**第十四届东南大学信息安全竞赛**

**作品报告**

**作品名称：**双重博弈：面向对抗样本演示攻击的生物指纹识别系统

**英文名称：** Liveness is not enough: when PADs meet PAEs

**题目类型：** 软件类

**指导老师：** 孙晨

**项目成员：** 朱鸿宇，胡文韬，周雅洁，赵思宇

**电子邮箱：** 213200421@seu.edu.cn

**提交日期：** 2023-4-30

目录

[摘要 1](#_Toc23734)

[第一章 研究背景及意义 2](#_Toc17969)

[1.1 指纹识别技术 2](#_Toc2475)

[1.1.1 生物认证技术 2](#_Toc8431)

[1.1.2 生物识别技术市场调研 3](#_Toc20923)

[1.1.3 指纹采集技术 4](#_Toc18562)

[1.1.4 指纹细节匹配 6](#_Toc16855)

[1.2 生物指纹识别技术的安全现状 7](#_Toc4426)

[1.2.1生物密码安全现状 7](#_Toc13133)

[1.2.2 指纹识别技术安全现状 8](#_Toc5150)

[1.2.2.1 对指纹的物理攻击 9](#_Toc30034)

[1.2.2.2 传统活体检测方案 9](#_Toc32305)

[1.2.2.3 基于深度神经网络的活体检测方法 10](#_Toc30674)

[1.3 对抗攻击研究现状 12](#_Toc16672)

[1.3.1 演示攻击 12](#_Toc10711)

[1.3.2 对抗攻击 13](#_Toc11330)

[1.3.3 物理可制作对抗样本 15](#_Toc4096)

[1.3.4 对抗样本防御现状 16](#_Toc441)

[1.3.4.1 对抗训练 17](#_Toc31714)

[1.3.4.2 防御式扰动 18](#_Toc27865)

[1.3.4.3 检测型防御 19](#_Toc21325)

[1.3.4.4 模型修复 19](#_Toc11477)

[1.3.5 防御挑战 20](#_Toc3779)

[1.4 研究意义 21](#_Toc7641)

[第二章 作品设计与实现 21](#_Toc4263)

[2.1 作品简介 21](#_Toc23547)

[2.1.1 设计思路 22](#_Toc17137)

[2.1.2 框架描述 24](#_Toc22185)

[2.2 决策树集成模型表示学习 25](#_Toc5990)

[2.2.1 深度模型架构 25](#_Toc8309)

[2.2.2 自动数据增强 28](#_Toc9890)

[2.2.3 CNN对树集成模型的知识蒸馏 30](#_Toc26236)

[2.3 对抗样本攻击方案 31](#_Toc19583)

[2.4 防御方案 33](#_Toc5042)

[2.4.1 可证明鲁棒决策树 33](#_Toc21656)

[2.4.2 集成学习 33](#_Toc2006)

[2.4.3 知识蒸馏 38](#_Toc8186)

[2.4.4 对抗训练 39](#_Toc18152)

[第三章 作品测试与分析 40](#_Toc13630)

[3.1 数据集简介 40](#_Toc27980)

[3.1.1 数据集内容 40](#_Toc30374)

[3.1.2 数据集获取过程 40](#_Toc22013)

[3.2 模型训练过程 42](#_Toc22003)

[3.2.1 模型参数设置 43](#_Toc2955)

[3.2.2 训练设置 43](#_Toc23563)

[3.2.3 测试设置 43](#_Toc8616)

[3.3 物理对抗样本制作 43](#_Toc19110)

[3.3.1 传感器选型 43](#_Toc17563)

[3.3.2 对抗样本制作材料与工具（暂无） 44](#_Toc4256)

[3.3.3 对抗样本制作过程（暂无） 44](#_Toc29908)

[3.4 防御效果测试 44](#_Toc19875)

[第四章 界面展示 45](#_Toc22629)

[第五章 创新性分析 45](#_Toc18494)

[第六章 总结与展望 46](#_Toc6469)

# 摘要

指纹识别由于技术成熟、成本低廉、操作方便，即便生物特征识别技术层出不穷，仍然在身份认证、取证等领域具有广泛应用，在个人、家庭和一般企业访问控制设备中占有最大的市场。但是，2016年的ISO生物特征识别标准（ISO/IEC 30107）指出，演示攻击（Presentation Attack，PA攻击者欺骗指纹识别传感器以冒充他人或逃避识别，如使用硅胶等材料制作假指纹通过身份认证）构成了指纹识别的严重威胁，合格的设备需要能够区分和拒绝来自演示攻击的假指纹。从此，针对演示攻击的防御成为研究热点，许多安全顶级会议和期刊（如USENIX Security、TIFS）都发表了关于指纹演示攻击检测（Presentation Attack Detection，PAD）的文章，并成立了两年举办一次的活体检测比赛LivDet。时至今日，指纹PAD领域的最新论文大都声称自己取得了极高的检测准确率，如97%以上。然而，指纹识别真的摆脱了安全威胁吗？

安全领域的攻防仿佛矛与盾的竞争，一方的进步往往唤起另一方的发展。2014年，针对深度学习的对抗样本被提出，人们发现，统计学习模型在高维输入空间中具有内在脆弱性，输入的微小变化（如向输入图像添加肉眼不可见的特制噪声扰动）可以完全颠覆算法输出。一旦信息系统部署了检测模型作为安全组件，检测模型所用算法也暴露成为了攻击面的一部分，指纹识别也不例外。基于对抗样本的演示攻击可以轻易绕过一般情况下准确的检测系统，同时因其卓越的隐蔽性（肉眼不可见）而不会产生任何异常。新一代对抗性演示攻击为指纹识别系统带来了极大威胁，同时，文献中广泛使用的深度神经网络检测模型由于庞大的计算开销难以部署于现实设备，无论可靠性还是可用性，现存演示攻击检测系统远远未达到要求。目前，指纹领域对于对抗性演示攻击（通过带有对抗扰动的假指纹样本绕过活体检测从而入侵系统）的防御研究仍然处于空白，我们将尝试弥补这一攻击面，同时设计更加轻量化、低成本的检测算法，构建可信、可靠、可用的强安全生物指纹识别系统。

关键词：指纹识别，生物指纹活体检测，对抗样本防御

# Abstract

Fingerprint recognition, due to its mature technology, low cost, and ease of operation, still has a wide range of applications in identity authentication, forensics, and other fields, and holds the largest market share in personal, household, and general enterprise access control devices, despite the emergence of various biometric recognition technologies. However, the ISO biometric recognition standard (ISO/IEC 30107) in 2016 pointed out that presentation attacks (PA attacks) pose a serious threat to fingerprint recognition, where attackers deceive the fingerprint recognition sensor to impersonate someone else or evade recognition by using materials such as silicone to create fake fingerprints for identity authentication. Qualified devices need to be able to distinguish and reject fake fingerprints from presentation attacks. Since then, defense against presentation attacks has become a research hot topic, and many top security conferences and journals (such as USENIX Security and TIFS) have published articles on fingerprint presentation attack detection (PAD) and established the LivDet, a biennial competition on liveness detection. To date, the latest papers in the fingerprint PAD field claim to have achieved extremely high detection accuracy, such as over 97%. However, has fingerprint recognition really overcome security threats?

The competition between attack and defense in the security field seems like a competition between a spear and a shield, where the progress of one often sparks the development of the other. In 2014, adversarial examples against deep learning were proposed, and people discovered that statistical learning models have inherent fragility in high-dimensional input spaces, where small changes in inputs (such as adding specially crafted noise to input images that are invisible to the human eye) can completely overturn the algorithm output. Once an information system deploys a detection model as a security component, the algorithm used by the detection model also becomes a part of the attack surface, and fingerprint recognition is no exception. Presentation attacks based on adversarial examples can easily bypass accurate detection systems under normal circumstances, while also remaining completely concealed due to their excellent stealth (invisible to the human eye). The new generation of adversarial presentation attacks poses a great threat to fingerprint recognition systems. Meanwhile, deep neural network detection models widely used in the literature are difficult to deploy on real devices due to their massive computational overheads. Existing presentation attack detection systems are far from meeting requirements in terms of reliability and availability. Currently, research on defense against adversarial presentation attacks (in which attackers invade systems by bypassing liveness detection through fake fingerprints samples with adversarial perturbations) in the fingerprint field is still blank, and we will attempt to fill this attack surface while designing more lightweight and low-cost detection algorithms to construct a trusted, reliable, and available strong security biometric fingerprint recognition system.

Keywords: fingerprint recognition, biometric liveness detection, defenses against adversarial example.

# 第一章 研究背景及意义

## 1.1 指纹识别技术

### 1.1.1 生物认证技术

认证技术通过检验消息传送过程中的某些参数防止伪造、篡改、冒名顶替等攻击，广泛应用于各类计算机通信网、办公自动化及各类密码管理系统中。

生物识别技术利用人体固有生理特性和行为特征进行身份验证，其通常具有可以测量或可自动识别和验证、遗传性或终身不变等特点，因此生物识别认证技术较传统认证技术存在较大的优势。

生物识别系统对生物特征进行取样，提取其特征并且转化成[数字](https://baike.baidu.com/item/%E6%95%B0%E5%AD%97" \t "_blank)[代码](https://baike.baidu.com/item/%E4%BB%A3%E7%A0%81" \t "_blank)，并进一步将这些代码组成特征模板。由于微处理器及各种电子元器件成本不断下降，精度逐渐提高，生物识别系统逐渐应用于商业上的授权控制如门禁、企业考勤管理系统安全认证等领域。用于生物识别的生物特征有手形、指纹、脸形、虹膜、视网膜、脉搏、耳廓等，行为特征有签字、声音、按键力度等。基于这些特征，人们已经发展了指纹识别、面部识别、虹膜识别、发音识别、手形识别等多种生物识别技术，具体如下：

(1)指纹识别，有多种实现方法，如匹配指纹局部细节，提取指纹的波纹边缘模式等。在所有生物识别技术中，指纹识别是当前应用最为广泛的一种。

(2)[人脸识别](https://baike.baidu.com/item/%E4%BA%BA%E8%84%B8%E8%AF%86%E5%88%AB" \t "_blank)，运用机器学习图像分类模型进行识别，具体流程包括人物图像采集、人物定位、人物识别预处理、身份确认以及身份查找等。

(3)虹膜识别，准确率高，比其余基于眼部特征的识别技术对人体产生干扰少，模板匹配性能高。

(4)声音识别，现今虽已有产品进入市场，但使用不方便，主要因为传感器和人的声音可变性都很大，且使用步骤复杂。

(5)手掌几何学识别，通过测量使用者的手掌和手指的物理特征来进行识别，甚至建模三维图像。性能好且使用比较方便，适用于人数较多场合，且容易集成到其他系统中。

表1-1 几种生物识别技术的对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 技术类别 | 人脸识别 | 指纹识别 | 虹膜识别 | 声纹识别 | 静脉识别 |
| 误识率 | 0.50% | 0.80% | 0.0001%，双眼识别误识率更高 | 1.00% | 0.00% |
| 稳定性 | 一般 | 易磨损 | 终生不变 | 不稳定 | 终生不变 |
| 应用场景 | 人脸联网核查、解锁、支付、智慧城市、智能家居等。 | 解锁、支付、智慧城市、智能家居、登录等。 | 银行金库加密、监控、支付、军队国防等。 | 社保、公安司法、刑侦、手机锁屏等。 | 机场身份认证安全管理、网络支付等。 |
| 便利程度 | 应用最广泛，识别距离不限，采集方便成本低。 | 接触式识别，门槛最低，成本低廉，使用便利。 | 识别速度快，识别距离为30cm。 | 获取方便自然，接受程度高。 | 两秒识别，采集距离有限，设备昂贵。 |
| 发展现状 | 光线等诸多条件影响结果，人脸识别不能应用在所有的银行场景中，高安全级别的应用场景需更谨慎。 | 指纹识别技术目前在考勤门禁银行、手机等领域已经得到应用，手机搭载指纹识别成为标配。 | 技术难度高，应用价格高，国内行业混乱，目前大多局限于工业领域的应用。 | 算法复杂度低，探索阶段，技术尚未成熟，易受身体年龄、情绪噪音等影响。 | 日韩1997年推出静脉识别产品，国内2004年开始研究，产品化程度差。 |
| 发展机构 | 支付宝、汉王、各大银行、天成盛业等。 | FPC、欧菲光、高通、天成盛业、汇顶、各大手机厂商等。 | Iridian、LG、松下、三呈、释码大华、中科虹霸等。 | 科大讯飞、联想、支付宝。 | 燕南科技、欧计斯软件、汉王等。 |
| 趋势 | 将推动在线金融业务的发展，是人工服务升级为智能自动服务的有效途径，未来将成为互联网金融领域的“基础设施”。 | 随着移动支付市场需求不断上升，指纹识别设备成本不断下降，指纹识别技术还将有爆发式发展。 | 中国移动购物市场规模已达万亿级别，一旦指纹无法满足安全级别，虹膜支付将成为未来移动支付主角。 | 目前声纹识别的全球市占率为15.8%，仅次于指纹和掌纹的生物特征识别，并由不断上升趋势。 | 目前市占率为2.4%，在网络支付、社保管理等领域，静脉识别将发挥革新性作用。 |

### 1.1.2 生物识别技术市场调研

全球市场对生物识别产品的需求在2010年达到71亿美元，生物识别设备的综合性年增长率达到21.3%。指纹识别技术是应用最多，也是应用最早的生物识别技术，此项识别技术是生物识别技术收入的主要贡献者，产值年增长率预计为16%。增长的主要原因是指纹识别设备价格的下降以及政府部门对指纹识别设备的推动和依赖[1][2]。另外，人脸识别、掌型识别以及其它生物识别技术如虹膜识别、声音识别、笔迹识别等的市场规模也有所增长。

中国生物特征识别行业最早发展的是指纹识别技术，基本与国外同步，早在80年代初就开始了研究，并掌握了核心技术，产业发展相对比较成熟。而中国对于静脉识别、人脸识别、虹膜识别等生物认证技术研究的开展则在1996年之后。

近年来，指纹传感器已被广泛集成到智能手机和平板电脑中。结合快速识别在线（FIDO）和其他协议，指纹传感器使移动银行等应用程序能够在不使用短密码的情形下本地认证终端用户。据估计，2017年全球配备指纹识别的智能手机出货量便已达到9.2亿部（约64%）。

目前在众多生物识别技术当中，市面上应用领域最为广泛的技术是指纹识别，其次是人脸识别，语音识别、虹膜识别、静脉识别等识别技术则使用较少。根据统计，指纹识别的市场占比为58%，而人脸识别占比18%，其次则是虹膜识别、掌纹识别、语音识别分别占比7%，7%，5%。

未来指纹识别的应用大方向，主要分为四个：一是涉密系统，即需要高度安全防护与保护的地方；二是针对大规模人群的身份鉴别的技术；三是针对城市公共事务的市民应用；四是从现实生活进入互联网虚拟世界的认证应用。

随着近几年生物特征识别技术场景的大量挖掘，指纹识别技术还会结合人脸识别、虹膜识别技术等构成集成式的生物特征识别技术，在金融、银行、安防等领域产生巨大的影响力。

### 1.1.3 指纹采集技术

指纹识别系统一般包括四个部分：指纹图像采集、指纹图像预处理、指纹图像特征提取，以及指纹图像识别。

指纹图像的采集是指纹识别系统中的重要组成部分，采集到的指纹图像的质量好坏直接影响到后续图像的处理。低质量的指纹图像会致使系统拒真，导致系统性能下降，而高质量的指纹图像可以简化指纹图像的处理算法，提高识别率，减小拒识率。

