

人脸伪造与检测中的对抗攻防综述

黄诗瑀¹, 叶锋^{1,2}, 黄添强^{1,2}, 李伟¹, 黄丽清^{1,2}, 罗海峰^{1,2}

(1. 福建师范大学计算机与网络空间安全学院, 福建 福州 350117; 2. 数字福建大数据安全技术研究所, 福建 福州 350117)

摘要: 人脸伪造和检测是当前的研究热点。通过人脸伪造方法可以制作虚假人脸图像和视频, 一些出于恶意目的而将名人虚假视频在社交网络上广泛传播, 不仅侵犯了受害者的声誉, 而且造成了不良的社会影响, 因此需要开发对应的检测方法用于鉴别虚假视频。近年来, 深度学习技术的发展与应用降低了人脸伪造与检测的难度。基于深度学习的人脸伪造方法能生成看起来更加真实的人脸, 而基于深度学习的虚假人脸检测方法比传统方法具有更高的准确度。大量研究表明, 深度学习模型容易受到对抗样本的影响而导致性能下降。近来在人脸伪造与检测的领域中, 出现了一些利用对抗样本进行博弈的工作。原先的博弈模式变得更加复杂, 伪造方和检测方在原先方法的基础上, 都需要更多考虑对抗安全性。将深度学习方法和对抗样本相结合, 是该研究领域未来的趋势。专注于人脸伪造与检测中的对抗攻防这一领域进行综述。介绍人脸伪造与检测的概念以及目前主流的方法; 回顾经典的对抗攻击和防御方法。阐述对抗攻击和防御方法在人脸伪造和检测上的应用, 分析目前的研究趋势; 总结对抗攻防对人脸伪造和检测带来的挑战, 并讨论未来发展方向。

关键词: 深度伪造; 虚假人脸检测; 对抗样本; 社交媒体取证

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023049

Survey on adversarial attacks and defense of face forgery and detection

HUANG Shiyu¹, YE Feng^{1,2}, HUANG Tianqiang^{1,2}, LI Wei¹, HUANG Liqing^{1,2}, LUO Haifeng^{1,2}

1. College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China

2. Digital Fujian Institute of Big Data Security Technology, Fuzhou 350117, China

Abstract: Face forgery and detection has become a research hotspot. Face forgery methods can produce fake face images and videos. Some malicious videos, often targeting celebrities, are widely circulated on social networks, damaging the reputation of victims and causing significant social harm. As a result, it is crucial to develop effective detection methods to identify fake videos. In recent years, deep learning technology has made the task of face

收稿日期: 2023-01-16; 修回日期: 2023-05-30

通信作者: 叶锋, yefeng@fjnu.edu.cn

基金项目: 国家自然科学基金(62072106); 福建省自然科学基金(2020J01168)

Foundation Items: The National Natural Science Foundation of China (62072106), The Natural Science Foundation of Fujian Province (2020J01168)

引用格式: 黄诗瑀, 叶锋, 黄添强, 等. 人脸伪造与检测中的对抗攻防综述[J]. 网络与信息安全学报, 2023, 9(4): 1-15.

Citation Format: HUANG S Y, YE F, HUANG T Q, et al. Survey on adversarial attacks and defense of face forgery and detection[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 1-15.

forgery and detection more accessible. Deep learning-based face forgery methods can generate highly realistic faces, while deep learning-based fake face detection methods demonstrate higher accuracy compared to traditional approaches. However, it has been shown that deep learning models are vulnerable to adversarial examples, which can lead to a degradation in performance. Consequently, games involving adversarial examples have emerged in the field of face forgery and detection, adding complexity to the original task. Both fakers and detectors now need to consider the adversarial security aspect of their methods. The combination of deep learning methods and adversarial examples is thus the future trend in this research field, particularly with a focus on adversarial attack and defense in face forgery and detection. The concept of face forgery and detection and the current mainstream methods were introduced. Classic adversarial attack and defense methods were reviewed. The application of adversarial attack and defense methods in face forgery and detection was described, and the current research trends were analyzed. Moreover, the challenges of adversarial attack and defense for face forgery and detection were summarized, and future development directions were discussed.

Keywords: deepfake, fake face detection, adversarial example, social media forensics

0 引言

人脸伪造是指使用深度学习技术,如自动编码器^[1] (auto coder) 和生成对抗网络^[2] (GAN, generative adversarial network),制作虚假人脸图像和视频。人脸视频伪造本质上属于人脸图像伪造,其做法是在视频中抽帧,逐帧对人脸进行伪造操作,再合并成视频。人脸伪造方法可分为4类:面部重现(face reenactment)、面部交换(face swap)、面部特征编辑(face attribute edit)和面部生成(face generate)。其中,面部重现是让目标人脸做出和源人脸(驱动人脸)一样的表情,常用工具包括Face2Face^[3]和NeuralTextures^[4];面部交换是将目标人脸粘贴至源人物的面部区域,现有的工具包括FaceSwap^[5]、Deepfakes^[6]和Faceshifter^[7];面部特征编辑是对目标人脸上的一个指定特征(如发色、年龄)做修改,常用的工具主要为条件控制的生成对抗网络(cGAN, conditional GAN)^[8]为基础的系列网络,如StarGAN^[9]、AttGAN (attribute GAN)^[10]和STGAN (selective transfer GAN)^[11];而面部生成包括伪造一张现实中不存在的人脸、对两张人脸进行特征融合以及进行人脸风格转换等,常见工具为StyleGAN系列^[12-14]网络。

上述人脸伪造方法在电影娱乐产业中得到广泛应用。然而它们被不法分子利用并制作出许多恶性的人脸伪造图片和视频。这些伪造信息通过社交媒体快速传播,引起了大众的信任危机。为了维护社会稳定,需要开发相应的虚

假人脸检测方法。现有的基于深度学习的虚假人脸检测(简称深伪检测)方法是先使用VGG^[15]、ResNet^[16]、Xception^[17]、MobileNet^[18]、EfficientNet^[19]、Vision Transformer^[20]等网络提取分析伪造特征,再使用分类算法进行真假分类。而虚假人脸视频的检测本质上属于虚假人脸图像检测,其从待检测视频中抽取部分帧,进行帧间和帧内特征提取,再进行真伪二分类。

虚假人脸越来越逼真,检测手段越来越先进,人脸伪造和检测之间形成博弈,二者在博弈中不断改进。虽然深度学习在人脸伪造和检测中都获得了比传统方法更好的结果,但Szegedy等^[21]发现深度学习模型存在缺陷,向其测试数据中加入微小对抗噪声,便可形成对抗样本,进而误导模型。许多人工智能应用,如人脸识别和语音识别,容易受到对抗样本的干扰,人脸伪造和检测亦是如此。伪造方通过向虚假人脸中加入不可见的对抗噪声以干扰检测器(如图1所示),或者利用经过对抗性训练的GAN,直接生成不带有伪影的虚假人脸图像,误导检测器将虚假人脸分类为真实的人脸(如图2所示)。而对于反伪造方,与图3中使用检测器进行取证的“人脸伪造后的被动取证”不同,他们事先在干净图像中加入肉眼不可见的对抗噪声,使得图像经过GAN等方法伪造后产生严重失真,实现从根源上抵御人脸伪造,这种策略也称为“人脸伪造前的主动干预”,如图4所示。因此,对抗样本的存在,改变了以往人脸伪造和检测的博弈模式。

近年来,国内外关于对抗样本^[22-23]和虚假人脸检测^[24-26]的综述,对这两个领域的内容已经做出较为完备的阐述与总结。本文在前人工作的基础上,进一步补充阐述了对抗攻防方法在人脸伪造和检测中的应用。

