计算机科学
COMPUTER SCIENCE

基于多模态特征融合的人脸物理对抗样本性能预测算法

周风帆，凌贺飞，张锦元，夏紫薇，史宇轩，李平

引用本文

周风帆，凌贺飞，张锦元，夏紫薇，史宇轩，李平基于多模态特征融合的人脸物理对抗样本性能预测算法[J]．计算机科学, 2023, 50(8): 280-285.

ZHOU Fengfan, LING Hefei, ZHANG Jinyuan, XIA Ziwei, SHI Yuxuan, LI Ping. Facial Physical Adversarial Example Performance Prediction Algorithm Based on Multi-modal Feature Fusion [J]. Computer Science, 2023, 50(8): 280-285.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100124

基于多模态特征融合的人脸物理对抗样本性能预测算法

周风帆[1]　凌贺飞[1]　张锦元[2]　夏紫薇[1]　史宇轩[1]　李　平[1]
1 华中科技大学计算机科学与技术学院　武汉 430074
2 中国工商银行软件开发中心　广东 珠海 519080
(ffzhou@hust.edu.cn)

摘　要　人脸物理对抗样本攻击(Facial Physical Adversarial Attack,FPAA)指攻击者通过粘贴或佩戴物理对抗样本,如打印的眼镜、纸片等,在摄像头下被识别成特定目标的人脸,或者让人脸识别系统无法识别的攻击方式。已有FPAA的性能评测会受到多种环境因素的影响,且需要多个人工操作的环节,导致性能评测效率非常低下。为了减少人脸物理对抗样本性能评测方面的工作量,结合数字图片和环境因素之间的多模态性,提出了多模态特征融合预测算法(Multimodal Feature Fusion Prediction Algorithm,MFFP)。具体地,使用不同的网络提取攻击者人脸图片、受害者人脸图片和人脸数字对抗样本图片的特征,使用环境特征网络来提取环境因素中的特征,然后使用一个多模态特征融合网络对这些特征进行融合,多模态特征融合网络的输出即为所预测的人脸物理对抗样本图片和受害者图片之间的余弦相似度。MFFP算法在未知环境、未知FPAA算法的实验场景下取得了0.003的回归均方误差,其性能优于对比算法,验证了MFFP算法对FPAA性能预测的准确性,可以对FPAA性能进行快速评估,同时大幅降低人工操作的工作量。

关键词:人工智能安全;对抗样本;人脸物理对抗样本攻击;性能预测;多模态特征融合
中图法分类号　TP391

Facial Physical Adversarial Example Performance Prediction Algorithm Based on Multi-modal Feature Fusion

ZHOU Fengfan[1],LING Hefei[1],ZHANG Jinyuan[2],XIA Ziwei[1],SHI Yuxuan[1] and LI Ping[1]
1 School of Computer Science and Technology,Huazhong University of Science and Technology,Wuhan 430074,China
2 Software Development Center,Industrial and Commercial Bank of China,Zhuhai,Guangdong 519080,China

Abstract　Facial physical adversarial attack(FPAA) refers to a method that an attacker pasting or wearing physical adversary examples,such as printed glasses,paper,to make the face recognition system to recognize his face as the face of a specific target,or make the face recognition system unable to recognize his face under the camera. The existing performance evaluation process of the FPAA can be affected by multiple environmental factors and require multiple manual operations,resulting in very low efficiency of performance evaluation. In order to reduce the workload of evaluating the performance of facial physical adversarial examples,combined with the multimodality between digital images and environmental factors,a multimodal feature fusion prediction algorithm(MFFP) is proposed. Specifically,different networks are used to extract the features of attacker's face images,victim's face images and facial digital adversarial example images,and the proposed environmental feature extraction network is used to extract the features of environmental factors. A multimodal feature fusion network is proposed to fuse these features. The output of the multimodal feature fusion network is the cosine similarity performance between the predicted facial physical adversarial example image and the victim image. MFFP algorithm achieves a regression mean square error of 0.003 under the experimental scenario of unknown environment and unknown FPAA,which is better than the performance of the baseline. It verifies the accuracy of MFFP algorithm for predicting of the performance of FPAA. Moreover,it verifies that MFFP can quickly evaluate the performance of FPAA,while greatly reduce the workload of manual operation.

