940 个短视频, 其中包含 1 980 个真实人脸视频以及 3 960 个欺骗人脸视频. 拍摄视频使用了 6 种不同的高清手机设备. 在捕获欺骗攻击样本时, 有意避免了屏幕和纸张边框等明显的欺骗伪影. 该数据集主要优势在于, 除了欺骗攻击数量丰富以外, 还模拟了复杂的欺骗场景(如光照和背景变化等), 有助于进行 FAS 的泛化研究.

- SiW[64]

该数据集收集了来自 165 名受测试者的 4 620 个视频, 涵盖了人脸到相机不同距离的变化、光线的变化、扭头角度的变化和表情的变化, 每个受测试者都有 8 个真实人脸视频和 20 个欺骗人脸视频, 其中, 欺骗人脸视频通过拍摄两种不同质量的打印照片以及 4 种不同的电子屏幕而来. 该数据集主要优点在于: 参与者基数大, 视频数量丰富; 参与个体来自于多种族, 考虑了人脸识别的种族差异; 包含了深度、光照、姿态和表情等多种因素.

- SiW-M[96]

该数据集总共从 493 个受测试者中收集了 1 628 个视频, 分为 660 个真实人脸视频以及 968 个欺骗人脸视频. 其中, 欺骗人脸视频采用 1080 P 高清录制, 共涵盖了 13 种不同的欺骗攻击类型, 包括: 1 种打印攻击、1 种重放攻击、5 种 3D 面具攻击、3 种化妆攻击和 3 种部分遮挡攻击. 此外, 该数据集还考虑到了人脸姿势和光照强度极端变化的情况. 与其他数据集相比, SiW-M 数据集具有最为丰富的欺骗攻击类型.

6.2 多模态数据集

- CASIA-SURF[16]

当前数据集大多数受测试者数量不超过 500, 特别是绝大多数数据集仅为单模态 RGB 数据集. 针对受测

试者数量少这一问题, CASIA-SURF 数据集选取了 1 000 个受测试者录制了 21 000 个多模态视频, 其中包含 3 000 个真实人脸视频和 18 000 个欺骗人脸视频. 欺骗人脸视频由打印攻击视频和重放攻击视频两种类型构成, 其中, 对于打印攻击视频, 要求对每个受测试者录制 6 种不同的打印攻击视频, 其通过首先对受测试者的照片进行不同的欺诈操作——如裁剪照片的关键区域(眼睛、鼻子、嘴巴)等以“伪造”不同的人, 或者弯曲照片以“伪造”真实人脸才具有的 3D 结构等, 然后再对欺诈照片进行拍摄得到. 针对多模态问题, CASIA-SURF 采用 Intel RealSense SR300 相机为每个受测试者拍摄了 RGB 图、深度图和红外图这 3 种模态图像.

- CASIA-SURF CeFA[17]

种族因素对人脸识别也有影响, 但是在 CASIA-SURF CeFA 之前, 几乎没有数据集考虑多种族因素, 由此导致模型容易对某些种族人群过拟合. 为此, CASIA-SURF CeFA 通过采集东亚、中亚和非洲等 3 个区域不同种族的人脸图像/视频对 CASIA-SURF 进行了扩展. 具体到 CASIA-SURF CeFA, 其从 1 607 个受测试者中收集了 23 538 个视频, 考虑了光照等环境因素, 涵盖 RGB、深度和红外这 3 种模态. 欺骗类型共有 4 种, 包括不同光照条件下的 2D 打印照片攻击、2D 重放视频攻击、3D 打印面具攻击和 3D 硅胶面具攻击. 对于 2D 攻击(打印攻击和重放攻击), 其从东亚、中亚和非洲每个区域各选取 500 人, 从 3 个模态的角度, 每人包含 1 张真实人脸、2 张分别从室内和室外捕获的虚假打印照片以及 1 份虚假重放视频, 共计 18 000 个样本. 对于 3D 攻击(打印面具和硅胶面具), 其基于 99 个受测试者在 6 种不同光照条件下拍摄了 5 346 个 3D 打印面具攻击视频, 基于 8 个受测试者在 4 种不同光照条件下拍摄了 192 个硅胶面具攻击视频, 共计 5 538 个 3D 面具攻击样本.

从以上数据集可以看出, 未来数据集采集的趋势是: 增加拍摄者数量、考虑多重因素(如种族、年龄、性别、亲属关系、载体材质等)、考虑环境变化(如光照条件、图像质量、深度信息、姿态变化、表情变化等)、考虑信息来源丰富(多模态)、考虑复杂或者新型攻击(如超真实 3D 面具攻击、部分遮挡攻击、化妆攻击等). 考虑得越完备, 所提供的数据集对于研发人员探索新的 FAS 思路和方法就越有益. 但是, 更加完备的数据集并非总是带来正面收益, 也可能会给 FAS 带来新的挑战. 例如, 由于数据集是在人们精心设定下所采集的, 数据集中数据所展现的某些表征在目标域中可能不会出现, 由此, 若 FAS 模型基于在目标域中不会出现的表征进行建模, 很可能会导致效果不理想. 为此, 需要研究泛化、可解释性、多表征融合等一般化的方法来解决上述问题. 总的来说, 更加完备的数据集对于 FAS 新思路新方法的探索、设计、训练和评估是更有益的.

7 实验分析

7.1 评估标准

每个数据集都有自己的评估协议, 在每个协议中, 都会按照统一的评估标准对 FAS 算法进行评估, 其中最常用的标准有 Anjos 等人[43]提出的平均错误率 HTER (half total error rate)[147]和等错误率 EER (equal error rate)[147]以及 ISO/IEC 提出的呈现攻击分类错误率 APCER (attack presentation classification error rate)、真实呈现分类错误率 BPCER (bona fide presentation classification error rate)和曲线下面积 AUC (area under roc curve).

以下对上述评价标准的含义进行说明.

在 FAS 系统中, 真假人脸分类错误有两种情况: 将欺骗人脸误认为是真实人脸和将真实人脸错误分类为欺骗人脸. 这两种错误分类的比率分别称为错误接受率 FAR (false acceptance rate)和错误拒绝率 FRR (false rejection rate), 公式(8)给出了计算公式.

$$FAR=\frac{FA}{NI},\ FRR=\frac{FR}{NC} \tag{8}$$

将 FAR 和 *FRR* 取均值即得到 HTER. 改变设定的检测阈值使得 FAR 和 FRR 两者相等即得到 EER, 其计算公式如公式(9)所示.

$$HTER=\frac{FAR+FRR}{2},\ EER=FAR=FRR \tag{9}$$

APCER 和 FAR 类似, 但是 FAR 表示的是各类欺骗攻击类型样本中被错误分类为真实人脸的比例, 而

APCER 表示某一特定欺骗攻击类型样本中被错误分类为真实人脸的比例. BPCER 和 FRR 含义相同. ACER 表示为 APCER 和 BPCER 的平均值.

$$APCER = \frac{1}{N_A}\sum_{i=1}^{N_A}(1 - Res_i) \tag{10}$$

$$BPCER = \frac{1}{N_R}\sum_{i=1}^{N_R} Res_i \tag{11}$$

$$ACER = \frac{APCER + BPCER}{2} \tag{12}$$

上述所有评估指标的值越小, 代表模型表现越好. 另一方面, AUC 采用接收者操作特征曲线 ROC (receiver operating characteristic curve)下的面积来评估模型的优劣, AUC 的数值越大, 代表模型表现越好.

