面向人脸属性编辑的三阶段对抗扰动生成主动防御算法

陈北京[1),2),3)] 张海涛[1),3)] 李玉茹[1),3)]

[1)] (南京信息工程大学教育部数字取证工程研究中心 南京 210044)

[2)] (南京信息工程大学 江苏省大气环境与装备技术协同创新中心 南京 210044)

[3)] (南京信息工程大学计算机学院、网络空间安全学院 南京 210044)

摘 要 针对恶意人脸属性编辑行为，基于取证的被动防御技术只能对篡改行为进行取证并不能防止其产生，从而难以消除恶意篡改行为已经造成的损失。因此，主动防御技术应运而生，其可以破坏属性编辑的输出从而避免人脸被篡改使用。然而，现有两阶段训练人脸篡改主动防御框架存在迁移性和扰动鲁棒性不足的问题，为此本文通过优化两阶段训练架构及损失函数和引入一个辅助分类器，提出一种三阶段对抗扰动主动防御框架。本文首先修改两阶段训练架构中的代理目标模型并基于此设计了训练扰动生成器的属性编辑损失，以提升代理模型的重建性能和属性约束能力，从而减少对代理模型的过拟合；其次，在训练阶段引入辅助分类器对代理模型提取的编码后特征进行源属性分类并基于此设计训练扰动生成器的辅助分类器损失，从而将原本的两阶段交替训练改为代理目标模型、辅助分类器和扰动生成器的三阶段交替训练，期望通过对抗攻击辅助分类器以促进对篡改模型的主动防御；最后，在扰动生成器的训练中，引入攻击层以促进对抗扰动对滤波和 JPEG 压缩的鲁棒性。实验结果验证，提出框架能够比现有框架更好地将主动防御从白盒的代理目标模型迁移到黑盒的属性编辑模型，黑盒性能提升 16.17%，且生成的对抗扰动较基线算法具有更强的鲁棒性，针对 JPEG 压缩的性能（PSNR）提升 13.91%，针对高斯滤波提升 17.76%。

关键词 人脸属性编辑；主动防御；对抗攻击；辅助分类器；交替训练

中图法分类号 TP391

Three-Stage Adversarial Perturbation Generation Active Defense Algorithm for Facial Attribute Editing

CHEN Bei-Jing[1)2)3)] ZHANG Hai-Tao[1)3)] LI Yu-Ru[1)3)]

[1)] (Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044)

[2)] (Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044)

[3)] (School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044)

Abstract With the gradual maturity of deep generation technology, the facial image generated by facial attribute editing technologies appear to mix the spurious with the genuine. Once these facial attribute editing technologies are maliciously used, such as infringing on personal privacy, and maliciously guiding public opinion, etc., they may trigger some moral, social, and security issues. Regarding the resolution of these malicious facial attribute editing behaviors, although the current passive defense technology based on forensics has achieved considerable performance, it can only provide evidence for tampering behavior and cannot prevent its occurrence, which is

收稿日期：2023-05-14；在线发布日期：2023-12-28.本课题得到国家自然科学基金 (62072251，62072250)资助.陈北京(**通信作者**)，博士，教授，主要研究领域为数字取证、彩色图像处理.E-mail: nbutimage@126.com.张海涛，硕士，主要研究领域为数字图像取证.E-mail: yzzhanghaitao@163.com. 李玉茹，硕士研究生，主要研究领域为数字图像取证.E-mail: 3246863022@qq.com.

difficult to eliminate the losses caused by malicious tampering behavior. Then, the active defense technology has emerged. It prevents face from being tampered with by disrupting the output of facial attribute editing. However, the existing two-stage training active defense framework for facial attribute editing has the issues of insufficient transferability and perturbation robustness. Therefore, this paper proposes a three-stage adversarial perturbation active defense framework for facial attribute editing by optimizing the two-stage training architecture and its loss function and introducing an auxiliary classifier. This paper first modifies the substitute target model in the two-stage training architecture and designs the attribute editing loss for the training of perturbation generator to improve the reconstruction performance and attribute constraint ability of the substitute model, thus reducing the overfitting issue of the substitute model; Secondly, the auxiliary classifier is introduced in the training phase to classify the source attributes of the encoded features extracted by the substitute model and the corresponding auxiliary classifier loss is designed for the training of perturbation generator. Then, the original two-stage alternate training is changed to the three-stage alternate training of substitute target model, auxiliary classifier and perturbation generator, so that it is expected to promote active defense against tampering model by countering auxiliary classifier; Finally, an attack layer is introduced in the training of the perturbation generator to enhance the robustness of the adversarial perturbation against filtering and joint photographic experts group (JPEG) compression. Experimental results on five facial attribute editing models (StarGAN, AttGAN with difference attribute vector input, AttGAN with target attribute vector input, STD-GAN, and style-aware model) show that the proposed framework can better migrate active defense from the white-box substitute model to the black-box attribute editing model than the existing frameworks, improving 16.17% in terms of peak signal-to-noise ratio (PSNR) in the case of black-box, and the generated adversarial perturbation has stronger robustness against JPEG compression and filtering than the baseline, improving 13.91% in terms of PSNR for JPEG compression, and 17.76% for the Gaussian filtering.

Key words facial attribute editing; active defense; adversarial attack; auxiliary classifier; alternate training

1 引言

伴随着卷积神经网络(Convolutional Neural Network, CNN)和深度生成模型如生成对抗网络(Generative Adversarial Network, GAN)[1]等深度学习技术的巨大成功，人脸属性编辑面部篡改技术近年来已经成为一个新兴话题。所谓人脸属性编辑，指的是对人脸图像的面部属性进行篡改，如发色、年龄、性别等属性。目前，属性编辑技术的发展使得篡改人脸越来越逼近自然人脸，最常用的技术主要包括：引入属性分类器剥离属性间关联性的AttGAN[2]与 StarGAN[3]。这些技术一旦被恶意使用，如侵犯个人隐私、恶意引导舆论等，可能会引发一些道德、社会和安全问题。

为了缓解恶意使用人脸属性编辑技术带来的风险，研究者们利用图像取证技术检测人脸图像是否被篡改，进行被动防御。虽然现有的取证检测器精度已经比较高，但是它们只能对篡改行为进行取证，难以消除恶意篡改已经造成的损失，因为由篡改人脸图像广泛传播造成的损失已经成为既定事实。于是，取证人员尝试从源头上阻止人脸被编辑以避免损失，这就是主动防御技术[4, 5]。

所谓主动防御，是指对即将发布的自然人脸作保护处理。当恶意伪造者使用属性编辑模型对处理后的自然人脸进行篡改时，生成的伪造人脸将与自然人脸生成的伪造人脸明显不同，从而从源头上降低人脸属性编辑模型的性能，在一定程度上阻止伪造行为。目前，主动防御技术主要都是采用对抗攻击手段[6]。具体来说，在自然人脸图像上添加人眼不易察觉的对抗扰动，使得属性编辑模型失效。根据扰动生成方式，现有主动防御算法可以分为基于梯度的算法[5,7-10]和基于生成模型的算法[4,11]。基于梯度的算法大多利用了投影梯度下降(Project Gradient Descent , PGD)对抗攻击算法[12]。这类算法存在以下两方面局限性，一方面主动防御效率较低，因为每一张图像都需要进行迭代梯度生成，另一方面主动防御性能方面仍然有待提升。基于生成模型的算法则采用了对抗变换网络(Adversarial Transformation Networks, ATNs)[13]或 AdvGAN[14]对抗扰动生成框架。Dong 等[11]基于 ATNs 提出了 TCA-GAN 架构，该架构首先破坏训练好的白盒人