到稿日期:2022-11-15　返修日期:2023-03-04
基金项目:国家自然科学基金(61972169);国家重点研发计划(2019QY(Y)0202,2022YFB2601802);湖北省重点研发计划(2022BAA046,2022BAA042);武汉基础研究知识创新项目(2020010601012182);中国博士后科学基金(2022M711251)
This work was supported by the Natural Natural Science Foundation of China(61972169),National Key Research and Development Program of China(2019QY(Y)0202,2022YFB2601802),Major Scientific and Technological Project of Hubei Province(2022BAA046,2022BAA042),Research Programme on Applied Fundamentals and Frontier Technologies of Wuhan(2020010601012182) and China Postdoctoral Science Foundation(2022M711251).
通信作者:凌贺飞(lhefei@hust.edu.cn)

Keywords Artificial intelligence security, Adversarial example, Facial physical adversarial attack, Performance prediction, Multi-modal feature fusion

1 引言

随着机器学习(尤其是深度学习)技术的不断发展，机器学习技术已经得到了广泛的应用。然而对抗样本[1]的存在却给这些机器学习系统的安全性带来了极大威胁。对抗样本是一种由攻击者生成的，会使机器学习系统产生符合攻击者预期的定向或者非定向错误的数据。使用对抗样本对机器学习系统进行的攻击被称为对抗样本攻击或者对抗攻击。

人脸识别是机器学习中的一个常见任务，同样，基于深度学习的人脸识别系统也具有对对抗样本攻击的脆弱性[2]。对人脸识别进行对抗样本攻击的对抗样本被称为人脸对抗样本。如果一个攻击者使用对抗样本对人脸识别系统进行攻击，只要对抗样本的性能足够好，理论上可以成功攻击任何未加防御措施的人脸识别系统。

因此，人脸对抗样本的存在给现有的人脸识别系统的安全性带来了很大的威胁。其中一个较为严重的威胁是人脸物理对抗样本带来的威胁。人脸物理对抗样本是一种存在于现实世界中的人脸对抗样本，相对于仅仅存在于数字世界的人脸对抗样本，人脸物理对抗样本的应用范围更广。

虽然人脸物理对抗样本的应用范围更广、危害性更大，但是各种环境因素(如光照强度、对抗样本的形变程度、拍摄对抗样本时的图片压缩程度等)可能会对人脸物理对抗样本的性能产生影响[3]。然而，目前对在不同环境因素影响下的人脸物理对抗样本的性能的研究极少。其中阻碍该研究进行的一个重要的原因是对人脸物理对抗样本在不同环境下的性能进行评测的人工工作量很大。具体地，人脸物理对抗样本性能评测过程可以分为人脸数字对抗样本的生成、人脸对抗样本的打印、人脸对抗样本的粘贴、人脸对抗样本的拍摄、人脸对抗样本的性能计算 5 部分。这 5 部分中包含大量的人工操作。根据本课题前期的实验数据统计，人工操作(如粘贴或者不佩戴人脸对抗样本等操作)占用了人脸物理对抗样本性能评测过程中的大部分的时间。

为了减轻人脸物理对抗样本性能评测工作的工作量，本文提出了人脸物理对抗样本攻击性能预测任务，该任务主要是在不粘贴或者不佩戴人脸物理对抗样本的情况下，通过环境因素、人脸数字对抗样本图片以及其他图片对该环境下的 FPAA 的性能进行预测。为了完成该任务，基于多模态特征融合的算法，本文提出了多模态特征融合预测算法(MFFP)。MFFP 算法使用不同的人脸特征提取网络提取攻击者人脸图片、受害者人脸图片和人脸数字对抗样本图片的特征，并使用环境特征提取网络提取环境因素的特征。之后，将这些特征进行融合，最后通过一个网络对融合后的特征进行回归，使网络的输出与真实的人脸物理对抗样本图片和受害者人脸图片之间的余弦相似度更加接近。MFFP 可以对某一个特定环境下的人脸物理对抗样本的性能进行预测，给科研人员对不同环境中的人脸物理对抗样本的性能评测提供了便利，减少了科研人员对不同环境中的人脸物理对抗样本的性能进行评测的工作量。

