计算机科学
COMPUTER SCIENCE

基于“隐形面具”的可逆人脸隐私保护方法

郑旭, 黄想杰, 杨杨

引用本文

郑旭, 黄想杰, 杨杨. 基于“隐形面具”的可逆人脸隐私保护方法[J]. 计算机科学, 2025, 52(5): 384-391.

ZHENG Xu, HUANG Xiangjie, YANG Yang. Reversible Facial Privacy Protection Method Based on “Invisible Masks” [J]. Computer Science, 2025, 52(5): 384-391.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100066

基于“隐形面具”的可逆人脸隐私保护方法

郑 旭[1] 黄想杰[1] 杨 杨[1,2]
1 安徽大学电子信息工程学院 合肥 230601
2 合肥综合性国家科学中心人工智能研究院 合肥 230026
(19276326064@163.com)

摘 要 随着人工智能和计算机视觉技术的快速进步,人脸信息已经被广泛应用于智能安防、金融支付和社交媒体等多个领域。这些采集的人脸信息一旦被泄露或被不法分子非法售卖,就会造成严重后果。因此,如何防止采集的原始人脸数据库被恶意窃取从而进行非法训练和非法识别,是亟待解决的问题。对此,提出了一种基于“隐形面具”的可逆人脸隐私保护方法。该对抗人脸若被恶意窃取,可使未授权人脸系统错误识别,对于被授权用户,可以在摘除“隐形面具”后恢复原始人脸信息,保证授权人脸系统正确识别,从而达到保护人脸数据库的目的。实验结果表明,该方法生成的对抗人脸具有更高的视觉质量,与原始人脸的平均 PSNR 在无攻击层下可以达到 55dB,并且使未授权系统错误识别率达到 99.6%。同时,该方法实现了可逆恢复人脸,恢复人脸具有更高的视觉质量,与原始人脸的平均 PSNR 达到 61dB,并且使授权系统正确识别率达到 99.8%。实验证明了该方法可以有效地保护人脸数据库。

关键词:深度学习;隐形面具;对抗样本;人脸数据库保护;视觉变换;可逆人脸隐私保护
中图分类号 TP311

Reversible Facial Privacy Protection Method Based on “Invisible Masks”

ZHENG Xu[1], HUANG Xiangjie[1] and YANG Yang[1,2]
1 School of Electronics and Information Engineering, Anhui University, Hefei 230601, China
2 Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230026, China

Abstract With the rapid progress of artificial intelligence and computer vision technology, facial information has been widely used in smart security, financial payment, and social media, etc. Once the collected facial information is leaked or illegally sold by unscrupulous individuals, it will cause adverse consequences. Therefore, how to prevent the original facial database from being illegally accessed and trained by malicious parties, and how to prevent illegal recognition, is an urgent issue that needs to be solved. Therefore, a reversible facial privacy protection method based on “invisible mask” is proposed. If the adversarial facial image is illegally accessed, it will cause the unauthorized facial recognition system to incorrectly recognize, and for authorized users, the original facial information can be recovered by removing the “invisible mask”, ensuring that the authorized facial recognition system can correctly recognize, thus achieving the purpose of protecting the facial database. Experimental results show that the method generates adversarial facial images with higher visual quality, the average PSNR between the adversarial facial image and the original facial image without attack layer can reach 55 dB, and the false recognition rate of the unauthorized system can reach 99.6%. At the same time, the method realizes reversible recovery of facial images, the average PSNR of the recovered facial image is 61dB, and the correct recognition rate of the authorized system can reach 99.8%. Therefore, the proposed method can effectively protect the facial database.

Keywords Deep learning, Invisible masks, Adversarial examples, Facial dataset protection, Visual transformation, Reversible facial privacy protection

1 引言

人脸信息作为一种独特且不可更改的生物特征数据,一旦被泄露或滥用,可能会引发严重的隐私侵犯和安全问题。美国人脸识别公司 Clearview AI[1] 曾因私自使用未授权的人脸进行非法训练,导致用户的人脸信息被滥用。因此,保护人

到稿日期:2024-11-11 返修日期:2024-12-28
基金项目:国家自然科学基金(62272003);安徽省高等学校自然科学基金(KJ2021A0016)
This work was supported by the National Natural Science Foundation of China(62272003) and Natural Science Foundation of Anhui Provincial Colleges and Universities(KJ2021A0016).
通信作者:杨杨(sky_yang@ahu.edu.cn)

脸隐私至关重要。随着网络技术的发展，人们热衷于将人脸图像上传至社交媒体平台，但是这些公开的人脸图像可能会被被未授权方窃取进行非法售卖和非法训练。因此，对于未授权方，人们希望即使他们窃取了人脸数据库也无法进行识别，而对于授权用户，则希望他们获取到原始人脸后进行合法识别和合法训练。针对上述需求，从人眼视觉角度来看，人们在追求保护人脸隐私的同时也在追求实现原始人脸高保真恢复。从机器视觉角度来看，人们希望对于未授权方，即使其窃取到人脸数据库也无法非法训练和非法识别，对于被授权方，希望通过高保真恢复原始人脸，使其可以被正确训练和正确识别。

传统的人脸隐私保护方法中，基于视觉混淆的技术是一种常见方法。该方法通过改变图像的视觉特征，使其难以被辨认或理解。该类技术经常被用于保护人脸隐私，通过混淆处理人脸面部信息，使原始的人脸特征难以被识别。Mritunjay 等[2]提出了一种可运行在低资源嵌入式平台的人脸去识别系统，该方法基于高斯模糊实现了近乎实时的身份去识别。Zhang 等[3]提出了另一种靠屏蔽来保护隐私的解决方案，通过摄像机内置的红外成像传感器捕获的热成像信息标识面部，可以实现准确、实时的人脸遮蔽。Letournel 等[4]提出了具有表情保留的面部去识别方法，基于变分自适应过滤机制，保留了关键的表情识别特征，同时隐藏了面部身份。基于视觉混淆的人脸隐私保护有效降低了人脸面部信息的可用性，但通常存在人脸可用性缺失的问题。因此，传统的人脸隐私保护无法满足从人眼视觉角度实现人眼视觉不可察觉性的需求。

