Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

视频深度伪造及检测技术——攻与防

博士研究生 张浩然

2023 年 02 月 19 日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 总结
- 参考文献

背景简介

- 预期收获
  - 了解视频伪造的常用手段
  - 理解视频深度伪造程序FaceSwap原理
  - 理解小样本条件下深度伪造图像生成方法
  - 了解视频伪造检测方法及对抗手段

背景简介

- 视频造假
  - 张冠李戴、挪用剪辑
  - 成本低廉、易产生误导，需仔细分辨
- 视频编辑
  - 使用软件进行逐帧修剪、加特效
  - 对伪造者水平要求高，普通用户很有经验才能分辨

一分成本一分效果

背景简介

- 深度视频伪造
  - 重现（reenactment）
    - 使原角色可以学习目标角色特定动作
    - 动画角色说话口型、电影后期制作表情微调
  - 编辑（editing）
    - 添加、更改或删除目标身份的属性
  - 合成（synthesis）
    - 没有目标身份为基础的情况下创建角色
  - 替换（replacement）
    - 人脸替换
    - 神态、眼神均可以较好的模仿

背景介绍

- AI Deepfakes：人工智能深度换脸技术
- FaceSwap
  - 可以实现低成本高效果脸部替换，可以将B的面部特征替换到A上，生成伪造视频
  - 与PS、AE等视频图像编辑工具不同，操作者并不需要懂得太多技术
  - 收集到足够素材，程序自动完成生成
  - FaceSwap是GitHub开源的多平台Deepfakes软件，基于Python开发

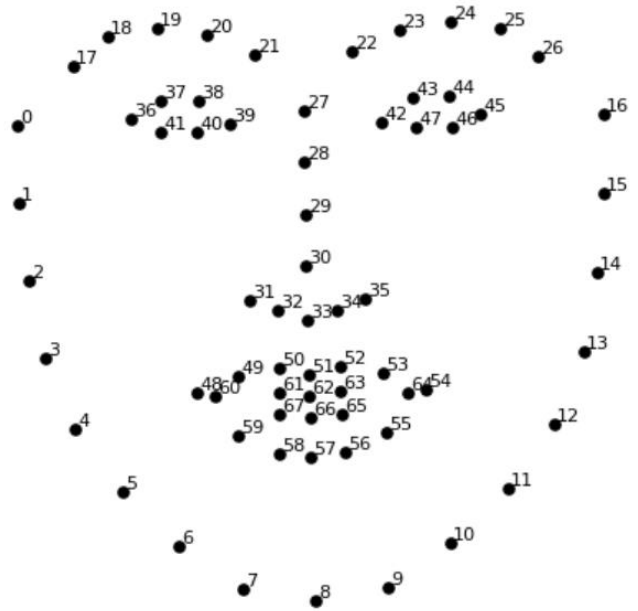
算法原理 FaceSwap

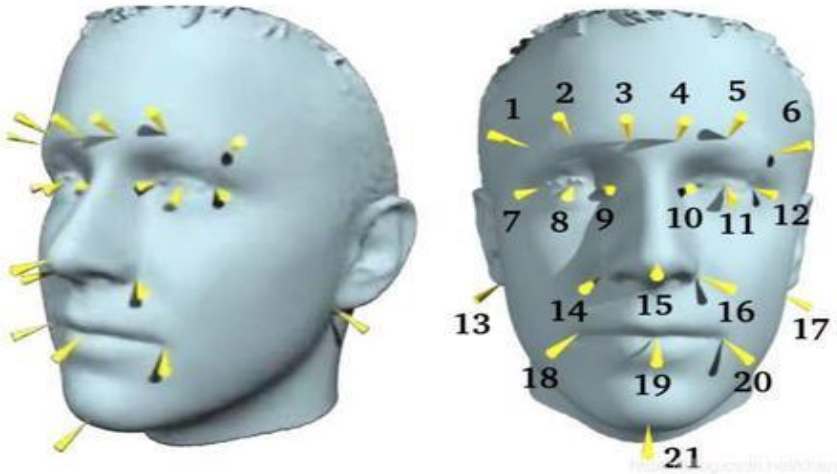
T	目标	将原视频中的人脸替换为目标人脸
I	输入	含有一定量原、目标人脸的视频
P	处理	1、提取原视频中原始人脸、人脸标记点 2、使用编码器对人脸编码，提取人脸潜在特征向量 3、使用解码器对特征向量解码，复原符合解码器特征的人脸图片 4、将原视频中人脸替换输出
O	输出	进行人脸替换后的伪造视频

