基于对抗攻击换脸的人脸保护系统设计与实现

- 答辩人：迟宇翔 MG1937004
- 导　师：申富饶 教授

目录
CONTENTS

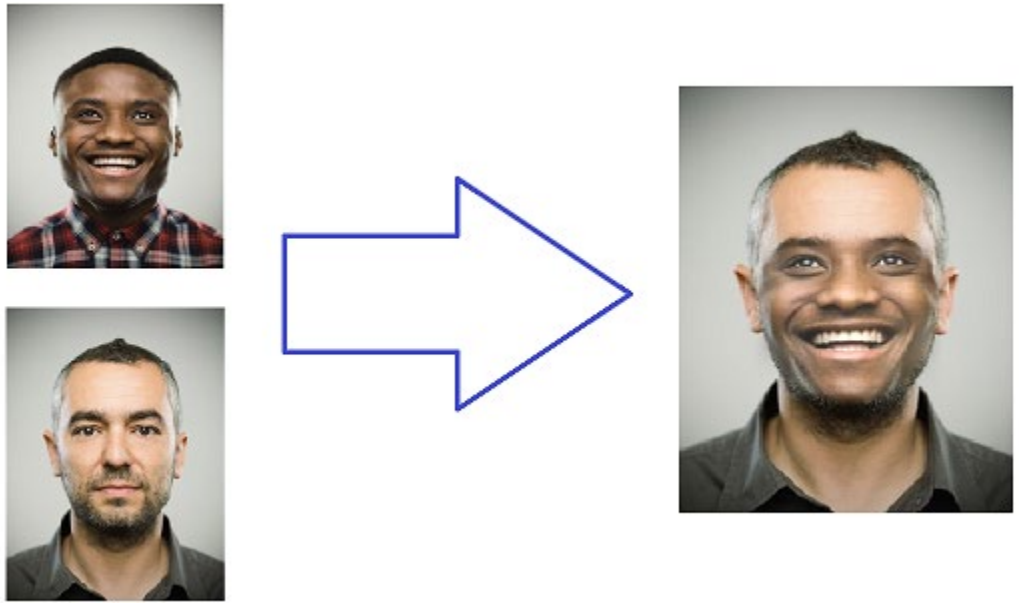
南京大学
NANJING UNIVERSITY

第一部分

研究背景

研究背景 | 研究现状

研究背景：智能化时代下的换脸技术



- 互联网中出现许多换脸应用

  FaceSwap, FaceApp, ZAO, …

- 脸部替换，交换两张图中的人脸

- 不良影响：众多名人、明星被换脸
- 滥用换脸技术将造成肖像权侵犯

研究背景：现有对策

- 换脸应用下架
  - 有效解决侵权问题，但使得换脸技术无法应用
- 对伪造图像进行检测
  - 无法解决侵权问题，且能被绕过
- 对模型攻击破坏换脸结果
  - 基于对抗攻击，在原图上添加扰动，破坏换脸图像
  - 现有工作可以一定程度上阻止换脸，但效果不稳定，不具有身份掩蔽作用

