计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(8):222−230 [doi: 10.15888/j.cnki.csa.009562]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于共性梯度的人脸识别通用对抗攻击①

段　伟[1], 高陈强[1,2], 李鹏程[1,2], 朱常杰[1]

[1](重庆邮电大学 通信与信息工程学院, 重庆 400065)
[2](信号与信息处理重庆市重点实验室, 重庆 400065)
通信作者: 段　伟, E-mail: duanwei99@foxmail.com

摘　要: 人脸识别技术的恶意运用可能会导致个人信息泄露, 对个人隐私安全构成巨大威胁, 通过通用对抗攻击保护人脸隐私具有重要的研究意义. 然而, 现有的通用对抗攻击算法多数专注于图像分类任务, 应用于人脸识别模型时, 常面临攻击成功率低和生成扰动明显等问题. 为解决这一挑战, 研究提出了一种基于共性梯度的人脸识别通用对抗攻击方法. 该方法通过多张人脸图像的对抗扰动的共性梯度优化通用对抗扰动, 并利用主导型特征损失提升扰动的攻击能力, 结合多阶段训练策略, 实现了攻击效果与视觉质量的均衡. 在公开数据集上的实验证明, 该方法在人脸识别模型上的攻击性能优于 Cos-UAP、SGA 等方法, 并且生成的对抗样本具有更好的视觉效果, 表明了所提方法的有效性.

关键词: 人脸识别; 对抗样本; 通用对抗攻击; 共性梯度; 个人隐私安全

引用格式: 段伟,高陈强,李鹏程,朱常杰.基于共性梯度的人脸识别通用对抗攻击.计算机系统应用,2024,33(8):222–230. http://www.c-s-a.org.cn/1003-3254/9562.html

Universal Adversarial Attack for Face Recognition Based on Commonality Gradient

DUAN Wei[1], GAO Chen-Qiang[1,2], LI Peng-Cheng[1,2], ZHU Chang-Jie[1]

[1](School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)
[2](Chongqing Key Laboratory of Signal and Information Processing, Chongqing 400065, China)

Abstract: The malicious use of facial recognition technology may lead to personal information leakage, posing a significant threat to individual privacy security. Safeguarding facial privacy through universal adversarial attacks holds crucial research significance. However, existing universal adversarial attack algorithms primarily focus on image classification tasks. When applied to facial recognition models, they often encounter challenges such as low attack success rates and noticeable perturbation generation. To address these challenges, this study proposes a universal adversarial attack method for face recognition based on commonality gradients. This method optimizes universal adversarial perturbation through the common gradient of the adversarial perturbations of multiple face images and uses dominant feature loss to improve the attack capability of the perturbation. Combined with the multi-stage training strategy, it achieves a balance between attack effect and visual quality. Experiments on public datasets prove that the method outperforms methods such as Cos-UAP and SGA in the attack performance on facial recognition models, and the generated adversarial samples have better visual effects, indicating the effectiveness of the proposed method.

Key words: face recognition; adversarial example; universal adversarial attack; commonality gradient; personal privacy security

① 基金项目: 国家自然科学基金 (62176035, 62201111); 重庆市教委科学技术研究计划 (KJZD-K202100606)
收稿时间: 2024-01-25; 修改时间: 2024-02-26; 采用时间: 2024-03-04; csa 在线出版时间: 2024-05-31
CNKI 网络首发时间: 2024-06-04

随着深度学习的快速发展, 人脸识别技术在登录、门禁、支付等领域得到了广泛应用. 然而, 其应用不仅局限于此, 还涵盖了用户行为分析和跟踪, 涉及社交媒体档案和个人简历等信息. 为了应对人脸识别模型在应用中可能面临的隐私安全问题, 当前主流方法主要通过在人脸图像中引入微小范围的扰动, 制作对抗样本, 从而使得人脸识别模型难以准确识别, 这些工作为解决人脸识别技术可能带来的隐私风险提供了一种有效途径[1–3].

对抗样本的核心思想是通过对输入数据进行微小扰动引导深度神经网络产生错误的结果. 当前在人脸识别领域, 生成对抗样本通常需要为每个图像迭代生成针对性的对抗扰动[4]. 然而, 在实际应用中处理大量图像时, 逐一生成对抗扰动成本极高. 因此, 研究适用于人脸识别任务的通用对抗扰动生成方法显得尤为重要. 通用对抗扰动, 即图像不可知对抗扰动, 是一种固定扰动, 不依赖于具体图像, 可以直接叠加到数据集中的图像上, 形成对抗样本, 并以较高的概率欺骗深度神经网络. 当前通用对抗扰动在图像分类[5,6]、目标检测[7] 、语音分类[8]等领域取得了显著进展. 然而, 目前生成通用对抗扰动的方法大多具有特定的应用场景, 如果直接将这些方法迁移至人脸识别领域攻击效果较差.

除此以外, 当前人脸识别的对抗样本还存在视觉质量不佳的问题. 目前人脸识别领域中的对抗样本主要通过在脸部图像上添加精心制作的装饰 (对抗扰动) 来进行攻击[9–11], 然而这些特定扰动的像素值范围并未受到限制. 添加这样的扰动后形成的对抗样本与原始图像差异较大, 导致视觉质量下降, 从而限制了在社交媒体上的实际应用.