目前的指纹采集器主要使用三种采集技术：光学全反射技术、硅晶体电容传感器技术、超声波扫描技术。

借助光学技术采集指纹是历史最久远、使用最广泛的技术。将手指放在光学镜片上，手指在内置光源照射下，用棱镜将其投射在电荷耦合器件(CCD)上，进而形成脊线，即指纹图像中具有一定宽度和走向的纹线，呈黑色；谷线，即纹线之间的凹陷部分，呈白色，是数字化的、可被指纹设备算法处理的多灰度指纹图像。

20世纪90年代后期，基于半导体硅电容效应的技术趋于成熟。硅传感器成为电容的一个极板，手指则是另一极板，利用手指纹线的脊和谷相对于平滑的硅传感器之间的电容差，形成8bit的灰度图像。电容传感器发出电子信号，电子信号将穿过手指的表面和死性皮肤层，直达手指皮肤的活体层(真皮层)，直接读取指纹图案。由于深入真皮层，传感器能够捕获更多真实数据，不易受手指表面尘污的影响，提高辨识准确率，有效防止辨识错误。

超声波指纹采集是一种新型技术，其原理是利用超声波具有穿透材料的能力，且随材料的不同产生大小不同的回波，即超声波到达不同材质表面时，被吸收、穿透与反射的程度不同。因此，利用皮肤与空气对于声波阻抗的差异，就可以区分指纹嵴与峪所在的位置。

各种技术都具有各自的优势与缺点，本项目在下面给出这三种主要技术的比较：

表1-2 三种指纹技术的对比

|  |  |  |  |
| --- | --- | --- | --- |
| 比较项目 | 光学全反射技术 | 硅晶体电容传感技术 | 超声波扫描技术 |
| 体积 | 大 | 小 | 中 |
| 耐用性 | 非常耐用 | 容易损坏 | 一般 |
| 成像能力 | 干手指交叉，汗多和稍脏的手指成像模糊 | 干手指好，汗多和稍脏的手指不能成像 | 非常好 |
| 耗电 | 较多 | 较少 | 较多 |
| 成本 | 低 | 低 | 很高 |

### 1.1.4 指纹细节匹配

指纹识别技术的传统方法是获得指纹图像，然后对图像进行处理，对里面的各类特征进行提取，将提取出来的指纹特征和预先保存的指纹数据进行比较。

除了基于DNN模型的端到端训练与认证，基于传统ML的提取特征再训练的指纹识别方案，指纹识别另一个流行方案是指纹匹配（fingerprint matching），通过分析指纹的局部特征，从中抽取详尽的特征点，以此来确定两枚指纹是否来自同一枚手指，从而较为可靠地进行身份鉴定。

本项目曾将指纹匹配技术融入CNN与决策树集成模型，并尝试使用细节提取（Minutiae Extrating）改进模型准确率，但实验效果并不理想，故作为研究背景在此略述：

虽然指纹识别已经研究多年，已有非常有效的解决方案，但指纹识别不能被认为是一个完全解决的问题，设计精确、互操作和计算光算法仍相当困难。

由于指纹的旋转、平移和变形、虚假细节的存在和缺乏真实的细节等因素，细节对应是非常模糊的。研究人员开发了各种局部特征（如图1），如脊（ridge）、方向（orientation）、细节（minutiae）等特征，以减少这种模糊性。

图1 指纹特征示例

在各种指纹匹配算法中，基于细节的匹配算法曾经是最流行的方法，因为人们普遍认为细节是最具鉴别性和最可靠的特征，这基于一种理想：虽然不同指纹间大的轮廓特征可能相同，但不同人指纹之间的细节不尽相同，如可能体现在不同的指纹纹路上：中间可能是连续的、平滑笔直的，也可能会有断裂、分支或者不规则的折线。这些与众不同的被认为十分重要的纹路，被定义为断点、分叉点和转折点等细节特征点，并利用局部方向曲率位置参数等组合构成唯一身份信息。

传统的细节匹配算法倾向于假设每个细节具有相同的可区分性，但这一假设至少受到了两个事实的挑战：其一是指纹的细节往往会形成簇，使得相似区域的细节具有较大的局部相似性，导致偏高的匹配细节对数，从而方便了冒名匹配。

此外，由于指纹图像质量较低，容易提取出虚假的细节，导致错误接受率高，错误拒绝率高。在细节提取过程中，由于干、湿手指等指纹图像质量较低，可能会提取出假细节，但删除假细节非常困难，因为真正有效的细节可能同时被删除。考虑到细节匹配算法倾向于认为每个细节对指纹匹配得分的贡献相等，故假细节的存在将十分有利于冒名匹配。

## 1.2 生物指纹识别技术的安全现状

### 1.2.1生物密码安全现状

现在的采用生物密码支付已成普遍应用，但生物密码安全吗？你知道你的生物信息到底放在了哪里吗？这是个重大而且敏感的问题。

理论上指纹、虹膜、人脸、声纹这些生物信息难以复制，但一旦经过采样数字化后便极易复制，故手机通过传感器采集生物特征并编码成生物密码的过程应该在一个黑盒子中封闭处理，以防止被木马程序劫持。这就是对生物密码处理的 “双系统原则”。

由此可见，生物识别技术现在依然存在风险，具体可分为：

(1)传输风险。即主动篡改与窃听风险，从这个角度看，生物指纹与普通口令在安全性上并无区别。

(2)不可撤销性。也是生物识别最大的弊端。钥匙被偷可以换锁，密码泄漏可以换密码，但指纹、虹膜、脸型扫描被窃取便无法撤销，甚至被转售。

(3)准确性和唯一性。目前，很多的生物识别应用都称准确性达到新高度，但从技术上看，一些人脸识别算法只是提取了人脸各处的高度，眉毛宽度，间距，眼睛定位，间距等信息进行三维建模。指纹识别算法即是取指纹中多个点的脊线和谷线，以及走向，经过方向图计算、图像增强、二值化和细化等步骤进行建模，可以高概率识别其实并不意味着伪造难度大。

(4)生物标识的乱用。一些企业除使用生物识别作为身份认证外，也把它当成了身份管理标识，其风险在于，如果把身份认证比喻为密码的话，身份管理就是身份证和银行卡，如果混淆了二者，就相当把身份证，银行卡和密码同时提供给对方。

(5)对生物识别的利用仍没有完善的技术标准和法规。例如企业是否可以存储用户的生物识别的特征。2014年3月22日，乌云平台披露了“携程网安全支付日志可被遍历下载，导致大量用户银行卡信息泄露，并称该漏洞已经过携程确认”。众所周知，得到信用卡卡号，有效期，CVV码，无需密码就可以完成支付。

在这些方面，生物识别技术仍存在许多未知的风险。

### 1.2.2 指纹识别技术安全现状

近年来，指纹传感器已被广泛集成到智能手机和平板电脑中，结合快速在线识别（FIDO）等协议，为移动银行等APP在本地认证终端用户提供技术支持。生物特征识别已广泛应用于安全领域，如身份认证、刑事调查、移动支付和访问控制等领域。在各种生物特征中，指纹因其普遍性、独特性和持久性，以及成熟轻便的技术，已占据生理特征身份验证市场的主导地位（58%的市场占有率），但它是否真的安全有待商榷，就像任何系统一样，指纹识别系统也面临着各种恶意攻击。

随着机器学习模型在生物识别领域的广泛部署应用，一种新型强大攻击，模型攻击——识别系统本身暴露为攻击面，将大范围引起安全威胁，其中对抗样本可以快速低成本制作，引导模型做出特定错误预测以达成目标，攻击成功率极高，造成损失极大，难以发现，难以有效防御。

另一方面，早在2001年便有研究表明，生物识别系统易受由PlayDoh、硅胶或乳胶等材料制成的人工指纹的演示攻击，其针对生物识别系统本身漏洞设计，可将仿制指纹识别为生物指纹，将非法生物指纹绑定到合法身份，或使合法指纹无法通过验证。在演示攻击活体检测算法相对成熟的今天，对抗样本的发现和使用将促进演示攻击的进化，并将给生物识别领域造成灾难。

#### 1.2.2.1 对指纹的物理攻击

* 破坏攻击

传统对指纹锁的攻击集中于破坏指纹锁的结构以及试图破解指纹密码。针对这一问题，现在市面上已有检测被动开锁的设备，其原理本质和上述指纹识别没有区别。采用验证指纹和指纹库是否匹配，在匹配不成功次数过多和强制撬锁开锁的情况下，利用指纹锁对用户手机等设备进行报警，在一定程度上提高了指纹锁的安全性。

* 仿制指纹攻击

指纹具有独一无二的特性，若非植皮或深达基底层的损伤，指纹通常不会改变，这样看来指纹识别似乎十分安全。但实际并非如此，一般来说，仿制指纹攻击的方法有以下几种：

(1)胶状手指：即被捏造仿造的手指、人造手指状物体，其可以精确模仿另一个人的指纹脊形谷结构；

(2) 2D或3D打印的指纹对象：即利用2D或3D打印指纹目标区域，以达到欺骗攻击的目的；

(3)篡改指纹：即故意篡改或损坏真实指纹图案以避免识别；

其中，第一种攻击形式最为常见，仿制指纹攻击可以使用多种制造工艺来实现，从基本的成型和铸造到利用复杂的2D和3D打印技术。明胶、橡皮泥、木胶和建筑用硅胶等容易获得且价格低廉的材料已被用于制造能够绕过指纹识别系统的高保真指纹欺骗。

例如，在2013年3月，一名医生使用硅胶做的手指来欺骗攻击医院生物识别系统，并因此被捕；2016年7月，研究人员使用2D打印指纹解锁了一款指纹安全智能手机来帮助警方侦察一起凶杀案。

此外，传统方式采集的指纹只有外部信息，很容易受到诸如灰尘、汗水、疤痕、伤口等环境的影响。加之随着人的年龄增长，皮肤褶皱逐渐增多，会导致指纹质量及识别精度较差。

#### 1.2.2.2 传统活体检测方案

ISO/IEC 30107标准定义了PAD（演示攻击检测），指的是所有旨在识别生物特征中错误或改变的特征的方法[3]。对应到指纹领域，演示攻击包括伪造手指、2D或3D打印指纹等（以下统称为假指纹），这两种方式是较为常见且最容易实现的。同时，由于攻击材料的多样性，指纹识别系统受到攻击的可能性大大增加。

在指纹识别系统中配备指纹演示攻击检测（PAD）方法可以在一定程度上有效抵御这种攻击。理想的PAD应当满足以下要求：非侵入性、用户友好、低成本、高效率和低错误率。PAD有两种实现模式，基于硬件和基于软件，下面将分别进行介绍。

基于硬件的PAD通常使用能够捕获指纹的额外特征的传感器，如电学特性、皮肤弹性、血流、脉搏等，这有助于区分真假指纹。指纹识别系统分析这些额外的特征，并将其与真指纹的特征进行比较，以确定是否发生了演示攻击。但是这种方式需要将传感器与指纹识别系统进行整合，成本较高且设备更新难度较大。

其他一些常见的传感器还有热传感器（检测指纹脊和谷之间的温度差异）、磁性传感器、多光谱传感器（使用多种波长来检测真假指纹的光学特性差异）[5]、超声波传感器[7]等。

基于软件的PAD依靠分析捕获的指纹图像的特征的算法来检测演示攻击。这些算法可以寻找图像中表明假指纹的特定模式或特征，例如表明使用了人造材料或缺乏自然特征的模式。

基于软件的PAD方法分为两类：动态方法和静态方法[8]。动态方法基于手指的一些动态信息，如出汗和皮肤变形等，需要连续采集指纹图像。静态方法基于手工设计特征或机器学习特征。早期基于软件的方法依赖于从指纹图像中提取的手工设计特征，将它们分类为真实或欺骗[4]。2015年之前的PAD方法主要利用手工设计的特征，如生理特征和纹理特征等[8]。2016年以来，基于软件的PAD方法研究聚焦于深度卷积神经网络（CNN），并且已被证明明显优于以前的方法[6][9]。

基于软件的PAD不需要任何额外的传感器，设备更新方便(通过更新软件以提升检测能力)且成本较低，得到了广泛的应用。

#### 1.2.2.3 基于深度神经网络的活体检测方法

自从Alexnet被提出[10]，深度卷积神经网络（CNN）在计算机视觉领域的分类问题中表现出了出色的性能。因此，研究人员也逐渐将CNN应用于活体检测领域。Arora等人使用VGG架构作为分类器，对经过对比度增强的指纹样本进行分类，奠定了深度学习指纹活体检测的基础[11]。同年，Nogueira等人使用预训练的VGG网络配合数据增强方法，并在Livdet 2009、2011和2013数据库上与使用LBP等手工特征的SVM对比，呈现出预训练深度网络在活体检测任务上的极大优越性。此外，Uliyan等人提出了基于DBM（Deep Boltzmann Machine）和RBM（Restricted Boltzmann Machine）的方法，用于提取和分析特征之间的关系，同样优于手工特征提取的方法[12]。不过，DBM逐渐被历史淘汰，随着硬件算力的增长，直接训练多层网络不再困难，后来的研究普遍集中于结构更加简洁清晰但富有潜力的深度卷积网络。

就准确率而言，采用深度学习技术的指纹活体检测取得了长足进展。然而，这些方法在面对使用未知材料制作的假指纹图像时性能较差，即由于概念漂移和域外泛化的困难性而难以应对零日材料攻击。当然，这是迄今为止绝大多数活体检测方法的通病，而非深度网络的弱点。为了提高深度网络对于未知攻击材料的泛化能力，研究者分别从指纹本身特性与深度学习技术入手，做出了许多尝试。Chugh等人根据指纹匹配任务中“具有辨识度的特征往往集中于指纹细节处”的先验知识，将指纹样本基于细节点分成多个小块并对齐，然后对所有小块进行活体检测并融合以生成全局的活体分数[13]。由于指纹细节处的辨识特征（比如气泡、裂纹形状）在活体指纹与各种伪造指纹间广泛存在，这种方法大大提升了对于未知攻击材料的检测能力。但是，基于小块的检测需要将每个样本分成约50个小块，在测试时造成了较大的计算负担。另一方面，Anusha等人提出了一种极为复杂的深度集成方法，利用LBP、Gabor滤波器将指纹图像补充为三通道，并采用通道注意力、空间注意力和块注意力三种注意力模块，使用两个不同配置的DenseNet分类器合作产生活体分数，将其融合作为全局活体分数[14]。这种方法可以算得上当时的集大成之作，使用传统特征提取方式作为输入补充，同时构造了先进的注意力机制用于DenseNet集成，但由于极高的计算复杂度与内存开销，完全无法在现实中部署。 相应地，Liu等人考虑在任意网络架构下，通过改进训练过程获取更好的泛化性能。他们提出在网络训练过程中抑制一部分通道，从而去除输入中对分类任务有害的噪声，保留更纯粹的特征用于识别；此外，他们联系域自适应（domain adaptation）的方法构造了新的损失函数（PA Adaptation loss），从而在特征空间中压缩活体指纹样本，同时使不同材料的假指纹样本更加分散，从而获得了泛化能力更强的决策边界，同时只需要单个mobilenet作为分类器[15]。

