ISSN 1003-501X CN 51-1346/O4

光电工程

Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

针对人脸识别卷积神经网络的局部背景区域对抗攻击

张晨晨，王帅，王文一，李迪然，李南，鲍华，李淑琪，高国庆

引用本文:

张晨晨，王帅，王文一，等. 针对人脸识别卷积神经网络的局部背景区域对抗攻击[J]. 光电工程，2023， **50**(1): 220266.

Zhang C C, Wang S, Wang W Y, et al. Adversarial background attacks in a limited area for CNN based face recognition[J]. *Opto-Electron Eng*, 2023, **50**(1): 220266.

https://doi.org/10.12086/oee.2023.220266

收稿日期: 2022-10-17; 修改日期: 2022-12-21; 录用日期: 2022-12-29

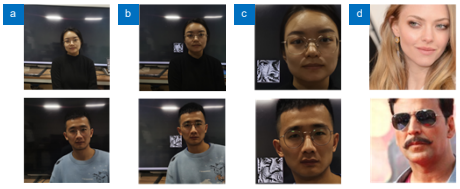
Opto-Electronic Engineering
光 电 工 程

Article

2023 年, 第 50 卷, 第 1 期

DOI: 10.12086/oee.2023.220266

针对人脸识别卷积神经网络的局部背景区域对抗攻击



张晨晨[1,2]，王　帅[1,2*]，王文一[2]，李迪然[2]，
李　南[1,2]，鲍　华[3,4]，李淑琪[3,4]，高国庆[3,4]

[1]电子科技大学长三角研究院，浙江 衢州 324000；
[2]电子科技大学，四川 成都 610000；
[3]中国科学院自适应光学重点实验室，四川 成都 610209；
[4]中国科学院光电技术研究所，四川 成都 610209

摘要：基于卷积神经网络 (CNN) 的识别器，由于其高识别率已经在人脸识别中广泛应用，但其滥用也带来隐私保护问题。本文提出了局部背景区域的人脸对抗攻击 (BALA)，可以作为一种针对 CNN 人脸识别器的隐私保护方案。局部背景区域添加扰动克服了现有方法在前景人脸区域添加扰动所导致的原始面部特征损失的缺点。BALA 使用了两阶段损失函数以及灰度化、均匀化方法，在更好地生成对抗块的同时提升了数字域到物理域的对抗效果。在照片重拍和场景实拍实验中，BALA 对 VGG-FACE 人脸识别器的攻击成功率 (ASR) 比现有方法分别提升 12% 和 3.8%。

关键词：人脸识别；CNN；对抗攻击；背景；物理域

中图分类号：TP391.41　　**文献标志码：**A

张晨晨，王帅，王文一，等. 针对人脸识别卷积神经网络的局部背景区域对抗攻击 [J]. 光电工程，2023，**50**(1): 220266
Zhang C C, Wang S, Wang W Y, et al. Adversarial background attacks in a limited area for CNN based face recognition[J]. *Opto-Electron Eng*, 2023, **50**(1): 220266

Adversarial background attacks in a limited area for CNN based face recognition

Zhang Chenchen[1,2], Wang Shuai[1,2*], Wang Wenyi[2], Li Diran[2], Li Nan[1,2], Bao Hua[3,4], Li Shuqi[3,4], Gao Guoqing[3,4]

[1]Yangtze Delta Region Institute of University of Electronic Science and Technology of China, Quzhou, Zhejiang 324000, China;
[2]University of Electronic Science and Technology of China, Chengdu, Sichuan 610000, China;
[3]Key Laboratory on Adaptive Optics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[4]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China

Abstract: Recognizers based on the convolutional neural networks (CNN) have been widely used in face recognition because of their high recognition rate. But its abuse also brings privacy protection problems. In this paper, we propose a local background area-based face confrontation attack (BALA), which can be used as a privacy protection scheme for CNN face recognizer. Adding disturbance in the local background region overcomes

收稿日期：2022-10-17；收到修改稿日期：2022-12-21
基金项目：衢州市财政专项资助科研项目 (2022D025)
*通信作者：王帅，wangshuai0601@uestc.edu.cn。

the loss of original facial features caused by adding disturbance to the foreground face region in existing methods. BALA uses a two-stage loss function, graying, and homogenization methods to better generate adversarial blocks and improve the adversarial effect after digital to physical domain conversion. In the photo retake and live shot experiments, BALA's attack success rate (ASR) against the VGG-FACE face recognizer is more than 12% and 3.8% higher than the current methods.

Keywords: face recognition; CNN; adversarial attack; background; physical domain

1 引 言

深度学习方法由于其突出的性能，已经广泛应用到各种人工智能 (AI) 系统中，比如自然语言处理[1]、人脸识别[2]、图像分类[3]、自动驾驶[4]、信号加密[5]、逆向设计[6] 等。人脸识别在融入到每个人生活的方方面面的同时，也带来了很多社会问题，比如人脸数据泄露[7]、非法出售人脸照片等。为此，很多平台都采取了保护人脸隐私的相关措施，例如，Facebook 最近宣布将关闭网络面部识别系统并删除其数据集，其中包括超过 10 亿人的面部扫描数据[8]。阿里巴巴、腾讯等多家中国企业也开始响应中国个人信息保护法。同时，许多学术研究人员也意识到了这个问题[9-10]。

众所周知，各种卷积学习网路 (CNN) 对带有精心设计的扰动的对抗样本非常脆弱[11]。这种对抗样本可干扰未经授权的 AI 面部识别，使其得到错误结果，从而保护隐私。

具体来说，对抗攻击是一种通过在数字图片上 (称为数字域攻击[12-15]) 或在真实场景中 (称为物理域攻击[16-18]) 添加对抗扰动来产生对抗样本[12] 的方法。在数字域攻击中，图像数据中的扰动在图像输入识别器之前添加。许多带有人眼不可见的扰动的数字对抗样本生成方法会导致识别器产生错误的输出[13]。然而，这些数字域方法带有明显的局限性。首先，添加的扰动是较轻微的，导致它们在数字域到物理域的转化，例如打印、重新拍摄照片、重新拍摄视频等过程中通常会失去对抗效果。其次，由于有些时候也需要进行必要的数字图像转换，例如压缩、采样等，这些过程通常会使数字域对抗样本失去对抗效果。在物理域攻击中，则只能在传感器成像前添加扰动，这使其具有更好的实际效果。物理域攻击也有一些对抗样本生成方法。例如，Mikhail 等人研究发现，在面部区域戴一些面部扰动饰品可以导致人脸识别器输出错误的识别结果[19]。除了这项工作外，还有一些研究使用在面部区域上添加扰动的方式来误导人脸识别器[20-21]。尽管这些方法有效，但在面部上粘贴奇怪的扰动块可能会妨碍正常的面部观察。此外，这些方法的效率也很低，因其首先要生成扰动补丁，然后再将其贴到对应的位置上。

本文通过在物理域添加背景对抗扰动块，从而在实现对抗非授权人脸识别的同时，还能保持所有原始面部特征。这种方案不仅可以误导不安全的人脸识别器，还能克服现有对抗样本生成方法在前景人脸区域添加显著扰动信号所导致的原始面部特征损失，保证不影响有授权的面部观察。Brown 等人已经展示了对于物体识别的物理域背景对抗攻击的可行性，称之为 Adv-patch[16]。然而，据我们所知，目前还没有关于面部物理域背景对抗攻击的工作。面部背景对抗攻击在物理域实现的困难主要是由于每张图片可添加背景扰动的区域是有限的，因为在人脸剪切过程中大部分背景扰动会被裁剪掉。为了在一定程度上解决以上提到的问题，本文提出了在局部区域的物理域背景面部对抗攻击的方法，称之为 BALA。BALA 的优势总结如下：

