Journal of Wuhan Polytechnic

基于卷积神经网络人脸识别系统的安全性研究

陈铿锵[1]，莫耀华[2]

（1. 汕尾职业技术学院 网络与实训中心，广东 汕尾 516600；
2. 广州大学 计算机科学与网络工程学院，广东 广州 510006）

摘 要：针对传统神经网络中的对抗样本问题，以人脸识别为例，讨论了基于传统神经网络的人脸识别技术的潜在安全风险，以及在口罩对抗样本生成并提供防御的方法。主要研究如下：第一，所提出的解决人脸识别技术安全问题的方法，产生了具有模糊和无偏差面具的图像，人脸识别模型将具有模糊面具的脸误判为目标人物。实验结果表明，对抗样本在一定程度上模仿了被隐藏者戴口罩的情形，成功率为 90.43%。第二，提出 DeFense-EC 保护方法来掩盖虚假图案，并使用图像重建技术来恢复虚假图案和抑制噪声。对不同数据集的测试证实了 DeFense-EC 重建的图像质量很高，其可靠性达到 92.32%。讨论传统的基于神经网络的人脸识别方法中的不利选择问题，并研究不利面具选择的风险和 DeFense-EC 保护方法的有效性。卷积网络的日益普及使逆向选择问题成为一个重要的研究领域，对逆向选择的彻底调查将促进卷积网络的发展和使用。

关键词：卷积神经网络；对抗样本；人脸识别；口罩攻击；图像修复

中图分类号：TP391.41;TP183　　**文献标识码：**A　　**文章编号：**1671-931X（2023）04-0099-13

DOI：10.19899/j.cnki.42-1669/Z.2023.04.016

随机测试可用于评估和提高人脸识别系统在现实世界应用中的可靠性[1]。目前有两种主要的方法来创建用于人脸识别的负面模型：一种是在整个脸部添加不可见的小变形，使其不容易被人眼看到。第二种类型是人眼可以看到但不太明显的脸部部位的扰动影响。提出了一种攻击者面具的生成方法，精心设计的攻击者面具通过模拟攻击者隐藏脸部的情况，在人脸识别模型中引起错误。

人脸识别模型根据标签之间的相似性（或距离）做出识别决定，而不是根据属于类别不同人脸图像的概率值。针对图像分类器的攻击如果是针对人脸识别网络的，是无效的。换句话说，人脸识别模型决定了它在应对有深度信号的反击时比有虚假信号的反击更有效[2]。

在现实世界中，一个错误的授权可能比拒绝一个错误的面部识别模式产生更严重的后果，例如当一个陌生人用面部识别打开手机时。下面几节分析了攻击对人脸识别模型的影响。由于人脸识别模型是现成的，本研究假设攻击者知道模型的内部参数。基于改进的人脸识别模型 Arcface[3]，在公共数据集上进行了实验，以证明蠕虫侵袭法的有效性。本文详细分析了用于开发攻击保护模型的方法，展示了

收稿日期：2022-09-21

作者简介：陈铿锵（1979— ），男，广东陆丰人，汕尾职业技术学院网络与实训中心讲师，研究方向：计算机科学与技术、网络安全技术、教育信息化；莫耀华（1996— ），男，广西来宾人，广州大学计算机科学与网络工程学院 2021 级硕士研究生，研究方向：深度学习。

算法的攻击保护能力，显示了算法的有效性，并提出了关于攻击保护的初步研究。

一、基于身份特征的口罩对抗样本生成算法

（一）MI-FGSM 攻击

Dong 等人[4]提出了 MI-FGSM 来提高保护效果。为了稳定优化并避免不必要的局部峰值，该算法使用了一种增益方法，在每次迭代中累积损失函数的梯度。MI-GSM 经常被用来攻击分类模型，由此产生的干扰没有被注意到。

对于分类模型，攻击方法的具体步骤如下：直接传播：首先通过直接传播将输入数据 x 到 y，加入分类模型 $F(\cdot)$，并计算损失函数 $J(x^{'}, y;\theta)$。反向传播：计算数据 $\nabla_x J(x^{'}, y;\theta)$ 的梯度与损失的关系。用式(1)计算堆栈梯度 M_t（t 为迭代次数）。生成对抗样本：为了创建梯度信息添加扰动 δ，添加一个基于梯度矩阵的扰动 $x^{'}$，并将像素值投射到式(2)中指定的区域。使用 L1、L2 和 l_∞ 约束扰动量。

$$M_{t+1} = \mu \cdot M_t + \frac{\nabla_x J(x_t^{'}, y;\theta)}{\left\| \nabla_x J(x_t^{'}, y;\theta) \right\|_2} \quad \text{式(1)}$$

$$x_{t+1}^{'} = Clip_\varepsilon(x_t^{'} + \alpha \cdot sign(M_{t+1})) \quad \text{式(2)}$$

其中，$Clip_\varepsilon(\cdot)$ 表明误差在可接受范围内的 ε 信号函数 $sign(\cdot)$，以及一个动量减少系数 μ。

（二）空间变换网络

如果猫的图像能够被正确地识别和分类，那么这个模型就是尺度和旋转不变的。但 CNN 并不经常这样做。为了解决这个问题，Yaderberg 等人提出了一个卷积神经网络的架构模型，称为空间转换网络(SpatialTransformerNetworks，STN)[5]。该模型将输入图像转化为完整的图像，并促进其他分类和识别任务（例如，将 1-a 图像转化为 1-b 图像，等等）。

STNs 由一个定位网络、一个网格发生器和一个扫描单元组成。具体转换过程如下：(1)预测工作由本地网络以图像转换函数为中心进行；(2)网格生成和采样器应根据上一步的预测来处理图像；(3)CNN 对被改变的图像进行分类和识别。STNs 的大小是微观的，因此它们可以在不改变 CNN 结构的情况下被用于内层、复合层或其他层的后面。STNs 的工作速度非常快，不会干扰 CNN 的整体训练计划。

a. 变换前　　b. 变换后

图 1　猫的图像示例

（三）口罩对抗样本生成算法

本文提出了一种面具攻击，模拟一个人戴着面具的情况，攻击者创建一个复合面具来伪装自己，迫使人脸识别模型将其识别为某张脸。攻击过程为：(1)确定扰动区域。将插值方法初始化为掩码方法，使用 STN 将插值掩码置于适当的面的位置，并相应地创建 M 掩码的插值掩码（将 M 的掩码范围对应的元素设为 1，其他元素设为 0）；(2)寻找最佳扰动。一旦确定了误差区域和误差形状，就应用迭代的 MI-GSM 攻击方法，根据给定的优化目标找到最佳误差；(3)最后，通过在 X 面的原始掩膜上添加掩膜变形来创建对抗样本 $x^{'}$：$x^{'} \leftarrow x \otimes (1-M) + \delta \otimes M$，如图 2 所示。

