ICS 35.240
L 70

团　　体　　标　　准

T/AII 001-2021

人脸识别安全技术规范

Technical Specification for Face Recognition Security

2021-11-02 发布　　2021-11-15 实施

深圳市人工智能行业协会　发布

目　　次

前　言

本文件按照GB/T 1.1-2020《标准化工作导则一第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由深圳市人工智能行业协会提出并归口。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件负责起草单位：腾讯科技（上海）有限公司、腾讯云计算(北京)有限责任公司、深圳市人工智能行业协会、京东科技控股股份有限公司、深圳市华赛睿飞智能科技有限公司、上海计算机软件技术开发中心、上海交通大学、华东师范大学。

本文件主要起草人：黄飞跃、李季檩、丁守鸿、吴双、李博、毕明伟、姚太平、刘海涛、邓莹婷、郎丽艳、王辉、钱丽玲、林必毅、孙凯、陈敏刚、盛斌、贺樑、李博、王启立。

本文件为首次发布。

引　言

人脸识别（Face Recognition）是一种以人的面部特征信息为核心进行身份识别的技术。近年来，随着人工智能、计算机视觉、云计算等技术的迅速发展，人脸识别技术获得了长足的进步并日臻趋于完善。简单易用的人脸识别技术伴随着互联网的发展越来越多地被应用于社会的各行各业。与此同时，人脸识别系统也面临着潜在的安全风险，攻击者尝试用各种手段破坏人脸识别系统的安全稳定运行，严重威胁到了系统使用者的生命与财产安全。目前市场上有些人脸识别系统中嵌入了如人脸活体检测技术的安全模组，但并没有统一的安全技术规范，各个厂家自成体系，系统的安全性参差不齐，同时没有统一衡量标准，一定程度上阻碍了人脸识别发展。

本规范以保障人脸识别系统的安全性为目标，提出了人脸识别技术安全规范，明确了人脸识别应用算法框架中内容安全和基础安全两大安全组件。通过对安全威胁的细化分类，可以引导从事相关业务的单位对不同安全风险进行有针对性地防范，以提升人脸识别系统的安全性。同时，本标准的提出也有利于上级监管部门与不同公司在统一的安全威胁下进行评测对比。引导并促进人脸识别技术安全健康地发展。

人脸识别安全技术规范

1 范围

本文件规定了人脸识别系统在活体检测、数字合成内容取证、数据安全和模型安全等方面的技术要求。

本文件适用于人脸识别系统的设计、技术开发、测试和管理。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 20271-2006 信息安全技术 信息系统通用安全技术要求

GB/T 20273-2006 信息安全技术 数据库管理系统安全技术要求

GB/T 26238-2010 信息技术 生物特征识别术语

GB/T 29268.1-2012 信息技术 生物特征识别性能测试和报告 原则与框架（ISO/IEC19795-1）

GB/T 38671-2020 信息安全技术 远程人脸识别系统技术要

GB/T 35678-2017 公共安全 人脸识别应用 图像技术要求

GB/T 31488-2015 安全防范视频监控人脸识别系统技术要求

3 术语和定义

GB/T 20271-2006、GB/T 26238-2010和GB/T 29268.1-2012确立的以及下列术语和定义适用于本文件。

3.1

生物特征识别 biometric recognition

基于个体的行为特征和生物学特征，对该个体进行的自动识别。

注：“个体”限指人。

3.2

人脸识别 face recognition

以人脸特征作为识别人体身份的一种人体生物特征识别方法。其通过分析提取用户人脸图像数字特征产生样本特征序列，并将该样本特征序列与已存储的模板特征序列进行比对，用以识别用户身份。

3.3

人脸识别系统 face recognition system

实现人脸识别功能的专用信息处理系统。人脸识别系统可以是一个由独立软硬件构成的独立系统，也可以是在信息系统已有平台上运行的计算机系统。

3.4

人脸识别身份认证系统 Face-based Identity Authentication System

指采用人脸识别技术，提供人脸用户身份认证的系统。

3.5

人脸图像采集模块 face image capture module

人脸识别系统的一个模块，用于进行人脸图像采集。

3.6

用户 user

指人脸识别系统用以识别的对象。

3.7

用户登记 user enrollment

分析提取用户人脸图像数字特征、产生并存储模板特征序列的过程。

3.8

人脸验证 face verification

人脸识别应用之一，将所产生的样本特征序列与按用户标识信息所给定的已存储的用户的模板特征序列进行比对（1：1比对），以确认用户是否为所声明的身份。

3.9

人脸辨认 face identification

人脸识别应用之一，将所产生的样本特征序列与已存储的指定范围内的所有模板特征序列进行比对（1：N比对），确定用户身份。

3.10

候选者 candidate

通过用户辨认所确定的用户。该用户是在已进行过用户登记的所有用户中选出的符合当前样本特征序列数据要求的用户。

3.11

相似度 similarity

两个生物特性相似程度的一个实数；数值越大两个生物特性越相似。

3.12

阈值 threshold

做出判定所依据的边界值（或值集）。

3.13

错误接受率（FAR） false accept rate(FAR)

