ICS 35.240
CCS L 70

# DB11

北　京　市　地　方　标　准

DB11/T 2252—2024

# 信息安全　人脸识别防对抗样本攻击测试要求

Information security—Testing requirements on defending adversarial attack against face recognition

2024 - 06 - 28 发布　　2024 - 10 - 01 实施

北京市市场监督管理局　发布

# 目　　次

前言 ..... II
1 范围 ..... 1
2 规范性引用文件 ..... 1
3 术语和定义 ..... 1
4 攻击方式分类 ..... 1
5 测试方法 ..... 2
5.1 测试条件 ..... 2
5.2 对抗样本制作方法 ..... 2
5.3 测试流程 ..... 4
5.4 测试结果计算方法 ..... 5
附录 A （资料性） ..... 6

# 前　言

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

本文件由北京市公安局提出并归口。

本文件由北京市公安局组织实施。

本文件起草单位：北京市公安局、北京市公安局人工智能安全研究中心、北京瑞莱智慧科技有限公司、北京百度网讯科技有限公司、中国科学院计算技术研究所、科大讯飞股份有限公司、公安部第三研究所、北京市标准化研究院。

本文件主要起草人：蔡瑜坤、曹奇、王崇鹏、张晓飞、孙毅、刘鸣、邓佳、马飞翔、孔凡真、杨彬、李晓波、王东明、辛铮、韦云霞、张旭东、孙空军、李连吉、萧子豪、张浩天、王海棠、郭建岭、曹娟、唐胜、方凌飞、朱莉莉、孔凡胜、程鸣、孙文琦、丁治国、周巧霖、樊子风。

# 信息安全 人脸识别防对抗样本攻击测试要求

## 1 范围

本文件规定了人脸识别防对抗样本攻击的攻击方式分类、测试方法。

本文件适用于人脸识别应用或系统的开发者、使用者及相关方开展防对抗样本攻击测试。

## 2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 38671 信息安全技术 远程人脸识别系统技术要求

GB/T 41987 公共安全 人脸识别应用 防假体呈现攻击测试方法

GA/T 1324 安全防范 人脸识别应用 静态人脸图像采集规范

## 3 术语和定义

GB/T 38671界定的以及下列术语和定义适用于本文件。

3.1

**对抗样本 adversarial examples**

在输入数据中通过故意添加细微干扰获得的、可导致机器学习算法模型以高置信度给出错误输出的样本。

3.2

**对抗补丁 adversarial patch**

通过替换特定区域图像内容构造的对抗样本，通常表现为特定形状的图像块。

3.3

**物理对抗补丁 physical adversarial patch**

基于对抗补丁制作出的物理攻击道具。

## 4 攻击方式分类

根据被测试的人脸识别应用或系统在进行识别时可接受的输入类型，攻击方式分为两大类，如表1所示。

表 1 攻击方式分类

| 攻击方式类型 | 攻击输入 | 适用测试对象 |
| --- | --- | --- |
| 注入攻击类 | 基于像素扰动的对抗样本（参见A.1）<br>基于对抗补丁的对抗样本（参见A.2） | 人脸识别应用或系统可通过输入人脸图像数据进行识别 |
| 呈现攻击类 | 物理对抗补丁 | 人脸识别应用或系统可通过摄像头采集人脸数据进行识别 |

## 5 测试方法

### 5.1 测试条件

#### 5.1.1 人脸识别准确率

测试开始前应按要求部署被测人脸识别应用或系统。被测人脸识别应用或系统的相似度阈值等各项设置应调整到默认状态。应确保被测应用人脸识别功能正常，可通过输入人脸图像或直接采集人脸图像等方式进行多次识别并统计人脸识别结果，记录正常人脸样本的识别次数、识别准确率。

#### 5.1.2 测试环境光线

进行呈现攻击测试时，测试环境光线可分别模拟室内环境、半室外环境及室外环境，应符合 GB/T 41987 中测试过程中环境光线的要求。

### 5.2 对抗样本制作方法

#### 5.2.1 人脸识别对象选取

在对抗样本攻击的测试对象样本中，男女比例应为1:1，年龄在16岁~60岁的占80%，小于16岁的占10%，大于60岁的占10%。测试人员确定后按照性别一致、年龄相仿的原则进行两两分组，随机选取其中一人作为模拟攻击者，另一人作为模拟攻击目标。