脸重建模型，然后被直接迁移到人脸属性编辑模型的主动防御；Huang 等[4]规范了破坏篡改任务的行为，提出了人脸篡改“主动防御”的概念，并将 AdvGAN 应用于该任务，他们通过实验分析发现直接对抗攻击训练好的白盒代理模型容易陷入局部最优，于是利用两阶段交替训练策略来训练代理模型和扰动生成器实现对人脸属性编辑和人脸重现模型的主动防御。相较于基于梯度的算法，此类基于生成模型的算法提升了效率并且拥有更好的主动防御性能。然而[4]采用的两阶段交替训练框架存在着过拟合于代理模型的问题，并且其在滤波、JPEG 压缩等攻击的情况下防御效果大幅下降。

因此，本文以 Huang 等的主动防御框架为基线框架，进一步改进基线框架以减少对代理目标模型的过拟合并尽量提升扰动鲁棒性：通过修改目标代理模型为以差异属性为输入的 AttGAN 并基于此提出了训练扰动生成器的属性编辑损失来保证对属性编辑模型的主动防御性能，引入一个辅助分类器对代理目标模型提取的潜在特征进行源属性分类，设计对抗攻击分类器的目标函数使分类器分类错误来促进对生成模型的主动防御，提出一种三阶段对抗扰动主动防御框架，此外，引入了攻击层来提升对抗扰动对滤波和 JPEG 压缩的鲁棒性。

2 相关知识

2.1 基线算法

Huang 等[4]在 2021 年除了首次将破坏人脸篡改的技术定义为主动防御之外，还提出了一个全新的主动防御框架来降低人脸篡改模型的性能。该框架并没有采用直接攻击训练好的伪造模型的方式，而是通过一个两阶段训练框架将代理目标模型和对抗扰动生成器进行交替训练。该交替训练不同于常规的 GAN 对抗训练。具体来说，如图 1 所示，该主动防御框架采用了 AdvGAN 的经典结构，主要包括代理目标模型、扰动生成器和判别器。代理模型采用的是 StarGAN。在每一次训练迭代中，首先是阶段 A 的训练，在该阶段不涉及任何扰动生成器 *PG* 和判别器 *PD* 的知识，从零开始采用 StarGAN 的损失函数定期更新代理模型生成器 *SG* 和判别器 *SD* 参数，同时获取了当前参数下的干净伪造人脸；随后是阶段 B 的训练，在该阶段，代理目标模型参数不发生变化，通过上一轮迭代中训练的扰动生成器生成对抗扰动添加到干净人脸得到防伪造人脸，将其输入到 *SG* 输出破坏伪造人脸，通过扩大干净伪造人脸和破坏伪造人脸的距离来训练 *PG*，当然，*PG* 和 *PD* 之间是 GAN 的对抗训练，这样可以保证防伪造人脸的视觉效果。Huang 等在实验中也证明两阶段训练得到的防伪造人脸主动防御效果优于直接攻击训练好的伪造模型的方式。

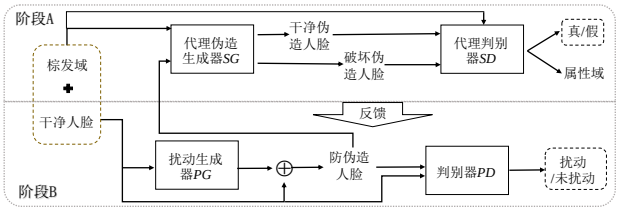

图 1 Huang 等的两阶段训练主动防御框架[4]

虽然该两阶段主动防御框架在人脸属性编辑模型 StarGAN 上取得了优异的主动防御性能，但仍然存在以下两方面局限性。一方面，该框架采用 StarGAN 架构作为代理目标模型，训练扰动生成器损失函数也是基于 StarGAN 设计的，从而导致该框架得到的防伪造人脸难以对未知的人脸属性编辑模型达成主动防御效果，性能过拟合于 StarGAN；另一方面，该框架生成的防伪造人脸在遇到滤波和 JPEG 压缩等攻击时，主动防御扰动会受到较大的影响。针对以上局限性，本文改进了该算法以增强主动防御性能和扰动鲁棒性。

2.2 AttGAN人脸属性编辑

He 等[2]提出的 AttGAN 框架是一个存在条件约束的 GAN 框架。该框架主要由采用编码器-解码器结构的生成器和包含真假判别器与属性分类器的判别器组成。其人脸属性编辑的过程可以描述为将自然人脸输入编码器得到编码后特征图，随后将条件约束即控制属性变化的目标属性向量与编码后特征图一起输入到解码器得到对应目标属性的篡改人脸。

在 AttGAN 中，为了使得生成器能够拥有根据条件约束控制属性编辑结果的能力，属性分类器在训练过程中起到了关键作用。具体来说，在对抗训练过程中，属性分类器能够正确预测输入自然人脸的原始属性标签，而生成器则约束篡改人脸经过属性分类器得到的预测属性为目标属性，如此通过属性分类约束使得生成器拥有属性编辑的能力。另外，为了要求对人脸目标属性编辑的同时保留原有的其他属性特征，He 等通过一个重建约束来实现，即当作为条件约束的目标属性为输入人脸的原始属性时，生成器得到的篡改人脸应当与输入人脸尽可能相似。除此之外，通过 WGAN_GP 对抗训练损失

[15]来稳定训练和保证篡改人脸的视觉效果。

3 本文算法

3.1 问题定义

本文的主动防御工作主要目标是在干净自然人脸图像中添加不易察觉的对抗扰动使得人脸属性编辑模型失效。给定一个原人脸属性为 a 的干净人脸图像 x 以及篡改目标属性 b，人脸属性编辑伪造过程可以描述为：

$$y = M(x, b) \tag{1}$$

其中，M 表示人脸属性编辑模型，y 为 x 经过属性编辑生成的属性为 b 的伪造人脸。

当干净自然人脸图像在被伪造者获得前，向图像注入对抗扰动进行主动防御可以表示为：

$$x' = x + \delta_x,\ \|\delta_x\|_\infty \le \varepsilon \tag{2}$$

其中，δ_x 为添加的对抗扰动，x' 为防伪造图像，ε 是控制扰动强度的阈值。

主动防御是为了破坏人脸属性编辑，换句话说，是想要使得防伪造图像经过伪造模型 M 生成的图像 y'和 y 的距离最大化。如此，主动防御过程可以描述为：

$$y' = M(x', b) \tag{3}$$

$$\delta_x = \underset{\delta_x}{\arg\max}\, D(y, y') \tag{4}$$

其中，$D(\cdot)$指距离函数。

3.2 总体框架

对于基线算法所采用的两阶段训练主动防御框架，其存在着过拟合于代理目标模型 StarGAN[3]且扰动鲁棒性不足的局限性。因此，本文对基线算法进行如下改进：

(1) 修改目标代理模型并基于此提出了训练扰动生成器的属性编辑损失。本文将 StarGAN 替换为 AttGAN 架构，而且将原 AttGAN 中的输入条件约束由目标属性修改为目标属性减去源属性的差异属性向量，即 b-a，如此两修改的主要原因为：① [14]发现对人脸重建模型的主动防御可以很好地迁移到人脸属性编辑模型，而 AttGAN 由于重建约束的存在拥有较强的人脸重建性能，其可以近似看作人脸属性编辑模型和重建模型的集成，因此，采用 AttGAN 可以减少对代理模型的过拟合，使得防伪造人脸能够较好地迁移到其他人脸属性编辑模型；② 本文通过对抗攻击引入的辅助分类器来促进对篡改生成模型的主动防御性能，并且 StarGAN 和 AttGAN 生成器都采用编码器-解码器架构，而 StarGAN 编码器的输出特征会受到输入目标属性约束的影响，AttGAN 编码器输出的特征只与源属性相关，由于辅助分类器输入利用了编码后特征进行源属性分类，因此 AttGAN 更有利于辅助分类器；③ [16]中实验表明差异属性输入 AttGAN 比目标属性输入 AttGAN 具有更优的生成性能。

(2) 引入辅助分类器优化主动防御性能并基于此提出了训练扰动生成器的辅助分类器损失，将原本的两阶段交替训练改为三阶段交替训练。具体来说，如果对抗扰动能够使得辅助分类器分类错误，那么意味着代理目标模型编码后特征被破坏，如此，解码器的输出属性编辑结果应该也是被破坏的。