在此背景下, 提出了一种利用共性梯度引导的人脸识别通用对抗扰动生成方法. 本文主要的贡献如下.

a) 提出一种利用多张人脸图像进行特定目标攻击, 获取人脸识别对抗扰动共性的方法, 其可以有效缓解通用对抗扰动在训练中出现的灾难性遗忘问题, 提升了扰动的跨图像迁移能力.

b) 利用主导性特征损失将特定身份信息嵌入至通用对抗扰动中, 有效提升扰动攻击成功率.

c) 利用多阶段训练策略, 均衡攻击成功率和视觉质量, 在多个人脸识别模型和数据集上达到人脸识别领域目前最优的攻击成功率和扰动视觉质量.

1 相关工作

1.1 对抗样本

对抗样本最早由 Szegedy 等人[12]提出, 目前基于梯度迭代生成对抗样本是应用最广泛的方法. Goodfellow 等人[13]提出的快速梯度符号法 (fast gradient sign method, FGSM) 是首次利用梯度迭代生成对抗样本, 其仅用一次迭代就能得到对抗样本, 该方法具有生成简单、迁移性好的特点. I-FGSM (iterative fast gradient sign method)[14]在 FGSM 的基础上, 利用梯度方向多次迭代, 解决了其攻击成功率较低的问题. I-FGSM 得到的扰动更小更精确, 然而其迁移性相比于 FGSM 会较差. MI-FGSM (momentum iterative fast gradient sign method)[15]主要是引入动量迭代的方法, 可以有效地避免对抗样本在迭代过程中陷入局部最优解, 具有更好的攻击性和迁移性.

在梯度迭代的基础上, 深度愚弄 (DeepFool)[16]基于超平面分类, 通过搜索原分类面与其他面之间的最小代价达到攻击效果. 该方法可以显著提高对抗样本的鲁棒性. Carlini 等人[17]提出一种将限制扰动大小与优化目标作为整体损失函数的攻击方法, 其通过调节参数增强对抗样本的迁移性, 但是迭代次数较多, 需消耗更多资源. 除此以外, 还有 ILCM (iteratice least-likely class method)、UPSET 和 ANGRI 等[18]对抗样本生成算法.

1.2 通用对抗扰动

近年来, 通用对抗扰动的研究备受关注, 众多学者从不同角度解释了通用对抗扰动的作用机理. Zhang 等人[19]证明通用对抗扰动相对于普通的对抗扰动 (依赖图像生成的对抗扰动) 具有独立的语义上的特征. Zhang 等人[20]认为不易被察觉的通用对抗扰动能够显著影响图像的分类的原因是深度神经网络对于高频信息敏感.

Moosavi-Dezfooli 等人[21]提出的通用对抗扰动 (universal adversarial perturbations, UAP) 的生成算法是通过对依赖图像的普通对抗扰动累加得到通用对抗样本. Kamath 等人[22]提出一种基于奇异向量计算通用对抗扰动的方法. Zhang 等人[23]采用自监督余弦相似度损失优化通用对抗扰动. Ud Din 等人[24]提出一种基于在变换域中计算扰动的通用对抗扰动生成方法. Liu 等人[25]提出随机梯度聚集, 该方法使用小批量训练进行预搜索, 将结果作为扰动的梯度更新, 可以增强梯度稳定性以及减少量化误差.

1.3 人脸识别领域中的对抗样本

目前人脸识别领域中的对抗攻击主要分为物理世界攻击和数字世界攻击. 物理攻击大多专注于攻击实际应用的人脸识别系统, 比如门禁、登录系统等, 通常不限制扰动大小, 常以面部配饰为手段. Komkov 等人[9]提出通过帽子制作对抗样本的方法. Yin 等人[10]提出一种将对抗扰动隐藏在妆容中的攻击算法. Zolfi 等人[11]提出一种可以在真实世界运用的人脸识别通用对抗面具, 尽管在佩戴口罩的人脸范围内不受扰动大小限制, 但在有监视的人脸识别环境下难以发挥作用. 在数字世界的人脸识别对抗攻击中, 注重对抗样本与实际图像保持高度相似性. Zhong 等人[26]提出一种基于随机失活的黑盒攻击算法. Jia 等人[27]通过扰动高级语义攻击人脸识别模型.

上述方法都没有从限制扰动的角度生成人脸识别通用对抗扰动. 本文提出的基于共性梯度的通用对抗扰动生成方法, 在保持视觉质量的前提下有效地保护人脸身份信息. 实验结果表明, 该方法在攻击成功率和扰动视觉质量上具有明显优势.

2 本文方法

本文所采用的方法框图如图 1 所示. 首先, 将扰动叠加至多张图像, 同时向身份 S 进行特定目标攻击, 获取梯度后融合得到共性梯度. 在此过程中, 利用相似度量模块, 按攻击程度对每个梯度的权重进行分配, 使得扰动中的特征主导人脸识别的判定过程. 最后, 采用多阶段训练策略更新权重ω, 以平衡攻击损失与均方误差损失之间的比例. 接下来将详细介绍具体的方法.

