多模态生物核验与防伪
AI安全实战

讲师简介

冯月-中关村科金
AI安全攻防实验室 总监

2015年，组建数据部数据挖掘团队，设计金融社交模型，惠及千万用户。2017年，筹建人工智能研究院产品团队，先后孵化智能营销机器人、智能呼叫中心质检平台、智慧双录电子合约平台，全面助力金融行业智能化升级。现主攻AI可信与AI安全方向，新产品多模态生物核验与防伪平台可最大程度提升企业核验能力的安全性、便捷性、可扩展性，荣获国家多项认证支撑，多项国际比赛冠军头衔，日调用量破百万。

案例简介

- 根据2020年11月份公安部发布的人脸识别安全性调研来看，现有**人脸技术存在严重的安全风险**，犯罪分子利用获取的用户照片，通过电子屏翻拍攻击、2D纸张打印攻击、3D面具攻击、深度伪造攻击等行为可以轻松突破该技术的防线，2000款知名APP在此沦陷。

- 我们希望通过我们的技术来解决这些问题，因此，我们在年初成立了专项AI安全攻防实验室，设计了**多模态生物核验与防伪算法融合体系**，基于音视频融合的技术方案，通过融入声音技术，并有机地与现有视觉技术进行结合，来大幅提高生物核验与**防伪**的能力，解决现有安全隐患的同时，击穿用户使用限制。该技术的广泛应用，不仅可以完成用户级的防伪需求，还可以给政府智能监管带来**更多可能**，如网络黑户的蛀虫发现、公安舆情稽查赋能、银监会金融机构敏感信息保护等。

案例背景

1. 攻击形式层出不穷
2. 模型不具备防伪能力
3. 模型防伪能力无法迁移
4. 产品不具备防伪能力
5. 防伪是把双刃剑，业务发展的矛盾调和

目录

msup

一. 防伪的价值--行业痛点
二. 防伪的关键技术&框架性设计
三. 实战示例分享
四. 总结与展望

一、防伪价值

** 《中国金融行业分析报告》

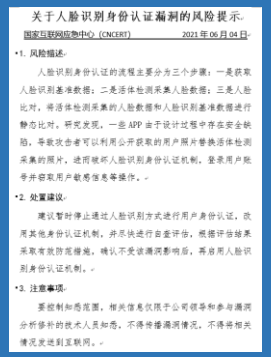
** 《公安部》官媒

1.1 国家监管部门重拳出击

- **公安部网安局** 2020年11月 发起《人脸识别与活体检测技术安全性评测要点》对200款知名app进行摸底，几乎全军覆没
- **工信部信通院** 2021年4月 发起《可信人脸应用守护计划》
- **国家互联网应急中心** 2021年6月 发布《人脸链路传输安全风险》
- **银监会** 2021年9月 发布《关于加强人脸识别技术应用安全管理的通知》

公安部网安局

网信办CNCERT


关于人脸识别身份认证漏洞的风险提示

国家互联网应急中心（CNCERT） 2021 年 06 月 04 日

•1. 风险描述

人脸识别身份认证的流程主要分为三个步骤：一是获取人脸识别基准数据；二是活体检测采集人脸数据；三是人脸比对，将活体检测采集的人脸数据和人脸识别基准数据进行静态比对。研究发现，一些 APP 由于设计过程中存在安全缺陷，导致攻击者可以利用公开获取的用户照片替换活体检测采集的照片，进而破坏人脸识别身份认证机制，登录用户账号并窃取用户敏感信息等操作。

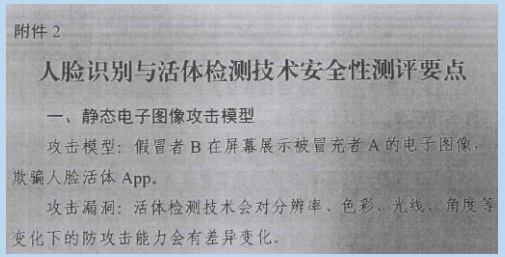
•2. 处置建议

建议暂时停止通过人脸识别方式进行用户身份认证，改用其他身份认证机制，并尽快进行自查评估，根据评估结果采取有效防范措施，确认不受该漏洞影响后，再启用人脸识别身份认证机制。

•3. 注意事项

要控制知悉范围，相关信息仅限于公司领导和参与漏洞分析修补的技术人员知悉，不得传播漏洞情况，不得将相关情况发送到互联网。

工信部信通院


附件 2

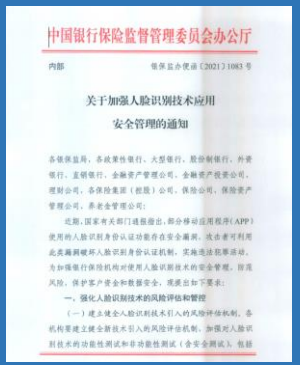
人脸识别与活体检测技术安全性测评要点

一、静态电子图像攻击模型

攻击模型：假冒者 B 在屏幕展示被冒充者 A 的电子图像，欺骗人脸活体 App。

攻击漏洞：活体检测技术会对分辨率、色彩、光线、角度等变化下的防攻击能力会有差异变化。

银监会


中国银行保险监督管理委员会办公厅

内部 银保监办便函〔2021〕1083 号

关于加强人脸识别技术应用安全管理的通知

各银保监局，各政策性银行、大型银行、股份制银行、外资银行、直销银行、金融资产管理公司、金融资产投资公司、理财公司，各保险集团（控股）公司、保险公司、保险资产管理公司，养老金管理公司：