2 相关工作

本节主要从对抗样本攻击、人脸数字对抗样本攻击和人脸物理对抗样本攻击这 3 个方面对本文的相关工作进行介绍。

2.1 对抗样本攻击

对抗样本攻击是使用攻击者精心制作的数据对机器学习系统进行攻击，以使机器学习系统产生定向或者非定向错误的过程。对抗样本攻击可以分为数字对抗样本攻击和物理对抗样本攻击两种。数字对抗样本攻击主要是使用数字形式的数据对机器学习系统进行的攻击，整个攻击过程均在数字世界中进行。而物理对抗样本攻击主要是使用物理形式的数据对机器学习系统进行的攻击，在攻击过程中需要将生成的对抗样本显现到现实世界中以实现对机器学习系统的攻击。其中显现到现实世界中的方式包括打印粘贴对抗样本[4]、使用激光光束修改相机拍摄的图像[5]等。典型的数字对抗样本攻击有动量迭代方法(Momentum Iterative Method, MIM)[6]、多样化输入方法(Diverse Input Method, DIM)[7]等；典型的物理对抗样本攻击有脏路补丁(Dirty Road Patch, DRP)[4]攻击方法和对抗激光束(Adversarial Laser Beam, AdvLB)攻击方法[5]等。

2.2 人脸数字对抗样本攻击

本节主要对最新的人脸数字对抗样本攻击方法进行介绍。

(1)Dropout 人脸攻击网络

一般生成人脸对抗样本的算法使用的人脸识别模型是训练好的人脸识别模型，Zhong 等[8]认为在生成人脸对抗样本的过程中依然可以改变训练好的模型的参数，从而提高用于生成人脸对抗样本的模型的多样性，进而提高生成的人脸对抗样本的可迁移性。具体来说，Zhong 等在生成人脸对抗样本的过程中使用了 dropout 算法以使生成的人脸对抗样本产生类似于集成的效果，且将该算法命名为 dropout 人脸攻击网络(Dropout Face Attacking Network, DFANet)[8]。因此，对于黑盒攻击，DFANet 在生成对抗样本的过程中可以有效防止对抗样本对代理模型的过拟合。

(2)定向个体保护迭代方法

Yang 等[9]提出了定向个体保护迭代方法(Targeted Identity-Protection Iterative Method, TIP-IM)。TIP-IM 是一种针对人脸分类任务的对抗样本攻击，而在此之前的大部分人脸对抗样本攻击工作是针对人脸验证任务的对抗样本攻击。该工作的主要目的是对人脸图片进行加密，使得不法分子的深度学习系统无法从经过该算法处理的人脸图片中识别出该图片所属的个体的真实身份。该算法的主要内容是在需要被加密的人脸图片上添加对抗噪声，使得不法分子的深度学习系统将添加对抗噪声之后的人脸图片识别成预先定义好的 gallery 数据集 G_1 中的图片(识别成其中的任意一张图片即算加密成功)。为了使生成的人脸对样本即加密后的图片

更加逼真，Yang 等[9]在生成对抗样本的过程中添加了最大平均差异(Maximum Mean Discrepancy，MMD)损失以提高生成的人脸对抗样本的自然性。为了使生成对抗样本的过程更加高效，他们还使用 Greedy Insertion 算法在 G_1 中选择合适的图片作为受害者人脸图片以提高最终生成的对抗样本的效率。

2.3 人脸物理对抗样本攻击

本节主要对最新的人脸物理对抗样本攻击方面的工作以及人脸物理对抗样本的评测过程进行介绍。

对抗眼镜[10]攻击是一种在眼镜镜框周围添加对抗噪声的攻击方法，该攻击方法采用基于优化的方法生成对抗样本。对抗帽子[11]攻击是一种在帽子上添加对抗噪声的攻击方法，该方法在生成对抗样本的过程中考虑了对抗样本的扭曲程度，以应对人脸物理对抗样本粘贴过程中产生的扭曲形变。对抗眼影(Adversarial Makeup，Adv-Makeup)[12]攻击是一种基于 GAN 的人脸物理对抗样本攻击方法。原先的人脸对抗样本攻击工作在对抗样本攻击时使用的物理对抗样本往往会引起别人的怀疑。为了解决该问题，Adv-Makeup 使用 GAN 在人的眼部部位生成类似于眼影的对抗样本以提高生成的人脸物理对抗样本的自然程度。为了提高对抗样本的可迁移性，Adv-Makeup 在生成对抗样本的过程中使用了元学习的方法。