(3) 引入攻击层增强对抗扰动对 JPEG 压缩和滤波操作的鲁棒性。

提出算法的整体框架如图 2 所示，其伪代码如算法 1 所示。本框架的三阶段训练过程如下：第一阶段是代理目标模型 *SM* 的生成器 *SG* 和判别器 *SD* 的训练，其中 *SG* 是编码器-解码器结构，即 SG_{enc}-SG_{dec}，*SD* 包含属性分类器 *SDC* 和真假判别器 *SDF*；第二阶段是添加的辅助分类器 *AC* 的训练，将 SG_{enc} 输出的包含人脸属性特征信息的 *z* 输入 *AC* 进行属性分类；第三阶段是扰动生成器 *PG* 和扰动判别器 *PD* 的训练，这里保留基线算法的第二阶段训练策略，即在 SG_{enc}-SG_{dec} 和 *AC* 参数不变的情况下进行 *PG* 和 *PD* 的对抗训练，而且基于代理目标模型和辅助分类器分别设计了使得防伪造人脸拥有主动防御效果的属性编辑损失和辅助分类器损失来训练 *PG*，并引入攻击层 *A*[17]以增强防伪造人脸在滤波和 JPEG 压缩等攻击下的鲁棒性。一旦训练完成，只需要将输入干净人脸输入到 *PG* 即可获得防伪造人脸。

算法 1. 三阶段训练主动防御框架(无攻击层).

输入：最大迭代次数 *n*，训练数据集 *T*。

输出：训练有素的扰动生成器 PG_{best}。

1 初始化得到 SM_0、AC_0、PG_0，$PG_{best} = PG_0$，
max_dist = *-inf*;

2 **FOR** *i* = 1 to *n* **DO**

3 从数据集获取小批量干净人脸数据得到
x = *SelectMiniBatch(T)*;

4 第一阶段训练，在 *x* 上训练代理目标模型并生成属性编辑后的伪造人脸得到
SM_i = *UpdateParameter*(SM_{i-1}, *x*)，*y* = SM_i(*x*);

5 第二阶段训练，冻结SG_{enc_i}参数训练辅助分类器得到
$AC_i = UpdateParameter(AC_{i-1}, SG_{enc_i}(x))$；

6 第三阶段训练，通过上一次迭代后的 PG_{i-1} 生成对抗扰动得到防伪造图像，并生成破坏伪造人脸，得到
$x' = x + \varepsilon \cdot PG_{i-1}(x)$，$y' = SM_i(x')$；
优化设计的目标损失来训练扰动生成器得到
$PG_i = UpdateParameter(PG_{i-1}, x)$；

7 最后与 maxdist 比较干净伪造人脸和破坏伪造人脸的距离得到

IF $max\_dist < D(y, y')$ **THEN**
$max\_dist = D(y, y')$;
$PG_{best} = PG_i$;
END
END
RETURN PG_{best}

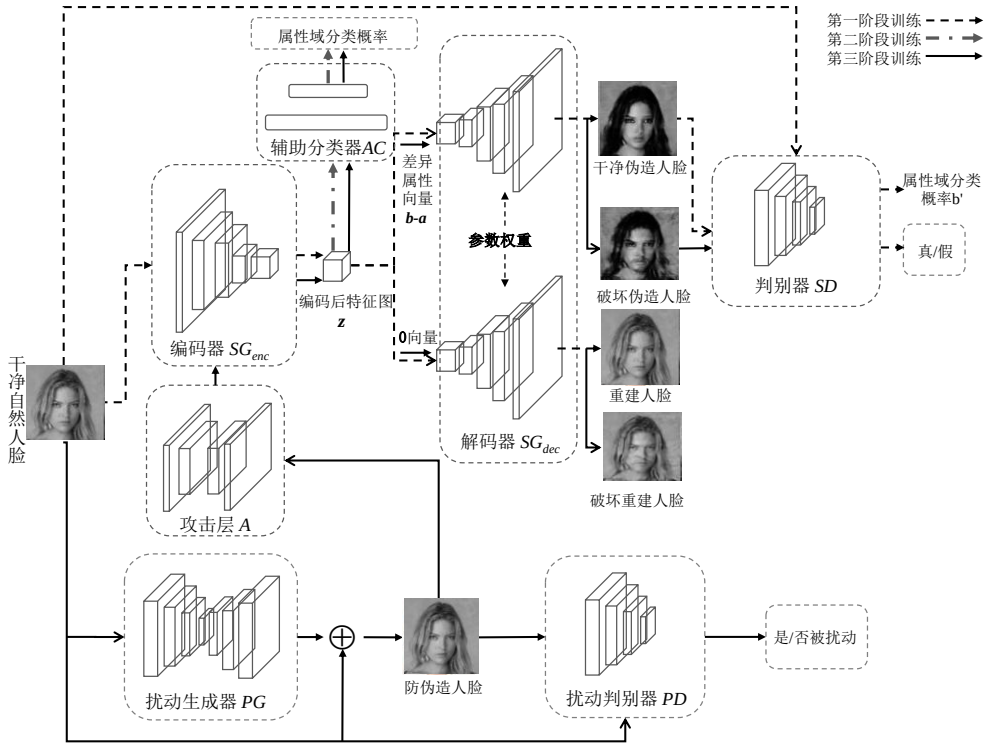

图 2 三阶段交替训练主动防御框架

3.3 模块架构

(1) 代理目标模型

如 3.2 中所述，本文框架第一阶段所使用的代理目标模型是以差异属性为输入的 AttGAN 架构。在该架构中，首先，源属性为 a 的干净人脸 x 输入编码器 SG_{enc} 得到潜在特征表示 z，即：

$$z = SG_{enc}(x) \tag{5}$$

然后，针对目标属性 b 的属性编辑，将原 AttGAN 中输入的目标属性向量替换为目标属性减去源属性的差异属性向量 $b - a$，并将 $b - a$ 和 z 一起输入解码器 SG_{dec} 得到属性编辑后的篡改人脸 y，即：

$$y = SG_{dec}(z, b - a) \tag{6}$$

针对重建约束，将 0 向量与 z 一起输入 SG_{dec} 获取重建人脸 x_{rec}，即：

$$x_{rec} = SG_{dec}(z, 0) \tag{7}$$

本文中 SG_{enc}、SG_{dec}、SD 的网络结构与原始 AttGAN 一致。

(2) 辅助分类器

本文框架的第二阶段训练引入了一个辅助分类器 AC。AC 对第一阶段训练得到的 SG_{enc} 从干净人脸提取的特征进行属性分类。该特征包含了干净人脸的原始属性特征，因为属性编辑的控制属性是直接输入 SG_{dec} 的。之所以设计这样的辅助分类器，主要是为了促进第三阶段训练中对代理伪造模型的主动防御。具体来说，在第二阶段训练中，干净自然人脸输入 AC，AC 正确进行属性分类，而在第三阶段训练过程中，防伪造人脸输入 AC，AC 受到对抗扰动的影响分类错误，这也就意味着主动防御扰动直接使得干净人脸和防伪造人脸编码后的潜在特征距离增大、属性分类错误，那么防伪造人脸潜在特征经过解码生成的伪造人脸应该也是被破坏的。如此，这阶段训练尝试将对抗攻击生成模型问题转化为对抗攻击分类模型。辅助分类器 AC 采用了两层全连接层，第一层全连接层的神经元个数

为 1024，第二层为 n，使用 leakyReLU 激活。

(3) 扰动生成器

本文框架的第三阶段训练是扰动生成器 *PG* 和 *PD* 的对抗训练，希望 *PG* 生成的扰动能够使得伪造模型属性编辑失败。在这一阶段训练过程中，代理目标模型 *SM* 和辅助分类器 *AC* 参数不发生改变。扰动生成器 *PG* 的网络架构如图 3 所示，采用 UNet-128[18]的结构。判别器 *PD* 使用常规的卷积神经网络，包括七层卷积层和一层全连接层。