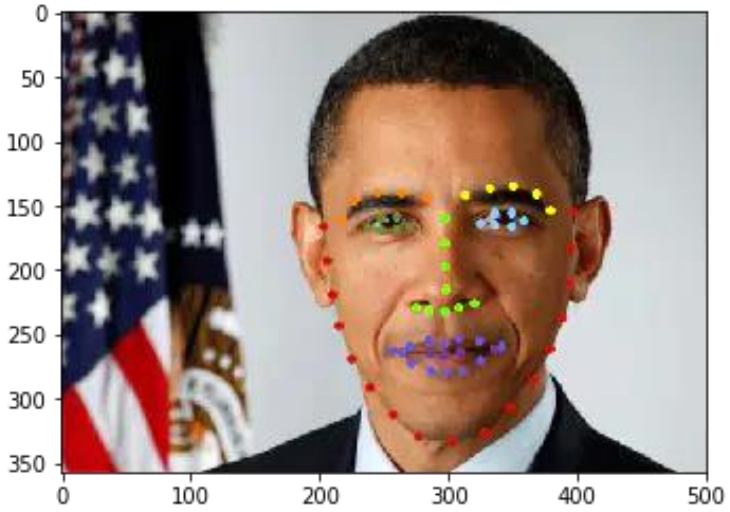
P	问题	生成目标人脸替换原人脸的图片用于视频伪造
C	条件	拥有一定量的原、目标人脸图片供模型训练
D	难点	1、如何有效学习输入人脸信息 2、如何生成特征组合的人脸
L	水平	GitHub 43.5k stars

算法原理

- 人脸提取
  - 视频分解为帧
    - 拆分视频为图片（帧）
  - 人脸特征提取（land-mark）
    - 检测：判别每一帧图片中是否存在人脸
    - 对齐：根据面部特定点，将面部与网格对齐
    - 蒙版生成：识别面部所在区域，屏蔽障碍物

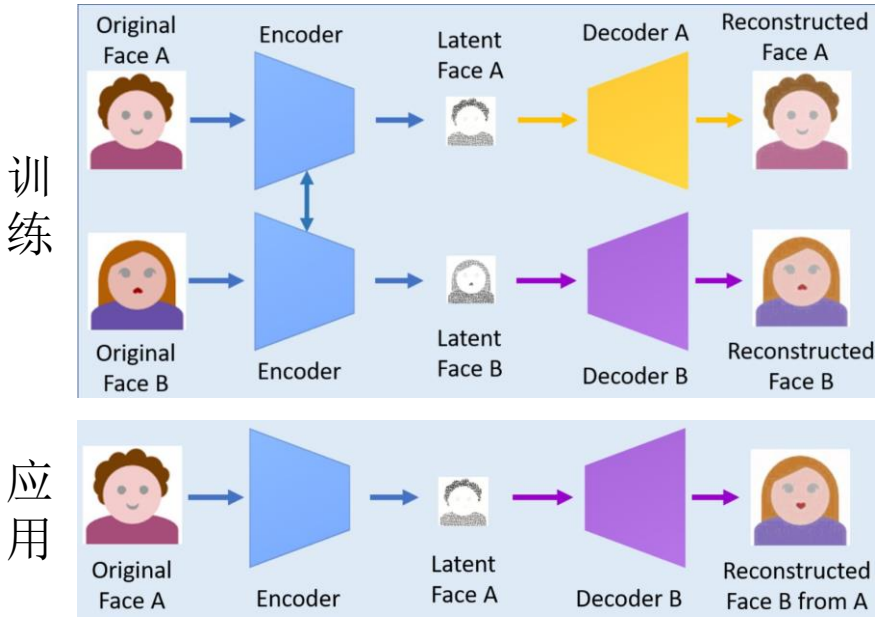






算法原理 自动编码器

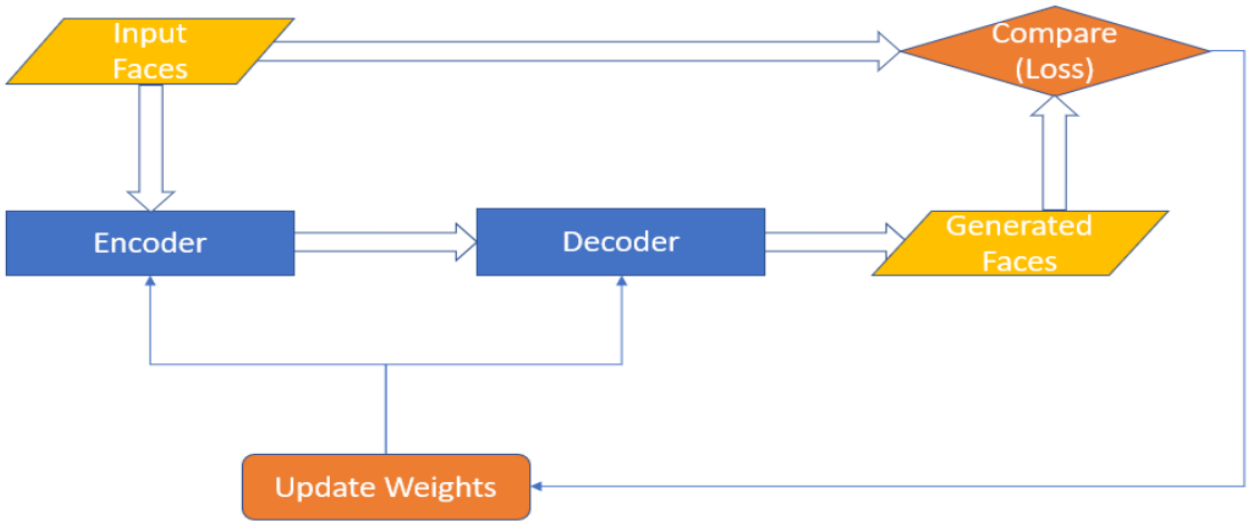
- 人脸特征学习
  - Encoder-Decoder自动编码器模型
  - 共享Encoder提取面部抽象特征
  - 切换Decoder生成面部具体特征
  - 训练3个模型并不断更新权重
  - 损失函数（SSIM）
    - 生成面部图像与原图像相似度