近期，国家有关部门通报指出，部分移动应用程序（APP）使用的人脸识别身份认证功能存在安全漏洞，攻击者可利用此类漏洞破坏人脸识别身份认证机制，实施违法犯罪活动。为加强银行保险机构对使用人脸识别技术的安全管理，防范风险，保护客户资金和数据安全，现提出如下要求：

一、强化人脸识别技术的风险评估和管理

（一）建立健全人脸识别技术引入的风险评估机制。各机构要建立健全新技术引入的风险评估机制，加强对人脸识别技术的功能性测试和非功能性测试（含安全测试），根据

1.2 实际案例屡见不鲜--2019年

- 人脸识别在没有活体检测支持的前提下存在极高的危险性。
- 采用照片、视频、人脸编辑软件、人脸面具等形式，可频繁实施线上欺诈。

关键案例：丰巢快递柜被小学生“破解”，刷脸取件功能已取消

只要用一张打印照片就能代替真人刷脸，骗过小区里的丰巢智能柜，取出父母的包裹！

1.3 攻击升级--2020年

3D 构建流程

3D建模

3D打印

喷漆上色

- **A** 单张身份证照片 224*224
- **B** 单张正面高清图 1000*1000
- **C** 多张图片 正1 侧5
- **D** 视频 环绕头部 20s