以上公式(8)中, FA 是错误接受的总次数, NI 是虚假人脸出现的总次数, FR 是错误拒绝的总次数, NI 是真实人脸出现的总次数. 公式(10)–公式(12)中, N_A (number of specific attack)表示某一类特定攻击类型的总攻击次数, N_R (number of real face)表示以真实人脸为样本的总检测次数. 在检测过程中, 如果第 i 次检测被分类为欺骗人脸, 则 Res_i 置 1; 否则置 0.

7.2 实验对比

本节对传统方法的性能(表 6)、深度学习方法的性能(表 7 和表 8)以及深度学习的泛化能力(表 9)进行实验对比.

表 6 传统方法在 Replay-Attack, CASIA-MFSD, MSU MFSD 数据集上的对比测试结果

方法	Replay-Attack		CASIA-MFSD	MSU MFSD
	EER (%)	HTER (%)	EER (%)	EER (%)
IQA[35]	–	15.2	32.4	–
Motion[43]	11.6	11.7	26.6	–
DMD[26]	5.3	3.8	21.8	–
LBP[3]	13.9	13.8	18.2	–
DOG[2]	–	–	17.0	–
LBP-TOP[38]	7.9	7.6	10.0	–
IDA[37]	–	7.4	–	8.5
GCT[29]	1.2	4.2	4.6	1.5
LBP+HOOF[42]	–	–	3.1	**0.0**
Color SURF[30]	**0.1**	**2.2**	2.8	2.2
Color texture[24]	0.4	2.8	**2.1**	4.9

注: 表中 ERR 和 HTER 的数值越小越好

表 7 基于深度学习的方法在 OULU-NPU 数据集上, 采用协议 4 的对比测试结果

方法	APCER (%)	BPCER (%)	ACER (%)
DeepPixBiS[102]	36.7±29.7	13.3±14.1	25.0±12.7
MILHP[70]	15.8±12.8	8.3±15.7	12.0±6.2
TSCNN[82]	11.3±3.9	9.7±4.8	9.8±4.2
Auxiliary[64]	9.3±5.6	10.4±6.0	9.5±6.0
SAPLC[106]	11.9±7.0	6.7±5.5	9.3±4.4
FAS-TD[57]	14.2±8.7	4.2±3.8	9.2±3.4
STASN[71]	6.7±10.6	8.3±8.4	7.5±4.7
BASN[63]	6.4±8.6	3.2±5.3	4.8±6.4
De-Spoofing[85]	5.1±6.3	6.1±5.1	5.6±5.7
STPM[15]	6.7±7.5	3.3±4.1	5.0±2.2
BCN[56]	2.9±4.0	7.5±6.9	5.2±3.7
CDCN++[55]	4.2±3.4	5.8±4.9	5.0±2.9
Disentanglement[61]	5.4±2.9	3.3±6.0	4.4±3.0
Spoof Trace[86]	**2.3±3.6**	**5.2±5.4**	**3.8±4.2**

注: 表中 APCER, BPCER 和 ACER 的数值越小越好. 表中加粗的行表示表内所有方法中性能对比最好的方法

表 8 基于深度学习的方法在 SiW 数据集上, 采用协议 1–协议 3 的对比测试结果

协议	方法	APCER (%)	BPCER (%)	ACER (%)
1	Auxiliary[64]	3.58	3.58	3.58
	STASN[71]	–	–	1.00
	FAS-TD[57]	0.96	0.50	0.73
	STPM[15]	0.64	0.17	0.40
	BCN[56]	0.55	0.17	0.36
	Disentanglement[61]	0.07	0.50	0.28
	CDCN[55]	0.07	0.17	0.12
	Spoof Trace[86]	**0.0**	**0.0**	**0.0**
2	Auxiliary[64]	0.57±0.69	0.57±0.69	0.57±0.69
	STASN[71]	–	–	0.28±0.05
	FAS-TD[57]	0.08±0.14	0.21±0.14	0.15±0.14
	STPM[15]	0.00±0.00	0.04±0.08	0.02±0.04
	BCN[56]	0.08±0.17	0.15±0.00	0.11±0.08
	Disentanglement[61]	0.08±0.17	0.13±0.09	0.10±0.04
	CDCN[55]	0.00±0.00	0.13±0.09	0.06±0.04
	Spoof Trace[86]	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
3	Auxiliary[64]	8.31±3.81	8.31±3.81	8.31±3.81
	STASN[71]	–	–	12.10±1.50
	FAS-TD[57]	3.10±0.81	3.09±0.81	3.10±0.81
	STPM[15]	2.63±3.72	2.92±3.42	2.78±3.57
	BCN[56]	2.55±0.89	2.34±0.47	2.45±0.68
	Disentanglement[61]	9.35±6.14	1.84±2.60	5.59±4.37
	CDCN[55]	**1.67±0.11**	**1.76±0.12**	**1.71±0.11**
	Spoof Trace[86]	8.3±3.3	7.5±3.3	7.9±3.3

注: 表中 APCER, BPCER 和 ACER 的数值越小越好. 表中加粗的行表示在不同评估协议下性能对比结果最好的方法

表 9 FAS 算法在不同数据集之间的交叉域泛化能力对比测试结果

类别	方法	[O,C,I]→M		[O,M,I]→C		[O,C,M]→I		[I,C,M]→O	
		HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
传统方法	MS_LBP[33]	29.76	78.50	54.28	44.98	50.30	51.64	50.29	49.31
	IDA[37]	66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
	CT[24]	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
	LBPTOP[32]	36.90	70.80	42.60	61.05	49.45	49.54	53.15	44.09
深度学习方法	CNN[13]	29.25	82.87	34.88	71.95	34.47	65.88	29.61	77.54
	Aux(Depth)[64]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	66.61
	Aux(All)[64]	–	–	28.4	–	27.6	–	–	–
	MMD-AAE[148]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
	MADDG[58]	17.69	88.06	24.5	84.51	22.19	84.99	27.98	80.02
	Cross-Domain[95]	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
	RFM[60]	13.89	93.98	20.27	88.16	17.3	90.48	16.45	91.16
	SSDG-R[91]	**7.38**	**97.17**	**10.44**	**95.94**	**11.71**	**96.59**	**15.61**	**91.54**