研究现状

- Yeh, etc. 基于PGD攻击CycleGAN
- 合成图像出现明显破坏痕迹
- 破坏形式不稳定，未掩蔽被换脸人物身份

南京大学
NANJING UNIVERSITY

第二部分

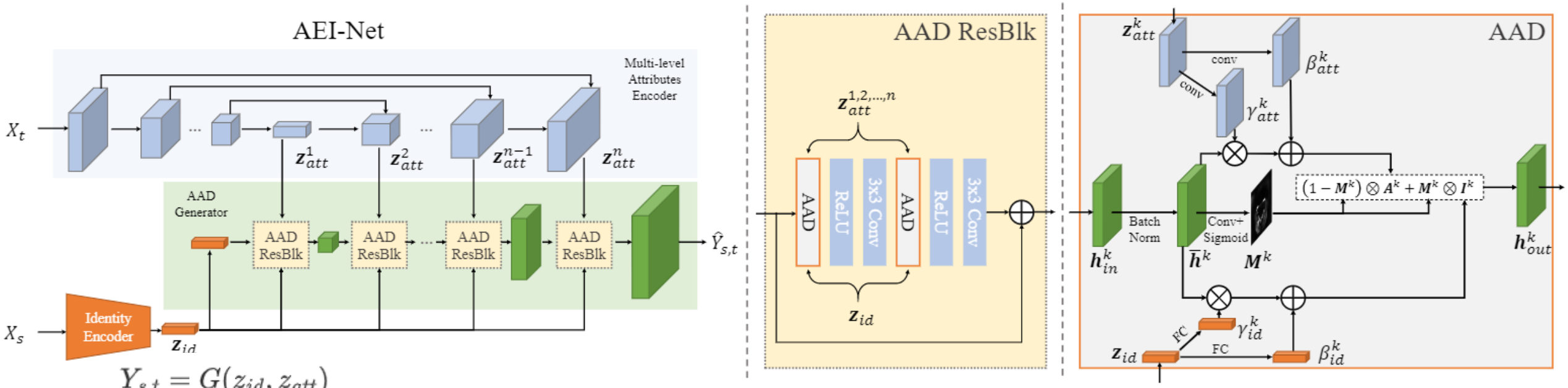
研究内容

研究目标 | 实现思路 | 攻击方法

研究目标

如何实现保护人脸：通过攻击换脸模型，在人脸图像中添加扰动，使其再次换脸后失败

所攻击的模型：FaceShifter换脸模型



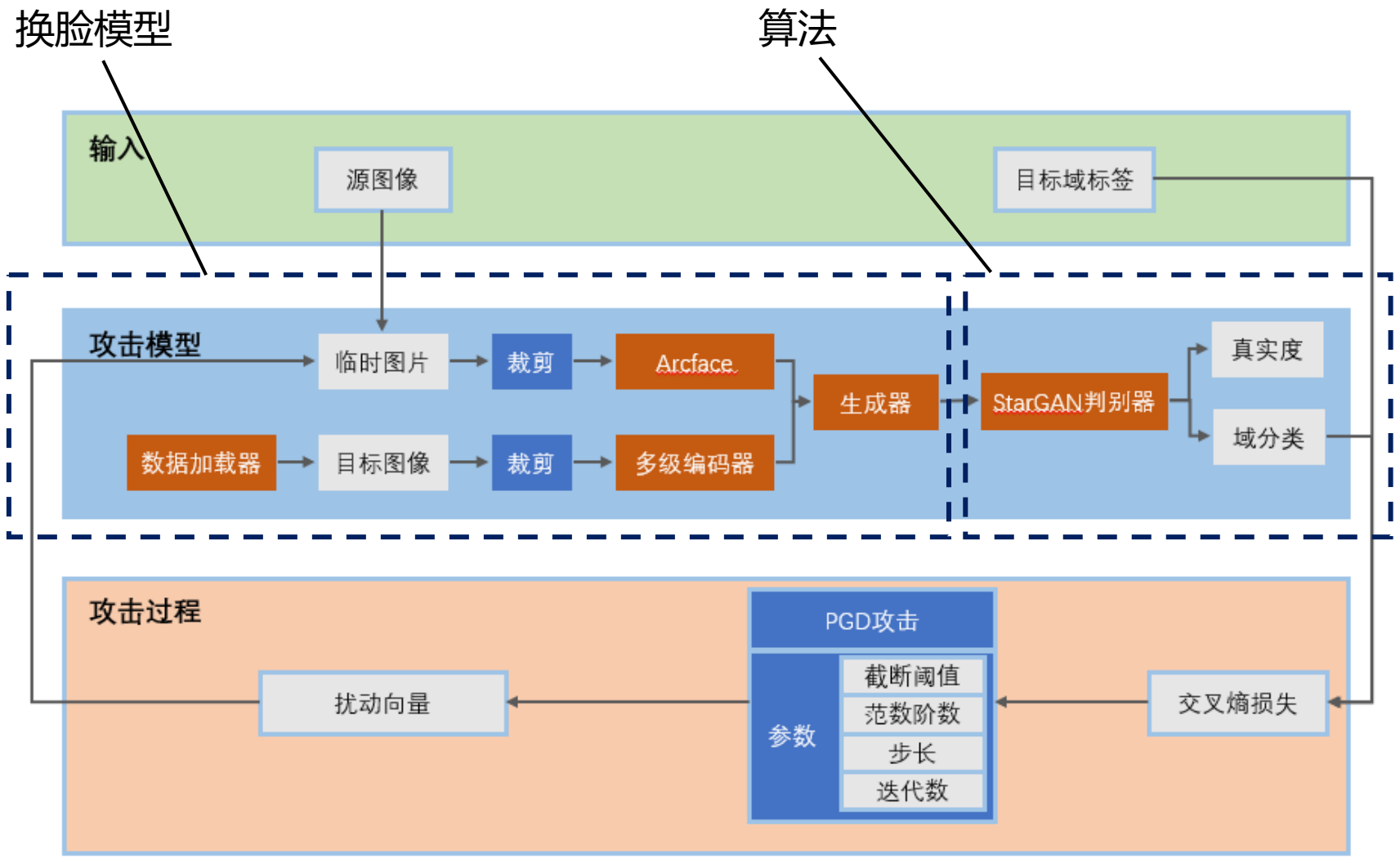
$$Y_{s,t} = G(z_{id}, z_{att})$$

z_{id}：源图像身份向量，由arcface身份编码器提取，包含人物身份特征

z_{att}：目标图像外观属性向量

实现思路

- 串联换脸模型和一算法
- 使用PGD攻击法对串联模型进行攻击
- 提出了常规攻击、定向攻击、身份攻击、可控语义攻击四种攻击方法



常规攻击

- 攻击目标：最小化换脸图片真实度
- 基于StarGAN判别器

$$L(\theta, X_s + r_{adv}, X_t) = -D_{src}(G(X_s + r_{adv}, X_t))$$