虽然这些人脸物理对抗样本方面的工作提高了人脸物理对抗样本的攻击性能，然而却忽视了人脸物理对抗样本研究过程中的一个很大的问题——人脸物理对抗样本性能评测过程的低效性。人脸物理对抗样本性能评测过程可以分为人脸数字对抗样本的生成、人脸对抗样本的打印、人脸对抗样本的粘贴、人脸对抗样本的拍摄、人脸对抗样本的性能计算 5 部分。人脸数字对抗样本的生成过程主要是使用某种人脸对抗样本攻击方法生成具有一定的攻击能力的人脸数字对抗样本。人脸对抗样本的打印的过程主要是将人脸数字对抗样本生成过程生成的人脸对抗样本调整到合适的尺寸，并通过打印机打印出来。由于不同攻击者的人脸的五官位置不同，调整到合适的尺寸较为困难，因此该过程可能较为耗时。人脸对抗样本的粘贴过程主要是将打印出的人脸对抗样本粘贴在志愿者脸上，或者将人脸对抗样本粘贴到眼镜上并使志愿者佩戴粘贴上人脸对抗样本的眼镜。人脸对抗样本的拍摄过程主要是通过摄像头将佩戴或者粘贴上人脸对抗样本的志愿者的人脸图片录入计算机中。人脸对抗样本的性能计算的过程主要是对物理对抗样本图片和受害者人脸图片的性能进行计算，一般使用人脸对抗样本图片和受害者人脸图片之间的余弦相似度作为性能的衡量指标。总体来说，人脸物理对抗样本性能评测过程存在人工工作量大的问题，该问题阻碍了人脸物理对抗样本研究的发展。

3 MFFP 算法

如第 2 节所述，原有的人脸物理对抗样本性能评测的主要流程是通过一系列的复杂物理操作获取到人脸物理对抗样本图片和受害者图片之间的余弦相似度值。因此，原有的人脸物理对抗样本性能评测流程存在着人工工作量大的问题。

为了解决该问题，本文提出了人脸物理对抗样本性能预测任务。该任务的目标主要是通过数字图片和环境因素对人脸物理对抗样本的性能即人脸物理对抗样本图片和受害者图片之间的余弦相似度进行预测。该任务的框架图如图 1 所示。图 1 中的黑色虚线表示人脸数字对抗样本攻击的过程，实线表示人脸物理对抗样本攻击的过程。图 1 中的问号人脸图片表示在该任务的执行过程中未获得真实的人脸物理对抗样本，表明该任务的完成无需在人脸上粘贴或者佩戴人脸物理对抗样本。

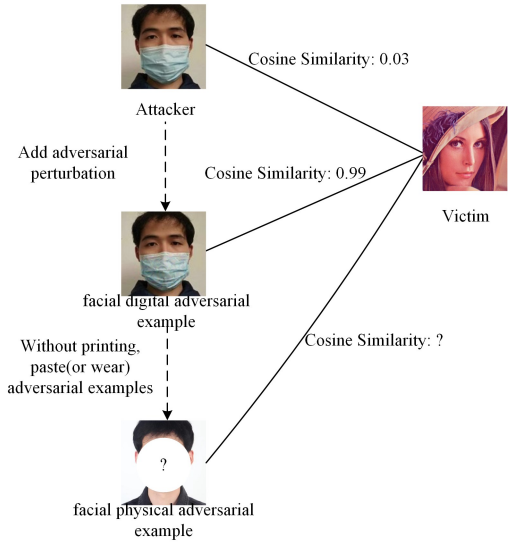

图 1 人脸物理对抗样本攻击性能预测任务

Fig. 1 Facial physical adversarial examples performance prediction task

由于预测人脸物理对抗样本的性能需要使用的信息不仅仅只是图像信息，或者只是表格信息，因此使用传统的图像数据回归网络(如视觉 Transformer(Vision Transformer)[13]、多层感知机混合器(MLP-Mixer)[14])或者表格数据回归网络(如表格网络(TabularNet)[15])不能对不同环境下的人脸物理对抗样本的性能进行很好的预测，即不能很好地完成人脸物理对抗样本性能预测任务。为了解决这个问题，本文提出了一种利用多模态特征融合来将预测人脸物理对抗样本性能时需要使用的图像信息和表格信息进行融合的算法，从而对不同环境下的人脸物理对抗样本的性能进行很好的预测。具体来说，本文定义了 3 种网络，分别是人脸特征提取网络、环境特征提取网络和多模态特征融合网络。人脸特征提取网络用来提取人脸图片的特征；环境特征提取网络用来提取人脸物理对抗样本的光照强度、粘贴位置、拍摄角度、形变程度和 JPEG 压缩程度等环境因素的特征；多模态特征融合网络将人脸特征提取网络以及环境特征提取网络提取到的特征进行多模态融合。算法的整体结构图如图 2 所示。

