

面向商用活体检测平台的鲁棒性评估

王鹏程¹, 郑海斌¹, 邹健飞¹, 庞玲², 李虎², 陈晋音^{1,3}

(1. 浙江工业大学信息工程学院, 浙江 杭州 310023;

2. 信息系统安全技术重点实验室, 北京 100101;

3. 浙江工业大学网络空间安全研究院, 浙江 杭州 310023)

摘要: 活体检测技术已经成为日常生活中的重要应用, 手机刷脸解锁、刷脸支付、远程身份验证等场景都会用到这一技术。但如果攻击者利用虚假视频生成技术生成逼真的换脸视频来攻击上述场景的活体检测系统, 将会对这些场景的安全性产生巨大的威胁。针对这个问题使用4种先进的Deepfake技术生成大量的换脸图片和视频作为测试样本, 用这些样本来对百度、腾讯等商用活体检测平台的在线API接口进行测试。测试实验结果显示常用的各大商用活体检测平台对Deepfake图像的检测成功率普遍很低, 并且对图像的质量较为敏感, 对真实图像的误检率也很高。其主要原因可能是这些平台设计时针对的是打印照片攻击、屏幕二次翻拍攻击、硅胶面具攻击等传统的活体检测攻击方法, 并未将先进的换脸检测技术集成到他们的活体检测算法中, 这些平台因此不能够有效应对Deepfake攻击。因此, 提出了一种集成活体检测方法Integranet, 该方法由4种针对不同图像特征的检测算法集成所得, 既能够有效检测出打印照片、屏幕二次翻拍等传统的攻击手段, 也能够有效应对先进的Deepfake攻击。在测试数据集上验证Integranet的检测效果, 结果显示Integranet检测方法相比较各大商用活体检测平台, 对Deepfake图像的检测成功率至少提高35%以上。

关键词: 深度伪造; 活体检测; 商用平台; 鲁棒性

中图分类号: TP391

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022010

Robustness evaluation of commercial liveness detection platform

WANG Pengcheng¹, ZHENG Haibin¹, ZOU Jianfei¹, PANG Ling², LI Hu², CHEN Jinyin^{1,3}

1. The College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

2. National Key Laboratory of Science and Technology on Information System Security, Beijing 100101, China

3. The Institute of Cybernetics Security, Zhejiang University of Technology, Hangzhou 310023, China

收稿日期: 2021-08-30; **修回日期:** 2021-11-03

通信作者: 陈晋音, chenjinyin@zjut.edu.cn

基金项目: 国家重点研发计划(2018AAA0100801); 国家自然科学基金(62072406); 信息系统安全技术重点实验室资助(61421110502); 浙江省重点研发计划项目(2021C01117); 2020年工业互联网创新发展工程项目(TC200H01V); 浙江省“万人计划”科技创新领军人才项目(2020R52011)

Foundation Items: The National Key R&D Program of China (2018AAA0100801), The National Natural Science Foundation of China (62072406), National Key Laboratory of Science and Technology on Information System Security (61421110502), The Key R&D Program of Zhejiang Province (2021C01117), The 2020 Industrial Internet Innovation Development Project (TC200H01V), “Ten Thousand Talents Program” of Zhejiang Province (2020R52011)

引用格式: 王鹏程, 郑海斌, 邹健飞, 等. 面向商用活体检测平台的鲁棒性评估[J]. 网络与信息安全学报, 2022, 8(1): 180-189.

Citation Format: WANG P C, ZHENG H B, ZOU J F, et al. Robustness evaluation of commercial liveness detection platform[J]. Chinese Journal of Network and Information Security, 2022, 8(1): 180-189.

Abstract: Liveness detection technology has become an important application in daily life, and it is used in scenarios including mobile phone face unlock, face payment, and remote authentication. However, if attackers use fake video generation technology to generate realistic face-swapping videos to attack the living body detection system in the above scenarios, it will pose a huge threat to the security of these scenarios. Aiming at this problem, four state-of-the-art Deepfake technologies were used to generate a large number of face-changing pictures and videos as test samples, and use these samples to test the online API interfaces of commercial live detection platforms such as Baidu and Tencent. The test results show that the detection success rate of Deepfake images is generally very low by the major commercial live detection platforms currently used, and they are more sensitive to the quality of images, and the false detection rate of real images is also high. The main reason for the analysis may be that these platforms were mainly designed for traditional living detection attack methods such as printing photo attacks, screen remake attacks, and silicone mask attacks, and did not integrate advanced face-changing detection technology into their liveness detection. In the algorithm, these platforms cannot effectively deal with Deepfake attacks. Therefore, an integrated live detection method Integranet was proposed, which was obtained by integrating four detection algorithms for different image features. It could effectively detect traditional attack methods such as printed photos and screen remakes. It could also effectively detect against advanced Deepfake attacks. The detection effect of Integranet was verified on the test data set. The results show that the detection success rate of Deepfake images by proposed Integranet detection method is at least 35% higher than that of major commercial live detection platforms.

