

基于多伪造模型对抗优化的鲁棒图像扰动

穆文鹏^{1,2}, 王金伟^{1,2}, 陈北京^{1,2}, 聂丽娜^{1,2}, 莫皓岚^{1,2}, 徐 菲^{1,2}

¹南京信息工程大学计算机学院、网络空间安全学院 南京 中国 210044

²南京信息工程大学数字取证教育部工程研究中心 南京 中国 210044

摘要 深度人脸伪造技术,即用于创造虚假面部特征或整体重塑面部的恶意手段,对个人声誉和社会安全构成重大威胁。为了应对这一挑战,部分研究者提出运用对抗性扰动来抵制深度人脸伪造模型的攻击,该方法旨在使这些深度人脸伪造模型的输出严重受损,以防其伪造能力。尽管加入对抗扰动的确导致生成的伪造面部图像与真实图像存在显著的视觉差异,但这些对抗扰动仅能对在训练过程中已知的特定深度人脸伪造模型提供较为有效的保护,难以涵盖多方面的威胁。为解决这一问题,本文提出训练多模型对抗扰动的方案,并且加入训练的多个模型内部结构千差万别,以使得到的对抗扰动可以通用地护卫面部图像免受多种已知和未知伪造模型的侵害。首先,我们建立了一种多模型并行攻击管道,同时对多个深度人脸伪造模型进行对抗攻击,不断强化对各模型的抗性。其次,我们创新性地设计了显著性偏向融合策略,旨在缓解不同管道分支造成的对抗性扰动冲突。此外,为提升模型的泛化能力和更好模拟实际图像拍摄情景,我们还在原始图像中采用数据增强技术。最后,在并行攻击管道分支融合对抗扰动后,采用启发式的树形 PARZEN 方法自动寻找最适攻击步长,以进一步缓解模型间扰动的兼容性问题。大量实验验证显示,我们所提出的多模型对抗性扰动能有效污损多种深度人脸伪造模型生成的伪造人脸图像,并在鲁棒性、高效性和泛化性上取得较好的平衡。

关键词 多模型通用图像扰动; 对抗攻击; 深度伪造; 机器学习

中图分类号 TP309.7 DOI号 10.19363/J.cnki.cn10-1380/tn.2025.07.04

Robust Image Perturbation Based on Multi-Forgery Model Adversarial Optimization

MU Wenpeng^{1,2}, WANG Jinwei^{1,2}, CHEN Beijin^{1,2}, NIE Lina^{1,2}, MO Haolan^{1,2}, XU Fei^{1,2}

¹ School of Computer Science and Cyberspace Security, Nanjing University of Information Science & Technology, Nanjing 210044, China

² Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science & Technology, Nanjing 210044, China

Abstract Deepfakes, malicious means used to create false facial features or completely reshape faces, pose a major threat to personal reputation and social security. In order to deal with this challenge, some researchers have proposed using adversarial perturbations to resist the attacks of deep face forgery models. This method aims to seriously taint the output of these deep face forgery models to prevent their forgery capabilities. Although adding adversarial perturbations does lead to significant visual differences between the generated forged facial images and real images, these adversarial perturbations can only provide relatively effective protection for specific depth face forgery models known during the training process, and are difficult to covers multiple threats. In order to solve this problem, this article proposes a scheme to train multiple models against perturbation, and the internal structures of the multiple models added to the training are widely different, so that the resulting adversarial perturbation can universally protect facial images from a variety of known and unknown forged models. First, we established a multi-model parallel attack pipeline to conduct adversarial attacks on multiple deep face forgery models at the same time, continuously strengthening the resistance to each model. Secondly, we innovatively design a saliency-biased fusion strategy, aiming to alleviate the adversarial perturbation conflicts caused by different pipeline branches. In addition, in order to improve the generalization ability of the model and better simulate actual image shooting scenarios, we also use data enhancement technology in the original images. Finally, after the parallel attack pipeline branches are fused to combat perturbations, the heuristic tree-shaped PARZEN method is used to automatically find the optimal attack step size to further alleviate the compatibility problem of perturbations between models. Ex-

通讯作者: 王金伟, 博士, 教授, Email: wjwei_2004@163.com。

本课题得到国家自然科学基金(No. 62072250, No. 62172435, No. U1804263, No. U20B2065, No. 61872203, No. 71802110, No. 61802212); 中原科技创新领军人才项目(No. 214200510019); 国家重点研发计划(No. 2021QY0700); 江苏省自然科学基金(No. BK20200750); 河南省网络空间态势感知重点实验室开放基金(No. HNTS2022002); 江苏省研究生研究与实践创新项目(No. KYCX200974); 山东省计算机网络重点实验室开放课题基金(No. SDKLCN-2022-05); 人文社会科学教育部项目(No. 19YJA630061); 国家级大学生创新创业训练计划支持项目(No. 202310300021Z)资助。

收稿日期: 2023-12-14; 修改日期: 2024-03-07; 定稿日期: 2025-06-10

tensive experimental verification shows that the multi-model adversarial perturbation we proposed can effectively deface forged face images generated by multiple deep face forgery models, and achieves a good balance in robustness, efficiency and generalization.

Key words multi-model general image perturbation; adversarial attacks; deepfake; machine learning

1 引言

最近,生成式对抗网络及其改进方案^[1-3]在虚拟内容生成领域取得了显著进展,为经济和娱乐创造了丰富的价值。然而,深度人脸伪造^[4-7],即基于深度学习的人脸修改网络,利用生成对抗网络(GAN)生成虚假的目标人脸或者目标人脸属性的内容,若使用不当,可能极大地危害个人隐私安全和社会声誉。例如,虚假伪造的图像或者视频^[8]可以展示一个人从未说过的话或从未做过的事,特别是涉及国家政治问题^[9],那么对他或她的声誉会造成严重的影响。同样,设计伪造出带有目标属性的虚假人脸图像,可能会欺骗商业应用的生物认证系统^[10],潜在破坏个人的财产安全。最新的人工智能(AI)诈骗^[11]可以使用个人图像伪造出任意的面部动作,在社交软件与他人进行面对面自如地聊天,骗取钱财。因此,为防御深度伪造来的威胁,不仅仅需要对伪造后的图像进行必要的污染,降低其视觉质量和可信度,帮助个人将其与真实的图像区分开,还需要保证该伪造的人脸图像无法通过生物认证系统的活体检测,以对个人财产安全进行有效保护。

目前主流且常见的方法是被动取证技术,它通过训练人脸伪造篡改检测网络^[12-14],来评估输入图像被深度伪造模型篡改的概率。然而,它的局限在于仅能对已经发生伪造的图像进行取证分析,属于一种事后补救措施,并不能从根本上阻止伪造行为的发生。近期,文献[15-16]提出使用对抗性扰动来对抗深度伪造模型的攻击,通过向人脸图像添加对抗性扰动,使得伪造后的图像呈现异常的输出污点。这种方法是一种主动保护人脸图像的方案,旨在提前防范恶意深度伪造。

在图 1 中展示了这两种技术的示意比较,主动防御可以防御潜在的伪造威胁,但是目前主要的研究仅仅集中在生成特定模型的对抗性扰动,意味着该种扰动只能保护面部图像免受特定深度人脸伪造模型的攻击。于是,本文提出一种有效的改进措施,它使用大量人脸图像进行多模型并行对抗训练,最终制作一种多模型对抗扰动(Multiple Model Adversarial Perturbation),简称为 MMA,以保护人脸图像免受多个伪造模型的影响。图 2 中,首先,为了使图像多样化且更模拟现实场景,对原始图像进行数据增强处理,接着,将处理后的图像并行输入到多模

型并行深度伪造模块中,使用基于小步迭代的模型攻击方法^[17](PGD)生成该轮迭代中各个伪造模型的对抗扰动,其次,为了削弱不同模型生成的对抗性扰动之间的冲突,提出多模型攻击过程中的显著性偏向融合策略,最后,为进一步削弱前后轮生成的对抗扰动之间的冲突,增强生成的对抗性扰动的泛化性,在每轮的迭代完成后,均使用树形 PARZEN 算法^[18]自动寻找适合不同模型的攻击步长。