注: 表中 HTER 的数值越小越好, AUC 的数值越大越好. 表中加粗的行表示表内所有方法中性能对比最好的方法

表 6 选取代表性的 Replay-Attack, CASIA-MFSD 和 MSU MFSD 这 3 个数据集对传统方法进行性能对比, 可见, 基于颜色纹理的方法在 3 个数据集上都要优于基于图像质量的方法和基于运动的方法. 我们认为主要原因有两个: 首先, 从数据集的角度来看, 上述 3 个数据集由于收集的时间较早, 因而数据规模相对较小、所包含的欺骗类型有限且图像的分辨率较低, 这使得基于图像质量和运动的方法能提取的特征较少; 其次, 从方法的角度来看, 基于图像质量的方法通过捕获真假人脸图像在质量上的差异来发现攻击, 但是在采集数据集样本时, 人们往往会尽量避免样本出现明显的图像质量问题, 因此加大了此类方法的难度; 另一方面, 基于运动的方法通过捕获真假人脸图像在时间维度上的运动差异来发现攻击, 但是一般情况下上述差异并不明显, 如何捕获精细的运动差异在方法上仍然存在提升空间. 比较而言, 基于颜色纹理的方法可从不同的颜色空间中捕获真假人脸的颜色纹理差异, 不仅可提取的特征更为丰富, 而且进行特征提取时更为容易, 因而在数据集中表现最优. 从表 6 还可以发现, 几乎所有的传统方法在 CASIA-MFSD 和 MSU MFSD 数据集上的性能都比 Replay-Attack 要差. 这是因为 CASIA-MFSD 和 MSU MFSD 不仅考虑的欺骗场景(如环境的变化、图

像的质量等)更为复杂, 而且受测试者的数量要远远多于 Replay-Attack, 故其所包含的欺骗人脸样本之间的差异更大, 这对 FAS 算法提出了更高的要求.

基于深度学习的方法已经全面超越了传统方法而一跃成为当前的主流方法. 针对基于深度学习 FAS 方法的性能对比问题, 我们选择较新的 OULU-NPU[146] (2017 年)和 SiW[64] (2018 年)两个数据集展示对比结果, 分别见表 7 和表 8. 对 FAS 方法进行评估时, 数据集需要设置“协议(protocol)”以评估方法的性能. OULU-NPU 数据集上共设有 4 种协议[146], 其中, 前 3 种分别从光照和背景、欺骗载体(如打印机、纸张和电子屏幕)、样本采集设备的角度对 FAS 方法的性能进行评估, 第 4 种协议则同时集成前 3 种协议以模拟真实的环境. 表 7 中列出了在 OULU-NPU 数据集上以最为逼近真实环境的第 4 种评估协议为标准的对比结果; 表 8 列出了在 SiW 数据集上以前 3 种评估协议为标准的对比结果. 表 7 和表 8 中加粗的行标示了数据集中当前协议下性能最优的方法.

从方法类别来看, 表 7 和表 8 不包括二元监督的方法, 这是因为仅仅将 FAS 看作是二分类问题并不能反映真实人脸和欺骗人脸之间的本质差别, 一般只能用于预处理阶段的“粗”分类. 从方法实现来看, 使用深度图对网络进行像素级监督学习的 FAS 方法[15,55,56,61]以及基于分类欺骗痕迹的 FAS 方法[85,86]在 OULU-NPU 和 SiW 数据集上均取得了优异的表现, 这表明深度图监督和分离欺骗痕迹可以让网络自发学习到真假人脸更为本质的特征, 同时更具泛化能力, 不会因为光照等环境的变化而大大降低性能, 也不会因为数据集的不同而导致明显的性能差异. 我们认为这是因为深度图本身对光照不敏感, 更重要的是, 分离欺骗痕迹以特征解耦的方式可以更准确地学习数据的本质表征[86].

除了在数据集内部进行比较外, 近年来, 越来越多的研究开始注重数据集之间的交叉泛化能力的比较. 表 9 展示了传统方法和基于深度学习方法在跨数据集测试上的泛化能力对比结果, 其中, O, C, I, M 分别表示数据集 OULU-NPU, CASIA-MFSD, Idiap Replay-Attack, MSU MFSD; 形式[A]→B 表示在数据集 A 上进行训练, 在数据集 B 上进行测试. 从表 9 可以看出, 传统人脸欺骗检测算法在各项评估标准中的结果和基于深度学习的人脸欺骗检测算法的结果差距较大. 这是因为传统方法通过手工设计的算子来提取特征, 而这些算子是数据集依赖的, 即根据数据集的特点而设计, 因而跨数据集测试的效果明显不如深度学习方法.

8 未来展望

由于基于深度学习的 FAS 方法已经全面超越传统方法(两者对比参见表 9), 因而本节对基于深度学习的 FAS 方法进行展望.

(1) FAS 域泛化

未来 FAS 域泛化的挑战在于:

- 一是如何进行连续域泛化. 当 FAS 用于流动场所时, FAS 系统输入数据的统计特征可能会不断变化, 此时需要进行连续的域泛化, 以适应不同统计特征的数据.
- 二是域泛化的可解释性和可信性. 在进行域泛化时, 不可避免地会引入新的特征或丢掉一些自有特征, 但这可能会导致总体分类准确率下降. 为此, 需要对域泛化进行解释, 以增强域泛化的可信性和鲁棒性.
- 三是面向新型 FAS 攻击的域泛化研究. 随着超真实 3D 面具、化妆攻击、整形攻击等新型攻击日渐成熟, 面向这些新型以及未知攻击的域泛化研究还不多见. 如何面向这类新型和未知攻击进行域泛化, 也是一个值得研究的问题.

(2) FAS 可解释性

未来 FAS 可解释性的挑战在于:

- 一是面向 FAS 的可解释深度学习模型. 一些研究采用形式化方法, 如基于规则的方法[10]将可解释性形式化为逻辑规则或者决策树, 这类方法具有坚实的数学基础, 且能够对模型同时给予全局和局部解释, 但是往往过于复杂而难以实际应用.

- 二是需要建立FAS可解释性的评估标准. 当前, 对于FAS算法的可解释性并没有统一的评估标准, 大多数情况下仍然依赖于人类的专家知识和经验, 这使得研究人员难以标准化地评估模型的可解释性程度.
- 三是需要研究FAS可解释性自身的可信性. 可解释性是增强深度学习可信性的重要手段, 但是, 可解释性自身也可以被攻击. 例如, 当采用显著图作为FAS解释工具时, 攻击者可以通过生成人类无法感知差异的对抗样本, 使得模型做出相同的分类, 但却生成截然不同的显著图, 从而做出错误的解释. 因此, 如何确保可解释性自身的可信性, 也是需要考虑的问题.