人脸验证过程中，发生错误接受的次数占冒充者比对总次数的比率，用百分比表示。

3.14

错误拒绝率（FRR） false reject rate(FRR)

人脸验证过程中，发生错误拒绝的次数占真实人比对总数的比率，用百分比表示。

3.15

误识 false alarm

人脸辨认过程中，非目标人被判为目标人。

3.16

辨认识别率 detection and identification rate

人脸辨认过程中，正确判定目标人身份次数占目标人出现总次数的比率，用百分比表示。

3.17

误识率 false alarm rate

人脸辨认过程中，错误识别次数占非目标人出现总次数的比率，用百分比表示。

4 人脸识别安全技术架构

4.1 人脸识别应用系统架构

人脸识别应用包含了人脸数据采集、人脸数据处理、人脸注册数据、人脸特征提取和人脸数据对比等重要模块，其中本规范涉及到的安全模块包括了内容安全和基础安全。人脸识别应用系统架构见图 1。

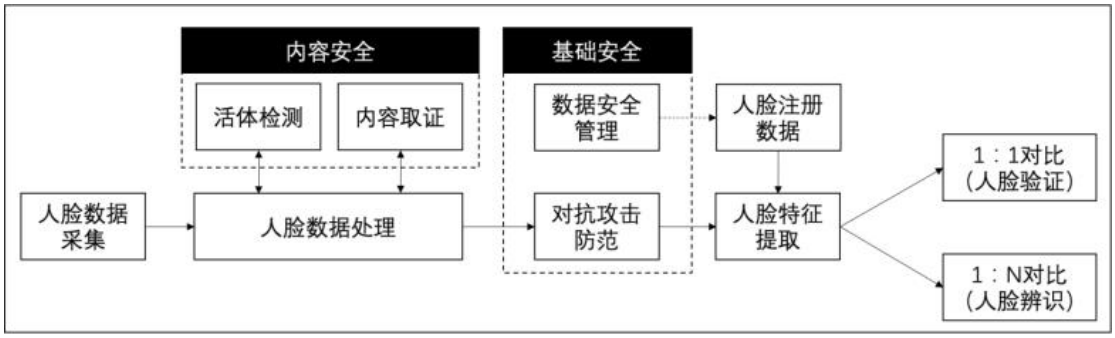

图 1 人脸识别应用系统框图

1） 人脸数据采集：此模块负责对人脸以及其它多因子生物信息进行采集；

2） 人脸数据处理：主要负责对人脸数据采集模块采集的信息进行处理，从而筛选出质量较高的人脸用于后续的识别，一般包含以图像处理等算法为基础的活体检测与内容取证；

3） 人脸特征提取：此模块负责对人脸及其它多因子生物信息进行特征的抽取与加工；

4） 人脸比对：此为服务端核心算法，对两张人脸提取到的特征进行相似度对比，一般包含 1:1 对比用于人脸验证和 1：N 对比用于人脸辨识。

5） 内容安全：应提供活体检测和内容取证功能；

6） 基础安全：应提供数据安全管理功能和对抗攻击防范功能。

4.2 人脸识别安全技术架构

本文涉及的人脸安全功能包含内容安全和基础安全两部分，其中内容安全部分包含了活体检测和内容取证两大模块；基础安全包含了数据安全和模型安全两大模块。人脸安全技术架构见图 2。