随着深度学习技术的迅速发展，人们开始采用对抗样本技术保护人脸隐私。对抗样本是指通过在原始样本中添加细微且难以被人类感官察觉的扰动，诱导深度学习模型产生错误的分类结果。Goodfellow 等[5]在 2015 年发现，人脸识别系统广泛使用的深度神经网络对对抗样本具有明显的脆弱性。此后，许多研究人员开始尝试向人脸图像中注入精心设计的微小噪声，生成看似正常实则能够欺骗深度学习模型的对抗样本，从而保护用户的面部信息不被非法识别或滥用。此外，Yin 等[6]提出了一种名为 Adv-Makeup 的对抗性人脸生成方法，通过在图像上添加眼影实现对人脸识别系统的不可察觉且可转移的攻击[7]。

随着网络技术的发展，人们在追求原始人脸视觉可读性的同时，也希望能够通过恢复原始人脸保证被授权用户的权益。随着 RDH 技术[8]的发展，Liu 等[9]提出的 RAE(Reversible Adversarial Example)在 RDH 基础上生成了可逆的对抗样本。随着可逆神经网络 INN 的发展，Chen 等[10]首次提出了一种基于可逆神经网络的方法 RAEG(Reversible Adversarial Example Generator)。该方法在 Liu 等[9]提出的 RAE 方法上进行改进，可以有效恢复原始人脸。然而，这些方法的对抗性和可逆性有待提高。同时，可逆神经网络要求前向阶段与后向阶段共享完全一样的参数，导致模型训练较为复杂。

综上所述，针对传统的基于视觉混淆的人脸隐私保护方法无法实现人眼视觉不可察觉性，以及目前基于对抗样本的人脸隐私保护方法对抗性和可逆性有待提高的问题，本研究旨在保证原始人脸数据库视觉不失真的前提下，达到对于未授权方会错误训练和错误识别，而对于被授权方则通过高保真可逆恢复原始人脸实现正确训练和正确识别的目的，如图 1 所示，故本文提出了基于“隐形面具”的可逆人脸隐私保护方法。所谓“隐形面具”即通过对抗样本技术生成一张“隐形面具”，将其佩戴于原始人脸上并生成对抗人脸，其中视觉无法区分对抗人脸与原始人脸。该方法使用 U-Net 和视觉变换[11]的网络结构，首先将原始人脸送入“隐形面具”生成网络生成“隐形面具”，然后通过对抗人脸生成网络将“隐形面具”戴在原始人脸上并生成对抗人脸。在戴“隐形面具”与摘除“隐形面具”之间模拟攻击者攻击，来提高本文模型的鲁棒性。最后通过原始人脸恢复网络摘除“隐形面具”恢复人脸。

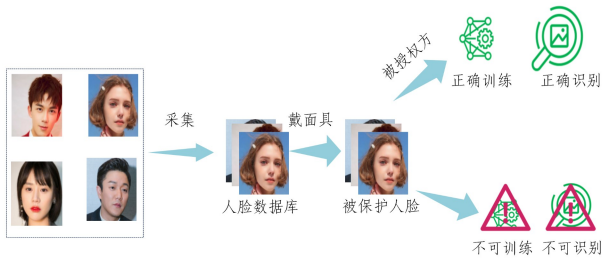

图 1　基于“隐形面具”的人脸隐私保护方法应用场景

Fig. 1　Application scenarios of face privacy protection method based on “invisible mask”

本文的主要创新点如下：

1)提出了一种基于“隐形面具”的可逆人脸隐私保护方法。可以在保护原始人脸数据库视觉不失真的前提下，防止其被非法训练和非法识别，从而保护了人脸数据库。

2)通过生成对抗网络技术，本文提出了对抗人脸生成网络。从机器视觉的角度来看，生成的对抗人脸可以使未授权方错误识别从而满足未授权方无法非法识别的特性；同时可以高保真恢复原始人脸信息，从而保证了授权模型的正确识别。

3)为了提高本文模型的鲁棒性，实验模拟攻击者进行攻击，如高斯噪声、JPEG 压缩[12]等，利用神经网络模拟这些操作，使本文模型能够抵抗这些攻击。因此，经过高斯噪声、JPEG 压缩等攻击的对抗人脸依然可以高保真恢复人脸，提高了该方法的鲁棒性；并且 U-Net 的结构编解码网络是独立的，解码网络不会过多受到噪声层的干扰。

2　相关工作

2.1　对抗样本

2015 年 Goodfellow 等[5]发现，人脸识别系统广泛使用的深度神经网络对对抗样本具有明显的脆弱性。此后，许多研究人员开始向人脸图像中加入微小扰动生成看似正常实则能够欺骗深度学习模型的对抗样本，从而保护用户的面部信息不被非法识别或滥用。对抗样本的概念最早是由 Szegedy 等[13]提出的，其可以向图像中添加人眼难以察觉的扰动使神经网络做出错误的判断。Carlini 等[14]提出了一种新的基于优化的对抗样本生成方法(C&W)，利用 tanh 和 arctanh 函数实现对样本空间的映射。为了优化扰动计算效率，Goodfellow 等[5]提出了一种基于优化的方法 FGSM，FGSM 根据损

失函数对当前输入图像的梯度方向更新图像，以生成对抗样本。但是这些方法都是不可逆的，无法恢复人脸，同时这些方法无法抵御攻击，从而失去了人脸隐私保护的有效性。

随着生成对抗网络的发展，Xiao 等[15]提出了基于生成的对抗攻击 AdvGAN，通过构建判别器、生成器和目标分类器学习原始样本与对抗样本分布之间的映射，生成更逼真的对抗样本。虽然 AdvGAN 可以抵御攻击，但是该方法不可逆，无法有效恢复人脸进行识别，从而无法保证授权模型的可用性。上述研究表明，虽然当前基于噪声扰动的对抗样本不可逆，但可以进一步使其具有可逆性，以满足不同应用需求的可逆人脸隐私保护方法。