#### 5.2.2 人脸数据采集

在呈现攻击类测试中，按照GA/T 1324采集规范采集人脸照片，针对每个测试组，需采集模拟攻击者、模拟攻击目标两人各1张照片。

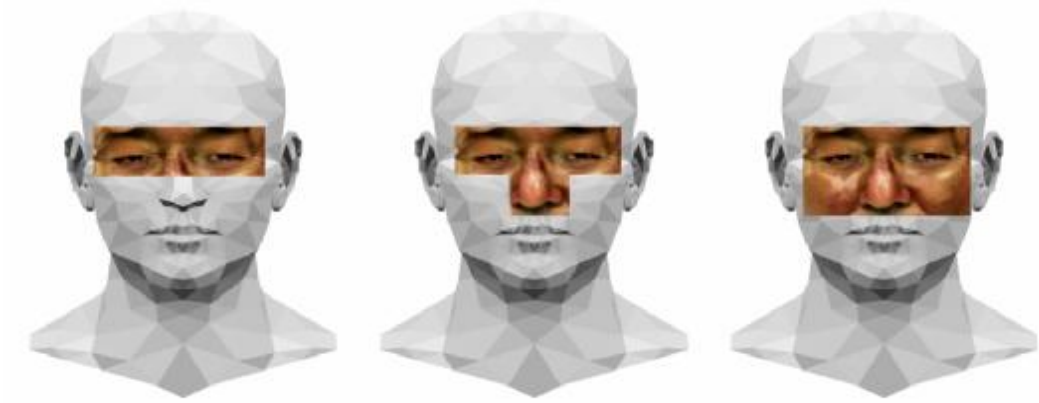
#### 5.2.3 对抗样本制作

##### 5.2.3.1 基于像素扰动的对抗样本制作

以被测人脸识别应用或系统为攻击对象，选取不少于200个测试组人脸数据，分别制作基于像素扰动的对抗样本。针对于每个测试组应生成不同扰动大小的对抗样本图像。记录测试组数量、对抗样本制作方法及扰动大小。

##### 5.2.3.2 基于对抗补丁的对抗样本制作

以被测人脸识别应用或系统为攻击对象，选取不少于200个测试组人脸数据，分别制作基于对抗补丁的对抗样本。如图1所示，针对每个测试组应分别生成覆盖眼部区域的对抗样本图像，如图1a)所示；生成覆盖眼部及鼻子区域的对抗图像，如图1b)所示；生成覆盖眼部鼻子及脸颊区域的对抗样本图像，如图1c)所示。记录测试组数量、对抗样本制作方法及扰动大小。