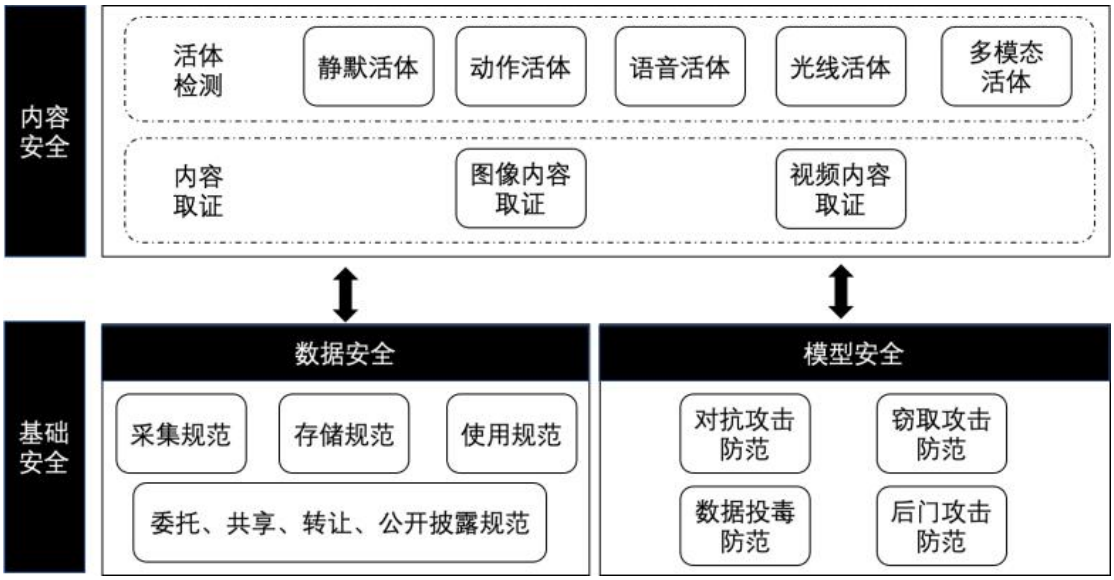

图 2　人脸识别安全技术架构

1）　活体检测：应提供静默活体、动作活体、语音活体、光纤活体和多模态活体检测功能；

2）　内容取证：应提供图像内容取证和视频内容取证功能；

3）　数据安全：应提供采集、存储、使用、委托、共享、转让、公开披露各个阶段的数据安全功能；

4）　模型安全：应提供对抗攻击防范、窃取攻击防范、数据投毒防范和后门攻击防范功能。

5　活体检测

5.1 活体检测概述

活体检测指在一些身份验证场景中确定对象真实生理特征的方法。在人脸识别应用中，一般使用摄像头拍摄及采集人脸图像的同时，结合云端人脸防伪检测技术，判断图像是否来自真人，有效防范纸质图片、屏幕翻拍、面具模型、合成换脸等类别的攻击，保证人脸信息的真实有效性。

5.2 活体检测手段

5.2.1 静默活体检测

静默活体是指无需检测主体做主动式反应，直接基于采集的图片或视频内容进行分析，包括但不限于背景、光线、色差等，最终作出活体判断。

人脸识别应用系统应支持静默活体检测。

5.2.2 动作活体检测

动作活体指根据检测主体的主动式反应进行活体检测的方式。一般通过指令要求客户进行相关操作并判断人脸的真实有效性，指令动作可以包括但不限于点头、抬头、左右转头、

张嘴、眨眼等。

人脸识别应用系统应支持动作活体检测。

5.2.3 光线活体检测

光线活体指通过驱动屏幕颜色变化并采集该过程中的人脸响应信息来检测是否为活体的技术。

人脸识别应用系统应支持光线活体检测。

5.2.4 多模态活体检测

多模态活体（深度，红外，RGB，声纹，唇语）是指同时利用多种模态信息进行组合式的活体检测，应支持但不限于以下模态：

a）3D 活体检测：根据检测主体的 3D 深度信息进行活体检测，检测主体的轮廓信息是否为真实有效的主体。

b）红外活体检测：根据检测主体的近红外成像进行活体检测，直接利用不同材质对近红外光的吸收/反射率的不同导致的成像差异进行判断。

c）热红外活体检测：根据检测主体的体温特征进行活体检测，直接利用热成像设备获取检测主体的面部温度分布图进行判断。

d）RGB 活体检测：根据检测主体的 RGB 图像进行活体检测，通过人脸的外观图像特征以及底层纹理特征进行判断。

e）唇语活体检测：根据检测主体的主动式反应进行活体检测，通过指令要求主体读出屏幕上显示的数字进行唇音一致性判断等方法来判断主体是否为真实有效的人脸。

f）声纹活体检测：根据检测主体的声纹特征进行活体检测，通过指令要求主体读出屏幕上显示的数字等交互方式提取主体的声纹特征进行判断。

g）超声波活体检测：根据检测主体的主动式反应进行活体检测，通过指令要求主体作出一些动作，如摇头、点头、张嘴等，利用终端设备发射出的超声波的反射变化进行判断。

以上这些活体检测可以根据需要进行优选择的组合使用，以期达到最优用户体验与检测效果。人脸识别应用系统应支持多模态活体检测。

6 数字合成内容取证

6.1 概述

数字合成内容取证是指在人脸识别系统中对输入的人脸图像或视频进行鉴别，判断数字内容的真实性、完整性和原始性，应支持但不限于对以下人脸编辑方法进行有效鉴别：

a）人脸生成：基于生成对抗网络生成不存在的人脸内容，常用的方法有 StyleGAN，PGGAN 以及人工 photoshop 编辑等;