2.2 人脸隐私保护

传统的人脸隐私保护通常通过模糊、像素化、乱置、掩蔽或加密等方式对整体面部信息进行混淆处理，使隐私侵犯者无法获取清晰的人脸图像。例如，图像混淆（Obfuscation）技术是一种保护隐私或隐藏敏感信息的手段，通过改变图像的视觉特征，使其难以被辨认或理解。该类技术经常被用于保护人脸隐私，通过混淆处理人脸面部信息，使原始的人脸特征难以被识别。再如，Chinomi 等[16]提出名为 Prisurv 的掩蔽方法，该方法基于预定义好的隐私策略自适应地通过各种类型的掩模、形状甚至背景替换视频中人脸的面部区域，从而实现隐私保护。但是这些方法都使得原始人脸无法使用。

随着深度学习的发展，You 等[17]提出了可逆的面部图像的加密方法，可以同时实现可逆恢复人脸以及面部隐私保护。但这些方法同样存在原始人脸可读性低的缺点，无法有效利用原始人脸。随着可逆神经网络的出现，基于可逆神经网络的方法可以增强原始人脸的可读性，同时实现可逆恢复原始人脸。例如，Chen[10]等通过可逆神经网络（INN）方法 RAEG 生成可逆的对抗样本，可以以较小的损失恢复原始图像。然而，可逆框架给结构施加了一些约束，要求编解码网络要完全一致，这给训练网络带来了一定困难。此外，Yang 等[18]利用可逆神经网络对人脸进行隐私保护，提出了 IMN 方法，但是由于 INN 要求编解码网络完全一致，因此该方法不具备鲁棒性。

综上所述，研究可逆的且具备鲁棒性的对抗人脸隐私保护方法是当前的重要课题。

3 算法

3.1 算法概述

本文提出的基于“隐形面具”的可逆人脸隐私保护方法流程图如图 2 所示，它由预训练好的“隐形面具”生成器、对抗人脸生成网络、模拟攻击层、原始人脸恢复网络和预训练好的目标分类网络组成。首先，原始人脸 X_0 通过“隐形面具”生成器生成一个与原始人脸 X_0 在视觉上看起来一样、但使分类网络实现错误分类的“隐形面具” X_s。“隐形面具” X_s 与原始人脸 X_0 作为输入，通过对抗人脸生成网络给原始人脸 X_0 戴上“隐形面具”，从而生成对抗人脸 X_t。为了增强该模型的鲁棒性，在发送方和接收方之间引入一个攻击层模拟攻击者对对抗人脸 X_t 进行攻击，得到受攻击的对抗人脸 X_d。对于被授权方，将对抗人脸 X_d 经过原始人脸恢复网络，摘除“隐形面具”，实现视觉上近似无损地恢复出原始人脸信息得到 X_{re}。为了监督“隐形面具”生成器的训练以及对对抗人脸 X_t、恢复人脸 X_d 进行识别，最后在该网络中引入人脸匹配器 FaceNet 作为人脸分类网络。

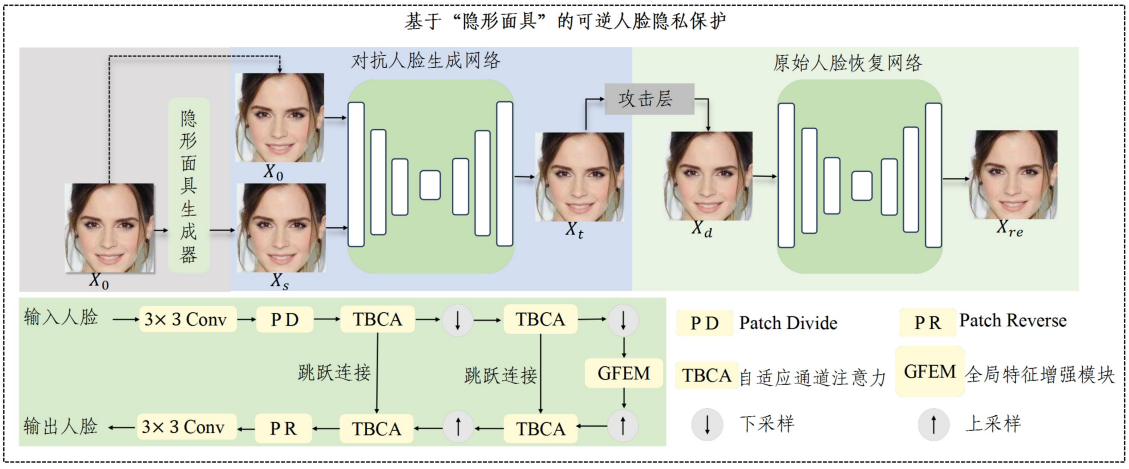

图 2 本文方法流程图

Fig. 2 Flowchart of our method

1）“隐形面具”生成器

生成对抗网络（GAN）在生成高度逼真的人脸图像方面取得了显著成就，因此本文用其生成具备对抗性隐私保护效果的“隐形面具”。具体结构如图 3 所示，给定一张原始人脸 X_0 作为输入，生成器输出一张噪声图作为对抗噪声 $G(X_0)$[19]。“隐形面具”可以使未授权分类网络错误分类，其计算式为：

$$X_s = X_0 + G(X_0) \tag{1}$$

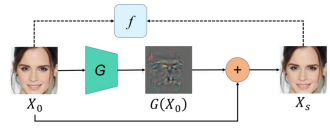

图 3 “隐形面具”生成器结构

Fig. 3 Generator structure of “invisible mask”