$$X_{adv} = X_s + r_{adv}$$

- 效果：换脸图片变得不真实
- 破坏程度有限

定向攻击

- 攻击目标：使换脸图片中心出现黑块
- 对输出图像作mask

$$M = \frac{1}{N}\begin{bmatrix} 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 1 & 1 & 1 & \dots & 0 \\ 0 & \dots & 1 & 1 & 1 & \dots & 0 \\ 0 & \dots & 1 & 1 & 1 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$L(\theta, X_s + r_{adv}, X_t) = M \circ G(X_s + r_{adv})$$

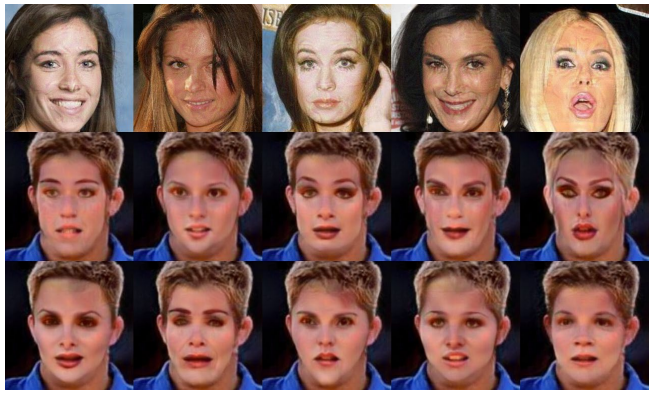
- 效果：图像中央明显变黑，破坏稳定

身份攻击

- 攻击目标：使换脸图片身份发生改变
- 攻击arcface提取的身份特征向量

$$L(\theta, X_s + r_{adv}, X_t) = \frac{1}{n}\sum_{k=1}^{n}\left\|z_{id} - z'_{id}\right\|_2^2$$