b）人脸替换：替换主体的人脸内容为另一主体的人脸内容，常用的方法有 Deepfake, FaceSwap 以及人工 photoshop 编辑等;

c）人脸属性编辑：对人脸的属性内容（肤色、性别、年龄等）进行修改，常用的方法有 StarGAN 以及人工 photoshop 编辑等;

d）人脸表情编辑：对主体的面部表情进行编辑，例如迁移另一主体的表情，常用的方法有 Face2Face 以及人工 photoshop 编辑等。

数字合成内容取证应支持图像内容取证和视频内容取证。

6.2 取证方法

6.2.1 图像内容取证

图像内容取证是指对输入的图像内容进行分析，应支持但不限于色彩、纹理和频域分布等，来判断该图像内容是否为原始真实内容

6.2.2 视频内容取证

视频内容取证是指对输入的视频内容进行分析，应支持利用多帧时序信息来判断该视频内容是否为原始真实内容。

7 模型安全防御规范

7.1 模型安全性概述

本章所指模型安全为人脸识别系统中的人工智能算法从模型训练到模型使用应具备的安全性能，需效防御投毒、后门、对抗、窃取等多种攻击，防范攻击能力等应至少满足以下安全要求。

7.2 常见攻击防范

7.2.1 对抗攻击防范

对抗样本攻击通过精心设计的输入样本误导人工智能算法。对抗样本为在正常人脸图像上添加人眼不易分辨的扰动，欺骗人工智能算法。对抗攻击可以发生在数字空间，也可以通过物理手段在真实世界构造。人脸识别系统中的人工智能算法应有效防御对抗样本攻击，包括但不限于 FGSM、DeepFool、C/W、PGD等算法生成的以及通过查询或者迁移方式生成的黑盒对抗样本攻击。可通过安全防御措施提升人工智能算法鲁棒性，有效抵御对抗攻击，包括但不限于以下方式：

a）人工智能算法模型训练阶段可将对抗样本加入训练数据集，构建高容忍扰动的人工智能算法， 有效抵抗对抗样本攻击；

b）人工智能算法模型使用阶段可添加对抗样本检测模块，有效过滤对抗样本；

c）人工智能算法模型可采用网络蒸馏技术降低模型对微小扰动的敏感度，提高模型鲁棒性；

d）应尽量避免直接使用常见的神经网络结构作为算法模型的骨干网络，提升模型对迁移性攻击的鲁棒性。

7.2.2 窃取攻击防范

人工智能算法应对模型窃取攻击进行有效防范，避免模型泄露，应满足以下安全性要求：

a）应加强人工智能算法关于训练数据、模型的传输、存储等方面的安全管理，防范模型被窃取；

b）应保障人工智能算法训练步骤安全，防范训练过程被窃取导致训练信息泄露，进而恶意构建相似模型；

c）应限制算法模型的查询次数，只输出模型结果，隐藏模型输出细节；

d）人工智能算法模型应具有辨识度，一旦出现恶意伪构模型，应能够通过合适途径辨别真伪，如模型添加水印等方法。

7.2.3 数据投毒防范

投毒攻击是指在模型训练时通过污染部分训练数据，从而达到恶意操纵模型的攻击。人工智能算法应对投毒攻击进行防范，有效挑拣出训练数据中的药饵数据，避免训练数据被投毒。人工智能算法为有效防范投毒攻击应满足以下安全性要求：

a）应确保训练数据来源可信、可靠，避免采集到污染数据；

b）应加强训练数据存储、使用等环节安全管理，防范训练数据被投毒；

c）应确保训练数据分布合理，防止污染攻击数据点混入，造成模型倾斜等错误状况；

d）宜采用回归分析等方法利用训练数据的特性检测与过滤数据集中的噪声和异常值，防止数据投毒攻击；

e）宜采用集成分析等方式通过采用独立训练的多个子模型综合结果提升人工智能算法抗投毒攻击的能力；

f）不宜使用完全公开的模型进行迁移学习，提升投毒攻击的难度。

7.2.4 后门攻击防范

后门攻击是一种新兴的针对人工智能算法模型的攻击方式攻击者会在模型中植入后门，使得模型被感染。通常情况下模型表现正常，但当后门被激活时，模型的输出将变为攻击者预先设置的恶意目标。人脸识别系统中的人工智能算法模型应有效防御后门攻击，包括但不限于以下手段：

