物理域中针对人脸识别系统的对抗样本攻击方法

蔡楚鑫[1,2], 王宇飞[1,2], 章烈剽[3], 卓思超[1], 张娟苗[4], 胡永健[1,2]

[1]华南理工大学电子与信息学院 广州 中国 510641
[2]中新国际联合研究院 广州 中国 511356
[3]广州广电卓识智能科技有限公司 广州 中国 510663
[4]广州广电运通金融电子股份有限公司 广州 中国 510663

摘要 对抗样本攻击揭示了人脸识别系统可能存在不安全性和被攻击的方式。现有针对人脸识别系统的对抗样本攻击大多在数字域进行, 然而从最近文献检索的结果来看, 越来越多的研究开始关注如何能把带有对抗扰动的实物添加到人脸及其周边区域上, 如眼镜、贴纸、帽子等, 以实现物理域的对抗攻击。这类新型的对抗样本攻击能够轻易突破市面上现有绝大部分人脸活体检测方法的拦截, 直接影响人脸识别系统的结果。尽管已有不少文献提出数字域的对抗攻击方法, 但在物理域中复现对抗样本的生成并不容易且成本高昂。本文提出一种可从数字域方便地推广到物理域的对抗样本生成方法, 通过在原始人脸样本中添加特定形状的对抗扰动来攻击人脸识别系统, 达到误导或扮演攻击的目的。主要贡献包括: 利用人脸关键点根据脸型构建特定形状掩膜来生成对抗扰动; 设计对抗损失函数, 通过训练生成器实现在数字域的对抗样本生成; 设计打印分数损失函数, 减小打印色差, 在物理域复现对抗样本的生成, 并通过模拟眼镜佩戴、真实场景光照变化等方式增强样本, 改善质量。实验结果表明, 所生成的对抗样本不仅能在数字域以高成功率攻破典型人脸识别系统 VGGFace10, 且可方便、大量地在物理域复现。本文方法揭示了人脸识别系统的潜在安全风险, 为设计人脸识别系统的防御体系提供了很好的帮助。

关键词 人脸识别; 对抗样本攻击; 数字域对抗样本; 物理域对抗样本; 打印分数损失函数
中图法分类号 TP309.1 DOI 号 10.19363/J.cnki.cn10-1380/tn.2023.03.10

Adversarial Attacks on Face Recognition System in Physical Domain

CAI Chuxin[1,2], WANG Yufei[1,2], ZHANG Liepiao[3], ZHUO Sichao[1], ZHANG Juanmiao[4], HU Yongjian[1,2]

[1]School of Electronic and Information Engineering, South China University of Technology,Guangzhou 510641, China
[2]Sino-Singapore International Joint Research Institute, Guangzhou 511356, China
[3]Guangzhou GRG Vision Co., Ltd., Guangzhou 510663, China
[4]Guangzhou GRG Banking Equipment Co., Ltd., Guangzhou 510663, China

Abstract Adversarial attacks exhibit both potential insecurity of face recognition systems and the way of performing attacks. Most current adversarial attacks on face recognition systems are carried out in digital domain. However, based on the recent reports in literature, more and more studies begin to concern about how to put the physical patches containing adversarial noise on human face and its neighboring regions, for example, eyeglass framework, paper sticker, and cap, so as to implement adversarial attacks in physical domain. Such a new type of attacks can easily break through most of current living face detection systems and thus affect the decision of face recognition systems. Although there are a few methods proposed for the generation of adversarial samples in digital domain, it is not easy or cheap to realize those methods in physical domain. This paper proposes a method of generating adversarial attack in digital domain which can be readily extended to physical domain. By adding adversarial perturbation of special shapes into an original face sample, we can fool the face recognition system and make it regard the face as someone else's face (i.e., dodging attack) or a specific person's face (i.e., impersonation attack). The major contributions of this paper include: First, we propose a method of using the face landmarks to construct a specific shape mask of the adversarial perturbation for individual face. Second, we design the adversarial loss function to train the generator to produce digital samples. Third, we design the printing score loss function to reduce the color difference between display and printer so as to reproduce those samples in physical domain. We improve the quality of adversarial samples by means of data enhancement which aims at simulating the way of wearing eyeglasses, illumination variations and other situations in real-world applications. Experimental results show that the proposed method can attack the face recognition system VggFace10 in a high success rate in digital domain. Moreover, it

通讯作者: 胡永健, 博士, 教授, Email:eeyjhu@scut.edu.cn。
本课题得到国家重点研发计划项目(No. 2019QY2202), 广州开发区国际合作项目(No. 2019GH16)和中新国际联合研究院项目(No. 206-A018001)资助。
收稿日期: 2021-10-14; 修改日期: 2021-12-26; 定稿日期: 2023-01-04

can be readily extended to physical domain and generates samples quickly and economically. Our study exposes the security risk of face recognition systems, which can provide us with useful information to design better face recognition systems against adversarial attacks in the future.

Key words face recognition; adversarial sample attack; digital adversarial samples; physical adversarial samples; printing score loss function

1 引言

人脸识别作为一种非侵入式身份验证方式, 因其良好的用户交互性, 在身份认证中应用得越来越广。近年来随着深度学习和人工智能技术的发展, 深度神经网络(Deep Neural Networks, DNN)在许多计算机视觉任务上取得了传统机器学习和图像处理算法无法企及的性能, 基于 DNN 的人脸识别(Face Recognition, FR)算法(例如 VGGFace[1]和 ArcFace[2])的性能远远高于传统基于手工特征的 EigenFace 和 FisherFace 等特征脸算法。然而, 针对 FR 系统的攻击手段一刻也没有停止发展, 高分辨率手机、高保真彩色打印机以及 3D 打印机等设备的日益普及, 使传统的视频回放/重放、照片打印、3D 面具等攻击方式的威力提升, 此外, 近年来还出现一种威胁更大且更为隐蔽的攻击方式, 这就是基于神经网络本身的对抗样本攻击。

最新的研究表明, 基于 DNN 的 FR 系统表现出一种耐人寻味的脆弱性, 可以用难以察觉或感知但看起来自然的对抗输入图像使得模型输出不正确的预测。简单来说, 有一个学习系统 S 及干净的输入样本(没有添加噪声的样本)x, 我们假设样本 x 被学习系统正确地分类, 即 $S(x)=y_{true}$, 建立一个几乎与样本 x 相同但是却被错误分类的样本 x', 使 $S(x')\neq y_{true}$, 这种样本 x' 我们称之为对抗样本[3]。