$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

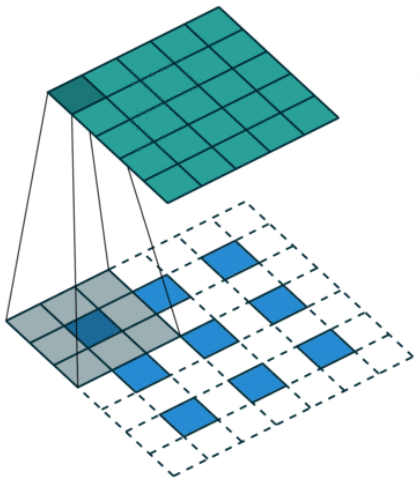
像素平均值

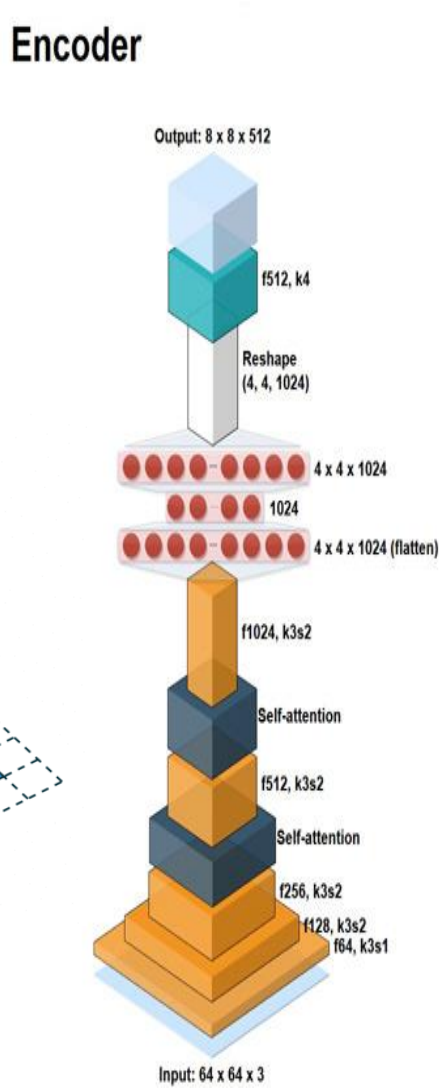
亮度相似性　　对比、结构相似性

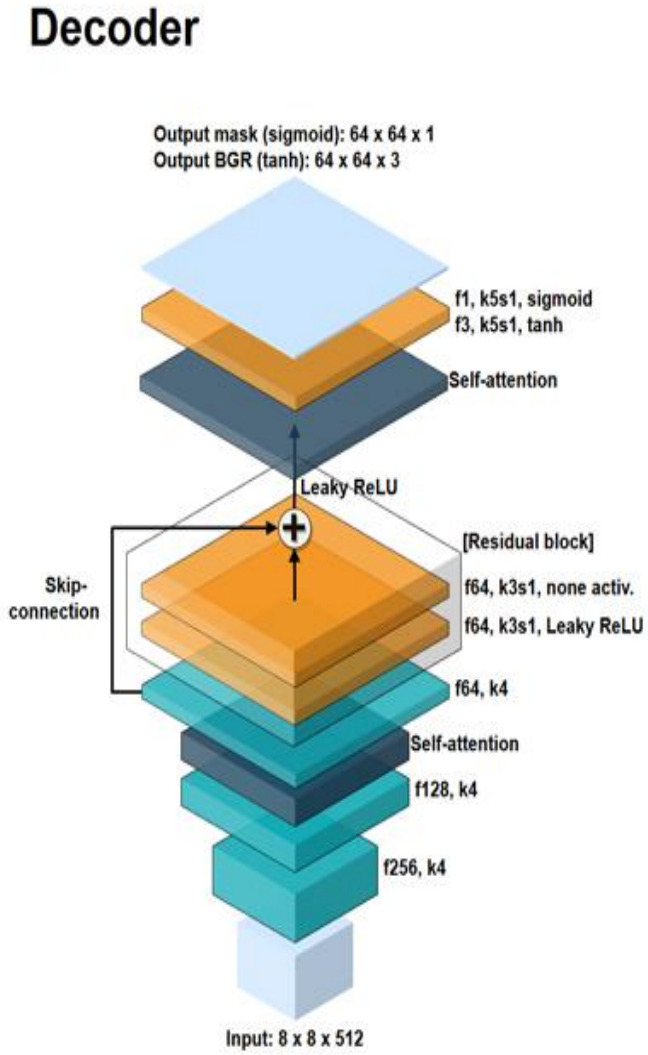


算法原理 自动编码器

- 特征提取-Encoder
  - 面部特征抽象为潜在向量
  - 潜在特征：面部特征、神态、表情、明暗
  - 卷积层、自注意力层进行特征提取
- 特征恢复-Decoder
  - 潜在向量重构为图像
  - 反卷积层：生成人脸
    - 特征映射到原始图像空间
    - 重建人脸数据
- 应用方式（B换给A）
  - 依据A的潜在特征，生成B的样子