生成对抗扰动得到防伪造人脸的整体过程可以如下描述，输入自然人脸经过扰动生成器得到对抗扰动，将该扰动经过阈值调整添加到自然人脸即可得到防伪造人脸，即：

$$\delta_x = PG(x) \tag{8}$$

$$x' = x + \varepsilon \cdot \delta_x \tag{9}$$

其中，δ_x为 *PG* 生成的扰动，ε为扰动强度的可调节阈值，x' 为防伪造人脸。

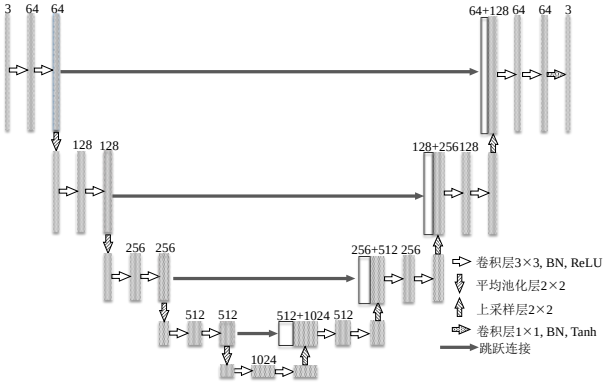

图 3 扰动生成器 *PG* 的网络架构

(4) 攻击层

为了加强防伪造人脸在现实情况下对 JPEG 有损压缩和滤波的鲁棒性，本文额外引入一个攻击层以实现对现实攻击的模拟。

JPEG 有损压缩通常包括色彩空间转换、采样、DCT 变换、量化和熵编码五个步骤。除量化之外的四个步骤均可微，可以直接处理。量化步骤因为含有不可微的取整运算，所以需要采用可微运算进行近似模拟。本文采用[19]中设计的如下式所示的方法来模拟取整操作[·]：

$$[y]_{approx} = [y] + (y - [y])^3 \tag{10}$$

对于滤波攻击，本文采用高斯滤波攻击来提高扰动对滤波的鲁棒性。实验中高斯滤波核为 3×3，均值为 0，标准差为 0.8。

攻击层的使用是在第三阶段训练中，即更新 *PG* 的参数时。每一次参数迭代时，随机从模拟 JPEG 压缩、高斯滤波和无攻击三种情况下选择一种对 x' 进行操作，然后再迭代参数。

3.4 训练目标函数

第一阶段训练，代理目标模型的整体训练逻辑与 AttGAN[2]类似，仅仅将原属性分类约束和重建约束中控制属性向量分别更改为差异属性向量 *b-a* 和 0 向量。

第二阶段训练，*AC* 的分类损失如下：

$$L_{ac\_cls} = \mathbb{E}_{x\sim p_{data},a\sim p_{attr}}[\sum_{i=1}^{n} -a_i \log AC_i(z) - (1-a_i)\log(1-AC_i(z))] \tag{11}$$

其中，p_{data}表示干净人脸 x 的分布；z 为干净人脸 x 输入编码器后得到潜在特征表示；p_{attr}为属性域的分布；原属性向量$a=[a_1,\cdots,a_n]$，目标属性向量$b=[b_1,\cdots,b_n]$，a_i和 b_i表示总共 n 个人脸属性中第 i 个属性的 0 或 1 标签，0 表示没有该属性，1 表示拥有该属性，例如黑发属性标签为 0 表示该人脸非黑发。

第三阶段训练，*PG* 的目标损失函数由三部分组成，如下所示：

$$L_{PG} = \beta_1 L_{pg\_adv} + \beta_2 L_{pg\_attr} + \beta_3 L_{pg\_ac} \tag{12}$$

其中，β_1、β_2、β_3 分别为三个可调节的超参数，L_{pg_adv}保证了 *PG* 和 *PD* 之间的对抗训练并迫使防伪造人脸与干净人脸相近，L_{pg_attr}保证添加的对抗扰动具有属性编辑主动防御性能，L_{pg_ac}用来增强主动防御效果。从而，为了与 *PG* 形成对抗训练，*PD* 的损失函数为：

$$L_{PD} = L_{pd\_adv} \tag{13}$$

下面详细介绍第三阶段训练的目标函数。

(1) *PG* 和 *PD* 的 GAN 训练损失L_{pg_adv} 和 L_{pd_adv}

这里考虑了 WGAN_GP 训练损失来稳定 *PG* 和 *PD* 的训练，具体表示为：

$$L_{pd\_adv} = -\mathbb{E}_{x\sim p_{data}}[PD(x)] + \mathbb{E}_{x\sim p_{data},b\sim p_{attr}}[PD(x')] + \lambda_{gp}\cdot \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{14}$$

其中，最后一项是利用 Wasserstein 距离来保证训练稳定的惩罚梯度项。而 *PG* 的 GAN 训练损失为：

$$L_{pg\_adv} = -\mathbb{E}_{x\sim p_{data},b\sim p_{attr}}[PD(x')] \tag{15}$$

(2) 训练 *PG* 的属性编辑损失L_{pg_attr}

该损失的目的是使得x'拥有主动防御能力，即在训练过程中尝试破坏 *SG* 的属性编辑。除了扩大干净伪造人脸和破坏伪造人脸的距离损失函数 L_{pg_comp} 外，L_{pg_attr} 还包括破坏属性分类约束的 *SD* 分类损失 L_{pg_sd} 和破坏重建损失L_{pg_rec}组成，即：

$$L_{pg\_attr} = \beta_4 L_{pg\_comp} + \beta_5 L_{pg\_rec} + L_{pg\_sd} \tag{16}$$

其中，β_4 和 β_5 为两个超参数。

L_{pg_comp} 的主要目标是使得防伪造人脸经过属性编辑的破坏伪造人脸 y' 和原始伪造人脸 y 之间距离最大化，即通过 y'和 y 的对比以使得防伪造人脸能够抵挡属性编辑，具体损失为：

$$L_{pg\_comp} = \mathbb{E}_{x\sim p_{data}}[-\sum_{j=1}^{n}\eta_j \left\|y - SG_{dec}(z', c_j - a)\right\|_1] \quad (17)$$

其中，z' 为防伪造人脸输入 SG_{enc} 得到的潜在特征表示，即：

$$z' = SG_{enc}(x') \quad (18)$$

c_j 表示从 x 的原属性向量 a 导出的一系列目标属性向量，例如原本黑发年轻男性向量[1,0,0,1,1]，衍生出的目标属性向量有重建向量[1,0,0,1,1]、金发年轻男性[0,1,0,1,1]、棕发年轻男性[0,0,1,1,1]、黑发年老男性[1,0,0,0,1]、黑发年轻女性[1,0,0,1,0]；η_j 是基线算法[4]采用的用来平衡不同编辑属性在图像像素级区域由于面积大小差异的问题而计算的损失函数大小差异，具体计算公式如下：

$$\eta_j = \frac{\left\|x - SG_{dec}(z, c_j - a)\right\|_1}{\sum_{j=1}^{n}\left\|x - SG_{dec}(z, c_j - a)\right\|_1} \quad (19)$$

L_{pg_rec} 的主要目标是破坏 SG 的重建过程。具体来说，差异属性向量如果为 0，那么 SG 可以看作是一个重建模型。如果对抗扰动能够破坏该重建，那么也应该能够很好达成对属性编辑模型的破坏。因此，L_{pg_rec} 重建损失能够保证防伪造人脸的主动防御效果能够迁移到其他属性编辑模型，而不仅仅是在代理目标模型中有效。L_{pg_rec} 的计算公式如下：

$$L_{pg\_rec} = \mathbb{E}_{x\sim p_{data}}[-\left\|y - SG_{dec}(z', 0)\right\|_1] \quad (20)$$