上述深度学习指纹活体检测方法在检测准确率方面比之传统特征提取取得了长足的进展，但仍然面临许多问题。深度学习由于其庞大的参数量，往往需要足够的训练数据才能获得令人满意的稳定效果，但是由于指纹相关任务的隐私问题以及假指纹的制作困难性，指纹活体检测的数据非常难收集。即使是世界最大的活体检测组织LivDet每两年组织的竞赛，每种传感器的训练数据也仅有约2000张，这对动辄使用百万张图像作为训练样本（如ImageNet）的深度学习显然是不够的。基于预训练-微调范式的迁移学习一定程度上缓解了训练样本的稀缺[11]，除此之外，在[13]的基础上，Chugh等人使用修改的VGG网络构造GAN生成合成指纹小块，将这些小块作为补充与原始数据一起用于训练最终的识别网络。这些方式弥补了小样本识别带来的负面影响，但训练数据稀缺仍然是构造指纹活体检测系统时不得不考虑的问题。同时，由于庞大的存储和计算开销，深度网络仍然很难应用于终端设备，即使是轻量化网络的代表mobilenet v2，光是参数存储空间就需要14MB（为了存储中间表示，需要远大于此的内存开销），对于单张指纹图像，在台式机单核CPU上的处理时延达到数百毫秒，算力更弱的终端设备可能需要几秒的时间才能完成整个处理流程，这对用户来说非常不方便。总的来说，深度学习方法在指纹活体检测中表现出了卓越的性能，但仍然面临数据稀缺与算力匮乏的困难，并且现在还不存在良好的解决方案。

## 1.3 对抗攻击研究现状

本项目选择基于机器学习的生物指纹活体检测方案，故对演示攻击和对抗攻击以及对二者的防御分别展开讨论。

### 1.3.1 演示攻击

ISO/IEC 30107是国际标准化组织（ISO）和国际电工委员会（IEC）制定的一个生物特征认证标准，其中包括了生物特征认证中的安全性要求和测试方法[3]。ISO/IEC 30107-1：2016是该系列标准中的一个重要部分，提供了一种名为“演示攻击”的测试方法，用于评估生物特征认证系统的安全性。

根据ISO/IEC 30107-1标准中的定义，演示攻击是指向生物特征捕获子系统展示生物特征的过程，旨在干扰生物特征识别系统的运行。这种攻击可以通过使用演示攻击工具（PAI）或人造物（artefact）来实现。

以指纹识别系统为例，演示攻击可以通过使用假指纹、印章或其他人造物来实现。这些人造物可以用于欺骗指纹捕获设备，并干扰指纹识别系统的运行。为了防止这种攻击，需要使用PAD技术来检测是否存在演示攻击，并采取相应措施来保护指纹识别系统的安全性和准确性。

生物识别攻击检测（PAD）旨在通过分析呈现的生物识别样本的各种特征，如纹理、颜色和深度等，来检测这些攻击。ISO/IEC 30107系列提供了PAD机制开发和测试的指南，以确保其准确性和可靠性。

### 1.3.2 对抗攻击

当人们在安全应用中部署机器学习相关技术时，不可避免地，机器学习检测模型本身也暴露成为了攻击面的一部分。不幸的是，尽管具有强大的统计学习能力，机器学习并非总能给出有意义和可信的结论，他们的弱点很容易被攻击者利用，使检测模型以很高的概率，甚至极高的置信度漏报，从而让对手兵不血刃地侵入系统。

近20年前，研究者已经开始设计绕过机器学习检测模型的技术，通常基于逃避[16]或投毒[17]。逃避是一种测试时攻击，早期的逃避技术依赖攻击者手动改变恶意样本特征，使其在特征空间中近似于良性样本，从而增大机器学习或规则检测器的漏报率。而投毒是一种训练时攻击，攻击者事先向训练集中注入恶意样本，如翻转训练标签，从而降低检测模型的测试时性能。

2014年，针对神经网络的对抗样本被发现[18]，机器学习大厦上空笼上了一朵浓厚的乌云。对抗样本强大的破坏能力可以轻易颠覆网络输出，同时，它们在输入空间中与原样本非常接近，无论人眼还是算法都很难发现。此外，对抗样本对于攻击者的开销微乎其微，因为他们只需要定制相应攻击样本的扰动，由于其优异的可转移性，甚至能够在与真实部署模型不同的代理数据集和模型架构上制作，完成黑盒攻击。此后，基于对抗扰动的攻击方法快速发展，研究者很快发现了相应安全隐患在不同模型家族中的普遍存在[19]，强大的多步梯度攻击[20][21]使得借助于正则化、数据增强以及启发式梯度掩蔽的防御不再有效[22]，人们开始认识到，对抗攻击似乎是一种根植于高维空间中统计学习方法的本源问题，无论对于传统分类器还是神经网络，天然算法都在对抗样本前暴露出严重的脆弱性[23]。随着对抗样本的出现，投毒攻击也迎来了新的发展，人们开始使用中毒样本在神经网络等模型中植入后门[24]，从而模型对正常样本分类结果正确，但遇到带有触发器的样本时，则会错误分类，甚至在多分类时给出攻击者期待的结果。此外，不同于对抗扰动，后门投毒攻击的触发器可由攻击者任意选定，因此获得了更大的灵活性，但相应地，投毒攻击需要攻击者对训练过程有一定控制能力，有一定实施困难，但在众包标记、互联网数据收集以及模型外包等日益丰富的场景下仍然非常危险。

对于对抗攻击的防御仍然在探索阶段。对抗样本仿佛随统计学习方法强大拟合与泛化能力一同到来的诅咒，攻击者非常容易利用模型的内在脆弱性，而防御者想要做出弥补却非常困难。目前的两种主流防御是对抗训练[21]和随机平滑[25]。对抗训练通过不断生成对抗样本参与训练弥补模型的脆弱性，但常常面临干净样本上泛化性能与对抗样本上鲁棒性的权衡；随机平滑是一种可证明鲁棒的防御技术，但难以应对极为常见和危险的无穷范数攻击。此外，这两种方法都引入了相当大的时间开销，且防御能力仍远远不能让人满意。相对而言，对于投毒攻击的防御容易一些，常见的bagging集成方法（引导聚合）就可以提供对任意投毒的可证明鲁棒性[26]，但是，当基本模型计算开销较大时（如深度神经网络），这也将带来难以承担（比无防御时高2-4个数量级）的训练与部署开销，严重影响模型的可用性。

### 1.3.3 物理可制作对抗样本

虽然数字域的对抗样本有强大的攻击能力，但在传感器捕捉图像后加入数字扰动需要改变相应设备内部处理逻辑，非常困难。相应的，研究人员发现，在物理域构造对抗样本也是可行的。物理对抗性攻击是指通过在物理空间中添加扰动来达到攻击效果，例如在物理域添加对抗性补丁[27,28]和对抗性贴纸[29,30]等。目前，物理对抗样本已对人脸识别、自动驾驶、语音识别等多个领域造成威胁。

（1）常见攻击任务

1）人脸识别

在人脸识别领域，对抗性攻击旨在通过修改人脸的配饰欺骗面部识别系统输出错误的人（冒充攻击）或未检测到的结果（躲避攻击），例如Sharif等人[31]通过生成一组物理对抗性眼镜架达到了对人脸识别系统的欺骗。

2）车辆检测

对车辆检测的对抗性攻击旨在通过车身上添加图案，从而隐藏车身，使其不被监控摄像头检测到。Zhang等人[32]使用克隆网络来模拟模拟器和探测器对3D车辆的联合响应，通过最小化克隆网络的输出来生成车辆伪装图案，利用游戏引擎模拟物理环境引起的复杂变换，在3D空间中进行了实验。

3）人员检测

对人员检测的物理对抗性攻击的目标是生成物理扰动（例如补丁、衣服等），从而将人隐藏在现实世界中的自动监控摄像头之外。Thys等人[33]设计了一种算法，采用对象性输出作为对抗性损失的目标，将对抗性补丁放置在检测对象的中心，使手持对抗性贴片纸板的人隐藏起来从而躲避检测。

（2）常见攻击介质

1）贴纸：由Zhang[30]等人提出，通过粘贴在目标物体表面进行攻击，它是目前最常用的对抗性媒介。贴纸在有效性和鲁棒性方面具有竞争力，但隐蔽性不佳。

2）补丁：由Brown[27]等人提出，通过将目标物体局部区域替换为图像补丁进行攻击，和贴纸类似，易于操作。

3）对抗性服装：由Xu[33]等人提出，在服装上添加物理扰动。由于可穿戴，其隐蔽性得到了提升，目前常用于欺骗人员探测器。

4）对抗性眼镜：由Sharif[31]等人提出，攻击者戴着一副带有特殊图案的眼镜来欺骗人脸识别算法，从而逃避被识别或冒充他人。对抗性眼镜具有足够的有效性和鲁棒性，但由于其奇怪而明显的图案，其隐蔽性较低。

5）对抗性图像：由Chen[34]等人提出，将物理扰动分布在整个图像上以导致DNN误判，但在保证攻击有效的同时有效隐藏扰动具有挑战性。

6）摄像头：由Sayles[35]等人提出，通过修改相机而不是目标对象来执行攻击。到目前为止，已经提出了两种类型：相机的滚动快门效果和相机的图像信号处理（ISP）。该方法隐蔽性较高。

7）光：由Nguyen[36]等人提出，不直接修改对象，而是利用照明设备或投影仪将特殊的光投射到目标对象上，攻击实现快、隐蔽性佳，但由于添加扰动的光是不可见的，故其在强光环境攻击效果不佳。

（3）指纹领域物理对抗性攻击

为分析可误导指纹呈现攻击检测（FPAD）模块决定的微小扰动是否可以被“打印”出来，并以人工材料指纹复制品的形式加以利用，Marrone等人[37]使用数字和打印的对抗性图像对配有FPAD模块的指纹身份验证系统进行了攻击。他们使用对抗性扰动算法DeepFool[38]对指纹图像进行约束，添加能误导FPAD而人眼不可见的扰动，再对数字图像进行建模、打印并制作成物理对抗样本来进行攻击，通过评估指纹活跃度得分，首次验证了指纹对抗性演示攻击的可行性和危害性。

综上所述，我们经过全面分析，建立了从指纹演示攻击到指纹物理对抗样本的攻击链条，同时概述了相关防御措施，形成攻击者与防御者间的不断升级的多重博弈。其中，指纹领域对于对抗性演示攻击（通过带有对抗扰动的假指纹样本绕过活体检测从而入侵系统）的防御研究仍然处于空白，我们将尝试弥补这一攻击面，构建更加可信可靠的强安全生物指纹识别系统。

### 1.3.4 对抗样本防御现状

对抗样本是针对机器学习模型的一种攻击方式，攻击者通过对原始数据进行微小的修改，使得机器学习模型在处理这些修改后的数据时产生误差，从而影响其预测结果。对抗样本攻击对机器学习的安全性产生了重大威胁，因此防御对抗样本攻击是机器学习领域的一个热点研究方向。

目前，对抗样本防御的方法可以大致分为以下几类：

1. 对抗训练（Adversarial Training）：对抗训练是目前应用最广泛的对抗样本防御方法之一，它通过在训练过程中，将原始数据加入一定的扰动，使得模型能够更好地适应对抗样本的攻击。
2. 防御式扰动（Defensive Distillation）：防御式扰动是一种在训练和测试过程中都引入一定噪声的方法，以降低攻击者的攻击效果。该方法可以通过对输入数据进行随机扰动、对模型输出进行平滑化等方式来实现。
3. 基于检测的方法（Detection-based）：基于检测的方法是一种通过检测模型对输入数据的异常行为来识别对抗样本的方法。该方法可以通过建立对模型行为的模型、对输入数据进行附加特征等方式实现。
4. 模型修复（Model Repairing）：模型修复是一种通过修复模型参数或者重训练模型来提高模型对抗样本攻击鲁棒性的方法。该方法可以通过加入正则化项、改变网络结构等方式来实现。

然而，目前没有一种对抗样本防御方法是完美的，每种方法都有其优缺点和适用范围。因此，在实际应用中，应该根据具体情况选择合适的防御方法，并在不断的研究中不断完善和优化防御策略。

#### 1.3.4.1 对抗训练

对抗训练（Adversarial Training）是一种针对对抗样本攻击的防御方法，最早由Goodfellow等人在2014年提出。该方法的基本思想是在训练过程中，不仅使用原始的训练数据，还向训练数据中注入一定的对抗样本，以增加模型对抗样本攻击的鲁棒性。具体来说，对抗训练方法通常包括以下步骤：

1. 生成对抗样本：利用一定的攻击算法（例如FGSM，PGD等）生成对抗样本，这些对抗样本是在原始训练数据的基础上添加一定扰动的样本。
2. 将对抗样本加入训练集：将生成的对抗样本与原始训练数据一起组成新的训练集，用于训练模型。
3. 训练模型：使用新的训练集训练模型，使其能够更好地对抗对抗样本攻击。
4. 测试模型：在测试阶段，使用正常的测试数据集对训练好的模型进行测试，以评估其鲁棒性和性能。

对抗训练方法的优点是可以提高模型对抗样本攻击的鲁棒性，同时可以应用于各种类型的模型，包括传统的机器学习模型和深度学习模型。不过，该方法也存在一些缺点，例如生成对抗样本需要耗费一定的计算资源和时间，同时注入对抗样本也可能会影响模型的预测准确率，并且对抗训练只能防止已知攻击，需要在实际应用中结合具体情况进行调整和优化。

#### 1.3.4.2 防御式扰动

Defensive Distillation是一种针对对抗样本攻击的防御方法，它最初由Google Brain团队的Papernot等人在2016年提出。

Defensive Distillation的基本思想是在训练和测试阶段中引入一定的随机扰动，以降低对抗样本攻击的效果。具体来说，Defensive Distillation方法将原始的训练数据通过一个新的模型（即“蒸馏模型”）进行转换，得到一个新的“蒸馏数据集”，然后再用这个“蒸馏数据集”训练目标模型。在训练过程中，对每个样本的输入和输出都添加一定的扰动，以增加模型对抗样本攻击的鲁棒性。在测试阶段，同样地对每个输入数据进行随机扰动，然后将扰动后的输入数据输入到训练好的目标模型中进行预测。

随机平滑另一种常见方法，其基本思想是对原始输入数据进行一定程度的随机扰动，从而使得攻击者无法准确地计算出对抗样本的梯度或者不同类别之间的决策边界，从而使得对抗攻击难以成功。

Defensive Distillation的主要优点是可以应用于各种类型的模型，包括传统的机器学习模型和深度学习模型。同时，该方法也比较容易实现，只需要在训练和测试阶段中引入一定的噪声即可。不过，Defensive Distillation方法也存在一些缺点，例如对于某些特定的攻击方法，该方法的防御效果可能不够理想，同时在引入噪声的同时，也可能降低了模型的预测准确率。

#### 1.3.4.3 检测型防御

检测型对抗样本防御（Detection-based Adversarial Defense）通过训练一个检测器，检测输入是否是对抗样本，并针对对抗样本进行处理或拒绝，从而提高模型的鲁棒性和对抗攻击的防御能力。

具体来说，检测型对抗样本防御可以分为两个步骤：训练检测器和对抗样本处理。在训练检测器时，可以使用正常样本和对抗样本对模型进行训练，通过比较模型在正常样本和对抗样本上的预测结果，构建一个二分类的检测器，用于判断输入是否为对抗样本。在对抗样本处理时，可以选择对对抗样本进行降噪或转换为正常样本，或者直接拒绝该样本，从而减少对抗攻击的影响。

检测型对抗样本防御的优点是可以使用任何模型和任何对抗样本生成方法进行防御，并且可以将其与其他对抗防御方法结合使用以提高模型的鲁棒性。同时，该方法也有一些缺点，例如检测器可能无法检测到新型的对抗样本，以及对抗样本处理可能会导致正常样本的损失，降低模型的性能。

总之，检测型对抗样本防御是一种重要的对抗攻击防御方法，可以有效提高模型的鲁棒性和对抗攻击的防御能力，但也需要注意其局限性和应用场景。

#### 1.3.4.4 模型修复

模型修复（Model Repairing）是一种用于解决深度学习模型存在缺陷或错误的方法。模型缺陷或错误可能是由于数据质量问题、训练过程中的问题或模型本身的设计问题等引起的。模型修复旨在通过修改模型的参数或架构，使其更加健壮、准确和可靠。