(3) 基于元学习的FAS

在新的深度学习范式中, 元学习迎来了爆发式的增长[137]. 元学习通过从机器学习模型的多个学习阶段(这些学习阶段通常涵盖相关任务的分布)中“学习”经验, 然后基于所“学习”到的经验进一步提高其未来的“学习”能力. 这种“学习如何学习”的方式更接近人类的思维方式[137]. 相比于传统深度学习实现的“特征-模型”学习, 元学习的目标是实现“特征-模型-算法”学习. 换句话说, 传统深度学习在面临一个新任务时由于采用固定算法而必须重新学习, 但元学习可以基于已有的学习根据学习策略而自我演化. 具体到 FAS, 一些基于零样本或者小样本的元学习在FAS领域已经成功应用[59,60], 充分显示了元学习的巨大潜力.

(4) 面向新型和未知攻击的FAS

随着科技的发展, 一些具有一定门槛的攻击技术逐步成熟走向实用, 典型的如超真实3D面具攻击、蜡像攻击、局部遮挡攻击、化妆攻击[149]等. 这类攻击从纹理、rPPG、时空、模态等常见欺骗线索采集的源头(参见第4.2节–第4.5节)发起攻击, 手段更有针对性, 因而更容易绕过人脸识别系统. 具体到上述4类攻击, 其中,

- 超真实 3D 面具攻击和蜡像攻击可以逼真地还原纹理信息和空间信息(深度信息), 也可以较好地反映时间信息(运动信息), rPPG 暂时对检测这类攻击具有一定的优势, 但 rPPG 信号自身很容易受到噪声干扰, 有时甚至会淹没在全局噪声中[48]; 特别地, 未来也不排除有透光性良好的材质出现.
- 局部遮挡攻击是一类特殊的打印照片攻击, 其利用打印照片对真实人脸进行遮挡, 以“替换”在识别时权重最高的真实人脸区域, 并保留人脸其他区域不变. 这种攻击较好地保留了人脸的纹理、时空、模态、rPPG 信号等信息, 因而具有较强的反 FAS 能力.
- 化妆攻击通过妆容“改变”人脸固有的特征, 也能够较好地保留人脸的时空、模态、rPPG 等信息.

由此可见, 单纯地从(即便是多个)传统欺骗线索采集源头来设计 FAS 方案在面对新型和未知攻击时将越来越困难. 针对上述趋势, 我们认为, 未来泛化和可解释性不应作为FAS设计时可有可无的可选因素, 而是必须要融入到 FAS 设计中: 应通过域泛化增强对(新型和未知攻击所位于的)不可见域的分类能力; 通过可解释性加强对(新型和未知攻击的)攻防理解能力, 从而从底层增强对FAS的本质规律认知. 除此以外, 新兴的学习范式, 如元学习、神经架构搜索 NAS[150,151]等, 也对新型和未知攻击有着良好的抗衡潜力.

9 总 结

随着人脸识别系统的广泛普及, 人脸反欺诈 FAS 成为研究的热点. 本文对 FAS 所面临的主要科学问题和相应的解决方法进行了介绍, 重点阐述了基于深度学习的主流 FAS 方法. 在此基础上, 对相关数据集和实验评估结果进行了对比总结. 最后展望了未来可能的研究方向.

References:

[1] Tan XY, Li Y, Liu J, Jiang L. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: Proc. of the European Conf. on Computer Vision (ECCV 2010). Crete: Springer, 2010. 504–517.

[2] Zhang ZW, Yan JJ, Liu SF, Lei Z, Yi D, Li SZ. A face antispoofing database with diverse attacks. In: Proc. of the IAPR Int'l Conf. on Biometrics (ICB 2012). New Dehli: IEEE, 2012. 26–31.

[3] Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing. In: Proc. of the Int'l Conf. of Biometrics Special Interest Group. Darmstadt: IEEE, 2012. 1–7.

[4] Patel K, Han H, Jain AK. Secure face unlock: Spoof detection on smartphones. IEEE Trans. on Information Forensics and Security (TIFS 2016), 2016, 11(10): 2268−2283.

[5] Nesli E, Marcel S. Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect. In: Proc. of the 6th IEEE Int'l Conf. on Biometrics: Theory, Applications and Systems. Arlington: IEEE, 2013. 1−8.

[6] Li XB, Komulainen J, Zhao GY, Yuen PC, Pietikäinen M. Generalized face anti-spoofing by detecting pulse from face videos. In: Proc. of the Int'l Conf. on Pattern Recognition (ICPR 2016). Cancun: IEEE, 2016. 4244−4249.

[7] Wang JD, Lan CL, Liu C, Ouyang YD, Qin T. Generalizing to unseen domains: A survey on domain generalization. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI 2021). Montreal: Morgan Kaufmann Press, 2021. 4627−4635.

[8] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations (ICLR 2014). Banff, 2014.

[9] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015). Boston: IEEE, 2015. 427−436.

[10] Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans. on Emerging Topics in Computational Intelligence, 2021, 5(5): 726−742.

[11] Feng LT, Po LM, Li YM, Xu XY, Yuan F, Cheung TC, Cheung KW. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. Journal of Visual Communication and Image Representation, 2016, 38: 451−460.

[12] Li L, Feng XY, Boulkenafet Z, Xia ZQ, Li MM, Hadid A. An original face anti-spoofing approach using partial convolutional neural network. In: Proc. of the Int'l Conf. on Image Processing Theory, Tools and Applications. Oulu: IEEE, 2016. 1−6.

[13] Yang JW, Lei Z, Li SZ. Learn convolutional neural network for face anti-spoofing. arXiv: 1408.5601, 2014.

[14] Zhu X, Li S, Zhang XP, Li HL, Kot AC. Detection of spoofing medium contours for face anti-spoofing. IEEE Trans. on Circuits and Systems for Video Technology, 2021, 31(5): 2039−2045.

[15] Wang ZZ, Yu ZT, Zhao CX, Zhu XY, Qin YX, Zhou QS, Zhou F, Lei Z. Deep spatial gradient and temporal depth learning for face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2020). Seattle: IEEE, 2020. 5042−5051.

[16] Zhang SF, Wang XB, Liu A, Zhao CX, Wan J, Escalera S, Shi HL, Wang ZZ, Li SZ. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach: IEEE, 2019. 919−928.

[17] Liu AJ, Tan ZC, Wan J, Escalera S, Guo GD, Li SZ. CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikola: IEEE, 2021. 1179−1187.