- 效果：身份改变，隐藏被换脸人身份

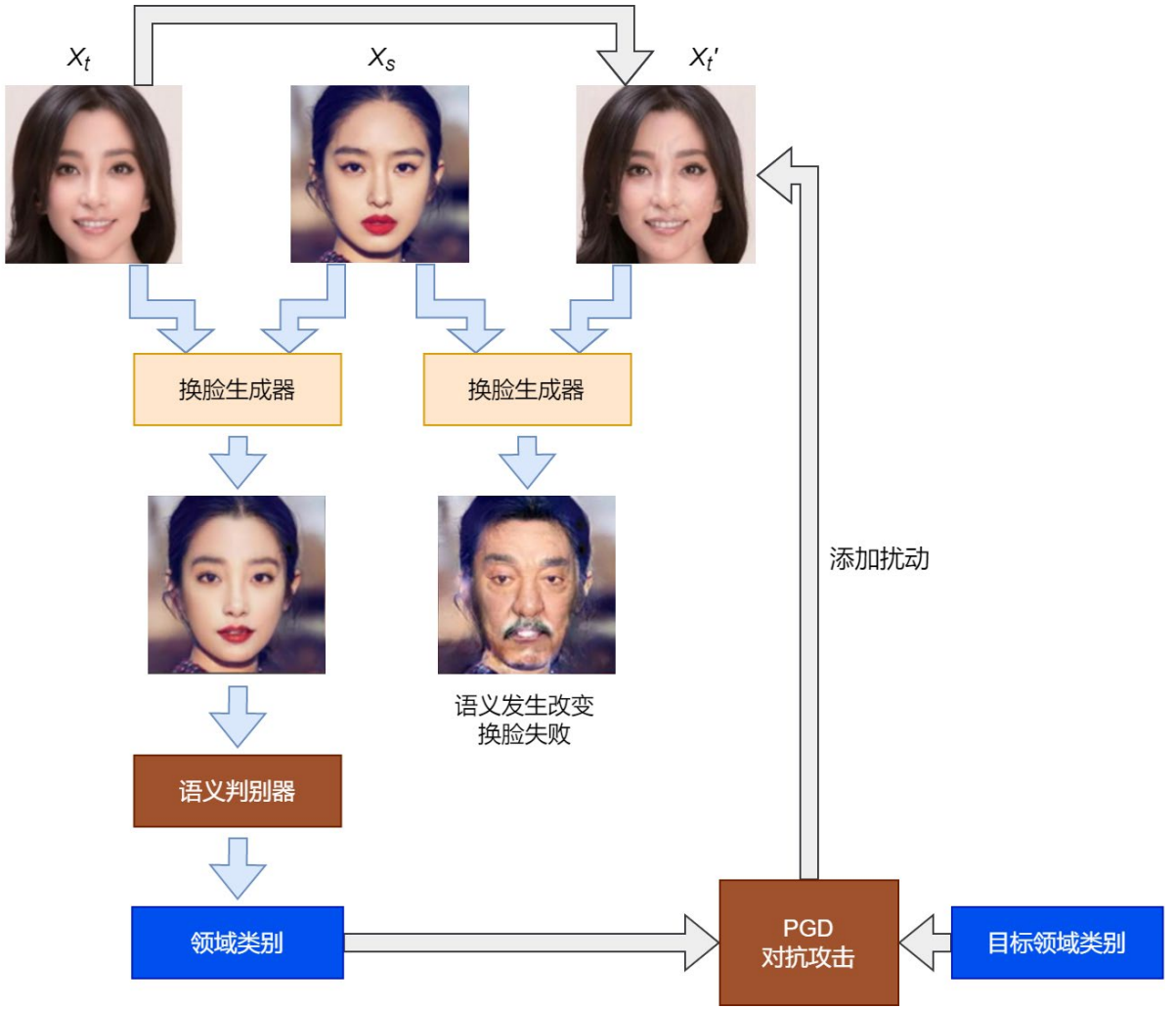


可控语义攻击

- 基于StarGAN判别器提取外观语义特征
- 攻击后改变换脸图片语义：年龄、性别、发色、衰老等

$$L(\theta, X_s + r_{adv}, X_t) = H(D_{cls}(Y_{adv}), c)$$

- 效果：可以控制语义发生变化的类型