Keywords: Deepfake, liveness detection, commercial platform, robustness

0 引言

深度神经网络广泛应用于图像识别^[1]、物体识别^[2]、自然语言处理^[3-5]、信号处理^[6]等领域。在视频图像领域,深度神经网络应用在一些对安全敏感的应用程序,如人脸识别等。人脸识别是基于人的脸部特征信息进行身份识别。人脸识别系统面临3种欺骗攻击:使用合法用户的照片、使用合法用户的视频、使用合法用户的三维模型或面具头套。其中,照片欺骗和视频欺骗是攻击人脸识别系统最频繁的方法^[7]。为了保护人脸识别系统免受上述攻击手段的影响,活体检测技术应运而生,成为日常重要应用。活体检测技术旨在判断系统采集的人脸数据是否为真实的人脸,以防止伪造人脸攻击。完善的活体检测技术是人脸身份认证系统进一步普及应用的前提条件,可广泛应用于各种场景或系统,包括刷脸支付、人脸解锁、公共安防以及各种在线人脸识别系统等^[8]。

近几年来,随着深度学习技术的不断发展,基于深度学习的视频伪造技术(简称 Deepfake^[9]技术)日趋成熟。不法分子可以通过获取公众人

物公开的视频或图片,轻松地对其进行换脸,进而做出散播谣言、编造虚假新闻、破坏他人名誉等恶劣行为,对我国网络环境的治理造成了不小的阻力^[10]。Deepfake 的出现不但给数字图像取证技术^[11-12]带来了巨大的挑战,同时给活体检测技术带来了巨大的威胁。据统计,2019 年年初网上有 7 964 段 Deepfake 视频,这个数字在 9 个月内几乎翻了一番,达到 14 678 段。除了数量大,Deepfake 视频质量在短时间内提高了很多。质量的提高使得人们越来越难检测出真实和虚假视频之间的差异^[13]。

针对 Deepfake 技术给活体检测系统带来的日益严峻的威胁。本文对常用的商用活体检测平台进行了鲁棒性评估,百度、腾讯、阿里巴巴、京东、科大讯飞、商汤科技、云从科技等互联网公司都提供商用活体检测平台 API 接口,主要用于对用户上传的图像进行检测,返回该图像中的人脸是否为真人。这些平台间的对比如表 1 所示。

其中,在百度、腾讯平台上集成有 liveness detection 和 spoofing detection 功能,而在其他平台上只有 liveness detection 功能。

表 1 商用活体检测平台对比
Table 1 Comparison of commercial liveness detection platforms

商用平台	输入格式	应用接口
阿里巴巴	图片	活体检测
百度	图片、视频	活体检测、伪造检测
科大讯飞	图片	活体检测
商汤科技	图片、视频	活体检测、伪造检测
腾讯	图片、视频	活体检测
云从科技	图片	活体检测
京东	图片	活体检测

本文使用 4 种较先进的 Deepfake 技术生成了大量的测试样本，并对测试样本进行图像压缩，模拟不同质量的图像，用这些样本来对科大讯飞、腾讯、百度、阿里巴巴、京东、云从科技、商汤科技这 7 种商用活体检测平台在线 API 端口进行测试。实验结果显示，Deepfake 攻击能够较好地通过商用活体检测平台上的 liveness detection，却很难通过 spoofing detection。因此百度、腾讯平台对于 Deepfake 图片或视频的检测成功率要比其他平台高。科大讯飞平台对真实图片的误检率很高，即容易把真实图片判断为假的，但是其对于 Deepfake 图像的检测成功率却很高，而在云从科技、京东平台上，对真实图像的误检率很低，但对于 Deepfake 图像的检测成功率同样很低，即不管是真实图像还是 Deepfake 图像都能够很好地绕过这些平台的检测。其主要原因可能是因为这些平台针对的是打印照片攻击、屏幕二次翻拍攻击、硅胶面具攻击等，并未将先进的换脸检测技术集成到他们的活体检测算法中。

因此，针对常用的商用活体检测平台对 Deepfake 图像的检测成功率较低的问题，本文提出了一种集成活体检测方法 Integrantet，该方法由 4 种针对不同的人脸图像特征的活体检测算法集成所得，分别是针对 Deepfake 过程产生的面部伪影进行检测的伪影检测算法^[14]，针对图像的介观属性进行检测的 Mesonet^[15]，针对纸张打印、屏幕翻拍、3D 模型等方式的欺骗行为进行检测的 HyperFAS 以及针对面部伪造过程中产生的伪造线索的 Spoof cues^[16]检测算法。本文在测试数据集上对集成活体检测方法 Integrantet 进行测试，发现 Integrantet 比各大商用活体检测平台对于