a）在算法应用时，宜对输入预处理，过滤掉能触发后门的输入，降低输入触发后门、改变模型判断的风险，防范后门攻击。

b）可采用模型剪枝、再训练技术对模型进行重建，破坏模型中可能埋藏的后门。

8 数据安全

8.1 数据安全概述及基本要求

本章所指的数据安全为人脸识别系统安全中的重要基础组成部分，描述了数据控制者在人脸相关数据的采集、存储、使用以及委托、共享转让和公开披露等过程中应该遵守的安全规范。数据安全的基本要求包括：

a）处理人脸识别数据时应遵循最小必要原则。

b）不应收集未授权自然人的人脸图像。

c）应具备与其所处理人脸识别数据的数量规模、处理方式等相适应的数据安全防护和个人信息保护能力。

8.2 处理规范

8.2.1 采集规范

a）采集人脸识别数据时，应向数据主体告知收集规则，包括但不限于收集目的、数据类型和数量、处理方式、存储时间等，并征得数据主体明示同意。

b）用于采集人脸识别数据的设备应遵循相关标准要求。

c）在满足应用场景安全要求前提下，应仅收集用于生成人脸特征所需的最小数量、最少图像类型的人脸图像。

d）采集得到的人脸识别数据应满足以下质量要求：

表1　人脸图像质量要素

序号	因素	类型
1	人脸姿态	包括但不限于大小、角度、完整度等
2	图像质量	包括但不限于分辨率、清晰度等
3	光线条件	包括但不限于正常光、强光、弱光、逆光等
4	场景条件	包括但不限于室内、室外等多场景等

表2　人脸视频质量要素

序号	因素	类型
1	人脸姿态	包括但不限于大小、角度、完整度等
2	视频质量	包括但不限于分辨率、清晰度、帧率、时长、稳定性等
3	光线条件	包括但不限于正常光、强光、弱光、逆光等
4	场景条件	包括但不限于室内、室外等多场景等
5	动作条件	包括但不限于静止、眨眼、张嘴、摇头、点头、晃动等

e）人脸图像、视频数据质量应满足包括但不限于表 3 所示要求:

表3 人脸数据质量要求

序号	因素	具体要求
1	图像尺寸	分辨率≥640*480
2	人脸大小	人脸区域大小≥100*100
		两眼瞳间距应≥60，宜≥90
3	清晰度	高斯模糊<0.24
		运动模糊<0.15
		拉普拉斯方差≥500
4	姿态	水平转动角（°）∈[-20，20]
		俯角（°）< 20
		仰角（°）> -20
		倾斜角（°）∈ [-20，20]
5	完整度	几何失真度≤5%
		眉毛可见度=100%
		眼睛可见度=100%
		鼻子可见度=100%
		嘴巴可见度=100%
		面颊皮肤可见度=100%

表3 人脸数据质量要求(续)

序号	因素	具体要求
6	保真度	无过渡化妆
7	光照	光照均匀，对比度适中
		无光斑和阴阳脸
		整体无过曝和欠曝
		灰度级=256
8	表情	无过渡夸张表情
	注 1：人脸姿态的定义参考 GB/T35678-2017,3.3。 注 2：特殊应用，如居民身份证数字相片、护照相片标准参考相关标准和规定。 注 3：人脸样本整体模糊程度的计算可参考 GB/T 33767.5-2018,7.4.7 或《Diatom autofocusing in brightfield microscopy: a comparative study》。	

8.2.2 存储规范

a）在未经过授权同意的情况下，不应该存储人脸图像等数据。

b）在授权到期或撤回授权的情况下，应删除人脸数据或进行匿名化处理。

c）应采取安全措施存储和传输人脸识别数据，包括但不限于加密存储和传输人脸识别数据，采用物理或逻辑隔离方式分别存储人脸识别数据和个人身份信息等。

8.2.3 使用规范

a）应在完成验证或辨识后立即删除人脸图像。

b）应对提取的人脸特征进行加密处理，提升特征的不可逆性，即难以从特征中逆向恢复出人脸图像。

8.2.4 委托、共享、转让、公开披露规范

1) 不应公开披露人脸识别数据，原则上不应共享、转让人脸识别数据。因业务需要，确需共享、转让的，应按照 GB/T CCCCC《个人信息安全影响评估指南》开展安全评估，并单独告知数据主体共享或转让的目的、接收方身份、接收方数据安全能力、数据类别、可能产生的影响等相关信息，并征得数据主体的书面授权。
2) 原则上不应进行委托处理，确需委托处理的，应在委托处理前审核受委托者的数据安全能力，并对委托处理行为开展个人信息安全影响评估。