1) 扰动块不会覆盖任何人脸区域，甚至被拍摄的前景人物都可以不知道它的存在；

2) BALA 只在背景的一小部分引入扰动，并且可以实现较高的错误分类概率；

3) 由于采用扰动块灰度化和相邻像素平均化的方法，所提出的对抗性扰动对数字域到物理域的转换过程具有鲁棒性；

4) 在扰动生成的过程中应用了两种不同的损失函数能够保证快速收敛和鲁棒性。

2 理论推导

2.1 相关工作

2.1.1 数字域攻击

目前，数字域对抗攻击法主要可以分为基于梯度[7,21] 和基于 GAN[22] 类方法，快速梯度符号法 (FGSM)[13] 是最具代表性的基于梯度的方法。此外，还有很多基于梯度更新的方法，例如多步迭代：投影

梯度下降法 (PGD)[12] 和跳跃梯度下降法 (SGD)[15]。梯度方法中的 Deep Fool[23] 可以计算最小扰动水平以误导卷积神经网络；One Pixel[24] 使用差分进化算法通过仅更改几个输入图像像素的值来实现高误导率；LaVAN[14] 则是使用尽可能小的矩形区域来误导识别器。

2.1.2 物理域攻击

Kurakin 等人研究发现数字域攻击的部分对抗样本能在打印或重拍照片中保留其对抗有效性[17]。期望转换法 (EOT)[18] 在一定程度上能够产生一个更加鲁棒的对抗样本来应对各种物理变化，比如高斯噪声，视角变化，和其他常见变化等。Adv-patch[16] 就是一种非常有代表性的使用 EOT 来产生一个相对高强度的扰动去抵抗一系列物理变化的方法。在针对人脸识别的前景物理域对抗攻击中，对抗的眼镜框架 (例如图 1 板块 A)[25]、T 恤[26]、在前额上的贴片[20]，都能以较高的成功率误导人脸识别系统，但会影响清晰的面部观察。Mikhail 等人使用生成的扰动补丁 (图 1 板块 A 中的补丁) 误导 ArcFace-100 人脸识别系统，该补丁可以打印并粘贴作为人脸的一种属性[19]。

2.2 本文方法

2.2.1 方法概述

本文提出的方法主要流程如图 1 的板块 B 或 C 所示，图 1 的板块 B 是物理域背景攻击的流程，板块 C 是数字域背景攻击的流程。根据能否获取攻击神经网络模型结构和网络模型权重参数可以将对抗攻击分为白盒攻击[19-20] 和黑盒攻击[27]。在本文中，我们使用白盒攻击方案，即已知 CNN 人脸识别器 (F) 的神经网络结构和权重参数。给定一张裁剪的人脸图片 $\boldsymbol{x}\in \mathrm{R}^{c*m*n}$，$m$和$n$是人脸图像的宽和高，$c$是颜色通道的数量 (一般$c=3$)，人脸图片的原始类别为$l_{\mathrm{org}}$。$F$ 能够将$\boldsymbol{x}$映射到一组向量$F(\boldsymbol{x})=[p_1,p_2,p_3,\cdots,p_k]$，其中$p_j(j=1,2,3,\cdots,k)$为$\boldsymbol{x}$属于每一类的概率，$k$是所有类的总数，$J$ 表示 F 对 $\boldsymbol{x}$ 映射的类别。如式 (1) 所示，实验中将人脸识别定义为函数$y(F,\boldsymbol{x})$：

$$\begin{aligned} &y(F,\boldsymbol{x})=J,\text{where}\\ &p_J=\max\{p_j,j=1,2,3,\cdots,k\}. \end{aligned} \tag{1}$$

当$\boldsymbol{x}$被识别为正确的类别时，$y_{\mathrm{pred}}(F,\boldsymbol{x})$将会等于$l_{\mathrm{org}}$。本文使用基于梯度下降的方法产生扰动$\boldsymbol{\delta}\in \mathrm{R}^{c*m*n}$，目标背景区域的对抗扰动块将通过在 2.2.2 节介绍生成掩膜$\boldsymbol{M}$的方法进行裁剪。图 2 显示了生成对抗扰动块的流程，该流程由三部分组成：第一部分是找到合适的位置并生成掩膜$\boldsymbol{M}$，第二部分是在数字域中基于反向梯度传播的方法不断迭代对抗样本，第三部分是使用灰度化和平均化进一步改进对抗扰动块，使其在物理域更加鲁棒。在实验生成扰动块的过程中考虑了常见的数字域到物理域的转换$T(\boldsymbol{x})$，例如亮度变化、饱和度变化、缩放和噪声的变化等。除此之外，第二部分和第三部分会迭代到对抗样本能够成功误导人脸识别网络 F 或达到指定迭代次数。

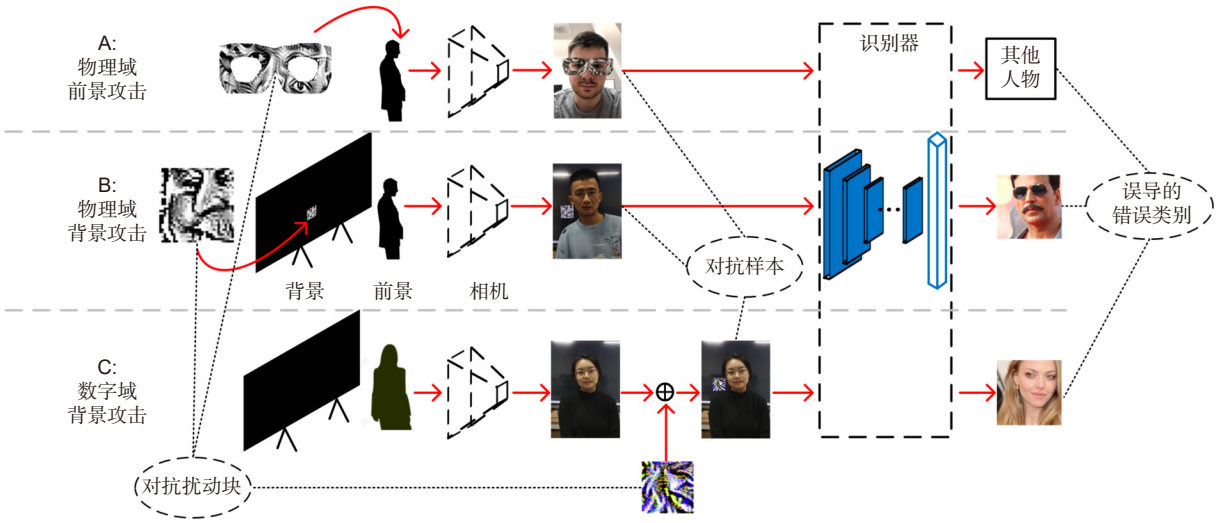


图 1 人脸对抗攻击流程图。板块 A 带有扰动块 (来自于 Mikhail 等人[19] 的补丁图片) 的物理域前景攻击；板块 B 为物理域背景攻击；板块 C 为数字域背景攻击。每种攻击方法目的在于使用对抗样本误导人脸识别器从而产生错误的分类结果