对抗样本攻击揭示了 FR 系统可能存在的不安全性和被攻击的方式。由于对抗样本攻击只需要以不易察觉的方式修改人脸的小部分区域便可实现攻击[4-9], 传统针对视频重放、照片打印和 3D 面具等攻击的活体检测算法很难检测出此类攻击[10]。通过研究对抗样本的生成和实现可为后续研究针对对抗样本的检测和提高 FR 系统安全性提供重要理论支撑。

现有针对 FR 系统的对抗样本攻击大多数在数字域进行, 然而从文献检索的结果来看, 越来越多的研究开始关注如何能把带有对抗扰动的实物添加或者通过光投影等方式呈现在人脸及其周边区域上, 如眼镜、贴纸、帽子等, 以实现物理域的对抗样本攻击[11]。物理域的对抗样本攻击可由个人在真实应用场景无需协助、成本低廉地实现, 因此, 对 FR 系统的危害尤为巨大。本文着眼于研究对抗样本物理域的实现研究, 首先在数字域提出一种快速对抗扰动生成方法, 可根据需要生成特定形状的扰动, 然后在物理域实现。本文以生成眼镜为物理域攻击载体, 也可以推广到其它形式的载体。主要贡献包括: (1)根据人脸关键点构建眼镜掩膜, 可随人脸自适应确定对抗扰动的形状、大小及位置; (2)设计对抗损失函数, 通过惩罚正确判断和鼓励攻击来生成数字域对抗样本; (3)设计打印分数损失函数来减小打印色差, 并通过数据增强, 模拟物理域佩戴攻击过程和光照变化, 将数字域样本更好地推广到物理域实现。

2 相关工作介绍

首先简要介绍标准术语[11]。对抗样本/图像(Adversarial Example/Image)是一种对干净图像(Clean Image)进行有意的修改(例如, 添加噪声)后得到的攻击图像, 用以欺骗机器学习模型, 如 FR 模型。对抗性训练(Adversarial Training)是使用对抗样本和干净图像的训练过程。敌手(Adversary)是一个创建对抗样本的代理人(Agent), 或者就是对抗样本本身。误导攻击(Dodging Attack)定义为攻击者试图将一张面孔错误地识别为任何其他任意面孔, 也被称为混淆攻击(Obfuscation Attack)。躲避攻击(Evasion Attack)定义为通过在测试阶段改变样本来逃避系统, 但不影响训练数据。扮演攻击(Impersonation Attack)定义为将一张脸伪装成一个特定的(经授权的)脸。下毒攻击(Poisoning Attack)发生在训练期间, 通过污染训练数据来实现攻击。

Szegedy 等人在文献[12]最早引入对抗样本的概念, 提出利用 L-BFGS 算法通过优化遍历流形网络表示在输入空间中发现对抗样本。Goodfellow 等人[13]利用 FGSM(Fast Gradient Sign Method)算法通过在原图片上沿神经网络的梯度变化最大方向等步长地添加对抗扰动来生成对抗样本。这两个经典方法均通过对原始样本添加对抗扰动来构建对抗样本, 其共同问题是生成速度慢, 计算复杂。为此, Goodfellow 等人提出的生成式对抗网络 GAN(Generative Adversarial Network)[14]被用来快速大量生成对抗扰动。文献[15]提出注意力对抗性攻击生成网络(A3GN), 通

过条件变分自动编码器和注意力模块来学习人脸之间的实例级对应关系，引入了 FR 网络作为第三方参与生成器和判别器之间的竞争以达到针对性攻击效果。上述方法均是在数字域进行。Sharif 等人在文献[16]和[17]开始讨论如何在物理域生成对抗样本，以眼镜形状为例，通过判别器判断生成的眼镜图片是否像眼镜，然后利用生成的眼镜攻击人脸识别器。该方法需要事先收集眼镜图片构建数据库来训练 GAN，生成的对抗样本的大小、位置与人脸可能存在不匹配，此外训练过程分步进行，非端到端实现。文献[18]利用余弦相似度损失来生成矩形对抗扰动，粘贴在帽子上攻击 FR 系统，但需知道攻击目标经 DNN 提取的嵌入特征，并需拉近对抗样本和攻击目标两个嵌入特征之间的余弦距离。文献[19]通过 FGSM 算法生成对抗性补丁进行物理域攻击，每次只能针对一个对抗补丁进行优化。文献[20]利用投影仪投影对抗扰动到人脸上实现物理域攻击，在攻击时需要进行精确的位置校准和颜色校准。上述物理域中的攻击方法普遍忽略了对抗扰动与现实人脸的适配问题，也未在训练时考虑物理域与数字域之间如光照变化等的环境差异，影响了其攻击成功率。此外，生成成本较高，生成形状、大小及位置不够灵活也是物理域对抗样本生成方法有待解决的问题。