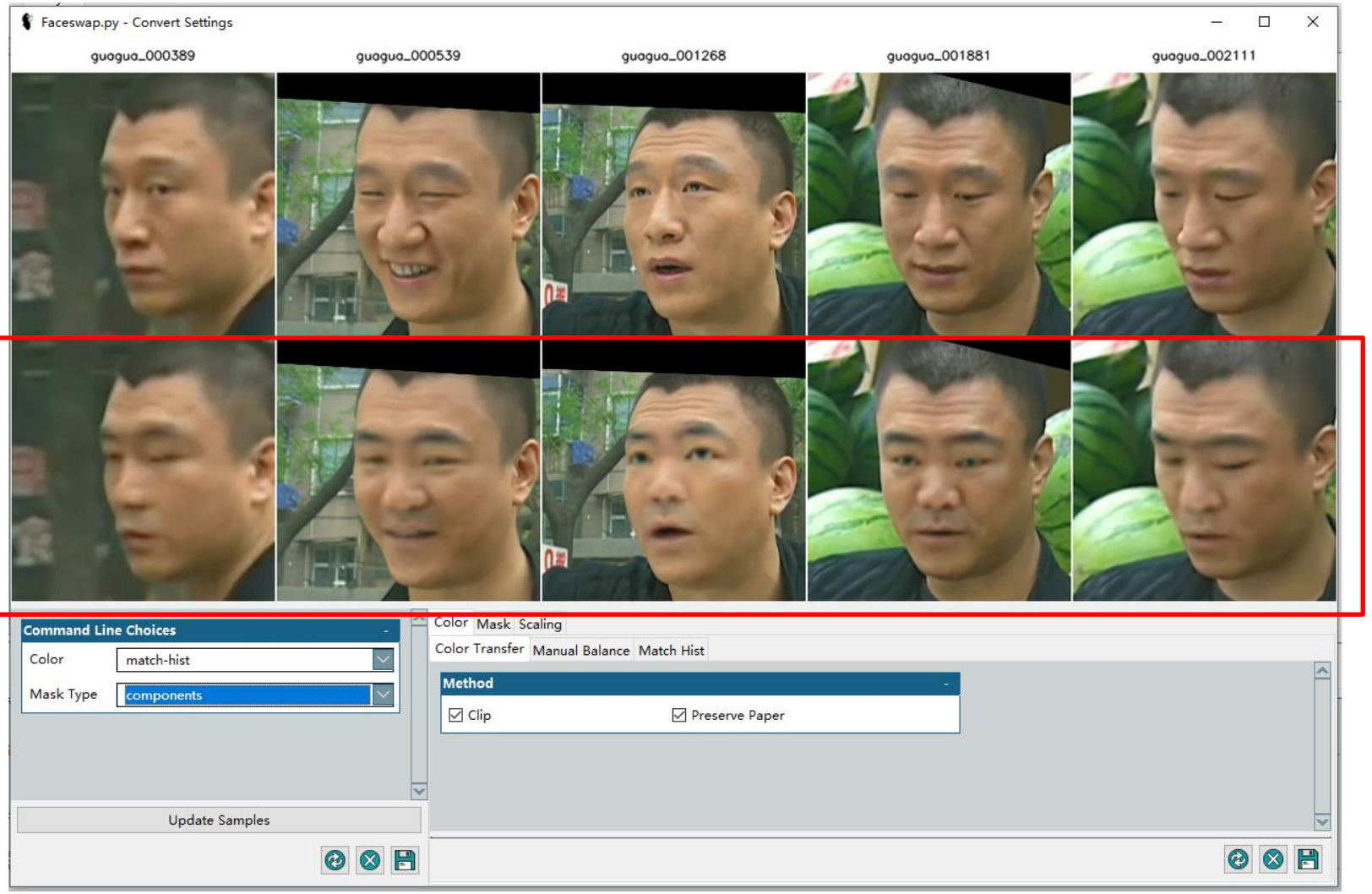


实验效果 FaceSwap

- FaceSwap实验

+

=



实验效果 FaceSwap

总结分析

- 基于自动编码器的FaceSwap方法
  - 具象 -> 抽象 -> 具象转化过程，提取潜在空间特征
  - 将原视频面部潜在特征与目标面部样貌特征相结合，生成伪造人脸图片及视频
  - 算法简单，噪声鲁棒性、可解释性较强
- 自动编码器应用场景
  - 文本：文本分类、情感分析、文本生成
  - 语音：音频降噪、音频伪造、语音识别
- 算法缺点
  - 依赖训练集，训练集质量、数量直接影响结果
  - 模型泛化性较差，仅能针对训练样本包含的人脸信息生成图片
  - 损失函数计算方式导致图片模糊