此处有演示视频

00:00:00

1.4 攻击再升级--2021年

单张图

制作唇语攻击视频

此处有演示视频

00:00:00

制作动作活体视频

此处有演示视频

00:00:00

** 以上技术在以 ZAO、Avatarify 为代表等全民娱乐类应用上唾手可得

1.5 风险就在身边

■ 客户案例代表

■ 技术供应商代表

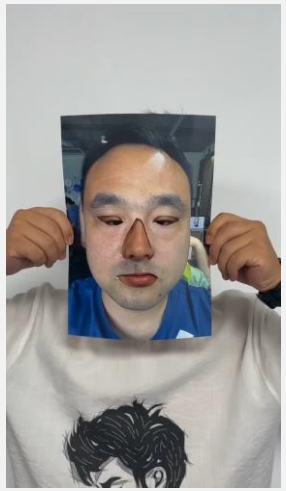
此处有演示视频

二、防伪的关键技术&框架性设计

1. 对攻击特质的了解
   a) 对攻击行为的了解
   b) 对正常客户的了解
2. 对操作空间的了解
   a) 挑战响应，音视频融合

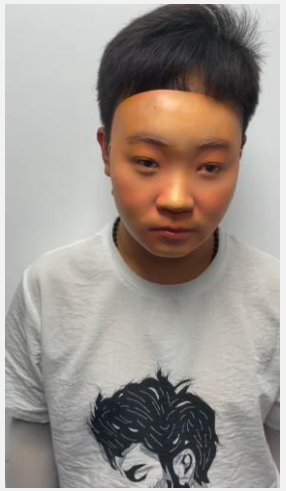
2.1 攻击行为的特质与弱点

■ 照片挖孔



■ 圆筒面具

■ 3D面具



■ 手机翻录

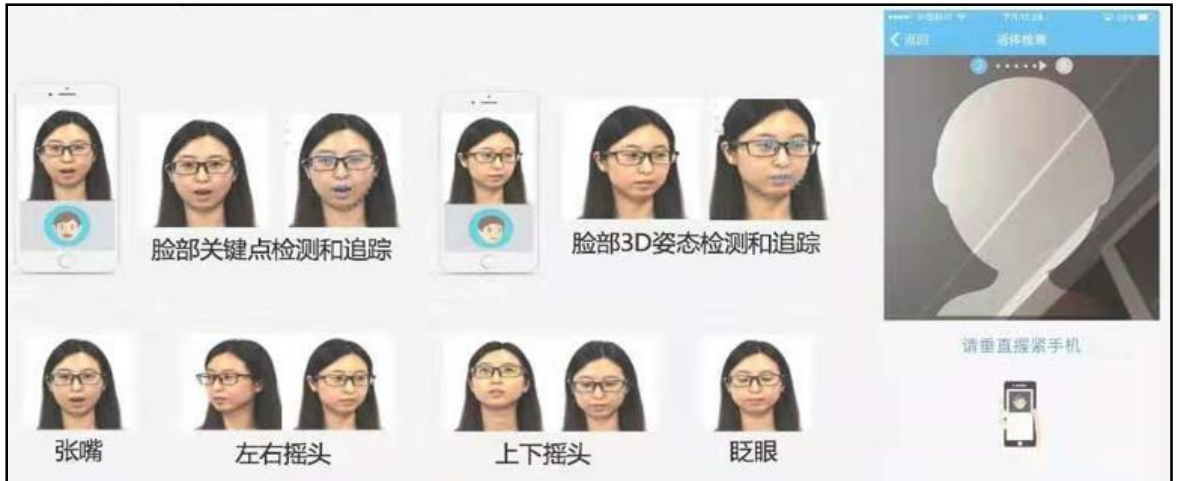
■ 深度伪造

① 攻击道具专业化
② 道具获取平民化
③ 呈现式欺骗性越来越强

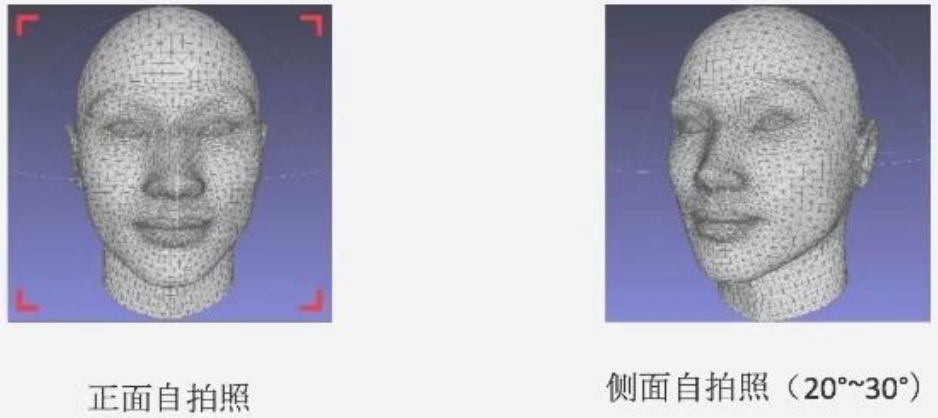
2.2 现有方案的特质与弱点

活体检测与人脸识别技术相结合

- 模型的脑容量只能接受特定的任务
- 该任务不包括多元复杂的伪造问题



动作活体检测
用户做出对应随机生成的动作



双角度活体检测
结合人脸多角度特征进行分析

炫彩活体检测
屏幕投影特殊光线或图案到用户脸部

2.3 关键技术与框架

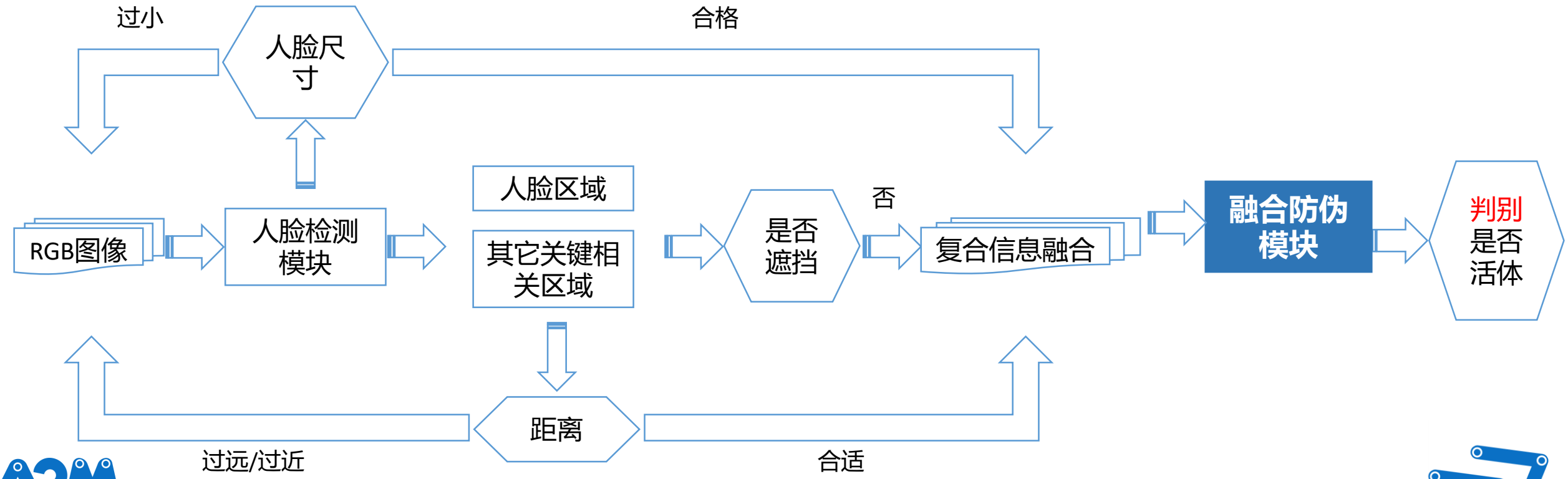
1. 找到显著的特征，构造相关数据
2. 找到能承载更多知识的神经网络，更有效的决策网路构建龙骨
3. 借助系统性的力量去覆盖模型的误识，让犯罪分子露出更多马脚