3 算法介绍

本文算法整体流程如图 1 所示，包括对抗样本

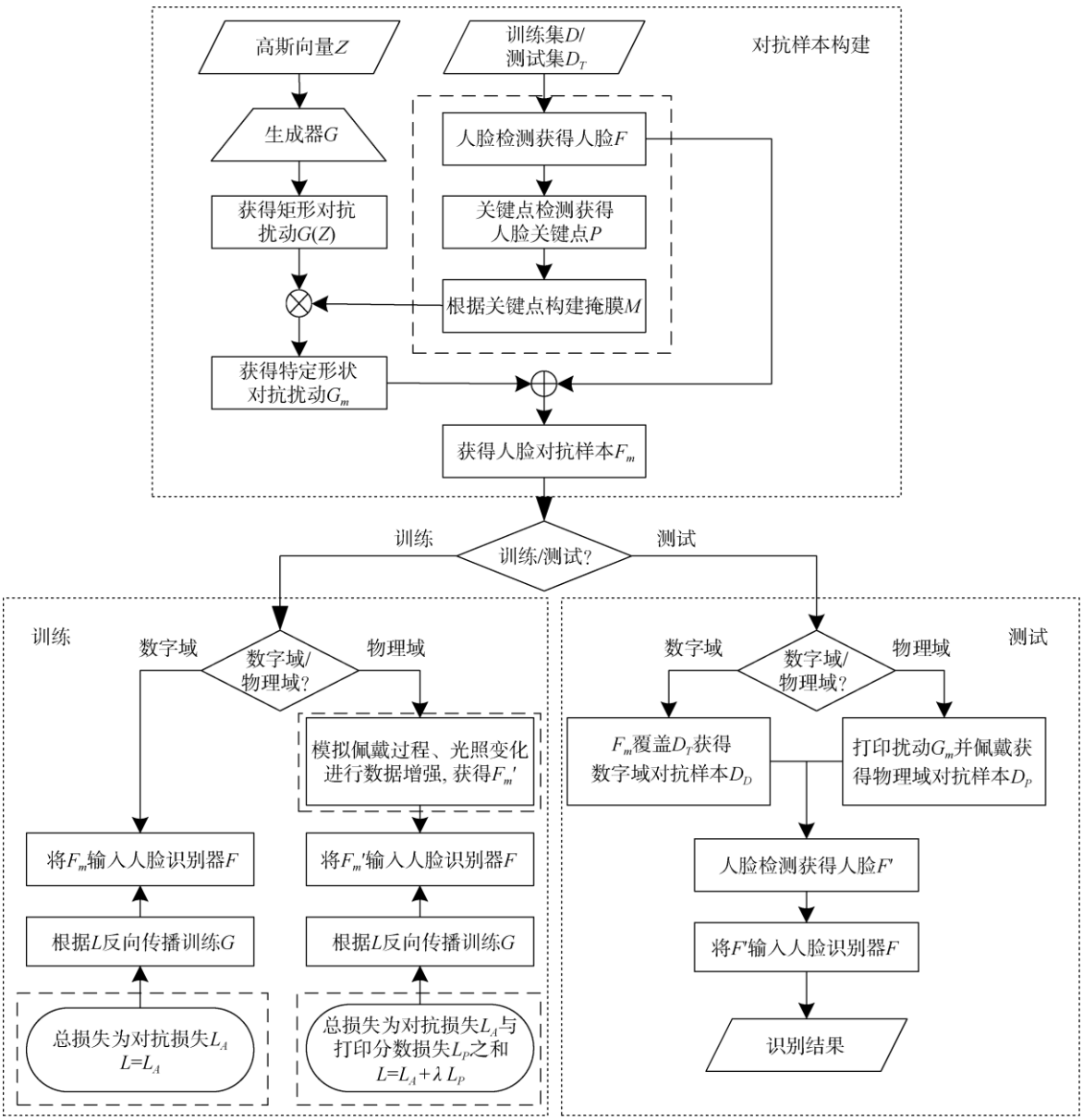


图 1　算法流程图

Figure 1　Flow chart of the proposed algorithm

构建、生成器网络训练和测试三部分。其中符号定义如下: $\boldsymbol{F}$ 表示人脸检测获得的人脸, $\boldsymbol{P}$ 表示人脸关键点, $\boldsymbol{M}$ 表示掩膜, $\boldsymbol{Z}$ 表示高斯向量, G 表示生成器, $G(\boldsymbol{Z})$表示生成器生成的矩形对抗扰动, $\boldsymbol{G}_m$ 表示掩膜形状的对抗扰动, $\boldsymbol{F}_m$ 表示构建的人脸对抗样本, $\boldsymbol{F}'_m$ 表示数据增强后的人脸对抗样本。

对抗样本构建分为两个阶段: 第一阶段先通过人脸检测获得人脸 $\boldsymbol{F}$, 然后获得人脸关键点 $\boldsymbol{P}$, 构建掩膜 $\boldsymbol{M}$; 第二阶段通过将 $\boldsymbol{Z}$ 输入 G 获得 $G(\boldsymbol{Z})$, 然后通过 $\boldsymbol{M}$ 与 $G(\boldsymbol{Z})$相乘获得 $\boldsymbol{G}_m$, 并将对抗扰动 $\boldsymbol{G}_m$ 在掩膜区域部分覆盖人脸 $\boldsymbol{F}$ 获得人脸对抗样本 $\boldsymbol{F}_m$。

生成器网络训练分成数字域和物理域攻击两部分。对于数字域攻击部分, 利用对抗损失函数 L_A 训练生成器 G, 通过生成器 G 和人脸识别系统的对抗迭代调整 G 的权重使得构建的 $\boldsymbol{F}_m$ 能够实现扮演攻击或误导攻击。对于物理域攻击部分, $\boldsymbol{F}_m$ 先通过模拟物理域场景数据增强获得 $\boldsymbol{F}'_m$, 再用于生成器 G 和人脸识别系统对抗迭代调整 G 的权重, 同时在数字域的基础上, 总损失函数增加了打印分数损失 L_P; 测试部分也分成数字域和物理域攻击两部分。对于数字域攻击部分, 利用训练好的生成器 G 构建人脸对抗样本 $\boldsymbol{F}_m$, 覆盖原始样本的人脸区域构成数字域对抗样本, 送入人脸识别系统测试结果。对于物理域攻击部分, 则需要将对抗扰动 $\boldsymbol{G}_m$ 打印进行佩戴构成物理域对抗样本, 然后对人脸识别器进行对抗样本攻击测试。

3.1 对抗样本的构建

现有算法往往通过收集数据集预训练生成器[16-17]来生成特定形状的对抗扰动,将其添加覆盖在人脸上, 但并未考虑脸型、大小与对抗扰动位置之间的协调性, 使得从数字域迁移复现到物理域的步骤较为繁琐, 且在物理域的攻击效果易受影响。为了简便有效地解决对抗扰动形状确定的问题, 本算法提出利用人脸关键点确定的掩模来构建具有特定形状的对抗扰动, 这样可实现眼镜大小与人脸自适应, 既提高了佩戴眼镜位置和脸型大小的协调性, 同时也简化了眼镜制作步骤和提高攻击成功率。构建对抗样本时, 首先对检测出的人脸进行人脸关键点检测, 并根据关键点构建掩膜。以常见的 68 个关键点为例(如图 2 所示), 根据 12 个眼睛关键点确定眼镜内、外边框和中间横梁的位置。其中, 内边框根据左右眼角以及上下眼皮的关键点往外扩展 L_1 个像素值确定; 外边框由眼镜内边框往外扩展 L_2 个像素确定; 中间横梁在上下外边框中间位置, 宽度为 L_2 个像素。L_1、L_2 根据输入人脸的尺寸决定。最终确定的眼镜掩膜 $\boldsymbol{M}$ 如图 2 中的框图所示。

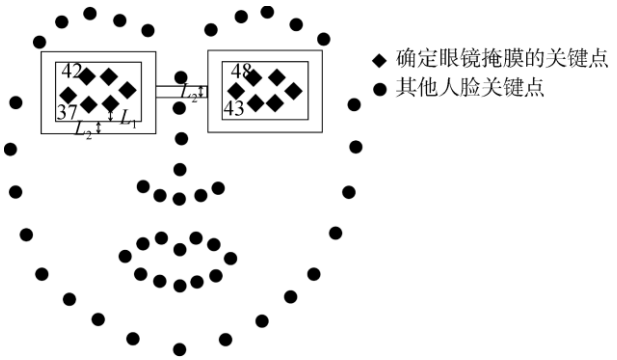


图 2 眼镜掩膜生成示意图。菱形为用于确定眼镜掩膜形状位置的人脸关键点, 圆形表示其他人脸关键点

Figure 2 Illustration of generating the mask of glasses. The diamonds indicate those face landmarks used to determine the shape and location of the mask, and the round dots indicate the rest face landmarks.