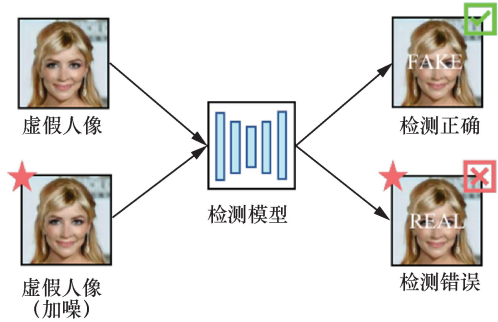


图1 针对假脸检测模型的噪声攻击

Figure 1 Noise attacks against fake face detection models

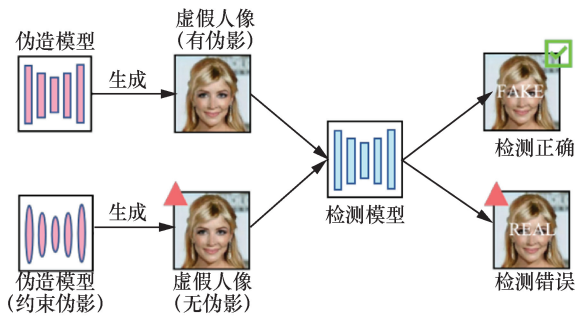


图2 针对假脸检测模型的去伪影攻击

Figure 2 Artifact-removing attacks against face detection models

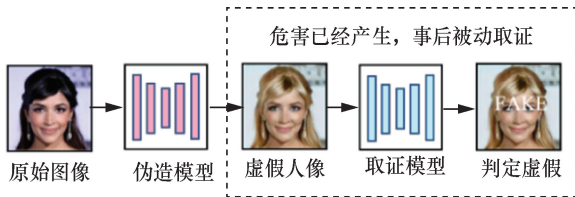


图3 人脸伪造后的被动取证

Figure 3 Passive forensics after face forgery

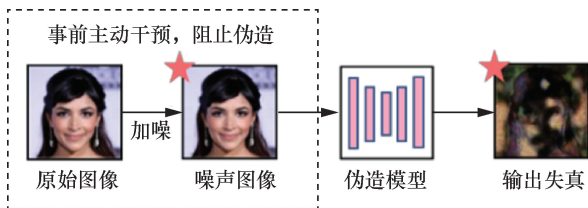


图4 人脸伪造前的主动干预

Figure 4 Active interference before face forgery

1 对抗攻击和防御

1.1 对抗攻击

通常情况下,对抗攻击算法首先在图像中加

入肉眼不可见的随机噪声,随后计算目标函数并求解梯度,最后利用梯度更新噪声分布,得到优化完毕的对抗样本。根据不同的攻击意图可将对抗攻击分为无目标攻击(untargeted attack)^[27]和有目标攻击(targeted attack)^[28]。无目标攻击根据被攻击模型的输出和真实类标签的距离,构建目标函数求得梯度,并在对抗噪声优化阶段进行梯度上升,让模型输出远离正确类别。而有目标攻击根据模型输出和目标类标签的距离,构建目标函数求得梯度,并在对抗噪声优化阶段进行梯度下降,让模型输出指定错误类别。相比有目标攻击,无目标攻击实现简单且成功率更高。

根据对抗样本训练阶段是否允许攻击者访问受害模型的信息,可将对抗攻击分为白盒攻击(white box attack)^[27]和黑盒攻击(black box attack)^[29]。由于攻击者能够获取模型的参数和结构信息,求得模型梯度并用于优化对抗噪声,因此白盒攻击能够达到较高的成功率。而在黑盒攻击中,模型对攻击者屏蔽了结构和参数信息,攻击者无法直接计算梯度用于优化对抗噪声,相比白盒攻击,黑盒攻击难度更大,且成功率更低。

白盒攻击最早可追溯至 Goodfellow 等^[27]提出的快速梯度符号法(FGSM, fast gradient sign method)。为了制造更强的对抗样本, Kurakin 等^[28]对输入数据进行多次迭代优化,提出 FGSM 的迭代版本——I-FGSM (iterative-FGSM)。基于迭代思想,后续出现了其他经典的白盒攻击算法,如基于决策边界的 Deepfool 算法^[30]、基于显著图的攻击方法(JSMA, jacobian-based saliency map attacks)^[31]、投影梯度下降(PGD, projected gradient descent)^[32]、可跨图像迁移的通用扰动(UAP, universal adversarial perturbation)攻击^[33]以及基于模型输出分数的(C&W, carlini&wagner)攻击^[34]。

在黑盒攻击中,攻击者无法直接求得梯度优化对抗噪声,需要使用两种梯度近似策略进行攻击:替代模型法和查询法。关于替代模型法,首先寻找和目标模型结构相似的替代模型,随后在替代模型上根据白盒攻击的做法,优化对抗噪声,最后将该对抗样本用于攻击黑盒模型。Goodfellow 等^[27]发现对抗样本能在不同模型之间

迁移,随后 PaperNot 等^[29]在现实场景下实施替代模型黑盒攻击。Liu 等^[35]发现迁移攻击效果很大程度上取决于目标模型和替代模型决策边界的相似性。实际场景下二者难免存在差异,现有的方法从 3 个角度解决上述问题:① 优化 I-FGSM^[36-37];② 增强数据分布^[38-40];③ 改进替代模型结构^[41]。而在查询法中,根据不同算法估计梯度方向,将估计得到的梯度用于优化对抗噪声。其中经典工作包括随机无梯度(RGF, random gradient-free)优化算法^[42]、自然进化策略(NES, natural evolution strategies)^[43]、Bandit 查询算法^[44]和简易黑盒攻击(SimBA, simple black-box attack)^[45],后来有关查询法的工作主要围绕减少查询次数和提高估计准确性^[46-47]。

1.2 对抗防御

对抗防御旨在消除对抗样本的干扰,使深度神经网络(DNN, deep neural network)模型恢复正常输出。现有的对抗方法分为 4 种:模型集成、梯度蒸馏、对抗训练和图像去噪。其中,模型集成^[48]会被各种可黑盒迁移的对抗样本攻破^[49],梯度蒸馏^[50]已经被 C&W 攻击^[34]证明无效。目前用于深伪领域的主流方法是对抗训练和图像去噪。

对抗训练最初由 Goodfellow 等^[27]提出,以 Mardy 等^[32]为代表,其做法是在模型训练阶段,引入干净的数据和带有对抗样本的数据,使得模型训练过程中学习对抗样本的分布模式,以至于在分类对抗样本时具有鲁棒性。对抗训练近年来的优化工作包括对抗训练加速^[51-52]和提升模型泛化性^[53-54]。而图像去噪防御是在对抗样本输入目标模型前用图像处理方法去除对抗噪声。事实表明,降低或者提高图像分辨率均可破坏对抗噪声,如 Dziugaite 等^[55]使用图像压缩进行对抗防御,以及 Mustafa 等^[56]提出图像超分辨率防御方法。此外,还有一些工作使用图像重建方法破坏对抗噪声实现对抗防御^[57-60]。

2 人脸伪造中的对抗攻击和防御

本节从攻击伪造模型和防御伪造模型两个角度,对现有工作展开叙述。攻击伪造模型是指在干净人脸图像中添加对抗噪声,导致伪造模型输出的虚假人脸图像失真,或者和原图相似(也称

图像锁定);防御伪造模型是指利用对抗防御方法,减轻对抗样本对伪造模型的影响,使得伪造模型重新输出视觉逼真的虚假人脸图像。

2.1 攻击伪造模型

本节根据 4 种常见的人脸伪造方法,利用对抗样本攻击伪造模型的工作。在介绍具体技术的同时,重点阐述这些工作的攻击特点,包括黑盒迁移性、对抗鲁棒性、数据迁移性和条件迁移性。关于攻击面部特征编辑的工作,本文还讨论其条件迁移性。接下来解释上述攻击特点。① 黑盒迁移性:对抗样本能够迁移至未知的黑盒伪造模型。② 对抗鲁棒性:对抗噪声不会被图像处理算法破坏,经过图像处理后依然保持攻击性。③ 数据迁移性:一次生成的对抗样本可以使多张图像上的人脸伪造产生失真。④ 条件迁移性:一张图像中的对抗样本可以攻击针对五官(眼睛、鼻子、嘴)的特征编辑。