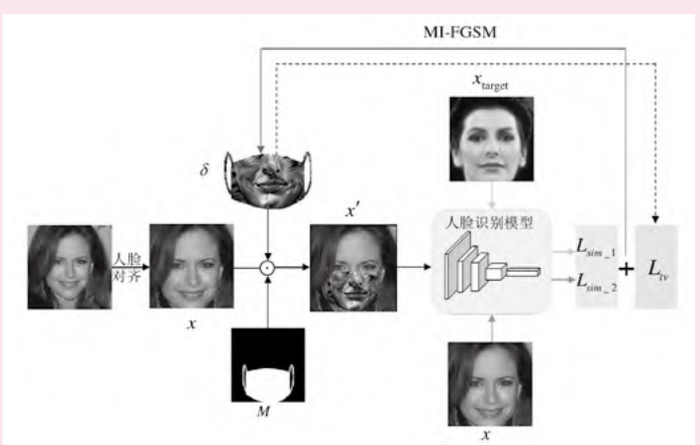


图 2　口罩攻击算法框架图

1. 确定扰动区域及优化目标

确定扰动区域。程序如下:(1) 使用 MTCNN 人脸检测器采集和对齐人脸;(2) 将扰动初始化为口罩的形状; (3) 面罩排列与脸部在 STN 上的正确位置和面罩支架的干扰。从分散注意力者的形状和位置的信息中,创建了一个面具矩阵 M,它与人脸图像的大小相等。

一旦确定了干扰量,就会根据优化目标确定最佳干扰。文献[4]指出,以前的造假方法建立的造假模型注重深度学习模型的最终输出(概率值或标签),但这种类型的方法一般不适用于使用深度学习的人脸识别,因为人脸识别依赖于根据识别字符之间的相似性(或距离)来估计识别结果,而不是预测标签。对于模式来说,情况并非如此。为此,我们开发了一个基于识别属性相似性的损失函数:

$$L_{sim\_1}=\cos(F(x')\ F(x_{target}))=\frac{F(x')\cdot F(x_{target})}{\|F(x')\|_2\times\|F(x_{target})\|_2}\quad 式(3)$$

其中, $\otimes$ 表示 Hadamard 输入, x' 即带有口罩扰动的人脸图像,即攻击样本, x_{target} 为攻击目标, F 是人脸识别模式, $\cos(\cdot)$ 表示余弦相似度,数值在 [-1,1] 范围内,数值越大意味着身份相似度越高。通过最大化这个损失函数,我们可以增加对抗样本 x' 和攻击目标 x_{target} 之间的余弦相似度。

一个有效的攻击方法必须考虑到对抗样本 x' 与攻击目标 x_{target} 之间的相似性,同时提高对抗样本 x' 与攻击模式 x 之间的相似性。为此,引入 L_{sim_2} 损失:

$$L_{sim\_2}=\cos(F(x')\ F(x))=\frac{F(x')\cdot F(x)}{\|F(x')\|_2\times\|F(x)\|_2}\quad 式(4)$$

余弦攻击算法通过最小化损失函数 L_{sim_2} 和减少攻击模式 x' 与原始 x 模式之间的余弦相似度来改进。为了确保生成的口罩扰动的空间均匀性,我们引入了一个额外的总损失(TV)[5]。

$$L_{tv}=\sum_{i,j}\left(\delta_{i,j}-\delta_{i+1,j}\right)^2+\left(\delta_{i,j}-\delta_{i,j+1}\right)^2)^{1/2}\quad 式(5)$$

其中, $\delta_{i,j}$ 是口罩扰动 δ 的像素值。综合式(3)、式(4)和式(5),口罩攻击算法的最终优化目标为:

$$\arg\max L=\arg\max(\lambda_1 L_{sim\_1}-\lambda_2 L_{sim\_2}-\lambda_3 L_{tv})\quad 式(6)$$

其中, λ_1 、λ_2 和 λ_3 为损失函数的权重。

2. 产生口罩扰动

一旦优化问题被定义,扰动就会用 MI-FGSM 生成。MI-FGSM 是一种用于分类模型的负模型生成方法,当与本工作中开发的优化问题相结合时,在攻击人脸识别模型时,可以在没有任何误差大小限制的情况下生成可见的面具错误。在每个迭代中,首先计算干扰掩码 δ 的梯度数据 $\nabla_\delta L$ 以优化 L,然后通过组合梯度数据获得当前梯度,最后利用当前梯度数据持续更新干扰掩码 δ 。程序如下:(1) 在迭代 T 期间,根据式 7 计算损失函数 L 相对于误差函数 δ_t 的梯度 $\nabla_\delta L$ 。(2) 通过使用公式 8 计算不同时期的斜率的指数移动平均值,得到新的梯度 M_{t+1} 。(3) 使用方程 9 中的梯度数据 M_{t+1} 更新误差 δ_t 。(4) 如果没有达到迭代次数,总是重复上述程序。

$$\nabla_{\delta_t}L\leftarrow\lambda_1\frac{\partial L_{sim\_1}}{\partial x_t'}\cdot\frac{\partial x_t'}{\partial\delta_t}-\lambda_2\frac{\partial L_{sim\_2}}{\partial x_t'}\cdot\frac{\partial x_t'}{\partial\delta_t}-\lambda_3\cdot\frac{L_v}{\partial\delta_t}\quad 式(7)$$

$$M_{t+1}\leftarrow\beta\cdot M_t+(1-\beta)\cdot\nabla_{\delta_t}L\quad 式(8)$$

$$\delta_{t+1}\leftarrow\delta_t+\alpha\cdot sign(M_{t+1})\quad 式(9)$$

经过一定次数的迭代,我们得到了最佳扰动 δ 和负样本,其中负样本可以表示为:

$$x'\leftarrow x\otimes(1-M)+\delta\otimes M\quad 式(10)$$

(四)攻击效果分析

图 3 显示了隐形攻击方法产生的一些响应模式和通知掩码,其中第一行显示了攻击的目标 X_{target} ,第二行是原始 X 模式,第三行是生成的对抗样本 x' ,最后一行是生成的扰动的口罩 δ 。我们看到,我们的参考模型对应的是脸部有一定程度隐藏的情况。优化口罩攻击的目标是使攻击者的识别特征 x' 和目标的模式 X_{target} 在特征空间中的距离最小。为了实现这一目标,面具的干扰图案类似于人脸的下部,就像被面具部分隐藏的脸被画在面具本身上一样,这表明下部的面部特征在人脸识别中起着重要作用。此外,不同的攻击对象有不同的口罩检测模式,这表明人格特征因人而异。