2.4 防伪方案龙骨

核心方案：

抽取不局限与人脸的相关特征复合式特征，借助深度神经网络的聚合能力，找出伪造和合法的明确边界

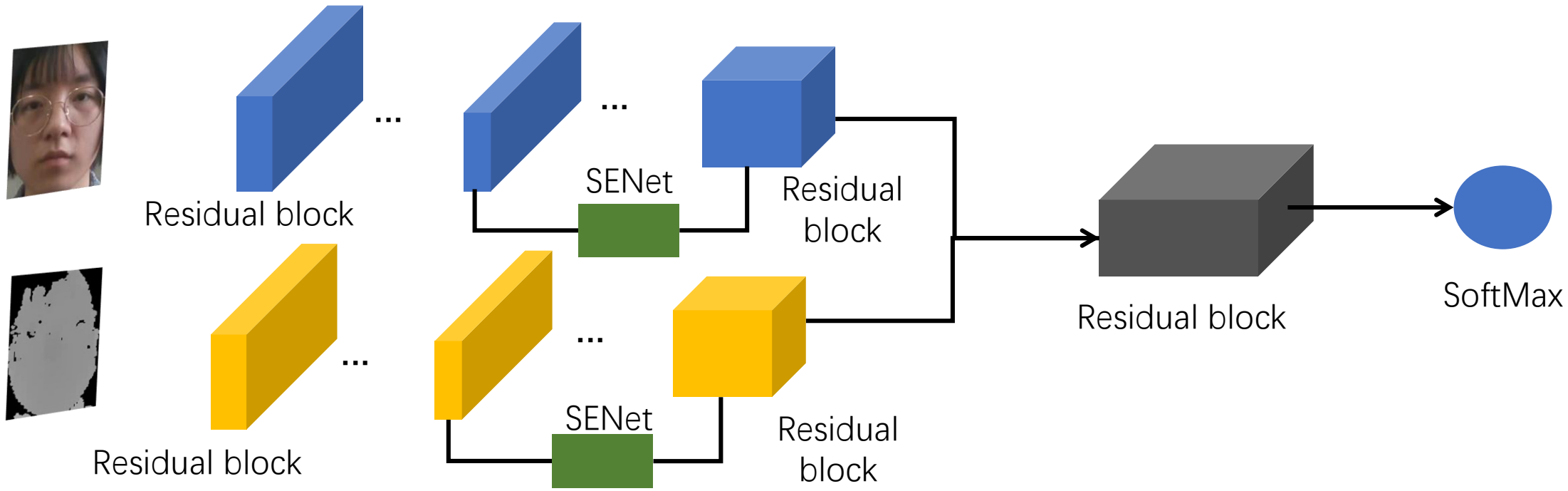
静默活体检测整体技术流程：



2.5.1 关键模块-核心知识承载网络选型1

防伪检测模块框架选择

基于深度学习的残差技术自主构建了以ResNet18网络为主题的框架1 → 评估结果：不合适



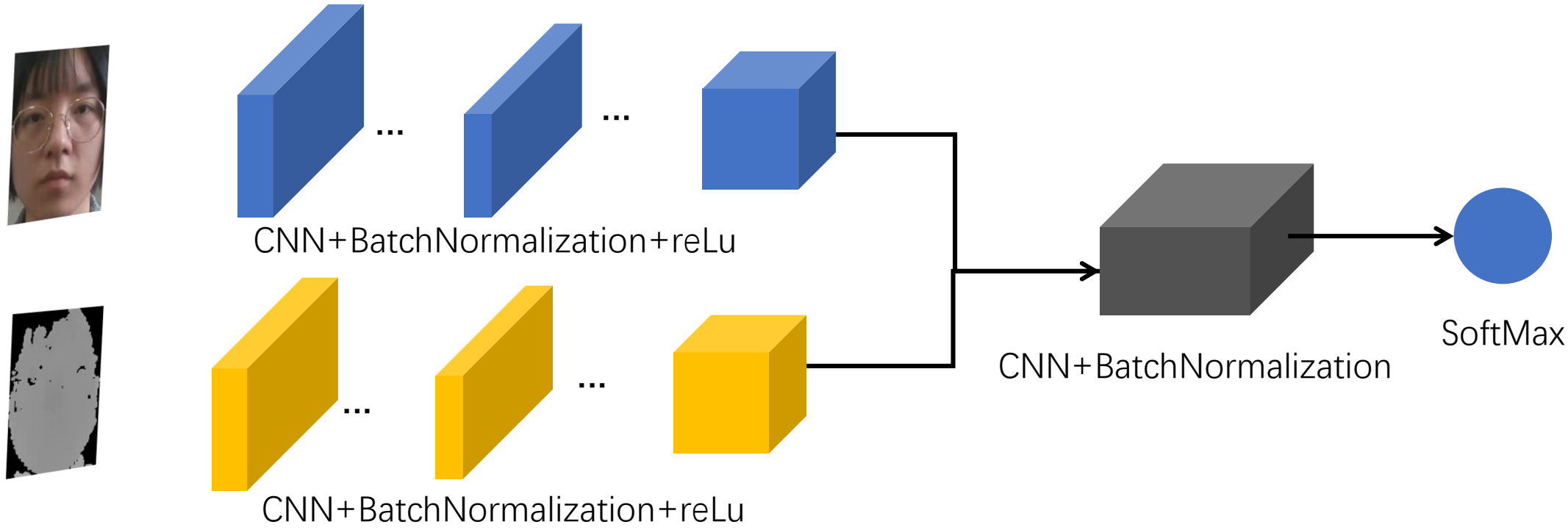
缺点：参数量过于庞大，不利于做嵌入式应用于移动端，泛化能力弱，识别速度慢。

模型在多分类情况下有过拟合迹象，故而研究对象还是转移到轻量级的网络。

2.5.2 关键模块-核心知识承载网络选型2