常见的模型修复方法包括：

1. 数据清洗和增强：通过对训练数据进行清洗和增强，去除不必要的噪声和干扰，同时增加数据集的多样性，提高模型的鲁棒性和泛化能力。
2. 参数微调：对已经训练好的模型进行微调，通过调整模型参数，优化模型性能，例如准确率、鲁棒性等。
3. 网络剪枝：通过去除冗余的神经元和连接，减少模型的复杂度和参数量，提高模型的训练效率和推理速度。
4. 模型蒸馏：利用较大、复杂的模型作为“教师模型”，将其知识传递给较小、简单的模型作为“学生模型”，从而提高模型的泛化能力和鲁棒性。
5. 模型结构搜索：通过自动化搜索技术，探索最优的模型结构，从而提高模型性能和效率。

需要注意的是，模型修复方法虽然可以提高模型性能和鲁棒性，但也可能会引入新的问题或风险。因此，在实际应用中，需要结合具体场景和需求，谨慎选择和优化模型修复方法。

### 1.3.5 防御挑战

对抗样本防御是机器学习领域中一个重要的研究方向，但是目前仍然存在着很多挑战和困难。以下是一些对抗样本防御面临的挑战：

1. 鲁棒性和准确性之间的权衡：防御方法通常会对模型的准确性和鲁棒性进行平衡。对抗样本防御需要在尽可能保持模型准确性的同时提高模型对抗样本的鲁棒性，这是一个相对困难的平衡问题。
2. 对抗样本生成方法的演化：对抗样本生成方法在不断演化，从最初的FGSM、PGD到现在的C&W、DeepFool等更加复杂和难以检测的对抗样本生成方法，防御方法也需要不断升级和改进。
3. 对抗样本的多样性：对抗样本可以具有不同的扰动方式、不同的攻击目标和不同的威胁等级，因此防御方法需要能够应对不同类型的对抗样本。
4. 攻击者的智能化程度：攻击者不断提高攻击的智能化程度，使用更加高级的对抗样本生成方法和攻击策略，对防御方法的挑战也越来越大。

总之，对抗样本防御面临着多个挑战和困难，需要不断进行研究和探索，开发出更加有效的防御方法，从而提高机器学习模型的鲁棒性和对抗攻击的防御能力。

## 1.4 研究意义

随着指纹识别系统在许多日常应用中的广泛部署，如金融、国际贸易安全、智能手机生物锁等，指纹识别系统的安全性以及对非活体指纹攻击的脆弱性日益受到关注，图像识别领域尤其严峻的是表示攻击（presentation attack），随着指纹验证系统在安全应用中的普及，我们迫切需要研究将基于软件的表示攻击检测算法（PAD，或称指纹活体检测算法）嵌入到指纹识别系统中。

指纹活体检测作为一种特殊的模式识别任务，受限于传感器扫描精度、指纹的生物细节特性以及表面噪声，直接使用通用机器学习模型难以达到较高的准确率，此外，该任务对算法实时性能和存储有较高要求，尽管有一定生物学的先验知识，做到高可信度与可用性的权衡依旧比较困难。

对此，本项目针对生物指纹活体检测，提出决策树集成与CNN协同检测方案，赋予可证明鲁棒决策树集成模型对感知数据的学习能力，并采用知识蒸馏，集成学习，对抗训练方案进一步增强模型的防御能力。本项目进行了大量实验以验证模型的准确性与鲁棒性，并针对指纹任务进行控制存储与运算成本的设计以满足实际部署要求，希冀促进生物指纹识别技术的安全应用。

# 第二章 作品设计与实现

## 2.1 作品简介

由于基于对抗样本的演示攻击存在，传统生物指纹活体检测的安全性与可靠性有待进一步研究，且指纹设备的轻量级部署和低资源消耗对模型架构提出较高要求，本项目研究了当前主要活体检测方法，包括硬件方案、传统特征提取方案以及深度学习方案，复现大量SOTA检测方案，通过知识蒸馏，变学习率计划和精调数据增强策略改进CNN模型的准确率稳定性以及训练成本，并总结针对检测模型的对抗攻击（对抗样本、投毒攻击）和物理可制作对抗样本研究现状，提出基于知识蒸馏，对抗训练与集成的防御方案，设计和实现了抵御对抗性指纹演示攻击的强安全生物指纹识别系统，此外，本项目还将对检测方案进行测试，包括购买指纹采样传感器读入成员指纹，通过UV打印技术制作物理对抗样本，再次读取对抗样本的图像，放入已完成部署的模型测试物理对抗样本。

### 2.1.1 设计思路

本项目采用DNN模型与决策树集成模型协作检测方案。同时享有CNN表示学习能力强，以及决策树集成方法天然抗投毒攻击，且模型开销小故方便部署的优势，但仍需考虑赋予决策树集成对感知数据的学习能力问题：

1. 树集成深度架构

我们提出的第一个解决方案是搭建树集成深度架构，模拟DNN的分布式表示学习，这也是DNN泛化能力好的基础，下图是周志华教授的深度森林模型，是这种架构最著名的例子。而我们设计了全新架构：

前端Feature Extractor（使用stacking集成方法）：

1）XGboost梯度提升分类树预测概率，输入原图，输出预测概率作为特征，使用K折交叉验证，将原图数据集分为k份，k棵子树每个子树在k-1份数据集上训练，在剩下1份数据集上测试，并将测试结果作为特征送往下一层，保证每棵树输出的特征彼此独立，减轻过拟合

2）XGboost梯度提升回归树预测可疑分数，输入原图和第一层提取的特征，输出预测可疑分数作为特征

3）用多层随机森林决策路径交互激活特征（提取路径信息，取多次实验多棵树在分类过程中稳定出现的路径作为高阶特征，是对第一步中输初的特征的改良）。

后端Classification Head：使用随机森林作为分类器

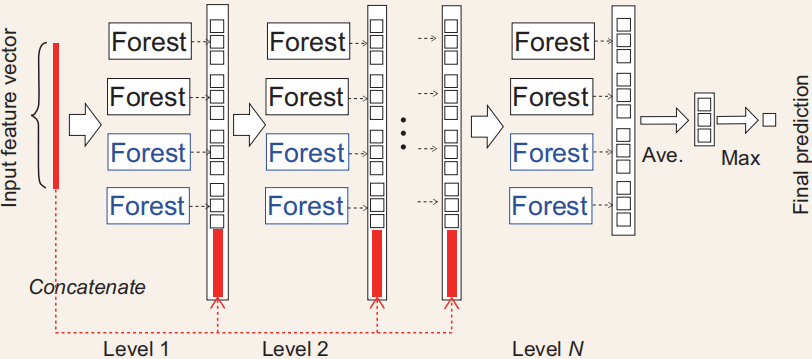


图2 深度森林架构

这种架构可以大幅缓解树模型严重的过拟合问题，模型准确率提升约5%。

1. 自动数据增强

但此时还远达不到实际部署安全目标，于是我们通过数据增强主动引导树模型去学习图像的旋转对称不变性，数据增强方法和参数的选用一反往日的手工测试选取，而是使用强化学习方法自动搜索适合，模型和数据集的增强顺序与强度。值得一提的是，这种搜索方法在神经网络上开销极大，搜索空间大10^30，而树模型的搜索空间大大降低，可使搜索更彻底。

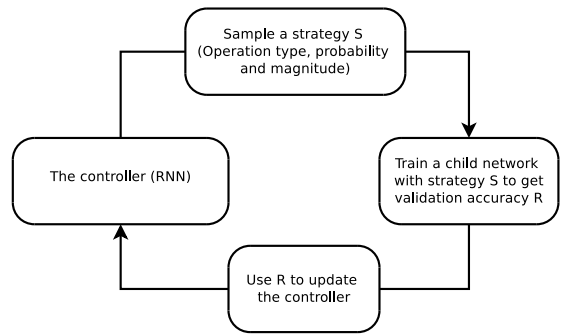


图3 强化学习数据增强搜索策略

1. 知识蒸馏

考虑到树模型可能有其它没考虑到的图像特性被丢失，我们通过改写XGBOOST的损失函数，实现神经网络向树模型的知识蒸馏，以学习CNN其余归纳偏置（即CNN如何解决问题），其中，分类树是交叉熵损失与KLD的线性加权，回归树是学生树与真实标签以及学生树与教师网络预测结果MSE的加权和。

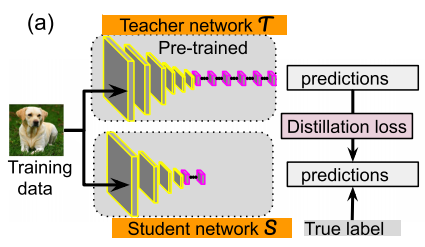


图4 知识蒸馏架构

第二个大问题是如何解决对抗样本带来的威胁，我们的解决方案是构建基于可证明鲁棒决策树的集成模型，即以可证明逃逸鲁棒决策树为基学习机，同时利用树集成对投毒的天然鲁棒性，构造可信检测模型。

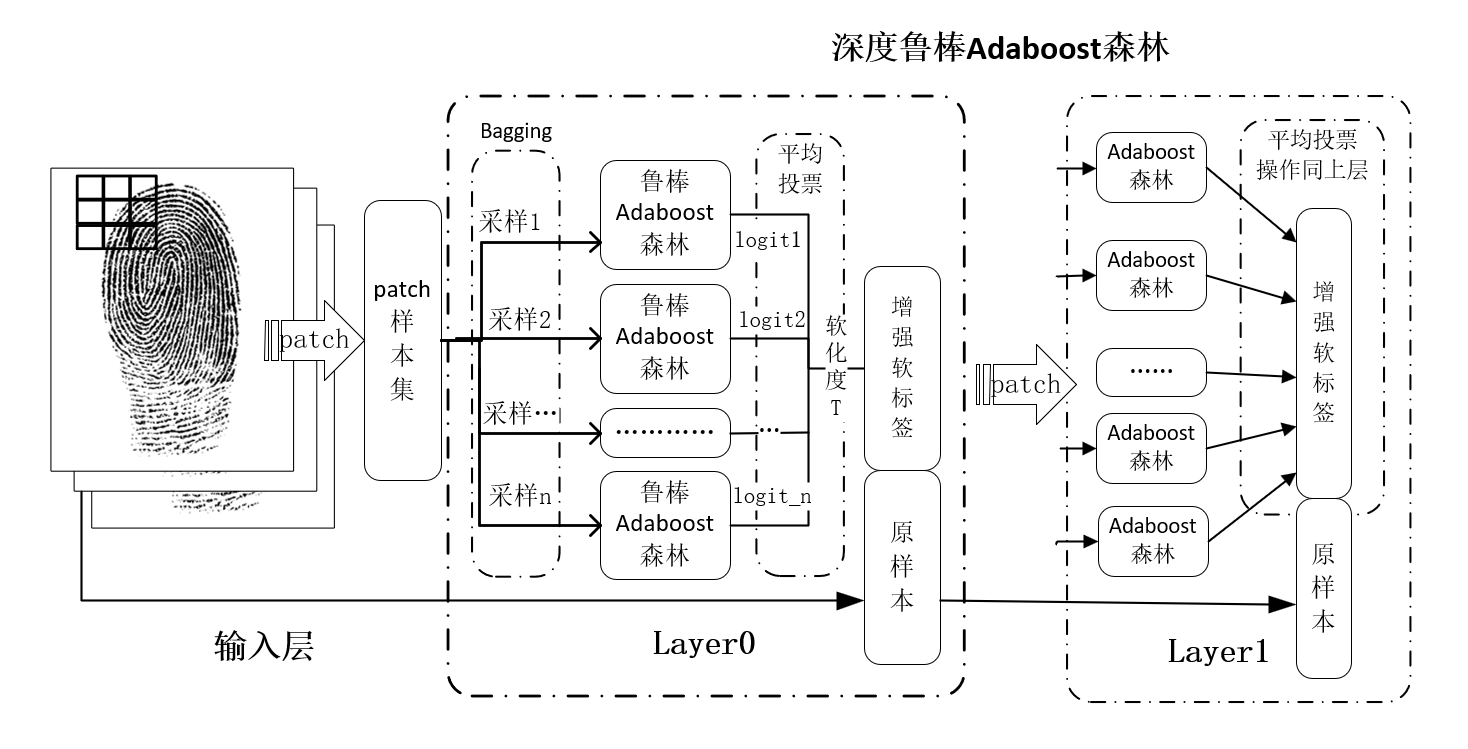


图5 深度鲁棒决策树集成模型

### 2.1.2 框架描述

本作品设计了一种基于鲁棒深度森林和CNN的指纹活体检测协同检测方法，应用于零信任场景下生物指纹认证中的伪造指纹识别，具体操作过程为：

1. 将所述检测模型部署在指纹认证设备后，当执行指纹认证时，从设备传感器（如光学、电容采集器）处收集指纹图像。
2. 利用所述深度鲁棒森林端到端处理指纹图像（无需特征提取），并同时使用CNN进行知识蒸馏，使用RNN执行自动数据增强方法搜索，得到识别结果（活性或伪造）和识别置信度。
3. 若识别结果为活体指纹，则对图像进一步做认证匹配处理；否则报警（疑似伪造指纹演示攻击）。

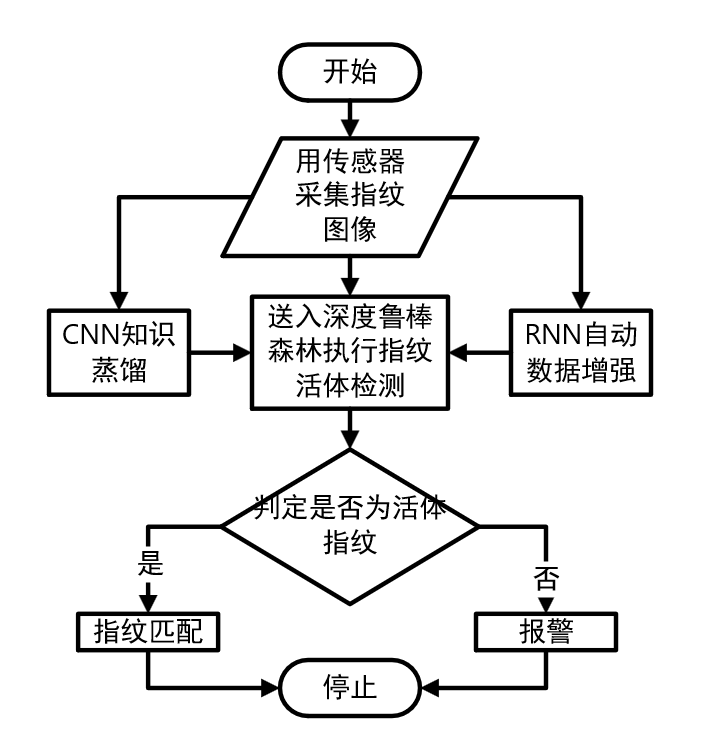


图6 总体框架图

## 2.2 决策树集成模型表示学习

### 2.2.1 深度模型架构

决策树集成模型（如随机森林和梯度提升树）是一种强大的预测模型，可以用于分类和回归问题。但是，决策树模型通常无法处理高维、非线性的数据，也容易过拟合。采用深度架构赋予决策树集成模型表示学习能力，可以提高模型的泛化能力和性能，处理高维、非线性数据，自适应地学习特征表示，同时保持一定的可解释性。

##### 2.2.2.1 XGBOOST架构

XGBoost（Extreme Gradient Boosting）是一种集成学习方法，它在Boosting框架下使用决策树作为基础模型，并通过梯度提升算法（Gradient Boosting）来训练模型。XGBoost具有高效、灵活、可扩展和高准确率等特点，因此在各种机器学习竞赛中都取得了优异的成绩。