2.1.1 攻击面部重现

面部重现是一种基于数据的伪造方法,经常用于伪造政治人物的演讲视频。这类方法需要根据驱动人脸和目标人脸来训练模型,因此在训练过程中,伪造模型会学习噪声分布模式,从而使模型具备一定对抗鲁棒性,导致攻击产生的输出失真不明显,经常需要借助检测器才能辨别真伪。现有关于攻击面部重现的工作较少。Yang 等^[61]利用 PGD 算法^[32]生成对抗样本,攻击基于 GAN 的面部重现方法 Mantra-Net^[62]。该工作考虑了对抗鲁棒性,但是没有考虑黑盒迁移性。Huang 等^[63]对此提出改进,利用替代模型进行黑盒迁移,并使用噪声生成器代替梯度优化算法,此外引入了一些辅助损失项,使得噪声攻击时更具有针对性,但是这项工作没有考虑对抗鲁棒性。不同于前者在 RGB 域攻击,Wang 等^[64]在 Lab 颜色空间的 a、b 通道中添加并优化噪声,这些噪声更加隐蔽且鲁棒,实验结果证明其能绕过较先进的防御方法 MagDR(mask-guided detection and reconstruction)^[65],实现对面部重现^[66]的攻击,但是这项工作没有考虑黑盒迁移性。

2.1.2 攻击面部交换

面部交换也是一种基于数据的伪造方法。面部交换模型也需要使用源人脸和目标人脸数据进

行训练,训练过程间接地学习了对抗样本的分布模式,因此针对这类伪造的攻击也较为困难,大部分攻击产生的失真不明显。Aneja 等^[67]提出针对人脸图像的有目标攻击(TAFIM, targeted attacks facial image manipulations)方法。TAFIM在通用噪声的基础上优化特定图像噪声,提高了攻击效率,同时在损失项中考虑了噪声的鲁棒性。实验表明,TAFIM能攻击SimSwap换脸^[68],并在伪造图像上产生大面积蓝色失真。Dong等^[69]使用FGSM和I-FGSM攻击deepfake。FGSM数据迁移性较好,但是特定图像上攻击性能较差,I-FGSM与之相反。该方法只能产生小幅失真,需要检测器鉴别真假,同时这项工作没有考虑黑盒攻击。后来他们改进了之前的工作^[70],提出可转移循环的对抗性GAN(TCA-GAN, transferable cycle adversary GAN),用以生成对抗样本,并在伪造的人脸图像上产生肉眼可见的模糊失真。TCA-GAN中使用了两对生成器-鉴别器,分别用于生成噪声和去除噪声,以此模拟实际场景中可能存在的防御,提高噪声鲁棒性。此外还使用对抗正则化,提高黑盒泛化能力。

2.1.3 攻击面部特征编辑

面部特征编辑是一种基于模型的伪造方法。不同于面部重现与面部交换,该方法对社会舆论的危害较小,主要应用于美颜娱乐软件。同时模型在产生伪造人脸时无须重新训练,因此更容易受到对抗样本的攻击而导致输出失真。现有关于攻击人脸伪造的工作,多数为攻击面部特征编辑,本文根据其对抗攻击特点,大致将这些工作划分为4个阶段。

第一阶段多为白盒攻击,没有考虑攻击的黑盒迁移能力。该阶段以Ruiz等^[71]和Yeh等^[72]的工作为代表,他们最初尝试攻击以cGAN^[8]为主干的面部特征编辑模型。前者使用I-FGSM^[28]攻击StarGAN模型^[9],同时利用扩谱干扰提高对抗噪声的鲁棒性,使得对抗噪声不会被图像压缩算法破坏。后者使用PGD算法^[32]攻击Pix2Pix^[73]和CycleGAN^[74]等模型,不仅考虑了图像失真攻击,还通过修改损失函数,使用图像锁定保护人脸图像免遭篡改。而裘昊轩等^[75]基于APGD(auto

PGD)算法^[76]提出DAPGD(dynamic auto PGD)算法,其核心在于动态设置学习率衰减的检查点,从而提高对抗样本的攻击性能。

现实场景下,大多数伪造模型是黑盒的,因此第一阶段的白盒攻击不具备应用意义。鉴于此,第二阶段主要围绕黑盒迁移性,改进第一阶段的工作,主要做法是先利用查询算法估计黑盒模型的大致梯度方向,再利用白盒优化算法得到对抗样本。Ruiz等^[77]提出LUP(leaking universal perturbations)黑盒查询攻击,该算法首先使用主成分分析法(PCA, principal component analysis)^[78]估计梯度的主要信息,再以此作为基础继续进行查询,查询结束后使用I-FGSM^[28]得到对抗样本。Yeh等^[79]提出LaSGSA(limit-aware self-guiding gradient sliding attack)方法。该方法首先使用限制感知的RGF^[42]和梯度步长滑动机制估计梯度,再将估计结果用于PGD攻击^[32]。与上述基于查询的黑盒攻击不同,Lyu等^[80]设计卷积网络,用于生成对抗水印,随后将带有水印的图像输入伪造模型,根据失真幅度设计损失函数,用以更新卷积网络的参数,多次重复以上步骤,从而实现迭代优化水印。此外,他们使用注意力机制确定人脸区域,使得水印优化过程集中在人脸区域。

虽然第二阶段工作成功攻击黑盒场景下的面部特征编辑,但是他们忽略了对抗样本的条件迁移性。为了改进这些不足,第三阶段综合考虑了对抗样本的黑盒迁移性和条件迁移性,该阶段的黑盒攻击主要是通过替代模型法实现的,替代模型法又可进一步分为单替代模型法和多替代模型加权法,前者只将单个白盒模型的梯度用于优化噪声,再将噪声迁移至黑盒模型;相比前者,后者将多个白盒模型梯度加权后再进行优化,能更有效地拟合大多数黑盒模型的梯度。Huang等^[63]的工作中,使用单替代模型法进行黑盒攻击,并将各条件域的联合损失作为优化目标,实现了条件迁移攻击。Huang等^[81]提出了跨模型的通用攻击水印(CMUA-watermark, cross models universal attack watermark),其通过加权联合损失实现数据迁移性和黑盒迁移性,同时通过多属性训练实现条件域之间迁移。Qiu等^[82]提出了跨条件域和模

型对抗攻击 (CDMAA, cross-domain and model adversarial attack)。他们认为, 各条件域之间具备梯度相似性, 因此 CDMAA 依然通过加权联合损失实现条件域迁移; 但是, 不同模型之间的梯度是不相似的, 不能直接通过加权联合损失进行迁移, 因此他们使用多重梯度下降算法 (MGDA, multiple gradient descent algorithm)^[83]进行梯度加权从而实现黑盒迁移。

随着图像去噪算法在该领域的应用, 一些用于攻击伪造模型的对抗噪声失去了原有的攻击性能, 因此第四阶段的工作主要考虑对抗噪声的鲁棒性。Dong 等^[70]提出的 TCA-GAN 不仅能够攻击面部交换, 还能攻击 StarGAN^[9]和 AttGAN^[10]等面部特征编辑模型。他们通过损失函数加权实现黑盒迁移性和条件迁移性, 还通过两个 GAN 互相博弈实现噪声的鲁棒性。同时, Wang 等^[64]的工作实现了攻击面部特征编辑。不同于其他鲁棒性工作, 他们考虑了对抗噪声如何抵御较先进 MagDR 图像重建算法^[65], 通过将图像 RGB 域上的噪声转换至 Lab 域以保证其鲁棒性, 但是他们没有考虑黑盒攻击以及条件域之间的泛化。

2.1.4 攻击人脸风格转换

面部生成是一种基于模型的伪造方法, 与面部特征编辑类似, 其主要应用于娱乐, 对社会的危害较小。现有关于攻击面部生成的工作较少。Aneja 等^[67]提出的 TAFIM 能够攻击人脸融合模型, 使得输出的伪造图像带有大面积白色失真。而 Qiu 等^[82]提出的 CDMAA 能够攻击用于动漫人脸生成的 U-GAT-IT 模型^[84], 使得输出的伪造图像上带有大面积黑色失真。