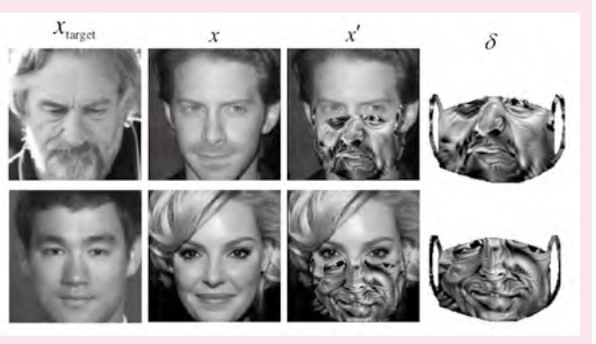


图 3　对抗样本示例

定性分析。为了进一步分析视觉面具反应的结果，从 CASIA WeFace 数据库的前 1000 名识别者中随机选择攻击目标，并生成反应面具。一些结果显示在图 4 中。

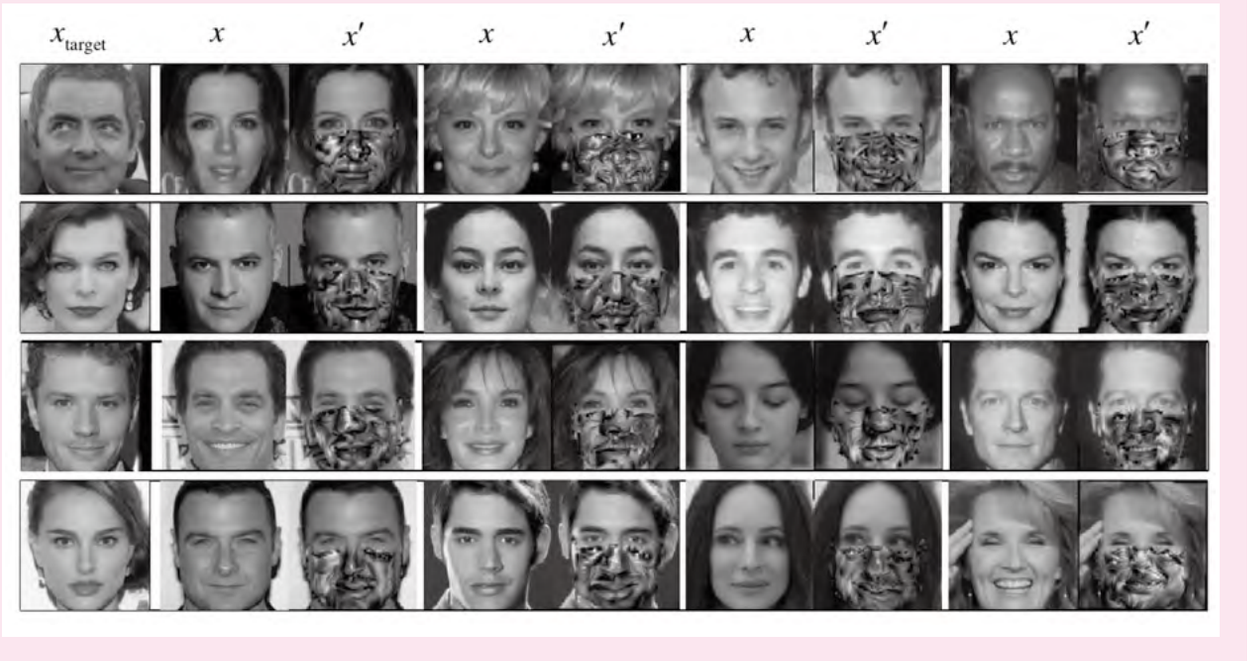


图 4　针对不同攻击目标生成的对抗样本示例

图 4 中的第一列显示了攻击的目标 X_{target}，下面几列显示了左边的原始 X 平面和右边的带干扰掩码的 X 平面。需要注意的是，编码后的面具类似于人的鼻子和嘴巴，根据文献[6]，鼻子和面具的其他部分含有重要的信息，可以提取出来进行识别，嘴巴也含有重要的信息，所以鼻子、嘴巴甚至眼睛都可以纳入构建的面具中。口罩变换算法的目的是提高生成的变换模型与目标脸的识别之间的余弦对应关系，并促进对鼻子和嘴的学习，因为这是目标脸的重要特征，所以生成的面具的鼻子和嘴与目标脸的特征相似。人脸识别网络的目标是识别人脸的面部特征，但面部特征之间的微小差异会使网络无法正确识别它们。本文提出的面具生成和背景检测算法可以成功骗过人脸识别模型。

二、基于图像修复的口罩对抗样本防御方法研究

（一）问题的建立

将应用了面具的脸部原始图像作为 X，并对其进行扰动 δ，以形成另一面的对抗样本 X'。本文提出了一种基于 CGAN 重建的保护方案，其中对抗样本 X' 被重建为原型 X，即 $G_\theta : X' \to X$，θ 生成器参数为 G。在测试阶段，人脸识别模型提取面部特征，计算面部特征之间的相似度和距离，并根据这个相似度和距离来确定身份，例如，如果距离大于阈值，就意味着该脸不是同一个人。鉴于人脸识别模型的这一特性，我们的防御问题是使重建的模型 $G_\theta(X)$ 与原始模型在 X 空间的身份函数之间的余弦距离最小，防御问题表述如下。

$$\theta^* = \arg_\theta \min L_d(F(G_\theta(X')), F(X) = 1 - \cos(F(G_\theta(X')), F(X) \quad \text{式(11)}$$

其中，余弦距离度量是 $L_d(\cdot)$，人脸识别模型是 $F(\cdot)$，F(X) 是由原始样本 X 和重建样本的人脸识别模型得出的 512 维身份（深度函数），$F(G_\theta(X)$ 通过将身份之间的重构样本 $G_\theta(X)$ 减少到 512 维身份，使重建样本 $G_\theta(X)$ 与原始样本 X 的身份相同，并具有更相似的深度函数。可以考虑到空间加权，以进一步提高图像质量。

（二）面向口罩对抗样本的人脸识别防御方法

“DeFense-EC” 生成一个轮廓重建网格和一个基于 CGAN 的人脸重建网格，首先从原始样本中重建“轮廓”，然后是“人脸”。为进一步提升生成图片

的质量，提出空间加权对抗损失。DeFense-EC 基于 CGAN 构建轮廓重构网络和人脸重构网络，以先轮廓后人脸的方式，将对抗样本重构为原始样本。DeFense-EC 的保护程序如下：在第一步中，恢复网络首先重新创建受损区域的轮廓；在第二步，人脸重建网络根据轮廓半径重建受干扰区域的人脸；最后，重建的人脸被确认。如图 5 所示。