由于本文提出的算法主要是通过多模态特征融合对人脸物理对抗样本图片和其对应的受害者人脸图片之间的余弦相似度进行预测，因此该算法被称为多模态特征融合预测算法(MFFP)。MFFP 算法使用的损失函数为均方误差(Mean Square Error，MSE)损失，其计算式如式(1)所示：

$$L=\frac{1}{N}\sum_i(sim_i^x-sim_i^y)^2 \tag{1}$$

其中，N 指当前 batch 的大小，sim_i^y 指人脸物理对抗样本图片和受害者图片之间的余弦相似度的 ground truth，sim_i^f 指多模态特征融合网络输出的人脸物理对抗样本图片和受害者图片之间的余弦相似度的预测值。下面将对 MFFP 算法的 3 个子网络，即环境特征提取网络、人脸特征提取网络和多模态特征融合网络进行介绍。

(1)环境特征提取网络

环境特征提取网络的主要目的是提取环境因素的特征。由于人脸物理对抗样本的环境因素之间的关系可能很复杂，并且可能存在相互制约和相互影响的因素，因此，MFFP 算法使用神经网络来提取这些环境因素的特征。使用神经网络提取特征的过程可以被看作是对影响人脸物理对抗样本的环境因素及其之间的关系进行建模的过程。MFFP 算法使用的环境特征提取网络的主要结构如图 3 所示。

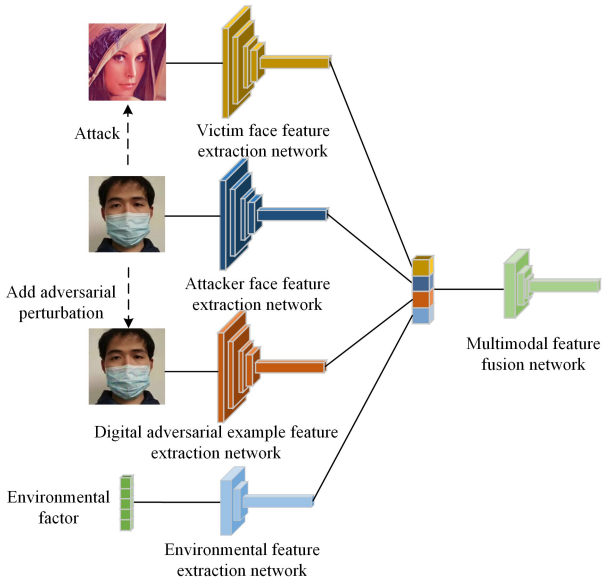

图 2 MFFP 算法的整体结构图

Fig. 2 Overall framework of MFFP algorithm

如图 3 所示，我们先将光照强度、对抗样本的形变程度、拍摄对抗样本时的图片压缩程度的数值拼接成一个向量，并将该向量输入环境特征提取网络。之后，依次将这些环境因素进行一次或者多次线性运算和激活(激活层在图中未画出)，最后再进行一次线性运算，即得到环境因素的特征。在实际应用过程中，一般的激活函数均可以被选择作为环境特征提取网络的激活函数。对于环境特征提取网络的训练，我们采用将环境特征提取网络作为整体网络的一个子网络的方式对其进行训练。

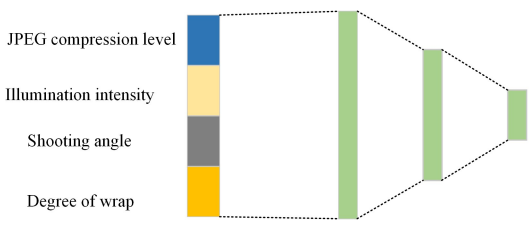

图 3 环境特征提取网络结构图

Fig. 3 Framework of environmental feature extraction network

(2)人脸特征提取网络

人脸特征提取网络分为攻击者人脸特征提取网络、受害者人脸特征提取网络和数字对抗样本特征提取网络，它们分别用于提取攻击者人脸图片的特征、受害者人脸图片的特征和人脸数字对抗样本图片的特征。

在实际应用的过程中，常用的图像识别网络(如 ResNet，DenseNet 和 EfficientNet[16]等)一般均可以作为攻击者人脸特征提取网络、受害者人脸特征提取网络和数字对抗样本特征提取网络。最终算法的性能和所选择的网络结构有关。