9 性能指标

9.1 活体性能指标

真人通过率：测试用例为真人，被正确判断为真人的概率
攻击误过率：测试用例为攻击，但被误判为真人的概率
活体性能指标要求如表4所示：

表 4　活体性能指标要求

活体项目	一级安全指标		二级安全指标	
	真人通过率	攻击误过率	真人通过率	攻击误过率
活体类型	≥95%	＜1%	≥95%	＜0.1%

9.2 内容取证性能指标

数字内容取证应能防范多种人脸内容篡改方法，包括但不限于人脸生成，人脸替换，人脸属性编辑和人脸表情编辑等。内容取证指标为对真人和攻击的整体鉴别准确率

数字内容取证性能指标要求如表 5 所示。

表 5　数字内容取证性能指标要求

数字内容取证	一级安全指标	二级安全指标
图像内容取证	准确率≥90%	准确率≥85%
视频内容取证	准确率≥95%	准确率≥90%

9.3 模型安全指标

a）对抗攻击防御指标

以人脸识别系统中 FAR 为一百万分之一时采用的相似度得分阈值为基准，如果攻击后的人脸与目标人脸之间的相似度大于该阈值，则认为攻击成功，反之则不成功，攻击成功的人脸对抗样本与全部测试样本之间的比值为攻击成功率 ASR (Attack Successful Rate)。

其中，数字域对抗攻击白盒成功率<50%；

黑盒物理攻击在现实人脸识别应用场景中威胁更大，具体安全指标见下表：

表 6　物理对抗防御性能指标要求

攻击区域	攻击区域示例	黑盒查询攻击（数值类型，查询次数小于 100）	黑盒查询攻击（决策类型，查询次数小于 100）	黑盒迁移攻击（以常见神经网络模型如 Resnet、VGG、Densenet 等作为替代模型）
额头（攻击面积在该区域占比<80%）		攻击成功率<1%	攻击成功率<0.5%	攻击成功率<0.1%

表 6　物理对抗防御性能指标要求（续）

攻击区域	攻击区域示例	黑盒查询攻击（数值类型，查询次数小于100）	黑盒查询攻击（决策类型，查询次数小于 100）	黑盒迁移攻击（以常见神经网络模型如 Resnet、VGG、Densenet 等作为替代模型）
眼睛（攻击面积在该区域占比<50%）		攻击成功率<1%	攻击成功率<0.5%	攻击成功率<0.15%
鼻子（攻击面积在该区域占比<100%）		攻击成功率<0.3%	攻击成功率<0.1%	攻击成功率<0.05%
面颊（攻击面积在该区域占比<80%）		攻击成功率<1%	攻击成功率<0.5%	攻击成功率<0.1%
人脸背景（攻击面积在该区域占比<100%）		攻击成功率<0.5%	攻击成功率<0.2%	攻击成功率<0.1%
鼻子，面颊组合区域（攻击面积在该区域占比<90%）		攻击成功率<3%	攻击成功率<1%	攻击成功率<0.5%

b）窃取攻击防御指标

被窃取模型性能下降>50%。

c）数据投毒攻击防御指标

数据投毒攻击成功率<20%。

d）后门攻击防御指标

后门攻击成功率<5%。

附 录 A
（资料性）
性能测试参考方案

A.1 活体测试

活体测试攻击类型包括但不限于二维假体攻击类型，三维假体攻击类型及劫持注入攻击类型。

测试方法：

真人通过率测试： N个真人，每个真人在正常环境下配合完成活体交互M次（测试过程可以选择不同的场景），活体通过次数记为P，即真人通过率TPR =P/（N*M），整体指标应符合真人通过率>95%的性能指标。实际测试过程中一般建议真人超过100人，测试总量超过1000次，真人图像应满足数据质量要求；

攻击误过率测试：构造二维假体素材并黑盒攻击N次（包含二维静态纸质图像攻击N1,二维静态电子图像N2,二维动态图像N3,等。 N = N1+N2+N3+...），构造三维假体素材并黑盒攻击M次（包含三维面具攻击M1次和三维头模攻击M2次等，M= M1+M2+...）,构造劫持注入类型素材并黑盒攻击P次（具体注入素材类型可参考下表内容及呈现方式等P1，P2，P3...,P=P1+P2+P3）;其中攻击总次数为N＋M＋P，记攻击成功的次数为F，则攻击误过率FAR＝F／（N＋M＋P），整体性能指标应符合一级安全指标误过低于 1 %，二级安全指标误过率低于0.1%的要求。考虑到攻击素材的成本差异较大（三维攻击素材成本高于注入攻击成本，远高于二维攻击素材），一般构造攻击类型的数量分布推荐为N：M：P=50：1：10，且总量不低于3000次。