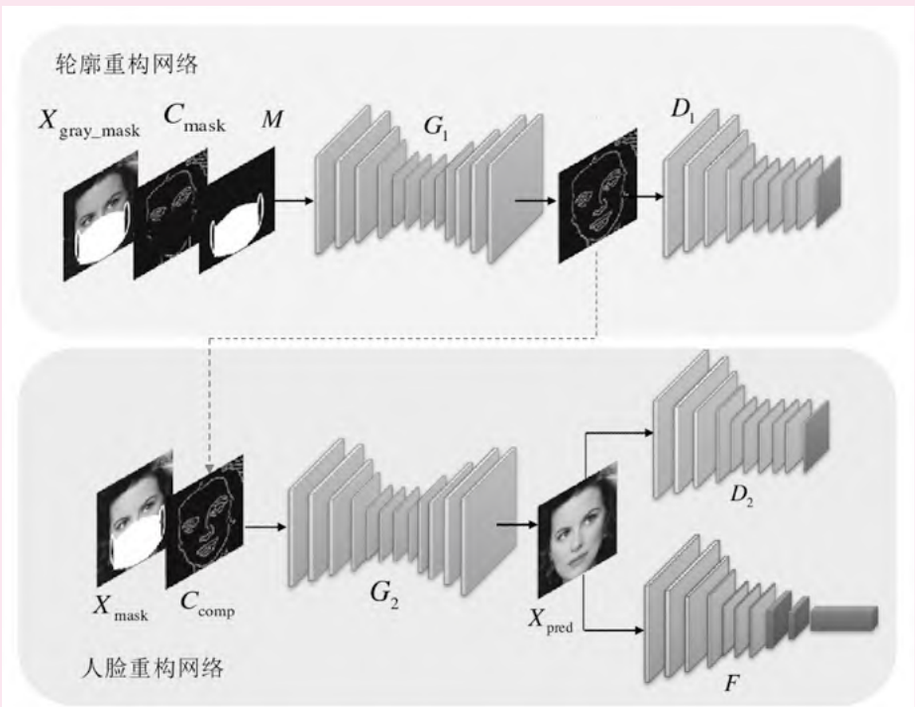


图 5　DeFense-EC 框架

轮廓重构和重构网络由生产器和判别器组成。为简单起见，我们指的是生成器和判别器分别用 G_1 和 D_1 表示以及人脸重构生成器和判别器分别用 G_2 和 D_2 表示。在灰度图 X_{gray} 和轮廓图 C_g 上标记原始脸部 X，X' 表示变形面具 M 标记敌人模型，在灰度图 X_{gray_mask} 和轮廓图 C_{mask} 上标记缺失变形部分的脸部图 X'。本研究中使用的不同原始脸部和它们各自的符号表示方法如图 6。

图 6　各种图像及其对应的符号标记

1.DeFense-EC

为了充分利用去除噪声掩码技术，必须将有噪声掩码的人脸转换为无噪声区域的人脸，并使用图 7 所示的噪声掩码 M 提取无噪声区域的人脸，即无噪声人脸掩码 $X_{mask} \leftarrow X' \otimes (1-M)$。然后通过在轮廓修复网格和表面修复网格之间的切换来恢复受

损区域的表面。

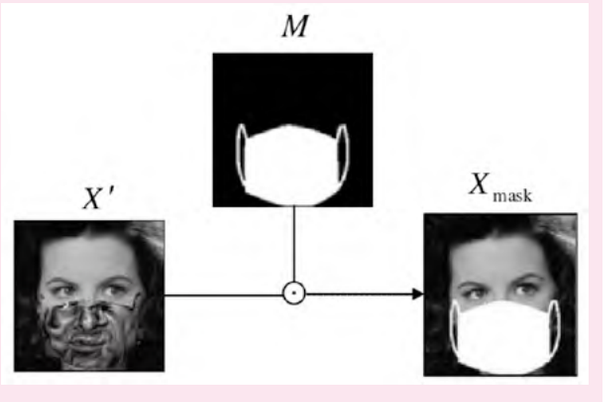


图 7　掩码操作示例

重建轮廓图。为了重建受干扰地区的轮廓，采用了图 8 所示的等高线网格。使用 Canny 轮廓检测器 X_{mask}，可以将轮廓图 C_{mask}、灰度图 X_{gray_mask} 和扰动掩码 M 这三个独立的通道合并为三个数据通道，并送至轮廓重建网络生成器 G_1，以获得完整的面部轮廓 $C_{pred}=G_1(X_{gray\_mask},C_{mask},M)$。分割器 D_1 必须能够尽可能准确地区分实际的原始轮廓图 C_g 和发生器产生的轮廓图 C_{pred}。它可以创建逼真而完整的轮廓图 C_{pred}。为了填补受损区域的轮廓图 C_{pred}，并保留未受损区域的轮廓图 C_g，应将创建的等高线的轮廓线合并成一张复合等高线图 $C_{comp}=C_g\otimes(1-M)+C_{pred}\otimes M$，其中包含原始表面轮廓线 G_1 中受损区域的轮廓线 C_{pred} 和原始表面轮廓线中剩余区域的轮廓线 C_g。

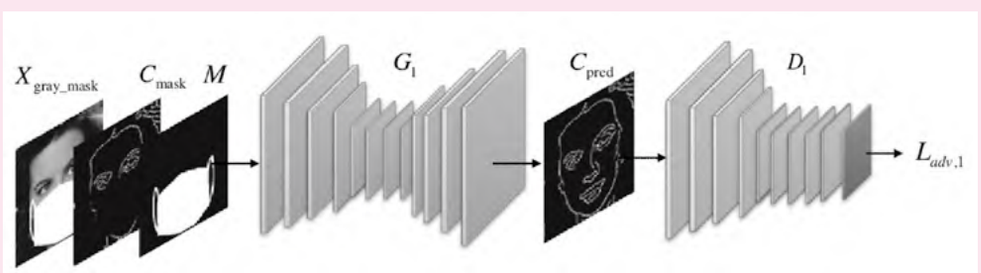


图 8　轮廓重构网络

重构人脸图。在创建脸部轮廓后，脸部重建网格根据轮廓图创建一个脸部，如图 9 所示。三通道人脸图 X_{mask} 和单通道轮廓图 C_{comp} 合并为四个数据通道，并送入生成器 G_2 进行面部重建，创建一个完整的面部 G_2，进行无对抗性扰动，即 $X_{pred}=G_2(X_{mask},C_{comp})$。判别器 D_2 可以更好地将原始 X 平面与生成器 G_2 的 X 平面分开。通过与 D_2 的博弈，G_2 被用来在原图 C_{pred} 旁边创建一个轮廓图 C_{comp}。最后，我们可以添加一个 X 与 X_{pred} 来填充编码区的人脸 $X_{comp}=X\otimes(1-M)+X_{pred}\otimes M$，而不对原始区的人脸进行编码，以创建一个单一的人脸图像 G_2，其中编码区的人脸来自生成器，其余的人脸图像来自原始人脸图像。一旦扰动得到纠正，脸部得到恢复，就可以被认出。