XGBoost的基本架构包括两部分：决策树模型和梯度提升算法。

决策树模型：

XGBoost使用CART（Classification and Regression Tree）算法来生成决策树模型。CART算法是一种二叉树模型，每个节点包括一个特征和一个阈值，根据特征和阈值来划分数据集，直到满足停止条件。XGBoost可以支持二分类、多分类和回归问题，并且能够自动处理缺失值和异常值。

梯度提升算法：

XGBoost使用梯度提升算法（Gradient Boosting）来训练模型，其中包括两个关键部分：损失函数和正则化项。损失函数用于评估模型的预测能力，常用的损失函数包括均方误差、对数损失函数等；正则化项用于控制模型的复杂度，常用的正则化项包括L1正则化和L2正则化等。

XGBoost通过优化损失函数和正则化项的组合来训练模型，并利用贪心算法来搜索最佳分裂点和特征，以减少模型的复杂度和提高模型的准确率。在模型训练过程中，XGBoost还使用了二阶梯度加速技术和特征重要性评估方法，以提高模型的训练速度和准确率。

总的来说，XGBoost具有以下特点：

1. 高效性：通过并行计算和缓存优化等技术，提高了模型的训练和预测速度；
2. 灵活性：支持多种损失函数和正则化项，适用于各种回归和分类问题；
3. 可扩展性：支持分布式计算和大规模数据处理；
4. 鲁棒性：能够处理缺失值和异常值，提高了模型的鲁棒性；
5. 高准确率：在各种机器学习竞赛中都取得了优异的成绩。

XGBoost的缺点在于，由于使用了大量的超参数，需要进行调参来优化模型的性能；同时，由于使用了决策树作为基础模型，可能会出现过拟合的情况，需要进行正则化和限制树的深度等操作来避免。

为了解决这些问题，XGBoost还引入了一些新的特性和优化策略，如：

1. 支持并行计算和缓存优化等技术，提高了模型的训练和预测速度；
2. 支持自动剪枝和缺失值处理等功能，提高了模型的泛化能力；
3. 引入了一些新的正则化和限制操作，如L1和L2正则化、权重衰减、最大深度等，提高了模型的稳定性和可解释性；
4. 引入了特征重要性评估方法，可以帮助用户选择重要的特征并进行特征工程；
5. 引入了一些新的损失函数和目标函数，如Hessian Boosting、LambdaMart等，可以处理更加复杂的问题。

总的来说，XGBoost是一种高效、灵活、可扩展和高准确率的集成学习方法，已经成为机器学习领域的重要工具之一。它在各种实际问题中都取得了优异的表现，如广告推荐、搜索排序、金融风控等领域，被广泛应用于工业界和学术界。

##### 2.2.2.2 深度森林架构

深度森林是一种新兴的集成学习方法，它将传统的决策树和深度学习相结合，能够在保持传统决策树的可解释性和鲁棒性的同时，获得深度学习的强大建模能力。深度森林的基本架构如下：

1. 对于每个输入样本，先将其经过一些传统的决策树模型，得到一组中间层的输出；
2. 将这些中间层的输出拼接在一起，作为输入，输入到一个神经网络模型中；
3. 神经网络模型可以是任意深度的模型，例如多层感知器（MLP）、卷积神经网络（CNN）等；神经网络模型的输出即为最终的分类结果。

深度森林的优点在于，它可以利用决策树的分层结构，从而对不同层次的特征进行提取和组合，同时利用深度学习模型的非线性映射能力，可以对高层次的抽象特征进行学习。此外，深度森林还具有以下特点：

1. 可解释性：决策树的结构可以被解释为特征的重要性、特征的交互关系等；
2. 鲁棒性：由于每个决策树只看到一部分数据，因此可以避免过拟合和噪声的影响；
3. 高效性：由于决策树的结构，深度森林可以快速地对新样本进行预测，适用于大规模数据集。

深度森林的训练过程与传统的深度学习模型相似，通常采用反向传播算法进行优化。由于深度森林的结构较为复杂，因此需要较大的计算资源和训练时间。

##### 2.2.2.3 HIDF改良深度森林

大多数关于深度学习的研究都是基于神经网络模型，其中许多层参数化的非线性可微模块都是通过反向传播来训练的。最近，有研究表明深度学习也可以通过非可微模块实现，而不需要反向传播训练。所开发的表示学习过程是基于决策树森林的级联级联，其中高内存需求和高时间成本抑制了大型模型的训练。

优化深度森林HIDF（Hierarchical Interaction-based Deep Forest）模型是一种基于交互的深度森林模型，将实例直接传递到最后阶段，而不是通过所有级别，随着级联级别的增加，从低到高改变模型复杂性，从而进一步降低了内存需求和时间成本。其架构包括两个主要部分：特征提取和交互层。在特征提取部分，HIDF使用了多个决策树森林来提取输入数据的不同方面的特征。在交互层中，HIDF使用了一种新颖的交互方式来捕捉不同特征之间的相互作用，并将这些相互作用信息传递给下一层。交互层使用了两个子网络，一个是特征子网络，另一个是交互子网络。在特征子网络中，每个输入特征都被映射到一个低维空间中，并与其他特征进行组合。在交互子网络中，每个输入特征都被映射到一个高维空间中，并与其他特征进行组合。然后，这两个子网络的输出被连接起来，并传递给下一层进行处理。

这种交互方式能够更好地学习数据之间的复杂关系，并在各种任务上实现更好的性能。

### 2.2.2 自动数据增强

##### 2.2.2.1 常用图像数据增强算法

数据增强方法，可以帮助提高模型的鲁棒性和泛化能力。以下是一些常用的PyTorch数据增强方法及其参数的简要介绍：

1. RandomHorizontalFlip：随机水平翻转图像

参数：p (float) - 翻转概率，默认值为0.5

1. RandomVerticalFlip：随机垂直翻转图像

参数：p (float) - 翻转概率，默认值为0.5

1. RandomResizedCrop：随机裁剪和缩放图像

参数：size (int或tuple) - 输出图像的大小，scale (tuple) - 随机缩放比例范围，ratio (tuple) - 随机长宽比范围，默认值为(0.08, 1.0)

1. ColorJitter：随机调整图像颜色

参数：brightness (float) - 亮度调整范围，默认值为0，contrast (float) - 对比度调整范围，默认值为0，saturation (float) - 饱和度调整范围，默认值为0，hue (float) - 色相调整范围，默认值为0

1. RandomRotation：随机旋转图像

参数：degrees (float或tuple) - 旋转角度范围，默认值为0

1. RandomAffine：随机仿射变换图像

参数：degrees (float或tuple) - 旋转角度范围，默认值为0，translate (tuple) - 平移范围，默认值为None，scale (tuple) - 缩放范围，默认值为None，shear (float或tuple) - 剪切范围，默认值为0

1. RandomPerspective：随机透视变换图像

参数：distortion\_scale (float) - 扭曲程度，默认值为0.5

1. Normalize：图像归一化

参数：mean (sequence) - 各通道均值，std (sequence) - 各通道标准差

##### 2.2.2.2 特殊图像数据增强算法

当训练数据量不足时，深度学习模型容易出现过拟合问题。为了缓解这个问题，可以使用另外两种数据增强方法：mixup和random erase。

1. mixup

mixup是一种在训练数据中对输入样本进行线性插值的方法，它可以在一定程度上缓解过拟合问题。具体来说，mixup算法会对输入样本进行如下处理：

随机从训练集中选取两个不同的样本；

为两个样本的特征和标签之间进行线性插值；

计算插值后的样本的特征和标签，并用这个样本进行训练。

1. random erase

random erase是一种随机擦除图像一部分矩形区域的方法，它可以引入一定的随机性，增加模型对数据的鲁棒性。random erase有3个重要参数：擦除区域的面积、擦除的强度、填充值，pytorch调用示例如下：

RandomErasing(p=0.5, scale=(0.02, 0.33), ratio=(0.3, 3.3), value=0, inplace=False)

其中，p是擦除的概率，scale是擦除区域面积的范围，ratio是擦除区域长宽比的范围，value是填充擦除区域的像素值，inplace表示是否原地修改输入数据。

这两种方法和深度森林架构配合非常好，在我们的数据集实验上提升了3%-5%的准确率。

##### 2.2.2.3 可学习的自动数据增强

自动数据增强（AutoAugment）使用强化学习算法来搜索最佳的数据增强策略。具体来说，它使用了一个由循环神经网络（RNN）组成的控制器和Proximal Policy Optimization算法作为训练算法，它的输入是当前图像和当前子策略，输出是下一个子策略。在每个步骤中，控制器会预测一个由softmax产生的决策，并将其作为嵌入传递到下一步。控制器总共有30个softmax预测，以预测5个子策略，每个子策略包含2个操作，每个操作包括三个参数：操作类型、幅度和概率。每个mini-batch中，AutoAugment会随机选择一个子策略来对图像进行增强，并用模型在测试集上的准确率作为reward。关键因素包括搜索空间的设计、训练算法的选择以及超参数的调整等。

### 2.2.3 CNN对树集成模型的知识蒸馏

知识蒸馏（knowledge distillation）是一种模型压缩技术，可以将一个复杂的模型压缩成一个简单的模型，同时保留其预测能力。具体来说，知识蒸馏是通过将一个大型、复杂的模型（称为“教师模型”）的知识“蒸馏”到一个小型、简单的模型（称为“学生模型”）中，从而让学生模型具有与教师模型相近的预测能力。知识蒸馏技术最早由Hinton等人在2015年提出，现已被广泛应用于各种模型压缩和加速领域，例如深度神经网络、树模型和强化学习模型等。。

具体来说，知识蒸馏技术通常分为两个步骤：

训练教师模型：首先，需要训练一个大型、复杂的模型作为教师模型，通常使用较深的神经网络，例如卷积神经网络（CNN）或循环神经网络（RNN），以及较大的数据集。教师模型需要具有较高的准确率和预测能力。

训练学生模型：接下来，需要训练一个小型、简单的模型作为学生模型，通常使用较浅的神经网络，例如浅层的CNN或全连接神经网络。学生模型需要尽可能地接近教师模型的预测能力。具体来说，可以使用教师模型的输出作为学生模型的“标签”，并将这些输出作为额外的信息传递给学生模型，从而帮助学生模型更好地学习教师模型的预测能力。此外，还可以通过一些正则化方法（例如温度缩放或平滑化）来控制学生模型对教师模型的“软目标”进行学习，从而更好地保留教师模型的预测能力。

CNN对树集成模型的知识蒸馏主要通过改写XGBoost的损失函数设计：

· XGboost分类树蒸馏：

· XGboost回归树蒸馏：

其中分类树是交叉熵损失与KLD的线性加权，回归树是学生树与真实标签以及学生树与教师网络预测结果MSE的加权和。α是一个超参数，用于控制每棵树的重要性，CrossEnTropy可以控制二者的预测软标签分布一致。

通过将教师模型的预测结果作为输入，学生模型可以尝试学习教师模型的预测能力，从而提高模型的性能和泛化能力。

## 2.3 对抗样本攻击方案

本项目使用FGSM（Fast Gradient Sign Method）和PGD（Projected Gradient Descent）两种常见的对抗性攻击方法。

FGSM：

FGSM是一种快速有效的攻击方法，通过对原始输入数据进行微小的修改来生成对抗性样本。它是一种基于梯度的攻击方法，其基本思想是利用目标模型的梯度信息来计算最小的扰动值，使得对抗性样本的误差最大化。具体来说，假设目标模型的损失函数为J(x,y)，其中x是输入数据，y是正确的标签。则对抗性样本x'可以通过以下公式计算：

x' = x + ε sign(∇x J(x,y))

其中ε是一个小的正数，sign是符号函数，表示在梯度方向上对x进行微小的扰动。这个公式的含义是将输入数据x沿着损失函数的梯度方向进行微小的扰动，从而得到一个新的对抗性样本x'。

PGD：

PGD是一种更加复杂的攻击方法，它可以通过多次迭代来生成对抗性样本，并在每次迭代中对输入数据进行微小的扰动，以增加攻击的隐蔽性和鲁棒性。与FGSM不同，PGD可以产生更具挑战性的对抗性样本，因为它允许在每次迭代中进行多次微小的扰动。

PGD攻击的基本思路是在每次迭代中，通过计算输入数据的梯度方向来确定微小的扰动，然后对输入数据进行微调以生成新的对抗性样本。由于攻击者不知道目标模型的内部参数，PGD通常是基于黑盒攻击的思想，即攻击者只能利用目标模型的输入输出信息进行攻击。

具体来说，假设目标模型的输入数据为x，输出为f(x)，损失函数为J(x,y)，则PGD攻击可以表示为以下迭代过程：

x\_0 = x

for i in range(T):

x\_i = x\_{i-1} + α sign(∇x J(x\_{i-1},y))

x\_i = clip(x\_i, x - ε, x + ε) # 将x\_i限制在一定范围内

return x\_T

其中T是迭代次数，α是每次迭代的步长，ε是扰动的范围。该迭代过程会不断生成新的对抗性样本，直到达到预设的迭代次数或达到一定的攻击效果为止。由于PGD攻击具有更高的鲁棒性和隐蔽性，因此它通常被认为是一种更加强大和有效的攻击方法。然而，由于其计算复杂度较高，因此通常需要更多的计算资源和时间。

## 2.4 防御方案

近年来，指纹活体检测的诸多出版物使得PAD算法性能得到大幅改进，MobileNet等模型的提出也使深度网络在资源受限的终端部署成为可能。但是，随着对抗扰动被发现，识别算法本身的漏洞逐渐被人们所熟知，附加在输入样本上的微小扰动便可能颠覆模型输出，即便在正常样本上性能优异，当面对有所准备的攻击者时，当前的指纹PAD算法已不再安全。针对神经网络的对抗样本，已经提出了许多防御方法，但要么面临识别准确率与鲁棒性的权衡，要么会大幅增加测试时开销，不适用于指纹PAD的场景。

相应的，决策树集成也是一种强大的学习算法，随机森林、Adaboost、梯度提升树等方法在表格问题上得到了广泛的应用。决策树集成同样受到对抗扰动的困扰，但由于树模型良好的可解释性，相对容易构造识别准确率仍然优秀且可证明鲁棒的算法变体，且树模型部署后样本可以以线性复杂度由根节点路由到叶节点，效率很高。比如，2022年提出的快速可证明鲁棒决策树（FPRDT）便非常适合指纹PAD场景。受限于树模型对空间依赖表示能力的欠缺，单棵树和普通集成（bagging或boosting）很难提供令人满意的识别能力。因此我们基于深度森林的架构，利用鲁棒森林模仿卷积运算提取图像块作为补充细化表示构建更为强大的模型。

### **2.4.1 可证明鲁棒决策树**

Fast Provably Robust Decision Tree (FPRDT)是一种快速算法，用于实现可证明鲁棒性的决策树。该算法采用贪心递归方法直接最小化对抗0/1损失，通过优化所有叶子节点和分裂子节点来实现。与全局优化所有内部和叶子节点以及局部优化单个分裂节点及其子节点不同，该方法的计算复杂度为O(n log n)。因此，FPRDT能够在保证可证明鲁棒性的同时，具有较低的计算复杂度。

### **2.4.2 集成学习**

集成学习可以增强模型的泛化能力，准确率，可扩展性等，当前主要有三种集成学习方法：bagging，boosting，stacking。

##### 2.4.2.1 bagging

Bagging是一种基于集成学习的方法，它通过在原始数据集中有放回地随机抽取样本，构建多个子数据集，并分别用子数据集训练多个基础模型，最终将基础模型的结果进行平均或投票等方式进行预测。