A.1.1 二维假体攻击

二维假体攻击包括但不限于二维静态纸质图像攻击、二维静态电子图像攻击和二维动态图像攻击。

a）防范二维静态纸质图像攻击时，应考虑的因素包括但不限于表A.1所示内容；

表A.1 二维静态纸质图像攻击

序号	因素	类型
1	人脸图像材质	包括但不限于打印纸、亚光相纸、高光相纸、绒面相纸、哑粉、光铜等
2	颜料	包含但不限于黑白、彩色、喷墨、激光、印刷、银盐冲印、绘画等
3	人脸图像质量	包括但不限于分辨率、清晰度、大小、角度、光照条件、完整度等
4	呈现方式	包括但不限于距离、角度、移动、弯曲、折叠等
5	裁剪方式	包括但不限于图像是否抠除眼部、鼻子、嘴巴等
6	光线条件	包括但不限于正常光、强光、弱光、逆光等

b)防范二维电子图像攻击时，应考虑的因素包括但不限于表A.2所示内容;

表A.2　二维电子图像攻击

序号	因素	类型
1	二维电子图像类型	包括但不限于拍摄数字图像、翻拍数字图像等
2	显示设备类型	包括但不限于手机、平板电脑、电脑等
3	显示设备能力	包括但不限于分辨率、亮度、对比度等
4	二维电子图像质量	包括但不限于分辨率、清晰度、人脸大小比例等
5	呈现方式	包括但不限于距离、角度、移动等
6	光线条件	包括但不限于正常光、强光、弱光、逆光等

c)防范二维动态图像攻击时，应考虑的因素包括但不限于表A.3所示内容。

表A.3　二维动态图像攻击

序号	因素	类型
1	二维动态图像类型	包括但不限于录制视频、合成视频等
2	显示设备类型	包括但不限于手机、平板电脑、电脑等
3	显示设备能力	包括但不限于分辨率、亮度、对比度等
4	二维动态图像质量	包括但不限于分辨率、清晰度、帧率、人脸大小比例、持续时间等
5	呈现方式	包括但不限于距离、角度、移动等
6	光线条件	包括但不限于正常光、强光、弱光、逆光等

A.1.2　三维防假体攻击

三维假体攻击包括但不限于三维面具攻击和三维头模攻击：

a）防范三维面具攻击时，应考虑的因素包括但不限于表A.4所示内容；

表A.4　三维面具攻击

序号	因素	类型
1	面具材质	包括但不限于塑料面具、3D纸张面具、硅胶面具等
2	呈现方式	包括但不限于距离、角度、移动等
3	光线条件	包括但不限于正常光、强光、弱光、逆光等
4	裁剪方式	包括但不限于面具是否抠除眼部、鼻子、嘴巴等
5	光线条件	包括但不限于正常光、强光、弱光、逆光等

b）防范三维头模攻击时，应考虑的因素包括但不限于表A.5所示内容。

表A.5　三维头模攻击

序号	因素	类型
1	头模材质	包括但不限于泡沫、树脂、全彩砂岩、石英砂等
2	呈现方式	包括但不限于距离、角度、移动等
3	光线条件	包括但不限于正常光、强光、弱光、逆光等

A.1.3　劫持注入类型攻击

劫持注入攻击类型一般是指通过hook手机劫持绕过摄像头注入一段伪造的视频

防范劫持注入类视频攻击包括但不限于表A.6所示内容：

表A.6　合成视频攻击

序号	因素	类型
1	视频内容	包括但不限于多端非实时拍摄真人视频，剪辑拼接视频，人脸融合视频，人脸驱动视频等
2	呈现方式	包括但不限于距离、角度、移动等
3	光线条件	包括但不限于正常光、强光、弱光、逆光等

A.2　内容取证测试

测试方法：

真人通过率测试： N个真人，每个真人在正常环境下配合拍摄完成M张图片（如果为视频测试，则录制 M个视频），其中达到质量要求（参考表A2.1.3）的数量为 S ，达到质量要求的真人通过次数记为 P，真人通过率TPR =P/S。实际测试过程中一般建议真人超过100人，有效测试总量超过1000次，真人图像应满足数据质量要求；

攻击误过率测试：利用篡改工具对真人图片或视频进行篡改，不限制生成结果质量。基于人脸生成方法攻击 A1 次， 基于人脸替换攻击 A2 次， 基于人脸属性编辑攻击 A3 次，基于人脸表情编辑攻击 A4 次， 累计攻击 A 次（A=A1+A2+A3+A4）,记攻击成功的次数为F，则攻击误过率FAR＝F/A。一般建议构造不同类型的攻击比例为1:1:1:1，整体攻击数量与真人数量比例为1:1，建议攻击的总量不低于4000次。