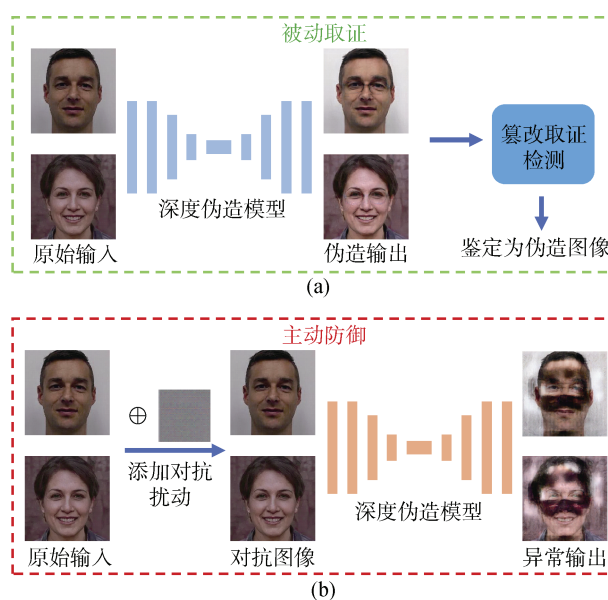


图 1 被动取证(图 a)和主动防御(图 b)流程比较
Figure 1 Comparison of passive forensics (Figure a) and active defense (Figure b) processes

此外,本文还在文献[19]的评估方法基础上进行了借鉴和改进。文献中的方法通过计算原始输出和异常输出之间的欧式距离来评估图像失真程度,这一方法的平均策略存在一个局限,即忽略了只修改少数属性的深度伪造方法,例如 HiSD^[20]中只添加眼镜的修改。这样的情况下,测量出的失真程度将会被其他未改变区域的平均值所掩盖。因此,论文进一步提出了使用修改掩模的方法,更加聚焦于实际被修改的区域。此外,本方法为确保伪造后的图像失真程度无法通过生物认证系统的活体检测,使用 FID^[21]图像生成伪造指标来衡量图像生成的质量,并结合置信度分数和活体检测模型的通过率来评估异常输出的生物学特征。这种综合的评估方法有助于更全面地评价伪造图像质量和生物学特征合法性。

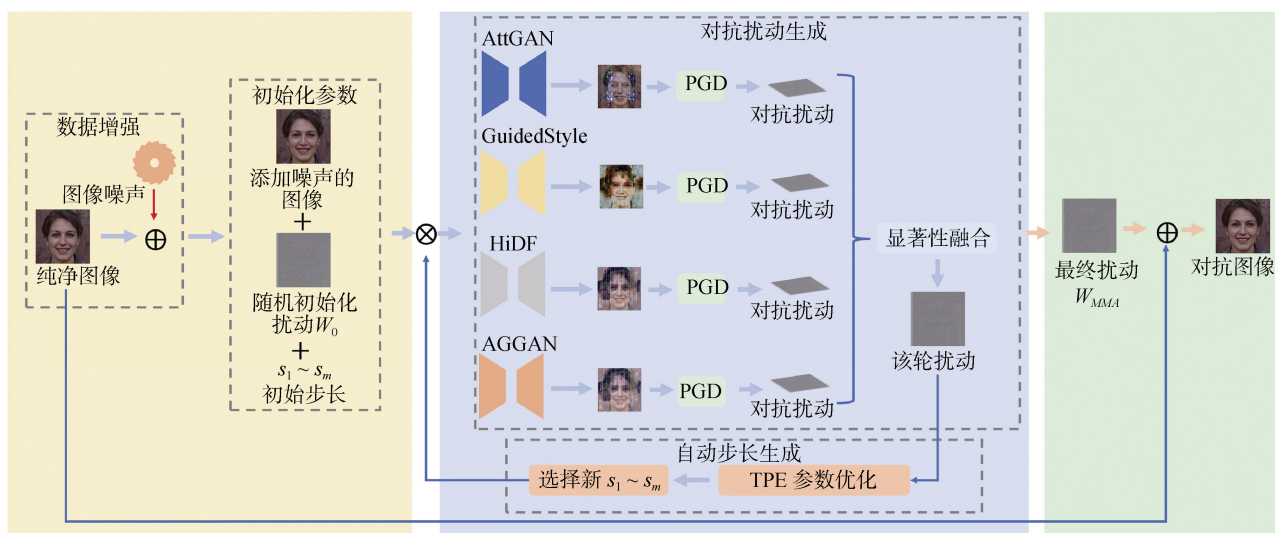


图2 在多个人脸修改网络上进行多模型对抗攻击的整体流程

Figure 2 The overall process of multi-model counterattack on multiple face modification networks

综上所述, 本文的主要贡献可以概括如下:

1) 引入了一种名为多模型对抗性扰动(MMA)的方法, 旨在保护人脸图像免受多种深度模型的伪造。通过大规模处理人脸图像并输入多模型并行训练网络, 得到了具备广泛适用性的最终对抗性扰动, 有效用于人脸图像的全面保护。

2) 使用了显著性偏向融合策略来缓解不同对抗性扰动之间的冲突, 进一步提高了 MMA 扰动的泛化性和鲁棒性。这一策略有助于确保扰动在不同伪造模型之间的兼容性, 并提升其在多种情况下的有效性。

3) 对多模型优化过程进行了深入细致的分析, 并采用了自动步长调整算法 TPE。这一优化手段旨在找到适合不同模型的攻击步长, 加速模型收敛到全局最优, 提升整体效率。

2 相关工作

2.1 生成模型对抗攻击

在以往的研究中, 对深度神经网络进行了广泛的对抗攻击探索。例如, Goodfellow 等人^[22]首先提出一步梯度攻击的快速梯度符号方法(FGSM), Dong 等人^[23]进一步发展了基于动量的快速梯度符号方法(M-FGSM), Xie 等人^[24]则提出的基于随机变化的 M-DI2-FGSM。此外, 针对对抗攻击中的超参数调整、梯度混淆和屏蔽问题, Croce 等人^[25]改进了投影梯度下降 PGD 攻击方法, 提出 AutoPGD。在图像场景转换和翻译任务上的研究中, Wang 等人^[26]和 YeH 等人^[27]对 CycleGAN^[28]和 pix2pixHD^[29]进行了测试, 然而, 这些任务主要涉及两个域的图像转换, 因此

攻击的成功率相对较高。最近的研究中, Ruiz 等人首次解决了条件图像翻译网络的攻击问题, 但所生成的对抗性扰动只能针对特定类型的图像免受特定深度伪造模型的攻击。这意味着为了应对每个深度伪造模型, 需要进行针对性的训练, 这增加了时间和人工操作成本。

2.2 多模型对抗性扰动

Moosavi-Dezfooli 等人^[30]在 2017 年首次提出了对抗性扰动的概念, 证明了一个良好训练的深度模型只需要一个对抗性扰动即可轻易失效。借鉴这项工作, Li 等人^[31]在图像检索系统中首次应用了通用对抗性攻击, 导致系统返回完全不相关的图像。另外, Ruiz 等人^[32]提出了泄露通用扰动算法(LUP), 旨在减少图像翻译网络所需的查询数量, 但其仅适用于破坏单一类型的图像翻译模型。针对单一模型带来的限制, 有学者提出了多模型对抗性扰动的概念。这种扰动指的是一个单一、经过良好训练的对抗性扰动能够添加到多个人脸图像中, 以欺骗多种深度伪造模型。Huang 等人首次探索了跨模型对抗性攻击, 可防御模型中植入的四种人脸图像深度伪造类型, 但采用了串行迭代融合策略, 存在计算成本高、需压缩训练集大小等问题, 并未能完全模拟真实场景。本研究提出了多模型并行训练方案, 并在训练过程中巧妙设计多种扰动融合策略, 以确保有效对抗多种攻击的同时, 大幅缩短训练时间。