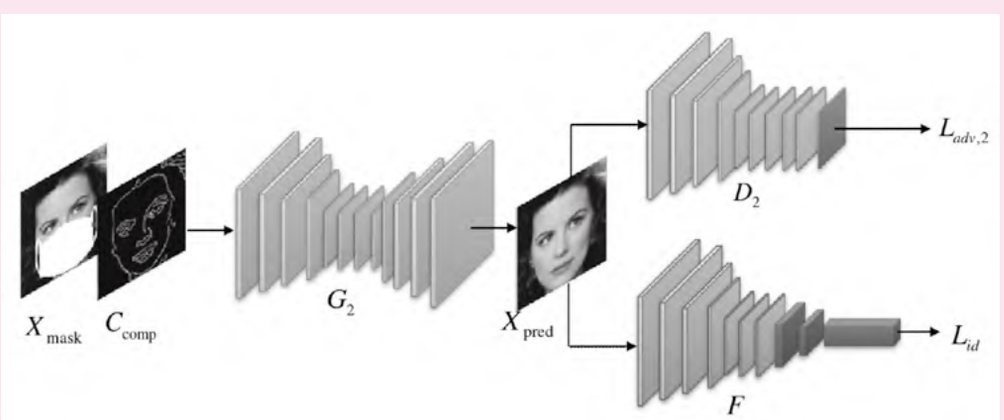


图 9　人脸重构网络

2. 损失函数

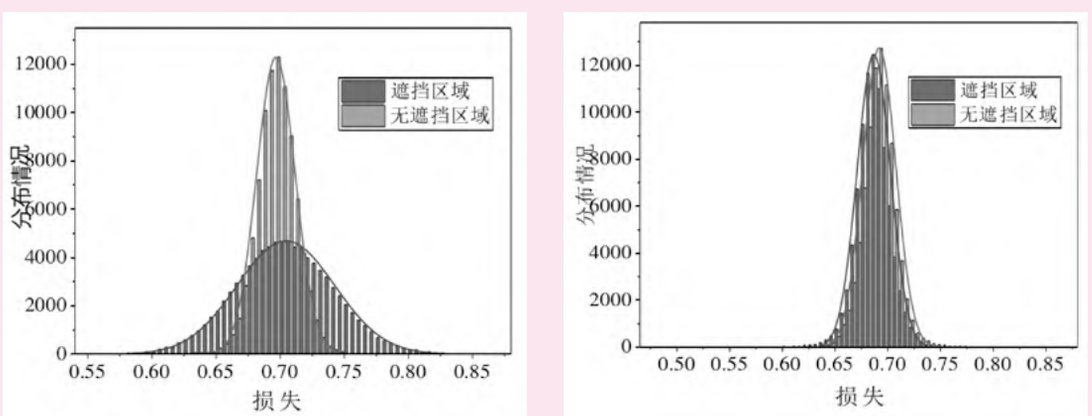
DeFense-EC 开发了几个缺失的轮廓训练功能和人脸重建网络，以更好地学习跟踪的第一个细节。

利用加权对抗损失，口罩扰动是一种噪声。如图 10a 所示，与其他部分相比，对面区块的噪声模式与区块中噪声最大部分的损失分布不一致。因此，原则上，这两个地区并不等同，这两部分的损失应分别处理。因此，建议使用空间加权负损失，其中干扰区域的负损失与其他区域的负损失加权不同，轮廓重建网和面孔重建网的空间加权负损失如下所示：

$$L_{adv,1}=W\cdot E_{(C_g,X_{gray})}[\log D_1(C_g,X_{gray}]+W\cdot E_{X_{gray}}\log[1-D_1(C_{pred},X_{gray}] \quad 式(12)$$

$$L_{adv,2}=W\cdot E_{(X,X_{gray})}[\log D_2(X,C_{comp}]+W\cdot E_{C_{comp}}\log[1-D_2(X_{pred},C_{comp}] \quad 式(13)$$

其中，$L_{adv,1}$ 和 $L_{adv,2}$ 是用于轮廓和脸部重建的空间加权负损失网，W 是对应于掩膜矩阵的空间加权矩阵，其中矩阵元素只有两个值：受干扰区域的相应元素的值为 0.75，其他区域的元素的值为 1。图 10b 显示了空间加权竞争损失函数确定后两个区域的分布情况，可以看出，这两个区域是收敛的，说明该损失函数是有效的。



a. 引入空间加权对抗损失前　　b. 引入空间加权对抗损失后

图 10　扰动区域与其他区域对抗损失函数的分布

为了提高训练周期重建的稳健性，引入了额外的特征适应损失 [7]。

$$L_{FM}=E\left[\sum_{i=1}^{L}\frac{1}{N_i}\left\|D_1^i(X)-D_1^i(X_{pred})\right\|_1\right] \quad 式(14)$$

其中，N_i 是第 i 层激活映射的元素个数，D_1^i 是判别器 D_1 的第 i 层映射。轮廓重构中的生成器与判别器损失是

$$L_{G_1}=L_{adv,1}+L_{FM} \quad 式(15)$$

$$L_{D_1}=-L_{adv,1} \quad 式(16)$$

人脸重建网络的损失包括特征重建的损失 [8]、风格重建的损失 [9] 和人脸重建的损失 [10]，以及空间加权连接的损失 l_1 损失以及 L_d。特征损失提取和风格损失提取是图像处理中常用的损失函数，其中特征损失提取确保图像的感知语义背景（内容和整体空间结构），风格损失提取确保图像的风格背景（颜色、纹理等）。

特征重建损失。传统的图像转换方法在像素级计算损失，即生成的图像与真实图像之间的差异。然而，像素级的损失并不能反映两幅图像之间的感知差异。如果两幅图像只相差一个像素，它们可能在感觉上是相似的，但却有很大的像素损失。因此，损失函数的重建根据其语义属性对差异进行比较。无损方法使用预定义的网格从图像的不同层中提取特征，并使用这个方程来减少特征之间的距离，更可靠地确定图像的相似度。公式如下：

$$L_{feature}=E\left[\sum_i\frac{1}{N_i}\left\|\phi_i(X)-\phi_i(X_{pred})\right\|_1\right] \quad 式(17)$$

其中，N_i 是第 i 层激活映射的元素个数，ϕ_i 是在 Imagenet 数据集上预训练的 VGG-19 网络的第 i 层的激活映射。在这种情况下，VGG-19 只被用来提取图像的语义特征，没有使用其他反向传播方法来更新参数。

风格重建损失。风格重建期间的风格损失超过了图像之间的风格差异。VGG-19 将再次被用来区分不同层次的图片元素，并计算图片元素的网格差异。如果元素是按维度 $C_j \times H_j \times W_j$ 描述的，重建类型的损失函数有以下形式：

$$L_{style} = E_j\left[\left\|G_j^{\phi}(X_{pred}) - G_j^{\phi}(X)\right\|_1\right] \quad 式(18)$$