对抗样本构建的第二部分是通过生成器生成矩形扰动矩阵, 并与眼镜掩膜相乘获得眼镜对抗扰动, 然后覆盖原始人脸样本由眼镜掩膜确定的位置, 形成人脸对抗样本 $\boldsymbol{F}_m$, 具体操作为:

$$\boldsymbol{F}_m = (1-\boldsymbol{M})\times\boldsymbol{F} + \boldsymbol{M}\times G(\boldsymbol{Z}) \tag{1}$$

3.2 数字域训练与对抗损失函数

现有算法往往通过更改标签并结合对应人脸识别系统的损失函数来生成对抗扰动, 因此难以扩展到使用其他损失函数的系统上。典型的人脸识别系统本质上是一个分类问题, 识别时把每个待测主体都输出为多个类别的概率分数, 将网络输出概率分数最大的对象作为识别结果。鉴于此, 本算法引入更能反映判决过程本质的概率分数损失(Score Loss)L_S, 与原本标签驱动的人脸识别系统损失(Face Loss)L_F 一起构成最终的对抗损失函数 L_A(Adversarial Loss):

$$L_A = L_S + \alpha L_F \tag{2}$$

其中 α 是 L_F 的加权系数。具体而言, 概率分数损失 L_S 与人脸识别器 F 最后一层输出的概率分数 $F(\bullet)$ 相关。对于误导攻击, L_S 定义为:

$$L_S = \frac{F_g(\boldsymbol{F}_m)}{\sum_{i\neq g} F_i(\boldsymbol{F}_m)+\varepsilon} \tag{3}$$

对于扮演攻击, L_S 定义为:

$$L_S = \frac{\sum_{i\neq t} F_i(\boldsymbol{F}_m)}{F_t(\boldsymbol{F}_m)+\varepsilon} \tag{4}$$

式中 $F_i(\bullet)$、$F_g(\bullet)$ 和 $F_t(\bullet)$ 分别表示人脸识别器 F 将输入判别为类别 i、正确类别(Ground Class)和目标类别(Target Class)的概率分数, ε 用于防止分

母出现 0 的情况以及避免初始分母值太小。通过最小化损失函数训练生成器，在惩罚正确判断的同时鼓励攻击，对于误导攻击，L_S 降低正确类别的概率分数 $F_g(\cdot)$，同时提高非正确类别的概率分数；对于扮演攻击，L_S 降低非目标概率分数，同时提高扮演目标类别的概率分数 $F_t(\cdot)$。相比于通过相减来达到上述目的，本文使用相除可很大程度上避免发生两项的值都很小的极端情况。

人脸识别器损失 L_F 表示训练人脸识别器 F 所用的损失函数。为了引导 G 生成成功攻击人脸识别器的对抗扰动，对于扮演攻击，将真实标签改为目标标签；对于误导攻击，将真实标签改为非正确标签的均分。

由此，本算法通过最小化损失函数迭代调整生成器 G 的权重，使得构建的人脸对抗样本能够欺骗所指定攻击的人脸识别器，且能以批量形式构建人脸对抗样本。在进行数字域攻击时，生成器网络的总损失函数即为 L_A:

$$L = L_A \tag{5}$$

3.3 物理域训练

将所生成的对抗扰动图案打印出来并佩戴以实现物理域攻击。然而，将数字域生成的对抗扰动复现到物理域面临两个挑战，一是物理域中的打印样本与数字域的生成样本之间不可避免存在较大色差，且该色差在不同的打印设备上还不完全一样，二是物理域中的各种环境因素(如光照)会影响摄像头重新捕获的人脸图像。上述两个问题对物理域攻击效果的影响均无法忽略，为此，本算法在数字域攻击的基础上引入打印分数损失，以及通过模拟物理域攻击实现数据增强，以此减小数字域到物理域的差异。

3.3.1 打印分数损失函数

首先考虑色差因素，由于数字域图像采用 RGB 光模式，而打印机采用 CMYK 色料模式，色域差异造成了物理域与数字域的色差。为了减小打印眼镜对抗扰动和数字域对抗眼镜扰动的色差，本算法通过构建打印机分数损失(Printing Score Loss)来减小对抗扰动颜色值与打印后二次重获的数字域颜色值的差异，具体定义为:

$$L_P = \sum_{p\in G_m} \min_{c\in \boldsymbol{C}} |p-c|^2 + \sum_{c\in \boldsymbol{C}} \min_{p\in G_m} |c-p|^2 \tag{6}$$

式中 p 表示眼镜形状对抗扰动 $\boldsymbol{G}_m$ 中的一个像素点颜色值，c 表示打印后二次重获的数字域颜色集 $\boldsymbol{C}$ 中的一种颜色值，其中加号左边用于减少数字域对抗扰动颜色值与打印后二次重获的数字域颜色值的差异，加号右边用于减少打印后二次重获的数字域颜色值与数字域对抗扰动颜色值的差异，两者结合可进一步减小差异。

为了解决打印设备适配的问题，本算法通过先打印后重新拍摄的方法获得可打印颜色集 $\boldsymbol{C}$，具体为: 首先在数字域上在 RGB 三个通道分别按照采样间隔 i 像素进行均匀采样，获得数字域 RGB 颜色值调色盘 $\boldsymbol{C}'$；接着使用打印机打印 $\boldsymbol{C}'$，获得对应的打印机打印 CMYK 颜色值调色盘 $\boldsymbol{C}''$；然后使用摄像头对 $\boldsymbol{C}''$ 进行拍摄，重新获得 RGB 颜色值调色盘，即为二次重获的数字域颜色集 $\boldsymbol{C}$，同时可得到二次重获的数字域到初始数字域颜色值的映射 $\boldsymbol{C}\rightarrow\boldsymbol{C}'$，通过打印初始数字域 $\boldsymbol{C}'$ 中的颜色值即可获得二次重获的数字域像素集 $\boldsymbol{C}$ 中对应的颜色值。