3 提出方法

在本节中, 我们首先将对本文提出的多伪造模型攻击算法的整体框架进行概述。接着, 将详细介绍

在原始图像中采用的数据增强技术。紧接着, 我们将深入探讨多伪造模型并行攻击的算法细节。随后, 将探讨多模型扰动优化中的关键问题, 即如何确保对抗扰动对多伪造模型的泛化性, 在这一部分, 我们将介绍每轮迭代中的显著性偏向融合策略。最后, 我们将提出一种自动搜索各个深度伪造模型攻击步长的调整优化算法。这些内容全面展示本文所提方法的框架和关键细节, 为对抗性扰动的多模型优化提供清晰的方法论支持。

3.1 模型整体框架

本文提出的 MMA 扰动使用并行融合训练的策略, 可以保护任意人脸图像免受多种深度伪造模型的攻击。整体的框架如图 2 所示, 主要分为四个部分:

(1) 受到数据增强^[33-35]策略的启发, 该策略已经被证明可以有效的通过对训练图像集应用一组标签保持变换(比如裁剪、缩放和旋转等)来防止训练网络的过拟合。此外, 鉴于现有数据集中图像无法充分模拟真实拍摄场景(例如光照、视角、摄像机差异等因素引起的噪点), 我们对原始图像进行数据增强, 以增加图像的多样性;

(2) 为考虑训练图像集大小对最终对抗性扰动效果及训练时间的影响, 我们采用了多个伪造模型并行分支的结构。通过并行对抗攻击, 每个分支制作相应的对抗扰动;

(3) 为实现对多伪造模型的泛化, 我们提出了显著性偏向融合策略, 用于评估和选择每个分支生成的对抗扰动;

(4) 每轮迭代的最后阶段, 对最终生成采用树形 PARZEN 算法(TPE)调整步长, 即自动选择新的攻击步长列 $s_1, s_2, \dots, s_m$ 。我们发现不同模型的攻击步长对最终生成的 MMA 扰动的泛化性具有重要影响, 因此进行了步长的自动优化选择。

3.2 纯净图像数据增强

数据增强已被证明是防止训练网络过度拟合的有效方法。同时, 文献[36-37]中的研究表明, 通过简单的图像变换操作, 对抗样本的恶意性可以被消除, 这表明经过变换的对抗图像可被视为良好样本进行优化。在这一启发下, 我们创造性地考虑到不同相机型号可能拥有不同传感器和图像处理引擎, 因而对光线敏感度存在差异, 导致长时间曝光产生不同的图像噪点。此外, 相机的感光度、曝光时间、温度等因素也会影响噪点的数量和分布。基于上述观察, 我们尝试对原始图像的光照强度、拍摄角度、图像硬件噪点以及 LAB 图像域的像素值进行随机数据增强处理。这样的处理可以更真实地模拟各种场景, 提高

训练模型的泛化性。我们的具体处理方法如算法 1 所示。

算法 1. 数据增强技术.

输入: n 批训练的人脸图像 $X_1, X_2, \dots, X_n$, 批处理大小 m , LAB 域内参数进行随机跳动方法 $\psi(\cdot)$, 随机对图像进行变换的方法 $TR(\cdot)$, 控制随机变换方法的图像操作参数取值集合 $choice \in \left\{ 'tr_lab', 'tr_light', 'tr_angle', 'add_noise' \right\}$

输出: 数据增强后的图像 $X'_1, X'_2, \dots, X'_n$

```

1  随机初始化初始扰动  $W_0$ 
2  FOR  $k = 1:n$  DO
3      FOR  $i = 1:m$  DO
4          IF  $choice == tr\_lab$  DO
5               $l, a, b = \text{rgb2lab}(X_i)$ 
6               $a' = a + \theta_a$ 
7               $b' = b + \theta_b$ 
8               $X'_i = \text{lab2rgb}(l, a', b')$ 
9               $\theta_a, \theta_b = \psi(\theta_a, \theta_b)$ 
10         ELSE
11              $X'_i = TR(X_i, choice)$ 
12         END IF
13     END FOR
14     随机对  $choice$  进行取值
15 END FOR
16 RETURN  $X'_1, X'_2, \dots, X'_n$ 

```

3.3 多伪造模型攻击

在并行训练的流程中, 首先将经过数据增强的一批图像 $X'_1, X'_2, \dots, X'_n$ 输入深度伪造模型 $P(\cdot)$ 中, 获得原始输出结果 $P(X'_1), P(X'_2), \dots, P(X'_n)$, 然后, 将初始扰动 W_0 和 $X'_1, X'_2, \dots, X'_n$ 一同输入到深度伪造模型 $P(\cdot)$ 中, 我们就得到初始的被污染输出 $P(X'_1 + W_0), P(X'_2 + W_0), \dots, P(X'_n + W_0)$, 接下来, 采用均方根误差来描述 $P(X'_1), P(X'_2), \dots, P(X'_n)$ 和 $P(X'_1 + W_0), P(X'_2 + W_0), \dots, P(X'_n + W_0)$ 之间的差异性, 目标是找到满足公式(1)成立的扰动 W_{MMA} 。

$$\max \sum_{i=1}^n RMSE(P(X'_i), P(X'_i + W_i)) \quad (1)$$

$$\text{s.t. } \|W_i\|_{\infty} \leq \epsilon$$

ϵ 值被设定为一个阈值, 目的在于避免 W_i 过大,

如果没有对 W_i 加以限制, 最终的扰动添加到原始图像中可能会导致人眼能够轻易地分辨出差异, 从而无法实现良好的隐蔽性。这里的 X' 表示经过数据增强后的人脸图像, 为了简化符号书写和表示, 在本文后续的公式中, 默认 X 代指 X' , 特此说明。

最后, 鉴于 PGD 迭代小步长攻击方法在攻击单个伪造模型时具有较高的成功率, 我们采用了该方法, 详见公式(2)。

$$\begin{aligned} X_0^{adv} &= X_0 + W_0 \\ X_{r+1}^{adv} &= \text{clip}_{X, \epsilon} \{X_r^{adv} + T\} \\ \text{where } T &= s \cdot \text{sign} \left(\nabla_X \text{Loss} \left(P(X_r^{adv}), P(X_r) \right) \right) \end{aligned} \quad (2)$$

在这里, X_r^{adv} 表示经过第 r 次迭代后的对抗人脸图像, s 是基础攻击步长序列, Loss 为损失函数, 这里选择的是均方根误差(RMSE), $P(\cdot)$ 表示深度模型伪造操作, 而 clip 操作则确保 X_r^{adv} 的范围约束在 $[X-\epsilon, X+\epsilon]$ 之间。

通过上述过程, 可以获得单一模型的对抗扰动, 用于保护面部图像免受特定深度伪造模型的影响。然而, 显而易见的是, 单一分支精心制作的对抗扰动并不充分, 因为它们缺乏泛化性。在本文的并行多伪造模型攻击背景下, 我们的目标是制作能够抵御多种伪造模型的对抗性扰动。因此, 在接下来的两个部分中, 我们将重点介绍如何实现泛化融合。首先, 我们给出了多伪造模型并行攻击的算法, 具体如算法 2 所示。

3.4 显著性偏向融合

本文提到的泛化融合主要涵盖两个方面, 方便起见, 我们将其称为横向融合和纵向融合。横向融合指的是利用显著性偏向融合策略对多个并行分支中的对抗性扰动进行融合。由于不同深度伪造模型的内部结构和参数各不相同, 因此它们生成的对抗性扰动可能存在冲突, 降低了 MMA 扰动的可移植性。为了减弱训练过程中的这种冲突和不平衡性, 我们提出了显著性偏向融合策略。具体来说, 在每轮的训练中, 从每个分支产生的对抗性扰动中选择损失函数 Loss 值小于设定阈值的部分, 认为这部分对抗性扰动效果不理想。然后, 逐轮进行异或操作, 得到该轮的最终融合结果, 详细内容见公式(3)。

算法 2. 多伪造模型并行攻击.