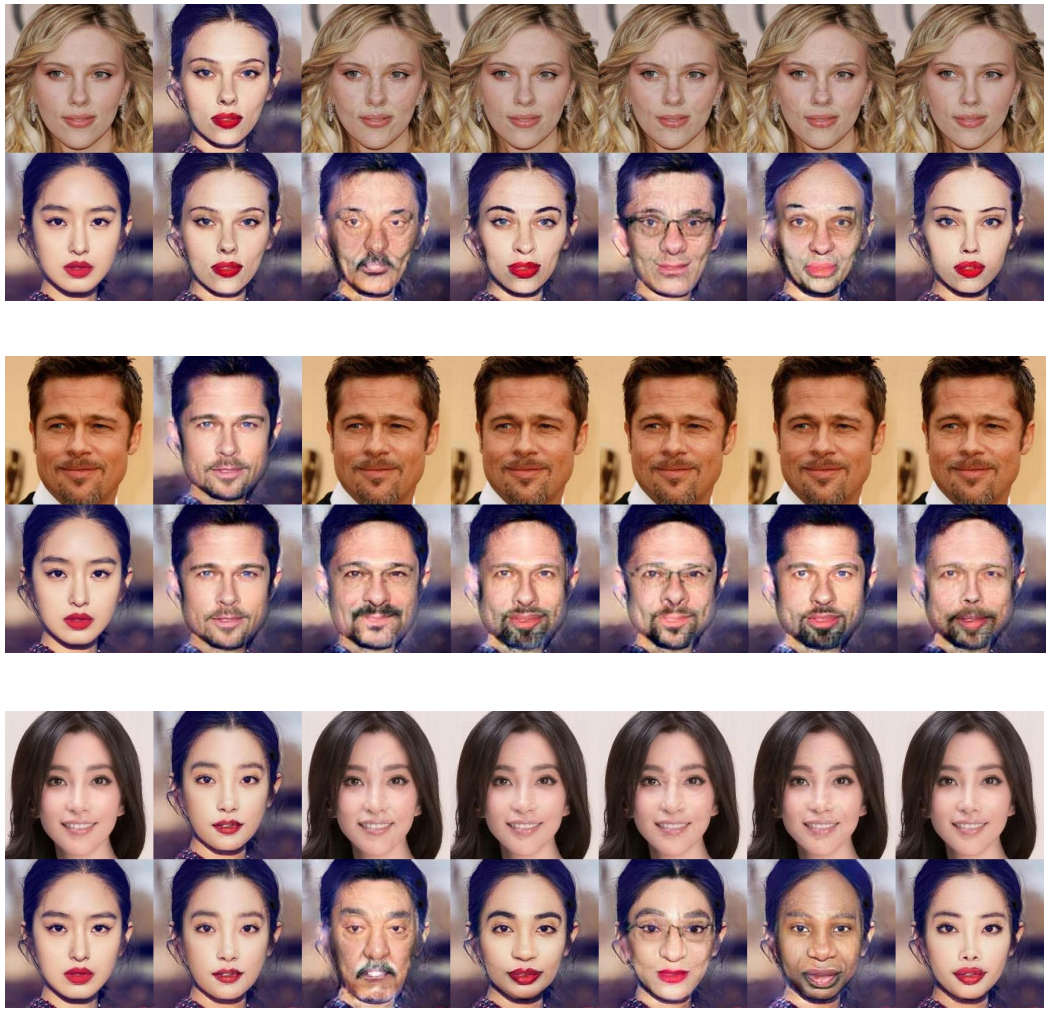


可控语义攻击

- 基于StarGAN判别器提取外观语义特征
- 攻击后改变换脸图片语义：年龄、性别、发色、衰老等

$$L(\theta, X_s + r_{adv}, X_t) = H(D_{cls}(Y_{adv}), c)$$

- 效果：可以控制语义发生变化的类型



南京大学
NANJING UNIVERSITY