(3)多模态特征融合网络

多模态特征融合网络的主要目的是对受害者人脸特征、攻击者人脸特征、数字对抗样本特征和环境特征进行特征融合。使用多模态特征融合网络进行特征融合的过程可以被看作是对影响人脸物理对抗样本的环境因素、受害者人脸特征、攻击者人脸特征、数字对抗样本特征及其之间的关系进行建模的过程。MFFP 算法使用的多模态特征融合网络的主要结构如图 4 所示。

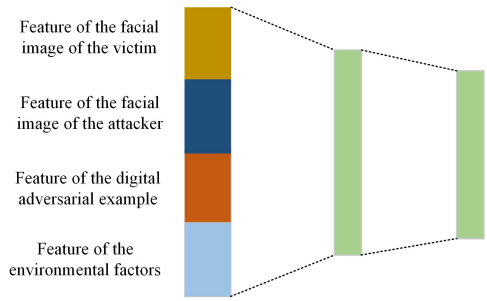

图 4 多模态特征融合网络结构图

Fig. 4 Framework of multimodal feature fusion network

4 实验验证

本节首先对人脸物理对抗样本性能预测数据集的制作进行阐述，接着对 MFFP 算法在不同的人脸识别网络下的性能进行论述。

4.1 人脸物理对抗样本性能预测任务数据集的制作

人脸物理对抗样本性能预测任务是通过人脸数字图片和环境因素对人脸物理对抗样本图片的性能进行预测的任务，因此该任务对应的数据集应该包含攻击者人脸图片、受害者人脸图片、人脸数字对抗样本图片、环境因素和人脸物理对抗样本图片 5 部分。对于该数据集的制作，我们先使用 Zhong 等在文献[8]中使用的 r50-webface-arc 模型(该模型对应的 ROC 曲线见文献[8])制作人脸数字对抗样本。然后，在多个不同的环境下将这些人脸数字对抗样本打印并粘贴在志愿者脸上。最后，我们对粘贴上人脸对抗样本的志愿者的人脸进行拍摄，并记录拍摄时的光照强度、粘贴位置、拍摄角度和形变程度等环境因素。我们共采集了约 46800 个不同环境下的人脸物理对抗样本。采集人脸物理对抗样本时使用的部分设备的图片如图 5 所示。图 5(a)是用于获取粘贴物理对抗样本时环境光照强度的照度仪的图片。图 5(b)是用于调整环境光照强度的台灯的图片。

(a) Illuminometer used to obtain the ambient light intensity when pasting the physical adversarial sample

(b) Desk lamp for adjusting ambient light intensity

图 5 采集人脸物理对抗样本时使用的部分设备的图片

Fig. 5 Pictures of some of the devices used to collect facial physical adversarial examples

4.2 MFFP 算法的实验验证

由于人脸物理对抗样本性能预测任务是本课题提出的，因此无法找出基准算法与 MFFP 算法的性能进行对比。为了验证 MFFP 算法的有效性，本课题定义了两种算法进行对比实验。这两种算法和 MFFP 算法的主要区别在于：第一种算法使用同一个人脸特征提取网络提取攻击者的人脸特征、受害者的人脸特征和数字对抗样本的特征，该算法被称为 Single Branch(SB)；第二种算法使用两个不同的人脸特征提取网络分别提取人脸对抗样本的特征和人脸干净样本的特征，该算法被称为 Double Branch(DB)。

MFFP 算法设计的目的是减少研究者对人脸物理对抗样本未知的人脸对抗样本生成算法在不同环境下进行性能评估的工作量。其中人脸物理对抗样本未知可以分为 3 种类型：第一种类型是环境未知，比如研究人员想获取到一个新的环境下的人脸物理对抗样本的性能，然而用于训练 MFFP 模型的数据集中的对抗样本所对应的环境集合中不包含研究人员期望测试的人脸物理对抗样本性能的环境，这种类型被称为“未知环境”(Unknown Environment，UE)；第二种类型是 FPAA 未知，比如研究人员提出了一个新的 FPAA，并希望用原先已经训练好的 MFFP 模型对该 FPAA 生成的人脸物理对抗样本的性能进行预测，然而用于训练 MFFP 模型的数据集中的对抗样本所对应的 FPAA 集合中不包含研究人员提出的该人脸对抗样本生成算法，这种类型被称为“未知算法”(Unknown Algorithm，UA)；第三种类型是环境未知并且人脸对抗样本生成算法未知，这种类型是 UE 和 UA 的结合，被称为 UE+UA。在 UE，UA 或 UE+UA 下对 MFFP 的算法的性能进行测试可以看作对 MFFP 的泛化性进行测试。计划从 3 个实验配置上对 MFFP 算法的泛化性进行测试。第一个实验配置是 UE，第二个实验配置是 UA，第三个实验配置是 UE+UA。