结合 3.2 节中的生成器对抗损失 L_A，在进行物理域攻击时，生成器网络的总损失函数为:

$$L = L_A + \lambda L_P \tag{7}$$

式中 λ 为打印分数损失 L_P 的加权系数。

3.3.2 模拟物理域攻击过程和光照变化数据增强

为了减小物理域佩戴攻击过程造成的差异，本算法对人脸对抗样本 $\boldsymbol{F}_m$ 进行模拟佩戴过程和光照变化数据增强获得 $\boldsymbol{F}_m'$。物理域攻击过程可定义为: 生成眼镜对抗扰动，打印眼镜对抗扰动，攻击者佩戴眼镜在未知的光照环境进行攻击并被摄像头重新捕获，因此可按下面固定顺序进行数据增强以模拟物理域攻击过程。

打印眼镜: 打印眼镜扰动过程中的尺寸变化会破坏像素值，本算法通过随机使用采样率 τ 对眼镜扰动下采样，然后通过上采样恢复原来的尺寸进行重采样模拟这个过程。

佩戴眼镜: 物理域中佩戴眼镜的位置可能会出现偏移与旋转，本算法通过对眼镜扰动进行随机水平平移 h 个像素、垂直平移 v 个像素以及中心旋转 d 度进行模拟。

模拟光照变化: 本算法通过两种方法模拟物理域光照变化。一是线性的对比度和亮度调整[21]:

$$p' = p\times a + b \tag{8}$$

二是非线性变化的伽马变换[22]:

$$p' = p^{\gamma} \tag{9}$$

式中 p 与 p' 分别表示原来像素值与变换后的像素值，a、b 分别是调整对比度和亮度的参数，γ 衡量非线性变化的程度。对于每一个人脸对抗样本，随机选择其中一种光照变化进行数据增强。

重新捕获人脸: 摄像头重新捕获人脸图像时在

数据传输过程中会引入高斯噪声，数据压缩过程中会引入压缩噪声，前者可通过添加均值为 μ、标准差为 σ 的高斯噪声进行模拟，后者可通过质量因子为 q 的 JPEG 压缩进行模拟[23]。

具体参数设置见第 4.5 节。

4 实验场景设置

4.1 人脸识别器与数据集

遵循文献[16]和[17]的设置，本文主要探索对以神经网络模型为基础的人脸识别分类器的攻击，这类人脸识别器在结构上是类似的，通常使用典型骨干网络进行特征提取，并使用全连接层进行分类。本文以 VGGNet 所构成的人脸识别器 VGGFace10 作为实验攻击对象，以此验证本文方法的有效性。VGGFace10 的输入是尺寸为 112×112×3 的人脸区域图像，输出人脸对应的 ID，人脸识别器损失 L_F 为交叉熵损失。根据人脸尺寸，眼镜掩膜的 L_1、L_2 取值分别为 3 和 5。

数据集包括 PubFig[24]中的 8 个 ID 和可在物理域实施攻击的来自本实验室的 2 个 ID，分别命名为 00 号至 09 号，平均每个 ID 包含有 300 张图片，如图 3 所示。每个 ID 图片数量按照 7∶3 随机划分为训练集和测试集。同时，将 Pubfig 剩余所有 ID 的测试数据作为库外陌生人脸数据，包含 34 个 ID 共 850 张人脸图片。

图 3　本实验中待 VGGFace10 识别的 10 个 ID(从左到右依次为 00 号-09 号)

Figure 3　The 10 IDs to be recognized by VGGFace10 (from left to right, No.00-No.09) in our experiments

数据预处理包括通过 Dlib 库进行人脸检测并裁剪、重新调整尺寸到 112×112×3，以及通过 2DFAN 网络进行人脸关键点检测。

4.2 生成器的构建

本文以修改的 DCGAN[25]生成器为例进行实验，其结构如表 1 所示，输入为 100 维的高斯向量，其后是全连接层、批量归一化层，随后 Reshape 成深度为 128 维的特征图，之后是四个反卷积层，其中最后一个反卷积层使用 Sigmoid 激活函数，其余使用 ReLu 激活函数，最后通过 Multiple 层乘以 255，输出尺寸为 112×112×3，取值为[0, 255]的矩形扰动矩阵。

表 1　本文采用的生成器结构

Table 1　The generator structure used in this work

Layer	输出尺寸	卷积核	步长	激活函数	输出通道数
Input	100	-	-	-	1
FC	6272	-	-	ReLu	1
Reshape	7×7	-	-	-	128
Deconv1	14×14	5	2	ReLu	128
Deconv2	28×28	5	2	ReLu	64
Deconv3	56×56	5	2	ReLu	32
Deconv4	112×112	5	2	Sigmoid	3
Multiple	112×112	-	-	-	3

4.3 数字域攻击实验设置

本文首先在数字域中攻击人脸识别系统，这是物理域攻击实现的前提。由于扮演攻击是特殊的误导攻击，实现了扮演攻击即实现了误导攻击，本文后续的攻击实验以扮演攻击的方式进行。为了展现本文提出的对抗样本攻击方法的有效性和通用性,设置了库内实验、库外实验和对比实验，同时将扮演成功的样本数占总样本数的比值定义为扮演成功率，作为评价扮演攻击效果的指标。

4.3.1 库内实验设置

使用 08 号 ID 的训练集数据训练生成器，使其能够分别扮演攻击其他 9 个 ID，然后使用训练好的生成器以 08 号 ID 的测试集数据构建数字域对抗样本并进行测试，进行数字域攻击实验。同时为了进一步验证本算法的有效性，调换攻击者和被攻击者的身份进行交叉实验，使用其他 9 个 ID 扮演攻击 08 号 ID

并进行测试。

4.3.2 库外实验设置

为了验证本文提出的对抗样本攻击方法的通用性, 即验证对抗扰动是否为通用扰动[26], 本文将生成器生成的对抗扰动添加到库外人脸图像中对VGGFace10进行攻击, 使其分别扮演00号-09号ID。生成器权重使用库内实验中的生成器权重, 具体而言, 扮演攻击08号使用其他9个ID扮演攻击08号所用的权重, 并对结果取平均值; 扮演攻击00号-07号和09号ID使用08号扮演攻击对应ID所用的权重。

4.3.3 对比实验设置

由于没有统一的对抗样本攻击指标, 且相关工作的实验设置存在较大差异, 为了与同领域相关工作进行对比, 本文仿真了CSS16[16]、AGNs[17]、AdvHat[18]论文, 并在相同的实验设置下进行08号ID分别扮演攻击其他9个ID的实验。