Fig. 1 Scheme of facial adversarial attacks. Panel A is a physical foreground attack with an adversarial patch (patch image from Mikhail et al.[19]); Panel B is a physical adversarial background attack, and panel C is a digital adversarial background attack. Every attack approach aims to mislead a face recognizer with an incorrect class using adversarial examples

2.2.2 掩膜生成

本文提出基于下颌线[28]生成扰动掩膜 ($\boldsymbol{M}$) 方法，如图 3 所示。该方案可以有效避免扰动在人脸裁剪过程中被裁剪掉。大多数现有的人脸检测器都会产生矩形锚点，其中包括人脸和部分周围背景[29]。很容易观察到，面部下巴两侧的背景区域相对较大，适合嵌入对抗扰动块。

具体来说，本文使用面部检测算法[28]获得下颌点 (图 3(b) 中的蓝色点) 的最大外接矩形 (图 3(b) 中的绿色矩形) 的四个角点作为候选点。根据扰动块的尺寸，在四个候选点中选择最合适的点 (图 3(b) 中的红色点)，并将掩膜中对应区域置为 1(图 3(c) 中白色区域)，掩膜的其他部分被置为 0(图 3(c) 中的黑色区域)。

2.2.3 扰动生成

令 $\boldsymbol{x}'$ 为迭代过程中的对抗样本，$\boldsymbol{x}^{\text{adv}}$ 为迭代结束后最终的对抗样本，$\boldsymbol{x}^{\text{adv}}$ 被识别器识别为 y_{anh}。在得到合适的掩膜后，使用式 (2) 计算对抗扰动，将扰动添加至掩膜区域后重新计算此时图像的分类结果，通过重复该过程，直到分类器分类错误或达到迭代的次数上限。对抗扰动的计算由梯度反向传播得到，如下所示：

$$\boldsymbol{\delta} = \varepsilon^* \operatorname{sign}(\partial L/\partial \boldsymbol{x}') \,, \tag{2}$$

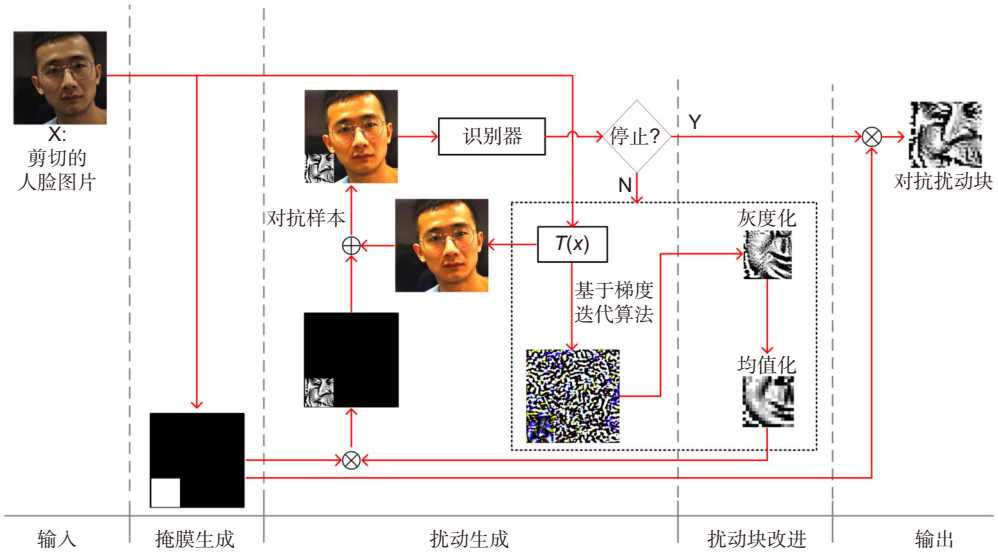


图 2 BALA 产生一个对抗扰动块的流程。主要包括三个部分，分别是掩膜生成、扰动生成和扰动块的改进。$T(\bullet)$ 表示一系列的物理变换

Fig. 2 The scheme of generating an adversarial patch in BALA. This scheme mainly consists of three parts, mask generation, perturbation generation, and perturbation further improving. $T(\bullet)$ represents a set of transformations

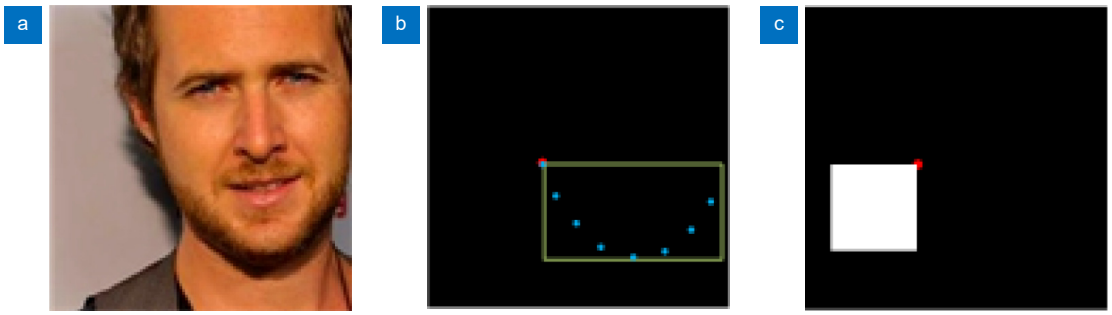


图 3 掩膜制作的示例。(a) 为一张裁剪的人脸图片；(b) 蓝色的点是脸部轮廓下颌线的采样点，绿色的线代表其最大外接矩形框；(c) 带有白色补丁的掩膜，白色补丁代表着对抗扰动块的位置，红色的点代表具体的候选点

Fig. 3 The illustration of mask generation. (a) Cropped face image; (b) Blue points are mandible points of the face, and the green lines represent the maximum outside rectangular; (c) The mask with a white patch represents the location of an adversarial patch, and the red point represents the specific candidate

其中，ε是控制每次迭代扰动强度的超参数，L 为损失函数。

如果希望加速上述迭代的收敛，可以选择较大的梯度损失函数。然而，这样将会导致很难收敛到精确的结果，而且人脸识别器预测的类别将会在临近分类阈值时反复跳变。与之相反，选择一个较小的梯度损失函数可能会收敛得更加精确，但收敛过程通常会花费更多的时间。因此，本文提出在不同的收敛阶段使用两个不同的损失函数，从而可以显著减少产生有效对抗扰动块的迭代次数，如图 4 所示。当人脸识别器对于$\boldsymbol{x}'$的输出仍为$l_{\rm org}$时，使用$L_{\rm s1}$作为式 (2) 中的 L：

$$L_{\rm s1} = \log[p_{y_{\rm org}}(F,\boldsymbol{x}')]\,. \tag{3}$$

去加速收敛过程，如图 4 的步骤 1 到步骤 2 所示。当$l_{\rm pred} \neq l_{\rm org}$时，使用$L_{\rm s2}$作为式 (2) 中的$L$：

$$L_{\rm s2} = p_{y_{\rm anh}}(F,\boldsymbol{x}')\,, \tag{4}$$

去微调$\boldsymbol{x}'$，并在点 2 处从 $p_{y\rm org}$ 到 $p_{y\rm anh}$ 跳变， 如图 4 的步骤 2 到步骤 3 所示。

2.2.4 扰动块的改进

为了使对抗扰动在从数字域到物理域转换的过程中继续有效，BALA 对对抗扰动块进一步改进。在本文的实验场景中，以普通电子屏幕作为物理背景。主要原因包括以下几个方面：首先，电子屏幕可以根据人脸的位置动态调整及显示对抗样本或扰动块；其次，该场景可以模拟在现实生活中可以自由设置背景的在线会议；第三，这个场景可以方便实现屏幕图像的重拍实验。

