计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300305

针对人脸检测对抗攻击风险的安全测评方法

景慧昀[1]　周　川[2,3]　贺　欣[4]
1 中国信息通信研究院　北京 100083
2 中国科学院信息工程研究所　北京 100097
3 中国科学院大学网络空间安全学院　北京 100049
4 国家计算机网络应急技术处理协调中心　北京 102209
(jinghuiyun@caict.ac.cn)

摘　要　人脸检测是计算机视觉领域的一个经典问题,其在人工智能大数据驱动的赋能下焕发出崭新生机,在刷脸支付、身份认证、摄像美颜、智能安防等领域均体现出重要的应用价值与广阔的应用前景。然而,随着人脸检测部署应用进程的全面加速,其安全风险与隐患也日益凸显。因此,文中分析总结了现行人脸检测模型在全生命周期的各阶段所面临的安全风险,其中对抗攻击因对人脸检测的可用性和可靠性构成严重威胁,并可能使人脸检测模块丧失基本功能性而受到了广泛关注。目前,面向人脸检测的对抗攻击算法主要集中于白盒攻击。但是,由于白盒对抗攻击需要充分理解特定人脸检测模型的内部结构和全部参数,而出于对保护商业机密和企业利益的考虑,现实物理世界中商业部署的人脸检测模型的结构与参数通常是不可访问的,这使得使用白盒攻击方法在现实世界中攻破商业人脸检测模型几乎不可能。针对上述问题,提出了一种面向人脸检测的黑盒物理域对抗攻击方法。通过集成学习的思想,提取众多人脸检测模型的公共注意力热力图,并针对获取到的公共注意力热力图发起攻击。实验结果表明,该方法能够成功逃逸部署于移动终端的黑盒人脸检测模型,包括移动终端自带相机软件、刷脸支付软件、美颜相机软件的人脸检测模块。这说明所提出的方法对评测人脸检测模型的安全性能够提供有益帮助。

关键词:人工智能安全;对抗攻击;人脸检测
中图法分类号　TP183

Security Evaluation Method for Risk of Adversarial Attack on Face Detection

JING Hui-yun[1],ZHOU Chuan[2,3] and HE Xin[4]
1 China Academy of Information and Communications Technology,Beijing 100083,China
2 Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100097,China
3 School of Cyber Security,University of Chinese Academy of Sciences,Beijing 100049,China
4 National Computer Network Emergency Response Technical Team/Coordination Center of China,Beijing 102209,China

Abstract　Face detection is a classic problem in the field of computer vision. With the power-driven by artificial intelligence and big data,it has displayed a new vitality. Face detection shows its important application value and great application prospect in the fields of face payment,identity authentication,beauty camera,intelligent security,and so on. However,with the overall acceleration of face detection deployment and application process,its security risks and hidden dangers have become increasingly prominent. Therefore,this paper analyzes and summarizes the security risks which the current face detection models face in each stage of their life cycle. Among them,adversarial attack has received extensive attention because it poses a serious threat to the availability and reliability of face detection,and may cause the dysfunction of the face detection module. The current adversarial attacks on face detection mainly focus on white-box adversarial attacks. However,because white-box adversarial attacks require a full understanding of the internal structure and all parameters of a specific face detection model,and for the protection of business secrets and corporate interests,the structure and parameters of a commercially deployed face detection model in the real physical world are usually inaccessible. This makes it almost impossible to use white-box adversarial methods to attack commercial face detection models in the real world. To solve the above problems,this paper proposes a black-box physical adversarial attack method for face detection. Through the idea of ensemble learning,the public attention heat map of many face detection models is extracted,then the obtained public attention heat map is attacked. Experiments show that our method realizes the successful escape of the black-box face detection model deployed on mobile terminals,including the face detection module of mobile terminal's

到稿日期:2021-03-21　返修日期:2021-04-23
基金项目:国家 242 信息安全计划(2018Q39)
This work was supported by the National 242 Information Security Program(2018Q39).
通信作者:周川(zhouchuan1@iie.ac.cn)

built-in camera software, face payment software, and beauty camera software. This demonstrates that our method will be helpful to evaluate the security of face detection models in the real world.

Keywords Artificial intelligence security, Adversarial attack, Face detection

1 引言

人脸检测是计算机视觉领域中的一个经典问题。随着人工智能时代的来临，人脸检测在人工智能大数据驱动的赋能下焕发出崭新生机，在刷脸支付、身份认证、摄像美颜、智能安防等领域均体现出重要的应用价值与广阔的应用前景。

然而，随着全球人脸检测和人脸识别的规模化建设与部署进程的全面加速，例如机场火车站等交通枢纽全面部署监控摄像头、“天网系统”依托于人脸检测和人脸识别实现对犯罪嫌疑人的千里追逃，人脸检测模型存在的安全风险与隐患也日益凸显，不容忽视。

为系统性分析人脸检测可能存在的全部安全风险与隐患，我们参照国际标准化组织(ISO)对于人工智能系统全生命周期的划分，将人脸检测全生命周期粗略的划分为设计研发、部署、运行监控以及废弃 4 个阶段。

在设计研发阶段，人脸检测的安全风险主要表现为机器学习开源框架平台等人工智能基础设施不完善，存在安全漏洞等问题。在部署阶段，人脸检测的安全风险主要表现为部署模型的软硬件环境不可信，人脸检测模型可能遭受非授权访问和非授权使用[1-2]。在运行监控阶段，恶意攻击者可能对人脸检测模型发起对抗样本[3]、算法后门[4]、模型窃取[5]、模型反馈误导、数据逆向还原[6]、成员推理[7]、属性推断、代码漏洞利用等安全攻击。在废弃阶段，可能存在人脸检测模型销毁不彻底，模型或人脸图像数据集等遭到恶意泄露的风险隐患[8]。