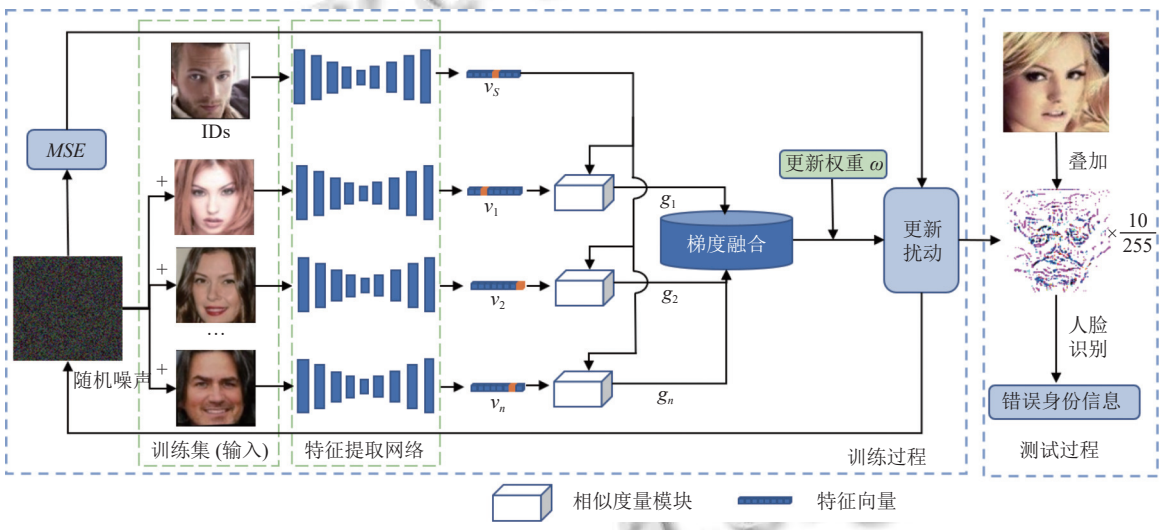

图 1 总体方法框图

2.1 对抗扰动共性计算

对人脸图像进行特征提取可表示为 $f(x): x \sim X$, 特征向量与人脸数据库中的身份映射可表示为 $D(f(x)) \to \varphi_y$, 通用对抗扰动 δ 需满足对于大多数 $x \sim X: D(f(x)) \neq D(f(x+\delta))$. 优化过程可表示为:

$$\Delta\delta = \arg\min_{r} \|r\|_2 \text{ s.t. } D(f(x_k+\delta+r)) \neq D(f(x_k)) \quad (1)$$

$$\delta_{k+1} = \delta_k + \Delta\delta \quad (2)$$

其中, r表示攻击第 k 张图像时在通用对抗扰动上叠加的分量, δ_k表示第 k 次迭代的通用对抗扰动. 而在训练过程中, 如果逐一使用训练集的图像优化通用对抗扰动, 可能出现灾难性遗忘现象[28]. 例如, 在攻击A图像时, 扰动在某一点p的像素值需减 1; 然而, 在随后攻击B图像时, 扰动在相同点p的像素值需加 1, 导致新的扰动对图像的攻击性降低. 为解决此问题, 本文提出了一种方法, 利用多张图像之间对抗扰动梯度的共性, 更有效地提取不同图像扰动之间的共同特征. 具体而言, 本文选取了 n 张图像, 并同时对它们进行向身份 S 的特定目标攻击, 通过获取这 n 张图像的共有梯度优化通用对抗扰动, 从而实现共性计算.

如图 2 所示, $A_1, A_2, \cdots, A_n$表示n张图像对应的对抗样本, $\varphi_1, \varphi_2, \cdots, \varphi_n$表示原始图像身份决策域, φ_S 表示特定身份S 的决策域. 在迭代过程中, 对抗扰动的演变由两个主要部分推动: 一是抵制对抗样本向原始决策域靠近的梯度g_{org} (黄线引导部分), 二是扰动中的特征引导着朝特定身份S 靠近的梯度g_S (蓝线引导部分).

共性计算攻击过程可表示为:

$$\delta_{k+1} = \delta_k + \alpha \cdot sign(g) \tag{3}$$

$$g = \nabla_\delta \frac{1}{N} \sum_{i=1}^{N} J(x_i, \delta, \varphi_S, f_\theta) \tag{4}$$

$$\nabla_\delta J(x_i, \delta, \varphi_S, f_\theta) = g_S + g_{\text{org}} \tag{5}$$

其中, $\|\delta\|_\infty < \varepsilon$, $J(x_i, \delta, \varphi_S, f_\theta)$表示$x_i$进行特定目标攻击的损失函数, N为参与训练人脸图像数量, $sign()$为符号函数. ε为限制扰动大小的超参, $\|\delta\|_\infty$表示对δ求L_∞范数, 即δ中最大的像素绝对值不超过ε. 在不同身份对应的对抗样本中, 它们的原始决策域之间存在较大差异, 即在远离原身份的过程中, 梯度g_{org}差异显著, 多个梯度融合后g_{org}均值趋于零而被抑制. 然而共同向特定身份 S 靠近的梯度g_S具有强相关性, 梯度融合后得以保留.