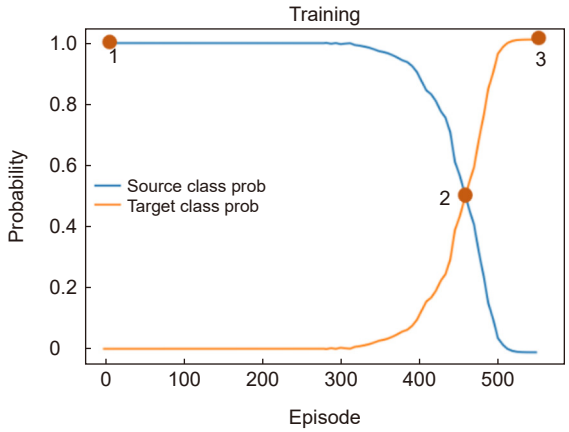


图 4　对抗扰动块的产生过程，在迭代过程中使用两种不同的损失函数。$L_{\rm s1}$ 和 $L_{\rm s2}$ 分别从步骤 1 到步骤 2($p_{y\rm org}$) 和从步骤 2 到步骤 3($p_{y\rm anh}$) 的过程中的损失函数，并在点 2 处从 $p_{y\rm org}$ 到 $p_{y\rm anh}$ 的跳变

Fig. 4　The generation process of an adversarial patch. Two different loss functions are used in iterations. $L_{\rm s1}$ and $L_{\rm s2}$ are used in stage of step 1 to step 2 ($p_{y\rm org}$) and step 2 to step 3 ($p_{y\rm anh}$), respectively. There is a change of $p_{y\rm org}$ to $p_{y\rm anh}$ at point 2

实验中的某个数字域的彩色对抗块如图 5(a) 所示，由于电子屏幕的显示特性 (电子噪声和屏幕不同的色域空间)，屏幕上的彩色扰动块在通过相机重拍之后

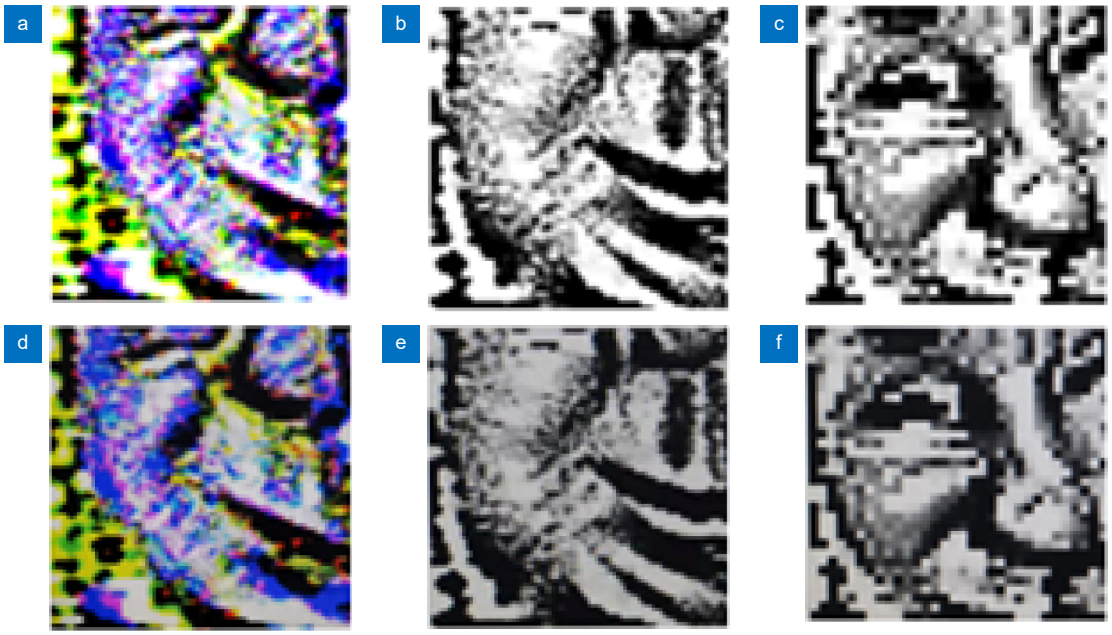


图 5　不同的 BALA 对抗扰动块。(a) 不带灰度化和均一化的彩色块；(b) 不带均一化的灰度块；(c) 带均一化的灰度块；(d), (e), (f) 分别对应 (a),(b),(c) 的屏幕重拍照片

Fig. 5　The diverse BALA adversarial patches. (a) Generated color patch without graying and averaging; (b) Generated gray patch without averaging; (c) Generated gray patch with averaging; Images (d), (e), (f) are corresponding re-taken images (a), (b), (c), respectively

会造成失真，如图 5(b) 所示。本文对原始图像块 ($\boldsymbol{patch}_{\rm dig}$) 与重拍之后的图像块 ($\boldsymbol{patch}_{\rm re}$) 之间的失真定义为$\omega$：

$$\omega = \text{mean}(|\boldsymbol{patch}_{\rm re} - \boldsymbol{patch}_{\rm dig}|) \tag{5}$$

通过实验发现，所有彩色图片的失真值ω，即图 5(a) 和图 5(b) 之间的失真值，远高于灰度的图 5(c) 和图 5(d) 的失真值ω。所以 BALA 使用灰度化过程来进一步提高扰动块的对抗性；虽然灰度化会损失部分扰动信息，但 BALA 可以使用灰度值直接计算对抗效果和迭代来消除这部分影响。另外，在迭代过程中，BALA 在对抗扰动块中也做了像素值均一化以保证在数字域 (图 5(e)) 和重拍的对抗样本 (图 5(f)) 中都能进一步提升对人脸识别器的误导成功率。值得注意的是，BALA 只在对抗扰动块，也就是 $\boldsymbol{M}=1$的区域中采用了灰度化和均一化。这里把裁剪出$\boldsymbol{M}=1$图像区域的操作定义为crop(.)，通过填充 0 值扩充该图像块区域并将其恢复为和$\boldsymbol{M}$相同尺寸大小的操作定义为 crop_reverse(.)。

综上所述，通过 BALA 产生对抗扰动块的详细内容如算法 1 中的伪代码所示。

算法1：BALA中生成对抗扰动块

输入：图片$\boldsymbol{x}$，模型F，掩膜$\boldsymbol{M}$，ε，概率阈值s，迭代次数上限C，变换集合T

输出：扰动块($\boldsymbol{patch}$)

counter $=0$, $l_{\rm org}=y(\mathrm{F},\boldsymbol{x})$, $\boldsymbol{\delta}=\boldsymbol{0}$, patch $=0.5*\text{crop}(\boldsymbol{M})$

$\boldsymbol{x}'=(\boldsymbol{1}-\boldsymbol{M})*T(\boldsymbol{x})+\boldsymbol{M}\cdot\boldsymbol{\delta}$, $l_{\rm pred}=y(\mathrm{F},\boldsymbol{x})$

while counter $<C$ do

  if $l_{\rm org}==l_{\rm pred}$ **then**

    $L=L_{s1}$

  else if $l_{\rm anh}==l_{\rm pred}$ **then**

    $L=L_{s2}$

  $\boldsymbol{\delta}=\varepsilon*\text{sign}(\partial L/\partial\boldsymbol{x}')$ $\boldsymbol{patch}'=\text{graying}(\boldsymbol{patch}+\text{crop}(\boldsymbol{M}\cdot\boldsymbol{\delta}))$

  patch $=$ averaging (patch $'$), $\boldsymbol{patch}'=\text{graying}(\boldsymbol{patch}+\text{crop}(\boldsymbol{M}\cdot\boldsymbol{\delta}))$