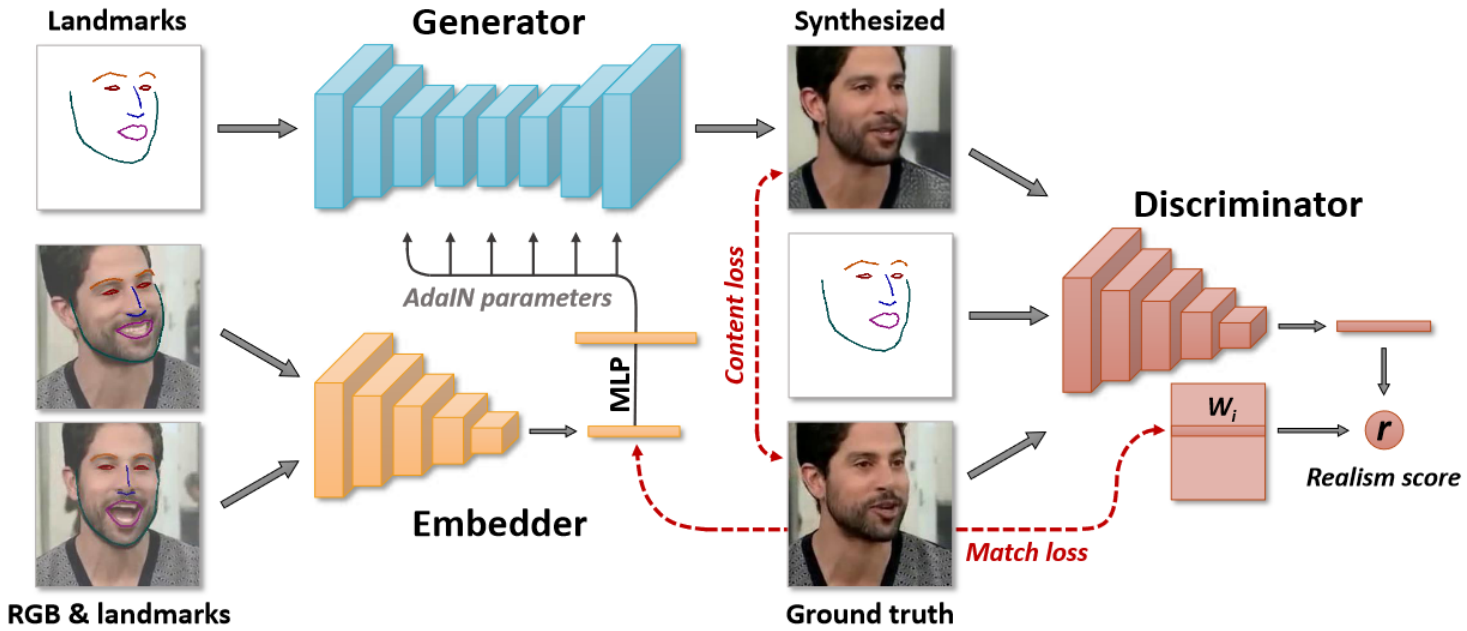
算法原理 Few-Shot Adversarial Learning

T	目标	使用少量图像进行人脸动作伪造
I	输入	少量或单张人脸图像、目标姿态的landmark标记
P	处理	1、建立模型针对大量人脸样本进行元学习 2、使用元学习模型参数对测试模型初始化 3、快速在线学习目标人脸图像，微调网络参数
O	输出	进行人脸姿态调整后的伪造图像

P	问题	可供训练的目标样本数量少
C	条件	目标图片面部特征清晰
D	难点	1、面部的光线、几何、运动较为复杂 2、人眼对人类头部外观建模的非常敏锐（恐怖谷现象） 3、针对小样本测试环境，如何快速完成迁移
L	水平	ICCV2019（计算机视觉顶会）