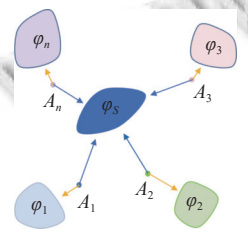

图 2 共性计算示意图

相比与现有方法将每一张图像的攻击梯度用于直接优化通用对抗扰动, 共性计算可以避免g_{org}梯度导致的扰动迭代过程中的震荡, 并且使优化扰动的梯度主要来源是特定目标攻击的增益, 减轻了原始图像对梯度的影响.

2.2 主导性特征损失

分类网络将特征向量输入分类头, 求最大值索引判断类别, 而人脸识别网络根据特征向量与人脸数据库中特征向量的相似度或者距离获取身份信息. 相对于分类任务, 人脸识别中的对抗样本需要改变特征向量多个维度的数值, 影响向量之间的相似度才能达到攻击效果[3]. 因此, 在相同大小的扰动限制下, 为分类任务设计的通用对抗扰动生成方法对人脸识别模型的攻击成功率较低. 为解决这一问题, 本文使用对抗样本经过特征提取网络输出的特征向量与身份S的特征向量之间的相似性反馈优化过程.

在相似性的约束下, 训练时扰动会不断地嵌入特定身份S的特征信息. 然而, 除了嵌入扰动内的特征外, 原始图像中的特征也对人脸识别身份判定起着重要影响. 为了成功误导人脸识别模型, 本文设计了一个相似度量模块Sim, 以确保扰动中的特征主导了人脸识别的判定过程, 即成为主导性特征. 结合共性计算攻击过程梯度求解可表示为:

$$g = \nabla_\delta \frac{1}{N} \sum_{i=1}^{N} Sim(f(S), f(x_i, \delta)) \tag{6}$$

$$Sim(a,b) = \begin{cases} 2\theta - \cos(a,b), & \text{if } \cos(a,b) \leqslant \theta \\ \max\left(0, \dfrac{2\theta - \cos(a,b)}{\theta}\right), & \text{else} \end{cases} \tag{7}$$

其中, $x_i \sim X, \|\delta\|_\infty < \varepsilon$, f为人脸识别特征提取网络, $\cos()$表示求余弦相似度, $f(S)$表示身份 S 的特征向量. Sim的核心在于增加与 S 相似度较低的图像梯度权重, 同时减少相似度已超过身份判定阈值的图像梯度权重. θ表示人脸识别模型将特征向量判定为同一身份的阈值.

由于扰动迭代过程会导致对抗样本与S的相似度增加, 会在视觉上不断显著表示S的特征, 从而影响对抗样本的视觉质量. 如果训练结束后裁剪超出限制部分的扰动, 对抗样本的攻击效果将降低, 所以本文结合均方误差 (mean square error, MSE) 限制扰动的迭代大小, 整体损失函数为:

$$L = \omega \sum_{i=1}^{N} Sim(f(S), f(x_i, \delta)) + MSE(\delta, 0) \tag{8}$$

其中, 通过多阶段训练策略得到的ω是控制攻击效果与视觉质量的比例权重.

2.3 多阶段攻击策略

为有效平衡通用对抗扰动的攻击成功率与视觉质量, 本文通过预热重启和余弦退火的策略对ω进行更新, 主要分为 3 个阶段.

第 1 阶段: 对抗扰动初始化. 该阶段的迭代目标是在增强攻击性的同时减少与特征表达无关的扰动. 因此, 本文在第 1 阶段采用预热重启的策略对ω进行更新, 即逐渐增加ω, 使得对抗扰动逐渐从关注减少扰动过渡到关注提升攻击性. ω更新过程可表示为:

$$\omega = \omega_{\min} + \frac{(\omega_{\max} - \omega_{\min})}{T_{\text{warm}}} \times t \tag{9}$$

其中, $\omega_{\max}$、$\omega_{\min}$对应着ω的最大权值与最小权值. T_{warm}为预热阶段的迭代次数, t为当前迭代次数.

第 2 阶段: 增强对抗扰动攻击. 在扰动控制在较小范围时, 保持ω的取值为$\omega_{\max}$不变, 从而持续增强扰动的攻击性.

第 3 阶段: 优化通用对抗扰动. 第 2 阶段生成的扰

动具有强大的攻击性, 同时包含丰富的 S 身份纹理信息. 为了缓解扰动中的纹理信息对对抗样本视觉质量的影响, 采用余弦退火策略, 即模拟余弦周期更新ω, 以平衡扰动大小与攻击性能之间的关系. ω更新过程可表示为:

$$\omega = \omega_{\min} + \frac{1}{2}(\omega_{\max} - \omega_{\min})\left(1 + \cos\left(\pi \frac{t_{\text{cur}}}{T_{\cos}}\right)\right) \tag{10}$$

其中, t_{cur}表示余弦退火阶段已经执行的迭代数, $T_{\cos}$表示余弦退火阶段总的迭代数. 第 3 段阶段结束将裁剪扰动至$(-\varepsilon, +\varepsilon)$. 本文基于共性梯度的人脸识别通用对抗攻击方法的完整描述如算法 1.