4.4 物理域攻击实验设置

通过08号ID扮演攻击其他ID进行物理域攻击实验, 具体而言, 08号ID在现实中佩戴使用惠普PageWide Pro 477dw彩色打印机打印的数字域眼镜对抗扰动, 然后通过谷客HD98电脑摄像头捕获图像, 输入人脸识别器进行识别。参照CSS16[16]、AGNs[17]的实验设置, 本文采样了48副数字域眼镜(涵盖了可生成的大部分样式的眼镜), 并选择其中扮演成功率最高的25%, 即12副眼镜打印进行物理域攻击, 计算其扮演成功率。为验证打印分数损失和模拟物理域数据增强的有效性, 进行了消融实验。同时在相同条件下进行了与CSS16[16]、AGNs[17]、AdvHat[18]等物理域攻击方法的对比实验。其中, AdvHat[18]采用的攻击方式是通过打印矩形对抗扰动粘贴到帽子上。

4.5 实验参数设置

概率分数损失 L_S 的参数 ε 取值为 1×10^{-2}, 对于物理域攻击, 制作颜色调色盘时的采样间隔 i 取5, 从此获得 51^3 种颜色值, 而各种数据增强的参数设置如表2所示。参照AGNs[17]的实验设置, 采用网格搜索来设置损失函数各个分量的权重 α, λ, 其中 α, $\lambda\in$ {0.1,0.5,1.0,1.5}。通过08号ID扮演攻击其他9个ID的数字域和物理域实验, 发现当 α=0.1, λ=0.1 时, 获得最高扮演攻击成功率。参数 α 的取值印证了3.2节中的分析, 即相比人脸识别损失 L_F, 概率分数损失 L_S(权重为1)更能反映判断过程的本质。参数 λ 的取值则体现了本算法以实现攻击为主要目的, 打印分数损失是为了模拟物理域条件引入的一种正则约束。

表2 数据增强操作及参数

Table 2 Data enhancement operations and parameters

模拟过程	数据增强操作	参数
打印眼镜	重采样	$\tau\in[1,2]$
佩戴眼镜	水平平移	$h\in[-10,+10]$
	垂直平移	$v\in[-10,+10]$
	中心旋转	$d\in[-5,+5]$
光照变化	线性变化	$a\in[0.5,1.5]$ $b\in[-20,+20]$
	伽马变换	$\gamma\in[0.5,1.5]$
捕获人脸	高斯噪声	$\mu\in[-0.5,+0.5]$ $\sigma\in[+0.5,+1.5]$
	压缩噪声	$q\in[50,100]$

5 实验结果

5.1 人脸识别器训练结果

使用训练集训练VGGFace10, 并用测试集进行测试, 每个ID测试结果如表3所示, 平均识别准确率为98.55%, 完成攻击对象的准备。

5.2 数字域攻击实验结果

5.2.1 库内实验结果

使用08号ID作为攻击者, 扮演攻击其他9个库内ID的实验结果如表4(左)所示, 其平均扮演成功率为97.85%, 表明08号ID能够以较高的成功率扮演其他9个库内ID, 表明了本文所提出的对抗样本攻击算法的有效性和高成功率。另外, 使用其他9个ID扮演攻击08号ID进行交叉验证, 其实验结果如表4(右)所示, 平均扮演成功率为99.49%, 进一步表明了本文提出方法的有效性和高成功率。

以08号ID扮演攻击02号ID和09号ID为例, 结果如图4左侧两列所示, 每幅图中人脸位置框上是识别结果ID, 下面是识别置信度, 蓝色表示VGGFace10正确识别, 红色表示成功扮演攻击目标对象。

5.2.2 库外实验结果

使用库外数据进行扮演攻击00号-09号ID的实验结果如表5所示, 库外人脸添加对抗扰动后, 也有一定概率能扮演攻击成功, 说明了本算法生成的眼镜形状对抗扰动属于通用对抗扰动, 也揭露了陌生人通过对抗样本攻破人脸识别系统的可能性。

5.2.3 对比实验结果

本算法与同领域典型工作CSS16[16]、AGNs[17]、AdvHat[18]进行对比的实验结果如表6所示, 可看出, 本文算法的扮演成功率相近甚至更高, 说明了本文

算法的有效性。

5.3 物理域攻击实验结果

08 号在物理域扮演攻击其他 ID 的结果如表 7 所示, 前 3 列分别展示本算法不使用打印分数损失(本算法-1)、只使用打印分数损失但不使用模拟物理域数据增强(本算法-2)、同时使用打印分数损失和模拟物理域数据增强(本算法-3)的实验结果, 其中最终结果(本算法-3)的扮演成功率最高, 验证了本文提出算法在物理域实现的有效性和可行性, 以及打印分数损失和模模物理域数据增强对物理域攻击的关键作用。后 3 列分别展示了 CSS16[16]、AGNs[17]、AdvHat[18]论文实验结果, 可以看出本算法与之相比有更高的扮演成功率, 进一步表明了本算法的有效性。另外, 与 5.2 节的数字域扮演攻击结果(表 6)相比, 各个算法的物理域结果均有所下降, 体现了物理域与数字域的差异, 证实了实现物理域攻击确实存在较大的挑战性。尽管本文算法通过打印分数损失和模拟物理域数据增强减小了这种差异, 但无法完全消除, 而且物理域和数字域还存在其他的差异难以一一模拟, 这一定程度上导致了物理域攻击成功率的下降, 同时也指明了未来的改进方向。

以 08 号 ID 扮演攻击 02 号 ID 和 09 号 ID 的结果为例, 物理域攻击结果如图 4 右侧三列所示, 可看出 08 号不佩戴对抗眼镜和佩戴空白眼镜, 人脸识别器均能正确识别, 说明了人脸识别器的精确性; 08 号佩戴了图 4(e)和图 4(j)展示的打印对抗眼镜后(如图 4(d)和图 4(i)所示), VGGFace10 分别错误地识别为 02 号和 09 号, 扮演攻击成功。

表 3 VGGFace10 对每个 ID 的测试准确率

Table 3 Test accuracy rate of VGGFace10 for each ID

ID	准确率	ID	准确率
00 号	97.75%	05 号	96.71%
01 号	97.44%	06 号	97.35%
02 号	100.00%	07 号	98.43%
03 号	97.83%	08 号	100.00%
04 号	100.00%	09 号	100.00%