2.2 防御伪造模型

对抗样本能够攻击伪造模型, 使其输出大面积失真的伪造人脸, 从而失去了舆论欺骗意义。上述方法看似解决了人脸伪造问题, 然而对抗防御可以让伪造模型重新输出逼真的伪造人脸, 对抵制伪造人脸带来了新挑战。现有的防御伪造模型方法包括对抗训练和图像去噪, 其原理如图 5 和图 6 所示。

在一些利用图像处理技术防御对抗样本的工作中^[57-60]中, 所提的防御方法未应用于人脸伪造模型的防御。对此, Chen 等^[65]做了系统性的工作,

将以上方法用于防御 PGD^[32]和 C&W^[34]攻击, 使 StarGAN^[9]输出视觉正常的伪造人脸, 他们提出新的图像重建方法 MagDR。首先训练掩模引导的一致性检测器和失真检测器, 从伪造模型的输出图像判断输入图像是否添加了对抗水印。如果输入图像有水印, 则利用所提的 RecNet 重建算法消除对抗噪声, 并恢复所需的输出。在实验部分, 将 MagDR 和之前的防御工作做对比, 将图像去噪之后再进行伪造。结果表明, 包括 MagDR 在内的一些方法重建的图像, 经过 StarGAN^[9]伪造之后没有留下大面积明显伪影并输出视觉正常的人脸伪造图像。

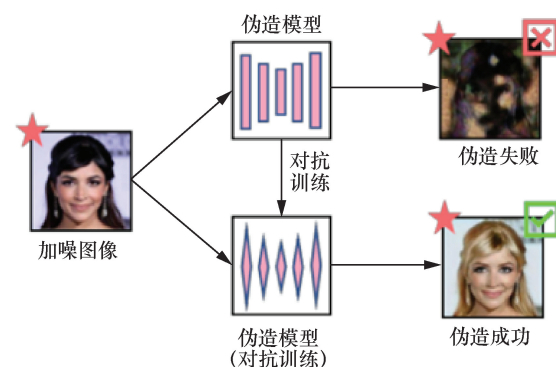


图 5 伪造模型的防御 (对抗训练)
Figure 5 Defense for forgery models (adversarial training)

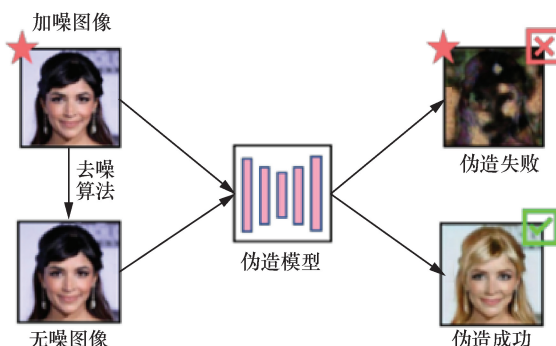


图 6 伪造模型的防御 (图像去噪)
Figure 6 Defense for forgery models (image denoise)

而 Ruiz 等^[71]结合对抗训练和图像处理防御人脸伪造模型, 将带有对抗噪声的图片对 GAN 中的生成器和判别器进行对抗性训练, 同时将图像进行高斯滤波去除对抗噪声, 实验表明, 经过防御的 StarGAN^[9]对 FGSM^[27]、I-FGSM^[28]和 PGD^[32]这 3 种攻击算法具备鲁棒性, 但是因为输入图像经过高斯滤波, 输出的伪造图像质量有所降低。

2.3 小结

对抗攻防在人脸伪造中的研究现状概览如图7所示。可以看出,攻击伪造工作主要集中在攻击面部特征编辑,而攻击面部重现和面部交换的工作较少。因为面部特征编辑是基于模型的伪造方式,攻击难度小且成功率高。面部重现和面部交换是基于数据的伪造方式,需要根据目标人脸训练模型,然而训练过程中,模型学习了对抗样本

的分布模式,导致模型能够抵御对抗干扰,使得攻击难度变大。现有的防御伪造工作主要是通过图像处理技术实现的,而对抗训练较少。用于人脸伪造的多半为GAN模型,相比一般的DNN模型,GAN模型较为复杂,其中包括生成器和鉴别器,导致基于模型的对抗训练方法难以部署。相比之下,图像处理防御兼容性强,无须考虑模型结构。

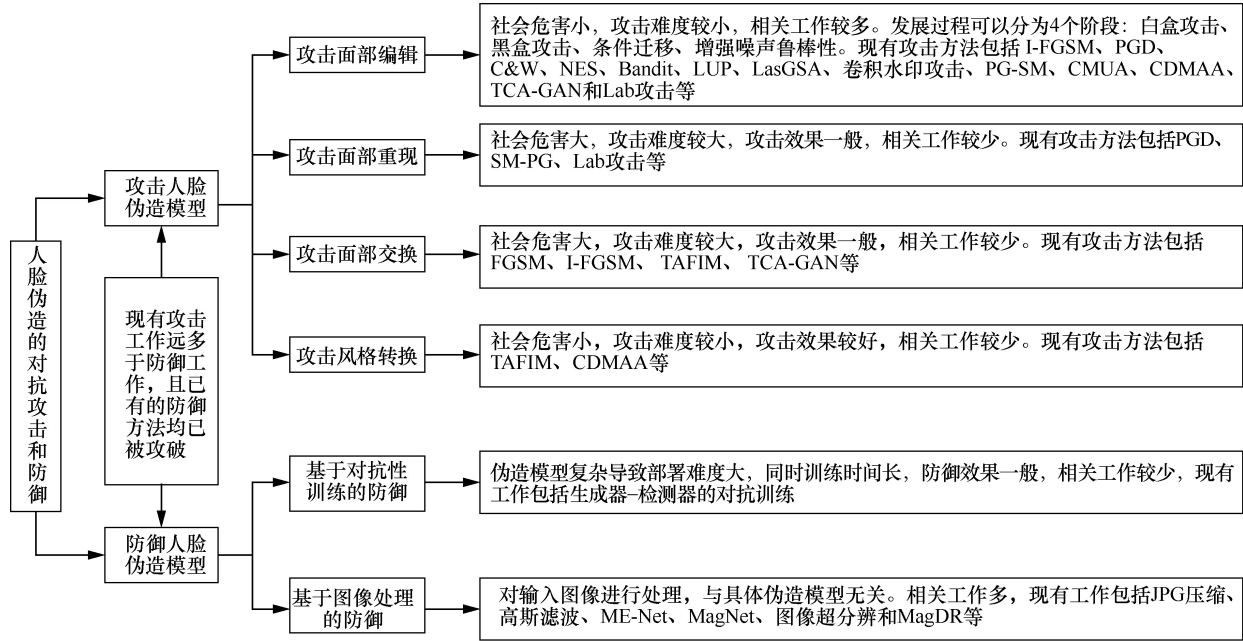


图7 对抗攻防在人脸伪造中的研究现状概览

Figure 7 An overview of adversarial attack and defense in face forgery

3 虚假人脸检测中的对抗攻击和防御

本节从攻击检测模型和防御检测模型两个角度阐述现有工作。攻击检测模型是将对抗性虚假人脸图像输入检测模型,使其误判为真;而防御检测模型是利用对抗防御方法,减轻对抗性虚假人脸图像对检测模型的影响,使得检测模型正确辨别对抗性图像的真伪。

3.1 攻击检测模型

本节根据几种对抗攻击的特点,对现有工作展开叙述。这些特点包括黑盒迁移性、对抗隐蔽性、对抗鲁棒性以及数据迁移性。

3.1.1 黑盒迁移性

现实对抗场景中,社交平台上用于鉴别真伪图像的检测模型往往不会公开内部细节,因此无法根据白盒攻击方法,生成对抗样本并攻击检测

模型。Hussain 等^[85]先用基于 NES 的黑盒查询法^[43]对 Xception^[17]和 MesoNet^[86]进行黑盒梯度查询,随后使用 I-FGSM 算法^[28]生成对抗样本。然而,攻击对象为检测模型时,基于查询的黑盒攻击会受到查询次数的限制。现有大部分工作^[87-92]使用替代模型法进行黑盒攻击,具体做法是先在白盒模型上,利用白盒攻击算法生成对抗样本,再将对抗样本用于未知黑盒模型的攻击。其中有部分工作改进了迁移性能,如 Li 等^[91]使用多模型 logit 输出分数加权和作为优化目标,而 Jia 等^[92]提出在频率域添加对抗噪声。