Bagging的基本流程如下：

1. 从原始数据集中有放回地随机抽取一部分样本，构建一个新的子数据集；
2. 使用子数据集训练一个基础模型，并对验证集进行预测；
3. 重复上述步骤，构建多个子数据集，训练多个基础模型，并对多个基础模型对测试集的预测结果进行平均或投票等方式进行预测。

Bagging的优点在于，它可以通过构建多个基础模型来降低过拟合的风险，提高模型的鲁棒性和泛化能力。此外，Bagging还具有以下特点：

1. 可扩展性：可以使用不同类型、不同参数的基础模型，并对它们进行组合；
2. 稳定性：由于基础模型是通过随机抽样构建的，因此可以减小训练集中的噪声和异常值的影响；
3. 高准确率：在一定条件下，Bagging可以达到较高的准确率。

Bagging的缺点在于，由于基础模型是相互独立的，因此可能无法捕捉到不同基础模型之间的相关性。此外，由于需要训练多个基础模型，因此Bagging可能需要较长的训练时间。

##### 2.4.2.2 boosting

Boosting是一种基于集成学习的方法，它通过反复训练一系列弱分类器，逐步提升整体模型的性能。Boosting的基本思路是，在每一轮训练中，给错误分类的样本增加权重，以便在下一轮训练中更加关注这些难以分类的样本，从而提高模型的准确率。

Boosting的具体实现方法有多种，其中最著名的是Adaboost算法。Adaboost算法的基本流程如下：

1. 初始化样本权重，每个样本权重为1/n，n为样本总数；
2. 对于每个基础模型，根据当前样本权重进行训练，并计算模型的错误率；
3. 根据模型的错误率，计算模型的权重系数，用于更新样本权重；
4. 根据样本权重，构建新的训练集，并重复上述步骤，直到达到指定的迭代次数或模型的准确率满足要求。

Boosting的优点在于，它可以利用多个基础模型的优势，通过增加错误分类样本的权重，来逐步提高整体模型的性能。此外，Boosting还具有以下特点：

1. 可解释性：基础模型通常是简单的分类器，因此模型具有较好的可解释性；
2. 鲁棒性：重点关注错误分类的样本，可以一定程度上缓解噪声和过拟合问题；
3. 高准确率：在一定条件下，Boosting可以达到较高的准确率。

Boosting的缺点在于，由于需要反复迭代，因此可能需要较长的训练时间。同时，Boosting的性能也取决于基础模型的选择和参数的调整。

##### 2.4.2.3 stacking

Stacking是一种基于模型融合的集成学习方法，它通过将多个基础模型的预测结果作为新的训练数据，再训练一个次级模型来对最终结果进行预测。

Stacking的基本思路如下：

1. 将原始训练集划分成两个部分，一部分用于训练基础模型，另一部分用于构建次级训练集；
2. 对于每个基础模型，利用第一步中的训练集进行训练，并对第二步中的验证集进行预测；
3. 将第二步中的所有基础模型的预测结果拼接在一起，作为次级训练集的输入特征；
4. 利用第一步中未参与训练的原始训练集标签，作为次级训练集的输出标签，对次级模型进行训练；
5. 对于测试集数据，先将其输入到第二步中的所有基础模型中进行预测，然后将预测结果输入到第四步中的次级模型中进行最终预测。

Stacking的优点在于，它可以利用多个基础模型的优势，来提高最终模型的准确性和泛化能力。此外，Stacking还具有以下特点：

1. 灵活性：可以使用不同类型、不同参数的基础模型，并对它们进行组合；
2. 可解释性：可以利用第三步中的特征重要性来解释最终模型的预测结果；
3. 鲁棒性：使用了多个基础模型，可以避免单个模型的局限性和过拟合问题。

Stacking的缺点在于，它需要较大的计算资源和训练时间，同时在构建次级训练集时可能存在过拟合问题。

##### 2.4.2.4 集成架构设计

深度鲁棒森林检测模型的构造过程如下：

1. 鲁棒子学习机为基于快速可证明鲁棒决策树模型（FPRDT，Fast Provably Robust Decision Trees），作为识别模型中最基本的子算法。
2. 进一步构造鲁棒Adaboost：基于提升（boosting）算法，由上述快速可证明鲁棒决策树构造，作为学习能力更强的子学习机。
3. 基于自主采样（bagging）算法，由鲁棒Adaboost构造鲁棒Adaboost森林，构成深度鲁棒森林模型基本处理模块，即集成的集成，作为较强的子学习机参与鲁棒深度森林构建。

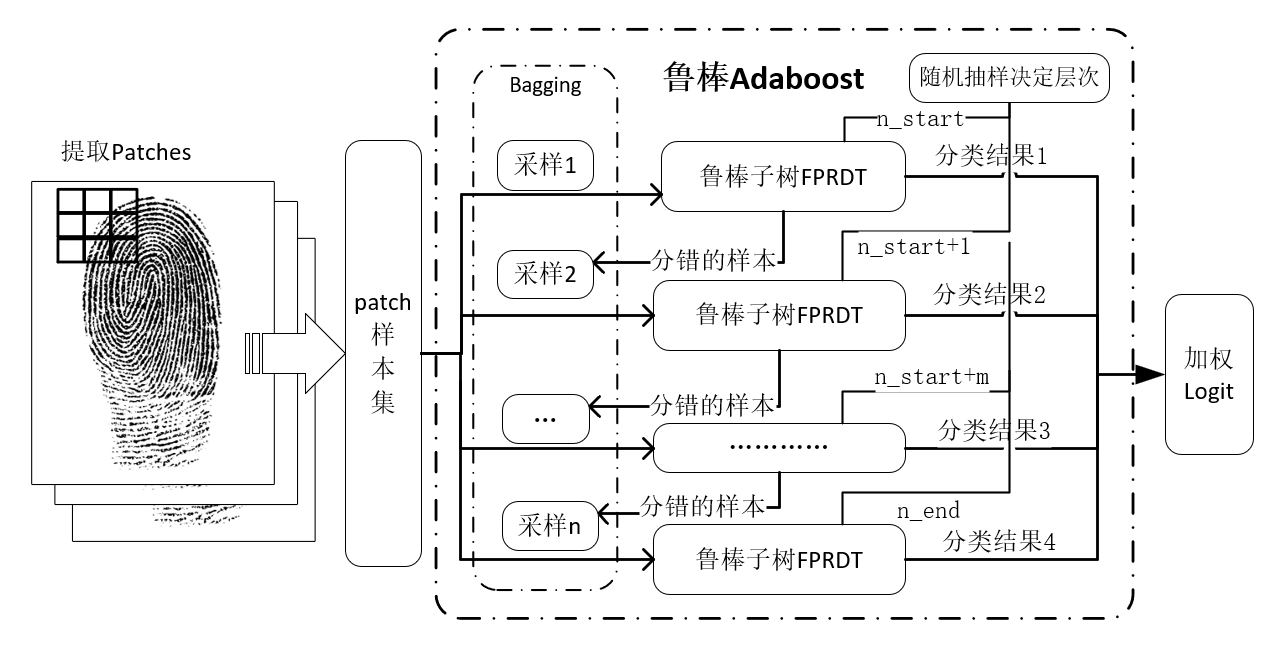


图7 深度鲁棒Adaboost架构

下面详细叙述深度鲁棒森林模型设计，即基于S2中鲁棒子学习机，改进原始深度森林模型架构，堆叠基本处理模块得到深度鲁棒森林模型；

1. 深度鲁棒森林layer 0构建：利用n \* n滑动窗口扫描训练样本，利用得到的图像块（patches）训练第一个基本处理模块（鲁棒Adaboost森林），并将每个块的输出概率作为增强表示与原样本特征向量拼接形成增强的特征向量；训练结束后，测试时对输入样本同样处理输出增强表示。
2. 深度鲁棒森林layer 1构建：在完整的增强样本特征向量上训练第二个基本处理模块，并将每个样本的输出概率作为增强表示与原样本向量（非layer 0得到的增强向量）拼接形成新的特征向量；训练结束后，测试时对输入样本同样处理输出增强表示。
3. 深度鲁棒森林layer 2构建：在完整的由layer 1得到的增强样本特征向量上训练第三个基本处理模块，对给定输入输出识别结果（活体或伪造）及置信度；训练结束后，测试时对给定样本给出识别结果与置信度。

下面详细叙述变层次提升策略，即随机决定鲁棒子学习机中鲁棒Adaboost模型的提升层次：当构建鲁棒Adaboost森林时，其中每个鲁棒Adaboost子学习机的提升层次并非直接给定，而是通过超参数n\_start与n\_end指明最低与最高提升层次，并随机抽样决定其最终提升层次，从而提高子学习机多样性，获得更好的泛化性能。

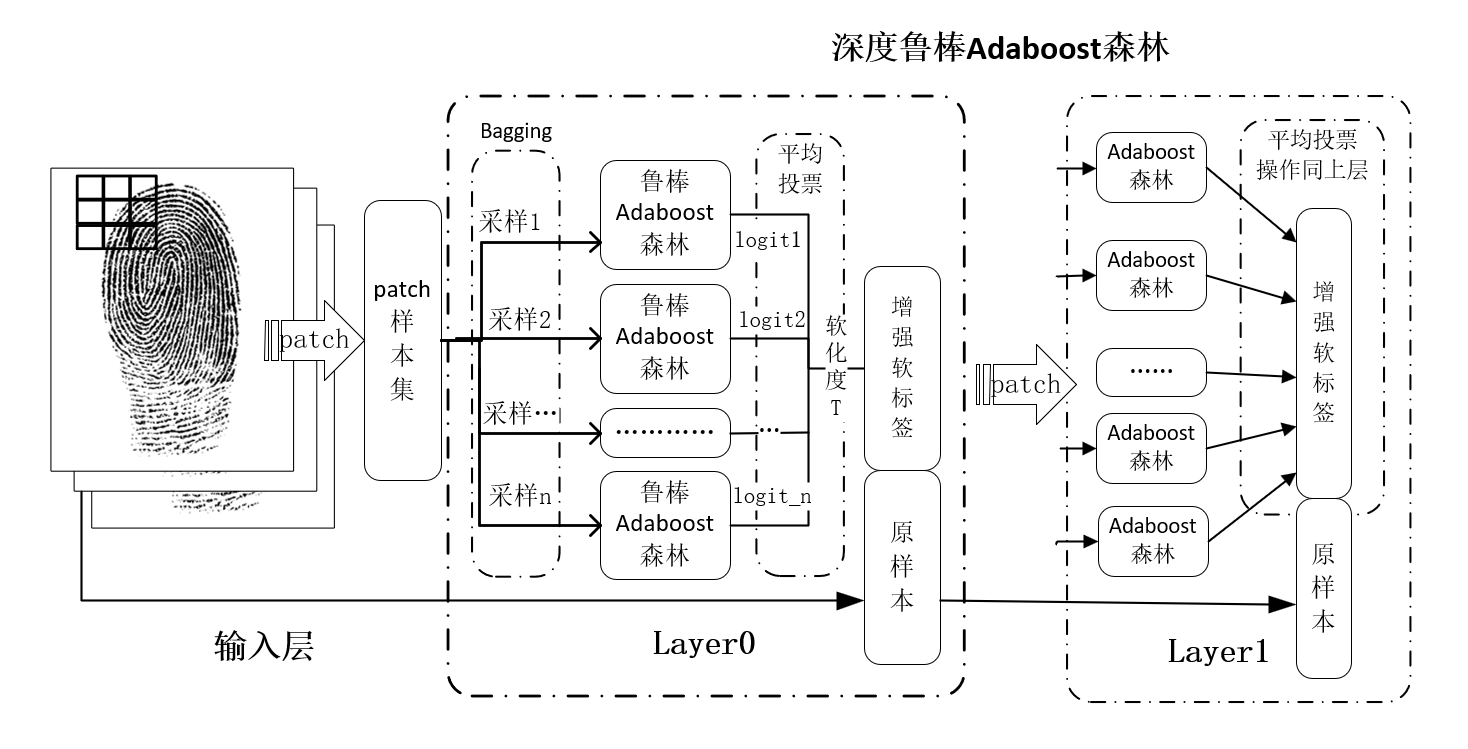


图8 深度鲁棒森林架构

下面详细叙述增强特征向量的制作，即通过鲁棒软标签投票策略，将鲁棒子学习机输出的二值结果（活体或伪造），转化为鲁棒软标签，从而使深度鲁棒森林中的多数投票与增强表示更灵活，获得更好的泛化性能。此步骤进一步包括：

（1）对于每个图像或图像块，由鲁棒Adaboost得到logit输出：

其中为鲁棒adaboost中每个鲁棒决策树的输出结果（0-1二值数据），为鲁棒Adaboost算法确定的组合系数。

（2）对于上一步中logit输出，由Sigmoid函数将其转化为概率：

其中T为超参数（temperature），控制所得概率的软化程度，当T越大时，所有样本输出的概率分布越接近均匀分布，probability即为所述鲁棒软标签，可进一步在鲁棒Adaboost森林的构建中参与平均投票，即：

confidence即基本处理模块（鲁棒Adaboost森林）输出置信度，在深度鲁棒森林layer 2中，可利用sign函数将其转化为0-1二值识别结果（活体或伪造）。

### 2.4.3 知识蒸馏

知识蒸馏如何将教师网络的对抗鲁棒性传递给学生网络？我们发现，即使只在干净的图像上进行训练，学生网络也可以吸收大量教师网络的对抗鲁棒性。具体来说，知识蒸馏通过让学生模仿教师在训练样本球内的输出来实现这一目标。这种方法是一种自然的对抗训练模式，但是应用于蒸馏过程中。

我们采用的Adversarially Robust Distillation (ARD)是一种新的知识蒸馏方法，旨在将对抗鲁棒性传递给学生网络。ARD的蒸馏过程如下：

1. 使用一个大型的教师网络进行训练，该网络在对抗攻击下表现出很高的鲁棒性。
2. 使用教师网络对训练数据进行预测，并将这些预测结果作为“软目标”传递给学生网络。
3. 在训练过程中，学生网络被要求模仿教师网络在对抗样本球内的输出。这意味着学生网络不仅需要在干净图像上表现良好，还需要在对抗攻击下具有鲁棒性。 为了加速训练过程，可以使用快速对抗训练方法来进行ARD蒸馏。
4. 在测试阶段评估学生网络在自然和对抗样本上的表现。

其中，快速对抗训练是一种加速对抗训练的方法，用于加速蒸馏过程。其具体过程如下：

1. 在每个训练迭代中，选择一个小批量的样本。
2. 对于每个样本，使用FGSM/PGD/Deepfool算法生成一个对抗性样本。
3. 使用对抗性样本和原始样本来更新学生网络的参数。
4. 重复步骤2和3多次，以增加对抗性样本的数量。

我们发现使用ARD方法训练的学生网络在对抗攻击下表现更加鲁棒，并且同时表现出更高的自然精度。有趣的是，ARD学生网络甚至可能比其教师网络表现出更高的对抗鲁棒性。

### 2.4.4 对抗训练

本项目使用的制作对抗性样本的方法包括：FGSM和PGD，通过最大化模型损失函数的线性近似来生成对抗性样本。

对抗性训练是一种通过在训练期间向模型中注入对抗性样本来提高模型鲁棒性的方法。具体而言，该方法通过在每个训练迭代中生成对抗性样本并将其添加到训练集中来增强模型的鲁棒性。这可以使模型更好地学习如何处理对抗性样本，并提高其在测试集上的准确率。

但是模型训练过程中的退化全局最小值会影响对抗性训练的有效性。在这种情况下，模型的损失函数无法可靠地近似为线性函数。具体而言，模型的决策面在数据点附近呈现出尖锐的曲率，从而降低了基于单个梯度计算的攻击效果。因此，模型学习生成弱扰动而不是防御强扰动，导致对黑盒攻击仍然容易受到影响。