在上述人脸检测面临的安全风险中，我们将重点关注人脸检测面临的对抗攻击安全风险，这是由于快速、准确地检测到人脸是人脸检测的核心功能性需求，恶意攻击者针对人脸检测模型发起的对抗样本攻击[9-11]可能会使人脸检测模块丧失基本功能性，直接导致基于人脸检测的相关应用意外崩溃。例如，测试者可以通过在面部或头饰上张贴对抗样本块实现对人脸检测的成功逃逸[11]，这意味着人脸检测模型将无法以矩形框框选的形式检测到人脸。

由于对抗样本攻击对人脸检测应用造成了巨大的安全威胁，因此针对人脸检测对抗样本攻击的研究迫在眉睫。通过面向人脸检测的对抗样本攻击方法能够有效评测现行人脸检测模型抵御对抗攻击的能力，测评现行人脸检测模型的安全性与鲁棒性。然而，目前面向人脸检测的对抗攻击方法相对较少。

Bose 等[9]通过求解约束优化问题生成对抗样本，从而使相应的人脸检测器无法检测到人脸。作为一种白盒攻击方法，其依赖于不受限制地访问 Faster R-CNN[12]人脸检测模型的全部结构与参数，这在实际应用场景中是不现实的。Zhou 等[10]使用附着在帽饰上的红外线 LED 灯来构造物理域对抗样本，并成功绕过人脸检测系统。然而，由于不同测试者的人脸特征点位置存在显著差异，因此攻击效果与测试者高度相关，更换测试者可能导致逃逸率的急剧下降[10]。Kaziа-khmedov 等[11]通过将对抗样本块直接张贴到人脸上实现了一种简单可复现的攻击 MTCNN[13]人脸检测模型的方法。

上述方法都需要直接访问人脸检测模型的结构和参数，因此无法实现面向人脸检测的黑盒对抗攻击。它们对现实世界中商业部署的人脸检测模型构成的威胁较小，因为在现实世界中，商业人脸检测模型的结构和参数无法访问。

为解决这些方法存在的上述问题，我们面向人脸检测提出了一种黑盒物理域对抗攻击方法。我们注意到注意力热力图 ahm(attention heat map)[14-15]可以反映关注于同一任务的多个未知 DNN 模型的共性弱点[14]。受此启发，我们将现有的多个开源人脸检测模型的注意力热力图平均求和，计算出公共注意力热力图 pahm(public attention heat map)，并用以定量反映人脸检测模型的共性弱点。该方法通过攻击公共注意力热力图所反映的人脸检测模型的共性弱点来生成对抗样本，克服了上述白盒攻击方法需要获取被攻击模型详细内部信息的严格限制，具有针对黑盒商业人脸检测模型的良好攻击效果。通过将打印好的对抗样本块直接张贴到人脸面颊区域，实现了面向人脸检测的黑盒物理域对抗攻击。

本文的具体贡献如下：

(1)参照国际标准化组织(ISO)对人工智能系统全生命周期的划分，从人脸检测全阶段出发，系统性分析总结了现行人脸检测模型面临的安全挑战，为提升人脸检测模型的安全性提供了有益参考。

(2)针对人脸检测模块在运行监控阶段面临的安全风险，本文设计并首次实现了面向人脸检测的黑盒物理域对抗攻击。该方法实现了针对三星、小米等手机终端的刷脸解锁模块和内置拍照软件的成功逃逸。同时，本文实现了针对商用软件的成功逃逸，例如支付宝的刷脸支付模块、B612 自拍美颜软件。

2 相关工作

2.1 人脸检测各阶段的安全风险

2.1.1 设计研发阶段的安全风险

设计研发阶段负责创建可部署人脸检测系统，主要包括确定设计方法、定义模型框架、编写软件代码和风险管理等过程。这个阶段的安全风险主要表现为机器学习开源框架平台等人工智能基础设施不完善。

2020 年 9 月，安全厂商 360 公开披露谷歌开源框架平台 TensorFlow 存在 24 个安全漏洞。机器学习开源框架平台的不完善和脆弱性将直接影响到基于其设计的人脸检测模型的安全性。

2.1.2 部署阶段的安全风险

部署阶段主要负责在目标环境中安装和配置人脸检测模型。该阶段的安全风险主要表现为部署模型的软硬件环境不可信，人脸检测模型可能遭受非授权访问和非授权使用[1-2]，这可能会直接导致模型窃取[5]、模型泄露以及随之而来的针对泄露人脸检测模型的白盒对抗样本攻击[9-11]。

2.1.3 运行监控阶段的安全风险

运行监控阶段主要包括针对人脸检测模型的运行监控、维护升级等过程。该阶段的安全风险主要表现为恶意攻击者对人脸检测模型发起的对抗样本[3]、算法后门[4]、模型窃取[5]、模型反馈误导、数据逆向还原[6]、成员推理[7]、属性推断、代码漏洞利用等安全攻击。

Fredrikson 等提出的模型逆向攻击(model inversion attack)[16]可以利用黑盒模型输出中的置信度等信息将训练集中的人脸恢复出来。

2.1.4 废弃阶段的安全风险

人脸检测模型的废弃阶段主要包括销毁功能性指标落后或者存在重大安全漏洞的人脸检测模型，包括人脸图像数据集、人脸检测模型以及整体部署环境。这个阶段的安全风险主要表现为销毁不彻底、模型泄露或人脸图像数据集等隐私信息[8]。