2）对抗人脸生成网络

上文介绍了“隐形”面具 X_s 的生成过程，接下来将介绍如

何把“隐形面具”戴在原始人脸X_0上，从而生成能抵抗非法分类网络分类且人眼可正确识别的对抗人脸X_t。本文中戴“隐形面具”的过程通过对抗人脸生成网络完成。如图2所示，对抗人脸生成网络首先将原始人脸$X_0 \in R^{3\times H\times W}$与“隐形面具”$X_s \in R^{3\times H\times W}$作为输入，在一维通道上将两个3通道图像拼接成6通道，然后采用3×3大小的卷积核来提取图像的低级特征X_f，通过Divide patch模块将X_f分成N×N大小的小块，小块的数量为$\frac{HW}{N^2}$，大小为N^2。对抗人脸生成网络是一个U型的网络结构，与U-Net的网络结构一致，由下采样和上采样组成。在下采样部分采用4×4大小的卷积核，步幅为2，通道数增加一倍，特征图分辨率降低一半。由于视觉变换会导致一些粗块的嵌入，进而造成局部信息的丢失，因此本文在每个下采样和上采样部分均嵌入了一个自适应通道注意力模块TBCA[20]，该模块的作用是捕获局部信息。引入自适应通道注意力模块来动态分配调整通道之间的权重，同时通过视觉变换利用窗口注意力机制W-MSA[20]对局部特征进行建模。由于本文方法通过视觉变换利用窗口注意力机制关注局部特征，这会限制全局信息捕获，因此在上采样恢复人脸信息前，本文嵌入了一个全局特征增强模块GFEM来增强全局信息。经过GFEM增强全局信息之后，在上采样阶段，特征图分辨率提高了一倍，然后将上采样特征与U-Net下采样阶段相应特征通过跳接连接并输入TBCA进行图像恢复，重建人脸信息。最后应用3×3大小的卷积核来提取高级特征并映射回对抗人脸X_t信息，最后输出得到对抗人脸X_t。

3)模拟攻击层

为了增加模型的鲁棒性，在发送方和接收方之间引入常见的图像处理操作(如高斯模糊、JPEG压缩[14]等)模拟攻击层。经过对抗人脸生成网络后，输出对抗人脸X_t，之后添加噪声层，模拟攻击者攻击来重建原始人脸，使模型抵御攻击，从而提高本文模型的鲁棒性。

4)原始人脸恢复网络

此阶段的目的是摘除“隐形面具”，恢复原始人脸信息，得到恢复人脸X_{re}，从而使被授权用户可以恢复原始人脸。本文通过原始人脸恢复网络摘除“隐形面具”。由于对抗人脸生成网络与原始人脸恢复网络的结构一样，是反过程，因此这部分不再赘述。下文将详细介绍本文模型中引入的自适应通道注意力TBCA与全局特征增强模块GFEM。

(1)自适应通道注意力结构：由于对抗人脸生成网络本质上是一个以图藏图的网络，该网络将视觉变换应用于图像隐藏，导致一些粗块嵌入，从而造成局部信息丢失，而局部信息对于图像隐藏来说至关重要，因此在该部分引入自适应通道注意力模块来动态分配调整通道之间的权重。

给定$N\times N$大小的小块，首先将这些小块重塑成特征图X_f，该部分采用两个1×1的卷积层来生成每个通道的权重，同时用一个非线性层来获取通道之间的信息X_{channel}，然后使用全局池化、激活函数、两层MLP来获得通道权重X_{weight}，之后将通道信息$X_{\text{channel}} \in R^{C\times H\times W}$与通道权重相乘得到通道偏置$X_{\text{bias}}$，并将其添加到原始特征图中得到新的令牌特征图。最后，通过偏置来动态调整通道之间的强度。由于视觉变换Swim Transfomer[20]采用窗口注意力机制关注局部特征，会限制全局信息的捕获，因此采用全局特征增强模块(Global Feature Enhancement Module，GFEM)对全局特征进行建模，最后经过上采样以及3×3大小的卷积核对输入人脸进行重建，从而得到输出的人脸。

(2)全局特征增强模块：上文中提到由于视觉变换利用窗口注意力机制关注局部特征，会限制全局信息的捕获，因此，在上采样之前嵌入一个全局特征增强模块来增强全局信息。

全局特征增强模块利用位置编码生成器和多个视觉Transfomer Blocks模块来构建全局特征。Chu等[21]认为卷积可以通过零填充捕获位置信息，因此在该模块中采用以深度卷积DWconv作为PEG生成位置偏码的特征提取方式。在条件位置编码嵌入之后，通过多个TransfomerBlocks来捕获全局特征。该过程计算式如下：

$$\bar{X}^l = MSA(LN(X^{l-1})) + X^{l-1} \tag{2}$$

$$X^l = MLP(LN(\bar{X}^l)) + \bar{X}^l \tag{3}$$

其中，$\bar{X}^l$和X^l分别表示的MSA模块和MLP模块的输出特征，MSA是多头注意力机制，MLP是多层感知器，l为块的数量。

5)目标分类网络

“隐形面具”生成器模块的目的是生成一张“隐形面具”，它要求在视觉上与原始人脸相似并且可以使人脸分类网络错误分类，因此采用人脸匹配器来监督生成器的训练，这里将FaceNet[22]中的主干特征提取网络用作特征提取器。与此同时，在戴“隐形面具”与摘除“隐形面具”的过程中，采用FaceNet来判断经过对抗人脸生成网络生成的受攻击的对抗人脸X_d使分类网络错误分类的效果，以及经过原始人脸恢复网络输出的恢复人脸X_{re}使分类网络正确分类的效果。本文在原始人脸X_0上预先训练好人脸分类网络，预训练的人脸分类网络也达到了预期的效果。