经过实验，得出算法在实验配置为 UE 的情况下的结果，如表 1 所列。从表 1 的实验结果可以看出，在 UE 实验配置的大部分情况下，MFFP 取得了比 SB 和 DB 更好的结果。即使训练集中不包含测试集中的环境，MFFP 依然可以对测试集中的人脸物理对抗样本的性能进行很好的预测。

表 1 不同算法在不同图像特征提取网络上的 UE 性能

Table 1 UE performance of different algorithms on different image feature extraction networks

Network	*MSE*			
	SB	DB	MFFP	Average
ResNet	46.1×10^{-4}	27.0×10^{-4}	62.9×10^{-4}	45.3×10^{-4}
DenseNet	30.3×10^{-4}	54.0×10^{-4}	24.7×10^{-4}	36.3×10^{-4}
MobileNetV2	95.4×10^{-4}	27.7×10^{-4}	23.9×10^{-4}	49.0×10^{-4}
WideResNet	109.7×10^{-4}	78.8×10^{-4}	45.7×10^{-4}	78.0×10^{-4}
EfficientNet	75.8×10^{-4}	32.4×10^{-4}	37.3×10^{-4}	48.5×10^{-4}
Average	71.5×10^{-4}	44.0×10^{-4}	38.9×10^{-4}	51.4×10^{-4}

算法在实验配置为 UA 的情况下的结果如表 2 所列。从表 2 的实验结果可以看出，在 UA 实验配置的大部分情况下，MFFP 取得了比 SB 和 DB 更好的结果。即使训练集中不包含测试集中的人脸对抗样本生成算法，MFFP 依然可以对测试集中的人脸物理对抗样本的性能进行很好的预测。

表 2 不同算法在不同图像特征提取网络上的 UA 性能

Table 2 UA performance of different algorithms on different image feature extraction networks

Network	*MSE*			
	SB	DB	MFFP	Average
ResNet	72.7×10^{-4}	47.1×10^{-4}	60.5×10^{-4}	60.1×10^{-4}
DenseNet	38.6×10^{-4}	55.8×10^{-4}	30.2×10^{-4}	41.5×10^{-4}
MobileNetV2	62.5×10^{-4}	112.4×10^{-4}	46.8×10^{-4}	73.9×10^{-4}
WideResNet	68.0×10^{-4}	49.5×10^{-4}	33.5×10^{-4}	50.3×10^{-4}
EfficientNet	64.1×10^{-4}	66.4×10^{-4}	40.1×10^{-4}	56.9×10^{-4}
Average	61.2×10^{-4}	66.2×10^{-4}	42.2×10^{-4}	56.6×10^{-4}

算法在实验配置为 UE+UA 的情况下的结果如表 3 所列。

表 3 不同算法在不同图像特征提取网络上的 UE+UA 性能

Table 3 UE+UA performance of different algorithms on different image feature extraction networks

Network	*MSE*			
	SB	DB	MFFP	Average
ResNet	50.9×10^{-4}	40.2×10^{-4}	34.5×10^{-4}	41.9×10^{-4}
DenseNet	74.5×10^{-4}	48.7×10^{-4}	44.2×10^{-4}	55.8×10^{-4}
MobileNetV2	48.4×10^{-4}	74.5×10^{-4}	32.9×10^{-4}	52.0×10^{-4}
WideResNet	56.2×10^{-4}	186.3×10^{-4}	44.5×10^{-4}	95.7×10^{-4}
EfficientNet	77.6×10^{-4}	55.3×10^{-4}	42.9×10^{-4}	58.6×10^{-4}
Average	61.5×10^{-4}	81.0×10^{-4}	39.8×10^{-4}	60.8×10^{-4}