2.2 现有人脸检测对抗攻击算法

根据能否将数字图像直接馈入机器学习算法，可以将面向人脸检测的攻击方法大致分为数字域对抗攻击方法和物理域对抗攻击方法。

2.2.1 面向人脸检测的数字域对抗攻击

数字域对抗攻击[14]是基于一个隐含假设，即攻击者可以直接向机器学习算法馈入数字图像形式的对抗样本。在攻击中，对抗扰动被直接添加到原始的数字图像中，增添了对抗扰动的数字图像，即对抗样本，随后被直接馈入分类模型[17]。

Bose 等[9]通过训练一个生成器实现了针对 Faster RCNN[12]人脸检测模型的数字域白盒攻击，其训练过程是一个类似于 C&W 攻击[18]的约束优化问题。作为一种典型的白盒攻击方法，其攻击成功率较高，但生成器的训练效率较低。同时，由于需要完全获知模型的结构信息和全部参数，致使其攻击能力十分有限，缺乏黑盒攻击能力。

2.2.2 面向人脸检测的物理域对抗攻击

物理域对抗攻击关注于在真实物体上部署对抗样本。例如，使用激光打印机将对抗样本块打印至 A4 纸上，随后将对抗样本块附着到物理世界中的真实目标上，例如现实中的人脸，以愚弄图像分类器，而非直接向机器学习算法馈入数字图像。

Kaziakhmedov 等[11]首先提出了面向人脸检测的物理域对抗攻击，该攻击能够成功逃逸摄像头对人脸的检测。该方法通过反向传播损失函数的梯度值来优化调整对抗样本块。作为一种基于梯度的对抗算法，其核心优势在于攻击速度快、效率高，但其白盒特性使其对真实场景下的商业人脸检测模型威胁较小，且攻击成功率也较低。

Zhou 等[10]提出了一种基于红外干扰的人脸检测逃逸方法。他们通过使用肉眼不可见的特殊红外光线干扰面部特征点区域，从而可以绕过或误导人脸识别系统。该方法具有较强的隐蔽性，但逃逸成功率相对较低，红外光源配置等步骤也比较复杂。另外，由于不同个体面部特征点存在巨大差异，当换用其他测试者进行逃逸实验时，该方法的逃逸能力可能会完全丧失。

3 面向人脸检测的黑盒物理域对抗攻击方法

本文重点关注人脸检测在运行监控阶段所面临的对抗样本攻击安全风险。我们设计了黑盒物理域对抗攻击方法用于评测现行人脸检测模型的安全性与鲁棒性。

虽然关注于同一任务的深度神经网络可能有着完全不同的模型结构和参数，但它们可能拥有共同的语义特征[14]，如相似的注意力热力图激活区域。这些相似的激活区域充分反映了模型的公共关注区域。面向人脸检测的黑盒物理域对抗攻击方法正是受到了上述思想的启发。

因此，本文通过平均求和的方式集成了多个人脸检测模型的注意力热图，突出了各模型在公共关注区域的公共注意力，融合了不同模型在特定人脸区域的特定注意力，得到公共注意力热力图。研究人员发现，通过攻击一个白盒深度神经网络的注意力热力图，能够使该神经网络注意力失去其注意力关注焦点，从而无法进行判断[14]。显然，当我们攻击公共注意力热力图时，由于其突出了各模型的公共关注区域并融合了不同模型的特定关注区域，因此我们可以更好地使人脸检测模型失去其关注焦点，从而无法检测到人脸。

由图 1 可见，4 种不同的人脸检测模型分别关注于人脸的不同区域，在将它们的注意力热力图平均相加后，我们获得了公共注意力热力图，该图关注于图像中的真实人脸区域。图 2 给出了在测试者面颊上张贴由本文方法生成的对抗样本块后的注意力热力图。各注意力热力图和计算得到的公共注意力热力图都充分表明，在张贴对抗样本块后，人脸检测模型不再能聚焦于图像中人脸的关键区域。

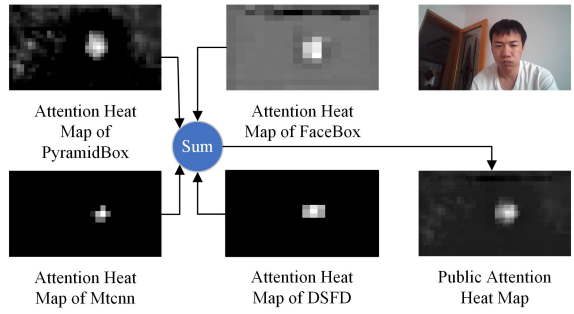

图 1 原始人脸的公共注意力热力图

Fig. 1 Public attention heat map of original face

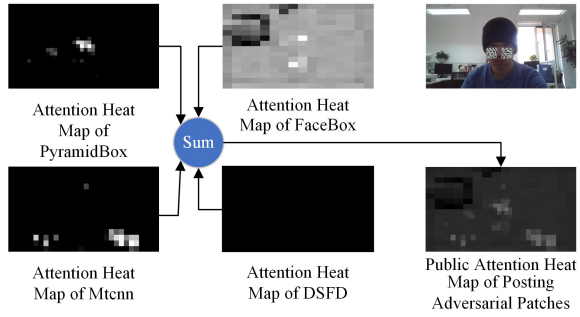

图 2 在测试者面颊张贴对抗样本块后的公共注意力热力图

Fig. 2 Public attention heat map of posting adversarial patches on tester's cheeks