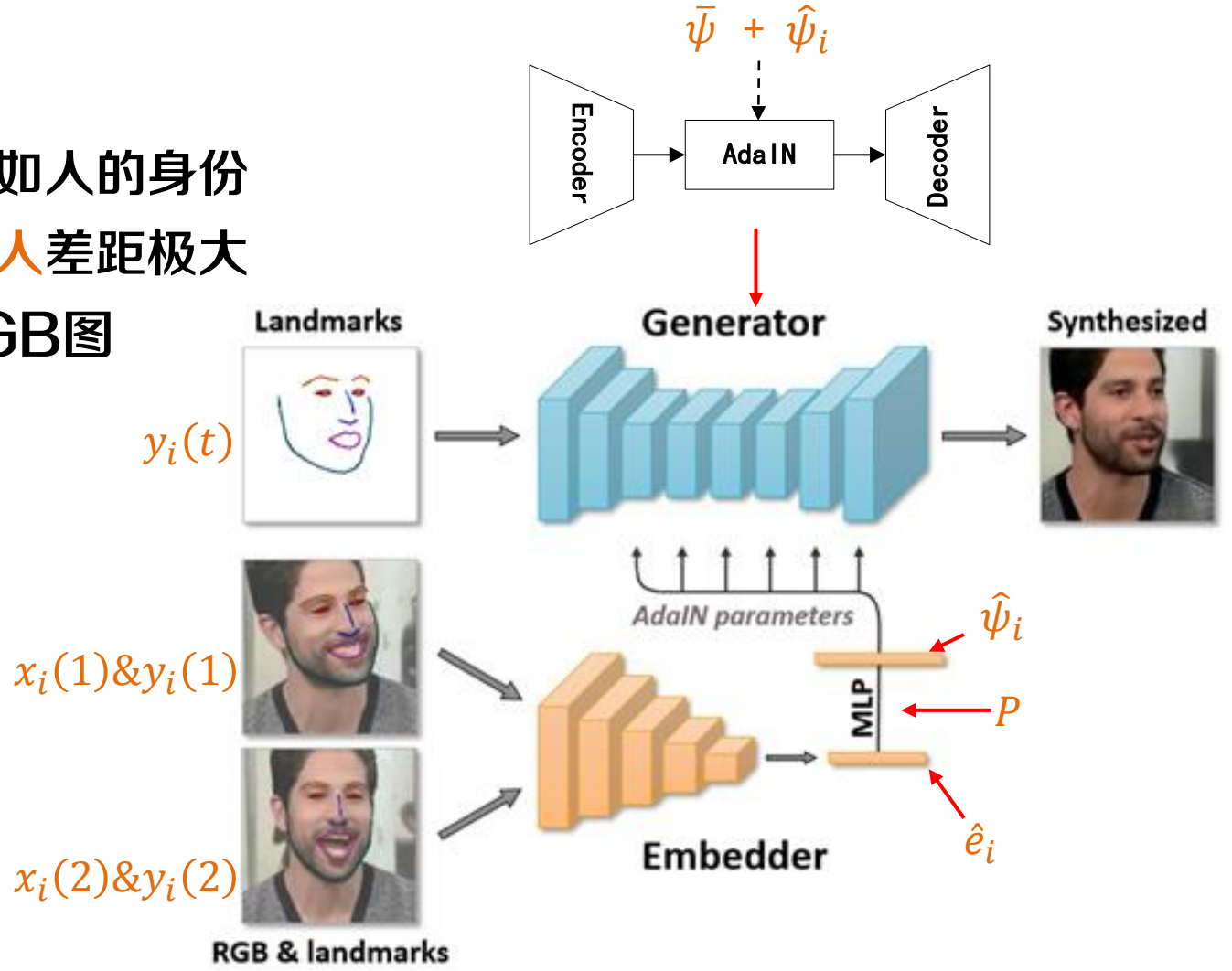
算法原理 Few-Shot Adversarial Learning

- 元学习阶段
  - 通过学习大量样本，构建元学习模型，可泛化到相似域
  - 提取输入图片有效的域特征，并生成伪造图片
- 模型结构
  - 嵌入器、生成器、判别器