其中，G_j^{ϕ} 是激活映射 ϕ_j 的 Gram 矩阵。

对抗修复方法与传统修复方法 l_1 结合使用时最为有效，因为它们更有效率[10]。

$$L_{l_1} = E\left[\left\|X - X_{pred}\right\|_1\right] \quad 式(19)$$

在这种情况下，分离器仍然必须检查样品的真实性，生成器不仅要确保生成的样品不会误导分离器，还要减少生成的样品与原始空白样品 l_1 之间的距离。分频器还负责保留高频数据，l_1 损失而回传分频器可以保留低频数据，从而减少所产生的图像模糊。

身份损失是指过滤方法的优化旨在使重建样本和原始样本的身份属性之间的余弦距离 L_d 最小。为了保证重建样本的身份特征与原始样本的身份特征完全一致，这个距离被用作生成器 G_2 损失函数的测量，这个距离测量被称为身份损失函数。

$$L_d = 1 - \cos(F(X)\ F(X_{pred})) = 1 - E\left[\frac{F(X)\cdot F(X_{pred})}{\left\|F(X)\right\|_2 \cdot \left\|F(X_{pred})\right\|_2}\right] \quad 式(20)$$

这里，F 是已经学习过的人脸识别模型 ArcFace，和无损识别一样，它只用来提取面部特征，不做进一步跟踪更新设置。脸部重建发生器和判别器的最终损失函数是。

$$L_{G_2} = \lambda_{adv,2}L_{adv,2} + \lambda_p L_{feature} + \lambda_s L_{style} + \lambda_{l_1}L_{l_1} + \lambda_d L_d \quad 式(21)$$

$$L_{D_2} = -L_{adv,2} \quad 式(22)$$

三、实验

在这一节中，CASIA-WebFace 和 LFW 数据库被用来创建一个初始样本，用于基于这两个数据集的口语实验。本文所述的所有实验都是在 GeForceRTX2080ti 服务器上进行的，使用的是 Pytorch1.4.0、Numpy1.14.3、Opencv3.4.0、Scikit-image0.14.0 和 Scipy1.0.1。DeFense-EC 脸部轮廓和重构网格生成器的结构取自文献[10]，由一个编码器和一个解码器组成，而分割器则采用 PatchGAN[11] 的思路。训练 GANs 一直是一个挑战。为了提高 DeFense CE 训练的鲁棒性并加速模型收敛，事件规范化(IN) 被应用于网络的各个层面。

(一) 去噪效果

在这项研究中，等高线图被用作代理图像，并使用颜色→等高线方法重建脸部，以消除面具区域的噪音。图 11 显示了修复网络的总体布局。

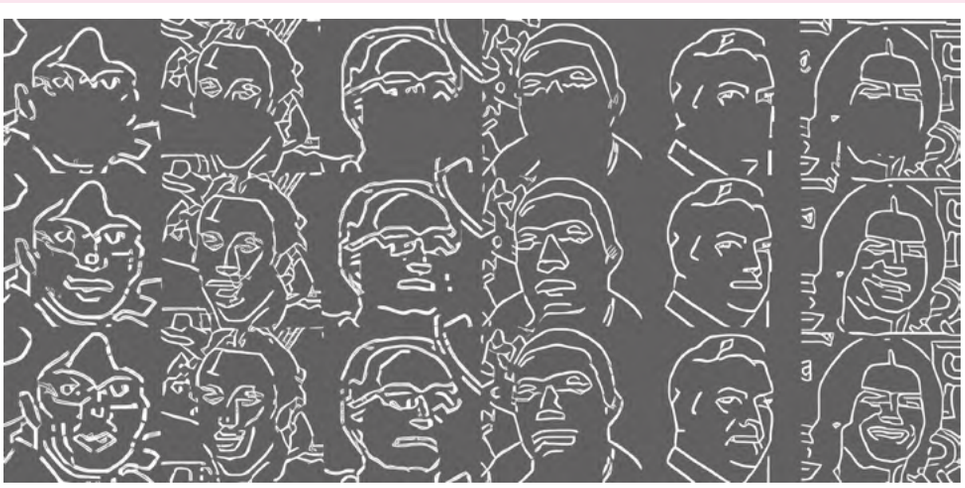


图 11　DeFense-EC 重构的部分人脸轮廓图

在图 11 中，第一条线是要填充的轮廓 C_{mask}，第二条是由轮廓反馈网络生成的轮廓 C_{comp}，第三条是原始脸部的轮廓 C_g。然后，人脸重建网络根据账户信息生成一张人脸图，如图 12 所示。

图 12 依次显示了要填充的人脸、由轮廓网重建的轮廓 C_{comp} 、由面网生成的人脸 X_{comp} 和实际的原始曲面 X。DeFense-EC 表明，使用等高线图作为中间图像能有效地恢复覆盖区的人脸，并保留覆盖区的人脸。

图 12　DeFense-EC 重构的部分人脸图

为了进一步分析 DeFense-EC 的面部重建，对 CASIA-WebFace 和 LFW 数据库进行了面部重建和面具去除测试。结果如图 13 和 14。

图 13　DeFense-EC 在 CASIA-WebFace 上消除扰动生成的部分人脸

图 14　DeFense-EC 在 LFW 上消除扰动生成的部分人脸

在图 13 和图 14 中，第一行显示了由遮蔽算法创建的对抗样本 X'，第二行显示了由 DeFense EC 降噪算法创建的人脸 X_{comp}，最后一行显示了原始 X 面。实验结果表明，DeFense-EC 成功地重建了变形的面，并保留了未变形的面，即使大部分变形在下层面可见。有趣的是，DeFense-EC 不仅在 CASIA WebFace 测试集上表现良好，而且在 LFW 数据集上也表现良好，该数据集的结构与训练集不同，而且仍然产生了逼真、不失真和非常漂亮的面孔，显示了出色的通用性。此外，当测试集和训练集中的身份不匹配时，DeFense-EC 为每张脸生成了不同的鼻子和嘴巴，这表明 DeFense-EC 不仅能在平均水平上预测训练集中的人脸，而且还能预测不同人脸的有意义的外观。有人提出，DeFense-EC 可以预测不同面孔的外观，而不仅仅是平均化它们。

（二）对比实验

“DeFense-EC” 生成了一个用于轮廓重建的网格和一个基于 CGAN 的网格用于人脸重建，轮廓图是人脸重建的间接表示。对于轮廓控制，我们在 $CGAN_1$ [12] 的基础上开发了一个新的模型，它由一个单一的网格组成，直接重建面部而不使用轮廓。在这项研究中，在 CASIA WebFace 数据库中比较了配置 CE 和 DeFense 的控制功效，结果如图 15。