  $l_{\rm pred}=y(\mathrm{F},\boldsymbol{x}')$, $counter++$

  if $p_{y_{\rm pec}}(\mathrm{F},\boldsymbol{x}')>s\,\&\,l_{\rm org}\neq l_{\rm pred}$ **then**

    break

3 实验结果

在如图 6 所示，实验使用一块竖立的尺寸为 60 chin (1 chin=2.54 cm)，分辨率为 1080 p，流明为 320 cd/m^2，刷新频率为 50 Hz 的电子屏幕 (图 6(c)) 作为背景，并设定从照相机到背景屏幕的距离约为 1 m；使用的是分辨率为 3850 pixels×2650 pixels 的普通智能手机的相机，将其固定在三脚架上进行拍摄。三脚架的高度能够根据前景人像进行人为的调整。电子屏幕设置为全黑背景是为了尽可能减少屏幕背光的影响。在计算中使用 NVIDIA RTX2080 GPU、E5-2600 CPU 和 16 GB RAM 服务器生成对抗扰动块。

3.1 实验设置

识别器：BALA 采用常用的 CNN 人脸识别网络，VGG-FACE[30] 作为识别器。采用 LibFace-Detection[31] 作为人脸检测器，该检测器的输出是包含面部区域的矩形框。

数据集：实验中使用的人脸数据集 (集合 A) 来自于 5 位本研究团队人员 (集合 A1) 以及 Oxford VGG 人脸数据集[30] 中的 1000 位人物 (集合 A2)，总计 1005 名人物。实验中使用了预训练的人脸识别器网络 VGG Face[30]，选用集合 A 中的各人物照片 100 张，构成 100500 张训练图片对 VGG Face 预训练模型进行微调，使其具备识别集合 A 中人物的能力。同时在集合 A 中随机挑选 500 个人物，分别获取他们各 100 张人脸照片 (不同于集合 A) 作为测试集。最终，经过微调训练的 VGG-FACE 模型能够在训练集中达到 96.2% 的识别准确率，在测试集中达到 95.1% 的识别准确率。此外，相关人员的人脸照片使用已通过其本人的授权。

对比方式：将 BALA 和目前最具代表性的两种对抗攻击方法进行对比，分别是 LaVAN 和 Adv-patch。LaVAN[14] 已经表明，可以通过改变有限区域的像素值，成功误导数字域中成功率非常高的识别器。

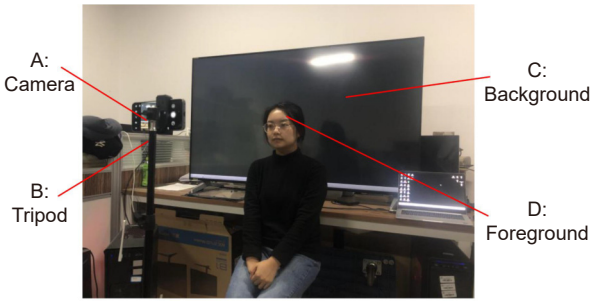


图 6 场景实拍实验设置。A 为相机，B 为三脚架，C 为电子背景屏幕，D 为前景人物

Fig. 6 The real-world experiment setup. A is a camera and B is the tripod; C is the electronic screen background and D is the foreground person

Adv-patch 可以通过在背景中放置对抗扰动块来误导识别器，但它更侧重于在静态场景中以物理域攻击的方式，误导物体分类器。与 Adv-patch 相比，BALA 能够在前景人物移动的场景中误导人脸识别器。

3.2 照片重拍和场景实拍实验

在本文中进行了重新拍摄照片以及模拟拍照过程的场景实拍实验。据了解，目前还没有针对电子屏幕重拍场景下的人脸对抗样本有效性研究。为了填补这一空白，本文在重新拍摄对抗样本后测试了其对抗有效性。图 7 的上半部分是照片重拍实验的方案，图 7 的下半部分是场景实拍实验的方案。这两个实验的步骤简要概述如下：

收集人脸数据：在照片重拍实验中，在集合 A2 中获取了每个人物 1 张，共计 1000 张不同的人脸照片。在场景实拍实验中，拍摄了集合 A1 中每一个人站在屏幕前的 6 张不同角度的正面照片，拍摄场景如图 6 所示。之后，利用人脸检测网络 LibFaceDetection 从拍摄的照片中截取了尺寸为 3×224×224 的 3 通道彩色人脸图片。

产生对抗扰动块：在给定 3×224×224 的面部图像后，利用 Adv-patch 和 LaVAN 分别生成一个 3×65×65 (约占 224×224 的 8.4%) 的彩色对抗扰动块；其中 $\varepsilon = 0.1$，迭代次数限制为 800 次[14]。通过 BALA 也产生一个 3×65×65 的灰色对抗扰动块，其中 $\varepsilon = 0.1$，$s = 0.9$，$C = 800$。具体来讲，BALA 在场景实拍实验中，使用 6 张照片来生成对抗扰动块，如图 7 绿色部分所示。此外，停止迭代的条件是对抗扰动块可以成功误导 VGG-Face 模型产生错误的分类结果或达到迭代上限。

显示及重拍：在照片重拍实验中，将由 Adv-patch、LaVAN 和 BALA 生成的对抗扰动块直接嵌入到原始数字图像中，产生对抗样本。三种方法产生的对抗样本都会显示在屏幕上并重新拍摄，以获得输入到 VGG-FACE 识别器的评估图像。在场景实拍实验中，Adv-patch、LaVAN 和 BALA 生成的对抗扰动块会直接显示在背景屏幕上，将带有对抗扰动块的背景与前景人物一起拍摄，以获得在场景实拍实验中输入到 VGG-FACE 识别器的评估图像。

3.3 实验结果

照片重拍：表 1 展示了 3 种方法产生的对抗样本经过重拍之后的攻击成功率 (ASR)。其中 ASR 定义为：在可被正确识别的原始人像图片集合中，添加对抗块后无法被正确识别的图片比例。BALA 在照片重拍实验中平均迭代次数为 483 次，平均每次迭代 0.3 s，

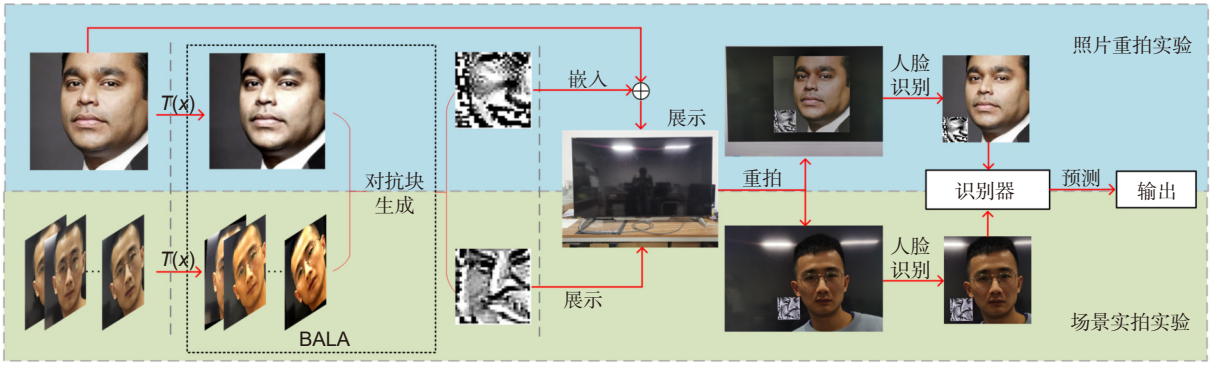


图 7 两种背景对抗攻击实验的流程图。蓝色部分是照片重拍实验，绿色部分是场景实拍实验

Fig. 7 The pipeline of two background adversarial attack experiments. The blue part is photo re-taken experiment and green part is the real-world experiment