防伪检测模块框架选择

以卷积网络串行链接的方式自主搭建了VGG-6为主体的框架2 → 评估结果：不合适

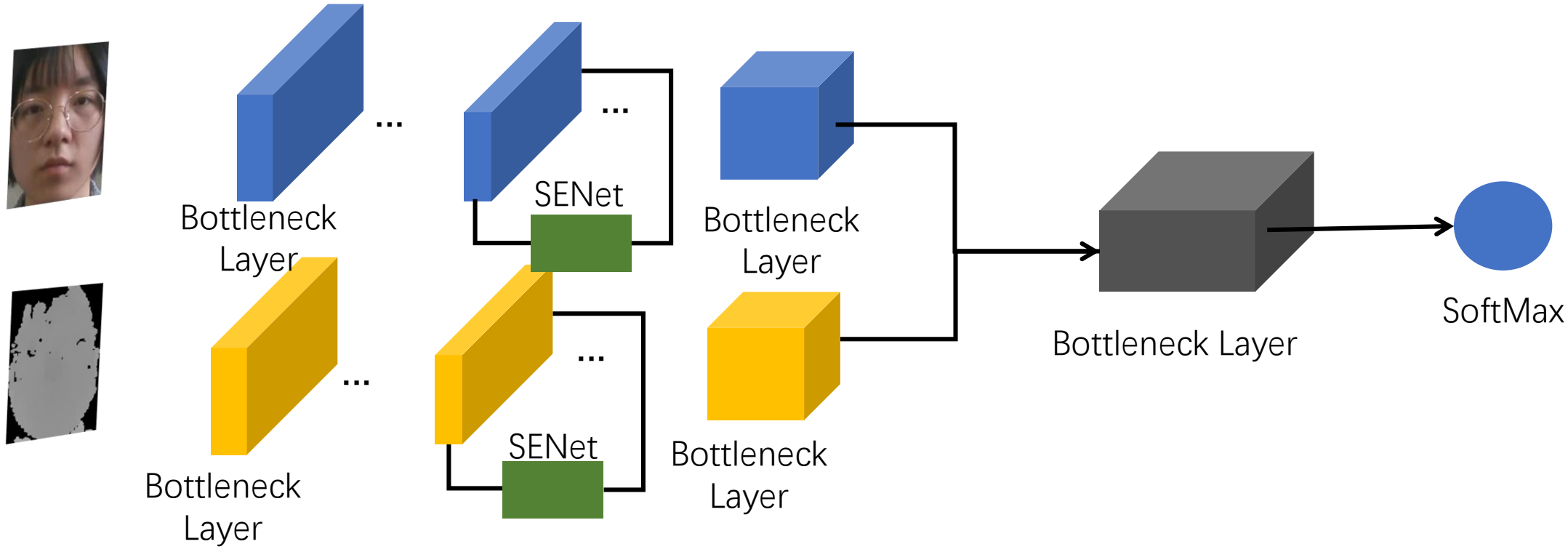


缺点：参数量虽小，可以进行嵌入式迁移，但是由于刻意改少卷积核的个数，导致学习的特征少，判断准确率低，网络泛化能力弱。

2.5.3 关键模块-核心知识承载网络选型3

防伪检测模块框架选择

基于深度可分离卷积自主构建以MobileNetV3网络为骨干的框架3 → 评估结果：比较合适



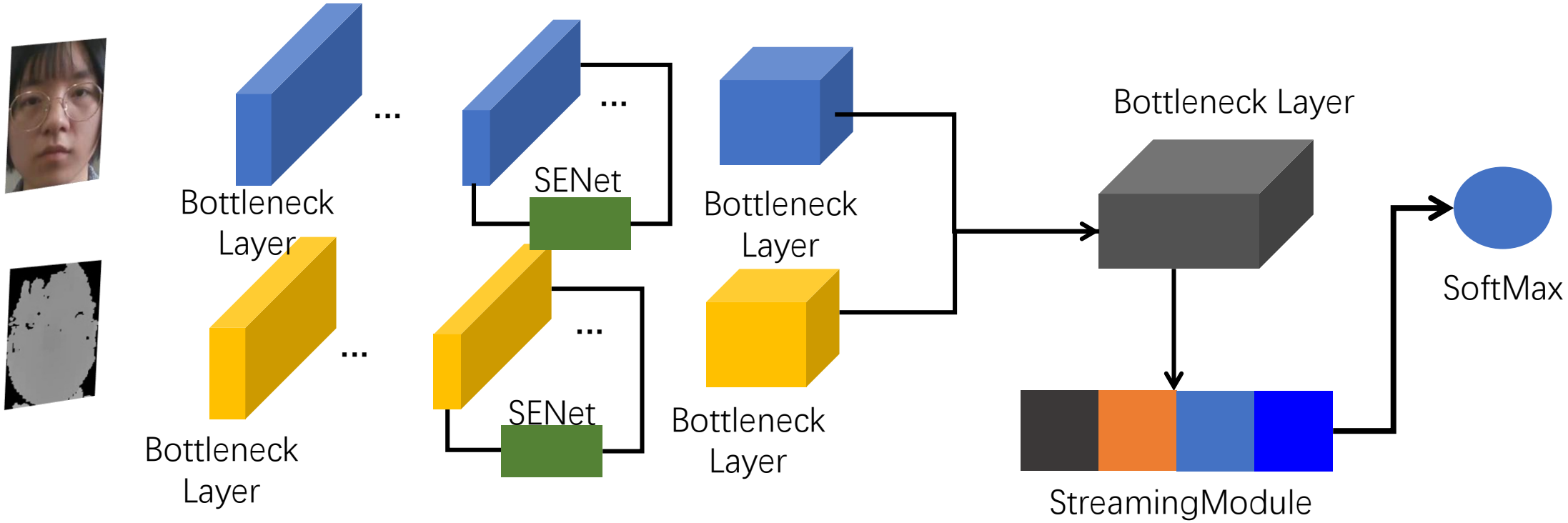
优点：参数量少，训练速度及预测结果的速度快，部署到嵌入式设备中能够有效降低延迟；