Deepfake 图像的检测成功率提高至少 35% 以上。

本文的主要贡献如下。

1) 对现有的常用商用活体检测平台进行鲁棒性评估，发现这些平台正面临着 Deepfake 技术带来的巨大威胁。

2) 利用目前较先进的 Deepfake 技术生成大量的测试样本，比较了不同 Deepfake 技术之间优劣。

3) 提出了一种活体检测集成方法 Integrantet，并在测试数据集上验证了该方法能够有效地检测出 Deepfake 攻击。

1 相关工作

1.1 Deepfake 生成方法

随着 Deepfake 技术的不断更新迭代，生成的图片或视频的质量也在持续提高。根据生成方式的不同，人脸 Deepfake 技术可以分为面部替换、面部重现、面部合成和面部编辑。面部替换就是将一个人的脸替换到另一个人的脸上，这是一种常见的 Deepfake 方式。在面部替换方面，根据生成方法的不同，又可以分为基于卷积神经网络 (CNN) 的方法和基于生成对抗网络 (GAN) 的方法。Korshunova^[17]等提出了一种基于卷积神经网络的面部替换方法 Faceswap，使用训练好的卷积神经网络从照片的非结构化集合中捕获目标身份的外观，通过在风格转换方面即以另一种风格渲染图像实现人脸交换，但这种方法人脸面部表情的变化处理得不够到位。为生成更加逼真的换脸视频，Nirkin^[18]等提出了一种面部交换和重现方法 FSGAN，这是一种基于生成对抗网络的方法，可以应用于多对人脸，而不需要在这些人脸上进行训练，但这种方法对于有面部遮挡的人脸图像生成的效果并不理想。因此 Li^[19]等提出了一个新的两阶段框架 FaceShift，用于高质量和遮挡感知的人脸交换。为了解决面部遮挡问题，FaceShift 在第一阶段也就是面部替换的基础上增加了第二阶段，包括一个新的启发式错误确认细化网络 (HEAR-Net, novel heuristic error acknowledging refinement network)。它被训练成以自监督的方式恢复遮挡区域，不需要任何人工注释。面部重现就是指修改一个人的面部表情，Siarohin^[20]等提出了一种基于 GAN 的视频驱动图片的面部重现方法 (FOMM, first order motion

model)。使用生成器网络对目标运动期间出现的遮挡进行建模,并组合从源图像提取的外观和从驱动视频导出的运动达到面部重现的目的。这种方法的效果依赖于驱动视频的运动幅度,一旦驱动视频的运动幅度过大,那么产生的视频的质量也不会很好。面部合成可以生成不存在的人脸图像,Karras^[21]等借鉴风格迁移的思想,改进了GAN中生成网络的结构,这种网络结构可以自动以无监督的方式学习高级特征(如人脸图像中的姿态、身份)和图像中的随机变化(如雀斑、头发),从而自动控制合成效果。这种新的生成器有效提升了图像生成质量,更好地解释了插值特性、解耦了变量的隐参数。面部编辑意味着可以添加、删除或更改面部的某些属性。这些属性可以是头发、年龄、衣服、种族、性别等。He^[22]等对生成的图像应用属性分类约束,而不是对潜在表示施加约束,以保证正确改变所需的属性。同时,引入重构学习来保留排除属性的细节,并将对抗性学习用于视觉上真实的编辑。属性分类约束、重构学习、对抗性学习这3个组件相互协作,形成了一个高质量面部属性编辑的有效框架,被称为AttGAN。此外,面部编辑的方法还有StarGAN^[23]和STGAN^[24]等。

1.2 Deepfake 检测方法

为了应对Deepfake技术对现实世界的危害,相关研究者提出了多种Deepfake检测技术,根据检测对象的不同可以分为基于图像检测和基于视频检测两大类。

图像检测方法主要可分为非深度神经网络方法和深度神经网络方法。在非深度神经网络方法中,Yang^[25]等发现Deepfake图像是通过将合成的人脸区域拼接到原始图像中而产生的,并且引入了从人脸图像估计3D头部姿态时可能会暴露的误差,基于这一特征训练一个SVM分类器来进行Deepfake图像的检测。在深度神经网络方法中,Feng^[16]等从异常检测的角度重新定义了人脸反欺骗,并提出了一个残差学习框架来学习区别性的实时欺骗差异,这些差异被定义为spoof cues,通过检测spoof cues来判断图片是否是伪造的。龚晓娟等^[26]针对Deepfake检测算法可解释性差等问题,提出融合双层注意力的神经网络模型,该模型利用通道注意力捕获假脸的异常特征,并结合空间注意力聚焦异常特征的位置,充分学习

假脸异常部分的上下文语义信息,从而提升换脸检测的有效性和准确性。并以热力图的形式有效地展示了真假脸的决策区域,使换脸检测模型具备一定限度的解释性。

相比于图片,视频中包含的信息更多,Li等^[27]提出了检测眨眼信息来有效判断真假,但由于现在生成技术的改进,该方法已经不能有效的检测。Li^[14]等通过观察发现Deepfake算法只能生成有限分辨率的图像,需要进一步扭曲以匹配源视频中的原始人脸。这种变换在生成的Deepfake视频中留下了独特的伪影,可以被卷积神经网络有效地捕获。因此,文献[14]将人脸仿射扭曲中的伪影作为区分真实和伪造图像的特征,训练一个CNN网络实现对Deepfake视频的检测。这种检测方法容易对某种Deepfake算法过拟合。Wang等^[28]结合多个视频帧面部的运动,通过残差空间梯度块(RSGB, residual spatial gradient block)和时空传播模块(STPM, spatio-temporal propagation module)准确地预测出视频的深度信息,并通过深度信息来判别视频的真假。

2 本文方法介绍

2.1 商用活体检测平台的鲁棒性评估方法

本文使用了几种较先进的Deepfake方法来生成大量的换脸视频和图片作为测试样本,并对测试样本进行压缩操作来降低其质量,最后将压缩前后的测试样本输入到商用活体检测平台上得到测试结果,其流程如图1所示。

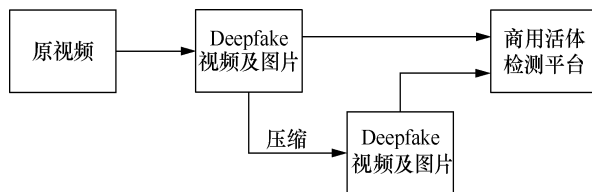


图1 商用活体检测平台鲁棒性评估流程
Figure 1 Robustness evaluation flowchart of commercial liveness detection platform

2.1.1 制作测试样本

制作测试样本是本文实验重要的一个环节,制作出的样本质量直接影响伪造检测环节的检测结果,为了获得好的伪造效果。本文使用了Faceswap、FSGAN、FaceShift、First Order Motion Model这4种Deepfake方法,再挑选检测通过率