表 1 在集合 A2 中照片重拍的平均 ASR(%)

Table 1 Photo re-taken experiment results over set-A1 in terms of average ASR(%)

	数字域攻击	照片重拍
LaVan[14]	96.7	58.4
Adv-patch[16]	**97.5**	65.1
BALA	94.6	**78.0**

实现了最高的 ASR(78.0%)，而在直接将扰动添加在数字照片上的数字域攻击中仅牺牲了 2.9% 的 ASR。Adv-patch 方法在数字域攻击中具有最佳性能(97.5%)。在图 8 中，展示了照片重拍的示例，包括原始人脸图像 (图 8(a))，通过 3 种方法重拍的带有对抗扰动块的照片 (图 8(b))，以及错误分类对应的图像 (图 8(c))。

在物理域的人脸照片重拍实验中，Adv-patch 的 ASR(65.1%) 要低于该原始研究中对物体分类的对抗攻击实验结果。在数字域的人脸图片攻击实验中，LaVAN 的 ASR(96.7%) 也要低于他们的原始研究 (LaVAN[14]) 中对物体分类的对抗攻击实验结果。这些性能的下降可能是由于物体和人脸识别任务之间的差异，以及 BALA 的实验场景更难攻击的设置所造成的。

场景实拍：本文为每个测试人员分别采集了 100 张带有轻微面部角度变化的图像。本实验在集合 A1 中的 5 位研究员上进行，对于同一人的全部 100 张照片都是使用相同的对抗扰动块。在实验中，将人脸到背景屏幕的距离分别设置为 10 cm、20 cm 和 50 cm。在 10 cm 的距离设置下，BALA 在物理域场景实拍实验中的 ASR(75.0%) 比 Adv-patch 方法高 13.8%。这种 BALA 优于 Adv-patch 的现象同样也可以在 20 cm 和 50 cm 的距离设置实验中观察到，如表 2 所示。带有对抗扰动块的图像比不带对抗扰动块的图像看起来更暗 (如图 9 所示)，是因为我们相机的焦点聚焦在背景和前景之间的位置，以便能清晰地拍摄背景对抗扰动块和前景人脸。图 9(b) 是将对抗扰动块显示在背景屏幕然后重新拍摄的对抗样本。图 9(c) 是检测的人脸图片，图 9(d) 是被误识别之后的类别图像。

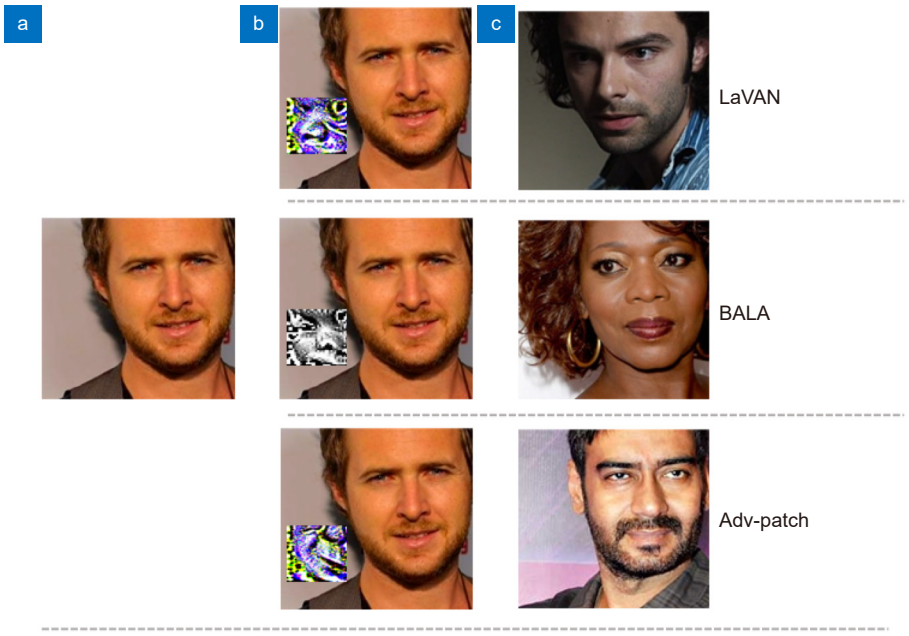


图 8　在照片重拍实验中，LaVAN，BALA 和 Adv-patch 三种方法生成的对抗样本。(a) 是原始人脸图片；(b) 表示将对抗扰动块显示到背景屏幕后重拍的照片；(c) 表示由 VGG-FACE 识别的错误类别对应的图像

Fig. 8　The adversarial examples generated by LaVAN, BALA, and Adv-patch in re-taken experiment. (a) is the original face image; (b) Present re-taking photos after displaying the adversarial patches on the background; (c) Present images of incorrect output classes from the VGG-FACE

表 2　在集合 A1 中，场景实拍实验中前景和背景屏幕之间不同距离的平均 ASR(%)

Table 2　Average ASR (%) of the different distance between foreground face and background screen in real-world experiments over set-A1

	10 cm	20 cm	50 cm
Adv-patch[16]	61.2	56.3	51.2
BALA	**75.0**	**69.2**	**62.4**

4 消融实验

4.1 BALA 中均一化的作用

本文分别使用无均值化、2×2 邻域平均和 4×4 邻域平均的方法生成了不同种类的对抗扰动块来验证 BALA 中均一化的作用。集合 A 中同一图像的对抗扰动块的示例如图 10 所示。相比于不做均值处理 (图 10(a)) 和做 4×4 邻域均值处理 (图 10(c)) 的对抗扰动块，采用 2×2 邻域均值处理 (图 10(b)) 方法产生的对抗扰动块具有明显的区别。

虽然没有做均值处理的对抗扰动块会拥有更多的扰动细节，但它的对抗效果会在从数字域到物理域转换的过程中损失巨大；与 2×2 邻域均值处理的方法相比，不做均值处理的对抗样本会在场景实拍实验中降低 6.8% 的 ASR，在照片重拍实验中降低 6.5% 的 ASR，如表 3 所示。与 2×2 邻域均值处理方法相比，4×4 的邻域均值处理后的对抗样本在两个实验中的 ASR 均要低 7% 以上，在数字域中要低约 20%，如表 3 所示。因此，考虑到上述结果，选择 2×2 邻域均值处理方法在 BALA 中生成对抗扰动块。

4.2 BALA 中灰度化的作用

本文分别研究了进行灰度化和不进行灰度化 (BALA-Color) 生成的对抗扰动块的效果。BALA-Color 在数字域攻击中实现了稍高的 ASR(增加 2.6%)，但在物理域照片重拍实验中的 ASR 会大幅下降 (下降 11.7%)，如表 3 所示。在场景实拍的实验中，