3.2　损失函数

Charbonnier Loss[23]是一种回归问题的损失函数，尤其在图像隐写、图像恢复和图像去噪等任务中表现良好。因此，对于戴“隐形面具”与摘除“隐形面具”这两个任务，本文采用Charbonnier Loss。Charbonnier Loss是一种改进的L_1损失与L_2损失的折中方法，目的是结合两种损失函数的优点，同时避免它们各自的缺点。Charbonnier Loss的标准定义如下：

$$Loss(x) = \sqrt{(x-y)^2 + \varepsilon^2} \tag{4}$$

其中，x和y代表输入的两幅图像，ε是超参数，避免损失函数为0。

整体损失函数包含对抗人脸生成的生成损失函数L_g和原始人脸恢复的恢复损失函数L_{re}。总损失函数定义如下：

$$L_{\text{total}} = \alpha L_g + \beta L_{re} \tag{5}$$

对于超参数α和β的比例问题对实验的影响，将在4.4节深入探究。

1)对抗人脸生成损失函数L_g

戴“隐形面具”(即对抗人脸生成这一任务)的目的是生成对抗人脸X_t，使得对抗人脸X_t在视觉上与原始人脸X_0尽可能

相似，同时能使分类网络错误识别。由于隐形面具X_s在视觉上与原始人脸X_0相似同时还能使分类网络错误分类，所以本文采用 Charbonnier Loss 来最小化对抗人脸X_t与隐形面具X_s之间的范数，完成对抗人脸生成的任务。对抗人脸生成损失函数计算式如下：

$$L_g = \sqrt{\| X_s - X_t \|^2 + \varepsilon^2} \tag{6}$$

其中，X_s为“隐形面具”，X_t为对抗人脸。

2）原始人脸恢复损失函数L_{re}

摘除“隐形面具”的目的是恢复原始人脸信息得到X_{re}，同时保证X_{re}在视觉上与原始人脸X_0相似并且能够被人脸分类网络正确分类。所以，在该任务中，本文仍采用 Charbonnier Loss 来最小化恢复人脸X_{re}与原始人脸X_0之间的范数，完成原始人脸恢复的任务。原始人脸恢复损失函数计算式如下：

$$L_{re} = \sqrt{\| X_o - X_{re} \|^2 + \varepsilon^2} \tag{7}$$

其中，X_o为原始人脸，X_{re}为恢复人脸，参数 ε 仍是为了避免损失函数为 0。

4 实验

4.1 实验设置

本文实验采用 CASIA-WebFace[24] 和 LFW（Labeled Faces in the Wild）[25] 这两个数据集。CASIA-WebFace 是公开的人脸数据集，由中科院 CASIA 创建，该数据集包含大量的人脸图像以及对应标签。本文随机从中选取 900 张不相关的人脸用于训练，另外选择 100 张人脸用于验证。LFW 是一个广泛使用的人脸识别数据库，其中包含大量的真实世界的人脸图像，本文在 LFW 上进行模型的测试，进而评估本文模型的优劣。为了更加客观公平地进行对比实验，所有的对比实验都是在相同人脸数据库中进行。

在评价指标方面，本文使用峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）、结构相似度指数（Structure Similarity Index Measure，SSIM）[26]、识别率（Attack Accuracy，ACC）来评价对抗人脸X_t和恢复人脸X_{re}的视觉质量、对抗人脸X_t的分类效果以及恢复人脸X_{re}的分类效果。PSNR 和 SSIM 越大，代表视觉质量越好。对抗人脸X_t的识别率 ACC[15] 越低，说明使分类网络错误分类的效果越好，从而说明人脸隐私保护的效果越好。恢复人脸X_{re}的识别率 ACC 越高，说明恢复人脸使分类网络正确分类的效果越好，从而说明本文模型的可逆性更好。识别率的计算式如下：

$$Attack\ accuracy = \frac{(No.\ of\ Comparisons > \tau)}{Total No.\ of\ Comparisons} \tag{8}$$

其中，τ是一个预先设置好的阈值，代表两张人脸相似度的分数阈值，本文实验中将阈值τ设置为 80，高于 80 代表人脸匹配属于同一个人。然后计算相似度分数高于 80 的人脸数量占总实验人脸数量的比值，即 ACC。

本文实验在 Linux 系统上进行，采用 NVIDIA RTX A6000 GPU 进行训练。优化器采用 Adam，批处理大小设置为 12，初始学习率设置为 2×10^{-4}，训练总轮数为 3 000。同时，在进行可逆对抗人脸网络训练之前，预训练好“隐形面具”生成器以及目标分类网络 FaceNet。预训练时使用的数据集仍是 CASIA-WebFace 和 LFW。

4.2 定性与定量分析

4.2.1 定性分析

为了更直观地表现本文方法在人脸隐私保护方面的优势，即本文生成的对抗人脸X_t可以保存原始人脸信息（“隐形面具”）且恢复人脸X_{re}的视觉质量更高，不仅可以在视觉上与原始人脸具有一致性，同时还能使未授权分类网络错误分类，我们进行了大量实验来展示本文方法的视觉质量，结果如图 4—图 6 所示。

为了从视觉上验证本文方法的优势，即生成的对抗人脸X_t、恢复人脸X_{re}与原始人脸X_o具有视觉一致性，图 4 给出了未加攻击层下，对抗人脸X_t、恢复人脸X_{re}与原始人脸X_o的主观视觉对比图。实验结果表明，生成的对抗人脸X_t与恢复的人脸X_{re}在视觉上与原始人脸X_o相似。

图 4 本文方法生成的对抗人脸和恢复人脸与原始人脸的主观对比

Fig. 4 Subjective comparison of adversarial faces, recovered faces generated by our method with original faces

为了验证本文方法在高斯噪声下的鲁棒性，图 5 展示了引入方差 $\sigma=0.05$ 的高斯噪声后生成的对抗人脸X_t、受攻击的对抗人脸X_d以及恢复人脸X_{re}与原始人脸X_o对比的主观视觉图。实验结果表明，生成的对抗人脸X_t、受攻击的对抗人脸X_d和恢复人脸X_o在视觉上与原始人脸X_o相似，从而说明本文方法可以抵御高斯噪声的攻击。