高的人脸图像作为输入,并根据不同的方法对人脸图片进行预处理以提升效果。

使用 Deepfake 方法制作测试样本的流程如下:首先,对源人脸图像进行预处理,对人脸进行识别,剪切并对齐摆正人脸,提取人脸中关键点的信息;然后,通过 GAN 和自动编码器等进行模型的训练;最后,通过训练获得的模型来实现人脸的交换。

2.1.2 视频和图像压缩

为了测试商用平台对不同质量的 Deepfake 视频和图片的检测效果,本文对生成的测试样本进行压缩处理,降低测试样本的质量。本文采用转码压缩法压缩视频,该方法是通过改变码率来达到降低视频质量的目的。本文采用的图像压缩方法有格式转换法、基于 SVD 的图压缩方法和基于 k-means 聚类算法的图压缩方法。

2.1.3 攻击商用活体检测平台

为了研究常用的商用平台对 Deepfake 伪造人脸攻击的鲁棒性效果,需要用伪造的人脸去攻击不同的商用检测平台。阿里巴巴、腾讯、百度等企业都有自己的 AI 开放平台,调用相关的人脸检测的接口,按照提供的文档编写相应的代码,就能实现人脸检测的功能并返回检测的结果。

当然,这些商用平台的检测方法都是不一样的,所以不同平台对于同一张人脸的攻击可能会有不同的结果,同一平台对于不同方法制作的伪造人脸的攻击也会有不同的结果,需要对每一种情况进行验证,才能有足够的结果进行最后的分析。在本文实验中,每种方法进行检测的伪造图片数量为 1 000 张。

2.2 集成活体检测方法

本文提供了一种集成活体检测方法 Integrant,该方法由伪影检测算法、Mesonet、HyperFAS、Spoof cues 检测算法这 4 种针对不同的人脸图像特征的活体检测算法集成所得。伪影检测算法基于 Deepfake 算法只能生成有限分辨率的图像,这些图像需要进一步的扭曲来匹配源视频中的原始面孔。这种变换会在得到的 Deepfake 视频中留下独特的伪影^[14],可以通过训练一个卷积神经网络来捕捉这些伪影。Mesonet 采用了两种层数都较低的网络,关注图像的介观特

性,不会因为视频或图像压缩过,质量下降从而检测准确率下降^[15]。与现有的商用活体检测平台一样,本文方法也集成了一种针对纸张打印、屏幕翻拍等攻击方法进行检测的 HyperFAS。Spoof cues^[16]检测算法使用了一个残差学习框架来学习区别性的实时欺骗差异,通过检测 Spoof cues 来判断图片是否为伪造。Integrant 将上述的 4 种算法进行了集成,通过 4 种检测算法从不同的角度对 Deepfake 攻击进行检测,共同对视频或图片的真假做决策,活体检测分数值的计算公式如式(1)所示。

$$F = \frac{1}{n} \sum_{i=1}^n S_i \quad (1)$$

其中, F 为最终检测分数值,为每种方法的检测分数值, n 为方法数,本文实验中设为 4, S_i 表示每种方法对该样本的检测分数。

在测试数据集上对所提出的集成活体检测方法 Integrant 进行测试,发现 Integrant 要比各大商用活体检测平台对于 Deepfake 图像的检测成功率至少提高 35%以上。

3 实验

3.1 实验设置

3.1.1 实验环境

实验操作系统主要为 Ubuntu 16.04.6,集成开发环境是 python3.6.0,采用 Python 中 TensorFlow 等深度学习框架。硬件条件:CPU 为 Intel(R) Xeon(R) Gold 5218 CPU @ 2.30 GHz, GPU 为 NVIDIA Tesla V100 SXM2 32 GB。

3.1.2 实验数据集

本文使用中国科学院人脸防伪检测数据集(CASIA-FASD)中的真实视频作为源视频进行换脸。CASIA-FASD 是中国科学院制作的用来做活体检测任务的数据集,分为训练集和测试集,训练集共 20 人,测试集共 30 人,每个人包含 12 个视频,其中有 3 段是真实视频,分为两段不同分辨率的和一段高质量的视频,其他的都是假视频,包括照片攻击、视频攻击等攻击方法。本文的实验只用到分辨率为 640×480 的真实视频。

FaceForensics++^[29]是 FaceForensic 的扩展版本,原视频都来自 Youtube,数据集包含使用 Face2Face、FaceSwap、DeepFakes 和 Neural

Textures 这 4 种 Deepfake 方法生成的两种不同质量的换脸视频。本文只用到该数据集中的原视频作为 FOMM 方法的驱动视频。

3.1.3 商用活体检测平台

本文的测试对象为商用活体检测平台,用到的商用活体检测平台有科大讯飞、腾讯、百度、腾讯、阿里巴巴、京东和云从科技,使用他们提供的在线检测 API 接口,上传图像进行测试。

3.1.4 Deepfake 方法

本文为评估商用活体检测平台的鲁棒性,采用 Deepfake 方法生成大量的测试样本对商用活体检测平台进行测试,用到的 Deepfake 方法有 Faceswap、FSGAN、FaceShift、FOMM。

3.1.5 评价指标

本文用到的评价指标为商用活体检测平台对上传图像的通过率 T ,其计算式如下:

$$T = R / N \quad (2)$$

其中, N 是上传的总图像数, R 是平台判断为真实图像的个数。对于真实图像, T 越高表示平台检测性能越好;但对于 Deepfake 图像, T 越低则表示平台检测性能越好。