缺点：识别精度不够高，有待进一步提升。

2.5.4 关键模块-核心知识承载网络选型4

防伪检测模块框架选择

最终框架:MobileNetV3+StreamingModule → 在框架3的基础上添加光流模型。评估结果：合适



优点：参数量更少，在误识率很低的情况下，精度有了显著提升。

在**TPR@FPR=10e-4**的指标上达到了**1.0**

2.6 知识输入-数据构造与增广

数据类型及规格：包含真实活体和活体攻击两种类型，采集视频时长大多在25-30s

数据采集模式：不同光线、多样化活体攻击方式、不同姿势和位置

- 私有数据集
  - 真实活体
    - 前后移动 $25s$
    - 上下左右摆动 $25s$
    - 前后移动 + 上下左右 + 冷暖光照 $30s$
    - ⋮
  - 活体攻击
    - 高清照片 + 面具
      - 前后移动 $25s$
      - 上下左右弯曲 $25s$
      - 前后移动 + 上下左右 + 光照 $30s$
    - 手机屏幕录制
    - 电脑屏幕录制

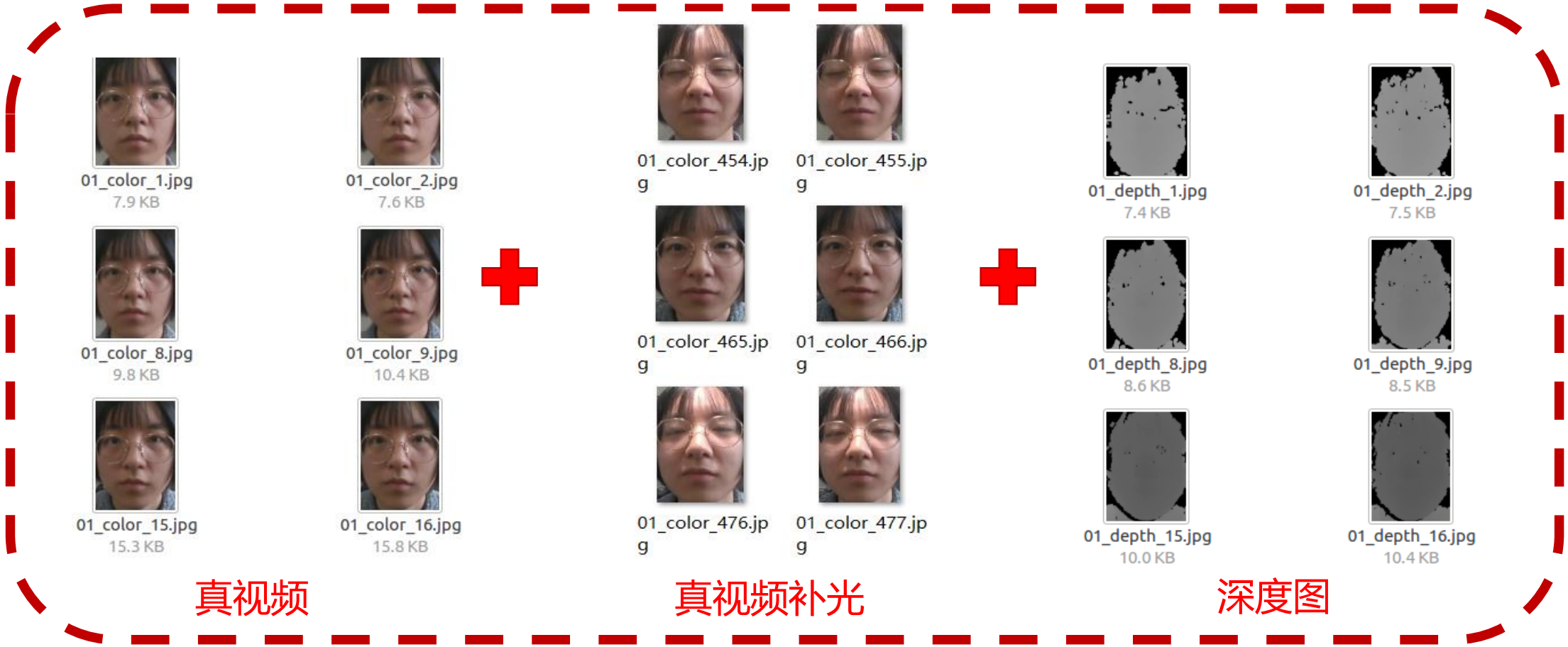
一方面，数据以25~30秒视频的形式采集能够极大程度的丰富数据集量级

另一方面，不同采集条件的综合能够全面覆盖所有活体攻击形式

2.6.1 知识输入-合法特征

私有数据构成-真人视频

活体数据采集样例



图片大小 112*112　　图片数量 约1万7千张 （11人、正脸、侧脸）

2.6.1 知识输入-非法特征

msup

私有数据构成-虚假视频

活体攻击数据采集样例


1-2-f-s_color_16.jpg


1-2-f-s_color_17.jpg


1-2-f-s_color_21.jpg


1-2-f-s_color_22.jpg


1-2-f-s_color_26.jpg


1-2-f-s_color_27.jpg

假视频


01-f-p_color_41.jpg


01-f-p_color_42.jpg


01-f-p_color_52.jpg


01-f-p_color_53.jpg


01-f-p_color_63.jpg


01-f-p_color_64.jpg

纸质打印


01-f-p_color_386.jpg


01-f-p_color_387.jpg


01-f-p_color_397.jpg


01-f-p_color_398.jpg


01-f-p_color_408.jpg


01-f-p_color_409.jpg

补光假视频


1-2-f-s_depth_166.jpg


1-2-f-s_depth_167.jpg


1-2-f-s_depth_175.jpg


1-2-f-s_depth_176.jpg


1-2-f-s_depth_184.jpg


1-2-f-s_depth_185.jpg

假的深度图

图片大小 112*112　图片数量 约4万张 （11人、正脸、侧脸）

2.7 效果验证-国际通用指标

防伪技术精度指标评估-**大数据、多样化、规范化**

- TP表示真阳性，表示判别模型True（成功地）判定出结果是Positive的
- TN表示真阴性，表示判别模型True（成功地）判定出结果是Negative的；
- FP表示假阳性，表示判别模型False（未能成功地）判定出结果是Positive的
- FN表示假阴性，表示判别模型False（未能成功地）判定出结果是Negative的。