算法 1. 基于共性梯度的通用对抗攻击方法

输入: 人脸识别模型f, N 张人脸图像集X, S 身份的图像 s, 迭代次数 T, 各阶段结束迭代的轮次T_{warm}, T_{aug}, $T_{\cos}$, 扰动距离ε, 攻击步长α.
输出: 通用对抗扰动δ.

1. $\delta \leftarrow$ randn$(-\varepsilon, +\varepsilon)$
2. for $t \leftarrow 1$ to T do
3. $\quad g_{\text{attack}} \leftarrow 0$
4. $\quad$ for $i \leftarrow 1$ to N do
5. $\quad\quad$ select $x_i \in X$
6. $\quad\quad g_i \leftarrow \nabla_\delta Sim(f(s), f(x_i, \delta))$ //利用Sim模块分配梯度权重
7. $\quad\quad g_{\text{attack}} \leftarrow g_{\text{attack}} + (g_i / N)$
8. $\quad$ end for
9. $\quad g_{mse} \leftarrow MSE(\delta, 0)$
10. $\quad$ if $t < T_{\text{warm}}$ do
11. $\quad\quad$ 根据式 (9) 更新ω
12. $\quad$ else if $t \geqslant T_{\text{warm}}$ and $t < T_{\text{aug}}$ do
13. $\quad\quad \omega \leftarrow \omega_{\max}$
14. $\quad$ else if $t \geqslant T_{\text{aug}}$ and $t < T_{\cos}$ do
15. $\quad\quad$ 根据式 (10) 更新ω
16. $\quad g_{\text{total}} \leftarrow g_{\text{attack}} + g_{mse}$
17. $\quad \delta \leftarrow \delta + \alpha \times sign(g_{\text{total}})$
18. end for
19. $\delta \leftarrow Clip_\varepsilon(\delta)$ //裁剪到有效区间
20. return δ

3 实验及结果分析

3.1 实验基础

为了全面验证所提方法的泛化性, 本文进行了一系列实验, 涵盖多个人脸识别数据集, 其中包括 LFW 数据集、AgeDB 数据集以及 CFP-FP 数据集.

LFW 数据集包含 5 749 个身份 ID、13 233 张人脸图像. 该数据集中的人脸图像是生活中的自然场景, 包含不同的姿势、光照、表情、年龄等, 有的人脸存在部分遮挡的情况, 识别难度较大. 因此在 LFW 数据集上的测试精度成为非受限情况下的人脸识别算法性能的重要体现.

AgeDB 数据集包含 3 万张图像, 其中包含亚洲的人脸图像 15 000 张, 欧美的人脸图像 15 000 张, 在同一个身份中的最低和最高年龄分别是 3 岁和 101 岁. 该数据集上的测评指标可以有效地衡量人脸识别模型在不同的种族以及不同的年龄段上面的性能.

CFP-FP 数据集中每一个身份 ID 包含 10 张正面图像以及 4 张侧脸图像, 其数据集上的测试指标可以有效地衡量人脸识别模型在不同角度获取人脸信息的性能.

实验过程中, 选用人脸识别领域内具有代表性的 ArcFace 和 SphereFace 为目标模型. 在 ArcFace 模型中, 采用了两种不同的主干网络, 分别是 ResNet50-IR 和 MobileFace, 以验证算法在不同结构下的鲁棒性. 利用攻击成功率 ASR 衡量生成的对抗扰动的攻击性能, 其计算公式为:

$$ASR = \frac{n_{\text{error}}}{n_{\text{total}}} \tag{11}$$

其中, n_{error}表示叠加通用扰动后识别错误的人脸图像数量, n_{total}表示总的测试样本数量. 利用均方误差与结构相似性 (structural similarity index, *SSIM*) 客观衡量对抗样本的视觉质量, 具体的计算公式为:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[x(i,j) - y(i,j)]^2 \tag{12}$$

$$SSIM = \frac{(2u_x u_y + C_1)(2\sigma_{xy} + C_2)}{(u_x^2 + u_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{13}$$

其中, $x(i,j)$和$y(i,j)$分别表示原始图像与对抗样本的第i行, 第j列像素值. u_x和u_y表示图像均值, σ_x^2和σ_y^2表示图像方差, σ_{xy}是两幅图像的协方差, C_1和C_2为常数, 用于避免分母为零.

为了全面评估生成算法在攻击成功率和扰动视觉质量等方面的性能, 本文选择了 SVD-UAP[22]、Cos-UAP[23]、SGA[25]作为基准方法. 这些基准是在分类任务上提出的通用扰动生成方法, 其中 SGA 是目前分类任务中攻击性能最好的方法.

本文实验中为与基准方法保持一致性, 限制扰动参数ε设置为 10/255. 一轮训练中人脸图像数量N为 24, 迭代总次数 T 取 6 000, 其中三阶段各占 2 000 次迭代, $\omega_{\min}$为 1, $\omega_{\max}$为 36, 同一身份相似度判定阈值θ为 0.42 (ArcFace)[29]、0.38 (SphereFace)[30], 攻击步长α为 0.01. 在攻击人脸识别的损失函数中, 调整参数

ω对于控制主导性特征嵌入程度和扰动大小至关重要. 当ω过大时, 生成扰动的视觉质量明显下降; 反之, 当ω过小时, 扰动的攻击成功率降低. 其中不同ω与 N 对实验结果的影响将在第 3.3 节详细阐述. 实验代码采用 PyTorch 框架编写, 硬件平台采用 4 块 GeForce RTX 3090.