第三部分

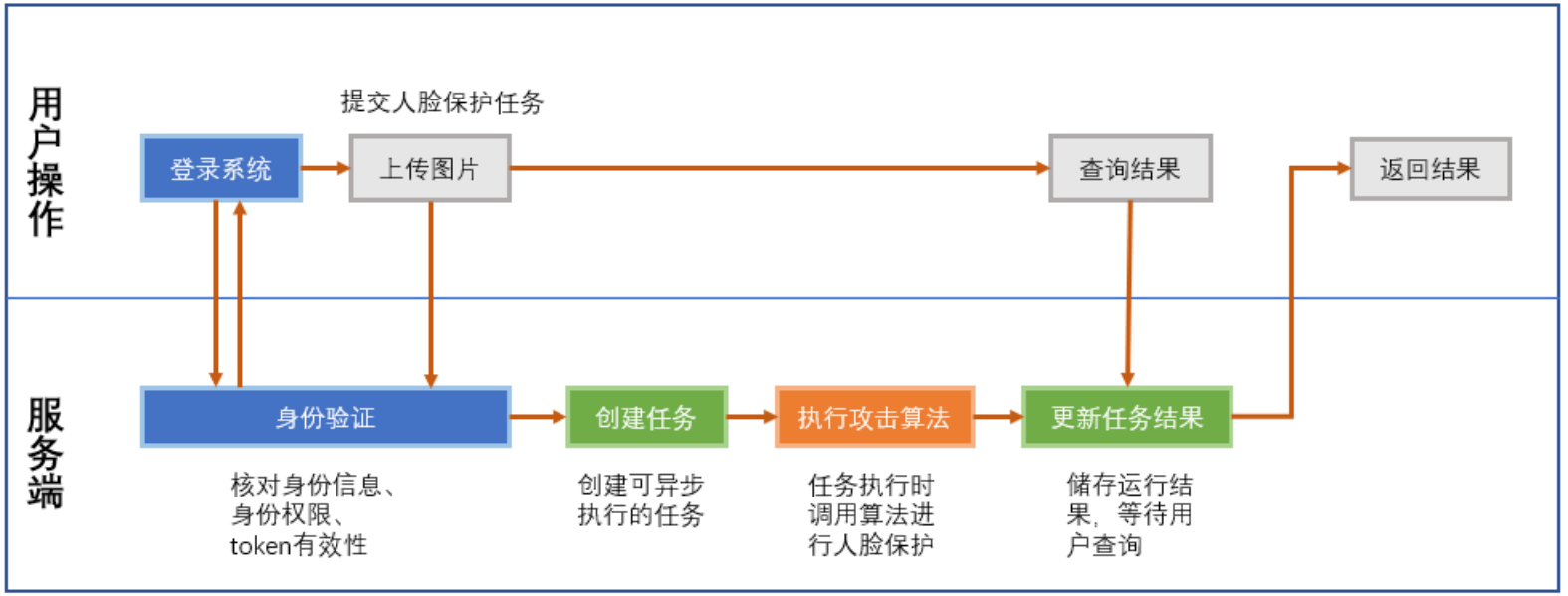
系统实现

系统需求 | 系统设计 | 系统实现

人脸保护系统：系统设计

系统需求：

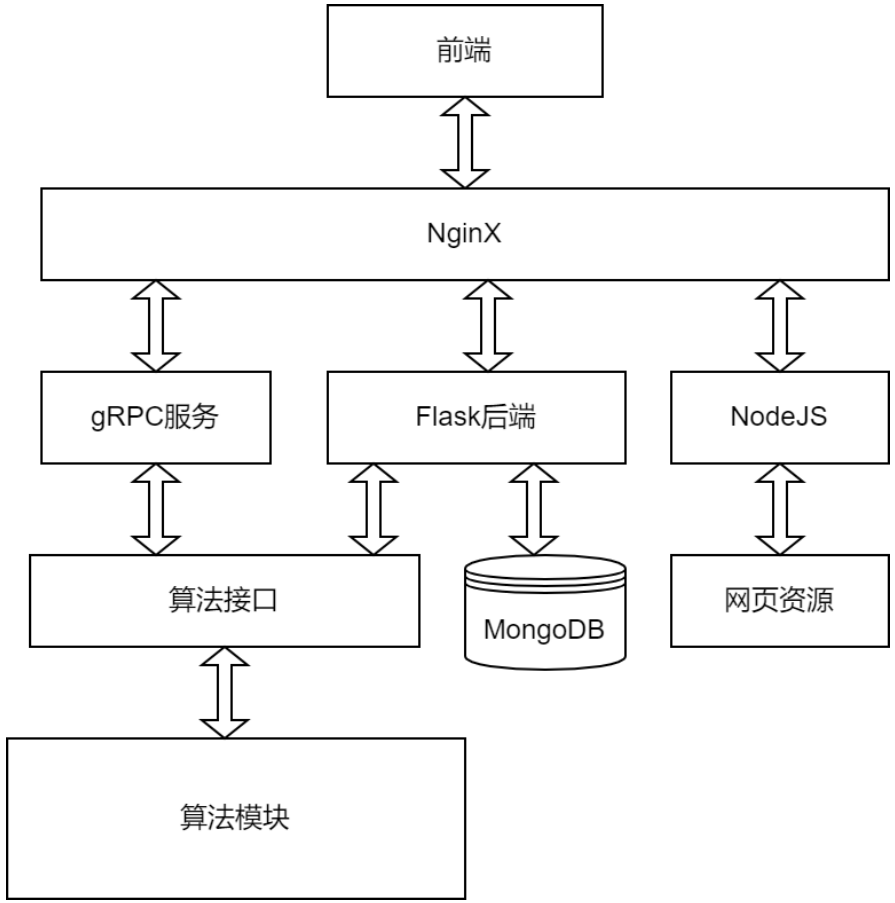
- 人脸保护功能
- 多种攻击方式
- 安全的身份验证与访问控制
- 异步任务机制
- 良好的历史查询机制



人脸保护系统：系统设计

系统架构设计：

- B/S架构，前后端分离
- Flask+NodeJS+React
- NginX反向代理
- MongoDB
- 对外提供RestFul API与gRPC服务



人脸保护系统：系统实现

身份验证模块

- 用户注册、登录、登出
- 基于双token验证机制
- 访问token和刷新token
- 基于flask_jwt_extended
- HTTP-only cookie