输入: n 批数据增强后训练后待训练的人脸图像 $X_1, X_2, \dots, X_n$, 给人脸图像添加对抗性扰动的方

法 $DK(\cdot)$, 添加了对抗扰动的人脸图像 X^{adv} , m 种伪造模型 $P_1, P_2, \dots, P_m$, 伪造模型攻击算法步长序列 $s_1, s_2, \dots, s_m$, 深度模型进行伪造并返回对抗性扰动的基础性方法 $F(\cdot)$ (详见公式(2)), 返回每一轮显著性偏向融合选择和处理结果方法 $R(\cdot)$ (详见公式(3)), 返回前后轮扰动的融合结果方法 $QH(\cdot)$ (详见公式(4))

输出: 最终的扰动 W_{MMA}

```

1  随机初始化初始  $W_0$ 
2  FOR  $k = 1 : n$  DO
3       $X_k^{adv} = DK(X_k, W_{k-1})$ 
4       $G_k = F(P_1(X_k^{adv}) \dots P_m(X_k^{adv}), s_1 \dots s_m)$ 
5       $G_k^{avg} = R(G_k)$ 
6       $W_k = QH(W_{k-1}, G_k^{adv})$ 
7  END FOR
8   $W_{MMA} = W_n$ 
9  RETURN  $W_{MMA}$ 

```

$$G_k^{avg} = R(G_k) = \frac{\sum_{i \in j} W_{k-1} \otimes X_k^i}{j_{size}} \quad (3)$$

$$\text{where } j = \arg(\text{Loss}(X_{k \rightarrow j}^{avg}, P(X_{k \rightarrow j}^{avg})) < \delta)$$

这里的 j_{size} 表示效果不佳序列 j 的规模, 符号 $\otimes$ 表示像素级别的异或操作, δ 是设定的损失函数阈值, $P(\cdot)$ 表示进行深度模型伪造, X^{adv} 代表添加了对抗扰动的人脸图像, G_k^{avg} 表示第 k 轮中得到对抗扰动的平均值。公式(3)的处理达到了两个关键目标, 一是使最终的扰动结果更多聚焦于多个伪造模型, 而非单一模型, 二是显著性地偏向选择伪造效果较差的扰动, 对它们进行融合, 以便在下一轮的迭代中更多地关注这些区域。

最后, 进行前后轮次间的对抗性扰动融合(即纵向融合)后, 得到的最终扰动 W_{MMA} 是之前每一轮融合叠加后的结果。公式(4)是纵向融合的代表形式。

$$\begin{aligned} G_0^{avg} &= W_0 \\ W_k^{mma} &= \alpha W_{k-1}^{mma} + (1 - \alpha) G_k^{avg} \end{aligned} \quad (4)$$

α 是权重系数, 这里设置为 0.8, 旨在避免某一轮结果对总体扰动的的影响过大, G_k^{avg} 是第 k 轮得到的对抗性扰动结果, W_{k-1}^{mma} 是前 $k-1$ 轮得到的扰动叠加结果, 通过加权融合, 得到下一轮的初始扰动 W_k^{mma} , 直至获得符合要求的 W_{MMA} (求解过程详见公式(1))。

3.5 树形 PAEZEN 自动步长优化算法

在显著性偏向融合过程中, 我们发现对抗性攻击 PGD 步长 s 的选择对最终得到的对抗性图像质量有显著影响, 我们引入了文献[6]中提到的树形 PARZEN 自动优化步长算法, 旨在自动寻找最适合的攻击步长。

首先, 因为我们选择的基础攻击方法 PGD 是属于 FGSM 族的, 梯度的 $\nabla_x Loss$ 是有符号函数归一化的, 具体的符号函数由公式(5)定义:

$$\|sign\ x\|_2 = \begin{cases} 1, & x \neq 0 \\ 0, & x = 0 \end{cases} \quad (5)$$

在实际的计算过程中, 我们的目的是让伪造后的人脸图像尽可能偏离正常范畴, 因此, 通常情况下, $\nabla_x Loss$ 的值几乎不会达到 0, 在这种情况下, 对于任何梯度, $\|\Delta P\|_2 = s \cdot \|sign(\nabla_x Loss)\|_2 \approx s$ 基本成立。

上述等式说明攻击过程中仅仅有 s 步长序列影响了梯度更新大小, 于是对 s 的选择对攻击成功率的影响很大。同样的道理适用于多模型攻击, 不同的是多模型攻击迭代中更新后的每轮扰动 ΔP_{more} 由多个模型 $\Delta P_1 \dots \Delta P_m$ 组合而成, 具体如公式(6)所示。

$$\Delta P_{more} = \sum_{i=1}^m w^{(m-i)} \Delta P_i = \sum_{i=1}^m w^{(m-i)} s \cdot sign(\nabla_x Loss) \quad (6)$$

在上式中, m 为模型个数, 权重参数 w 为初始设定的常数, 符号 $\nabla_x Loss$ 为 P_i 的优化方向。因此, 整体的优化方向受到 $s_1, s_2, \dots, s_m$ 的支配。如何在多种不同的模型中找到理想的优化步长是多模型对抗扰动的问题。

所以, 本文引入 TPE 方法来解决这个问题, 通过自动搜索合适的 $s_1, s_2, \dots, s_m$, 以平衡多个模型计算出的不同方向。TPE 是一种基于顺序模型优化的超参数优化方法, 该方法可以基于历史的参数测量结果, 依次构建模型来近似模拟超参数的性能, 然后根据该模型选择新的超参数进行测试。在对抗训练任务中, 我们将 $s_1, s_2, \dots, s_m$ 统一考虑为超参数 x , 攻击成功率设置为 TPE 的关联质量分数 y , TPE 方法使用两个密度函数来定义 $P(x|y)$, 详见公式(7):

$$P(x|y) = \begin{cases} \zeta(x), & \text{if } y < y^* \\ g(x), & \text{if } y > y^* \end{cases} \quad (7)$$

其中 y^* 由历史上最好的测量值确定, 也即为阈值, 密度函数 $\zeta(x)$ 是由观测空间 $\{x^{(i)}\}$ 建立的, 该观测空间对应的是 $f(x^{(i)})$ 小于 y^* , 再使用剩余观测空间建

立 $g(x)$ 。基于 TPE 的方法取决于比当前所观察到的最好的 $f(x)$, 以得到更大的 y^* , 这样的话我们使用一些观测点就可以用来建立 $\zeta(x)$ 。

TPE 中使用期望改进(EI)作为采集函数, 其他的基准相比, 是更加直观的, 而且被证明具有更加优越的性能, 公式(8)定义:

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y|x) dy \\ = \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)p(y)}{p(y)} dy \quad (8)$$

对 $p(x)$ 的定义如下公式(9):

$$p(x) = \int p(x|y)p(y)dy = \gamma\zeta(x) + (1-\gamma)g(x) \quad (9)$$

其中 $\gamma = p(y < y^*)$, $\gamma \in (0,1)$, 主要用于划分 $\zeta(x)$ 和 $g(x)$, 实际的含义表示为构建好分布的最佳观测值的占总观测值的比例为 γ 。

将 $p(x)$ 带入, 可以将期望改进 EI 化简为如公式(10):

$$EI_{y^*}(x) = \frac{\gamma y^* \zeta(x) - \zeta(x) \int_{-\infty}^{y^*} p(y) dy}{\gamma \zeta(x) + (1-\gamma)g(x)} \\ \propto \left(\gamma + \frac{g(x)}{\zeta(x)}(1-\gamma) \right)^{-1} \quad (10)$$

公式(10)表明, 为了最大化 EI, 我们希望在点 x 下 $\zeta(x)$ 高的同时 $g(x)$ 的概率低, 在每次迭代中, 算法返回具有最大 EI 的候选 x^* , 定义在公式(11):

$$x^* = \arg \max EI_{y^*}(x) \quad (11)$$

4 实验结果及分析

在这一部分, 我们将首先详细描述所使用的数据集, 包括其来源、规模和特征。接着, 我们将详细列出实验的具体实现步骤, 包括数据处理、模型训练以及扰动生成的过程。之后, 我们会介绍选定的评估指标, 这些指标将评估对抗扰动的效果和性能。最后, 我们会全面呈现本文提出的 MMA 扰动的实验结果, 包括对多模型对抗性的效果评估和对比分析。

4.1 数据集和实验细节

实验中, 主要采用了 CelebA^[38] 和 CASIA - WebFace^[39] 两个数据集, 其中 CelebA 包含了近两万张人脸图像, 涵盖几乎所有人种和地域的人脸类别, 而 CASIA-WebFace 包含了五十万张互联网公共网站上的人脸图像, 覆盖了更多姿势变化和杂乱场景。我