本文所提出的方法的流程图如图 3 所示。该方法主要包括 3 个部分：准备工作、计算多个人脸检测模型的公共注意力热力图以及利用获得的公共注意力热力图更新对抗样本块。在准备工作部分，我们对输入图像进行初始配置，将初始化的

对抗样本块通过仿射变换应用到配置好的输入图像上，为首次训练做准备，并选择不同结构的人脸检测模型。在获取公共注意力热力图部分，将准备工作部分处理完成的图像馈入多个预先选择好的人脸检测模型中，计算出各模型的注意力热力图。然后，对各模型的注意力热力图进行平均求和以得到公共注意力热力图。在更新对抗样本块部分配置损失函数，通过 MI-FGSM[11] 更新对抗样本块，并将更新后的对抗样本块再次通过仿射变换应用到准备工作部分配置好的输入图像上。至此，已经完成了对对抗样本块的初次训练。之后，我们将迭代执行后两个部分，直至达到相应的迭代训练次数。

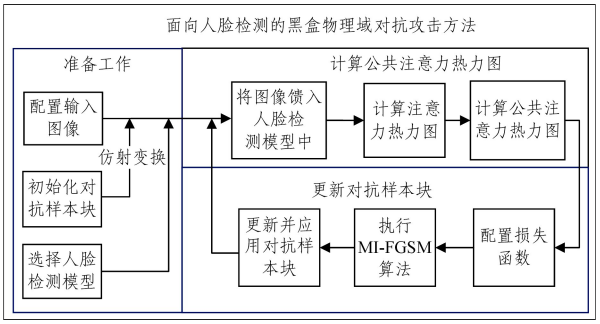

图 3 所提出方法流程图

Fig. 3 Pipeline of our method

3.1 准备工作

(1)配置输入图像并仿射变换初始化的对抗样本块

使用激光打印机打印出黑白棋盘格并张贴至测试者的面颊上。为了增强攻击鲁棒性，我们使用相机在不同光照、距离、角度等因素下多次采集人脸图像。然后将收集到的图像的黑白棋盘格边缘进行人为标定。最后，我们初始化对抗样本块，并将初始化的对抗样本块仿射变换到识别标记边缘而得到的投影区域上。

(2)选择人脸检测模型

由于官方人脸检测模型和非官方人脸检测模型的结构与参数存在一定差异，我们专门收集了几种具有不同结构的非官方人脸检测模型用于训练对抗样本块，以便使用相应的官方人脸检测模型进行黑盒逃逸测试。所选模型为 Mtcnn[13]1)，PyramidBox[19]2)，FaceBox[20]3) 和 DSFD[21]4)。通过上述人脸检测模型，我们得到了用于判断图像某一特定区域是否包含人脸的概率矩阵。利用该概率矩阵可以计算出相应模型的注意力热力图。

3.2 获取公共注意力热力图

公共注意力热力图的计算包括 3 个部分。首先，将所有完成仿射变换的输入图像馈入 4 个预先选定的人脸检测模型中。然后，对于每个人脸检测模型，将某一区域判定为包含人脸的概率矩阵中的最大值 $y_{\max}$ 输入 Grad-CAM 方法[22] 中并获得相应的注意力热力图 ahm，其公式如下：

$$\alpha_k=\frac{1}{Z}\sum_i\sum_j\frac{\partial y_{\max}}{\partial A_{i,j}^k} \tag{1}$$

$$ahm=\mathrm{ReLu}(\sum_k\alpha_k A^k) \tag{2}$$

其中，$A_{i,j}^k$ 是第 k 个通道的特征图中 i,j 位置处的像素值，Z 是第 k 个通道的特征图的像素总数。我们首先计算概率矩阵中的最大值 $y_{\max}$ 关于像素值 $A_{i,j}^k$ 的偏导数。通过全局平均池化，我们得到了第 k 个通道特征图的权重 α_k。通过对各通道的特征图进行加权求和并通过 $ReLU$ 函数激活，最终获得了其注意力热力图 ahm。最后，将 4 个选定人脸检测模型生成的 4 个注意力热力图 ahm 平均，我们得到一幅训练图像的公共注意力热力图 $pahm$。

$$pahm=\frac{1}{4}\times(ahm_{\mathrm{mtcnn}}+ahm_{\mathrm{pyramidbox}}+ahm_{\mathrm{yoloface}}+ahm_{\mathrm{dsfd}}) \tag{3}$$

3.3 更新对抗样本块

更新对抗样本块共包括 3 个部分。首先，我们配置损失函数来表示优化目标；然后，利用 MI-FGSM 算法[23] 更新对抗样本块；最后，将更新的对抗样本块仿射变换至准备工作部分配置好的输入图像上。

(1)配置损失函数

公共注意力热力图 $pahm$ 本质上是一种特征，可以通过该热力图衡量一幅图像上的不同区域对多种人脸检测模型共同关注的人脸检测任务的不同贡献程度。对于这一特征，我们选择同样能反映出对人脸检测任务贡献程度的 L_{clf} 损失函数。将所选 4 种模型的 L_{clf} 损失相加，可以得到损失函数 $L_{clf_{\mathrm{total}}}$。

$$L_{clf}=\sum_{i=1}^{N}\sum_m\max(p_m-0.5,0)^2 \tag{4}$$

$$L_{clf_{total}}=L_{clf_{\mathrm{mtcnn}}}+L_{clf_{\mathrm{pyramidbox}}}+L_{clf_{\mathrm{yoloface}}}+L_{clf_{\mathrm{dsfd}}} \tag{5}$$