图 15　各个模型消除扰动的效果展示

在图 15 中，AdvMask 是由遮蔽算法创建的不需要的模型，$CGAN_1$ 是使用没有轮廓线的网格直接重建面部的结果，GT 是原始面部。研究发现，$CGAN_1$ 没有轮廓的面部重建不够真实和自然，而且在鼻子和嘴等部位与原脸有很大差别。然而，用 EC 建模的有轮廓的脸，包括鼻子和嘴，看起来更加真实和自然，更接近真实的脸。这表明，从轮廓线上进行面部重建可以使面部表情和情绪更加真实和详细，因为轮廓线包含面部纹理和空间结构等重要信息，而丰富的纹理信息对生成细节很有帮助。如果提议的损失在 EC 中得到实施，由 DeFense-ES 生成的人脸将更加完整，因为该模型将包含所有的身份数据，便于识别。

然后通过使用 PSNR 和 SSIM 对不同数据集的图像进行重建来评估每个模型的质量。表 1 列出了在 CASIA-WebFace 和 LFW 数据集上测试的 CGAN1、EC 和 DeFense-EC 模型以及每个模型的重建图像的 PSNR 和 SSIM 结果。

表 1　各个模型重构图像的 PSNR 和 SSIM 得分

数据	CASIA-WebFace			LFW		
方法	$CGAN_1$	EC	DeFense-EC	$CGAN_1$	EC	DeFense-EC
SSIM	89.79	94.21	95.58	94.41	95.34	97.57
PSNR	26.58	29.45	31.20	28.62	30.90	31.68

表 1 显示，两个数据集的 EC 的 PSRN 和 SSIM 值都很高，这表明使用轮廓梁进行面部重建可以提高图像质量。 此外，DeFense 的 EC 值高于 $CGAN_1$ 和 EC，表明身份特征损失和空间加权也影响了图像质量，与上述定性分析一致。

以前的研究表明，使用轮廓信息可以提高图像质量。为了研究 Canny 参数值 σ 如何影响图像质量，我们为每条轮廓线选择了不同的 Canny 参数值 σ，为每条轮廓线使用了不同的人脸重建网格，并评估了所得图像的质量。

图 16 显示了用不同数值 σ 创建的图像的 PSNR 值。随着数值 σ 的增加，图像质量先好后坏；σ 在 1.5–2.5 的范围内，图像质量变得更好。为了进一步分析这种情况的原因，图 17 显示了各个数值 σ 的轮廓。请注意，σ 用不同的值创建的轮廓图包含不同的结构和纹理信息。如果 σ 数值为 1，等高线图包含更多的纹理和结构信息；如果 σ 数值为 3 或 4，等高线图包含较少的纹理和结构信息。这意味着当 σ 取较小的值时(0 除外)，不仅是降低图像质量的噪声源，而且使用小的非零值限制了模型捕捉图像的能力。σ 在更高的数值下，可用的轮廓信息是有限的，不能完全用于提高图像质量。因此，σ 的值被设定为 2，以获得更多可接受和可用的轮廓信息。

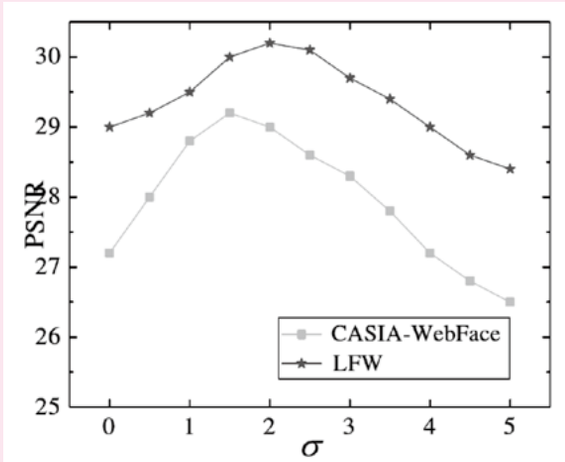


图 16　Canny 检测器参数 σ 取值对生成图片质量的影响

图 17　参数 σ 不同取值下的轮廓图

(三)防御效果

在这一节中，分析了 DeFense–EC 针对 Muzzle 攻击的保护算法：在 CASIA–WebFace 和 LFW 数据库中测试了 CGAN1、EC 和 DeFense–EC 的性能，结果见表 2。

表 2　各个模型的防御成功率

方法	CASIA–WebFace	LFW
$CGAN_1$	89.13%	90.21%
EC	90.78%	91.85%
DeFense–EC	91.24%	92.32%

见表 2 所示，CGAN1、EC 和 DeFense–EC 的性能水平在 CASIA–WebFace 语言算法中为 89.13%、90.78% 和 91.24%，在 LFW32 数据集中为 90.21%、91.85% 和 92%。三种模型保护这两个数据集的能力随着时间的推移而提高，DeFense–EC 在所有情况下都能提供更好的保护，而语言算法的影响则有所下降，这表明 DeFense–EC 是有效的。

为了进一步分析这三个模型的保护特性，对这三个人脸识别模型的人脸识别准确率进行了测试。为了测试 LFW 的差异效应，我们选择了 6000 对同一人的脸和 6000 对同一人的不同脸，结果见表 3 所示。两张相同的图像中，一张是原始脸，另一张是重建的脸。在两张具有不同标识符的图片中，一张是复原的肖像，另一张是攻击的目标。见表 3 所示，CGAN1、EC 和 DeFense–EC 去除干扰掩码后的 ArcFace 检测准确率分别为 95.89%、97.30% 和 98.21%，表明 DeFense–EC 对干扰样本的保护效果

更好。

表 3　各个模型消除扰动下的人脸识别准确率

方法	$CGAN_1$	EC	DeFense-EC
人脸识别准确率(%)	95.89	97.30	98.21

图 18 显示了去除干扰物后每个模型中人脸检测的 ROC 曲线，图 GT 显示了没有负干扰物的人脸的 ROC 曲线。对于 DeFense-EC、EC 和 DeFense-EC 三个模型，曲线下的面积逐渐增大，其中这个面积代表 AUC 值，这个值的增大意味着声压人脸识别效果逐渐提高，DeFense-EC 的声压人脸识别效果更好，DeFense-EC 的保护效果很明显。为了找出 DeFense-EC 受到更好保护的原因，我们通过消除干扰来测试三种模型—CGAN1、EC 和 DeFense-EC 的人脸识别性能。“ArcFace”是用来检索身份属性的。在去除每个模型的噪声后，用 t-SNE 将身份属性还原为两个维度，用它们的方差分布来获得相同身份的人的身份属性的相关程度和不同身份的人的身份属性的分布，如图 19 所示。