们从这两个数据集中各选择了 1 万张图像, 其中 7,000 张用于训练图像的制作对抗性扰动, 其余的图像用于测试验证结果的说服力。此外, 我们还使用了索尼 A7 相机拍摄了 500 张真实场景下的人脸图像, 作为额外的实验图像, 验证 MMA 扰动在真实场景中的鲁棒性。

我们选取了开源人脸伪造网络, 包括 AGGAN、AttGAN、HiSD 以及 GuidedStyle。它们均在 CelebA 上进行了训练, 其中, HiSD 针对性训练了一种给人脸添加眼镜的模型, 针对性比较强, 而 GuidedStyle 采用了新的学习框架, 利用从人脸属性分类器中学习到的知识来指导 StyleGAN 的图像生成过程, 因此网络结构更复杂, 难度更大。

在步长的搜索过程中, 设定了最大迭代次数为 2000, 并将每个模型的步长搜索空间设置为[0,10]。首先使用 batch size 为 16 的步长, 然后对多模型扩展到 64 进行搜索。

4.2 评估指标

我们设计了一种更加合理且全面的评估方法, 该方法主要专注于三个方面的指标。

首先, 针对原始输出和异常输出之间的差异分析, 文献[4]中使用 L^2 范数来衡量原始输出 $G(X)$ 和异常输出 $G(X+W)$ 之间的差异, 当 $\|G(X+W) - G(X)\|_2 > 0.1$ 时, 认为人脸图像成功被对抗性扰动保护。然而, 我们对该评估方法进行了研究发现, 在某些情况下其泛化性不够强, 因为它并未充分关注修改区域。因此, 引入了下述公式(12)的掩模矩阵。

$$Mask_{(i,j)} = \begin{cases} 1, & \text{if } \|G(x)_{(i,j)} - X_{(i,j)}\|_2 > 0.5 \\ 0, & \text{ELSE} \end{cases} \quad (12)$$

这里的 (i,j) 表示图像中的像素点坐标, 使用掩模矩阵之后, 计算 L_{mask}^2 只会计算变化较大的区域, 而忽略其他区域。

$$L_{mask}^2 = \frac{\sum_i \sum_j Mask_{(i,j)} \cdot \|G(X+W_{MMA})_{(i,j)} - G(X)_{(i,j)}\|_2}{\sum_i \sum_j Mask_{(i,j)}} \quad (13)$$

在实验中, 使用 L_{mask}^2 来判断图像是否被成功保护, 若其值大于 0.1, 则确定图像保护成功, 并采用成功率(Success rate, SR)来表示人脸图像保护的成功率。

此外, 使用 FID 来评估伪造人脸图像的生成质量, Fréchet Inception Distance(FID)是一种评价生成图片的质量的指标, 它综合考量了生成图片分布与

真实图片分布之间的距离。 FID 值越小, 表示生成图片的真实性越高。

最后, 为了验证伪造图像是否能够通过生物活体检测系统认证, 我们使用了开源的 HyperFAS^[40]进行测试。如果系统给出 HyperFAS 分数($HFAS-S$)大于 99%, 则可以认为该伪造图像通过了活体检测。

4.3 实验结果

4.3.1 鲁棒性实验

在本研究中, 我们通过一系列充分的实验验证了提出的 MMA 扰动的鲁棒性。

首先, 为了验证 MMA 扰动在伪造模型上的鲁棒性, 我们将对 CelebA、CASIA - WebFace 以及我们团队自建的 Photos500 图像集添加对抗性扰动。随后, 我们将使用多个深度人脸伪造模型对处理后的人脸图像进行深度伪造, 并对伪造后的结果进行定性和定量分析。

MMA 扰动的定性和定量结果分别呈现在图 3 和表 1 中。我们发现, 在 CelebA、CASIA-WebFace 以及我们自行采集的 Photos500 图像集上, 所提出的 MMA 扰动呈现出相似的性能表现。这些结果的呈现强调了 MMA 扰动在各种数据集上的一致性效果, 并突显了其鲁棒性。这些发现在验证我们提出的方法的可靠性和有效性方面起着关键作用。

在图 3 中, 通过虚线分隔, 左侧为经过人脸伪造模型处理后的原始图像结果, 右侧为经过人脸伪造模型处理后的对抗图像结果。具体而言, A 行是输入模型的原始图像和对抗图像, B 行至 E 行分别代表了使用不同的人脸伪造模型(AttGAN、AGGAN、GuidedStyle、HiSD)进行的伪造。观察图 3, 我们可以直观地发现, 在添加对抗性扰动 MMA 后, 对抗图像在视觉上与原始图像保持了高度的一致性, 然而, 这些图像却成功地防御了多种人脸伪造模型的伪造攻击。

从表 1 中可知, 在三个数据集上计算得到的 L_{mask}^2 均超过 0.1, 这表明它们都成功地进行了保护。观察保护成功率 SR , 对于 AGGAN 和 HiSD 这两种方法上效果良好, 接近 100%, 然而在 GuidedStyle 和 AttGAN 上略低, 这一现象的原因可能在于 GuidedStyle 对人脸的属性特征进行了精准编辑, 而我们更多关注的是共性特征。另外, AttGAN 包含多达 14 种人脸属性, 这种多样性特征对准确率有一定的影响。从图像失真输出指标 FID 来看, 在 AttGAN 和 HiSD 上值较高, 说明生成图像的质量较差, 导致 $HFAS-S$ 较低, 活体检测通过率大幅降低。而在 AGGAN 和 GuidedStyle 上, FID 稍高, 但使得图像

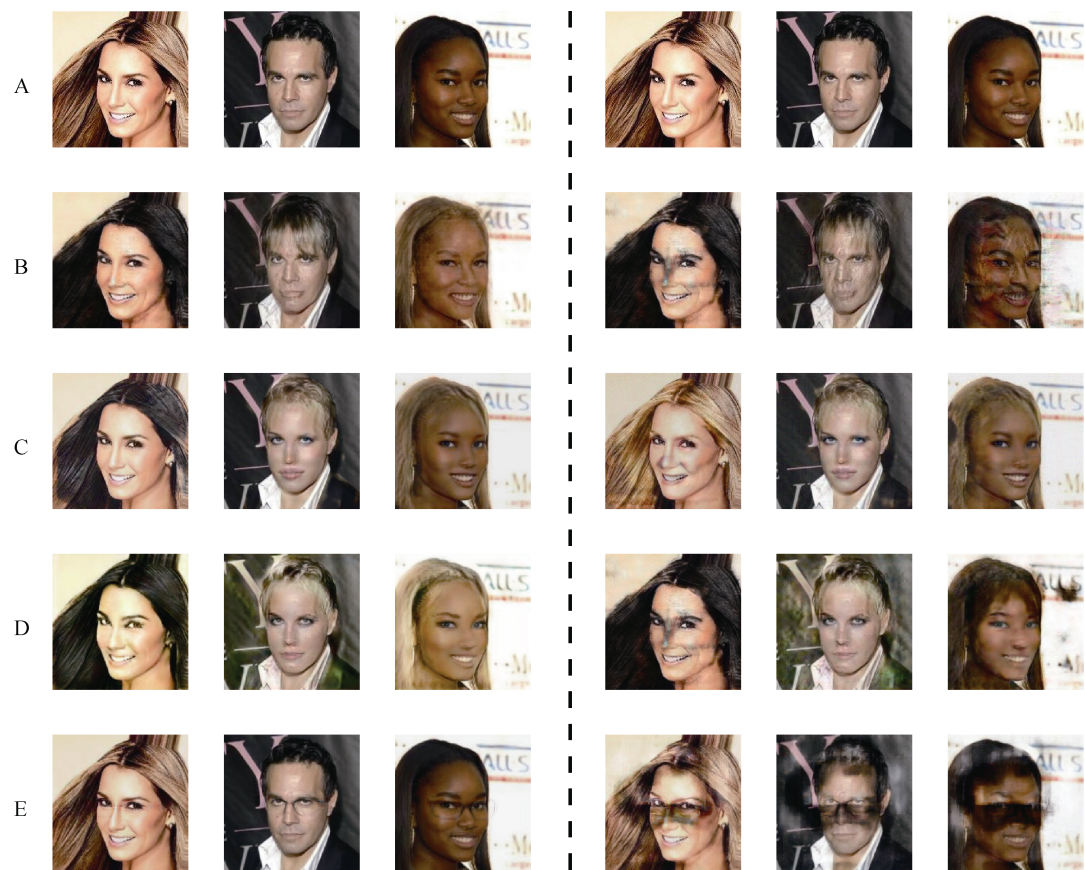


图 3 原始图像和对抗图像经过多种人脸伪造模型后的结果比较(其中 A 行左侧为原始图像, 右侧为对抗图像, B-E 行依次为使用 AttGAN、AGGAN、GuidedStyle 和 HiSD 对 A 行伪造后的结果)