其中，p_m 表示人脸图像的区域 m 被相应的人脸检测模型判断为包含待检测人脸的置信概率。N 为参与训练对抗样本块的输入图像数量。由于公共注意力热力图 $pahm$ 和 $L_{clf_{\mathrm{total}}}$ 在现实物理意义层面具有高度一致性，我们有理由假设，通过优化 $L_{clf_{\mathrm{total}}}$ 损失函数，公共注意力热力图 $pahm$ 也将朝着模型不再能关注于图像中人脸关键区域的方向改变。图 1 和图 2 充分证明了我们的假设是合理的。同时，我们利用训练过程中人脸区域中激活区域的变化来进一步证明上述假设成立。图 4 可视化地展示了随着训练的不断进行，激活区域的面积(像素数)在波动过程中迅速减少，最后所有的像素均无法超过设定的阈值，而真实人脸区域的激活区域面积稳定于 0，这意味着此时人脸检测模型不再能关注于图像的人脸关键区域。

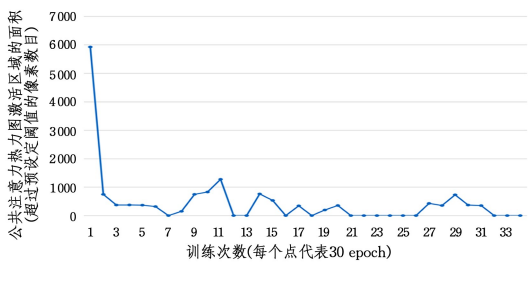

图 4 公共注意力热力图的激活区域面积在波动中迅速下降

Fig. 4 Public attention heat map's activation area rapidly decreases in fluctuation

1) http://github.com/edosedgar/mtcnnattack/tree/master/mtcnn

2) http://github.com/EricZgw/PyramidBox

3) http://github.com/610265158/faceboxes-tensorflow/tree/tf1

4) http://github.com/610265158/DSFD-tensorflow

接着，我们人为定义了全变差损失 L_{TV}，使对抗样本块的优化过程更偏向于产生没有尖锐色彩突变和噪声的视觉感观较好的图像模式。我们通过位置 i,j 处的像素值计算 L_{TV}：

$$L_{TV}=\sqrt{(p_{i,j}-p_{i+1,j})^2+(p_{i,j}-p_{i,j+1})^2} \tag{6}$$

通过对两种损失函数加权求和，得到总损失函数 L_{total}：

$$L_{\text{total}}=L_{clf_{\text{total}}}+\alpha L_{TV} \tag{7}$$

2）更新并应用对抗样本块

在计算出总损失 L_{total} 后，我们将 L_{total} 作为优化目标馈到 MI-FGSM 算法中，并使用算法的输出更新对抗样本块。最后，将更新后的对抗样本块通过仿射变换应用到准备工作部分配置好的图像上。后续，我们可以利用获得的图像再次训练并更新对抗样本块，直至达到迭代训练的预设置次数。

4 人脸检测逃逸实验和结果分析

根据测试内容的不同，我们将实验分为两部分：手机终端刷脸解锁模块逃逸实验和手机终端内嵌人脸检测功能相关软件的逃逸实验。本文所有实验均设置比较以下 3 种情况：测试者张贴对抗样本块、张贴随机生成的扰动块（服从 0～255 均匀分布）和不进行张贴。

4.1 手机终端刷脸解锁模块逃逸实验

由于部署在手机终端上的人脸检测模型的结构与参数信息不可访问，这部分的逃逸实验在本质上属于黑盒攻击。我们收集了主流的商业手机品牌，最终选择了 iPhone 11、三星 Galaxy S10 5G、小米 Redmi K20 Pro，它们都支持通过人脸识别功能解锁手机。为控制变量，实验使用相同的照明条件、背景和测试角度。测试者提前注册好其面部信息，在人脸上没有张贴任何附着物的正常情况下，测试者可以通过“刷脸”快速解锁手机。

测试者采用在面颊区域分别张贴对抗样本块、随机扰动块以及不进行张贴三种方式。按下电源键，向上滑动屏幕以触发人脸识别解锁功能。当手机终端未检测到人脸时，屏幕上会提示“未检测到人脸”；当手机终端认为检测到人脸，但与注册的人脸不匹配时，会在屏幕上提示“人脸不匹配”，并拒绝解锁；当手机终端认为人脸检测成功且检测到的人脸与注册人脸相匹配时，手机成功解锁。

对于三星 Galaxy S10 5G 和小米 Redmi K20 Pro，当测试者张贴对抗样本块时，手机终端经过很长时间的人脸搜索，最终仍无法检测到人脸，并在屏幕上给出提示“未检测到人脸”；当测试者张贴服从均匀分布的随机扰动块时，手机终端能迅速检测到人脸并给出提示“人脸不匹配”；当测试者不进行张贴时，手机终端可以很快被解锁。这充分展示了我们的对抗样本块具有良好的人脸检测逃逸效果。但当我们对 iPhone 11 手机终端进行人脸检测逃逸测试时，无论实际情况是“未检测到人脸”还是“人脸不匹配”，均显示屏幕顶部的“锁”没有打开，手机仍处于锁定状态。因此，我们无法通过 iPhone 11 验证生成的对抗样本块针对人脸检测模块的逃逸效果。