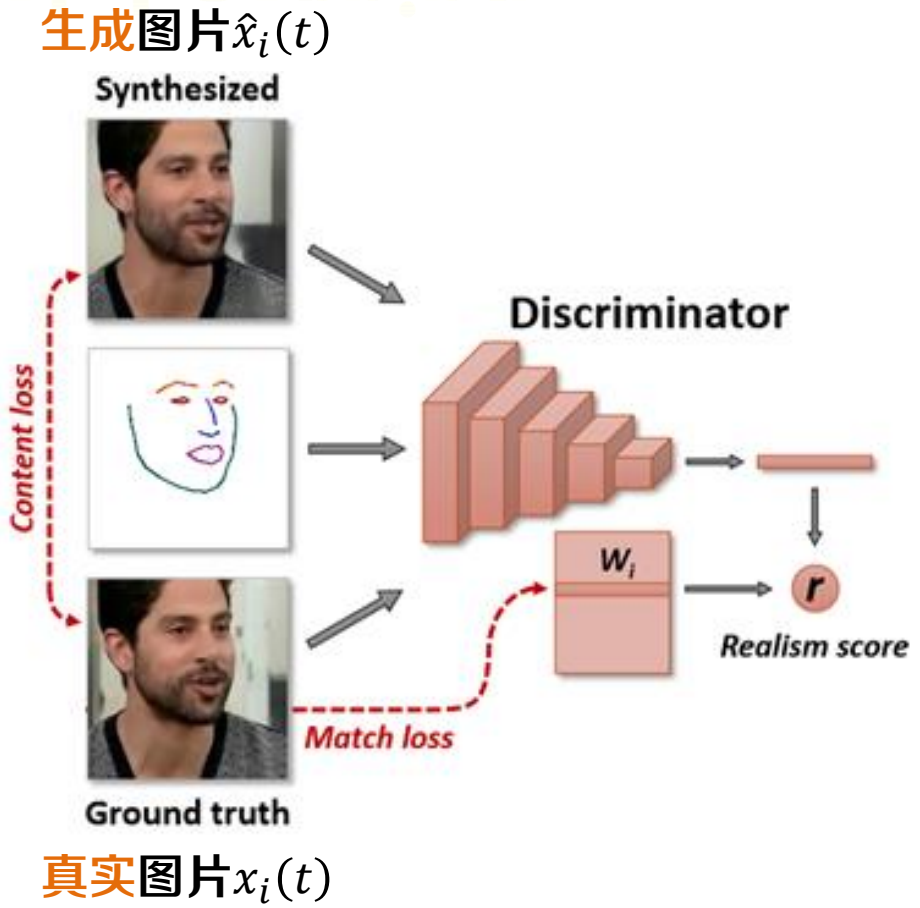
算法原理 Few-Shot Adversarial Learning

- 嵌入器（Embedder）
  - 提取域不变特征$\hat{e}_i(s)$
    - 一个视频（域）中不变的特征，如人的身份
    - 不同姿态、帧，差别极小；不同人差距极大
  - $E(x_i(s), y_i(s); \phi) = \hat{e}_i(s)$，$x_i$为RGB图
- 生成器（Generator）
  - 生成伪造人脸（自动编码器）
  - $G(y_i(t), \hat{e}_i; \psi, P) = \hat{x}_i(t)$伪造图像
  - AdaIN风格迁移：添加域特征
    - 指导生成器产生图像
    - P为投影矩阵，映射特征向量$\hat{e}_i$
    - 特征参数$\hat{\psi}_i = P\hat{e}_i$，通用参数$\bar{\psi}$



算法原理 Few-Shot Adversarial Learning

- 判别器（Discriminator）
  - 判别生成器的图像是否真实
    - 生成器：生成真实的图片骗过判别器
    - 判别器：高准确的识别生成的假图片
  - 真实性分数$r = D(\hat{x}_i(t), y_i(t), i; \theta, W, w_0, b)$
    - 真、假图相似度
    - 生成图和landmark拟合程度
  - W矩阵：记录训练集所有域的特征
    - 每一行W_i代表了一个人的域特征
  - W_i和$\hat{e}_i$区别：整体和局部
    - 某个人整体的域特征
    - 某个输入batch下的域特征



结构复杂、参数众多，如何训练?

算法原理 元学习阶段

- 嵌入器&生成器损失函数

评价生成器　　判别器进行打分　　评价嵌入器

$$\mathcal{L}(\phi,\psi,P,\theta,W,w_0,b)=\mathcal{L}_{CNT}(\phi,\psi,P)+\mathcal{L}_{ADV}(\phi,\psi,P,\theta,W,w_0,b)+\mathcal{L}_{MCH}(\phi,W)$$

$$-D(\hat{x}_i(t),y_i(t),i;\theta,W,w_0,b)+\mathcal{L}_{FM}$$　稳定项

判别器针对生成图像打分，越真实越高

$$V(\hat{x}_i(t),y_i(t);\theta)^T(W_i+w_0)+b$$

  - 评价损失
    - $\mathcal{L}_{CNT}$：$x_i(t)$与$\hat{x}_i(t)$相似度
    - $\mathcal{L}_{MCH}$：$\hat{e}_i$和W_i的相似度
  - 损失函数越小，生成能力越强
  - 训练目标：生成器重构图像真实，判别器不能判别，准确提取域特征
- 判别器损失函数
  - $\mathcal{L}_{DSC}(\phi,\psi,P,\theta,W,w_0,b)=max\left(0,1+D\left(\hat{x}_i(t)\right)\right)+max\left(0,1-D\left(x_i(t)\right)\right)$
  - 损失函数越小，检测能力越强
  - 训练目标：精准区分生成的伪造图片和真实图片