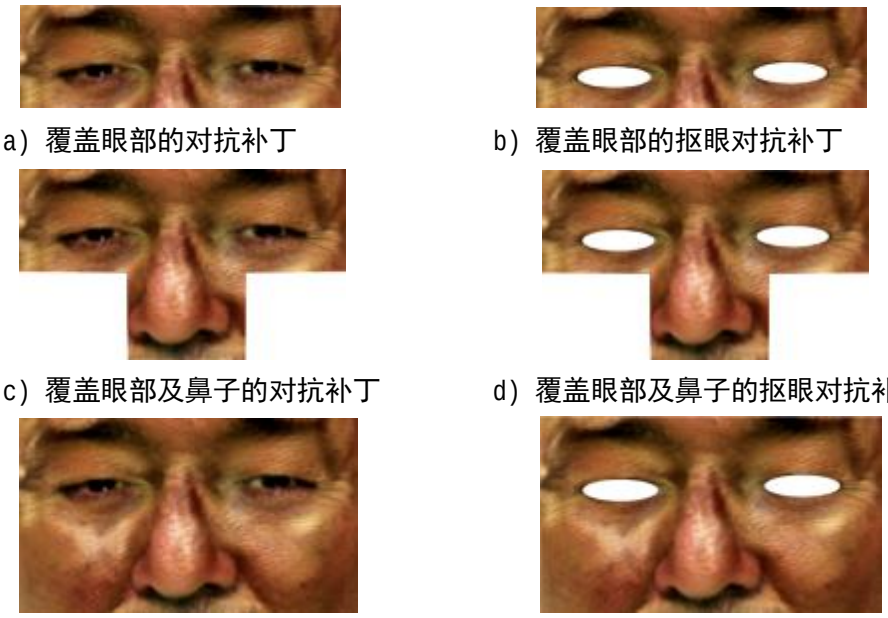


a）覆盖眼部的对抗样本　　b）覆盖眼部及鼻子的对抗样本 c）覆盖眼部鼻子及脸颊的对抗样本

图 1 基于对抗补丁的对抗样本规格示例

5.2.3.3 物理对抗补丁制作

以被测人脸识别应用或系统为攻击对象，选取不少于 10 个测试组人脸数据，分别制作物理对抗补丁。针对每个测试组应分别制作覆盖眼部区域的物理对抗补丁，如图 2a)和图 2b)所示，其中图 2b)为扣眼的物理对抗补丁；制作覆盖眼部及鼻子区域的物理对抗补丁，如图 2c)和图 2d)所示，其中图 2d)为覆盖眼部及鼻子的抠眼对抗补丁；制作覆盖眼部鼻子及脸颊区域的物理对抗补丁，如图 2e)和图 2f)所示，其中图 2f)为覆盖眼部鼻子及脸颊的抠眼对抗补丁。针对于带有眨眼动作配合式活体的人脸识别应用或系统应分别制作扣眼区域类型的物理对抗补丁，针对于不带有眨眼动作配合式活体的人脸识别应用或系统应制作带有眼部区域特征类型的物理对抗补丁。记录测试组数量、对抗样本制作方法及扰动大小。



a）覆盖眼部的对抗补丁　　b）覆盖眼部的抠眼对抗补丁

c）覆盖眼部及鼻子的对抗补丁　　d）覆盖眼部及鼻子的抠眼对抗补丁

e）覆盖眼部鼻子及脸颊的对抗补丁　　f）覆盖眼部鼻子及脸颊的抠眼对抗补丁

图 2 物理对抗补丁规格示例

物理对抗补丁可选用纸张、硅胶、塑料等材质进行制作，所生成的对抗补丁瞳距需与模拟攻击者瞳

距保持一致。

### 5.3 测试流程

#### 5.3.1 防基于像素扰动的对抗样本攻击测试

##### 5.3.1.1 针对人脸验证的测试流程

测试流程如下：

a) 输入准备好的模拟攻击目标图像及对应的基于像素扰动的对抗样本测试组，并依次获取识别结果；
b) 若应用将模拟攻击者对抗样本及模拟攻击目标识别为同一人，则判定应用防对抗样本攻击失败；反之则判断为防对抗样本攻击成功。

##### 5.3.1.2 针对人脸辨识的测试流程

测试流程如下：

a) 将模拟攻击目标图像标识注册到人脸辨识应用中；
b) 将基于像素扰动的对抗样本输入人脸辨识应用中；
c) 若应用将模拟攻击者对抗样本及模拟攻击目标识别为同一人，则判定防对抗样本攻击失败，反之则判断为防对抗样本攻击成功。

#### 5.3.2 防基于对抗补丁的对抗样本攻击测试

##### 5.3.2.1 针对人脸验证测试流程

测试流程如下：

a) 输入准备好的模拟攻击目标图像及对应的基于对抗补丁的对抗样本测试组，并依次获取识别结果。
b) 若应用将模拟攻击者对抗样本及模拟攻击目标识别为同一人，则判定防对抗样本攻击失败；反之则判断为防对抗样本攻击成功。

##### 5.3.2.2 针对人脸辨识测试流程

测试流程如下：

a) 将模拟攻击目标图像标识注册到人脸辨识应用中；
b) 将基于对抗补丁的对抗样本输入人脸辨识应用中；
c) 若应用将模拟攻击者对抗样本及模拟攻击目标识别为同一人，则判定防对抗样本攻击失败，反之则判断为防对抗样本攻击成功。

#### 5.3.3 防物理对抗补丁攻击测试

##### 5.3.3.1 针对人脸验证测试流程

测试流程如下：

a) 输入准备好的模拟攻击目标图像后，各测试组模拟攻击者应依次佩戴上制作好的物理对抗补丁，并按照人脸识别应用或系统的指示进行测试。
b) 各测试组测试者面对摄像机，摄像机采集到的正面人脸图像两眼间距不小于 60 像素。佩戴物理对抗补丁人脸水平转动±90°，转动过程中俯仰角不超过 30°，倾斜角不超过 20°，转动一次耗时 1.5s~3s。人脸验证应用或系统输出一次结果记为一次测试过程。每个物理对抗补丁