FPR：

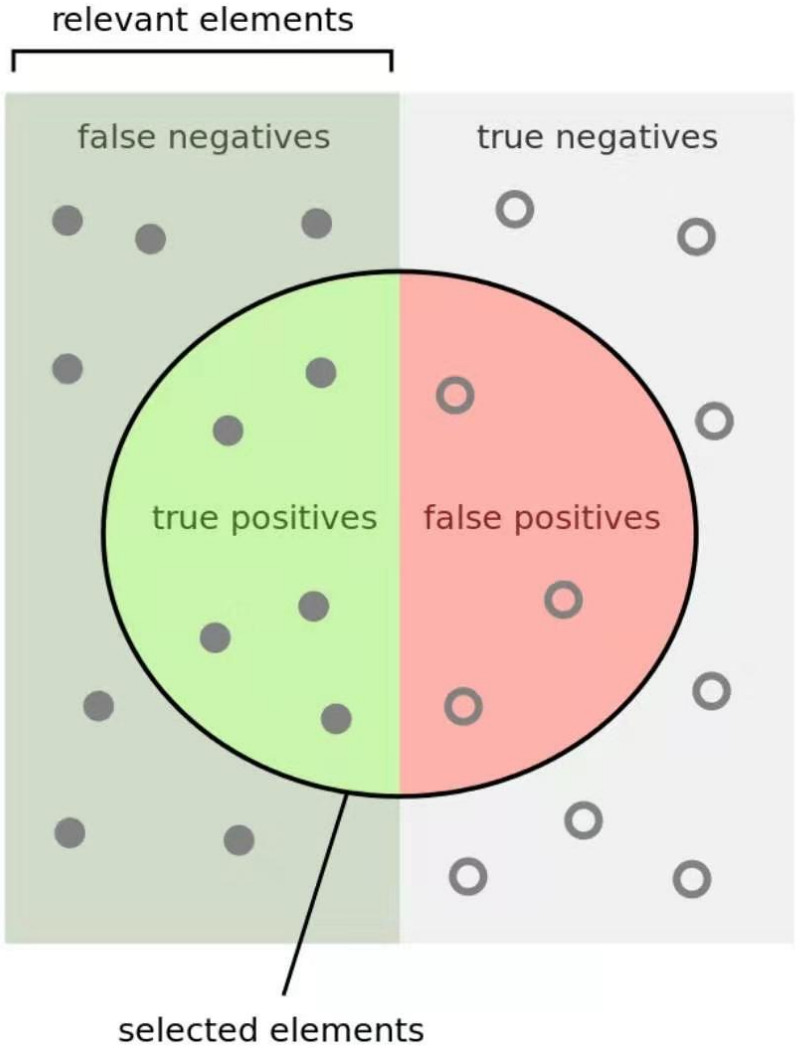
$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

召回率（TPR）：

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

准确度（ACC）：

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FP + TN + FN}$$



FPR表示当前被错误分到正样本类别中真实的负样本所占所有负样本总数的比例

2.7.1 量化结果

msup

评估结果分析-**覆盖多样化数据，高精度、低误识，平均准确率超过99%**

评估结果						
数据集	等错误率(EER)	TPR@FPR=10e-1	TPR@FPR=10e-2	TPR@FPR=10e-3	TPR@FPR=10e-4	准确率(ACC)
私有数据	0.00658	1.000	0.99983	0.996722	0.953697	0.99342
CASIA-SURF	0.00168	1.000	1.000	0.997662	0.961590	0.99832

TPR@FPR 表示FPR为某值的情况下召回率(TPR)的对应值

2.7.2 可视化结果

防伪检测系统

- 基于**单目摄像头**（可见光）
- 利用前后文信息
- 针对屏幕翻拍、打印照片、打印3D面具等攻击手段可以达到**99%以上的检出准确率**，反欺诈能力强
- 应用领域广泛，可应用到金融核身、基于人脸识别的门禁、闸机、卡口、自助酒店入住等场景
- 该系统**能够满足H5、移动端**线上活体检测的功能需求

静默活体检测示例

2.8 99%的实验室准确率在实用生产时无法达到监管要求

在无试错成本的情况下，攻击者仍能找到突破口

攻击次数	1次	2次	3次	…	20次	100次	500次
全部防御成功的概率	99%	98%	97%	…	92%	37%	1%

综合防控

1. 引入挑战-响应模式，通过数字的互动、炫彩的互动、动作的互动
2. 限制互动复杂度，如数字需大于6位、动作需多于2组
3. 限制设备、账户在规定时间内的尝试次数
4. 限制高危机型
5. 引入国密
6. 强制前后端分离，在server端进行判断
7. 引入多渠道认证机制
8. 建立人工审核机制