从表 3 的实验结果可以看出，在 UE+UA 实验配置的情况下，MFFP 取得了比 SB 和 DB 更好的结果。即使训练集中不包含测试集中的环境和人脸对抗样本生成算法，MFFP 依然可以对测试集中的人脸物理对抗样本的性能进行很好的预测。

4.3 通过数字图片和环境因素可以对人脸物理对抗样本性能进行预测的原因分析

虽然 MFFP 取得了很好的人脸物理对抗样本性能预测性能，然而通过数字图片和环境因素可以对人脸物理对抗样本性能进行预测的原因是未知的。对于该原因的分析，我们首先需要理解人脸物理对抗样本的性能主要受到哪些因素的影响。人脸物理对抗样本的性能主要指受害者人脸识别模型提取的人脸物理对抗样本图片和受害者图片之间的余弦相似度。由于受害者图片是数据集中已经固定下来的图片，因此

人脸物理对抗样本图片是该余弦相似度值的主要影响因素；而人脸物理对抗样本是由数字图片在环境因素的作用下经过一系列的变换产生的，该变换相当于一个输入为数字图片和环境因素的函数。原有的工作表明，深度神经网络可以拟合任意的函数[17]，因此 MFFP 方法对该函数的拟合是可行的，即仅仅通过数字图片和环境因素可以对物理对抗样本的性能进行预测。

结束语 为了减少在 FPAA 评测过程中粘贴或者佩戴人脸物理对抗样本的工作量，提出了 FPAA 性能预测任务。针对该任务，基于多模态特征融合方法，提出了多模态特征融合预测算法。MFFP 通过环境因素、人脸数字对抗样本图片以及其他数字图片对该环境下的 FPAA 的性能进行预测。实验结果表明，相较于对比方法，MFFP 取得了更低的回归均方误差，验证了 MFFP 对性能预测的准确性与 FPAA 性能预测任务的可行性。

参 考 文 献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]// International Conference on Learning Representations. 2014: 1-10.

[2] QIU H N, XIAO C W, YANG L, et al. SemanticAdv: Generating Adversarial Examples via Attribute-conditional Image Editing [C] // European Conference on Computer Vision. Springer. 2020: 19-37.

[3] SHEN M, YU H, ZHU L H, et al. Effective and Robust Physical－World Attacks on Deep Learning Face Recognition Systems [J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4063-4077.

[4] SATO T, SHEN J J, WANG N F, et al. Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack[C]// USENIX Security Symposium. USENIX Association. 2021: 3309-3326.

[5] DUAN R J, MAO X F, QIN K. A, et al. Adversarial laser beam: Effective physical-world attack to DNNs in a blink[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 16062-16071.

[6] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 9185-9193.

[7] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 2730-2739.

[8] ZHONG Y Y, DENG W H. Towards transferable adversarial attack against deep face recognition [J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 1452-1466.

[9] YANG X, DONG Y P, PANG T Y, et al. Towards face encryption by generating adversarial identity masks[C]// International Conference on Computer Vision. IEEE, 2021: 3897-3907.

[10] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition[C]// {ACM} {SIGSAC} Conference on Computer and Communications Security. ACM, 2016: 1528-1540.

[11] KOMKOV S, PETIUSHKO A. AdvHat: Real-World Adversarial Attack on ArcFace Face {ID} System[C]// International Conference on Pattern Recognition. IEEE, 2020: 819-826.

[12] YIN B J, WANG W X. YAO T P, et al. Adv-Makeup: A New Imperceptible and Transferable Attack on Face Recognition [C]// International Joint Conference on Artificial Intelligence. 2021: 1252-1258

[13] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]// International Conference on Learning Representations. 2021: 1-21.

[14] TOLSTIKHIN I, HOULSBY N, KOLESNIKOV A, et al. Mlp-mixer: An all-mlp architecture for vision[C]// Advances in Neural Information Processing Systems. MIT Press, 2021: 24261-24272.

[15] DU L, GAO F, CHEN X, et al. TabularNet: A Neural Network Architecture for Understanding Semantic Structures of Tabular Data[C]// ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, 2021: 322-331.

[16] TANM X, QUOC V L E. Efficientnet: Rethinking model scaling for convolutional neural networks[C]// International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[17] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators [J]. Neural Networks, 1989, 2: 359-366.

ZHOU Fengfan, born in 1998, Ph.D. His main research interest is adversarial attacks on face recognition.

LING Hefei, born in 1976, Ph.D supervisor. His main research interest is computer vision.

（责任编辑：何杨）