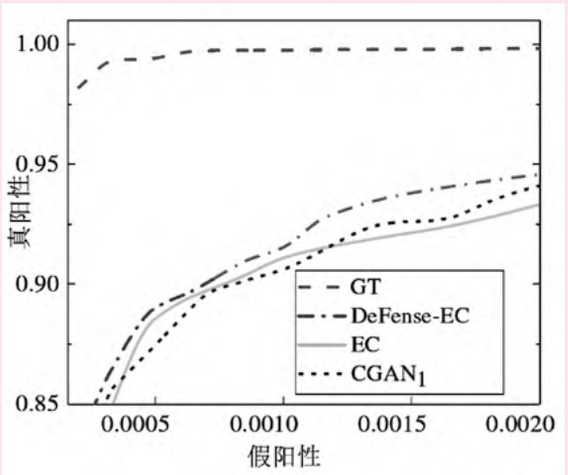


图 18　人脸识别 ROC 曲线

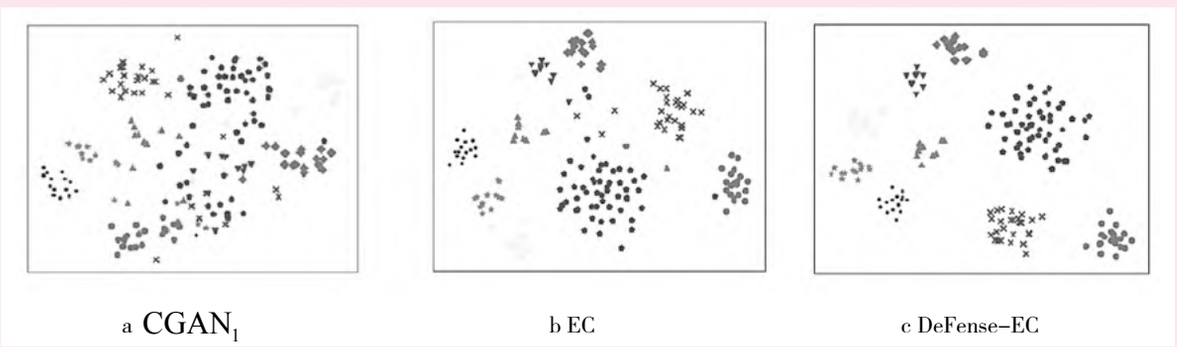


图 19　身份特征可视化

图 19abc 显示了去除干扰后 $CGAN_1$、EC 和 DeFense-EC 的人脸识别性能，当同一点形代表同一人脸时。图 19a 显示了去除 $CGAN_1$ 病变后，脸部在景深中不再可见。在去除 EF 分心效应后，面部图像仍然没有完全解决，尽管混合程度有所提高，如图 19b 所示。此外，DeFense-EC 在降噪后获得了最大的过剩的面部特征，有利于基于特征的面部模式识别聚类，表明 DeFense-EC 可以有效地保护面部识别模式免受口罩对抗样本攻击。

四、总结

本文提出了一个稳健的 DeFense-EC 模型，该模型通过使用第一和第二人脸重建方法在人脸干扰区域重建人脸来消除面具干扰并验证面具碰撞。该模型实现了基于 CE 图像修复模型的人脸识别网络，该模型计算了特征的身份损失，以提高掩蔽性能，并改善所得到的图像的质量，同时损失空间权重。用 CASIA-WebFace 和 LFW 数据集进行的实验表明，DeFense-EC 能以更真实自然的方式从加密区域提取人脸，生成高质量的未加密人脸图像，并有效地保护人脸识别模式免受口罩对抗样本攻击的影响。

对由口罩攻击生成的对抗样本进行了类似的实验，成功率为 92.32%，表明该防御方法对言语攻击是有效的。

第一，重构人脸以消除对抗扰动。使用轮廓图作为中间步骤的两步可见度恢复：在第一步，受影响区域的轮廓被恢复。 在第二步，从轮廓图中还原扰动区域的人脸。

第二，为了提高生成图像的质量，对光圈区域分配不同的损失权重，并对损失函数进行空间加权。实验结果证实，损失函数提供了更真实和自然的面孔。

第三,为了确保重建的人脸中的鉴别性信息不发生变化,在损失函数中加入鉴别性特征的损失,使用经过训练的人脸识别模型提取鉴别性特征,并计算特征之间的距离作为损失函数的量度。实验结果表明,这种损失函数能更好地保留重建的人脸的身份,并提高识别的准确性。

参考文献:

[1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv,2014:1409-1556.

[2] Zamir S W, Arora A, Khan S, et al. Multi-stage progressive image restoration[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:14821-14831.

[3] Zhang W N , Zhu Q , Wang Y , et al. Neural personalized response generation as domain adaptation[J]. World Wide Web, 2019(4):1427-1446.

[4] 杨帆,李阳,苗壮,等.基于特征加权聚合的图像检索目标对抗攻击方法[J].计算机应用研究,2021(12):3760-3764.

[5] 程旭,王莹莹,张年杰,等.基于空间感知的多级损失目标跟踪对抗攻击方法[J].通信学报,2021(11):242-254.

[6] 魏忠诚,冯浩,张新秋,等.基于注意力机制的物理对抗样本检测方法研究[J].计算机应用研究,2021:1-6.

[7] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust[C]. Proceedings of the IEEE conference on computer vision and pattern recognition,2015:2892-2900.

[8] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]. Proceedings of the IEEE conference on computer vision and pattern recognition,2015: 815-823.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman. Deep Face Recognition[C].British Machine Vision Conference, 2015:5-8.

[10] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:4690-4699.

[11] 赵立怡. 基于生成式对抗网络的图像修复算法研究[D].西安:西安理工大学,2018:11-12.

[12] Song L, Cao J, Song L, et al. Geometry-aware face completion and editing[C]. Proceedings of the AAAI Conference on Artificial Intelligence,2019(1):2506-2513.

[责任编辑 :胡大威]

Research of the Security of Face Recognition System based on Convolutional Neural Network

CHEN Kengqiang[1], MO Yaohua[2]

(1. Network and Training Center, Shanwei Polytechnic, Shanwei, Guangdong, 516600, China; 2. School of Computer Science and Network Engineering, Guangzhou University, Guangzhou, Guangdong, 510006, China)

Abstract: In response to the problem of adversarial samples in traditional neural networks, using face recognition as an example,the potential security risks of face recognition techniquesbased on traditional neural networksis discussed as well as methods to generate and provide defense against adversarial samples in masks. The main research of this paper is as follows: first, the proposed approach to solve the security problem of face recognition technology generates images with fuzzy and unbiased masks, and the face recognition model misidentifies the face with fuzzy mask as the target person. The experimental results show that the adversarial sample mimics the situation of the concealed person wearing a mask to some extent, with a success rate of 90.43%. Second, the DeFense-EC protection method is proposedto mask the false patterns and use image reconstruction techniques to recover the false patterns and suppress the noise. Tests on different datasets confirm the high image quality of the DeFense-EC reconstruction with a reliability of 92.32%. This paper discusses the problem of unfavorable selection in traditional neural network-based face recognition methods and investigates the risk of unfavorable mask selection and the effectiveness of the DeFense-EC protection method. The increasing popularity of convolutional networks has made the problem of adverse selection an important area of research, and a thorough investigation of adverse selection will facilitate the development and use of convolutional networks.

Key words: convolutional neural networks; adversarial samples; face recognition; maskattack; image restoration