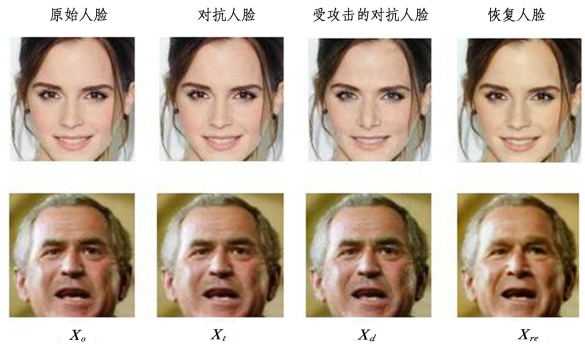

图 5 $\sigma=0.05$ 高斯噪声下，本文方法生成的对抗人脸、恢复人脸和受攻击的对抗人脸与原始人脸的主观对比

Fig. 5 Subjective comparison of the adversarial faces, recovered faces, and attacked adversarial faces generated by our method with the original faces with $\sigma=0.05$ Gaussian noise

同样，为了验证本文方法在 JPEG 压缩下的鲁棒性，图 6 展示了引入压缩因子 $Q=50$ 的 JPEG 压缩之后的主观视觉对比图。

图 6　$Q=50$ JPEG 压缩下，本文方法生成的对抗人脸、恢复人脸和受攻击的对抗人脸与原始人脸的主观对比

Fig. 6　Subjective comparison of the adversarial faces, recovered faces, and attacked adversarial faces generated by our method with the original faces with $Q=50$ JPEG compression

4.2.2　定量分析

为了更加客观地分析本文方法的优势，下面将用客观评价指标来定量分析本文方法的性能。表 1 中展示了两种攻击层以及不加攻击层下的 7 个定量指标。在视觉质量方面，在不引入噪声层下的情况，受保护人脸的视觉质量指标 $PSNR\approx 55$ dB，恢复人脸的视觉质量指标 $PSNR\approx 61$ dB。可以看出，本文方法在可逆性和视觉质量方面具有优势。与此同时，本文在引入攻击层后，虽然高斯噪声影响了生成对抗人脸的视觉质量，但是恢复的人脸视觉质量仍然良好。这也说明了本文方法可以抵御噪声攻击，具备良好的鲁棒性。在识别率方面，生成的对抗人脸识别率仅有 0.4%，这充分说明了本文方法生成的对抗人脸几乎完全可以使分类网络错误分类。与此同时，恢复人脸识别率高达 98.8%，这也充分说明了本文方法可以有效恢复原始人脸信息，使人脸分类网络正确分类。

表 1　在不同攻击层下，原始人脸、受攻击的对抗人脸、恢复人脸的识别率和视觉质量

Table 1　Recognition rate and visual quality of original face, attacked adversarial face and recovered face under different attack layers

攻击层	ACC_{ori} (X_o)	ACC_{prt} (X_d)	ACC_{rev} (X_{re})	$PSNR_{prt}$ (X_o, X_d)/dB	$PSNR_{rev}$ (X_o, X_{re})/dB	$SSIM_{prt}$ (X_o, X_d)	$SSIM_{rev}$ (X_o, X_{re})
无攻击	0.996	0.004	0.988	55.461	60.722	0.994	0.998
高斯噪声	0.996	0.004	0.988	29.241	41.040	0.920	0.969
JPEG 压缩	0.996	0.004	0.988	52.610	52.344	0.992	0.994

注：ACC_{ori} 是原始人脸识别率，ACC_{prt} 是受保护人脸的识别率，ACC_{rev} 是恢复人脸识别率，$PSNR_{prt}$ 和 $PSNR_{rev}$ 是受保护人脸与恢复人脸的 $PSNR$，$SSIM_{prt}$ 和 $SSIM_{rev}$ 是原始人脸与恢复人脸的 $SSIM$。

4.3　对比实验

考虑到本文提出的人脸隐私保护方法的两大优势，一是能确保被授权用户可用，即具有可逆性，二是生成的对抗人脸视觉质量更好，即达到所谓“隐形面具”的效果。我们将本文方法与目前主流的人脸隐私保护方法进行可逆性和鲁棒性以及视觉质量的对比，结果如表 2—表 4 所列。

更好，即达到所谓“隐形面具”的效果。我们将本文方法与目前主流的人脸隐私保护方法进行可逆性和鲁棒性以及视觉质量的对比，结果如表 2—表 4 所列。

4.3.1　可逆性和鲁棒性对比

本文所有的对比实验都是在 LFW 数据集上进行的。与本文方法进行对比的目前几种主流的隐私保护方法如下。

Yang 等提出了一种基于可逆面具网络(Invertible Mask Network，IMN)的面部隐私保护方法[18]。首先，通过所提出的掩膜生成器模型生成高分辨率的掩膜面部。然后，将掩膜面部覆盖到受保护的面部上，得到伪装面部，其中伪装面部在视觉上与掩膜面部无法区分。最后，授权用户可以从掩膜面部中去除掩膜，获得恢复后的面部，其中恢复后的面部在视觉上与受保护面部无法区分。基于噪声扰动的对抗样本 AdvFace 方法[15]对人脸添加细微扰动，生成在视觉上与原始人脸一致的对抗人脸，同时该对抗人脸可以使分类网络错误分类。同样，基于可逆神经网络的 RAEG 生成可逆的对抗样本[10]，通过轻微改变图像，使分类网络错误分类。从表 2 中可以看出，IMN 虽然可以恢复人脸但容易受到噪声层的干扰，在引入噪声层后，无法提取原始人脸。同样，AdvFace 生成的对抗人脸虽然可以抵御噪声攻击，但是该方法不可逆，无法恢复人脸。本文方法与 RAEG(可逆的对抗样本生成)都具有鲁棒性和可逆性。因此，接下来会着重讨论本文方法与 RAEG 生成的对抗人脸的视觉质量、恢复人脸的视觉质量对比。表 4 中列出了与 RAEG 所生成的受保护人脸与恢复人脸的视觉质量对比。