[18] Pan G, Sun L, Wu ZH, Lao SH. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: Proc. of the 11th IEEE Int'l Conf. on Computer Vision (ICCV 2007). Rio de Janeiro: IEEE, 2007. 1−8.

[19] Sun L, Pan G, Wu ZH, Lao SH. Blinking-based live face detection using conditional random fields. In: Proc. of the Int'l Conf. on Biometrics (ICB 2007). Seoul: Springer, 2007. 252−260.

[20] Kollreider K, Fronthaler H, Faraj MI, Bigun J. Real-time face detection and motion analysis with application in "liveness" assessment. IEEE Trans. on Information Forensics and Security (TIFS 2007), 2007, 2(3): 548−558.

[21] Peixoto B, Michelassi C, Rocha A. Face liveness detection under bad illumination conditions. In: Proc. of the IEEE Int'l Conf. on Image Processing (ICIP 2011). Brussels: IEEE, 2011. 3557−3560.

[22] Komulainen J, Hadid A, Pietikäinen M. Context based face anti-spoofing. In: Proc. of the 6th IEEE Int'l Conf. on Biometrics: Theory, Applications and Systems. Madrid: IEEE, 2013. 1−8.

[23] Määttä J, Hadid A, Pietikäinen M. Face spoofing detection from single images using texture and local shape analysis. IET Biometrics, 2012, 1(1): 3−10.

[24] Boulkenafet Z, Komulainen J, Hadid A. Face spoofing detection using colour texture analysis. IEEE Trans. on Information Forensics and Security (TIFS 2016), 2016, 11(8): 1818−1830.

[25] De Freitas Pereira T, Anjos A, De Martino JM, Marcel S. Can face anti-spoofing countermeasures work in a real world scenario? In: Proc. of the Int'l Conf. on Biometrics (ICB 2013). Madrid: IEEE, 2013. 1−8.

[26] Tirunagari S, Poh N, Windridge D, Iorliam A, Suki N, Ho ATS. Detection of face spoofing using visual dynamics. IEEE Trans. on Information Forensics and Security (TIFS 2015), 2015, 10(4): 762−777.

[27] Chan PPK, Liu WW, Chen DN, Yeung DS, Zhang F, Wang XZ, Hsu CC. Face liveness detection using a flash against 2D spoofing attack. IEEE Trans. on Information Forensics and Security (TIFS 2017), 2017, 13(2): 521−534.

[28] Arashloo SR, Kittler J, Christmas W. Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical image features. IEEE Trans. on Information Forensics and Security (TIFS 2015), 2015, 10(11): 2396−2407.

[29] Boulkenafet Z, Komulainen J, Hadid A. On the generalization of color texture-based face anti-spoofing. Image and Vision Computing, 2018, 77: 1−9.

[30] Boulkenafet Z, Komulainen J, Hadid A. Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters, 2016, 24(2): 141−145.

[31] De Freitas Pereira T, Anjos A, De Martino JM, Marcel S. LBP—TOP based countermeasure against face spoofing attacks. In: Proc. of the Asian Conf. on Computer Vision. Daejeon: Springer, 2012. 121−132.

[32] Komulainen J, Hadid A, Pietikäinen M. Face spoofing detection using dynamic texture. In: Proc. of the Asian Conf. on Computer Vision. Daejeon: Springer, 2012. 146−157.

[33] Määttä J, Hadid A, Pietikäinen M. Face spoofing detection from single images using micro-texture analysis. In: Proc. of the Int'l Joint Conf. on Biometrics (IJCB 2011). Washington: IEEE, 2011. 1−7.

[34] Yang JW, Lei Z, Yi D, Li SZ. Person-specific face antispoofing with subject domain adaptation. IEEE Trans. on Information Forensics and Security, 2015, 10(4): 797−809.

[35] Galbally J, Marcel S. Face anti-spoofing based on general image quality assessment. In: Proc. of the Int'l Conf. on Pattern Recognition (ICPR 2014). Stockholm: IEEE, 2014. 1173−1178.

[36] Galbally J, Marcel S, Fierrez J. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. IEEE Trans. on Image Processing (TIP 2014), 2014, 23(2): 710−724.

[37] Di W, Hu H, Jain AK. Face spoof detection with image distortion analysis. IEEE Trans. on Information Forensics and Security (TIFS 2015), 2015, 10(4): 746−761.

[38] De Freitas Pereira T, Komulainen J, Anjos A, De Martino JM, Hadid A, Pietikäinen M, Marcel S. Face liveness detection using dynamic texture. EURASIP Journal on Image and Video Processing, 2014, 2014(1): 1−15.

[39] Shao R, Lan XY, Yuen PC. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. IEEE Trans. on Information Forensics and Security (TIFS 2018), 2018, 14(4): 923−938.

[40] Chingovska I, Anjos A. On the use of client identity information for face anti-spoofing. IEEE Trans. on Information Forensics and Security (TIFS 2015), 2015, 10(4): 787−796.

[41] Agarwal A, Singh R, Vatsa M. Face anti-spoofing using Haralick features. In: Proc. of the 8th IEEE Int'l Conf. on Biometrics Theory, Applications and Systems. Niagara Falls: IEEE, 2016. 1−6.

[42] Siddiqui TA, Bharadwaj S, Dhamecha TI, Agarwal A, Vatsa M, Singh R, Ratha N. Face anti-spoofing with multifeature videolet aggregation. In: Proc. of the Int'l Conf. on Pattern Recognition (ICPR 2016). Cancun: IEEE, 2016. 1035−1040.

[43] Anjos A, Marcel S. Counter-measures to photo attacks in face recognition: A public database and a baseline. In: Proc. of the Int'l Joint Conf. on Biometrics (IJCB 2011). Washington: IEEE, 2011. 1−7.

[44] Bao W, Li H, Li N, Jiang W. A liveness detection method for face recognition based on optical flow field. In: Proc. of the Int'l Conf. on Image Analysis and Signal Processing. Linhai: IEEE, 2009. 233−236.

[45] Anjos A, Chakka MM, Marcel S. Motion-based counter-measures to photo attacks in face recognition. IET Biometrics, 2014, 3(3): 147−158.

[46] Nowara EM, Sabharwal A, Veeraraghavan A. PPGSecure: Biometric presentation attack detection using photopletysmograms. In: Proc. of the IEEE Int'l Conf. on Automatic Face & Gesture Recognition. Washington: IEEE, 2017. 56−62.

[47] Liu SQ, Yuen PC, Zhang SP, Zhao GY. 3D mask face anti-spoofing with remote photoplethysmography. In: Proc. of the European Conf. on Computer Vision (ECCV 2016). Amsterdam: Springer, 2016. 85−100.

[48] Liu SQ, Lan XY, Yuen PC. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In: Proc. of the European Conf. on Computer Vision (ECCV 2018). Munich: Springer, 2018. 558−573.