此外，可以采用对抗性训练集的多样性增强：这是一种通过使用不同的扰动策略和参数来生成对抗性样本，例如可以使用FGSM、BIM、DeepFool等不同的攻击算法来生成对抗性样本，并将它们添加到训练集中。此外，还可以使用不同的扰动大小和方向来生成对抗性样本，以增加数据集的多样性。在使用对抗训练集进行训练时，需要确保在每个训练迭代中都会随机选择一个对抗生成器，并从该生成器中选择一个随机扰动进行训练。这有助于确保模型能够学习到各种类型和大小的对抗性扰动。

# 第三章 作品测试与分析

## 3.1 数据集简介

测试分析中主要使用ATVS\_FFP和LivDet系列数据集。数据集由采集正常指纹和各种材料、各种场景（隐指纹摘录或配合获取）下的攻击假指纹组成。具体信息可见https://livdet.diee.unica.it/。

表2 LivDet 2015-2021数据集参数说明

### **数据集内容**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Liv Det 2015 | | | | Liv Det 2017 | | |
| GreenBit | Biometrika | Digital Persona | CrossMatch | Green Bit | Orcanthus | Digital Persona |
| DactyScan26 | HiScan-PRO | U.are.U 5160 | L Scan Guardian | DactyScan84C | Certis2 Image | U.are.U 5160 |
| 500×500 | 1000x1000 | 252×324 | 640x480 | 500x500 | 300xn | 252x324 |
| 500 | 1000 | 500 | 500 | 500 | 500 | 500 |
| --- | --- | --- | --- | 1000/1700 | 1000/1700 | 999/1700 |
| --- | --- | --- | --- | 1200/2400 | 1200/1678 | 1199/2028 |
| Ecoflex, Gelatine, Latex, Wood Glue, Liquid Ecoflex, RTV | | | Body Double, Ecoflex, PlayDoh, OOMOO, Gelatin | Wood Glue, Ecoflex, Body Double, Gelatine, Latex, Liquid Ecoflex | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Liv Det 2019[6] | | | Liv Det 2021[7] | |
| Green Bit | Orcanthus | Digital Persona | Green Bit | Dermalog |
| DactyScan84C | Certis2 Image | U.are.U 5160 | DactyScan84C | LF10 |
| 500×500 | 300xn | 252x324 | 500x500 | 500x500 |
| 500 | 500 | 500 | 500 | 500 |
| 1000/1020 | 1000/990 | 1000/1019 | 1250/(2050/2050) | 1250/(2050/2050) |
| 1200/1224 | 1200/1088 | 1000/1224 | 1500/(2640/2640) | 1500/(2460/2460) |
| Ecoflex, Gelatine, Latex, Wood Glue, Liquid Ecoflex, Body Double | | | Latex, BodyDouble, ElmersGlue, etc. | |

LivDet参数整理如表2：

### **3.1.2 数据集获取过程**

小组成员1月初首先按照LivDet旧网址（09-15数据集申请信息在旧网址，17年之后的大赛数据集申请改到了新网址）数据集申请页面添好申请信息向指定邮件地址寄信，但可能因言辞生硬没有得到回信。

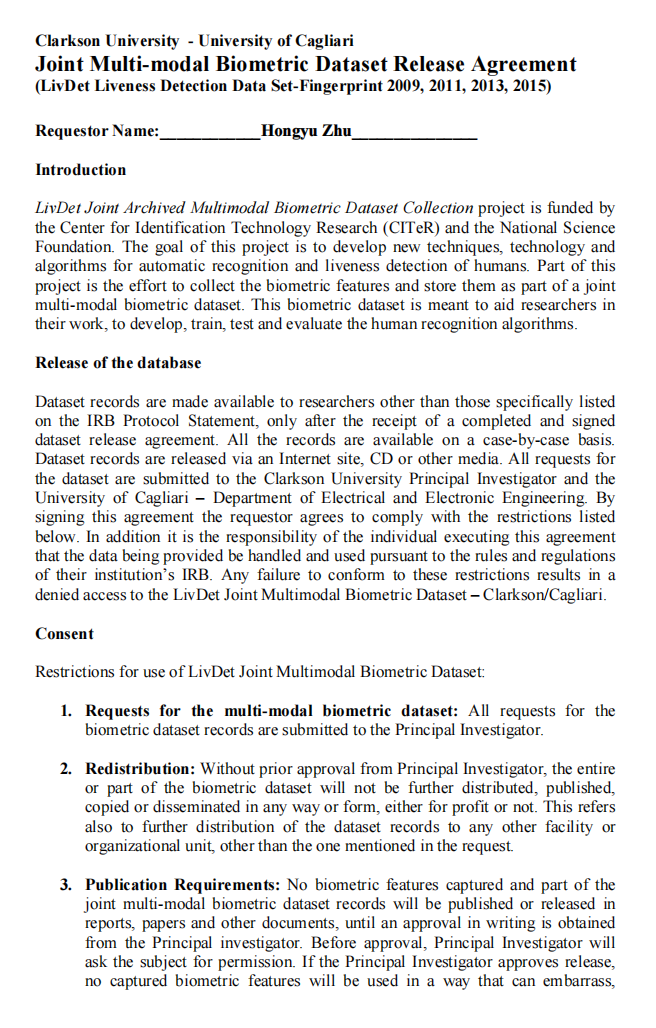


图9 LivDet官网09-15数据集申请信息



图10 LivDet17-21数据集申请信息

后小组成员于两周后再次寄出言辞诚恳的申请信，得到回信告知需要签一份数据集申请隐私保护与合法使用协议，需要教学院长或所在组织法律代表人的签名，学院公章等，签好后邮寄纸质版或含有数字签名的电子版。



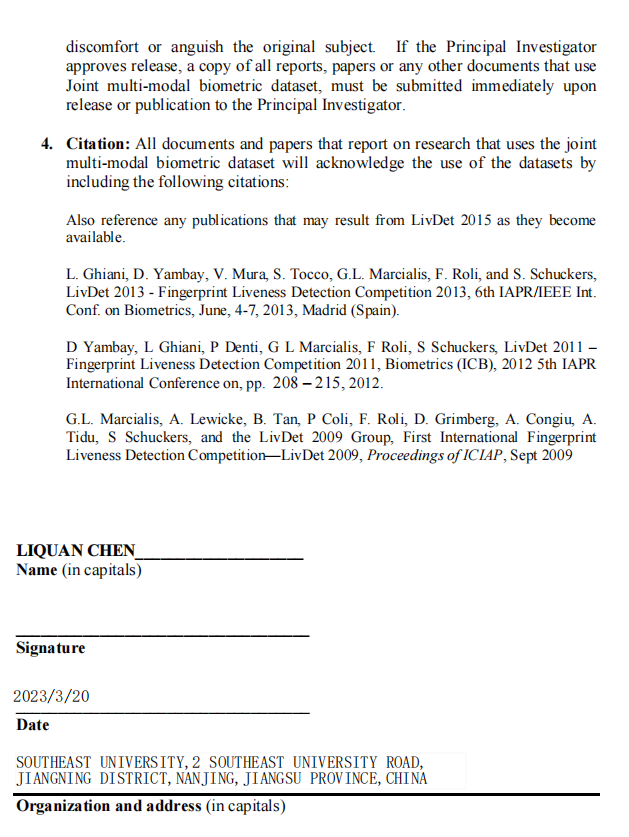


图11 LivDet09-15数据集申请协议（内容包括隐私保护与合法使用）

考虑到此时距离春节较近，不宜打扰学院老师合家团聚，于是先申请国内2010年的ATVS\_FFP数据集用来做实验暂缓燃眉之急（在中科大官网注册即可申请），在春节后一周向教学院长提出申请签证。数字签名和学院盖章的申请文件放在学校随心签上，但考虑到开学即考试比较忙碌，于3月中旬得到办理。

Livdet协议中要求邮件必须要法定代表人发送，然而申请邮件前后请陈院长发了两次，皆没有音信，但距离中期答辩已经很近，于是小组成员还是决定自己再发一封，才发现附件太大时163邮件系统不允许跨国寄信，缩减附件后终于拿到了心心念念的数据集。

在艰难得等待签章以及回信的过程中，小组考虑过申请领域内另一数据集：加拿大MSU的指纹数据集，组员申请到指导老师的数字签名签好协议，查到MSU作者已经博士毕业，目前在亚马逊工作，于是给作者所在实验室和作者个人的gmail又各发了一封邮件，但都没有回复。

巧妇难为无米之炊，3.16日小组拿到了所需的LivDet全套数据集。高质量的数据集于机器学习项目来说重要性不在话下：拿到数据集前小组在老数据集ATVS\_FFp上因各模型准确率都极高无法调整实验策略以及测试自己的创新，拿到数据集后小组发现了此前实验的大量缺陷，且不同模型架构间差异显著（如随机森林与卷积神经网络mobilenet），使用各种技巧的效果对比显著，对改进与提高准确率有极大帮助，进展飞快。

## 3.2 模型训练过程

尽管理论准备完善，实验上困难重重：

1. 相关领域开源论文太少，大部分文章没有源码（且联系不上作者），自己实现的版本无法达到论文中声称效果；少量开源论文按默认参数（或微调后）同样无法验证论文声称效果，从而难以比较和改进。

解决措施：目前已复现6篇最新论文，将持续实验和调整，之后会尝试复现更多方法。

1. 目前基于决策树集成设计全新的活体检测方法，需要对已有树集成算法做出较大改变，但XGboost等软件库集成度很高，接口几乎写死，改动较难进行；其余开源代码接口和文档做的很差，有些只支持命令行运行，同时底层优化不够，速度很慢，带来了额外的开销。

解决措施：目前在利用XGboost少量自定义特性（如损失函数）完成部分交互，同时修改HiDF等软件库源代码，写出新的接口便于后期开发。

### **3.2.1 模型参数设置**

深度鲁棒森林每层鲁棒Adaboost数量为64，提升层次为3-5（变长提升），鲁棒决策树最大深度设置为5。对抗扰动大小ε设置为0.03或0.05，投毒（标签翻转）比例为10%。

### **3.2.2 训练设置**

在Linux主机（Ubuntu 22.04，32核CPU，4 块RTX 3090 GPU），python 3.9.12环境下训练。ATVS-FFp数据集按8:2比例划分训练集和测试集，训练集内部进行5折交叉验证；livdet数据集使用提供方划分的训练集和测试集，训练集内部进行5折交叉验证。

### **3.2.3 测试设置**

与训练环境相同。

## 3.3 物理对抗样本制作

实体演示方面的困难主要包括指纹扫描传感器的购买以及物理对抗样本的制作。考虑到不同传感器采集指纹图片的精度与风格对检测效果的影响很大，需使用与训练数据集LivDet系列相近规格的传感器，但经充分调研筛选出符合条件的传感器价格过于昂贵。而市面大多数指纹传感器已经内置不可更改的硬件检测部件，无法录入假指纹，给我们的软件检测算法的部署验证增大困难。目前几经周折，已经购入适用传感器与假指纹制作模具，但真实物理对抗样本的制作仍处于尝试阶段。

### **3.3.1 传感器选型**

考虑到不同传感器采集指纹图片的精度与风格对检测效果的影响很大，需使用与训练数据集LivDet系列相近规格的传感器，但经充分调研筛选出符合条件的GreenBit 、 DactyScan84c、HiScan-pro、Guardian Lscan、LumidigmV302、MSO 330等 价格过于昂贵，同时市面大多数指纹传感器无法录入假指纹。指纹传感器的购买历程较为曲折：

项目需求：

（1）传感器需要能够采集明胶、硅胶等材料制作成的假指纹，故最好选择光学传感器（以避免使用电容式传感器内置活体检测机制可能遇到的假指纹采集困难情况）。

（2）传感器和指纹图像处理设备之间能够保持良好的通信。

通过阅读论文，我们发现GreenBit的传感器在各方面都有较好的表现，且在LivDet 15-19的数据集中都出现了GreenBit采集的指纹图像。确定品牌后我们先后找到了国内外购买途径，由于国外沟通较为不便且运输可能存在的问题，我们选择了国内的代理进行咨询，得知其售价为4800人民币，远超预算。

我们又搜索联系了HiScan-pro、Guardian Lscan、LumidigmV302、MSO 330等在权威论文中被使用过的传感器，均因预算不足而放弃；在京东上我们先后找到了ZKTeco、URU450、FRT610等光学传感器，但是经过咨询得知均由于内置不可更改活体检测算法难以采集假指纹。最后我们联系到菲格特指纹仪公司，经电话核实基本参数和其他信息等，符合我们的基础和特殊需求，终于敲定方案，完成购买。



图 12 GreenBit DS84C 指纹扫描仪



图 13 菲格特 RT808 光学指纹仪

### **3.3.2 对抗样本制作材料与工具**

我们首先对购置的RT808光学指纹仪进行了测试，用2D打印图案和3D的印章分别进行了录入假指纹的模拟。我们发现，使用纸张2D打印图案时，光学指纹仪会把透光时显示出的纸张纹路也录入进去，导致录入图案背景噪声过大；而使用3D印章时则呈现了较好的效果。考虑到2D打印更易呈现微小扰动，我们又尝试了在光滑、柔软且不透光的材料上标记图案进行录入，效果较好。因此，我们目前的首选思路是挑选合适的光滑、柔软、透光度低的片材（例如硅胶等），在上面2D打印指纹，测试其效果；3D打印指纹将作为备选方案。

材料与工具：

2D：硅胶（或其他软胶）、硅胶打印机

3D：UV树脂、乳胶、UV树脂打印机

### **3.3.3 对抗样本制作过程**

步骤一：在指纹图片上添加几乎不可见的微小扰动

步骤二：

方案一：在光滑、柔软、透光度低的片材（例如硅胶等）上2D打印指纹

方案二：设置合适的指纹脊高，3D打印出模具，再将乳胶等材料注入模具中制作出假 指纹

### 3.4 防御效果测试

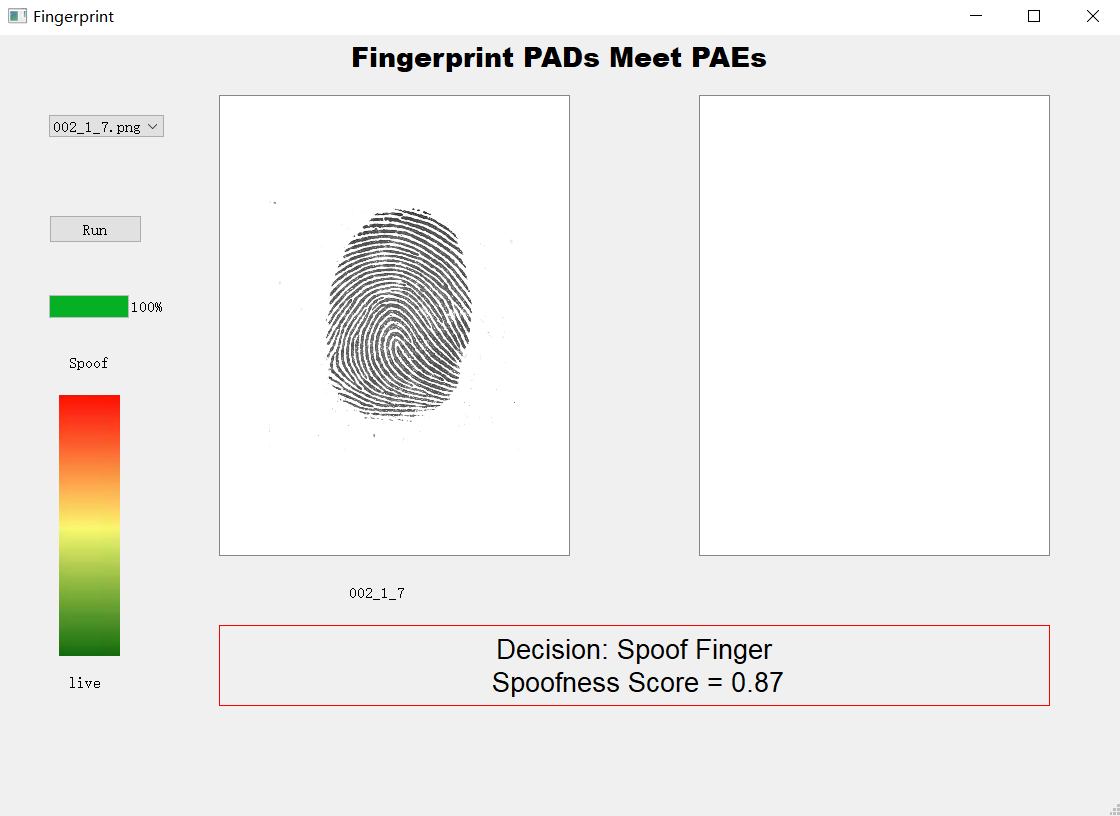
在ATVS-FFp活体检测数据集上的对比实验表明，本项目选择的深度森林架构能有效防御对抗样本和数据投毒，以及二者的混合攻击。当对抗扰动大小epsilon=0.05时，深度鲁棒森林准确率为96.94%；epsilon=0.03时，准确率为97.55%，而深度神经网络（resnet18）相应攻击下准确率分别为0%与12.25%。当执行10%的标签翻转投毒攻击时，深度鲁棒森林准确率为96.94%，即使对抗扰动（epsilon=0.05）与标签翻转协同攻击，准确率仍然能够保持在95%以上，而resnet18的准确率分别为94.50%与0%。