表 2　本文方法与其他隐私保护方法鲁棒性和可逆性对比实验

Table 2　Robustness and reversibility comparison experiment of our method with other privacy protection methods

方法	鲁棒性	可逆性
You 等[17]	×	⊙
IMN[18]	×	⊙
AdvFace[15]	⊙	×
RAEG[10]	⊙	⊙
Ours	⊙	⊙

4.3.2　视觉质量对比

本文与主流的可逆人脸隐私保护方法[17]、基于可逆神经网络的 RAEG[10] 和 IMN[18] 进行受保护人脸、恢复人脸的主观视觉对比。由于可逆人脸隐私保护方法与 IMN 不具备鲁棒性，因此本文与上述两种方法对比时不引入攻击层。同时，本文与 RAEG 对比了生成的对抗人脸 X_t、恢复人脸 X_{re} 的客观视觉质量及攻击率。为了更加客观地与 RAEG 进行对比，本文在相同的模拟攻击下，进行大量的实验，并记录下 PSNR，SSIM 和 ACC 指标值。

图 7 展示了本文方法与 You 等[17]提出的方法、IMN[18]中生成的受保护人脸视觉质量的对比结果。可以看出，本文方法生成的受保护人脸上具有更高的视觉质量，可以更加有效地进行人脸隐私保护。

如图 8 所示，为了展现本文方法在恢复人脸及可逆性上的优势，本文与上述两种可逆的人脸隐私保护方法进行了可逆性对比。

同时，在图 8 中的两张人脸图像中，本文计算了恢复人脸

与原始人脸的 PSNR 及 SSIM 并进行了客观对比。对比结果如表 3 所列，相比其他两种可逆的人脸隐私保护方法，本文方法在可逆性上具有明显优势。

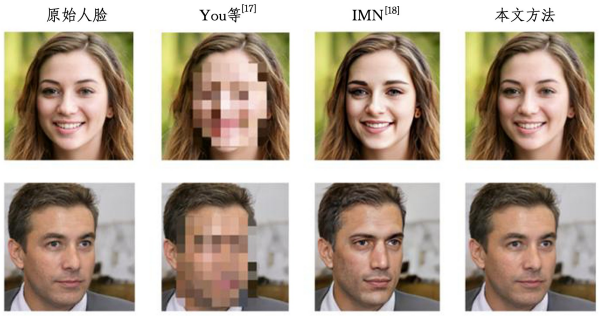

图 7 不同方法生成的受保护人脸视觉质量对比

Fig. 7 Comparison of visual quality of protected faces generated by different methods

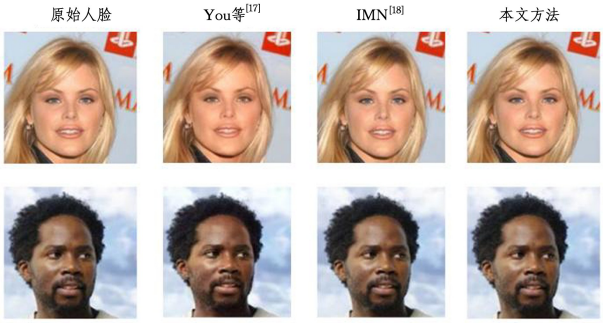

图 8 不同方法的恢复人脸视觉质量对比

Fig. 8 Comparison of visual quality of recovered faces generated by different methods

表 3 不同方法可逆性对比

Table 3 Comparison of reversibility of different methods

方法	人脸 1		人脸 2	
	$PSNR$ (X_o, X_{re})/dB	$SSIM$ (X_o, X_{re})	$PSNR$ (X_o, X_{re})/dB	$SSIM$ (X_o, X_{re})
You 等[17]	36.62	0.989	35.64	0.989
IMN[18]	50.91	0.997	52.72	0.997
Ours	59.61	0.998	58.51	0.998

由于上述两种可逆人脸隐私保护方法不具备鲁棒性，因此本文与 RAEG 进行视觉质量和识别率的客观对比。实验结果如表 4 所列，可以看出，本文方法得到的受攻击对抗人脸的视觉质量会更高，$PSNR \approx 29$ dB，$SSIM \approx 0.92$。然而，为了提高受攻击对抗人脸的视觉质量和鲁棒性，本文牺牲了一部分恢复人脸的视觉质量，从而获取了更好的视觉质量和更高的错误识别率。综上所述，本文提出的人脸隐私保护方法具有更好的视觉质量，同时还能使对抗人脸被未经授权模型错误识别。

表 4 本文方法与 RAEG 的视觉质量对比

Table 4 Comparison of visual quality of our method and RAEG

方法	$PSNR_{prt}$ (X_o, X_d)/dB	$PSNR_{rev}$ (X_o, X_{re})/dB	$SSIM_{prt}$ (X_o, X_d)	$SSIM_{rev}$ (X_o, X_{re})	ACC_{prt} (X_d)
RAEG[10]	27.120	43.542	0.901	0.996	0.093
Ours	29.241	41.040	0.920	0.969	0.004

4.4 参数设置

在损失函数中，超参数 $\alpha:\beta$ 用来平衡不同损失函数的权值。这里讨论参数 α 和 β 对模型性能的影响。通过调整 α 到合适的值，可以使受攻击的对抗人脸生成的视觉质量更好，在视觉层面上难以区分。考虑到首先应该使受攻击的对抗人脸先达到良好的视觉质量，所以 α 对实验结果的影响更大。如表 5 所列，当超参数 $\alpha:\beta$ 比例为 3:1 时，可以让受攻击的对抗人脸与恢复人脸的视觉质量达到平衡。因此，在实验中本文将超参数 $\alpha:\beta$ 的比例设置为 3:1。