Figure 3 Comparison of the results of the original image and the adversarial image after multiple face forgery models (where the left side of row A is the original image, the right side is the adversarial image, and rows B-E are the forged results of row A using AttGAN, AGGAN, GuidedStyle and HiSD in turn)

表 1 添加 MMA 扰动后对各个伪造模型的破坏效果

Table 1 Destruction effect on each forged model after adding MMA perturbation					
数据集	模型	$L^2_{mask} \uparrow$	$SR \uparrow$	$FID \uparrow$	$HFAS-S \downarrow$
CelebA	GuidedStyle	0.11	95.36%	75.22	36.85%
	AGGAN	0.19	99.83%	56.32	56.24%
	AttGAN	0.14	93.20%	78.73	35.27%
	HiSD	0.24	99.96%	101.41	4.21%
CASIA-WebFace	GuidedStyle	0.13	95.45%	86.31	31.87%
	AGGAN	0.20	99.88%	78.64	47.22%
	AttGAN	0.15	95.24%	143.86	31.27%
	HiSD	0.27	100.00%	197.12	8.01%
Photos500	GuidedStyle	0.14	96.62%	113.67	30.63%
	AGGAN	0.21	99.97%	96.70	66.21%
	AttGAN	0.17	96.34%	217.88	21.38%
	HiSD	0.29	100.00%	244.54	10.07%

具备欺骗活体检测系统的能力。此外, 在 CelebA 数据集上的表现相对于 CASIA-WebFace 和 Photos100 数据集较差, 因为这些伪造方法均基于 CelebA 进行

训练, 对该数据集具有更强的鲁棒性。尽管如此, 所有数据集上的异常输出均显示出较高的 FID , 表明它们的伪造结果较差。

定性和定量的结果均表明 MMA 扰动能够成功保护人脸图像免受多种伪造模型的攻击。

接着, 为了评估训练得到的对抗扰动对不同图像处理攻击的鲁棒性, 我们先从 CelebA 数据集中随机选取 1000 张测试图像, 再将对抗扰动添加到测试图像中, 然后将添加了扰动的测试图像分别进行五种常见的图像处理攻击, 包括 JPEG 压缩(质量因子为 100,75,50), 高斯噪声(方差为 3,5,7), 高斯模糊(模糊核大小为 3,5,7), 缩放(缩放比例为 75%, 50%, 25%)和中心裁剪(裁剪比例为 10%, 30%, 50%)。最终得到添加了 MMA 扰动的测试图像在经过图像处理攻击后对四种伪造模型(AttGAN、AGGAN、GuidedStyle、HiSD)的平均保护成功率 SR_{avg} 和平均图像失真损失 FID_{avg} 。

测试结果如表 2 所示, 平均保护成功率 SR_{avg} 和平均图像失真 FID_{avg} 在所有的攻击强度下的表现均较好。但是, 在高损失压缩, 高强度噪声以及中心裁剪比例过大时, 它们的表现有所降低, 我们认为高损失压缩和高强度噪声对含有对抗扰动的图像的统计分布特征造成干扰, 使图像的真实性遭到破坏。另外, 分析 CelebA 数据集可知, 其中的人脸区域大部分集中在整个图像的中心, 当过度的中心裁剪对中心人脸的关键部分破坏后, 会导致结果的下降。这些观察结果为我们进一步理解对抗性扰动在不同条件下的鲁棒性提供了宝贵的参考。

尽管表现存在轻微的下降, 但平均保护成功率 SR_{avg} 仍然保持不低于 85%, 平均图像失真 FID_{avg} 低于 40, 这表明能够对面脸图像进行有效的保护。我们分析对抗扰动可以保持对不同图像处理攻击鲁棒性的原因是在训练阶段已经对原始训练数据进行

增强, 这使得模型训练得到的对抗扰动对图像处理具有抗性。此外, 在训练阶段采用的并行分支结构也为模型提供了更大的图像维度扩展, 使得对抗扰动能够学习到更多复杂的图像特征, 从而能够在图像攻击时具有更好的鲁棒性。

最后, 为了验证本实验得到的对抗扰动在当前流行的对抗攻击方法中的优越性, 我们进一步将 MMA 扰动与在 CelebA 上流行的保护方法进行了比较, 包括 PGD, MIN, M-DI2-FGSM, AutoPGD, CMUA, 将这些方法调整为相同的硬件基础设置。表 3 列举了 MMA 扰动的保护成功率 SR 和图像失真输出指标 FID 与这些方法的比较结果。从保护成功率 SR 来看, 我们发现除了 CMUA 和我们的方法外, 其他的方法只能很好的针对单一模型进行保护, 说明它们缺乏很好的泛化性, 而 CMUA 和我们的方法均采用了多模型训练, 可以有效针对不同的伪造模型进行很好的保护, 但因为我们的方法中使用数据增强和显著性偏向融合等策略, 使结果比 CMUA 更加出色。再从图像失真输出指标 FID 来看, 我们提出的方法 FID 的值大多都较大, 说明添加了我们制作的对抗扰动后再进行伪造, 真实性得到了极大破坏, 值得关注的是, MIM 方法对于 AGGAN 的破坏效果更加好, 这是因为 AGGAN 的模型参数更为针对性地被 MIM 学习到, 而我们的方法更多关注共性特征, 会忽略一些特殊性质。最后, 从 SR 和 FID 的关系来看, 基本呈现出数值正比对应的关系, 但也存在一些特殊情况, 例如, PGD 制作的对抗扰动在 HiSD 伪造下的保护成功率只有 0.003, 但是 PGD 的方法却可以使 HiSD 模型生成相对虚假的人输出, 这说明我们需要综合考虑两个指标, 从不同的角度来判断一个方法的好坏, 而我们的方法在两个指标下均表现较好。

表 2 在 CelebA 数据集上 MMA 扰动对不同图像处理的鲁棒性评测结果

Table 2 Robustness evaluation results of MMA perturbation on different image processing on CelebA dataset				
JPEG 压缩	未处理	质量因子=100	质量因子=75	质量因子=50
SR_{avg} / FID_{avg}	97.09% / 77.92	95.19% / 70.84	92.52% / 65.71	86.20% / 60.66
高斯噪声	未处理	方差=3	方差=5	方差=7
SR_{avg} / FID_{avg}	97.09% / 77.92	94.81% / 67.22	91.35% / 61.04	85.34% / 56.38
高斯模糊	未处理	模糊核大小=3	模糊核大小=5	模糊核大小=7
SR_{avg} / FID_{avg}	97.09% / 77.92	96.12% / 65.81	92.40% / 59.22	88.63% / 55.70
缩放	未处理	缩放比例=75%	缩放比例=50%	缩放比例=25%
SR_{avg} / FID_{avg}	97.09% / 77.92	96.54% / 72.58	94.38% / 68.22	90.07% / 61.36
中心裁剪	未处理	裁剪比例=10%	裁剪比例=30%	裁剪比例=50%
SR_{avg} / FID_{avg}	97.09% / 77.92	95.26% / 73.25	93.47% / 67.61	85.03% / 59.10