3.1.2 对抗隐蔽性

成功攻击虚假人脸检测的前提是检测方没有发现图像中的对抗痕迹,直接将其输入检测模型,得到错误结果。如果对抗痕迹过于明显,那么会引起检测方的警觉,将采取防御手段消除对抗攻

击带来的影响,从而导致攻击失败。在图像中加入噪声的对抗攻击,会在一定程度上导致图像失真或模糊。为了减少失真度,前期的工作^[85,87-88,93]使用目标函数优化方法,但是经过这些方法优化之后的噪声,依然是分布于全图的对抗噪声,存在冗余。Neekhara 等^[89]获取检测模型注意力区域,由于检测模型主要根据这些区域判定人脸图像真伪,所以这些区域中加入对抗噪声即可实现攻击检测模型。类似地,Liao 等^[90]提出了多层语义特征关键区域选择 (MLSKRS, multi-layers semantic key region selection),用于确定对抗关键区域。以上方法是在 RGB 图像域中添加噪声,而 Jia 等^[92]和 Wang 等^[94]分别将对抗噪声加入频率域和 YCbCr 颜色域以降低人眼可察觉程度。

除了在图像中减少冗余噪声,还有一些方法没有在意图像中加噪,而是直接使用 GAN 生成无伪影的虚假人脸图像欺骗检测器。Li 等^[91]提出一种攻击虚假人脸检测器的方法,不在图像中添加噪声,而是利用对抗优化算法,对 StyleGAN^[12]网络中的潜层向量进行梯度更新,使其生成的图像能够直接欺骗检测模型。而 Zhao 等^[95]和 Liu 等^[96]通过 GAN 生成器和鉴别器的对抗训练,消除虚假人脸图像中的伪造痕迹,以实现攻击检测器。

3.1.3 对抗鲁棒性

图像和视频上传至社交平台前会经过一次压缩,压缩有可能破坏图像中的对抗噪声,对抗噪声失效后,会导致虚假的人脸图像无法逃避检测。为了提高对抗噪声的鲁棒性,需要在用于生成对抗噪声的损失函数中考虑各种可能的图像变换,并构建联合损失以优化对抗噪声。现有工作中,Hussain 等^[85]考虑了加性高斯噪声、高斯模糊、图像上采样以及下采样;而 Neekhara 等^[89]则考虑了图像平移、加性高斯噪声以及图像压缩。此外,上述对抗性 GAN 生成图像的攻击方法^[95-96],由于没有在意虚假人脸图像加入对抗噪声,不会受到图像压缩方法的影响。

3.1.4 数据迁移性

在一些涉及虚假人脸视频的对抗攻击中,若对视频的每帧都加噪,则时间成本和计算开销高。为了更有效地制作对抗性虚假人脸视频,需要生

成通用的噪声,迁移至视频的每一帧。现有工作多为制作对抗性虚假人脸图像,因此没有考虑数据迁移性。而 Hussain 等^[85]的工作中,涉及制作对抗性虚假视频,但是他们没有提出生成通用噪声的方法。作为改进,Neekhara 等^[89]的工作中,使用联合数据分布的目标函数以生成通用噪声,提高制作对抗性虚假人脸视频的效率。

3.2 防御检测模型

虚假人脸检测模型容易受到各类干扰而导致真假分类错误。本节介绍一些用于增强检测器对抗鲁棒性的防御方法,根据防御方式可以划分为梯度正则、模型集成、图像处理和对抗性训练。

3.2.1 基于梯度正则的防御

在对抗样本生成阶段,对检测模型进行梯度屏蔽,使得攻击算法无法获得其梯度,从而生成噪声失败,现有的工作是:Gandhi 等^[88]使用 Lipschitz 正则化^[86]限制了检测器相对输入数据的梯度,但是这种梯度屏蔽方法对于提升检测模型的鲁棒性没有明显效果。此外,该方法不适用于各种模型架构,无法防御黑盒攻击,在实际场景中不适用。

3.2.2 基于模型集成的防御

该方法借鉴了机器学习中集成学习的思想,其原理如图 8 所示,每一种虚假人脸检测方法,都有其独特的优势。利用这一点,Khan 等^[97]集成了 VGG16^[15]、InceptionV3^[98]、Xception^[17]构建集成分类器,但是他们并没有在意对抗样本上测试准确度,只是在 DFDC 和 TIMIT 数据集上进行评估,证明集成模型的分类型准确性高于各集成部分。而 Dutta 等^[99]提出 EnsembleDet 算法,集成 Xception^[17]和 MesoNet^[86]模型以抵御 FGSM^[27]和 I-FGSM^[28]在 Face2Face^[3]上的反取证攻击。上述集成模型的防御方法容易被可模型间黑盒迁移的对抗攻击所攻破,因此 Hooda 等^[100]集成了多个基于频谱的虚假人脸分类模型,并对各集成模型执行梯度正交化,从而降低对抗样本在这些模型之间的迁移性。

3.2.3 基于图像处理的防御

添加至图像中的细微噪声,会因各种图像处理方法而遭到破坏,因此可以利用这些方法防御检测模型,其原理如图 9 所示。基于此,Gandhi

等^[88]使用深度图像先验 (DIP, deep image prior)^[59]进行对抗防御。他们在数据输入分类器之前对其进行预处理来消除噪声。但是其迭代 6 000 次后才能达到最优精度, 处理耗时较长。Luo 等^[101]对此提出改进, 使用双边滤波图像处理方法, 将对抗性虚假人脸图像去噪后再放入检测器进行分类。

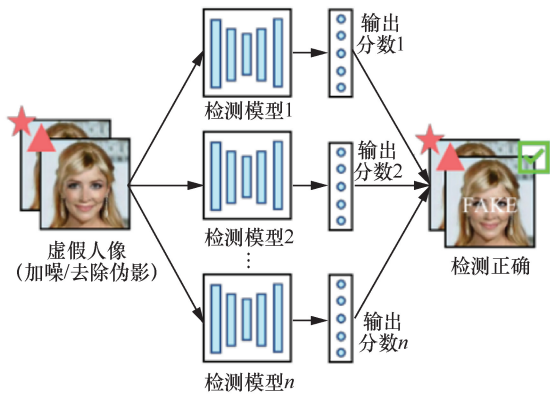


图8 假脸检测模型的防御 (模型集成)
Figure 8 Defense method for fake face detection models (model ensemble)

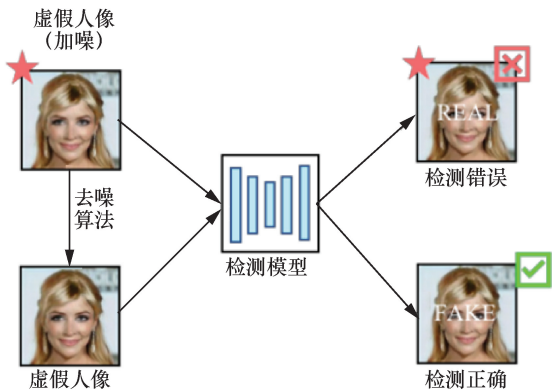


图9 假脸检测模型的防御 (图像处理)
Figure 9 Defense method for fake face detection models (image processing)

3.2.4 基于对抗训练的防御

对抗训练是让检测模型在各种对抗性数据上训练分类, 从而提高其鲁棒性和泛化性, 其原理如图 10 所示。如果只用一种对抗数据训练检测模型, 那么模型可能无法防御其他类型的对抗数据。鉴于此, Luo 等^[101]提出联合对抗训练, 将 PGD、I-FGSM 和 FGSM 这 3 种对抗攻击算法所生成的对抗性图像引入 Xception 的训练过程, 使模型鲁棒性得到提升。