3.2 对比实验

本文对生成的通用扰动攻击成功率进行了评估, 涵盖多个人脸识别模型以及不同的人脸数据集. 详细的实验结果请参见表 1. 可以看出, 本文方法使用 3 个人脸识别模型进行 9 次实验取得了 83.43% 的平均攻击成功率, 相比于 SGA 方法提升了 7.12%. 证明本文方法在人脸识别领域中相比于现有的通用对抗攻击方法在攻击成功率上具有明显优势. 除了攻击成功率之外, 有关数据指标的具体数据详见表 2. 本文方法在多个方面取得了显著的成果, 其中包括平均结构相似度得分达到 0.91 和最小均方误差结果为 26.54. 本文方法在损失函数中引入了对扰动大小的限制, 结合多阶段训练策略有效地避免了对非关键区域的过度扰动. 相较于将每张图像的攻击梯度直接应用于优化扰动的方法, 本文方法生成的扰动在视觉质量上具有更优越的表现. 此外, 通过同时获取多张图像的对抗样本共性再进行扰动优化, 本文方法在训练时间上取得了与现有方法相当或更优的性能.

表 1 通用对抗扰动的攻击成功率 (%)

方法	ArcFace (ResNet50-IR)			ArcFace (MobileFace)			SphereFace		
	LFW	AgeDB	CFP-FD	LFW	AgeDB	CFP-FD	LFW	AgeDB	CFP-FD
SVD-UAP	67.41	56.70	58.80	60.71	52.47	47.27	73.63	71.05	65.36
Cos-UAP	76.49	62.35	63.81	70.68	58.55	53.57	79.79	76.11	69.96
SGA	86.32	76.81	78.45	80.96	65.71	64.26	83.49	81.24	69.55
$\text{Ours}_{\text{wo-dc}}$	84.29	78.83	80.68	79.09	67.67	63.32	78.95	80.93	71.27
$\text{Ours}_{\text{wo-sc}}$	74.92	67.50	60.77	68.65	51.50	45.32	80.16	78.33	66.98
Ours	**91.58**	**86.31**	**83.14**	**86.07**	**75.08**	**73.70**	**94.46**	**86.43**	**74.16**

表 2 各项指标对比

方法	训练时间	平均MSE	平均结构相似度
SVD-UAP	12 h 33 min	41.18	0.89
Cos-UAP	1 h 25 min	87.15	0.78
SGA	2 h 11 min	69.54	0.84
本文方法	1 h 32 min	26.54	0.91

3.3 消融实验

1) 验证共性计算算法与主导性特征损失组合的有效性. 选择 UAP 方法中的利用 DeepFool 攻击对每一张图像搜索最小扰动, 并依次叠加至通用扰动的方法替换共性计算算法, 表示为$\text{Ours}_{\text{wo-sc}}$. 选择 Cos-UAP 中的损失替换主导性特征损失, 表示为$\text{Ours}_{\text{wo-dc}}$, 实验结果见表 1. 当替换掉共性计算算法后, 平均攻击成功率下降至 66.01%, 由此可见, 通过融合多个攻击梯度进行迭代出的扰动因受原始图像影响较小, 具有更佳的跨图像性能. 当替换主导性特征损失后, 平均攻击成功率下降至 76.11%, 证明主导性特征损失模块对攻击成功率有重要的提升作用.

2) 多阶段训练策略的优势. 为了验证多阶段策略训练的有效性, 本文在训练时选取固定的ω值, 观察其与多阶段策略训练得到的扰动在攻击成功率上的差异. 以 ArcFace (MobileFace) 为例, 由图 3 可以看出, 随着ω从 1 开始增加, 攻击成功率不断增加, 后趋于平稳. 增长至 36 后, 攻击成功率达到峰值开始呈下降趋势, 下降的主要原因是ω越大扰动中超过ε的像素值越多, 而超过的部分在第 3 阶段训练结束后被裁剪. 在固定取值中其对应的最高攻击成功率为 89.08%, 低于使用多阶段策略的攻击成功率 91.58%, 证明多阶段训练策略对提升攻击性能的有效性.

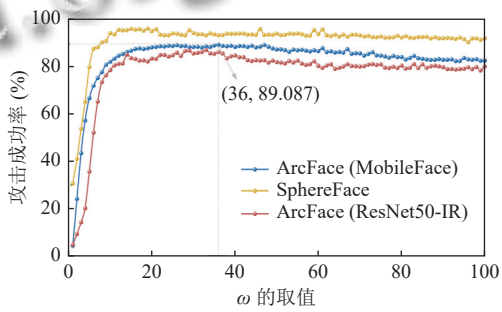

图 3 LFW 数据集上不同ω对应的攻击成功率

3) 训练图像数量对攻击效果的影响. 本文方法中参与共性计算对抗扰动的图像数量直接影响扰动攻击效果. 详细列举共性计算中不同的人脸图像数量 N 对于最终得到的扰动攻击成功率的影响, 实验结果如图 4 所示.