表 5 超参数比例设置

Table 5 Hyperparameter ratio setting

$\alpha:\beta$	$PSNR_{prt}$ (X_o, X_d)/dB	$PSNR_{rev}$ (X_o, X_{re})/dB	$SSIM_{prt}$ (X_o, X_d)	$SSIM_{rev}$ (X_o, X_{re})
1:1	23.596	40.540	0.912	0.960
3:1	29.241	41.040	0.920	0.969
5:1	29.720	35.441	0.924	0.930

结束语 本文提出了一种基于“隐形面具”的可逆人脸隐私保护方法，即生成一张“隐形面具”戴在原始人脸上并生成对抗人脸，其中人眼无法区分对抗人脸与原始人脸，并且生成的对抗人脸可以使分类网络错误分类。对抗人脸可以在摘除“隐形面具”后恢复原始人脸信息，保证授权系统的正确识别。同时，本文在训练过程利用常见的图像处理操作为模拟攻击方进行攻击，并要求经过这些攻击的人脸依然能使分类网络错误分类，提高了对抗人脸的鲁棒性。大量的实验证明，本文方法可以在保护人脸数据库的同时确保被授权用户的使用。

参 考 文 献

[1] BORENSTEIN J, AYANNA H. Emerging challenges in AI and the need for AI ethics education[J]. AI and Ethics, 2021, 1(1): 61-65.

[2] MRIT M, NARAYANAN P. The de-identification camera [C]// Proceedings of the 2011 Third National Conference on Computer Vision and PatternRecognition. 2011: 192-195.

[3] ZHANG Y, LU Y, NAGAHARA H, et al. Anonymous camera for privacy protection[C]// Proceedings of the 22nd International Conference on Pattern Recognition. 2014: 4170-4175.

[4] LETOURNEL G, BUGEAU A, DOMENGER J P. Face de-identification with expressions preservation[C]// ProcEedings of the Internationak Conference on Image Processing. 2015: 4366-4370.

[5] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2015.

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

[7] JIA X J, WEI X X, CAO X C, et al. Comdefend: An efficient image compressionmodel to defend adversarial examples[C]// Proceedingsof the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6084-6092.

[8] ZHANG X P. Reversible data hiding with optimalvalue transfer [J]. IEEE Transactions on Multimedia, 2012, 15(2): 316-325.

[9] LIU J Y，HOU D D，ZHANG W M，et al. Reversible adversarial examples[J]. arXiv：1811. 00189，2018.

[10] CHEN K J，CHEN K J，ZENG X H，et al. Invertible image dataset protection[J]. arXiv：2021，14420，2021.

[11] KE X，WU H Q，GUO W Z. StegFormer：Rebuilding the Glory of Autoencoder-Based Steganography [C] // Proceedings of theAAAI Conference on Artificial Intelligence. Vancouver，Canada，2024：2723-2731.

[12] ZHU J，RUSSELL K，JUSTIN J，et al. Hidden：Hiding data with deep networks [C] // European Conference on Computer Vision. Munich，Germany，2018：657-672.

[13] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. In-triguing properties of neural networks[J]. arXiv：1312. 6199，2013.

[14] CARLININ，WAGNER D. Towards evaluating the robustness of neural networks[J]. IEEE Symposium on Security and Privacy. San Francisco. USA，2017：39-57.

[15] XIAO C，LI B，ZHU J，et al. Generating adversarial examples with adversarial networks[J]. arXiv：1801. 02610，2018.

[16] CHINOMI K，NITTA N，ITO Y. PriSurv：Privacy protected video surveillance system using adaptive visual abstraction[C] // Proceedings of the 14th International Conference on Advances in Multimedia Modeling. Berlin：Springer，2008：144-154.

[17] YOU Z，LI S，QIAN Z，et al. Reversible privacy-preserving recognition[C] // 2021 IEEE International Conference on Multimedia and Expo(ICME). IEEE，2021：1-6.

[18] YANG Y，HUANG Y，SHI M，et al. Invertible Mask Network for Face Privacy Preservation[J]. Information Sciences，2023，629：566-579.

[19] DEBAYAND，ZHANG J B，JAIN A. Advfaces：adversarial face synthesis[J]arXiv：1908. 05008，2019.

[20] LIN Y，CAO Y，HU H. Swin transformer：Hierarchical vision transformer using shifted windows [C] // Proceedings ofthe IEEE/CVFInternational Conference on ComputerVision. Montreal，Canada，2021：10012-10022.

[21] CHU X，TIAN Z，ZHANG B，et al. Conditional positional encodings for vision transformers[J]. arXiv：2102. 10882，2021.

[22] SCHROFF F，KALENICHENKO D，PHILBIN J. Facenet：A unified embedding for face recognition and clustering. [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition. Santa Barbara，USA，2015：815-823.

[23] CHARBONNIE R，BLANC-FERAUD L ，AUBERT G，et al. Two deterministic half-quadratic regularization algorithms for computed imaging[C] // Proceedings of 1st International Conference on Image Processing. 1994：168-172.

[24] YI D，YANG M，WU Y M. CASIA-WebFace：A Web Face Database for Face Recognition[C] // IEEE Conference on ComputerVision and Pattern Recognition (CVPR). Columbus，USA，2014.

[25] HUANG G B，RAMESH M，LEARNED E. Labeled Faces in the Wild：A Survey of Face Recognition in Unconstrained Environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)，2008，12(30)：2127-2140.

[26] WANG Z，BOVIK A，SHEIKH H R. Image quality assessment：From error visibility to structural similarity[J]. IEEE Transactions on Image Processing. 2004，13(4)：600-612.

ZHENG Xu，born in 2000，postgraduate. His main research interest is information hiding.

YANG Yang，born in 1980，professor，is a member of CCF(No. H3489M). Her main research interests include information hiding，quantum artificial intelligenceand image quality assessment.

（责任编辑：何杨）