3.2 不同 Deepfake 方法生成结果对比

本文主要比较了 4 种不同的 Deepfake 方法生成的测试样本的效果。首先,使用 CASIA-FASD 数据集中的 46 段真实视频作为源视频,使用 Faceswap、FSGAN、FaceShift、FOMM 这 4 种 Deepfake 方法生成 1 000 段 Deepfake 视频;再从每段视频中取一帧图片,共 1 000 张图片,一起作为测试数据集。其中使用 FOMM 方法时,为了达到好的效果,需要对输入图像进行预处理,本文实验采用 Python 中的 Dlib 库对原图像中人脸进行识别并裁剪成 256×256 像素,将处理后的人脸作为该方法的输入。

图 2 显示了人脸交换方法的生成效果。可以看到,FaceSwap 生成的效果很好,极大地保留了源人脸的人脸特征,FaceShifter 则极大地保留了目标人脸的信息,因此看起来与源人脸有一定差距,FSGAN 中生成的效果最好,因为它极大地保留了源人脸和目标人脸的信息。对不同伪造方法进行分析发现,眼镜、厚重刘海、浓密胡须等对生成结果的效果有一定的负面影响。

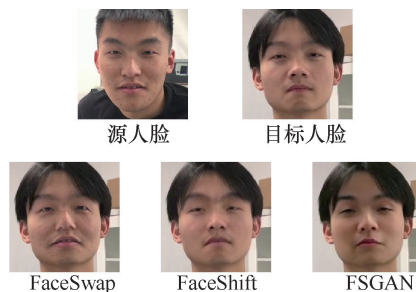


图 2 人脸交换效果
Figure 2 Face swap rendering

图 3 显示了 FOMM 方法的生成效果,上面一行显示的是驱动视频的帧,下面一行显示对应的生成效果。从图 3 可知,对于预处理过的输入数据,生成效果较好,并且当人脸动作幅度较小时,靠肉眼基本无法判断视频真伪,但当人脸动作幅度较大时,会产生明显的扭曲。



图 3 FOMM 生成效果
Figure3 FOMM generates renderings

3.3 不同商业平台活体检测的鲁棒性评估

本节主要介绍了不同商业平台活体检测平台对于 Deepfake 伪造人脸攻击的鲁棒性评估。将 3.2 节中生成的 Deepfake 图像和原图像作为输入,在科大讯飞、腾讯、百度、腾讯、阿里巴巴、京东、云从科技的活体检测 API 接口上进行测试,得到的测试结果如图 4 所示。

同一方法在不同的平台,不同方法在同一平台的结果都不一样。从 Deepfake 攻击的角度来看,面部交换方法中 FaceSwap 和 FaceShift 除在科大讯飞平台上的通过率较低,在其他平台上都有较高的通过率。FSGAN 除在百度和腾讯平台上的通过率较低,在其他平台上也都有较高的通过率。而作为面部重现方法的 FOMM 则所有平台上都有着很高的通过率,分析其原因可能是因为

FOMM 采用的是视频驱动图片的方法, 这将极大地保留住原图像的面部生物特征, 有效地绕过检测。从科技平台的角度来看, 京东和云从科技平台对所有的 Deepfake 方法都有着很高的通过率, 也即它们对 Deepfake 攻击的检测效果很差。而科大讯飞平台看似对 FaceSwap 和 FaceShift 有着相对较好的检测率, 但其对真实图像却有着很大的误检率, 即容易将真实图像判断为假图像。总的来说, 常用的各大商用活体检测平台对 Deepfake 攻击的检测成功率普遍很低。

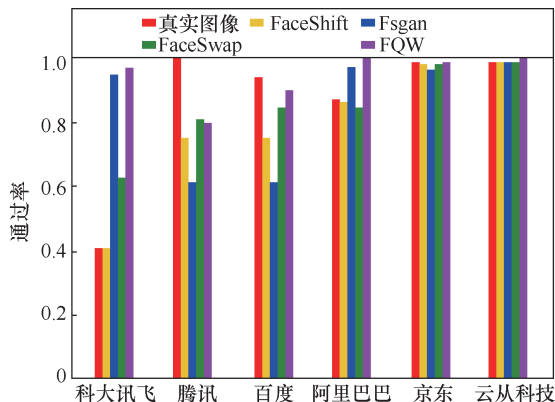


图 4 商用活体检测平台 Deepfake 图像检测结果

Figure 4 Deepfake image detection results of each commercial liveness detection platform

图 5 显示了百度平台上 liveness detection 与 spoofing detection 结果的对比, 可以看出 spoofing detection 对于 Deepfake 图像有着比 liveness detection 更好的检测效果, 可见 spoofing detection 要更加针对 Deepfake 攻击, 这一检测功能除了百度和腾讯平台外, 其他的平台目前都没有。

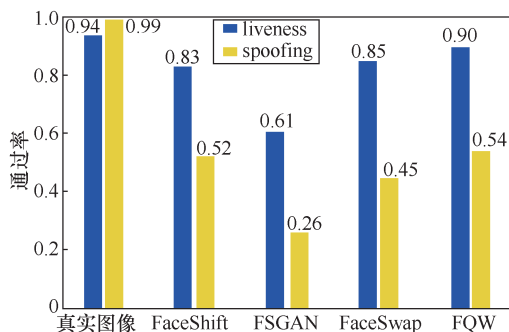


图 5 百度平台上 liveness detection 与 spoofing detection 结果对比

Figure 5 Comparison chart of liveness detection and spoofing detection results on Baidu platform