图 4 显示: 逐渐增加参与扰动共性计算的图像数

量, 初始阶段扰动的跨图像攻击能力迅速提升, 达到瓶颈后呈现微弱地增长. 随着参与扰动计算的图像数量增加, 训练时间与显存资源需求逐步增加, 所以本文选择在 3 个人脸识别模型均取得良好的攻击成功率且值偏小的: $N=24$.

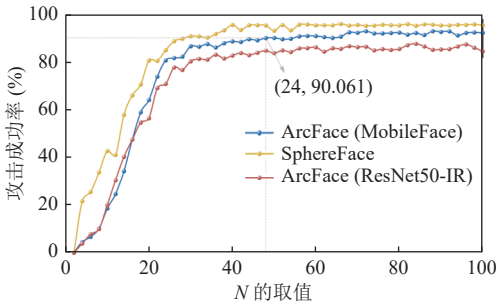

图 4 LFW 数据集上不同N对应的攻击成功率

4) 不同初始化对攻击效果的影响. 为了验证不同的初始化扰动对攻击的影响, 本文选取 3 种不同初始化结果对比攻击成功率, 分别是将扰动置为 0 值初始化、扰动在$(-\varepsilon,+\varepsilon)$随机值初始化以及使用不同的标准人脸图像$(S_1,S_2,S_3)$初始化. 对比实验结果如表 3 所示. 使用特定身份的人脸图像初始化时, 攻击成功率较低. 原因是扰动携带了除身份信息外的冗余信息, 且在迭代结束时仍包含超过 ε 的扰动, 裁剪后导致攻击成功率下降. 相比之下, 随机初始化扰动的攻击效果稍好于 0 值初始化. 由于受到 PGD (projected gradient descent) 方法[31]启发, 随机初始化的对抗扰动更具攻击鲁棒性. 因此, 本文选择在限制区间$(-\varepsilon,+\varepsilon)$内随机初始化通用对抗扰动.

表 3 扰动初始化方式其攻击成功率 (%)

初始化方式	攻击成功率
使用0值初始化	90.25
使用$(-\varepsilon,+\varepsilon)$ 中随机值初始化	91.58
使用S_1 初始化	88.96
使用S_2初始化	87.20
使用S_3初始化	87.72

3.4 对抗样本可视化

将通用对抗扰动经过处理后的可视化结果如图 5 所示. 观察可知, 本文方法中的扰动从随机初始化到训练完成, 更多地学习到人脸特征的细节信息. 叠加通用对抗扰动的人脸图像可视化结果如图 6 所示, 其中 SVD-UAP、SGA 以及本文方法进行特定目标攻击, 目标特定身份为 John, Cos-UAP 进行非特定目标攻击. 图 6 中, 每一张对抗样本给出人脸识别模型判定可能性最高的身份、特征向量之间的相似度, 以及对抗样本与原图的结构相似度.

图 5 对抗扰动可视化结果

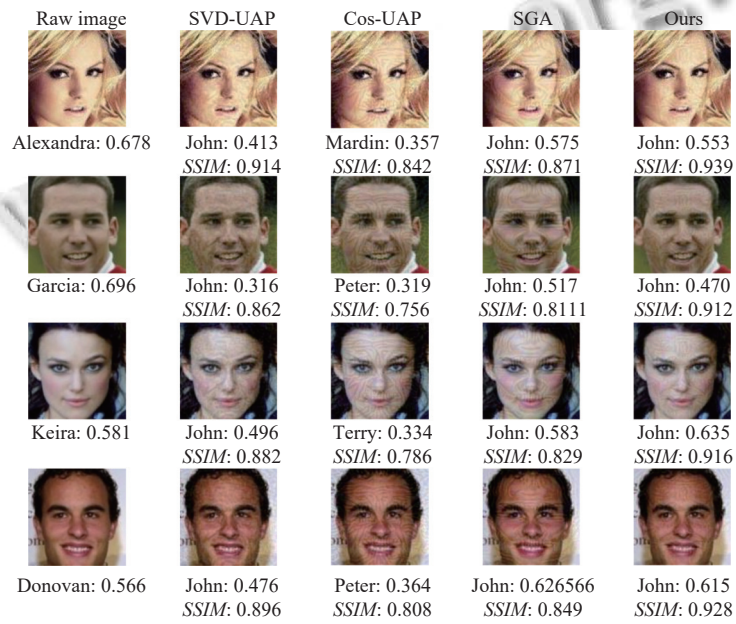

图 6 生成的对抗样本

4 结束语

本文提出了一种适用于人脸识别网络的通用对抗扰动生成方法. 该方法通过迭代梯度的共性计算增强了对抗扰动的跨图像迁移能力, 并通过引入主导性特征损失来提高扰动的攻击成功率. 最后, 通过采用多阶段训练策略, 实现了攻击效果与视觉质量的均衡. 相对于当前在分类领域表现最佳的 SGA 生成方法, 该方法在攻击成功率和扰动视觉质量方面都表现出显著的优势. 然而, 本文方法依然存在一些局限性, 特别是在无法获取人脸识别模型特征向量的黑盒环境中, 如何有效地嵌入主导性特征仍然是一个需要深入探讨的问题.