表 3 本方法与其他流行方法比较结果

Table 3 Comparative results of this method with other popular methods

方法	SR ↑				FID ↑			
	GuidedStyle	AGGAN	AttGAN	HiSD	GuidedStyle	AGGAN	AttGAN	HiSD
PGD	0.175	0.998	0.217	0.003	15.32	45.07	17.58	34.72
MIM	0.059	1	0.017	0.044	18.14	69.74	16.09	33.15
M-DI2-FGSM	0.004	0.999	0.023	0.005	17.01	46.57	16.35	17.75
AutoPGD	0.001	0.996	0.040	0.005	14.52	48.51	18.89	18.78
CMUA	0.713	1	0.871	1	54.89	50.96	65.06	92.73
我们的方法	0.954	1	0.932	1	75.22	56.32	78.73	101.41

4.3.2 训练时间实验

在训练时间方面，我们的方法显著优于目前已知在多模型保护上表现良好的 CMUA 方法。CMUA 方法采用串行多轮迭代的训练方式，需要对每个伪造模型的对抗性扰动输出进行融合。对于 n 轮训练 m 个伪造模型，时间复杂度为 $O(mn)$ 。相比之下，我们提出的方法的时间复杂度仅为 $O(n)$ 。而且，通过式(3)可知，我们的方法采用的是差额选取融合，而不是像 CMUA 那样全部融合。所以，考虑到训练时间，CMUA 方法仅使用了 128 张图像来制作对抗性图像，因此对人脸边缘关键区域的捕捉不够详尽，在多属性伪造方法(比如 AttGAN)下表现不佳。相比之下，我们采用并行方法，在相同的环境下，对 7000 张图像进行训练的时间仅为 CMUA 的 1.46 倍。这意味着我们的方法可以在更短的时间内，针对更多的图像进行更全面的保护。

4.3.3 泛化性实验

在本节中，我们使用几种开源和商业应用的深度人脸伪造项目及软件对含有我们对抗扰动的图像进行攻击伪造，其中包含以 GAN 学习生成技术为基础的 StarGAN、PGAN^[41]和 Conditional-GAN^[42]，对人脸进行风格属性迁移的 Face2Face^[43]以及商业化使用的 APP 软件 Zao^[44]和 FakeAPP^[45]。

在实验中，我们首先从三个数据集上各选取了 100 张纯净的图像，并将这些图像添加前期已经训练较好的对抗扰动，再将添加了扰动的载体图像分别通过这六种人脸伪造的项目和软件。最后，使用破坏成功率对伪造后的图片结果进行分析，具体来说，当 SR 的值大于 85%且 FID 不低于 40 时，我们认为 是成功攻击的一个样本。

如表 6 所示，我们对 6 种人脸伪造项目和软件进行了 300 组样本的测试，得到了破坏成功率。我们发现，在风格属性迁移和以 GAN 为基础的攻击手段中，取得了尤其突出的表现。而对于 Zao 和 FakeAPP，由

于其商业化的使用，我们暂未能完全了解其内部的具体技术。我们推测破坏成功率略低的原因可能是它们内部网络结构中存在平滑和加噪预处理步骤，对我们预先加入的对抗噪声存在微量抑制作用。尽管如此，对于未知的人脸伪造项目和软件，我们的对抗扰动均能达到超过 85%的破坏成功率，这充分证明了我们对抗扰动的泛化性。

4.3.4 消融实验

在本节中，首先研究我们提出的数据增强技术对图像保护有效性，如表 4 所示，当不使用数据增强技术时，保护成功率的结果有明显的下降，对活性检测成功率也有一定的影响。这表明采用数据强技术确实有助于提高图像保护的有效性，对我们提出的方法来说是一项重要的预处理步骤，值得采用。

此外，我们研究了显著性偏向融合策略中损失函数阈值 δ 对平均保护成功率 SR_{avg} 和训练时间的影响。如表 5 所示。为了直观地表示时间的大小对比，将 $\delta = 0.01$ 时的时间设置为理论量上时间的最小值 a (分钟)。从表 5 的结果可以看出，随着损失函数阈值的增加，平均保护成功率呈现先上升后下降的趋势，这是因为阈值过小时，训练到的数据量非常有限，无法学习到有价值的内容，但是过大的阈值则几乎不对数据进行筛选，不能充分考虑到训练效果不太好的那一部分模型，表现平平。此外，值得关注的是

表 4 在 CelebA 数据集上应用数据增强技术的性能指标对比

Table 4 Comparison of performance indicators using data enhancement techniques on CelebA datasets

伪造模型	性能指标变化(箭头左侧未应用, 右侧应用)	
	SR ↑	HFAS - S ↓
GuidedStyle	87.69%→95.36%	41.31%→36.85%
AGGAN	97.35%→99.83%	67.73%→56.24%
AttGAN	90.67%→93.20%	40.52%→35.27%
HiSD	98.46%→99.96%	16.38%→4.21%

表 5 在 CelebA 上不同损失函数阈值 δ 对模型性能影响

Table 5 Effect of different loss function thresholds on model performance on CelebA

评价指标	损失函数阈值 δ				
	0.01	0.05	0.08	0.15	0.3
$SR_{avg} \uparrow$	56.38%	90.36%	97.09%	94.17%	88.62%
Time(min) $\downarrow$	a	$23a$	$107a$	$368a$	$819a$

表 6 未知伪造攻击下的破坏成功率

Table 6 Destruction success rate under unknown forgery attack

伪造项目和软件	破坏成功率(%)
StarGAN	98.33%
PAGN	96.67%
Condition-GAN	94.00%
Face2Face	95.33%
Zao	88.67%
FakeAPP	86.00%

训练时间的变化。我们惊奇地发现损失函数阈值微小的增加会导致运行时间迅速上升, 因此损失阈值的设置对运行时间的影响极大。这证明我们提出的方法中采用显著性偏向融合策略是必要的, 它对提高保护成功率和缩短训练时间都至关重要。

最后, 我们研究了采用扰动融合和自动优化步长的有效性。通过 3.5 节分析得知, 自动步长的调整会对于融合扰动而言类似于一个权重, 影响下一轮迭代对抗扰动对各个深度伪造模型的关注程度。所以非常值得探究 TPE 自动优化步长的必要性。

表 7 在 CelebA 上融合扰动和自动优化步长操作对模型性能的影响

Table 7 Effects of fusing perturbations and automatically optimizing step size operations on model performance on CelebA

方法	基础 PGD	对照 1	对照 2	本方法
融合扰动		√		√
自动优化步长			√	√
SR (GuidedStyle)	5.2%	34.7%	91.3%	95.36%
SR (AGGAN)	99.6%	89.6%	99.8%	99.83%
SR (AttGAN)	3.5%	81.2%	47.4%	93.20%
SR (HiSD)	0.1%	3.4%	10.8%	99.96%

如表 7 所示, 基础的 PGD 算法在 GuidedStyle, AttGAN 和 HiSD 上的性能非常差, 这说明其对模型的可迁移性较弱。在对照 1 中, 单独使用融合扰动策略, 模型在 AttGAN 上的性能有所提高, 但是, 在

AGGAN 上则下降 10%, 这表明为了实现少部分的泛化, 牺牲了部分模型的高性能。另一方面, 在对照 2 中, 如果只采用自动优化步长的策略, 而不在每轮迭代中将扰动进行融合, 每轮需要维护至少 4 种不同的对抗扰动, 这将导致处理时间的成本成倍增长, 而且只采用自动优化步长策略在 AttGAN 和 HiSD 上得到的结果也不够好。但是, 将两者结合, 我们发现最终的得到的 MMA 扰动在所有的深度伪造模型中都会有显著的提高。这表明, 融合扰动和自动优化步长策略对本方法都至关重要, 任何一个都是必不可少的。

5 结论

本文提出了一种多模型分支攻击网络结构, 用于制作能够保护大量人脸图像免受多种伪造模型攻击的对抗性扰动 MMA。具体来说, 我们引入了显著性偏向融合策略, 旨在缓解不同并行分支产生的对抗性扰动之间可能存在的冲突。在此基础上, 采用树形 PARZEN 自动调整步长算法, 灵活调整不同分支结构的优化趋势。除此之外, 在预处理阶段, 我们采用了精心设计的数据增强策略, 以增强原始数据的多样性。实验结果显示, 我们的 MMA 扰动能有效破坏深度伪造模型的恶意篡改操作, 显著降低了模型生成的伪造图像质量, 使得伪造后的人脸图像失去了实用价值。结论表明我们的方法能够在多模型攻击背景下取得有效成果。未来的工作可以进一步改进这种扰动方法, 探索其在更多应用领域的潜在价值和实际应用。