4.2 手机终端内嵌人脸检测功能相关软件的逃逸实验

为了提高本文面向人脸检测的对抗攻击方法的可信度和说服力，我们使用该方法攻击商业手机终端内置的相机模块。同时，我们针对支付宝刷脸支付模块和美颜自拍软件 B612 也进行了人脸检测逃逸实验。同样，由于手机终端和商业软件的不可访问特性，这一部分内容在本质上也属于黑盒攻击。

（1）内置相机模块

测试者分别在面颊区域张贴对抗样本块、随机扰动块（服从 0～255 均匀分布）以及不进行张贴，打开手机终端自带的相机应用并调整为“人像模式”，切换至前置摄像且将面部正对摄像头。当相机应用检测到人脸时，它会以矩形框的形式框选出人脸；当矩形框无法正确框选出人脸区域或矩形框消失，我们则认为针对人脸检测模块的对抗攻击取得成功。我们针对 Xiaomi，Samsung，Iphone 等手机终端均进行了测试，实验结果如图 5 所示。

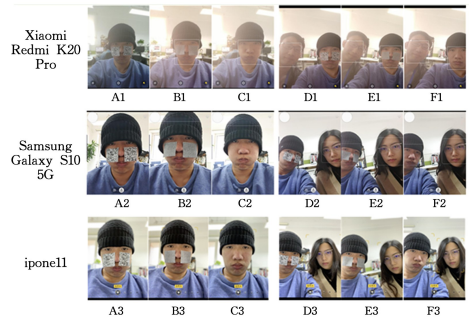

图 5 商业手机终端内置相机的人脸检测逃逸效果

Fig. 5 Face detection escape effect of commercial mobile phones' built-in camera application

由图 5 可以看出，仅在测试者张贴由本文方法生成的对抗样本块时，商业手机终端内置的拍照模块无法以矩形框框选人脸，此时实现了对手机终端自带相机应用的人脸检测模块的成功逃逸，而张贴随机扰动块和不进行张贴均不能实现成功逃逸。

同时，A1 图像明显暗于 B1 和 C1 图像，这是由于测试者张贴对抗样本块后，商业手机终端由于无法检测到人脸而导致自动调光功能失效。这从另一角度验证了本文所提方法对人脸检测模型良好的逃逸能力。

（2）支付宝刷脸支付模块

测试者分别在面颊区域张贴对抗样本块、随机扰动块，以及不进行张贴，在手机终端上打开支付宝并搜索官方服务“支付宝刷脸生活”，在“刷脸设置”中选择“体验刷脸支付”以针对支付宝刷脸支付模块进行人脸检测逃逸实验。实验结果如图 6 所示。

图 6 支付宝刷脸支付人脸检测模块的逃逸效果

Fig. 6 Face detection escape effect of alipay face payment

图 6 给出了使用本文所提出的方法针对支付宝刷脸支付应用的逃逸效果。其中，A4 提示“未检测到人脸”，证明了张贴由本文方法生成的对抗样本块能够实现对支付宝人脸检测模块的成功逃逸；B4 提示“请露出正脸”，这意味着此时该模块认为检测到了人脸，但部分区域被遮挡；C4 提示“请眨眨眼”，这意味着此时刷脸支付模块已检测到人脸并完成了人脸比对，进入活体检测环节。图 6 充分证明了仅在张贴对抗样本块的条件下能够实现对支付宝刷脸支付的成功逃逸，张贴随机扰动块或不进行张贴则不能实现逃逸，即本文方法针对

支付宝刷脸支付的人脸检测模块具备良好的黑盒对抗攻击逃逸效果。

(3)B612 美颜自拍软件

测试者分别在面颊区域张贴对抗样本块、随机扰动块,以及不进行张贴,在每种情况下,实验在两种设置下进行,即打开美颜功能和关闭美颜功能。实验结果如图 7 所示。

图 7 B612 美颜相机软件的人脸检测逃逸效果

Fig. 7 Face detection escape effect of B612

图 7 给出了针对 B612 美颜相机软件的逃逸效果。A5,C5,E5 都没有开启美颜功能,而 B5,D5,F5 则开启美颜功能。实验表明,B612 美颜相机软件的“大眼”和“瘦脸”等美颜功能只有在测试者张贴由本文方法生成的对抗样本块时才会失效(即 B5),这是由于模型无法检测到人脸,后续人脸定位以及相应部位的美颜自然相应失效,充分证明了本文方法针对 B612 美颜相机软件的人脸检测模块具备良好的黑盒对抗攻击逃逸效果(即 B5)。

5 对比实验

为从量化角度研究,分析在 $L_{clf_{\text{total}}}$ 中融合不同数量与组合的人脸检测模型对于现实场景下官方人脸检测模型的黑盒逃逸效果的影响,我们设计了如下对比实验。

5.1 实验说明

实验关注于在现实场景中使用由本文方法生成的对抗样本块攻击各大 IT 厂商所发布的官方人脸检测模型。3.1 节提到,我们在选择参训人脸检测模型时统一选用了非官方的人脸检测模型,正是由于我们在训练对抗样本块的过程中所使用的非官方模型同官方模型在模型结构与参数上均存在较大差异,这部分的实验在本质上仍属于黑盒攻击实验。

5.2 待测官方人脸检测模型选择

我们选用百度官方人脸检测模型 PyramidBox[1)]、腾讯优图官方人脸检测模型 light-DSFD[2)],以及 Awesome 开源推荐人脸检测算法 Yoloface[3)]。