参考文献

1 Liu JY, Zhang WM, Fukuchi K, *et al*. Unauthorized AI cannot recognize me: Reversible adversarial example. Pattern Recognition, 2023, 134: 109048. [doi: 10.1016/j.patcog.2022.109048]

2 Hasan MR, Guest R, Deravi F. Presentation-level privacy protection techniques for automated face recognition —A survey. ACM Computing Surveys, 2023, 55(13s): 286.

3 Ren M, Zhu YH, Wang YL, *et al*. Perturbation inactivation based adversarial defense for face recognition. IEEE Transactions on Information Forensics and Security, 2022, 17: 2947–2962. [doi: 10.1109/TIFS.2022.3195384]

4 Zheng X, Fan YB, Wu BY, *et al*. Robust physical-world attacks on face recognition. Pattern Recognition, 2023, 133: 109009. [doi: 10.1016/j.patcog.2022.109009]

5 Deng YP, Karam LJ. Frequency-tuned universal adversarial attacks on texture recognition. IEEE Transactions on Image Processing, 2022, 31: 5856–5868. [doi: 10.1109/TIP.2022.3202366]

6 Xu YH, Ghamisi P. Universal adversarial examples in remote sensing: Methodology and benchmark. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5619815.

7 Mi JX, Wang XD, Zhou LF, *et al*. Adversarial examples based on object detection tasks: A survey. Neurocomputing, 2023, 519: 114–126. [doi: 10.1016/j.neucom.2022.10.046]

8 Kim H, Park J, Lee J. Generating transferable adversarial examples for speech classification. Pattern Recognition, 2023, 137: 109286. [doi: 10.1016/j.patcog.2022.109286]

9 Komkov S, Petiushko A. AdvHat: Real-world adversarial attack on ArcFace face ID system. Proceedings of the 25th International Conference on Pattern Recognition. Milan: IEEE, 2021. 819–826.

10 Yin BJ, Wang WX, Yao TP, *et al*. Adv-Makeup: A new imperceptible and transferable attack on face recognition. Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal: IJCAI, 2021. 1252–1258.

11 Zolfi A, Avidan S, Elovici Y, *et al*. Adversarial mask: Real-world universal adversarial attack on face recognition models. Proceedings of the 2023 European Conference on Machine Learning and Knowledge Discovery in Databases. Grenoble: Springer, 2023. 304–320.

12 Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. Proceedings of the 2nd International Conference on Learning Representations. Banff: ICLR, 2014. 1–10.

13 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

14 Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017. 99–112.

15 Dong YP, Liao FZ, Pang TY, *et al*. Boosting adversarial attacks with momentum. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9185–9193.

16 Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2574–2582.

17 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose: IEEE, 2017. 39–57.

18 白祉旭, 王衡军, 郭可翔. 基于深度神经网络的对抗样本技术综述. 计算机工程与应用, 2021, 57(23): 61–70.

19 Zhang DH, Zhang TY, Lu YP, *et al*. You only propagate once: Accelerating adversarial training via maximal principle. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: ACM, 2019. 21.

20 Zhang CN, Benz P, Karjauv A, *et al*. Universal adversarial perturbations through the lens of deep steganography: Towards a Fourier perspective. Proceedings of the 2021 AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 3296–3304.

21 Moosavi-Dezfooli SM, Fawzi A, Fawzi O, *et al*. Universal adversarial perturbations. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017. 86–94.

22 Kamath S, Deshpande A, Subrahmanyam KV, *et al*. Universalization of any adversarial attack using very few test examples. Proceedings of the 5th Joint International Conference on Data Science & Management of Data. Bangalore: ACM, 2022. 72–80.

23 Zhang CN, Benz P, Karjauv A, *et al*. Data-free universal adversarial perturbation and black-box attack. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE Press, 2021. 7848–7857.

24 Ud Din S, Akhtar N, Younis S, *et al*. Steganographic universal adversarial perturbations. Pattern Recognition Letters, 2020, 135: 146–152. [doi: 10.1016/j.patrec.2020.04.025]

25 Liu XN, Zhong YY, Zhang YH, *et al*. Enhancing generalization of universal adversarial perturbation through gradient aggregation. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE Press, 2023. 4412–4421.

26 Zhong YY, Deng WH. Towards transferable adversarial attack against deep face recognition. IEEE Transactions on Information Forensics and Security, 2020, 16: 1452–1466.

27 Jia S, Yin BJ, Yao TP, *et al*. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, 2022. 34136–34147.

28 Jia XJ, Zhang Y, Wei XX, *et al*. Prior-guided adversarial initialization for fast adversarial training. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 567–584.

29 Deng JK, Guo J, Xue NN, *et al*. ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4685–4694.

30 Liu WY, Wen YD, Yu ZD, *et al*. SphereFace: Deep hypersphere embedding for face recognition. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6738–6746.

31 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: ICLR, 2018. 1–28.

(校对责编: 孙君艳)