L_{pg_sd} 主要是为了破坏 SG 训练过程中的属性约束。如果破坏伪造人脸在 SD 的属性分类中被分类为非目标属性，那么破坏伪造人脸的很大程度上是没有属性编辑成功的。具体损失如下：

$$L_{pg_{sd}} = \mathbb{E}_{x\sim p_{data}}\left[\sum_{j=1}^{n}\left(SDF(y_j') - \sum_{i=1}^{n}\left(c_{rj_i}\log SDC_i(y_j') + (1 - c_{rj_i})log\left(1 - SDC_i(y_j')\right)\right)\right)\right] \quad (21)$$

其中，c_{rj} 指的是将 c_j 中的 n 个属性的二元标签取反，即标签 0 变 1，1 变 0；y_j' 对应于目标属性向量为c_j 的破坏伪造人脸，即：

$$y_j' = SG_{dec}(z', c_j - a) \quad (22)$$

(3) 促进主动防御的辅助分类器损失L_{pg_ac}

该损失的主要目的是通过对抗攻击辅助分类器来促进对代理伪造模型的主动防御，将对抗攻击生成器的任务尝试转化为更常见的对抗攻击分类器任务，如果添加的扰动能够使得第二阶段训练的 AC 分类错误，那么意味着扰动能够在 z' 输入 SG_{dec} 之前成功破坏了相关属性特征，导致防伪造人脸无法被正确预测为源属性，那么在属性编辑前可能达成主动防御。此外，由于攻击潜在特征能够增强对抗样本的迁移能力[20, 21]，还尝试扩大潜在特征表示 z 和 z' 的距离损失函数来增强破坏效果。具体的损失函数为：

$$L_{pg\_rec} = \mathbb{E}_{x\sim p_{data}}\left[\sum_{i=1}^{n}[a_i\log AC_i(z') + (1 - a_i)\log\left(1 - AC_i(z')\right)] - \left\|z - z'\right\|_1\right] \quad (23)$$

除此之外，在具体训练过程中，本文参考了基线算法的训练策略，即在进行第三阶段训练时，不仅仅使用了当前轮次 SG 和 AC 的参数，还集成了上一次迭代的 SG 和 AC 参数下计算的L_{pg_attr} 和 L_{pg_ac}。通过这样设计，保证了当前迭代过程参数优化的稳定性，并且一定程度上可以保证生成扰动的主动防御能力。

4 实验结果与分析

4.1 实验基础设置

数据集。本文中的实验采用 CelebA[22]作为基础数据集。它包含 10177 人的 202599 张 178×218 自然人脸图像。根据主动防御需要，数据集被分为三部分，100000 张用于训练本文的三阶段训练主动防御框架，100000 张用于训练所有的人脸属性编辑模型，剩下的 2599 张用于测试主动防御性能。在数据预处理阶段，训练集所有图像都经过数据随机水平翻转、从中心扩散裁剪为 178×178、缩放为 128×128 以及常用的标准化处理。

人脸属性编辑模型。在训练和评估主动防御之前，首先需要训练防御对象--目标伪造模型。本文以基线算法和本文所采用的代理目标模型作为主动防御性能评估的重点，即 StarGAN 属性编辑模型[3]和采用差异属性向量输入的 AttGAN 属性编辑模型[16]。此外，正常训练的采用目标属性向量输入的 AttGAN[2]、STD-GAN[23]、注重风格生成模型[24]被用来测试主动防御的迁移效果。训练细节如下：数据集与上述数据集一致，目标函数均遵循原文设置，StarGAN 学习率设置为 0.0001，其余模型学习率设置为 0.0002，批量大小统一为 64，选择 Adam 作为优化器。属性编辑图像示例如图 4 所示。

实验评价指标。本文采用四个指标来评估防伪造人脸视觉质量和主动防御性能，即：结构相似性(Structural Similarity, SSIM)[25]、峰值信噪比(Peak Signal to Noise Ratio, PSNR)、学习感知图像块相似

性 (Learned Perceptual Image Patch Similarity, LPIPS)[26]以及常用的 VGG 感知损失[27]均值。对于防伪造人脸视觉质量评估，由于防伪造人脸与干净自然人脸越相似质量越高，因此它们之间的 SSIM 和 PSNR 越大越好，LPIPS 和 VGG 感知损失越小越好。对于主动防御性能评估，由于破坏伪造人脸与用干净伪造人脸差异越大防御性能越高，因此与防伪造人脸视觉质量评估刚好相反，它们之间的 SSIM 和 PSNR 越小越好，LPIPS 和 VGG 感知损失越大越好

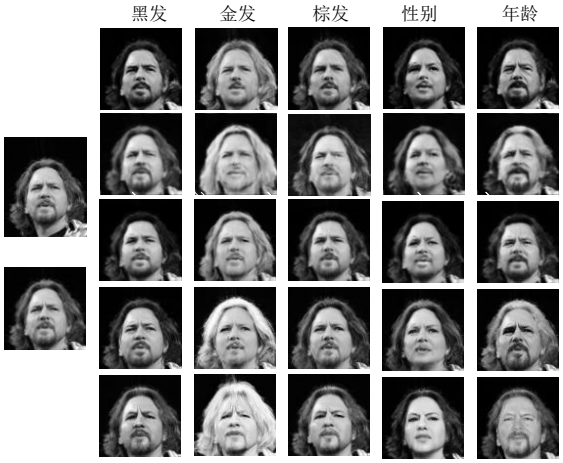

图 4 属性编辑示例，第一列为数据集原始自然人脸和其裁剪缩放后的版本，其余列为属性编辑后人脸，行从上到下依次为 StarGAN、以差异属性为输入 AttGAN、以目标属性为输入的 AttGAN、STD-GAN、注重风格生成模型

给定两张人脸 U 和 V，它们之间 SSIM 定义为：

$$\mathrm{SSIM}(U,V)=\frac{(2\mu_U\mu_V+C_1)(2\sigma_{UV}+C_2)}{(\mu_U^2\mu_V^2+C_1)(\sigma_U^2\sigma_V^2+C_2)} \quad (24)$$

其中，其中 μ_U 和 μ_V 是均值，σ_U 和 σ_V 是标准差，σ_{UV} 是互协方差，C_1 = 0.0001 和 C_2=0.0009 是避免分母为零的两个常数。

它们之间的 PSNR 为：

$$\mathrm{PSNR}(U,V)=10\log_{10}\left(\frac{M^2}{MSE(U,V)}\right) \quad (25)$$

其中，M 是最大像素值，MSE 是均方误差，定义为：

$$MSE=\sum_{m=1}^{p}(U_m-V_m)^2 \quad (26)$$

其中，p 是像素数量。

LPIPS 是一种流行的基于 CNN 的图像质量评估度量，用于语义相似性测量。算法细节详见[26]。

VGG 感知损失是风格迁移任务中常用的语义相似性感知损失。相关细节详见[27]。

实验环境。本文中的所有代码均通过 PyTorch 实现，三阶段训练主动防御框架和基线框架均在两块 24GB GeForce RTX 3090、3.80GHz i7-10700KF CPU 和 32GB RAM 上运行。

实验超参数设置。本文算法的具体细节描述如下：公式(11)和(15) 中的 β_1、β_2、β_3、β_4 和 β_5 分别设为 0.001、10、1、1 和 10。优化器采用 Adam，代理目标模型学习率被设置为 0.0002，辅助分类器和扰动生成器学习率设为 0.0001，批量大小设置为 48，训练属性为随机选的黑发、金发、棕发、性别和年龄。关于对比算法的超参数设置，与其原始文献一致。此外，参照基线算法，综合考虑防伪造人脸主动防御性能和图像视觉质量，本文实验中测试不同的对抗扰动阈值ε，即从 0.01 到 0.05，间隔 0.01。本文算法代码已在 GitHub 开源：https://github.com/imagecbj/Three-Stage-adversarial-perturbation-Initiative-Defense-for-Face-Attribute-Editing。