最终准确率为真人通过数量P以及攻击拦截数量（A-F） 占总测试数量的比例，Acc=(P+A-F)/(S+A)。

A.2.1　攻击数据

人脸图像内容攻击应考虑的因素包括但不限于表A.7所示内容:

表A.7　图像攻击

序号	因素	类型
1	人脸生成	包括但不限于StyleGAN， PGGAN等整脸生成方法
2	人脸替换	包括但不限于Deepfake、FaceSwap等人脸替换方法

表A.7　图像攻击(续)

序号	因素	类型
3	属性编辑	包括但不限于StarGAN, STGAN等人脸属性编辑方法
		属性编辑内容包括但不限于头发，年龄，眼镜，肤色等
4	表情编辑	包括但不限于Face2Face、NeuralTextures等人脸表情编辑方法
		表情编辑包括但不限于整体表情、局部嘴巴、眼睛编辑
5	攻击图像质量	包括但不限于分辨率、清晰度等

人脸视频内容攻击应考虑的因素包括但不限于表A.8所示内容：

表A.8　视频攻击

序号	因素	类型
1	人脸生成	包括但不限于StyleGAN， PGGAN等整脸生成方法
2	人脸替换	包括但不限于Deepfake、FaceSwap等人脸替换方法
3	属性编辑	包括但不限于StarGAN, STGAN等人脸属性编辑方法
		属性编辑内容包括但不限于头发，年龄，眼镜，肤色等
4	表情编辑	包括但不限于Face2Face、NeuralTextures等人脸表情编辑方法
		表情编辑包括但不限于整体表情、局部嘴巴、眼睛编辑
5	视频篡改帧数	包括但不限于所有帧篡改，部分帧篡改等
6	攻击视频质量	包括但不限于分辨率、清晰度、帧率、时长等

A.3　模型安全测试

对抗、投毒集后门攻击测试方法：选定N个自然人，并准备M个攻击目标ID，每个ID对应一张人脸图像。每个自然人需要定向攻击到M个目标ID，进而可以构造出N*M组攻对。利用下述各类攻击方法，为每组攻击对构造P个攻击实例。以人脸识别系统中FAR为一千万分之一时采用的相似度相似度得分阈值为基准，如果攻击后的人脸与目标人脸之间的相似度大于该阈值，则认为攻击成功，反之则不成功。攻击成功的人脸对抗样本总数记为S，全部测试样本为P*N*M，则攻击成功率为ASR = S / (P*N*M)，实际测试过程中一般建议自然人超过100人，目标ID数量超过100人，测试总量超过50000次。

a）对抗攻击：

表A.9　对抗攻击

序号	方式	类型
1	数字白盒攻击	包括但不限于FGSM、DeepFool、C/W、PGD；数字域攻击约束扰动大小为8/255

表A.9　对抗攻击（续）

序号	方式	类型
2	物理黑盒攻击	包括但不限于查询攻击和迁移性攻击进行黑盒物理攻击，黑盒攻击单个样本最大查询次数限制在100次内

b）数据投毒攻击：

表A.10　数据投毒攻击

序号	方式	类型
1	数据投毒攻击	使用包括但不限于LTC、WB等代表性数据投毒攻击算法，投毒数据比例限制在1%以内

c）后门攻击：

表A.11　后门攻击

序号	方式	类型
1	后门攻击	使用包括但不限于BadNets、Blened Attack 、Consistent Attack 等代表性后门攻击算法

窃取测试方法：选定 N 组人脸数据作为模型窃取所用查询数据，另准备 M 组人脸数据作为原模型与窃取后模型性能的测试数据。先使用 N 组查询数据对原模型 F 进行窃取攻击，得到窃取模型 Q。用 M 测试数据分别对原模型 F 和窃取模型 Q 进行人脸识别测试，测试时均使用 FAR 为一千万分之一时采用的相似度相似度得分阈值，同时比较两个模型此时的 FFR 数据，模型性能下降率=（Q 模型 FFR-F 模型 FFR）/F 模型 FFR。实际测试过程中一般建议查询数据超过 1000 组，测试人脸数据超过 10000 组。

d）模型窃取攻击：

表A.12　模型窃取攻击

序号	方式	类型
1	模型窃取攻击	使用包括但不限于 JBDA、Knockoff 等代表性窃取算法，单样本最大查询次数限制在 1000 次。