上述基于图像处理的防御方法, 虽然模型结

构独立且简单易用, 但是会降低图像质量。而一些人脸取证方法需要根据 GAN 伪影来判别图像真伪, 图像质量较低会导致这些伪影消失, 从而引起检测器判别错误。此外, 经过社交媒体平台压缩处理后的虚假人脸图像, 也会带来这些问题。鉴于此, Wang 等^[102]在检测模型的训练阶段, 引入加性对抗样本和高斯对抗样本, 使检测模型学习伪影以外的其他伪造特征, 在面对低分辨率的伪造图像时, 也能表现良好。

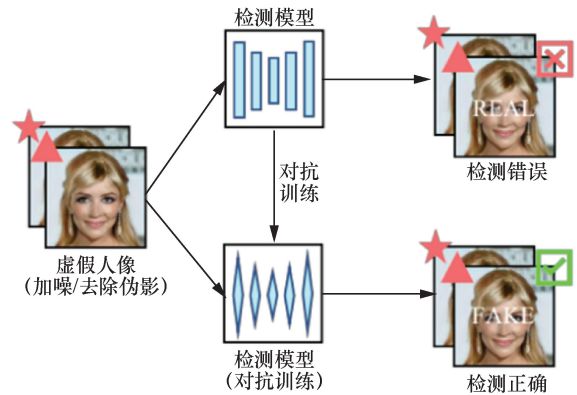


图10 假脸检测模型的防御 (对抗训练)
Figure 10 Defense method for fake face detection models (adversarial training)

一些检测模型容易在特定的伪造方法上过拟合, 在未知方法生成的虚假人脸上检测效果不佳, 鉴于此, Chen 等^[103]使用伪造区域以及混合参数约束 GAN, 并生成指定类型的虚假人脸图像。随后将这些虚假人脸图像作为训练数据, 让检测模型学习更一般的伪造特征, 从而提高其泛化性。Jeong 等^[104]也利用对抗性训练增强模型鲁棒性, 他们指出, 若检测方法考虑 GAN 的频率域伪影, 虽然有利于提高检测模型在特定伪造方法上的性能, 但是泛化性会有所降低, 而若只考虑一般性特征, 又会导致特定的伪造方法上效果不佳, 因此提出利用 GAN 循序渐进地生成带有频率域扰动的虚假人脸, 用其作为训练数据, 使检测模型先学习频域率伪影而后学习一般性特征, 提高检测模型的特异性能和一般性能。

3.3 小结

攻击虚假人脸检测模型研究工作的概览如表 1 所示。总体来看, 现有的攻击工作主要解决 4 个方面的问题: 提高对抗样本的黑盒迁移性、隐蔽性、鲁棒性以及数据迁移性。黑盒迁移主要通过

替代模型迁移法实现,该方法有一定效果,但是其取决于模型之间的结构相似性。对抗样本的鲁棒性和隐蔽性是一对互斥的属性,鲁棒性越高,对抗强度也越高,导致对抗样本隐蔽性降低。目前的主要解决方法是目标函数优化和对抗性GAN。前者通过目标函数中的平衡因子控制优化偏好,但其本质上无法同时提升鲁棒性和隐蔽性,而后者可以较好地解决上述问题,但是需要额外的训练成本。此外,图像核心区域和其他域中加噪是一种成本相对低廉且有效的攻击方式,但是前者效果取决于注意力区域确定算法的精确度,而后者在黑盒攻击时效果一般。

防御虚假人脸检测模型研究工作的概览如表2所示。基于梯度正则的防御限制多而不能投入实际使用,具备实用意义的包括基于模型集成的防御、基于图像处理的防御和基于对抗训练的防御。基于模型集成的防御和基于图像处理的防御均属于使用较为灵活的防御方式,可以不用修改模型

细节,但是它们都存在安全漏洞,基于模型集成的防御会被可迁移的对抗样本攻破,导致投票后选出错误类别,而基于图像处理的防御无法防御GAN痕迹抹除攻击。基于对抗训练的防御可以通过训练阶段学习数据分布从而防御多种类型的攻击,是相对而言最全面的取证防御策略,但是其存在训练数据处理成本较高的问题。

4 结束语

随着社交媒体的快速发展,人脸伪造带来的舆论安全问题日益突出,对虚假人脸进行鉴别和防范是当下的研究热点。对抗样本的应用改变了原有的伪造和取证模式,使得二者之间的博弈充满更多干扰因素。本文从人脸伪造和虚假人脸检测两个方面出发,展示了近年来对抗样本在该领域的研究进展与应用。关于人脸伪造方面,现有的对抗攻击多用于破坏舆论影响较小的人脸特征编辑,针对危害严重的人脸交换和面部重现,攻

表1 攻击虚假人脸检测模型研究工作的概览
Table 1 The overview of attacking fake face detection models

对抗特性	实现方法	代表工作	优点	缺点
黑盒迁移性	查询法	文献[85]	无须训练替代模型	受查询次数限制
	替代模型法	文献[87-92]	不受查询次数限制	要训练结构相似的替代模型
对抗隐蔽性	目标函数优化噪声	文献[85,87-88,93]	无须额外处理成本	效果有限,噪声依然分布于全图
	核心区域加噪	文献[89-90]	减少加噪区域	注意力区域偏差,导致攻击效果不佳
	图像其他域中加噪	文献[92,94]	噪声较难察觉	黑盒攻击能力一般
	对抗性GAN	文献[91,95-96]	无须添加噪声	训练GAN需要计算成本
对抗鲁棒性	目标函数优化噪声	文献[85,89]	无须额外处理成本	无法考虑所有可能的图像处理
	对抗性GAN	文献[91,95-96]	不受图像处理防御影响	训练GAN需要计算成本
数据迁移性	目标函数优化噪声	文献[89]	提升对抗噪声生成效率	会降低特定图像攻击能力

表2 防御虚假人脸检测模型研究工作的概览
Table 2 The overview of defending fake face detection models

防御方式	具体技术	代表工作	优点	缺点
梯度正则	Lipschitz 正则化 ^[85]	文献[88]	没有额外的处理开销	防御效果有限,仅有轻微的性能提升
模型集成	模型输出分数集成	文献[97,99]	集成各模型的优势进行投票,简单有效	需要训练多个用于集成的子模型,且无法防御可迁移的对抗方法
	正交梯度的模型集成	文献[100]		
图像处理	深度图像先验 ^[59]	文献[88]	与具体模型结构无关,能兼容各种检测模型	无法防御对抗性GAN的指纹去除攻击,且会导致图像质量降低而影响取证精度
	双边滤波	文献[101]		
对抗训练	联合对抗训练	文献[101]		
	噪声对抗训练	文献[103]	通用性强,能够防御多种形式的对抗攻击	需要额外生成对抗样本以供模型学习,计算开销较大,且要控制训练数据比例
	频率域伪影对抗训练	文献[104]		

击效果较为一般。同时,一些先进对抗防御方法能够让伪造模型抵御对抗样本的干扰,重新输出逼真的伪造人脸。至于虚拟人脸检测方面,现有检测模型大多是面向特定场景的虚假人脸检测,在特定数据上的检测效果良好,但是其容易受到对抗样本的威胁,在噪声数据上效果不佳。此外,近来扩散模型^[105-106]的应用使得AI生成的图像更加自然与真实,导致检测模型更难发现虚假图像中存在的瑕疵。总体来看,现阶段和虚假人脸的斗争形势依然严峻。为了解决上述问题,未来反伪造的工作可从以下角度考虑。

1) 攻击角度: 利用对抗样本攻击人脸伪造的多个子过程。完整的人脸伪造包括人脸区域确定、五官标定、伪造、抹除痕迹等多个子过程,如果仅攻击伪造过程,那么可能导致一些伪造方法逃避攻击。确定人脸区域是成功伪造的前提,现有工作^[107]中有针对人脸区域定位的攻击,但是与攻击伪造模型一样,这些方法仅攻击单个过程。未来的工作中可以尝试设计攻击多个子过程的对抗噪声,从而最大程度地破坏人脸伪造。