6 结论

人脸识别系统的安全性是学术界和产业界共同关注的热点问题之一。现有研究大多着眼于数字域的对抗样本攻击, 本文突破此局限性, 将数字域生成的对抗扰动复现至物理域, 成功实现对人脸识别系统的物理域对抗样本攻击。具体地, 本文提出了一种根据人脸关键点构造特定形状的对抗掩膜的算法, 首先设计了对抗损失函数生成数字域的对抗扰动, 然后引入打印机分数损失函数将数字域的对抗扰动迁移到物理域, 为更好解决光照等实际环境因素对攻击效果的影响, 提出通过模拟物理佩戴和光照变化的方式进行数据增强。实验结果表明本文方法不但可在数字域以高成功率扮演攻击人脸识别系统 VGGFace10, 还可成功地在物理域实现攻击, 且实现过程只需一台普通商用彩色打印机, 无需其他复杂的专业设备。本文在成功实施对人脸识别系统的物理域攻击的同时, 较大提高了将数字域对抗样本进行实物化的可操作性。除了眼镜, 采用本文方法还可方便地生成其他形式的物理攻击载体。将来的改进方向包括考虑在数据增强过程中考虑更多类型的物理域差异, 进一步提高物理域攻击的成功率, 以及改进眼镜掩膜的构建方法使其外观更接近现实生活中的眼镜。

表 4 数字域库内实验结果

Table 4 Intra-dataset experimental results of digital domain attacks

扮演攻击	扮演成功率	扮演攻击	扮演成功率
08 号-->00 号	96.48%	00 号-->08 号	100.00%
08 号-->01 号	97.21%	01 号-->08 号	97.44%
08 号-->02 号	98.00%	02 号-->08 号	98.27%
08 号-->03 号	100.00%	03 号-->08 号	100.00%
08 号-->04 号	97.66%	04 号-->08 号	100.00%
08 号-->05 号	98.88%	05 号-->08 号	100.00%
08 号-->06 号	95.54%	06 号-->08 号	100.00%
08 号-->07 号	98.21%	07 号-->08 号	100.00%
08 号-->09 号	98.66%	09 号-->08 号	99.72%
平均值	**97.85%**	平均值	**99.49%**

表 5 数字域库外实验结果

Table 5 Out-of-database experimental results of digital domain attacks

扮演攻击	扮演成功率	扮演攻击	扮演成功率
陌生人-->00 号	34.06%	陌生人-->05 号	42.19%
陌生人-->01 号	50.94%	陌生人-->06 号	30.15%
陌生人-->02 号	99.38%	陌生人-->07 号	37.50%
陌生人-->03 号	30.75%	陌生人-->08 号	94.96%
陌生人-->04 号	47.50%	陌生人-->09 号	30.31%

表 6 数字域对比实验结果

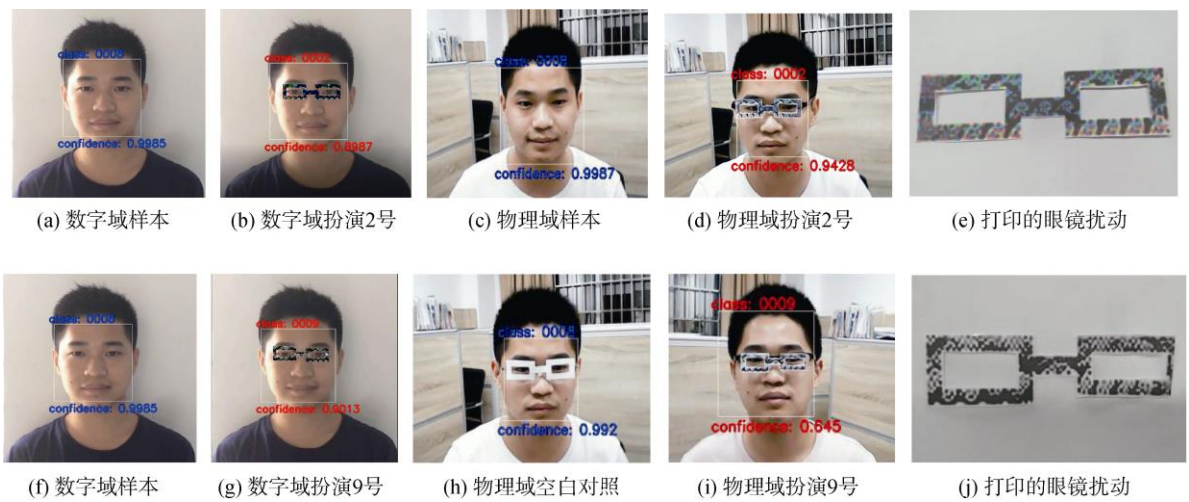
Table 6 Comparison of experimental results of digital domain attacks

扮演攻击	扮演成功率			
	CCS16[16]	AGNs[17]	AdvHat[18]	本算法
08 号-->00 号	96.12%	95.32%	94.00%	96.48%
08 号-->01 号	97.32%	96.76%	98.00%	97.21%
08 号-->02 号	96.43%	95.63%	97.00%	98.00%
08 号-->03 号	95.83%	96.43%	96.00%	100.00%
08 号-->04 号	97.24%	98.33%	98.00%	97.66%
08 号-->05 号	97.65%	97.97%	93.00%	98.88%
08 号-->06 号	95.87%	96.23%	96.00%	95.54%
08 号-->07 号	95.71%	96.87%	99.00%	98.21%
08 号-->09 号	96.86%	97.96%	97.00%	98.66%
平均值	96.55%	96.82%	96.44%	**97.85%**

表 7 物理域 08 号 ID 扮演攻击其他 ID 实验结果

Table 7 Experimental results of ID 08 impersonating and attacking other IDs in physical domain

扮演攻击	扮演成功率					
	本算法-1	本算法-2	本算法-3(最终)	CCS16[16]	AGNs[17]	AdvHat[18]
08 号-->00 号	66.67%	75.00%	75.00%	75.00%	100.00%	66.67%
08 号-->01 号	75.00%	75.00%	83.33%	75.00%	66.67%	100.00%
08 号-->02 号	83.33%	83.33%	91.67%	83.33%	75.00%	91.67%
08 号-->03 号	83.33%	91.67%	91.67%	91.67%	83.33%	83.33%
08 号-->04 号	75.00%	83.33%	83.33%	75.00%	91.67%	66.67%
08 号-->05 号	91.67%	100.00%	100.00%	100%	83.33%	75.00%
08 号-->06 号	75.00%	75.00%	66.67%	75.00%	75.00%	83.33%
08 号-->07 号	75.00%	83.33%	83.33%	66.67%	75.00%	75.00%
08 号-->09 号	91.67%	91.67%	100.00%	83.33%	83.33%	66.67%
平均值	79.63%	84.25%	**86.11%**	80.56%	81.49%	78.70%



(a) 数字域样本 (b) 数字域扮演2号 (c) 物理域样本 (d) 物理域扮演2号 (e) 打印的眼镜扰动

(f) 数字域样本 (g) 数字域扮演9号 (h) 物理域空白对照 (i) 物理域扮演9号 (j) 打印的眼镜扰动

图 4 08 号 ID 的数字域与物理域扮演攻击结果展示

Figure 5 Illustration of digital domain and physical domain impersonation attacks of ID 08

参考文献

[1] Parkhi O M, Vedaldi A, Zisserman A. Deep Face Recognition[C]. *Proceedings of the British Machine Vision Conference 2015*, 2015, 1(3): 6.