参考文献

- [1] Zhang B, Zhu J, Su H. Toward the Third Generation Artificial Intelligence[J]. *Science China Information Sciences*, 2023, 66(2): 121101.
- [2] Zador A, Escola S, Richards B, et al. Catalyzing Next-Generation Artificial Intelligence through NeuroAI[J]. *Nature Communications*, 2023, 14: 1597.
- [3] Sohail S S, Farhat F, Himeur Y, et al. Decoding ChatGPT: A Taxonomy of Existing Research, Current Challenges, and Possible Future Directions[J]. *Journal of King Saud University - Computer and Information Sciences*, 2023, 35(8): 101675.
- [4] Tang H, Xu D, Sebe N, et al. Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation[C]. *2019 International Joint Conference on Neural Networks*, 2019: 1-8.
- [5] 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.
- [6] Choi Y, Choi M, Kim M, et al. StarGAN: Unified Generative Ad-

- versarial Networks for Multi-Domain Image-to-Image Translation[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8789-8797.
- [7] Hou X X, Zhang X K, Liang H B, et al. GuidedStyle: Attribute Knowledge Guided Style Manipulation for Semantic Face Editing[J]. *Neural Networks*, 2022, 145: 209-220.
- [8] Cheng H, Guo Y Y, Wang T Y, et al. Voice-Face Homogeneity Tells Deepfake[J]. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2024, 20(3): 1-22.
- [9] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection[J]. *Information Fusion*, 2020, 64: 131-148.
- [10] Nirkin Y, Wolf L, Keller Y, et al. DeepFake Detection Based on Discrepancies between Faces and Their Context[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(10): 6111-6121.
- [11] Yang W Y, Zhou X Y, Chen Z K, et al. AVoid-DF: Audio-Visual Joint Learning for Detecting Deepfake[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 2015-2029.
- [12] Jia S, Ma C, Yao T P, et al. Exploring Frequency Adversarial Attacks for Face Forgery Detection[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 4093-4102.
- [13] Epstein D C, Jain I, Wang O, et al. Online Detection of AI-Generated Images[C]. *2023 IEEE/CVF International Conference on Computer Vision Workshops*, 2023: 382-392.
- [14] Ying Q C, Zhou H, Qian Z X, et al. Learning to Immunize Images for Tamper Localization and Self-Recovery[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(11): 13814-13830.
- [15] Ruiz N, Bargal S A, Sclaroff S. Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems[M]. *Computer Vision – ECCV 2020 Workshops*. Cham: Springer International Publishing, 2020: 236-251.
- [16] Li M J, Wu H Z, Wang Z C, et al. Generating Watermarked Adversarial Texts[J]. *Journal of Electronic Imaging*, 2023, 32(2).
- [17] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C]. *International Conference on Learning Representations*, 2018.
- [18] Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for hyper-parameter optimization[J]. *Advances in neural information processing systems*, 2011: 24.
- [19] Huang H, Wang Y T, Chen Z Y, et al. CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(1): 989-997.
- [20] Li X Y, Zhang S C, Hu J, et al. Image-to-Image Translation via Hierarchical Style Disentanglement[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 8635-8644.
- [21] Heusel M, Ramsauer H, Unterthiner T, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium[EB/OL]. 2017: 1706.08500. <https://arxiv.org/abs/1706.08500v6>.
- [22] Kurakin A, Goodfellow I J, Bengio S. Adversarial Examples in the Physical World[M]. *Artificial Intelligence Safety and Security*. First edition. | Boca Raton, FL: CRC Press/Taylor & Francis Group, 2018. Chapman and Hall/CRC, 2018: 99-112.
- [23] Dong Y P, Liao F Z, Pang T Y, et al. Boosting Adversarial Attacks with Momentum[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 9185-9193.
- [24] Xie C H, Zhang Z S, Zhou Y Y, et al. Improving Transferability of Adversarial Examples with Input Diversity[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 2725-2734.
- [25] Croce F, Hein M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks[EB/OL]. 2020: 2003.01690. <https://arxiv.org/abs/2003.01690v2>.
- [26] Wang L, Cho W, Yoon K J. Deceiving Image-to-Image Translation Networks for Autonomous Driving with Adversarial Perturbations[J]. *IEEE Robotics and Automation Letters*, 2020, 5(2): 1421-1428.
- [27] Yeh C Y, Chen H W, Tsai S L, et al. Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks[C]. *2020 IEEE Winter Applications of Computer Vision Workshops*, 2020: 53-62.
- [28] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2242-2251.
- [29] Wang T C, Liu M Y, Zhu J Y, et al. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8798-8807.
- [30] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.
- [31] Li J, Ji R R, Liu H, et al. Universal Perturbation Attack Against Image Retrieval[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 4898-4907.
- [32] Ruiz N, Bargal S A, Sclaroff S. Protecting Against Image Translation Deepfakes by Leaking Universal Perturbations from Black-Box Neural Networks[EB/OL]. 2020: 2006.06493. <https://arxiv.org/abs/2006.06493v1>.
- [33] He K M, Zhang X Y, Ren S Q, et al. Identity Mappings in Deep Residual Networks[M]. *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016: 630-645.
- [34] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.
- [35] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. *3rd International Conference on Learning Representations*, 2015.
- [36] Guo C, Rana M, Cisse M, et al. Countering Adversarial Images Using Input Transformations[EB/OL]. 2017: 1711.00117. <https://arxiv.org/abs/1711.00117v3>.
- [37] Xie C H, Wang J Y, Zhang Z S, et al. Mitigating Adversarial Effects through Randomization[EB/OL]. 2017: 1711.01991. <https://arxiv.org/abs/1711.01991v3>.
- [38] Liu Z, Luo P, Wang X, et al. Large-scale celebfaces attributes (celeba) dataset[J]. *Retrieved August*, 2018, 15(2018): 11.
- [39] Yi D, Lei Z, Liao S C, et al. Learning Face Representation from Scratch[EB/OL]. 2014: 1411.7923. <https://arxiv.org/abs/1411.7923v1>.

- [40] Howard A G, Zhu M L, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [EB/OL]. 2017: 1704.04861. <https://arxiv.org/abs/1704.04861v1>.
- [41] Karras T, Aila T M, Laine S, et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation[EB/OL]. 2017: 1710.10196. <https://arxiv.org/abs/1710.10196v3>.
- [42] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.
- [43] Thies J, Zollhöfer M, Stamminger M, et al. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2387-2395.
- [44] Zao app. 2019. <https://zao-app.com/>.
- [45] FakeApp. 2019. <https://www.deepfakescn.com>.



穆文鹏 现为南京信息工程大学计算机学院、网络空间安全学院本科生, CCF 学生会员。主要研究方向为: 信息安全, 多媒体取证。Email: imwp2002@163.com



王金伟 于 2007 年在南京理工大学自动化学院获得博士学位。现任南京信息工程大学教授, 博士生导师。研究领域为多媒体版权保护、多媒体取证、多媒体加密和数据认证。Email: wjwei_2004@163.com



陈北京 于 2011 年在东南大学获得工学博士学位。现任南京信息工程大学教授, 博士生导师。研究领域为多媒体内容安全(数字图像篡改检测、伪脸检测、水印和加密)、彩色图像处理以及模式识别等。Email: nbutimage@126.com



聂丽娜 现为南京信息工程大学计算机学院、网络空间安全学院本科生。主要研究方向为: 计算机视觉, 信息安全。Email: PeterNzz@outlook.com



莫皓岚 现为南京信息工程大学计算机学院、网络空间安全学院本科生。主要研究方向为: 信息安全, 多媒体取证。Email: molancci@163.com



徐菲 现为南京信息工程大学计算机学院、网络空间安全学院本科生。主要研究方向: 信息安全。Email: xf493915052@163.com