5.3 实验环境

为控制变量,在实验过程中我们保持相同的光照条件、背景图像以及测试角度。

5.4 对比实验

对比实验考虑如下 4 种情况,即在 $L_{clf_{\text{total}}}$ 中分别仅包含 1 个、2 个、3 个、4 个人脸检测模型的 L_{clf} 时,探究人脸检测模型的数量与组合对于在现实场景下攻击官方人脸检测模型所得到的黑盒逃逸效果的影响。

实验在 3 种不同距离条件下完成,即测试者与摄像头的距离分别采用近距离、中等距离和远距离。对上述距离的定义在实验全过程中保持统一且一致。

在每种距离条件下,测试者以恒定速度偏转头部,每种情况均使用屏幕录制软件录制 5 s 视频。使用 FFmpeg 工具逐帧截取视频。逃逸成功率被定义为图像帧中无法被人脸检测模型框选出人脸的帧占全部图像帧的百分比。

在近、中和远 3 种不同距离条件下攻击各官方人脸检测模型的逃逸成功率如表 1 所列。

表 1 参训模型的不同数量与组合对不同距离下逃逸效果的影响

Table 1 Influence of different number and combination of training models on escape effect of different distances

(单位:%)

逃逸率		近距离	中距离	远距离	平均值
攻击官方 Pyramidbox 模型	单模型(MTCNN)	9.27	3.33	6.67	6.42
	双模型(Mtcnn+Pyramidbox)	62.67	52.67	56.00	57.11
	三模型(Mtcnn+Pyramidbox+Facebox)	70.33	76.00	73.33	72.22
	四模型(Mtcnn+Pyramidbox+Facebox+DSFD)	**74.67**	**92.00**	**75.33**	**80.67**
攻击官方 Light-DSFD 模型	单模型(MTCNN)	4.27	15.58	1.19	7.01
	双模型(Mtcnn+Pyramidbox)	35.33	56.67	48.00	46.67
	三模型(Mtcnn+Pyramidbox+Facebox)	72.67	74.00	71.33	72.67
	四模型(Mtcnn+Pyramidbox+Facebox+DSFD)	**100.00**	**92.67**	**84.67**	**92.45**
攻击官方 Yoloface 模型	单模型(MTCNN)	1.32	20.67	19.34	13.78
	双模型(Mtcnn+Pyramidbox)	61.33	72.85	83.33	72.50
	三模型(Mtcnn+Pyramidbox+Facebox)	86.67	78.67	90.67	85.34
	四模型(Mtcnn+Pyramidbox+Facebox+DSFD)	**88.00**	**80.67**	**96.67**	**88.45**

1) http://github.com/EricZgw/PyramidBox

2) http://github.com/610265158/faceboxes-tensorflow/tree/tf1

3) http://github.com/610265158/DSFD-tensorflow

通过平均近、中、远 3 种不同距离条件下攻击官方人脸检测模型的逃逸成功率，我们得到了平均逃逸成功率，如表 2 所列。通过进一步对所攻击的 3 个官方人脸检测模型求取平均值，我们获得不同参训模型数量条件下的平均逃逸成功率，如表 2 的最后 1 列所示。

表 2 参训模型的不同数量与组合对平均逃逸效果的影响

Table 2 Influence of different number and combination of training models on average escape effect

（单位：%）

三距离平均逃逸成功率	攻击官方 Pyramidbox 模型	攻击官方 Light-DSFD 模型	攻击官方 Yoloface 模型	平均值
单模型(MTCNN)	6.42	7.01	13.78	9.07
双模型(Mtcnn+Pyramidbox)	57.11	46.67	72.50	58.76
三模型(Mtcnn+Pyramidbox+Facebox)	72.22	72.67	85.34	76.74
四模型(Mtcnn+Pyramidbox+Facebox+DSFD)	**80.67**	**92.45**	**88.45**	**87.19**

5.5 实验结果分析说明

表 2 充分说明，当参训模型采用单模型、双模型、三模型、四模型进行融合训练时，其对各官方人脸检测模型的平均逃逸成功率分别是 9.07%，58.76%，76.74%，87.19%。在当前实验条件与实验环境下，融合 4 种人脸检测模型相较于融合单模型、双模型、三模型，人脸检测模型的逃逸率分别提高了 78%，28%，10%。上述结果说明融合 4 种人脸检测模型实现了当前最优的人脸检测黑盒逃逸效果，同时证明了本文所提出的方法的有效性和可靠性。

6 展望

本文分析总结了现行人脸检测模型所面临的安全挑战，重点关注于人脸检测在运行监控阶段面临的对抗样本攻击安全风险，并提出一种面向人脸检测的黑盒物理域对抗攻击方法，其能够有效测评现行人脸检测模型存在的安全风险。

人脸检测作为人脸识别、安全认证以及众多人脸相关应用的前序模块，其安全性直接关系到后序模块能否正常运行。实验中针对支付宝刷脸支付和 B612 美颜相机软件的成功逃逸充分说明了人脸相关应用存在着巨大的安全漏洞，而随着人脸检测和人脸识别相关应用的大规模部署，这一风险将愈发严重，因此从人脸检测的安全性与鲁棒性出发来设计更安全的人脸检测模型将受到广泛关注。

受限于所研究领域的庞大繁杂和个人精力有限，我们没能针对人脸检测各阶段的安全风险进行全面评测。同时，针对各个安全风险的防御方法也应有着重要的研究意义与研究价值，这里仅从几个角度对未来的研究方向加以具体展望。