除了对 Deepfake 图像的检测, 本文还测试了

百度、腾讯、商汤科技平台对 Deepfake 视频的检测通过率, 用到的 Deepfake 方法有 FSGAN、FaceShift 和 FOMM 这 3 种, 测试结果如图 6 所示。可以看出, Deepfake 视频在活体检测平台上的测试结果同图像类似, 对于真实视频, 活体检测的通过率要比图像更高, 而同样作为面部重建方法的 FOMM 生成的视频在所有测试平台上都有着较高的通过率。

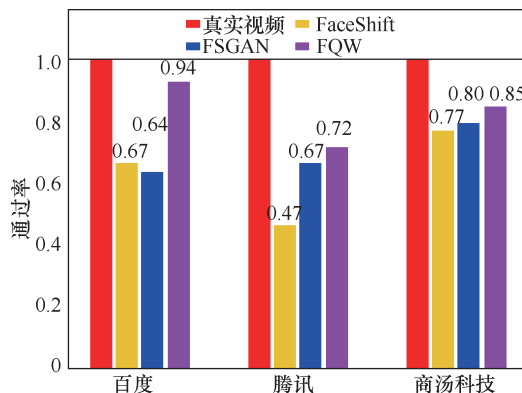


图 6 视频活体检测结果

Figure 6 Video liveness detection result chart

3.4 不同质量的 Deepfake 图像的检测效果评估

为了测试商用平台对不同质量的 Deepfake 视频和图片的检测效果, 本文对生成的测试样本进行压缩处理再上传到科大讯飞、腾讯、百度、腾讯、阿里巴巴、京东、云从科技的活体检测 API 接口上进行测试, 本文采用的压缩方法为基于 SVD 的图压缩方法, 得到的测试结果如图 7 所示。

从图 7 可以看出, 图像压缩对科大讯飞和腾讯平台的测试结果有着较大的影响, 而且效果相反, 在科大讯飞平台上, 图像压缩后通过率会有提升, 而在腾讯平台上则会下降, 这说明以上两个平台对输入图像的质量较为敏感, 也即它们的鲁棒性较差。

表 2 显示了 3 种不同压缩方法在科大讯飞平台上的活体检测通过率对比, 本文用到的 Deepfake 方法是 FaceShift。可以看到基于 SVD 的图压缩方法和基于格式转换的图压缩方法都能提高 Deepfake 图像的通过率, 而基于 k-means 聚类的图压缩方法则降低了 Deepfake 图像的通过率, 进一步说明了科大讯飞平台对输入图像的质量非常敏感。