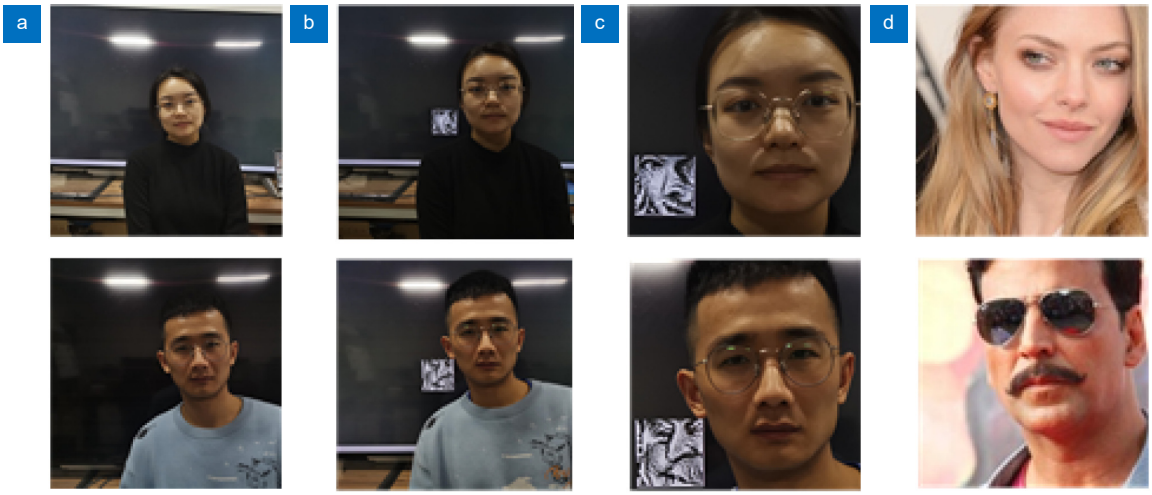


图 9 通过 BALA 在场景实拍实验中产生的对抗样本。(a) 原始照片；(b) 将扰动块添加到背景后重新拍摄的照片；(c) 剪切人脸之后的对抗样本；(d) VGG-FACE 网络输出的错误分类对应的人脸

Fig. 9 The adversarial examples generated by BALA in the real-world experiment. (a) The original photos; (b) Present re-taking photos after displaying the adversarial patches on the background; (c) Present adversarial examples from the cropped faces; (d) Present face images of incorrect output classes of the VGG-FACE network

图 10 采用灰度化 BALA 生成不同的均值对抗扰动块。(a) 无均值法生成的扰动块；(b) 采用 2×2 邻域均值化生成的扰动块；(c) 采用 4×4 邻域均值化生成的扰动块

Fig. 10 The diverse averaging images of BALA with graying. (a) Generated patch using no averaging approach; (b) Generated patch by averaging pixels in 2 × 2 region; (c) Generated patch by averaging pixels in 4 × 4 region

表 3 集合 A 中 BALA 的均一化作用下的平均 ASR(%)

Table 3 The results of averaging effect of BALA over set-A in terms of average ASR(%)

	BALA			BALA-Color
	无均值	2×2	4×4	2×2
数字攻击	**98.7**	94.6	75.4	97.2
照片重拍	71.5	**78.0**	70.2	66.3
场景实拍攻击	62.4	**69.2**	60.3	55.4

BALA-Color 的ASR(55.4%) 比BALA(69.2%) 低13.8%。这些结果表明了灰度化在物理域攻击的优势。

4.3 讨 论

上述实验表明，在场景实拍的黑色背景区域上放置有限大小的对抗扰动块来攻击 VGG-FACE 神经网络模型是可行的。如果进一步使用对抗扰动块的模糊化处理[32] 策略，BALA 还可以产生更具隐藏性的对抗扰动块，这对各种自然背景来说就显得不那么明显。通过以上实验可以看出，生成的对抗扰动块实际上就是一些面部特征 (图 8 和图 10)，例如鼻子、眼睛和额头等。Mikhail 等人也已经发现了相似的结论[19]。其次，通过实验发现，如果在人脸检测过程中，对抗扰动块受到一定程度的切割，当切割面积小于总面积的 5%，不会影响对抗扰动块攻击 VGG-FACE 模型的有效性。

对抗扰动块生成的过程中，在数字域中使用无损压缩格式保存的对抗扰动块 (例如 PNG 格式) 比其他压缩图像格式 (例如 JPG 等)，在物理域重新拍摄对抗样本的过程中会表现得更加鲁棒。

本文在实验设置中考虑了实际应用场景问题，场景实拍攻击实验中参考了在线会议或直播的场景设置，例如固定的背景板和拍摄相机、适当的前背景距离、跟随人脸产生的对抗扰动块、防剪裁的设定等。随着在线会议、直播的兴起，人脸数据隐私保护愈发成为必须要考虑的安全因素；在未来的线上会议、直播中，使用本文的方法，只需要在后面放一块电子屏幕即可避免自己的人脸数据被恶意用做人脸识别器的训练图片。

5 结 论

本文提出了一种生成对抗扰动块的方法，通过将其显示在背景设备上可以在物理域中攻击人脸识别系统。根据实验结论，BALA 在 ASR 方面优于其他最先进的背景攻击方法。具体来说，在照片重拍实验中，与 Adv-patch 和 LaVAN 相比，BALA 实现了超过其 12% 的 ASR 性能。在场景实拍实验中，与 Adv-patch 方法相比，BALA 将 ASR 提高了 3.8%。本文提出的 BALA 方法在物理域屏幕中添加背景扰动块不仅可以对抗非法人脸识别器，而且还能保证清晰的人脸面部观察。在保护面部数据不被泄露的过程中，前景人物甚至可以不知道扰动块的存在。

参考文献

[1] Lee S, Woo T, Lee S H. SBNet: segmentation-based network for natural language-based vehicle search[C]//*2021IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021: 4049−4055. https://doi.org/10.1109/CVPRW53098.2021.00457.

[2] Sun R, Shan X Q, Sun Q J, et al. NIR-VIS face image translation method with dual contrastive learning framework[J]. *Opto-Electron Eng*, 2022, **49**(4): 210317.
孙锐, 单晓全, 孙琦景, 等. 双重对比学习框架下近红外[J]. 光电工程, 2022, **49**(4): 210317.

[3] Meng Q E, Shin'ichi S. ADINet: attribute driven incremental network for retinal image classification[C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 4032–4041. https://doi.org/10.1109/CVPR42600.2020.00409.

[4] Singh V, Hari S K S, Tsai T, et al. Simulation driven design and test for safety of AI based autonomous vehicles[C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021: 122−128. https://doi.org/10.1109/CVPRW53098.2021.00022.

[5] Liao M H, Zheng S S, Pan S X, et al. Deep-learning-based ciphertext-only attack on optical double random phase encryption[J]. *Opto-Electron Adv*, 2021, **4**(5): 200016.

[6] Ma T G, Tobah M, Wang H Z, et al. Benchmarking deep learning-based models on nanophotonic inverse design problems[J]. *Opto-Electron Sci*, 2022, **1**(1): 210012.

[7] Raji I D, Fried G. About face: a survey of facial recognition evaluation[Z]. arXiv: 2102.00813, 2021. https://arxiv.org/abs/2102.00813.

[8] Pesent J. An update on our use of face recognition[EB/OL]. (2021-11-02). https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/.