2) 防御角度: 结合图像去噪和对抗训练防御检测模型。未来的工作可尝试根据攻击类型(如加噪攻击、痕迹抹除攻击等)进行划分,再将这此方法生成的对抗性图像根据一定比例组合成数据集,作为模型的训练样本,使得模型能够在不同的对抗环境中学习到伪造人脸的高级语义特征,从而提高检测模型的泛化性。此外,用于防御对抗样本的图像去噪是一种数据防御方法,与具体检测模型无关。先利用图像处理破坏对抗噪声的结构,降低对抗图像的攻击性,再将这些图像输入上述经过对抗训练的模型,可以进一步提高检测模型在对抗图像上的精度。

参考文献:

- [1] KINGMA D P, WELLING M. Auto-encoding variational bayes[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2014.
- [2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [3] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2387-2395.
- [4] THIES J, ZOLLHÖFER M, NIEBNER M. Deferred neural rendering: image synthesis using neural textures[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.
- [5] FaceSwap[EB].
- [6] DeepFakes[EB].
- [7] LI L, BAO J, YANG H, et al. Faceshifter: towards high fidelity and occlusion aware face swapping[J]. arXiv Preprint arXiv:1912.13457, 2019.
- [8] MIRZA M, OSINDERO S. Conditional generative adversarial nets[C]//Proceedings of the Conference on Neural Information Processing Systems (NIPS). 2014.
- [9] CHOI Y, CHOI M, KIM M, et al. Stargan: unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 8789-8797.
- [10] HE Z, ZUO W, KAN M, et al. Attgan: Facial attribute editing by only changing what you want[J]. IEEE transactions on image processing, 2019, 28(11): 5464-5478.
- [11] LIU M, DING Y, XIA M, et al. Stgan: a unified selective transfer network for arbitrary image attribute editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3673-3682.
- [12] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 4401-4410.
- [13] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). 2020: 8110-8119.
- [14] KARRAS T, AITTALA M, LAINE S, et al. Alias-free generative adversarial networks[C]//Advances in Neural Information Processing Systems. 2021: 852-863.
- [15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2015.
- [16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern recognition (CVPR). 2016: 770-778.
- [17] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer vision and Pattern recognition (CVPR). 2017: 1251-1258.
- [18] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[J]. arXiv Preprint arXiv:1704.04861, 2017.
- [19] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning (ICML). 2019: 6105-6114.
- [20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J].

- arXiv Preprint arXiv:2010.11929, 2020.
- [21] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2014.
 - [22] 林点, 潘理, 易平. 面向图像识别的卷积神经网络鲁棒性研究进展[J]. 网络与信息安全学报, 2022, 8(3): 111-122.
LIN D, PAN L, YI P. Research on the robustness of convolutional neural networks in image recognition[J]. Chinese Journal of Network and Information Security, 2022, 8(3): 111-122.
 - [23] ZHANG J, LI C. Adversarial examples: opportunities and challenges[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(7): 2578-2593.
 - [24] 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. Research progress of digital image forensic techniques based on deep learning[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 13-28.
 - [25] 董琳, 黄丽清, 叶锋, 等. 人脸伪造检测泛化性方法综述[J]. 计算机科学, 2022, 49(2): 12-30.
DONG L, HUANG L Q, YE F, et al. Survey on generalization methods of face forgery detection[J]. Computer Science, 2022, 49(2): 12-30.
 - [26] NGUYEN T T, NGUYEN Q V H, NGUYEN D T, et al. Deep learning for deepfakes creation and detection: a survey[J]. Computer Vision and Image Understanding, 2022, 223: 103525.
 - [27] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2015.
 - [28] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]//Artificial Intelligence Safety and Security. 2018: 99-112.
 - [29] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.
 - [30] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.
 - [31] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//2016 IEEE European Symposium on Security and Privacy (EuroS&P). 2016: 372-387.
 - [32] MARDY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the International Conference on Learning Representations(ICLR). 2018.
 - [33] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017: 1765-1773.
 - [34] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (S&P). 2017: 39-57.
 - [35] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2017: 1-24.
 - [36] 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 (CVPR). 2018: 9185-9193.
 - [37] SHI Y, WANG S, HAN Y. Curls & whey: boosting black-box adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019: 6519-6527.
 - [38] XIE C, ZHANG Z, ZHOU Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 2730-2739.
 - [39] DONG Y, PANG T, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 4312-4321.
 - [40] WANG W, YIN B, YAO T, et al. Delving into data: effectively substitute training for black-box attack[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 4761-4770.
 - [41] FAN M, GUO W, YU S, et al. Enhance transferability of adversarial examples with model architecture[J]. arXiv Preprint arXiv: 2202.13625, 2022.
 - [42] NESTEROV Y, SPOKOINY V. Random gradient-free minimization of convex functions[J]. Foundations of Computational Mathematics, 2017, 17(2): 527-566.
 - [43] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[C]//International Conference on Machine Learning (ICML). 2018: 2137-2146.
 - [44] ILYAS A, ENGSTROM L, MADRY A. Prior convictions: Black-box adversarial attacks with bandits and priors[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2019.
 - [45] GUO C, GARDNER J, YOU Y, et al. Simple black-box adversarial attacks[C]//International Conference on Machine Learning (ICML). 2019: 2484-2493.
 - [46] LI H, XU X, ZHANG X, et al. Qeba: Query-efficient boundary-based black box attack[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 1221-1230.
 - [47] MAHO T, FURON T, LE MERRER E. Surfree: a fast surrogate-free black-box attack[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 10430-10439.
 - [48] ABBASI M, GAGNÉ C. Robustness to adversarial examples through an ensemble of specialists[C]//Proceedings of the International Conference on Learning Representations Workshops (ICLR). 2017.
 - [49] HE W, WEI J, CHEN X, et al. Adversarial example defense: ensembles of weak defenses are not strong[C]//11th USENIX Workshop on Offensive Technologies (WOOT 17). 2017.
 - [50] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//2016 IEEE Symposium on Security and Privacy (S&P).