[2] Deng J K, Guo J, Yang J, et al. ArcFace: Additive Angular Margin Loss for Deep Face Recognition[C]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021: 5962-5979.

[3] Zhang S S, Zuo X, Liu J W. The Problem of the Adversarial Examples in Deep Learning[J]. *Chinese Journal of Computers*, 2019, 42(8): 1886-1904.
(张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. *计算机学报*, 2019, 42(8): 1886-1904.)

[4] Huang L F, Gao C Y, Zhou Y Y, et al. Universal Physical Camouflage Attacks on Object Detectors[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 717-726.

[5] Li H L, Wang Y F, Xie X F, et al. Light can Hack your Face! Black-Box Backdoor Attack on Face Recognition Systems[EB/OL]. 2020: arXiv: 2009.06996. https://arxiv.org/abs/2009.06996

[6] Xiao Z H, Gao X F, Fu C L, et al. Improving Transferability of Adversarial Patches on Face Recognition with Generative Models[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 11840-11849.

[7] Andriushchenko M, Croce F, Flammarion N, et al. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search[C]. *Computer Vision - ECCV 2020*: *16th European Conference*, *Glasgow*, *UK*, *August 23-28*, *2020*, *Proceedings*, *Part XXIII*, 2020: 484-501.

[8] Tong L, Chen Z Z, Ni J C, et al. FACESEC: A Fine-Grained Robustness Evaluation Framework for Face Recognition Systems[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 13249-13258.

[9] Yin B J, Wang W X, Yao T P, et al. Adv-Makeup: A New Imperceptible and Transferable Attack on Face Recognition[C]. *The Thirtieth International Joint Conference on Artificial Intelligence*, 2021: 1252-1258.

[10] Zhang B W, Tondi B, Barni M. Adversarial Examples for Replay Attacks Against CNN-Based Face Recognition with Anti-Sp fing Capability[J]. Computer Vision and Image Understanding, 2020, 197/198: 102988.

[11] Vakhshiteh F, Nickabadi A, Ramachandra R. Adversarial Attacks Against Face Recognition: A Comprehensive Study[J]. *IEEE Access*, 2021, 9: 92735-92756.

[12] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[EB/OL]. 2013: arXiv: 1312.6199. https://arxiv.org/abs/1312.6199

[13] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. https://arxiv.org/abs/1412.6572

[14] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. *Communications of the ACM*, 2020, 63(11): 139-144.

[15] Yang L, Song Q, Wu Y Q. Attacks on State-of-the-Art Face Recognition Using Attentional Adversarial Attack Generative Network[J]. *Multimedia Tools and Applications*, 2021, 80(1): 855-875.

[16] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 1528-1540.

[17] Sharif M, Bhagavatula S, Bauer L, et al. A General Framework for Adversarial Examples with Objectives[J]. *ACM Transactions on Privacy and Security*, 2019, 22(3): 1-30.

[18] Komkov S, Petiushko A. AdvHat: Real-World Adversarial Attack on ArcFace Face ID System[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 819-826.

[19] Pautov M, Melnikov G, Kaziakhmedov E, et al. On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System[C]. *2019 International Multi-Conference on Engineering*, *Computer and Information Sciences*, 2020: 391-396.

[20] Nguyen D L, Arora S S, Wu Y H, et al. Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020: 3548-3556.

[21] Bai J S, Li Y C, Lin L, et al. Mobile Terminal Implementation of Image Filtering and Edge Detection Based on OpenCV[C]. *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications*, 2020: 214-218.

[22] Luo Y, Chen L. Research on the Influence of Gamma Correction Method on Local Feature Descriptors[C]. *2020 IEEE 4th Information Technology*, *Networking*, *Electronic and Automation Control Conference*, 2020: 2584-2588.

[23] Hendrycks D, Dietterich T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations[EB/OL]. 2019: arXiv: 1903.12261. https://arxiv.org/abs/1903.12261

[24] Kumar N, Berg A C, Belhumeur P N, et al. Attribute and Simile Classifiers for Face Verification[C]. *2009 IEEE 12th International Conference on Computer Vision*, 2010: 365-372.

[25] Ono H, Suzuki S. Data Augmentation for GrossMotor- ActivityRecognition Using DCGAN[C]. *2020 IEEE/SICE International Symposium on System Integration*, 2020: 440-443.

[26] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

蔡楚鑫 于 2019 年在华南理工大学信息工程专业获得学士学位。现在华南理工大学信息与通信工程专业攻读硕士学位。研究领域为多媒体信息安全。研究兴趣包括: 人脸对抗攻击、人脸欺诈检测。Email: eechuxincai@mail.scut.edu.cn

王宇飞 于 2018 年在华南理工大学信息与通信工程专业获得博士学位。现任职于中新国际联合研究院。研究领域为多媒体信息安全。研究兴趣包括: 人脸欺诈检测、人工智能应用。Email: w.yf05@mail.scut.edu.cn

章烈剽 于 2007 年在武汉理工大学控制理论与控制工程专业获得硕士学位。现任广州广电卓识智能科技有限公司技术总监。研究领域为计算机应用。研究兴趣包括: 人脸欺诈检测、图像处理及人工智能在金融行业的工程应用。Email: zlpiao@grgbanking.com

卓思超 于 2020 年在华南理工大学光电信息科学与工程专业获得学士学位。现在华南理工大学电子信息专业攻读硕士学位。研究领域为多媒体信息安全。研究兴趣包括: 篡改人脸视频检测。Email: 202021012813@mail.scut.edu.cn

张娟苗 于 2008 年在中北大学测控技术与仪器专业获得学士学位。现任广州广电运通金融电子股份有限公司高级项目经理。研究领域为计算机应用。研究兴趣包括: 计算机在金融领域的应用。Email: zjmiao@grgbanking.com

胡永健 于 2002 年在华南理工大学通信与信息系统专业获得博士学位。现在华南理工大学电子与信息学院教授。研究领域为多媒体信息安全、图像处理、人工智能及其应用。研究兴趣包括: 信息隐藏、人脸欺诈检测。Email: eeyjhu@scut.edu.cn