[49] Hernandez-Ortega J, Fierrez J, Morales A, Tome P. Time analysis of pulse-based face anti-spoofing in visible and NIR. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2018). Salt Lake: IEEE, 2018. 544−552.

[50] Shao R, Lan XY, Yuen PC. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3D mask face anti-spoofing. In: Proc. of the IEEE Int'l Joint Conf. on Biometrics (IJCB 2017). Denver: IEEE, 2017. 748−755.

[51] Lucena O, Junior A, Moia V, Souza R, Valle E, Lotufo R. Transfer learning using convolutional neural networks for face anti-spoofing. In: Proc. of the Int'l Conf. on Image Analysis and Recognition. Montreal: Springer, 2017. 27−34.

[52] Nagpal C, Dubey SR. A performance evaluation of convolutional neural networks for face anti spoofing. In: Proc. of the Int'l Joint Conf. on Neural Networks (IJCNN 2019). Budapest: IEEE, 2019. 1−8.

[53] Guo JZ, Zhu XY, Xiao JC, Lei Z, Wan GX, Li SZ. Improving face anti-spoofing by 3D virtual synthesis. In: Proc. of the Int'l Conf. on Biometrics (ICB 2019). Crete: IEEE, 2019. 1−8.

[54] Atoum Y, Liu YJ, Jourabloo A, Liu XM. Face anti-spoofing using patch and depth-based CNNs. In: Proc. of the IEEE Int'l Joint Conf. on Biometrics (IJCB 2017). Denver: IEEE, 2017. 319−328.

[55] Yu ZT, Zhao CX, Wang ZZ, Qin YX, Su Z, Li XB, Zhou F, Zhao GY. Searching central difference convolutional networks for face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2020). Seattle: IEEE, 2020. 5295−5305.

[56] Yu ZT, Li XB, Niu XS, Shi JG, Zhao GY. Face anti-spoofing with human material perception. In: Proc. of the European Conf. on Computer Vision (ECCV 2020). Glasgow: Springer, 2020. 557−575.

[57] Wang ZZ, Zhao CX, Qin YX, Zhou QS, Qi GJ, Wan J, Lei Z. Exploiting temporal and depth information for multi-frame face anti-spoofing. arXiv: 1811.05118, 2018.

[58] Shao R, Lan XY, Li JW, Yuen PC. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach: IEEE, 2019. 10015−10023.

[59] Qin YX, Zhao CX, Zhu XY, Wang ZZ, Yu ZT, Fu TY, Zhou F, Shi JP, Lei Z. Learning meta model for zero-and few-shot face anti-spoofing. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI 2020). New York: AAAI, 2020. 11916−11923.

[60] Shao R, Lan XY, Yuen PC. Regularized fine-grained meta face anti-spoofing. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI 2020). New York: AAAI, 2020. 11974−11981.

[61] Zhang KY, Yao TP, Zhang J, Tai Y, Ding SH, Li JL, Huang FY, Song HC, Ma LZ. Face anti-spoofing via disentangled representation learning. In: Proc. of the European Conf. on Computer Vision (ECCV 2020). Glasgow: Springer, 2020. 641−657.

[62] Zhang YH, Yin ZF, Li YD, Yin GJ, Yan JJ, Shao J, Liu ZW. CelebA-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: Proc. of the European Conf. on Computer Vision (ECCV 2020). Glasgow: Springer, 2020. 70−85.

[63] Kim T, Kim Y, Kim I, Kim D. BASN: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision Workshops (ICCVW 2019). Seoul: IEEE, 2019. 494−503.

[64] Liu YJ, Jourabloo A, Liu XM. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake: IEEE, 2018. 389−398.

[65] Lin BF, Li XB, Yu ZT, Zhao GY. Face liveness detection by rPPG features and contextual patch-based CNN. In: Proc. of the Int'l Conf. on Biometric Engineering and Applications. Stockholm: ACM, 2019. 61−68.

[66] Yu ZT, Li XB, Wang PC, Zhao GY. TransRPPG: Remote photoplethysmography transformer for 3d mask face presentation attack detection. IEEE Signal Processing Letters, 2021, 28: 1290−1294.

[67] Xu ZQ, Li S, Deng WH. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In: Proc. of the IAPR Asian Conf. on Pattern Recognition. Kuala Lumpur: IEEE, 2015. 141−145.

[68] Gan JY, Li SL, Zhai YK, Liu CY. 3D convolutional neural network based on face anti-spoofing. In: Proc. of the Int'l Conf. on Multimedia and Image Processing. Wuhan: IEEE, 2017. 1−5.

[69] Li HL, He PS, Wang SQ, Rocha A, Jiang XH, Kot AC. Learning generalized deep feature representation for face anti-spoofing. IEEE Trans. on Information Forensics and Security (TIFS 2018), 2018, 13(10): 2639−2652.

[70] Lin C, Liao ZYC, Zhou P, Hu JG, Ni BB. Live face verification with multiple instantialized local homographic parameterization. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI 2018). Macao: Morgan Kaufmann Publishers, 2018. 814−820.

[71] Yang X, Luo WH, Bao LC, Gao Y, Gong DH, Zheng SB, Li ZF, Liu W. Face anti-spoofing: Model matters, so does data. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach: IEEE, 2019. 3502−3511.

[72] Cai RZ, Li HL, Wang SQ, Chen CS, Kot AC. DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing. IEEE Trans. on Information Forensics and Security (TIFS 2020), 2020, 16: 937−951.

[73] Parkin A, Grinchuk O. Recognizing multi-modal face spoofing with face recognition networks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2019). Long Beach: IEEE, 2019. 1617−1623.

[74] Shen T, Huang YY, Tong ZJ. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2019). Long Beach: IEEE, 2019. 1611−1616.

[75] Zhang P, Zou FH, Wu ZW, Dai NL, Mark S, Fu M, Zhao J, Li K. FeatherNets: Convolutional neural networks as light as feather for face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2019). Long Beach: IEEE, 2019. 1574−1583.

[76] Liu AJ, Tan ZC, Li X, Wan J, Escalera S, Guo GD, Li SZ. Static and dynamic fusion for multi-modal cross-ethnicity face anti-spoofing. arXiv: 1912.02340, 2019.

[77] Yu ZT, Qin YX, Li XB, Wang ZZ, Zhao CX, Lei Z, Zhao GY. Multi-modal face anti-spoofing based on central difference networks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2020). Seattle: IEEE, 2020. 650−651.

[78] Yang Q, Zhu X, Fwu JK, Ye Y, You GM, Zhu Y. PipeNet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2020). Seattle: IEEE, 2020. 644−645.

[79] George A, Mostaani Z, Geissenbuhler D, Nikisins O, Anjos A, Marcel S. Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans. on Information Forensics and Security (TIFS 2020), 2020, 15: 42−55.