4.2 主动防御实验

(1) 防伪造人脸视觉质量评估

本文框架与对比算法不同扰动阈值下的防伪造人脸视觉质量对比如图 5 所示，可以发现在同等阈值下，本文算法和基线算法[4]以及基于 PGD 的算法[5, 10]视觉效果相当。这也就说明了 *PG* 和 *PD* 的 GAN 训练损失 L_{pg_adv} 和 L_{pd_adv} 的有效性。该损失使得防伪造人脸和干净自然人脸相近，导致通过扰动生成器和扰动判别器的对抗训练得到的防伪造人脸图像视觉质量更多地取决于扰动阈值的大小。当然，扰动阈值越大，防伪造人脸与干净人脸相比的视觉质量受到的影响越大。扰动阈值 ε =0.05 下防伪造人脸示例如图 6 所示。从图 6 可以发现对比的五种算法从人眼视觉上来看均具有较强的扰动不可见性。这主要是因为阈值为 0.05 的扰动情况下，单个像素点的像素值变化幅度不超过 6.4，且在 GAN 训练损失的作用下，防伪造人脸的视觉效果拥有一定的保证。

(2) 主动防御性能评估

首先在基线算法[4]以及 Ruiz 等[5]使用的代理目标模型 StarGAN 以及本文算法和 Huang 等[10]所使用的代理目标模型差异属性输入 AttGAN 上测试主动防御性能，随后在对比算法和本文算法训练中均未使用的三个生成模型上进一步测试迁移性能，即目标属性向量输入 AttGAN、STD-GAN 和注重风格生成模型。不同扰动阈值下的各算法主动防御性能评估结果如图 7 所示。同时，图 8 也展示了扰动阈值为 0.05 的可视化结果。由于篇幅原因，不同发色的示例只展示了黑发下的效果。从图 7 和图

8 可以发现：

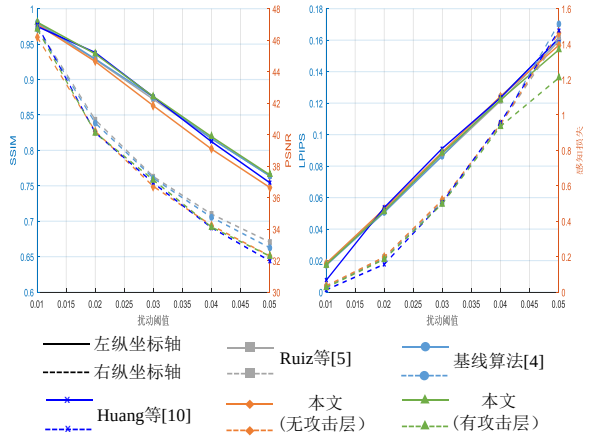

图 5 不同阈值下各算法防伪造人脸视觉质量对比（SSIM 和 PSNR 越大质量越高，LPIPS 和感知损失越小质量越高）

(a) **在本文和 Huang 等[10]所使用的代理目标模型“以差异属性向量为输入的 AttGAN”上**，基线算法和 Ruiz 等采用的 PGD 算法在迁移到该模型时，几乎失去了主动防御性能，破坏篡改的程度很小；而对于本文算法，其破坏伪造人脸相对于干净伪造人脸有明显的失真；Huang 等尽管和本文算法采用相同的代理目标模型，但防御性能还不强。此外，本文算法主动防御后生成的篡改人脸的性别或年龄可能会发生改变，这主要是 L_{pg_sd} 和 L_{pg_ac} 的作用，L_{pg_sd} 迫使 SG 对防伪造人脸属性编辑时往非目标属性去编辑，而 L_{pg_ac} 直接破坏 SG_{dec} 提取的人

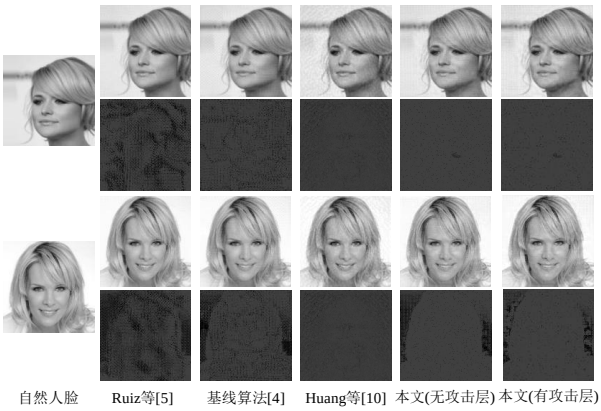

图 6 ε =0.05 下干净自然人脸、防伪造人脸示例。同一张人脸，第一行为相应算法得到的防伪造人脸，第二行为其防伪造人脸和自然人脸的十倍放大差异图

脸属性相关特征使得辅助分类器属性分类错误，而在训练过程选择的属性中，性别和年龄为固定的二元标签，而发色是多样的，那么破坏伪造人脸相对于干净伪造人脸的性别和年龄发生改变是符合直觉的，同时这也说明了将主动防御转为对属性分类器的破坏的有效性。从破坏程度上来说，针对差异属性输入 AttGAN 的破坏程度不如针对 StarGAN 的，这主要是受到模型本身的影响，AttGAN 因为有重建约束而具有强大的重建能力，这一定程度上加大了主动防御的难度，当然也保证了本文算法的主动防御功能迁移到相对容易模型的效果。

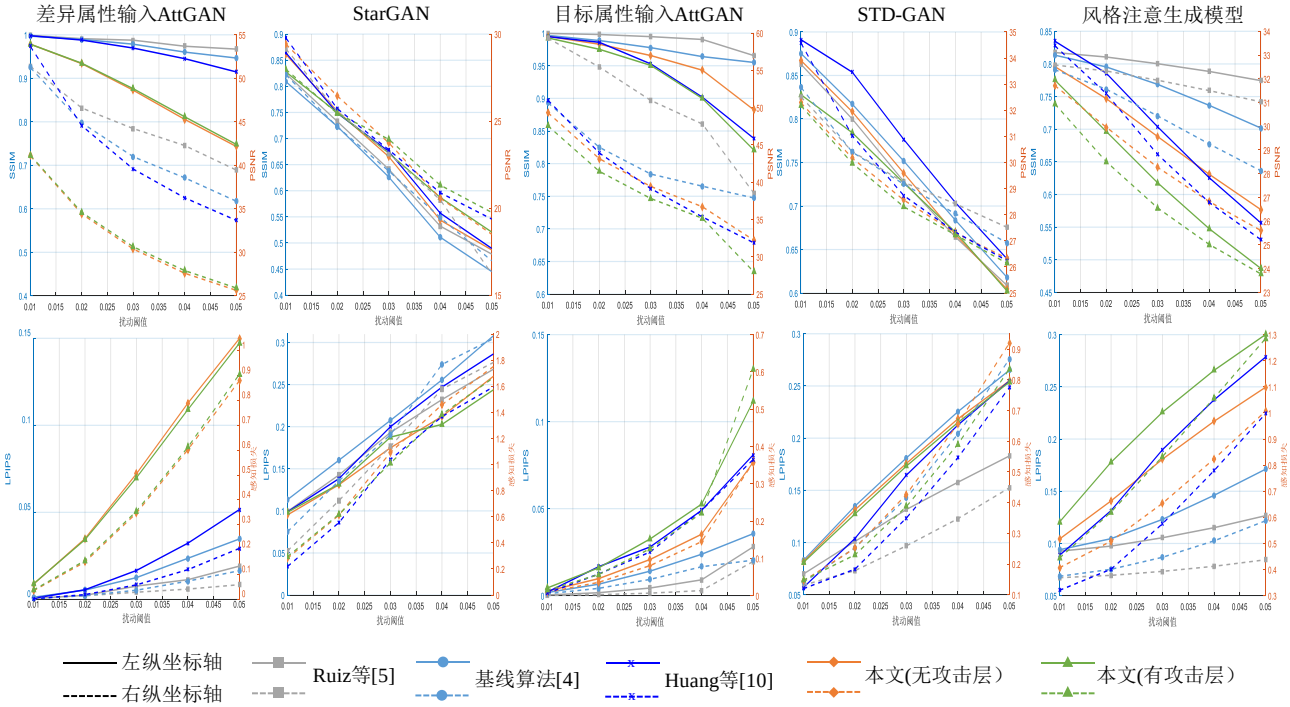

图 7 不同扰动阈值下各算法主动防御性能对比（SSIM 和 PSNR 越小防御性能越高，LPIPS 和感知损失越大防御性能越高）

(b) **在基线算法以及 Ruiz 等使用的代理目标模型“StarGAN”上**，本文算法和 Huang 等的算法在人脸属性编辑的破坏程度上不如对比算法，这主要是因为对于本文算法和 Huang 等的算法来说，