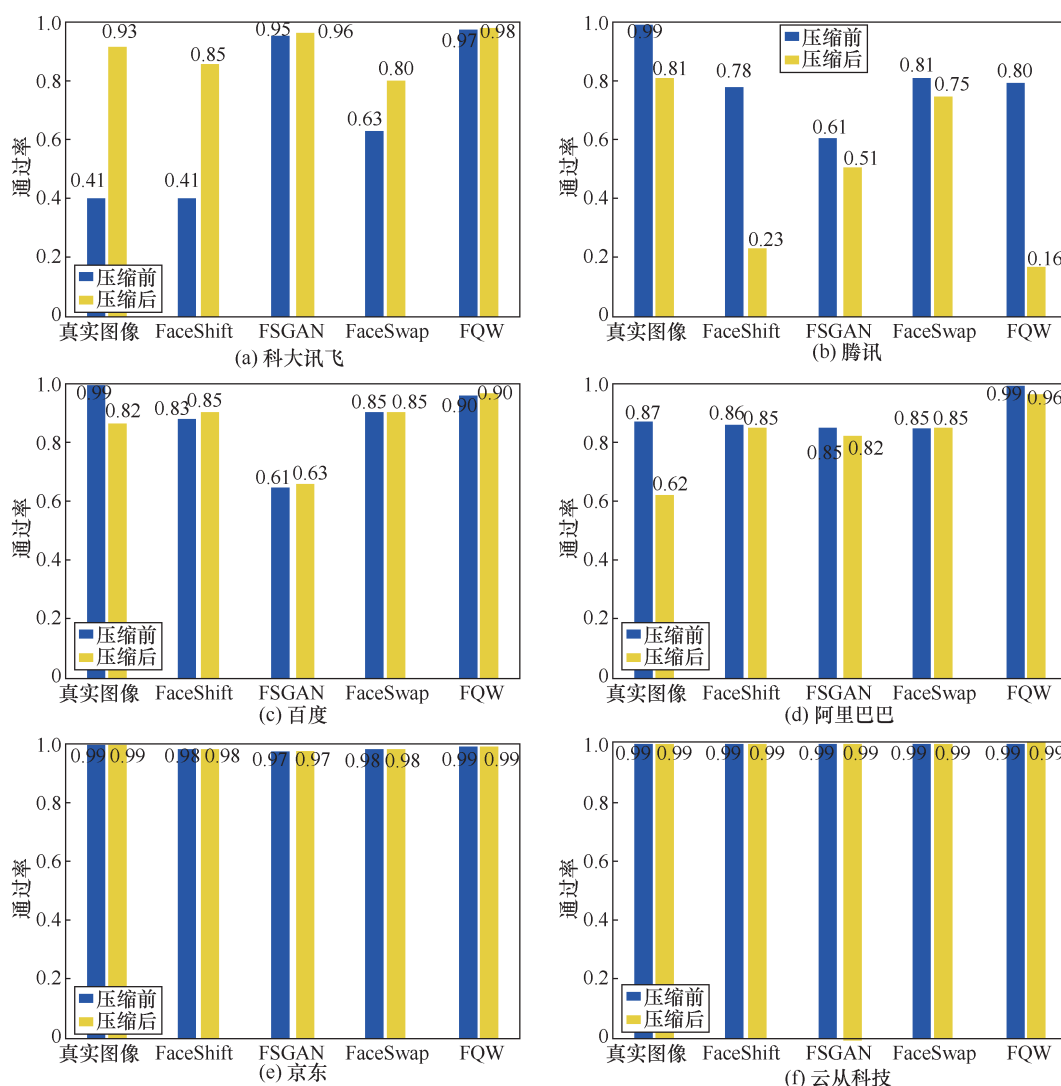


图 7 图像压缩前后检测结果对比

Figure 7 Comparison of test results before and after image compression

表 2 3 种不同压缩方法在科大讯飞平台上的活体检测通过率对比

Table 2 Comparison of in vivo detection pass rates of three different compression methods on the KedaXunfei platform

压缩方法	压缩前通过率	压缩后通过率
SVD	41%	85%
k-means 聚类	41%	21%
格式转换	41%	66%

3.5 集成活体检测方法检测结果分析

本节主要测试了集成活体检测方法 Integrantet 对于 Deepfake 图像的检测效果, 结果如表 3 所示。

从表 3 可以看出 Integrantet 对于 Deepfake 图

像的检测通过率要比常用的各大商用活体检测平台低很多, 这说明 Integrantet 集成活体检测方法对于 Deepfake 图像的检测效果要更好, 这是因为 Integrantet 与其他只关注单一特征点的方法不同, 它集成了 3 种不同的 Deepfake 检测方法, 及一种针对纸张打印、屏幕翻拍、3D 模型等攻击手段的活体检测方法。3 种 Deepfake 方法分别针对的是图像的介观属性以及 Deepfake 过程中产生的面部伪影和伪造线索进行检测。这 4 种方法的集成使得该方法能够有效检测出各种人脸伪造攻击。

4 结束语

本文对常见的几种商用活体检测平台进行了鲁棒性评估, 发现了目前的商用平台并不能很好地

表 3 Integrantet 检测通过率对比
Table 3 Integrantet test pass rate comparison

平台	Integrantet	科大讯飞	腾讯	百度	阿里巴巴	京东	云从科技
FaceSwap	2.12%	41.23%	78.26%	83.42%	86.26%	98.19%	99.34%
FaceShift	10.34%	95.24%	61.41%	61.19%	85.39%	97.38%	99.42%
FSGAN	9.68%	63.14%	81.26%	85.27%	85.24%	98.17%	99.29%
FOMM	7.41%	97.29%	80.31%	90.18%	99.17%	99.42%	99.14%

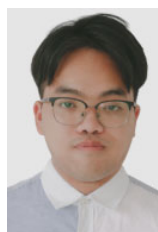
抵御 Deepfake 的攻击,为此集成了一种能够从多角度检测伪造图像的活体检测方法 Integrantet,并验证了其针对 Deepfake 图像有非常好的检测效果。但是这种方法也存在着复杂度较高的问题,因为它是由 4 种活体检测方法集成所得。希望通过本文的研究工作能有更好的针对 Deepfake 检测方法可以集成到这些商用活体检测平台上。

参考文献:

- [1] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.
- [2] KRISHNAMURTHY B, SARKAR M. Deep-learning network architecture for object detection: U.S. Patent 10,152,655[P]. 2018-12-11.
- [3] DENG L, LIU Y. A joint introduction to natural language pro-cessing and to deep learning[C]//Proceedings of Deep Learning in Natural Language Processing. 2018: 1-22.
- [4] COLLOBERT R, WESTON J, BOTTOU L, et al. Naturallang-uage processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(8): 2493-2537.
- [5] XIONG W, DROPO J, HUANG X, et al. Achieving human parity in conversational speech recognition[J]. arXiv Preprint Arxiv: 1610.05256, 2016.
- [6] MOUSAVI A, BARANIUK R G. Learning to invert: signal recovery via deep convolutional networks[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. 2017: 2272-2276.
- [7] 王宇龙,刘开元.基于面部特征点运动的活体识别方法[J]. 网络与信息安全学报,2018,4(6):36-44.
WANG Y L, LIU K Y. Liveness recognition method based on facial feature point motion[J]. Chinese Journal of Network and Information Security, 2018,4(6):36-44
- [8] 马玉琨,徐姚文,赵欣,等.人脸识别系统的活体检测综述[J]. 计算机科学与探索, 2021, 15(7): 1195-1206.
MA Y K, XU Y W, ZHAO X, et al. Review of presentation attack detection in face recognition system[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(7): 1195-1206.
- [9] Deepfakes[EB].2019.
- [10] 谢永江,姜淑丽.我国网络立法现状与问题分析[J]. 网络与信息安全学报, 2015, 1(1): 24-30.
XIE Y J, JIANG S L. Analysis of the situation and problem on the legislation of cyberspace in China[J]. Chinese Journal of Network and Information Security, 2015, 1(1): 24-30
- [11] 乔通,姚宏伟,潘彬民,等.基于深度学习的数字图像取证技术研究进展[J]. 网络与信息安全学报, 2021, 7(5): 13-28.
QIAO T, YAO H W, PAN B M, et al. Current study of digital image forensic techniques based on deep learning[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 13-28.
- [12] 黄惠芬,王志红,常玉红.数学形态学在数字伪造模糊检测中的应用[J]. 网络与信息安全学报, 2017, 3(4): 20-25.
HUANG H F, WANG Z H, CHANG Y H. Blur detection of digital forgery using mathematical morphology[J]. Chinese Journal of Network and Information Security, 2017, 3(4): 20-25.
- [13] CHEN H S, ROUHSEDAGHAT M, GHANI H, et al. DefakeHop: a light-weight high-performance deepfake detector[J]. arXiv: 2103.06929, 2021.
- [14] LI Y, LYU S. Exposing DeepFake videos by detecting face- warping artifacts[C]//Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2018.
- [15] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. Piscataway: IEEE Press, 2018: 1-7.
- [16] FENG H, HONG Z, YUE H, et al. Learning generalized spoofcues for face anti-spoofing[J]. arXiv:2005.03922, 2020.
- [17] KORSHUNOVA I, SHI W Z, DAMBRE J, et al. Fast face-swap using convolutional neural networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. 2017: 3697-3705.
- [18] NIRKIN Y, KELLER Y, HASSNER T. FSGAN: subject agnostic face swapping and reenactment[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 7183-7192.
- [19] LI L,BAO J, YANG H, et al. FaceShifter: towards high fidelity and occlusion aware face swapping[J]. arXiv:1912.13457, 2019.
- [20] SIAROHIN A, S LATHUILIÈRE, TULYAKOV S, et al. first order motion model for image animation[J]. arXiv: 2003.00196, 2020.
- [21] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of 2019