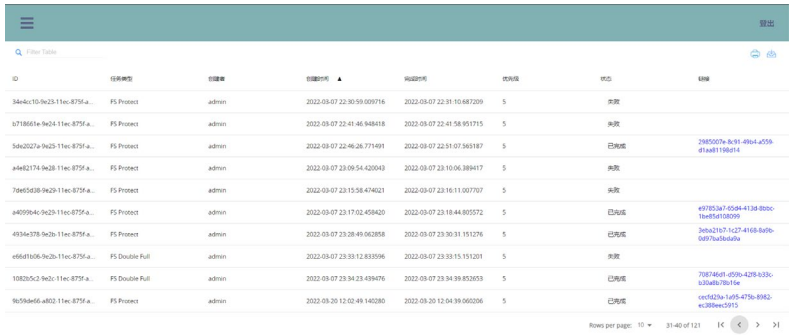
算法模块

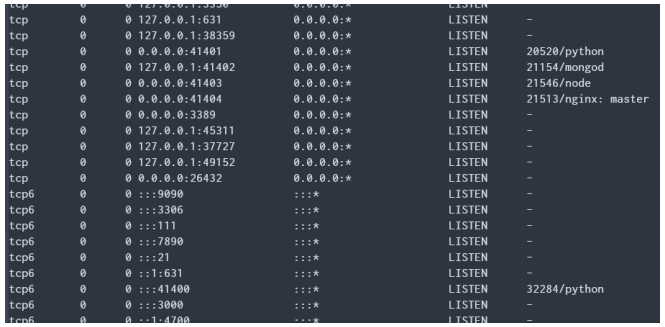
- 带参数选择的攻击
- 常规攻击/定向攻击/身份攻击/可控语义攻击
- epsilon/步长/步数
- 换脸算法

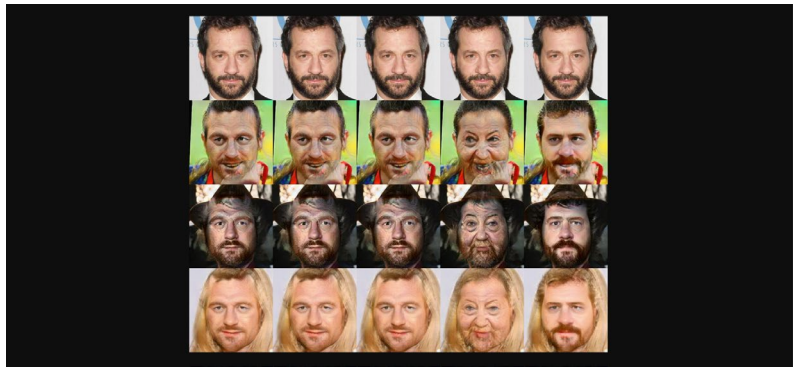
任务管理模块

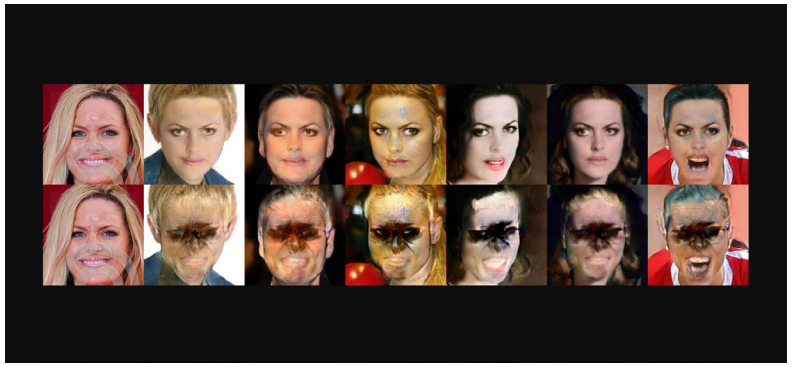
- 异步任务机制
- 基于Multiprocessing多进程框架
- 多任务多GPU并行
- 结果与历史记录查询

人脸保护系统：系统运行效果









南京大学
NANJING UNIVERSITY

第四部分

研究生期间成果

工作成果

相关成果

专利

- 迟宇翔，范彧. 基于强化学习的多目标复杂交通场景下自动驾驶解决方法. 专利申请号：202210370991.7
- 范彧，迟宇翔, 俞扬. 一种基于对抗学习的数据隐私保护方法. 专利申请号：202210372873.X

南京大学
NANJING UNIVERSITY

第五部分

全文总结

全文总结

全文总结

可控制图像语义改变的攻击

- 结合PGD攻击与StarGAN判别器进行对抗攻击
- 攻击所造成的改变类型可人为控制
- 实现稳定、可靠的人脸保护

三种不同类型的攻击

- 常规攻击/定向攻击/身份攻击
- 真实度降低/中央出现黑块/身份改变
- 实现多角度多样性的稳定人脸保护

人脸保护系统

- 将所提出的攻击方法应用其中
- 健壮的系统框架，提供可靠的人脸保护服务
- 异步任务机制提升系统性能

南京大学
NANJING UNIVERSITY

谢谢！