StarGAN 模型是完全黑盒的篡改模型，而对于对比算法来说，StarGAN 是在训练中直接使用的白盒模型。但根据图 8 所示，本文算法生成的主动防御扰动是可以成功迁移到 StarGAN 上的，其篡改人脸存在着明显的失真，已经阻止了属性编辑行为，并且从图 7 可知其破坏伪造人脸的视觉评估指标仅仅略低于对比算法。

(c) **在对比算法和本文算法训练中均未使用的“目标属性向量输入 AttGAN”、“STD-GAN”和“注重风格生成模型”上**，各个算法的主动防御性能有差异，本文算法基本上全面占优，特别是加入攻击层的本文算法。这主要是因为攻击层的加入类似于对迭代中的输入数据进行了增强，从而有助于提升对抗扰动的迁移性。针对三个黑盒生成模型来说，各算法针对注重风格生成模型的防御性能都相对较好。这主要是因为注重风格生成模型自身的属性编辑能力不是太强，在未防御下其生成图像都不是那么成功，如图 8 第一行所示。

(d) **在整体上**，本文算法在白盒的差异属性输入 AttGAN 和黑盒的 StarGAN 上都能达成主动防御效果，而基线算法以及 Ruiz 等采用的 PGD 算法对其白盒的 StarGAN 具有强大的主动防御效果，但对其黑盒的差异属性输入 AttGAN 却几乎失去了主动防御效果；在对比算法和本文均未知的三个黑盒模型上，针对注重风格生成模型，所有算法都具有比较好的防御性能，但针对目标属性输入 AttGAN 和 STD-GAN 上，所有算法的主动防御性能都不是那么理想，不过本文算法具有最好的效果。当扰动强度阈值为ε=0.05 时，本文算法相对于基线算法[4]针对各自白盒共两个目标模型上的性能（PSNR）提升 13.62%，针对共同的三个黑盒上提升 16.17%。

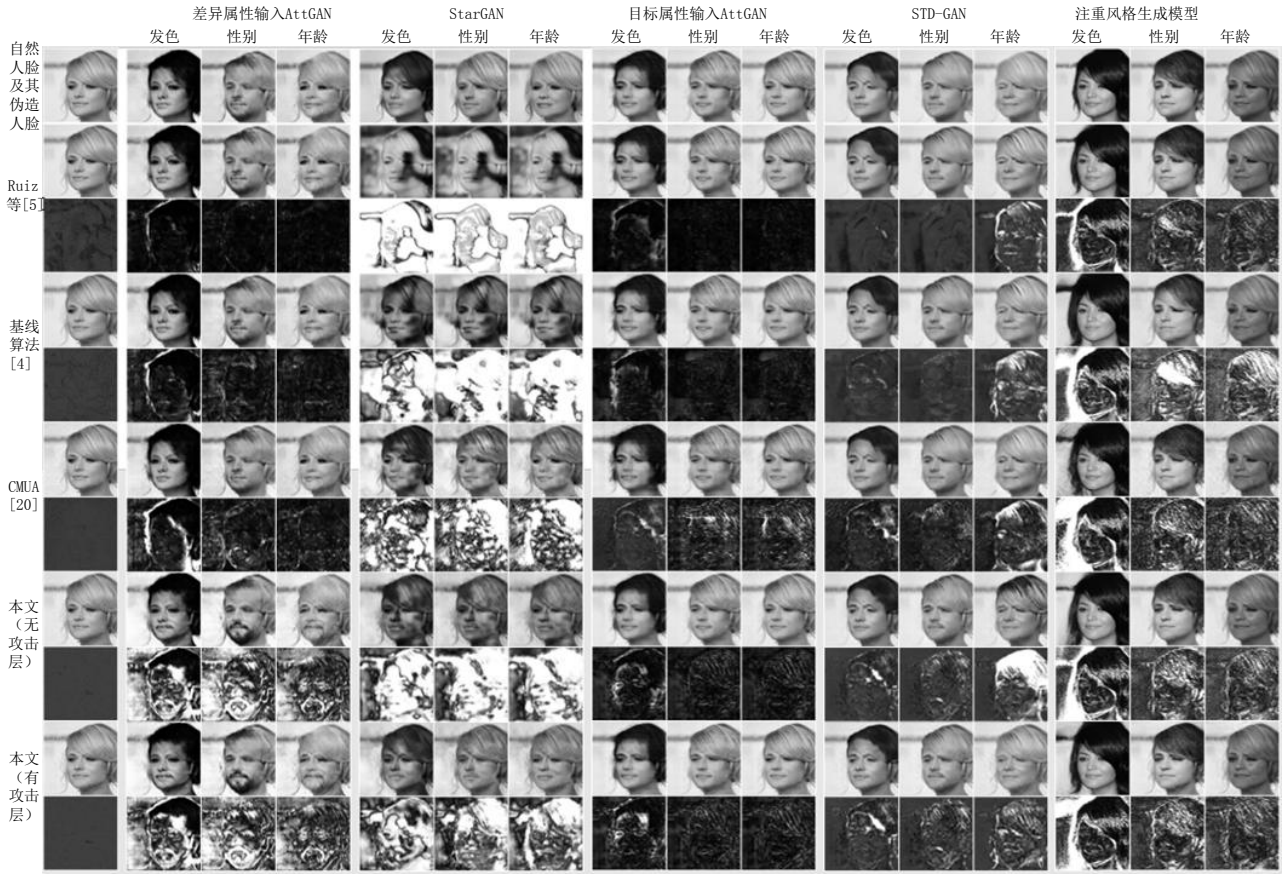

图 8 ε =0.05 下在各算法针对三种人脸属性编辑模型的主动防御效果示例。同一方法，第一行为防伪造人脸及其破坏伪造人脸，第二行为各自与自然人脸及其干净伪造人脸的十倍放大差异图

(3) 消融实验

如小节 3.2 所述，本文对基线算法的主要改进为：修改代理目标模型 *SM*，引入了辅助分类器 *AC* 和攻击层 *A*。因此，本小节将进行相关消融实验。为了更好的体现引入攻击层的性能，这里考虑在 JPEG 压缩攻击和高斯滤波攻击下进行实验。相关攻击参数为：JPEG 压缩因子 85，高斯滤波的核大小为 3×3、均值为 0、标准差为 0.8。两种攻击下的

消融实验结果分别如表 1 和表 2 所示。由两个表中的结果可知，(a) 在两种攻击情况下三个改进措施均有效的提升了防伪造人脸的主动防御性能。因为这里本实验考虑的是有攻击的场景，因此攻击层的引入提升最大；(b) 提出算法（基线算法+*SM*+*A*+*AC*）较基线算法具有更强的鲁棒性，针对 JPEG 压缩的性能（PSNR）提升 13.91%，针对高斯滤波提升 17.76%。