[80] George A, Marcel S. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. IEEE Trans. on Information Forensics and Security (TIFS 2020), 2020, 16: 361−375.

[81] Pinto A, Goldenstein S, Ferreira A, Carvalho T, Pedrini H, Rocha A. Leveraging shape, reflectance and albedo from shading for face presentation attack detection. IEEE Trans. on Information Forensics and Security (TIFS 2020), 2020, 15: 3347−3358.

[82] Chen HN, Hu GS, Lei Z, Chen YW, Robertson NM, Li SZ. Attention-based two-stream convolutional networks for face spoofing detection. IEEE Trans. on Information Forensics and Security (TIFS 2020), 2020, 15: 578−593.

[83] Wang GQ, Lan CX, Han H, Shan SG, Chen XL. Multi-modal face presentation attack detection via spatial and channel attentions. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2019). Long Beach: IEEE, 2019. 1584−1590.

[84] Liu A, Tan ZC, Wan J, Liang YY, Lei Z, Guo GD, Li SZ. Face anti-spoofing via adversarial cross-modality translation. IEEE Trans. on Information Forensics and Security (TIFS 2021), 2021, 16: 2759−2772.

[85] Jourabloo A, Liu YJ, Liu XM. Face de-spoofing: Anti-spoofing via noise modeling. In: Proc. of the European Conf. on Computer Vision (ECCV 2018). Munich: Springer, 2018. 290−306.

[86] Liu YJ, Stehouwer J, Liu XM. On disentangling spoof trace for generic face anti-spoofing. In: Proc. of the European Conf. on Computer Vision (ECCV 2020). Glasgow: Springer, 2020. 406−422.

[87] Feng HC, Hong ZB, Yue HX, Chen Y, Wang KY, Han JY, Liu JT, Ding ER. Learning generalized spoof cues for face anti-spoofing. arXiv: 2005.03922, 2020.

[88] Liu YJ, Liu XM. Physics-guided spoof trace disentanglement for generic face anti-spoofing. arXiv: 2012.05185, 2020.

[89] Li HL, Li W, Cao H, Wang SQ, Huang FY, Kot AC. Unsupervised domain adaptation for face anti-spoofing. IEEE Trans. on Information Forensics and Security (TIFS 2018), 2018, 13(7): 1794−1809.

[90] Tu XG, Zhang HS, Xie M, Luo Y, Zhang YF, Ma Z. Deep transfer across domains for face antispoofing. Journal of Electronic Imaging, 2019, 28(4): 043001.

[91] Jia YP, Zhang J, Shan SG, Chen XL. Single-side domain generalization for face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2020). Seattle: IEEE, 2020. 8481−8490.

[92] Wang GQ, Han H, Shan SG, Chen XL. Improving cross-database face presentation attack detection via adversarial domain adaptation. In: Proc. of the Int'l Conf. on Biometrics (ICB 2019). Crete: IEEE, 2019. 1−8.

[93] Wang GQ, Han H, Shan SG, Chen XL. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. IEEE Trans. on Information Forensics and Security (TIFS 2021), 2021, 16: 56−69.

[94] Saha S, Xu WH, Kanakis M, Georgoulis S, Chen YH, Paudel DP, Gool LV. Domain agnostic feature learning for image and video based face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2020). Seattle: IEEE, 2020. 802−803.

[95] Wang GQ, Han H, Shan SG, Chen XL. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2020). Seattle: IEEE, 2020. 6677−6686.

[96] Liu YJ, Stehouwer J, Jourabloo A, Liu XM. Deep tree learning for zero-shot face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach: IEEE, 2019. 4680−4689.

[97] Pérez-Cabo D, Jiménez-Cabello D, Costa-Pazo A, López-Sastre RJ. Deep anomaly detection for generalized face anti-spoofing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2019). Long Beach: IEEE, 2019. 1591−1600.

[98] Chen ZH, Yao TP, Sheng KK, Ding SH, Tai Y, Li JL, Huang FY, Jin XY. Generalizable representation learning for mixture domain face anti-spoofing. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI 2021). 2021. 1132−1139.

[99] Wang JJ, Zhang JY, Bian Y, Cai YY, Wang CM, Pu SL. Self-domain adaptation for face anti-spoofing. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI 2021). 2021. 2746−2754.

[100] Jia YP, Zhang J, Shan SG, Chen XL. Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. Pattern Recognition (PR 2021), 2021, 115: Article No.107888.

[101] Qin YX, Yu ZT, Yan LB, Wang ZZ, Zhao CX, Lei Z. Meta-teacher for face anti-spoofing. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI 2021), 2021. [doi: 10.1109/TPAMI.2021.3091167]

[102] George A, Marcel S. Deep pixel-wise binary supervision for face presentation attack detection. In: Proc. of the Int'l Conf. on Biometrics (ICB). Crete: IEEE, 2019. 1−8.

[103] Wu XJ, Zhou JH, Liu J, Ni FY, Fan HQ. Single-shot face anti-spoofing for dual pixel camera. IEEE Trans. on Information Forensics and Security (TIFS 2021), 2021, 16: 1440−1451.

[104] Deb D, Jain AK. Look locally infer globally: A generalizable face anti-spoofing approach. IEEE Trans. on Information Forensics and Security (TIFS 2020), 2020, 16: 1143−1157.

[105] Chen BL, Yang WH, Li HL, Wang SQ, Kwong S. Camera invariant feature learning for generalized face anti-spoofing. IEEE Trans. on Information Forensics and Security (TIFS 2021), 2021, 16: 2477−2492.

[106] Sun WY, Song Y, Chen CS, Huang JW, Kot AC. Face spoofing detection based on local ternary label supervision in fully convolutional networks. IEEE Trans. on Information Forensics and Security (TIFS 2020), 2020, 15: 3181−3196.

[107] Lafferty J, Mccallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). 2001. 282−289.

[108] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55(1): 119−139.

[109] Kollreider K, Fronthaler H, Bigun J. Evaluating liveness by face images and the structure tensor. In: Proc. of the 4th IEEE Workshop on Automatic Identification Advanced Technologies. Buffalo: IEEE, 2005. 75−80.

[110] Turaga P, Chellappa R, Veeraraghavan A. Advances in Video-based Human Activity Analysis: Challenges and Approaches. Amsterdam: Elsevier, 2010. 237−290.

[111] Poh MZ, Mcduff DJ, Picard RW. Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. on Biomedical Engineering, 2011, 58(1): 7−11.

[112] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations (ICLR 2015). San Diego, 2015.

[113] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas: IEEE, 2016. 2818−2826.

[114] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas: IEEE, 2016. 770−778.