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 4396-4405.
- [22] HE Z L, ZUO W M, KAN M N, et al. AttGAN: facial attribute editing by only changing what you want[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5464-5478.
- [23] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8789-8797.
- [24] LIU M, DING Y K, XIA M, et al. STGAN: a unified selective transfer network for arbitrary image attribute editing[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3668-3677.
- [25] YANG X, LI Y Z, LYU S W. Exposing deep fakes using inconsistent head poses[C]//Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. 2019: 8261-8265.
- [26] 龚晓娟, 黄添强, 翁彬, 等. 基于双层注意力的 Deepfake 换脸检测[J]. 网络与信息安全学报, 2021, 7(2): 151-160.
GONG X J, HUANG T Q, WENG B, et al. Deepfake swapped face detection based on double attention[J]. Chinese Journal of Network and Information Security, 2021, 7(2): 151-160.
- [27] LI Y Z, CHANG M C, LYU S W. In ICTU OCULI: exposing AI created fake videos by detecting eye blinking[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. Piscataway: IEEE Press, 2018: 1-7.
- [28] WANG Z Z, YU Z T, ZHAO C X, et al. Deep spatial gradient and temporal depth learning for face anti-spoofing[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 5041-5050.
- [29] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: learning to detect manipulated facial images[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 1-11.

[作者简介]



王鹏程(1997-), 男, 湖南衡阳人, 浙江工业大学硕士生, 主要研究方向为深度学习和人工智能安全。



郑海斌(1995-), 男, 浙江台州人, 浙江工业大学博士生, 主要研究方向为深度学习、安全可信人工智能。



邹健飞(1995-), 女, 江西赣州人, 浙江工业大学硕士生, 主要研究方向为深度学习和人工智能安全。



庞玲(1979-), 女, 河北石家庄人, 硕士, 信息系统安全技术重点实验室高级工程师, 主要研究方向为网络安全。



李虎(1987-), 男, 甘肃定西人, 博士, 信息系统安全技术重点实验室高级工程师, 主要研究方向为人工智能安全、机器学习等。



陈晋音(1982-), 女, 浙江象山人, 浙江工业大学教授, 主要研究方向为人工智能安全、图数据挖掘和进化计算。

《网络与信息安全学报》

第二届编辑委员会

主 任 委 员：方滨兴

常务副主任委员：李风华

副 主 任 委 员：李 晖 俞能海 Kui Ren[USA] 刘建伟

邹德清 杜学绘 刘华鲁

委 员：（按姓氏笔画排序）

丁 勇 马建峰 王 标 王 滨 王布宏

王连海 王良民 王国胤 王明生 王美琴

牛 犇 仇洪冰 田有亮 田俊峰 史国振

付章杰 冯 涛 朱建明 延志伟 仲 盛

向 涛 刘 哲 刘 渊 刘西蒙 刘哲理

闫 峥 孙东红 杨宏宇 肖 甫 肖 亮

吴 铤 吴晓平 邱卫东 何道敬 沈 剑

沈 超 张小松 张云勇 张红旗 张宏莉

张建标 张振江 陆月明 陈立全 陈兴蜀

陈性元 陈雷霆 苗春雨 林 鹏 易东山

赵 波 赵琉涛 咸鹤群 禹 勇 姚志强

秦 拯 贾 焰 殷丽华 翁 健 黄欣沂

曹春杰 彭国军 蒋文保 韩伟力 韩劲松

程 光 阚海斌

David Naccache [FRA] Gang Qu [USA] Jinjun Chen [AUS]

Willy Susilo [AUS] Yier Jin [USA]

网络 信息安全学报

Chinese Journal of
Network and Information Security



2022年2月15日出版 定价：60.00元