需要进行不少于 10 次的测试。

c) 若在一次测试过程中将模拟攻击者对抗样本及模拟攻击目标识别为同一人，则判定为防对抗样本攻击失败，反之则判断为防对抗样本攻击成功。

#### 5.3.3.2 针对人脸辨识测试流程

测试流程如下：

a) 将模拟攻击目标图像标识注册到人脸辨识应用中，而后各测试组测试者应依次佩戴上制作好的物理对抗补丁，按照人脸识别应用或系统的指示进行测试。

b) 各测试组测试者面对摄像机，摄像机采集到的正面人脸图像两眼间距不小于 60 像素。人脸水平转动±90°，转动过程中俯仰角不超过 30°，倾斜角不超过 20°，佩戴物理对抗补丁从-90°到+90°转动一次耗时 1.5s~3s。重复“正面脸→左侧脸→右侧脸”动作 3 次，记为一次测试过程。每个物理对抗补丁需要进行不少于 10 次的测试，并记录测试结果。

c) 若在一次测试过程中将模拟攻击者对抗样本及模拟攻击目标识别为同一人，则判定防对抗样本攻击失败，反之则判断为防对抗样本攻击成功。

### 5.4 测试结果计算方法

#### 5.4.1 防像素扰动对抗样本攻击失败率

按式（1）计算得到防像素扰动对抗样本攻击失败率$F_1$。$F_1$越低表明人脸识别应用或系统抵御像素扰动对抗样本攻击的能力越强。

$$F_1=X/ZX\times 100\% \quad (1)$$

式中：

$F_1$ ——防像素扰动对抗样本攻击失败率；

$X$ ——防像素扰动对抗样本攻击失败的次数；

$ZX$——防像素扰动对抗样本攻击测试总次数。

#### 5.4.2 防对抗补丁攻击失败率

按式（2）计算得到防对抗补丁攻击失败率。$F_2$越低表明人脸识别应用或系统抵御对抗补丁攻击的能力越强。

$$F_2=B/ZB\times 100\% \quad (2)$$

式中：

$F_2$ ——防对抗补丁攻击失败率；

$B$ ——防对抗样本攻击失败的次数；

$ZB$——防对抗补丁攻击测试总次数。

#### 5.4.3 防物理对抗补丁攻击失败率

按式（3）计算得到防物理对抗补丁攻击失败率。$F_3$越低表明人脸识别应用或系统抵御物理对抗补丁攻击的能力越强。

$$F_3=W/ZW\times 100\% \quad (3)$$

式中：

$F_3$ —— 防物理对抗补丁攻击失败率；

$W$ —— 防物理对抗补丁攻击失败的次数；

$ZW$—— 防物理对抗补丁攻击测试总次数。

# 附 录 A
（资料性）
对抗样本描述与计算方式

## A.1 基于像素扰动的对抗样本描述

像素扰动的对抗样本的具体描述见式（A.1）

$$x' = x + \delta, \|\delta\|_p \leq \epsilon \quad \text{.....(A.1)}$$

式中：

$x'$—— 对抗样本，图像中每像素的取值范围为 0 到 255；

$x$—— 真实样本，图像中每像素的取值范围为 0 到 255；

$\| \|_p$——$p$范数，通常取 0、2、∞；

$\delta$—— 对抗扰动，其维度与样本维度相同，扰动在$p$范数度量下的大小不超过$\epsilon$；

$\epsilon$—— 扰动大小约束上界，扰动在$p$范数度量下的大小不超过$\epsilon$。

## A.2 基于对抗补丁的对抗样本描述

基于对抗补丁的对抗样本具体描述见式（A.2）

$$x' = (1 - m) * x + m * \delta \quad \text{.....(A.2)}$$

式中：

$x'$—— 对抗样本，图像中每像素的取值范围为 0 到 255；

$x$—— 真实样本，图像中每像素的取值范围为 0 到 255；

$m$—— 添加对抗补丁区域的掩膜，类别为矩阵，集中于某一区域，取值为 0 或 1；

$\delta$—— 对抗补丁，集中于$m$=1 的区域，图像中每像素的取值范围为 0 到 255 之间的整数。