5 结语

针对现有两阶段训练人脸伪造主动防御框架迁移性和鲁棒性不强的问题，本文通过优化基线框架的两阶段训练架构及损失函数和引入一个辅助分类器，提出一种三阶段对抗扰动主动防御框架。提出的框架在主动防御性能、扰动鲁棒性上均优于基线算法，这主要归因于：(1)采用的代理伪造模型为以差异属性为输入的 AttGAN 架构，其具有强大的重建性能和属性约束能力；(2)对抗攻击引入的辅助分类器能够破坏编码后特征从而促进对整个生成模型的主动防御；(3)训练扰动生成器的属性编辑损失从破坏编辑、破坏重建、破坏属性分类约束三方面使得防伪造人脸具有较强的主动防御能力；(4)引入的攻击层能够增强防伪造人脸在遇到滤波和 JPEG 压缩等攻击下的鲁棒性。在未来工作中，可以通过优化代理目标模型的选择如通过集成模型等手段进一步提升对抗扰动的迁移能力，在攻击层中考虑更多的变换或者动态的攻击强度来进一步提升扰动鲁棒性。

表 1 ε =0.05 和压缩因子为 85 的 JPEG 压缩下的消融实验

方法	StarGAN				差异属性输入 AttGAN			
	SSIM	PSNR(dB)	LPIPS	感知损失	SSIM	PSNR(dB)	LPIPS	感知损失
基线算法[4]	0.8719	31.3081	0.0694	0.3002	0.9605	37.6289	0.0243	0.0854
+*SM*	0.8465	30.9292	0.0879	0.4269	0.8886	31.7464	0.0655	0.3234
+*SM*+*AC*	0.8426	30.8636	0.0887	0.4268	0.8834	31.2435	0.0661	0.3258
+*SM*+*A*	0.8203	30.7672	**0.1099**	**0.5372**	0.8483	29.5498	0.0921	0.4529
+*SM*+*A*+*AC*	**0.8139**	**30.0721**	0.1090	0.5280	**0.8398**	**29.2750**	**0.0937**	**0.4978**

表 2 ε =0.05 和核大小为 3×3、均值为 0、标准差为 0.8 的高斯滤波下的消融实验

方法	StarGAN				差异属性输入 AttGAN			
	SSIM	PSNR(dB)	LPIPS	感知损失	SSIM	PSNR(dB)	LPIPS	感知损失
基线算法[4]	0.9330	34.6800	0.0312	0.1136	0.9747	39.8078	0.0147	0.0452
+*SM*	0.9275	35.5805	0.0461	0.1860	0.9192	33.5009	0.0467	0.2154
+*SM*+*AC*	0.9237	35.1633	0.0471	0.1906	0.9122	32.6768	0.0482	0.2287
+*SM*+*A*	0.8678	32.6330	0.0898	**0.3928**	0.8491	29.8456	0.0934	0.4576
+*SM*+*A*+*AC*	**0.8571**	**32.1573**	**0.0973**	0.3869	**0.8288**	**29.1016**	**0.1078**	**0.5119**

参 考 文 献

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144.

[2] He Z, Zuo W, Kan M, et al. Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 2019, 28(11): 5464-5478.

[3] Choi Y, Choi M, Kim M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8789-8797.

[4] Huang Q, Zhang J, Zhou W, et al. Initiative defense against facial manipulation//Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual, 2021, 35(2): 1619-1627.

[5] Ruiz N, Bargal S A, Sclaroff S. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems//Proceedings of Computer Vision–ECCV 2020 Workshops, Part IV 16, 2020: 236-251.

[6] Ji S L, Du T Y, Deng S G, et al. Robustness Certification Research on Deep Learning Models: A Survey. Chinese Journal of Computers, 2022, 45(1): 190-206. (in Chinese)
(纪守领, 杜天宇, 邓水光, 等. 深度学习模型鲁棒性研究综述. 计算机学报, 2022, 45(1): 190-206.)

[7] Yeh C Y, Chen H W, Tsai S L, et al. Disrupting image-translation-based deepfake algorithms with adversarial attacks//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, Snowmass Village, USA, 2020: 53-62.

[8] Fang Z, Yang Y, Lin J, et al. Adversarial attacks for multi target image translation networks//2020 IEEE International Conference on Progress in Informatics and Computing, Shanghai, China, 2020: 179-184.

[9] Qiu H, Du Y, Lu T. The framework of cross-domain and model adversarial attack against deepfake. Future Internet, 2022, 14(2): 46.

[10] Huang H, Wang Y, Chen Z, et al. CMUA-watermark: A cross-model universal adversarial watermark for combating deepfakes//Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, Virtual, 2022, 36(1): 989-997.

[11]Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations, Vancouver, Canada, 2018.

[12]Baluja S, Fischer I. Adversarial transformation networks: Learning to generate adversarial examples. arXiv preprint arXiv:1703.09387, 2017.

[13]Xiao C, Li B, Zhu J Y, et al. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 3905-3911.

[14]Dong J, Wang Y, Lai J, et al. Restricted black-box adversarial attack against deepfake face swapping. IEEE Transactions on Information Forensics and Security, 2023, 18: 2596-2608.

[15]Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans. Advances in Neural Information Processing Systems, 2017, 30: 1-11.

[16]Liu M, Ding Y, Xia M, et al. STGAN: A unified selective transfer network for arbitrary image attribute editing//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 3673-3682.

[17]Ying Q, Zhou H, Zeng X, et al. Hiding images into images with real-world robustness//Proceedings of 2022 IEEE International Conference on Image Processing, Bordeaux, France, 2022: 111-115.

[18]Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//Proceedings of 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Part III 18. Munich, Germany, 2015: 234-241.

[19]Shin R, Song D. JPEG-resistant adversarial images//Proceedings of NIPS 2017 Workshop on Machine Learning and Computer Security, Long Beach, USA, 2017, 1: 8-14.

[20]Che Z, Borji A, Zhai G, et al. A new ensemble adversarial attack powered by long-term gradient memories//Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York City, USA, 2020, 34(04): 3405-3413.

[21]Yu Y, Gao X, Xu C Z. Lafeat: Piercing through adversarial defenses with latent features//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 5735-5745.

[22]Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3730-3738.

[23] Guo X, Kan M, He Z, et al. Image style disentangling for instance-level facial attribute transfer. Computer Vision and Image Understanding, 2021, 207: 103205.

[24]Kim K, Park S, Jeon E, et al. A style-aware discriminator for controllable image translation//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 18239-18248.

[25]Wang Z, Bovik A C, Sheikh H R, et al., Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[26]Zhang R, Isola P, Efros A A, et al., The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 586-595.

[27]Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2414-2423.

Chen Beijing, Ph. D., Professor. His research interests include digital forensics, color image processing.

Zhang Haitao, M.S. His research interest is digital image forensics.

Li Yuru, M.S. Candidate. Her research interest is digital image anti-forensics.

Background

With the maturation of deep generation technology, the face images generated by the facial attribute editing models appear to mix the spurious with the genuine. To alleviate the risk of malicious use of face tempering techniques, researchers have proposed passive image forensic methods to detect the authenticity of face images. Although the current passive forensic detectors based on deep neural network have achieved considerable performance, they are too passive to eliminate the losses already caused by maliciously tempered faces. In order to avoid losses caused by fake faces, forensic researchers attempt to actively defend against existing tempering models, disrupting face tempering through active defense methods.

This paper belongs to the field of digital image forensics and focuses on solving the problem of passive forensics being difficult to eliminate the losses caused by facial image tampering. At present, there is only several studies on active

defense, and most of these works have used the gradient based adversarial attack algorithm PGD attack, whose defense performance is inefficient and unsatisfactory. Although the generation model based algorithm such as the baseline framework of this paper has improved the efficiency, but it is still insufficient in terms of the transferability and robustness of adversarial perturbation. Therefore, by optimizing the two-stage training architecture and its loss function and introducing an auxiliary classifier, this paper improves the transferability and robustness of the perturbation on the facial attribute editing model.

Our research group has already worked on the fields of digital forensics, image watermarking, and color image processing. These works have been published in some international journals, such as IEEE TIP, IEEE TSP, IEEE TCSVT, IEEE TMM, etc.

This research is supported by the National Natural Science Foundation of China (62072251，62072250).