[9] Sun Q R, Ma L Q, Oh S J, et al. Natural and Effective Obfuscation by Head Inpainting[C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 5050–5059. https://doi.org/10.1109/CVPR.2018.00530.

[10] Wright E. The future of facial recognition is not fully known: developing privacy and security regulatory mechanisms for

facial recognition in the retail sector[J]. *Fordham Intell Prop Media Ent L J*, 2019, **29**: 611.

[11] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]//*2nd International Conference on Learning Representations*, 2014.

[12] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]//*6th International Conference on Learning Representations*, 2018.

[13] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//*3rd International Conference on Learning Representations*, 2015.

[14] Karmon D, Zoran D, Goldberg Y. LaVAN: localized and visible adversarial noise[C]//*Proceedings of the 35th International Conference on Machine Learning*, 2018: 2512–2520.

[15] Wu D X, Wang Y S, Xia S T, et al. Skip connections matter: on the transferability of adversarial examples generated with ResNets[C]//*8th International Conference on Learning Representations*, 2020.

[16] Brown T B, Mané D, Roy A, et al. Adversarial patch[Z]. arXiv: 1712.09665, 2017. https://arxiv.org/abs/1712.09665.

[17] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[C]//*5th International Conference on Learning Representations*, 2017.

[18] Athalye A, Engstrom L, Ilyas A. Synthesizing robust adversarial examples[C]//*Proceedings of the 35th International Conference on Machine Learning*, 2018: 284–293.

[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*, 2019: 391–396.

[20] 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. https://doi.org/10.1109/ICPR48806.2021.9412236.

[21] 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. https://doi.org/10.1109/CVPRW50498.2020.00415.

[22] Jan S T K, Messou J, Lin Y C, et al. Connecting the digital and physical world: improving the robustness of adversarial attacks[C]//*Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2019: 119. https://doi.org/10.1609/aaai.v33i01.3301962.

[23] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574−2582. https://doi.org/10.1109/CVPR.2016.282.

[24] Su J W, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. *IEEE Trans Evol Comput*, 2019, **23**(5): 828−841.

[25] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//*Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 1528–1540. https://doi.org/10.1145/2976749.2978392.

[26] Xu K D, Zhang G Y, Liu S J, et al. Adversarial T-shirt! Evading person detectors in a physical world[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 665–681. https://doi.org/10.1007/978-3-030-58558-7_39.

[27] Rahmati A, Moosavi-Dezfooli S M, Frossard P, et al. GeoDA: a geometric framework for black-box adversarial attacks[C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8443–8452. https://doi.org/10.1109/CVPR42600.2020.00847.

[28] Sun Y, Wang X G, Tang X O. Deep convolutional network cascade for facial point detection[C]//*2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3476–3483. https://doi.org/10.1109/CVPR.2013.446.

[29] Wang J F, Yuan Y, Yu G. Face attention network: an effective face detector for the occluded faces[Z]. arXiv: 1711.07246, 2017. https://arxiv.org/abs/1711.07246.

[30] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[C]//*Proceedings of the British Machine Vision Conference 2015*, 2015: 41.1–41.12.

[31] Peng H Y, Yu S Q. A systematic IoU-related method: beyond simplified regression for better localization[J]. *IEEE Trans Image Process*, 2021, **30**: 5032−5044.

[32] Duan R J, Ma X J, Wang Y S, et al. Adversarial camouflage: hiding physical-world attacks with natural styles[C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 997−1005. https://doi.org/10.1109/CVPR42600.2020.00108.

作者简介

张晨晨 (1995-)，男，硕士研究生，主要研究方向为多传感器融合、卷积神经网络的对抗攻击和机器学习。

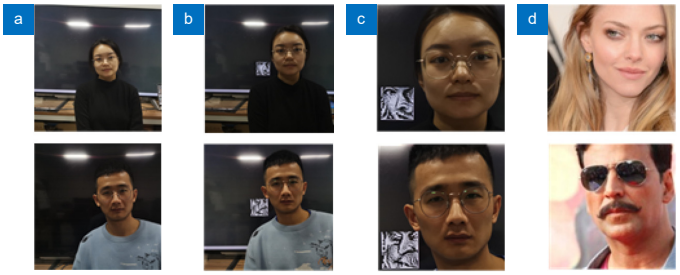
E-mail：202022012134@std.uestc.edu.cn

【通信作者】王帅 (1980-)，男，博士，高级工程师，主要从事图像处理、模式识别、机器学习、多源信息融合研究。

E-mail: wangshuai0601@uestc.edu.cn

Adversarial background attacks in a limited area for CNN based face recognition

Zhang Chenchen[1,2], Wang Shuai[1,2*], Wang Wenyi[2], Li Diran[2], Li Nan[1,2], Bao Hua[3,4], Li Shuqi[3,4], Gao Guoqing[3,4]



The adversarial examples generated by BALA in the real-world experiment. Column (a) are the original photos. Column (b) present re-taking photos after displaying the adversarial patches on background. Column (c) present adversarial examples from the cropped faces. Column (d) present face images of incorrect output classes of the VGG-FACE network

Overview: At present, face recognition has been integrated into every aspect of everyone's life, which makes face privacy protection an important topic. Face image recognizers based on convolutional neural networks (CNN) have been widely used, but CNN-based image classifiers are easily misled by adversarial examples with special human perturbations, resulting in the false label. However, many physical facial adversarial generation methods used obvious perturbation patterns in the foreground leading to hampering a clear face observation. Therefore, using physical background perturbation patches may be a suitable way to not only against illegal face recognizers but also keep clear face observation.

Taking advantage of the fact that adversarial examples can interfere with CNN face recognizers, this paper proposes a privacy protection scheme for intelligent face recognizers. In order to overcome the loss of original facial features caused by the addition of significant perturbation patches in the foreground face area by existing adversarial example generation methods, this paper adds background adversarial perturbation blocks in the physical domain, so as to achieve anti-unauthorized face recognition while maintaining all original facial features. Specifically, this paper proposes a novel adversarial perturbation patches generation and addition method, called Facial Adversarial Background Attack in a Limited Area (BALA). To the best of our knowledge, BALA is the first method to mislead face recognizers by modifying real local background regions.

The innovation of this paper is that (1) BALA optimizes the gradient back-propagation efficiency by using different loss functions in two iterative stages, so as to better generate perturbation patches; (2) BALA uses adversarial region grayscale and averaging to strengthen adversarial effects after digital to physical domain conversion. In experiments, the adversarial patches displayed on the background screen can mislead the VGG-FACE face recognizer without covering any face area, and the adversarial patch area is only 8.4% of the face detection area. In the photo retake experiment, BALA improves the attack success rate (ASR) by 12% compared with Adv-patch and LaVAN methods, and in the live shot experiment, BALA's ASR is 3.8% higher than Adv-patch. These results demonstrate that our proposed BALA has leading performance in adversarial attacks against background faces in the physical domain.

Zhang C C, Wang S, Wang W Y, et al. Adversarial background attacks in a limited area for CNN based face recognition[J]. *Opto-Electron Eng*, 2023, **50**(1): 220266; DOI: 10.12086/oee.2023.220266

Foundation item: Municipal Government of Quzhou (2022D025)

[1]Yangtze Delta Region Institute of University of Electronic Science and Technology of China, Quzhou, Zhejiang 324000, China; [2]University of Electronic Science and Technology of China, Chengdu, Sichuan 610000, China; [3]Key Laboratory on Adaptive Optics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China; [4]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China

* E-mail: wangshuai0601@uestc.edu.cn