[115] Sharan L, Liu C, Rosenholtz R, Adelson EH. Recognizing materials using perceptually inspired features. Int'l Journal of Computer Vision, 2013, 103(3): 348−371.

[116] Zhang XE, Ng R, Chen QF. Single image reflection separation with perceptual losses. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake: IEEE, 2018. 4786−4794.

[117] Heusch G, Marcel S. Pulse-based features for face presentation attack detection. In: Proc. of the 9th IEEE Int'l Conf. on Biometrics Theory, Applications and Systems. Redondo Beach: IEEE, 2018. 1−8.

[118] Niu XS, Yu ZT, Han H, Li XB, Shan SG, Zhao GY. Video-based remote physiological measurement via cross-verified feature disentangling. In: Proc. of the European Conf. on Computer Vision (ECCV 2020). Glasgow: Springer, 2020. 295−310.

[119] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations (ICLR 2021). Virtual Event, 2021.

[120] Sun SY, Kuang ZH, Sheng L, Ouyang WL, Zhang W. Optical flow guided feature: A fast and robust motion representation for video action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu: IEEE, 2017. 1390−1399.

[121] Liu AJ, Wan J, Escalera S, Jair Escalante H, Tan ZC, Yuan Q, Wang K, Lin C, Guo GD, Guyon I, Li SZ. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2019). Long Beach: IEEE, 2019. 1601−1610.

[122] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu: IEEE, 2017. 1251−1258.

[123] Xie SN, Girshick R, Dollár P, Tu ZW, He KM. Aggregated residual transformations for deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu: IEEE, 2017. 1492−1500.

[124] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake: IEEE, 2018. 7132−7141.

[125] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proc. of the Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234−241.

[126] Zhou KY, Liu ZW, Qiao Y, Xiang T, Loy CC. Domain generalization: A survey. arXiv: 2103.02503, 2021.

[127] Yu ZT, Qin YX, Li XB, Zhao CX, Lei Z, Zhao GY. Deep learning for face anti-spoofing: A survey. arXiv: 2106. 14948, 2021.

[128] Fan FL, Xiong JJ, Li MZ, Wang G. On interpretability of artificial neural networks: A survey. IEEE Trans. on Radiation and Plasma Medical Sciences, 2021, 5(6): 741−760.

[129] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas: IEEE, 2016. 2574−2582.

[130] Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. Ruan Jian Xue Bao/Journal of Software, 2020, 31(1): 67−81 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]

[131] Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Machine Learning, 2010, 79(1): 151−175.

[132] Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. In: Advances in Neural Information Processing Systems (NIPS 2007). Vancouver: MIT, 2007. 137−144.

[133] Prakash A, Boochoon S, Brophy M, Acuna D, Cameracci E, State G, Shapira O, Birchfield S. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In: Proc. of the Int'l Conf. on Robotics and Automation. Montreal: IEEE, 2019. 7249−7255.

[134] Volpi R, Namkoong H, Sener O, Duchi J, Murino V, Savarese S. Generalizing to unseen domains via adversarial data augmentation. In: Advances in Neural Information Processing Systems (NIPS 2018). Montreal: MIT, 2018. 5339−5349.

[135] Zhou KY, Yang YX, Hospedales T, Xiang T. Deep domain-adversarial image generation for domain generalisation. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI 2020). New York: AAAI, 2020. 13025−13032.

[136] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI 2013), 2013, 35(8): 1798−1828.

[137] Hospedales T, Antoniou A, Micaelli P, Storkey A. Meta-learning in neural networks: A survey. arXiv: 2004.05439, 2020.

[138] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. In: Advances in Neural Information Processing Systems (NIPS 2018). Montreal: MIT, 2018. 5339−5349.

[139] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: Proc. of the Int'l Conf. on Machine Learning (ICML 2015). Lille: ACM, 2015. 1180−1189.

[140] Gretton A, Borgwardt KM, Rasch MJ, Schlkopf B, Smola A. A kernel two-sample test. Journal of Machine Learning Research (JMLR 2012), 2012, 13: 723−773.

[141] Xie QZ, Dai ZH, Du YL, Hovy E, Neubig G. Controllable invariance through adversarial feature learning. In: Advances in Neural Information Processing Systems (NIPS 2017). Long Beach: MIT, 2017. 585−596.

[142] Akuzawa K, Iwasawa Y, Matsuo Y. Adversarial invariant feature learning with accuracy constraint for domain generalization. In: Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Würzburg: Springer, 2019. 315−331.

[143] Motiian S, Piccirilli M, Adjeroh DA, Doretto G. Unified deep supervised domain adaptation and generalization. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV 2017). Venice: IEEE, 2017. 5715−5725.

[144] Chen WF, Fu Z, Yang DW, Deng J. Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems (NIPS 2016). Barcelona: MIT, 2016. 730−738.

[145] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV 2017). Venice: IEEE, 2017. 618−626.

[146] Boulkenafet Z, Komulainen J, Li L, Feng XY, Hadid A. OULU-NPU: A mobile face presentation attack database with real-world variations. In: Proc. of the IEEE Int'l Conf. on Automatic Face & Gesture Recognition. Washington: IEEE Press, 2017. 612−618.

[147] Bengio S, Mariéthoz J. A statistical significance test for person authentication. In: Proc. of the Odyssey 2004: The Speaker and Language Recognition Workshop. 2004. 237−244.

[148] Li HL, Pan SJ, Wang SQ, Kot AC. Domain generalization with adversarial feature learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake: IEEE, 2018. 5400−5409.

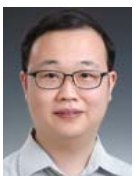
[149] Li Y, Song LX, Wu X, He R, Tan TN. Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI 2018). Louisiana: AAAI, 2018. 7057−7064.

[150] Zoph B, Le QV. Neural architecture search with reinforcement learning. In: Proc. of the 5th Int'l Conf. on Learning Representations (ICLR 2017). Toulon, 2017.

[151] Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake: IEEE, 2018. 8697−8710.

附中文参考文献:

[130] 潘文雯, 王新宇, 宋明黎, 陈纯. 对抗样本生成技术综述. 软件学报, 2020, 31(1): 67−81. http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]



张帆(1977−), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为信息系统安全, 机器学习及其安全.

陈伟(1979−), 男, 博士, 教授, CCF 专业会员, 主要研究领域为网络安全, 机器学习.

赵世坤(1997−), 男, 硕士生, CCF 学生会员, 主要研究领域为机器学习及安全.

刘小丽(1981−), 女, 博士, 讲师, 主要研究领域为信息系统安全, 形式化.

袁操(1980−), 男, 博士, 副教授, 主要研究领域为机器学习.

赵涵捷(1963−), 男, 博士, 教授, 主要研究领域为移动计算, 云计算, 物联网, 量子计算, 网络及信息安全.