- 2016: 582-597.
- [51] YE N, LI Q, ZHOU X Y, et al. Amata: an annealing mechanism for adversarial training acceleration[C]//Proceedings of the AAAI Conference on Artificial Intelligence(AAAI). 2021, 35(12): 10691-10699.
 - [52] SHAFABI A, NAJIBI M, GHASI M A, et al. Adversarial training for free![J]. Advances in Neural Information Processing Systems, 2019, 32.
 - [53] YAN Z, GUO Y, ZHANG C. Deep defense: training DNNs with improved adversarial robustness[J]. Advances in Neural Information Processing Systems, 2018, 31.
 - [54] SONG C, HE K, WANG L, et al. Improving the generalization of adversarial training with domain adaptation[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2019: 1-14.
 - [55] DZIUGAITE G K, GHAHRAMANI Z, ROY D M. A study of the effect of jpg compression on adversarial images[J]. arXiv Preprint arXiv:1608.00853, 2016.
 - [56] MUSTAFA A, KHAN S H, HAYAT M, et al. Image super-resolution as a defense against adversarial attacks[J]. IEEE Transactions on Image Processing, 2019, 29: 1711-1724.
 - [57] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 135-147.
 - [58] YANG Y, ZHANG G, KATABI D, et al. Me-net: towards effective adversarial robustness with matrix estimation[C]//International Conference on Machine Learning (ICML). 2019: 1-22.
 - [59] ULYANOV D, VEDALDI A, LEMPITSKY V. Deep image prior[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern recognition (CVPR). 2018: 9446-9454.
 - [60] JIA X, WEI X, CAO X, et al. Comdefend: an efficient image compression model to defend adversarial examples[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 6084-6092.
 - [61] YANG C, DING L, CHEN Y, et al. Defending against gan-based deepfake attacks via transformation-aware adversarial faces[C]//2021 International Joint Conference on Neural Networks (IJCNN). 2021: 1-8.
 - [62] WU Y, ABDALMAGEED W, NATARAJAN P. Mantra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9543-9552.
 - [63] HUANG Q, ZHANG J, ZHOU W, et al. Initiative defense against facial manipulation[C]//Proceedings of the AAAI Conference on Artificial intelligence (AAAI). 2021, 35(2): 1619-1627.
 - [64] WANG R, HUANG Z, CHEN Z, et al. Anti-forgery: towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations[J]. arXiv Preprint arXiv: 2206.00477, 2022.
 - [65] CHEN Z, XIE L, PANG S, et al. Magdr: mask-guided detection and reconstruction for defending deepfakes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 9014-9023.
 - [66] TRIPATHY S, KANNALA J, RAHTU E. Icfac: interpretable and controllable face reenactment using GANs[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2020: 3385-3394.
 - [67] ANEJA S, MARKHASIN L, NIEBNER M. TAFIM: targeted adversarial attacks against facial image manipulations[C]//European Conference on Computer Vision (ECCV). 2022: 58-75.
 - [68] CHEN R, CHEN X, NI B, et al. Simswap: an efficient framework for high fidelity face swapping[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 2003-2011.
 - [69] DONG J, XIE X. Visually maintained image disturbance against deepfake face swapping[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). 2021: 1-6.
 - [70] DONG J, WANG Y, LAI J, et al. Restricted black-box adversarial attack against deepfake face swapping[J]. arXiv Preprint arXiv:2204.12347, 2022.
 - [71] RUIZ N, BARGAL S A, SCLAROFF S. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems[C]//European Conference on Computer Vision (ECCV). 2020: 236-251.
 - [72] YE H C Y, CHEN H W, TSAI S L, et al. Disrupting image-translation-based deepfake algorithms with adversarial attacks[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACV). 2020: 53-62.
 - [73] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1125-1134.
 - [74] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision (CVPR). 2017: 2223-2232.
 - [75] 裴昊轩, 杜彦辉, 芦天亮. 针对深度伪造的对抗攻击算法动态 APGD 设计[J]. 计算机工程与应用, 2022, 58(24): 97-106.
 - [76] QIU H X, DU Y H, LU T L. Design of DAPGD of adversarial attack algorithm against deepfake[J]. Computer Engineering and Applications, 2022, 58(24): 97-106.
 - [76] CROCE F, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]//International Conference on Machine Learning. 2020: 2206-2216.
 - [77] RUIZ N, BARGAL S A, SCLAROFF S. Protecting against image translation deepfakes by leaking universal perturbations from black-box neural networks[J]. arXiv Preprint arXiv:2006.06493, 2020.
 - [78] PEARSON K. LIII. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901, 2(11): 559-572.
 - [79] YE H C Y, CHEN H W, SHUAI H H, et al. Attack as the best defense: nullifying image-to-image translation GANs via limit-aware

- adversarial attack[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 16188-16197.
- [80] LYU L. Smart watermark to defend against deepfake image manipulation[C]//2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). 2021: 380-384.
- [81] HUANG H, WANG Y, CHEN Z, et al. Cmu-a-watermark: a cross-model universal adversarial watermark for combating deepfakes[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2022, 36(1): 989-997.
- [82] QIU H, DU Y, LU T. The framework of cross-domain and model adversarial attack against deepfake[J]. *Future Internet*, 2022, 14(2): 46.
- [83] DÉSIDÉRI J A. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization[J]. *Comptes Rendus Mathématique*, 2012, 350(5-6): 313-318.
- [84] KIM J, KIM M, KANG H, et al. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2020.
- [85] HUSSAIN S, NEEKHARA P, JERE M, et al. Adversarial deepfakes: evaluating vulnerability of deepfake detectors to adversarial examples[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2021: 3348-3357.
- [86] AFCHAR D, NOZICK V, YAMAGISHI J, et al. Mesonet: a compact facial video forgery detection network[C]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018: 1-7.
- [87] CARLINI N, FARID H. Evading deepfake-image detectors with white-and black-box attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR). 2020: 658-659.
- [88] GANDHI A, JAIN S. Adversarial perturbations fool deepfake detectors[C]//2020 International Joint Conference on Neural Networks (IJCNN). 2020: 1-8.
- [89] NEEKHARA P, DOLHANSKY B, BITTON J, et al. Adversarial threats to deepfake detection: a practical perspective[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition (CVPR). 2021: 923-932.
- [90] LIAO Q, LI Y, WANG X, et al. Imperceptible adversarial examples for fake image detection[C]//Proceedings of the IEEE International Conference on Image Processing (ICIP). 2021.
- [91] LI D, WANG W, FAN H, et al. Exploring adversarial fake images on face manifold[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2021: 5789-5798.
- [92] JIA S, MA C, YAO T, et al. Exploring frequency adversarial attacks for face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2022: 4103-4112.
- [93] FAN L, LI W, CUI X. Deepfake-image anti-forensics with adversarial examples attacks[J]. *Future Internet*, 2021, 13(11): 288.
- [94] WANG Y, DING X, YANG Y, et al. Perception matters: exploring imperceptible and transferable anti-forensics for GAN-generated fake face imagery detection[J]. *Pattern Recognition Letters*, 2021, 146: 15-22.
- [95] ZHAO X, STAMM M C. Making GAN-generated images difficult to spot: a new attack against synthetic image detectors[J]. *arXiv preprint arXiv:2104.12069*, 2021.
- [96] LIU C, CHEN H, ZHU T, et al. Making deepfakes more spurious: evading deep face forgery detection via trace removal attack[J]. *IEEE Transactions on Dependable and Secure Computing*, 2023.
- [97] KHAN S A, ARTUSI A, DAI H. Adversarially robust deepfake media detection using fused convolutional neural network predictions[J]. *arXiv Preprint arXiv:2102.05950*, 2021.
- [98] SZEGEDY C, VANHOUCHE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.
- [99] DUTTA H, PANDEY A, BILGAIYAN S. EnsembleDet: ensembling against adversarial attack on deepfake detection[J]. *Journal of Electronic Imaging*, 2021, 30(6): 063030.
- [100] HOODA A, MANGAOKAR N, FENG R, et al. Towards adversarially robust deepfake detection: an ensemble approach[J]. *arXiv preprint arXiv:2202.05687*, 2022.
- [101] LUO Y, YE F, WENG B, et al. A novel defensive strategy for facial manipulation detection combining bilateral filtering and joint adversarial training[J]. *Security and Communication Networks*, 2021, 2021.
- [102] WANG Z, GUO Y, ZUO W. Deepfake forensics via an adversarial game[J]. *IEEE Transactions on Image Processing*, 2022.
- [103] CHEN L, ZHANG Y, SONG Y, et al. Self-supervised learning of adversarial example: towards good generalizations for deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 18710-18719.
- [104] JEONG Y, KIM D, RO Y, et al. FrePGAN: robust deepfake detection using frequency-level perturbations[J]. *arXiv Preprint arXiv:2202.03347*, 2022.
- [105] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems. 2020: 6840-6851.
- [106] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10684-10695.
- [107] SUN P, LI Y, QI H, et al. Landmark breaker: obstructing deepfake by disturbing landmark extraction[C]//2020 IEEE International Workshop on Information Forensics and Security (WIFS). 2020: 1-6.

[作者简介]



黄诗瑀（1999- ），男，福建厦门人，福建师范大学硕士生，主要研究方向为对抗性深度学习和数字媒体取证。



李伟（1997- ），男，湖北十堰人，福建师范大学硕士生，主要研究方向为对抗性深度学习和数字媒体取证。



叶锋（1978- ），男，福建福州人，博士，福建师范大学副教授，主要研究方向为计算机视觉和视频图像编码。



黄丽清（1991- ），女，福建莆田人，博士，福建师范大学讲师，主要研究方向为视频图像超分辨、去模糊处理和数字媒体取证。



黄添强（1971- ），男，福建莆田人，博士，福建师范大学教授、博士生导师，主要研究方向为机器学习安全和数字媒体取证。



罗海峰（1990- ），男，福建龙岩人，博士，福建师范大学讲师，主要研究方向为三维点云数据处理。