针对部署模型的软硬件环境不可信的安全风险，或可制定软硬件环境可信度评测机制并通过身份验证、限制访问次数和限制访问频率等措施防止用户非法访问或恶意频繁访问模型。

针对废弃阶段模型销毁风险，可设计高可信数据销毁算法，保证模型以及参训人脸图像数据集被妥善处理以防止隐私泄露。同时，可以针对人脸图像集等隐私数据采用脱敏处理，仅存储照片特征，即使被盗，攻击者也无法还原出图像集本身。

针对人脸检测面临的对抗攻击风险，通过进行对抗训练可大幅提升模型应对对抗样本攻击的表现，进而提升模型的鲁棒性。基于本文思路，通过检测常规图像和对抗样本图像在公共注意力热力图层面的显著特征差异，可使分辨面向人脸检测的对抗样本攻击成为可能。

从系统可靠性的角度出发，作为前序模块的人脸检测一旦受到恶意攻击，下流环节的应用可能会意外崩溃。针对这一潜在风险，可以设置模型安全冗余，使得单个人脸检测模型在遭受攻击时不会影响到整个应用的最终功性能，从而提升应用的可靠性。

最后，本文为物理场景下面向人脸检测模型的黑盒攻击提供了有益思路。事实上，本文后续仍有许多可研究工作，未来将从针对对抗样本块所粘贴位置对逃逸效果的影响、利用动态优化算法实现自动寻找最优的对抗样本块粘贴位置、采用全新媒介以生成具备更好逃逸效果的对抗样本块等方面继续开展研究工作。

结束语 人脸检测作为计算机视觉领域的典型应用，在刷脸支付、身份认证、摄像美颜、智能安防等领域均体现出重要的应用价值与广阔的应用前景。本文从人脸检测全阶段出发，系统性分析总结了现行人脸检测模型面临的安全挑战，为增强人脸检测模型的安全性与鲁棒性提供了有益参考。

本文进一步关注于人脸检测面临的对抗样本攻击安全风险，提出了一种面向人脸检测的黑盒物理域对抗攻击方法，用以测评现行人脸检测的安全性与鲁棒性。通过计算公共注意力热力图，该方法成功提取出人脸检测模型的共性弱点。基于对抗攻击的基本思想，通过对公共注意力热力图构造对抗攻击，我们成功地实现了针对支付宝刷脸支付模块、B612 美颜自拍模块的高成功率黑盒对抗攻击。同时，我们实现了针对主流商业手机终端人脸检测模块的成功逃逸，并使其刷脸解锁、自动对焦、自动调光等功能因无法检测到人脸而失效。通过本文提出的攻击方法，可以评估当前人脸检测模型的物理域安全性，有助于进一步了解人脸检测深度神经网络的脆弱性，推动人脸检测深度学习模型向更安全的方向迈进。

参 考 文 献

[1] ZHANG T, HE Z, LEE R B. Privacy-preserving machine learning through data obfuscation[J]. arXiv: 1807.01860, 2018.

[2] XU K, CAO T, SHAH S, et al. Cleaning the null space: A privacy mechanism for predictors[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv: 1312.6199, 2013.

[4] GEIGEL A. Neural network trojan[J]. Journal of Computer Security, 2013, 21(2): 191-232.

[5] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction apis[C]// 25th USENIX Conference on Security. USENIX Association. 2016: 601-618.

[6] SHOKRI R, STRONATI M, SONG C, et al. Membership infe-

rence attacks against machine learning models[C]// 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017: 3-18.
[7] FREDRIKSON M, LANTZ E, JHA S, et al. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing[C]// 23rd USENIX Conference on Security. USENIX Association. 2014: 17-32.
[8] CHEN D, ZHAO H. Data security and privacy protection issues in cloud computing[C]// 2012 International Conference on Computer Science and Electronics Engineering. IEEE, 2012: 647-651.
[9] BOSE A J, AARABI P. Adversarial attacks on face detectors using neural net based constrained optimization[C]// 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2018: 1-6.
[10] ZHOU Z, TANG D, WANG X, et al. Invisible mask: Practical attacks on face recognition with infrared [J]. arXiv: 1803.04683, 2018.
[11] KAZIAKHMEDOV E, KIREEV K, MELNIKOV G, et al. Real-world attack on MTCNN face detection system[C]// 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). IEEE, 2019: 0422-0427.
[12] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. arXiv: 1506.01497, 2015.
[13] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.
[14] CHEN S, HE Z, SUN C, et al. Universal adversarial attack on attention and the resulting dataset damagenet[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020.
[15] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.
[16] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]// Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 1322-1333.
[17] ZHAO Y, ZHU H, LIANG R, et al. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors[C]// Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 1989-2004.
[18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017: 39-57.
[19] TANG X, DU D K, HE Z, et al. Pyramidbox: A context-assisted single shot face detector[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 797-813.
[20] ZHANG S, ZHU X, LEI Z, et al. Faceboxes: A CPU real-time face detector with high accuracy[C]// 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017: 1-9.
[21] LI J, WANG Y, WANG C, et al. DSFD: dual shot face detector [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5060-5069.
[22] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.
[23] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.

JING Hui-yun, born in 1987, Ph.D, senior engineer. Her main research interests include artificial intelligence security and data security.

ZHOU Chuan, born in 1997, postgraduate, is a student member of China Computer Federation. His main research interests include artificial intelligence security and cloud computing security.