此外，变步长提升与鲁棒软标签策略对于准确率的提高有较大帮助，当每个鲁棒Adaboost中有64棵可证明鲁棒性决策树时，无投毒无对抗样本存在的情境下，不施加这两种策略的准确率为96.63%，施加后为97.24%；调整对抗样本对抗扰动的大小epsilon=0.03时，不施加这两种策略的准确率为96.33%，施加后为97.55%。

最后，本项目还比较了深度鲁棒森林与一般随机森林在对抗场景下的活体检测效果。同样设置500棵决策树作为子学习机，随机森林在无攻击情况下准确率为96.63%，而在对抗攻击（epsilon=0.05）存在的情况下降至86.24%，在10%标签翻转投毒攻击情况下为96.32%。与之相比，深度鲁棒森林在无攻击时准确率达97.24%，相同对抗样本存在的情况下准确率为96.94%，10%标签翻转投毒攻击后准确率为96.94%。进一步地，鲁棒深度森林可能通过增大规模获得更优秀的识别效果。

# 界面展示

本项目初步设计了用户图形界面，主要用于通过我们训练出的模型进行指纹的真假检测并输出检测分数，页面展示如下：



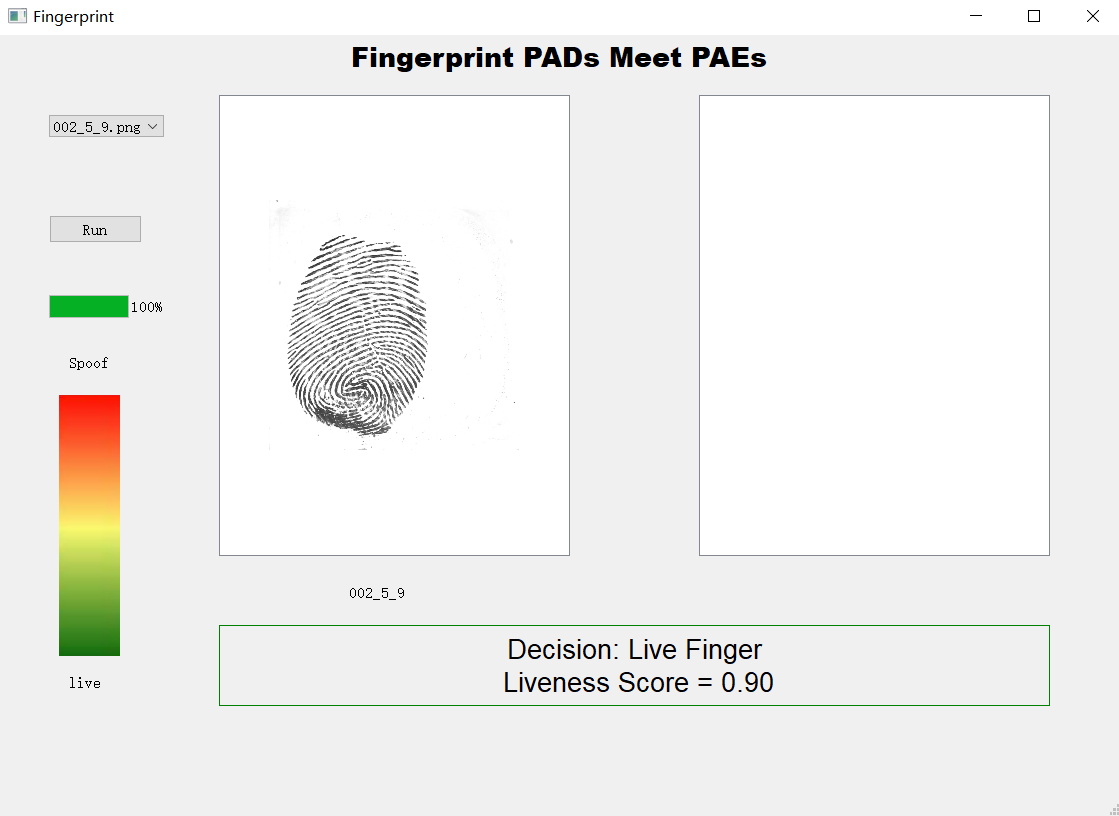


图14 初步用户图形界面展示

界面右侧的图形展示框还未进行开发，后期计划在此展示对抗样本检测情况（是否检测为物理对抗样本）。

# 创新性分析

与现有技术相比，本项目设计了一种基于鲁棒深度森林的指纹活体检测方法，应用于零信任场景下生物指纹认证中的伪造指纹识别，包括整体处理架构；鲁棒子学习机构造；深度鲁棒森林模型设计；变层次提升策略；鲁棒软标签投票策略。上述方法可以部署于现存指纹认证传感设备，具有良好的活体检测精度与处理开销权衡，同时对于包括逃逸和投毒的模型对抗攻击具备极高的鲁棒性，在各种威胁场景下展现出良好的生存能力，从而弥补了当前监测方案中算法本身的漏洞，能够更好地协助指纹认证方法保护设备完整性、可用性与用户数据保密性。本项目的创新点总结如下：

针对生物指纹活体检测，提出决策树集成与CNN协同检测方案，赋予决策树集成模型对感知数据的学习能力：

1.从DNN泛化能力基础——分布式表示学习获得启发，改进深度学习架构，设计全新决策树深度表示学习架构：前端是特征提取器，每一层同时使用XGBOOST梯度提升分类树与回归树，同时，每一层的输出特征还包括提取的决策路径信息，也即是取多次实验多棵树在分类过程中稳定出现的路径作为高阶特征。而后端的分类头使用随机森林模型。这种架构可以大幅缓解树模型严重的过拟合问题，模型准确率稳定在94%左右。

2.通过自动数据增强主动引导树模型去学习图像的旋转对称等不变性。数据增强方法和参数的选用并非传统手工测试选取，而是使用强化学习方法自动搜索适合模型和数据集特性的增强顺序与强度。此搜索方法在CNN上存储与时间开销极大，而树模型的搜索空间大大降低，可使搜索更彻底。

3.为防止树模型在第二步之外还丢失别的图像特性，通过改写XGBOOST的损失函数，实现神经网络向树模型的知识蒸馏，以学习CNN其余归纳偏置，其中分类树是交叉熵损失与KLD的线性加权，回归树是学生树与真实标签以及学生树与教师网络预测结果MSE的加权和。

针对对抗样本演示攻击采用知识蒸馏，集成学习，对抗训练方案。模型设计上：构建基于可证明鲁棒决策树的集成模型，即以可证明逃逸鲁棒决策树为基学习机，同时利用树集成对投毒的天然鲁棒性，构造可信检测模型。

# 总结与展望

本项目针对生物指纹活体检测，提出决策树集成与CNN协同检测方案，赋予可证明鲁棒决策树集成模型对感知数据的学习能力，并采用知识蒸馏，集成学习，对抗训练方案进一步增强模型的防御能力。本项目进行了大量实验以验证模型的准确性与鲁棒性，并针对指纹任务进行控制存储与运算成本的设计以满足实际部署要求，希冀促进生物指纹识别技术的安全应用。

本项目目前已完成一篇专利《一种基于深度鲁棒森林的指纹活体检测方法》投递中，SRTP项目（校级重大）中期答辩优秀，正在申请国省创。

未来，项目仍有大量改善工作有待完成：

首要的是提升树模型性能，需要改进森林表示学习框架，使之更加轻量准确。同时，考虑改良增强策略搜索方法，以求降低时间复杂度，并设计更有效的知识转移方法（包括设计更好的蒸馏损失函数，设计知识的其它编码表示）

其次，需要进一步测试和改善模型的鲁棒性，如结合多样化目的（入侵/拒绝服务/破坏取证）测试更先进的攻击，降低对抗训练或可证明鲁棒树开销。

参考文献

[1] Maltoni D, Maio D, Jain A K, et al. Handbook of fingerprint recognition[M]. London: springer, 2009.

[2] Yin X, Zhu Y, Hu J. Contactless fingerprint recognition based on global minutia topology and loose genetic algorithm[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 28-41.

[3] ISO/IEC 30107-3. Information technology-Biometric presentation attack detection-Part 3:Testing and reporting, 2016.

[4] Marasco E, Ross A. A survey on antispoofing schemes for fingerprint recognition systems[J]. ACM Computing Surveys (CSUR), 2014, 47(2): 1-36.

[5] Rowe R K, Nixon K A, Butler P W. Multispectral fingerprint image acquisition[J]. Advances in biometrics: Sensors, algorithms and systems, 2008: 3-23.

[6] Nogueira R F, de Alencar Lotufo R, Machado R C. Fingerprint liveness detection using convolutional neural networks[J]. IEEE transactions on information forensics and security, 2016, 11(6): 1206-1213.

[7] Agassy M, Castro B, Lerner A, et al. Liveness and spoof detection for ultrasonic fingerprint sensors: U.S. Patent 10,262,188[P]. 2019-4-16.

[8] Marasco E, Ross A. A survey on antispoofing schemes for fingerprint recognition systems[J]. ACM Computing Surveys (CSUR), 2014, 47(2): 1-36.

[9] Popli A, Tandon S, Engelsma J J, et al. A Unified Model for Fingerprint Authentication and Presentation Attack Detection[C]//2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2021: 1-8.

[10] Krizhevsky, A. , I. Sutskever , and G. Hinton . "ImageNet Classification with Deep Convolutional Neural Networks." Advances in neural information processing systems 25.2 (2012).

[11] Nogueira, R.F., de Alencar Lotufo, R., Campos Machado, R.: Fingerprint liveness detection using convolutional neural networks. IEEE Transactions on Information Forensics and Security 11(6), 1206–1213 (2016)

[12] Uliyan, D.M., Sadeghi, S., Jalab, H.A.: Anti-spoofing method for fingerprint recognition using patch based deep learning machine. Engineering Science and Technology, an International Journal 23(2), 264–273 (2020).

[13]Chugh, T., Cao, K., Jain, A.K.: Fingerprint spoof buster: Use of minutiae-centered patches. IEEE Transactions on Information Forensics and Security 13(9), 2190–2202 (2018). https://doi.org/10.1109/TIFS.2018.2812193

[14] Anusha, B., Banerjee, S., Chaudhuri, S.: Defraudnet:end2end fingerprint spoof detection using patch level attention. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2684–2693. IEEE Computer Society, Los Alamitos, CA, USA (mar 2020). https://doi.org/10.1109/WACV45572.2020.9093397

[15] F. Liu, Z. Kong, H. Liu, W. Zhang and L. Shen, "Fingerprint Presentation Attack Detection by Channel-Wise Feature Denoising," in IEEE Transactions on Information Forensics and Security, vol. 17, pp. 2963-2976, 2022, doi: 10.1109/TIFS.2022.3197058.

[16] Lowd, D. , and C. Meek . "Adversarial learning." Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005 ACM, 2005.

[17] Barreno, M. , et al. "Can machine learning be secure." Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS 2006, Taipei, Taiwan, March 21-24, 2006 ACM, 2006.

[18] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. Paper presented at 2nd International Conference on Learning Representations, ICLR 2014.

[19] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. Paper presented at 3nd International Conference on Learning Representations, ICLR 2015.

[20] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017, pp. 39-57, doi: 10.1109/SP.2017.49.

[21] Madry, Aleksander Makelov, Aleksandar Schmidt, Ludwig Tsipras, Dimitris Vladu, Adrian. Towards Deep Learning Models Resistant to Adversarial Attacks. Paper presented at 6nd International Conference on Learning Representations, ICLR 2018.

[22] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2016, pp. 582-597, doi: 10.1109/SP.2016.41.

[23] Generalized No Free Lunch Theorem for Adversarial Robustness, Elvis Dohmatob, Proceedings of the 36th International Conference on Machine Learning, PMLR 97:1646-1654, 2019.

[24] Badnets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. CoRR abs/1708.06733, 2017. 1, 4, 13

[25] Certified Adversarial Robustness via Randomized Smoothing, Jeremy Cohen, Elan Rosenfeld, Zico Kolter, Proceedings of the 36th International Conference on Machine Learning, PMLR 97:1310-1320, 2019.

[26] Jia, J., Cao, X., & Gong, N. Z. (2021). Intrinsic Certified Robustness of Bagging against Data Poisoning Attacks. Proceedings of the AAAI Conference on Artificial Intelligence, 35(9), 7961-7969. https://doi.org/10.1609/aaai.v35i9.16971

[27] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. arXiv preprint arXiv:1712.09665 (2017).

[28] Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. 2021. Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World. In Proceedings of the 29th ACM International Conference on Multimedia. 5307–5315.

[29] Stepan Komkov and Aleksandr Petiushko. 2021. Advhat: Real-world adversarial attack on arcface face id system. In 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 819–826.

[30] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. 2018. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In International Conference on Learning Representations.

[31] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in Proceedings of the 2016 acm sigsac conference on computer and communications security, 2016, pp. 1528–1540.

[32] S. Thys, W. V an Ranst, and T. Goedemé, “Fooling automated surveillance cameras: adversarial patches to attack person detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 0–0.

[33] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. 2020. Adversarial t-shirt! evading person detectors in a physical world. In European conference on computer vision. Springer, 665–681.

[34] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. 2018. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 52–68.

[35] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. 2021. Invisible Perturbations: Physical Adversarial Examples Exploiting the Rolling Shutter Effect. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14666–14675.

[36] Dinh-Luan Nguyen, Sunpreet S Arora, Yuhang Wu, and Hao Yang. 2020. Adversarial light projection attacks on face recognition systems: A feasibility study. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 814–815.

[37] Marrone, S., Casula, R., Orrù, G., Marcialis, G., Sansone, C. (2021). Fingerprint Adversarial Presentation Attack in the Physical Domain. In ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12666. Springer, Cham. https://doi.org/10.1007/978-3-030-68780-9\_42

[38] Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)