2.8.1 综合防御模块，不同融合方案效果

分项		组合方式	
		动作+炫彩活体	数字唇语
攻击方式 - 模拟炫光	纸质照片	全部拦截	全部拦截
	纸质照片挖孔	部分拦截	全部拦截
	电子图像	全部拦截	全部拦截
	电子视频	全部拦截	全部拦截
	面具头模	部分拦截	全部拦截
	面具头模挖孔	部分拦截	全部拦截
易用指标	通过性	高	高
	操作复杂度	低	中
	单次验证时长	8s	7s
	端侧资源消耗	1核170M	0.3核100M
	云侧资源消耗	暂无数据	TESLA M40:利用率55%/显存5.5G
	适配接入难度	低	低

三、实战示例分享1

攻击方案

- 收集被攻击者面部信息
- 制作3D头模
- 任意第三人攻击

防守方

招某银行 -- 失败

活体：面部动作
核验：人脸

农某银行 -- 失败

活体：面部动作
核验：人脸

我们-- 成功

活体：语音+唇语
核验：声纹+人脸

招某银行

此处有演示视频

00:00:00

农某银行

此处有演示视频

00:00:00

猫头鹰

此处有演示视频

00:00:00

实战示例分享2

真实用户 **吕小东** 男 26岁

攻击者 **胡美丽** 女 22岁

攻击方案

- 趁用户熟睡
- 使用纸片模拟唇部动作
- 任意第三人模仿声音

中某银行 -- 防守失败

活体：语音+张张嘴

核验：人脸

我们-- 防守成功

活体：语音+唇语

核验：声纹+人脸

中某银行

此处有演示视频

00:00:00

猫头鹰

此处有演示视频

00:00:00

实战示例分享3

msup

场景介绍

- 本人夜晚在家

Or 本人行走在黑夜的晚上

Or 本人演唱会或电影院

被试者

招某银行 -- 失败

活体：面部动作

核验：人脸

农某银行 -- 失败

活体：面部动作

核验：人脸

我们-- 成功

活体：语音 声纹 or 唇语 人脸

核验：

此处有演示视频

00:00:00

此处有演示视频

00:00:00

此处有演示视频

00:00:00

四、总结与展望

1. 监管趋势会从建议走向巡查
2. 机构对防伪态度从抗拒走向标配
3. 黑客技术魔高一丈，对抗样本攻击、数据投毒
4. 防伪不仅仅是活体的问题，合同签章、身份证上传、信贷抵赖、征信上报、电信诈骗 等问题中都有具体场景

4.1 从实验室到投产

msup

简化流程

- 无需用户做任何动作，简单拍摄2秒视频

活体攻击全方位可防控性

- 不单单可对视频、照片翻拍的2D攻击模式进行检测，3D面具、头套等攻击模式的防控同样支持

加强抗干扰能力

- 加强在光线、噪音、模糊等复杂条件下的抗干扰能力

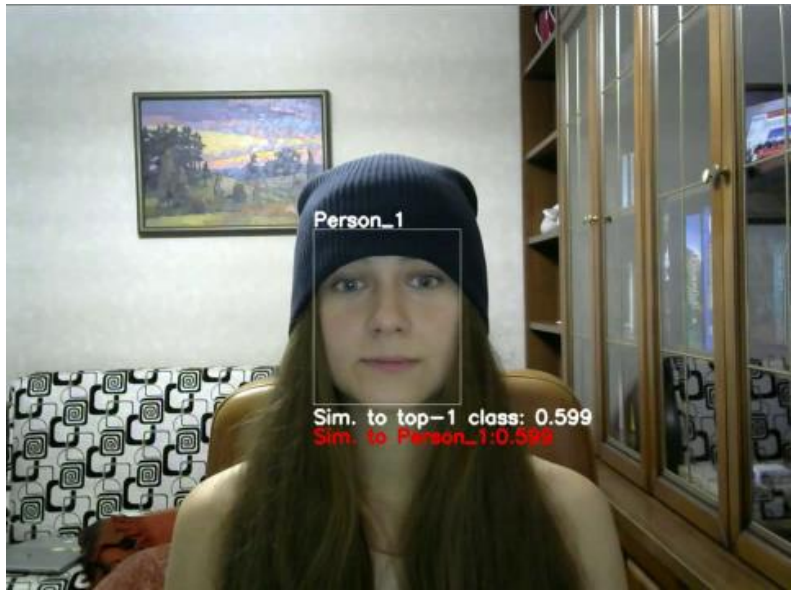
高精度识别

- 图像2D特征与3D特征相结合，增强特征表达能力

4.2.1 更多防伪问题

活体检测对抗攻击

- **伪造的特殊图案：在彩色打印机上打印特定的对抗样本，将其贴到的帽子上，对人脸识别系统进行攻击。**
- **伪造的干扰像素：通过对识别系统的对抗攻击，生成干扰方法**



2019年华为莫斯科实验室和莫斯科国立大学攻关中

2014年谷歌Szegedy等提出

2017年日本ATR和九州大学等提出

4.2.2 更多防伪问题

身份证防伪

4.2.3 抵赖防伪问题

语音伪造

关注msup公众号

获取更多AI落地实践

麦思博(msup)有限公司是一家面向技术型企业的培训咨询机构，携手2000余位中外客座导师，服务于技术团队的能力提升、软件工程效能和产品创新迭代，超过3000余家企业续约学习，是科技领域占有率第1的客座导师品牌，msup以整合全球领先经验实践为己任，为中国产业快速发展提供智库。