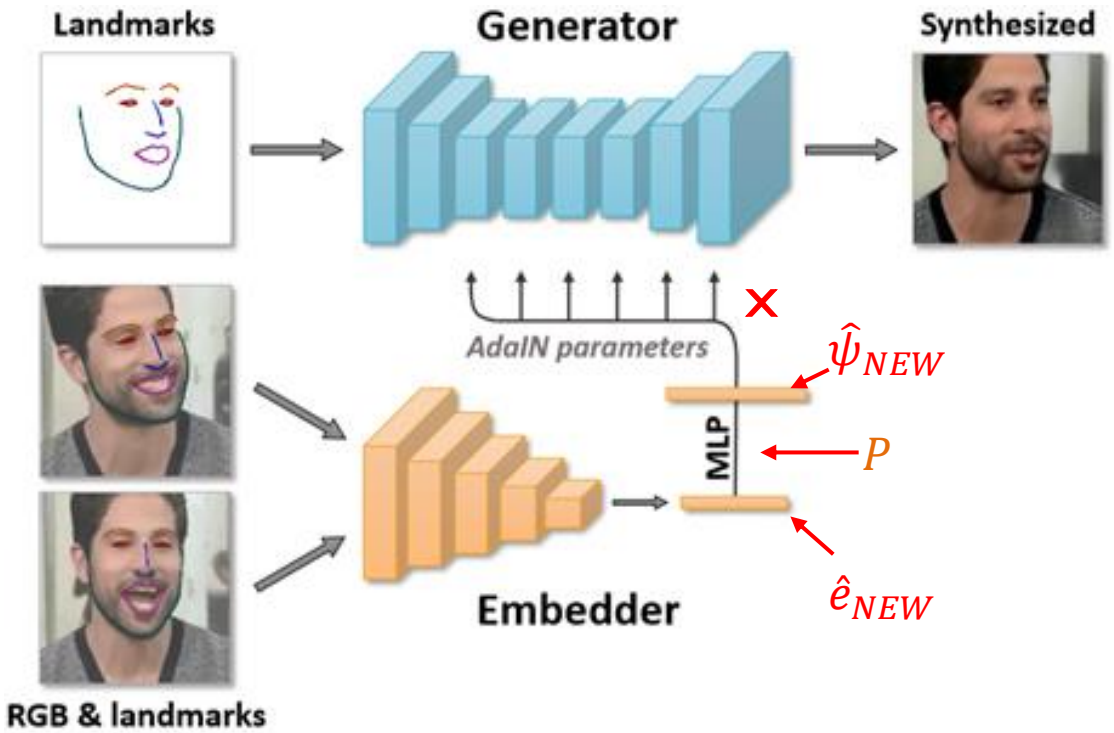
交替训练，对抗优化

生成更逼真的图像

算法原理 初步测试阶段

- 收敛元学习模型状态
  - 嵌入器：根据少量的输入，可以很好的提取目标域特征
  - 生成器：根据给定目标landmark和图片域特征，生成目标人脸逼真图像
- 初步测试阶段：目标T张图片
  - $\hat{e}_{NEW} = \frac{1}{T}\sum_{t=1}^{T} E(x(t), y(t); \phi)$
  - 直接将$\hat{e}_{NEW}$输入生成器
- 域分布差异影响结果
  - 人物、风格不同导致域特征提取欠佳
  - 嵌入器域特征$\hat{e}_{NEW}$提取偏差
  - 导致$\hat{\psi}_{NEW}$不精确，影响图像生成

不同域产生误差影响图像质量



算法原理 小样本在线学习

- 生成器
  - 针对测试域特征，微调模型参数
  - 原本$\hat{x}_i(t) = G(y_i(t), \hat{e}_i; \psi, \mathrm{P})$生成图像，特征参数$\hat{\psi}_i = P\hat{e}_i$
  - P投影矩阵：联系嵌入器和生成器的桥梁
    - 元学习阶段直接参与训练
    - 小样本在线学习阶段仅初始化$\psi' = \hat{\psi}_{NEW} = P\hat{e}_{NEW}$，$\psi'$直接参与训练
  - 现在$\hat{x}_{NEW}(t) = G'(y(t); \psi, \psi')$代替，接近真实值训练收敛快
- 检测器
  - 原本分数：$V(\hat{x}_i(t), y_i(t); \theta)^T(W_i + w_0) + b$
  - 由于$\hat{e}_{NEW} \approx W_{NEW}$，初始化$w' = \hat{e}_{NEW} + w_0$，现在分数：$V(\hat{x}(t), y(t); \theta)^T w' + b$
- 嵌入器：不参与训练

元学习参数指导在线学习初始化参数，加快模型收敛

实验效果 Few-Shot Adversarial Learning

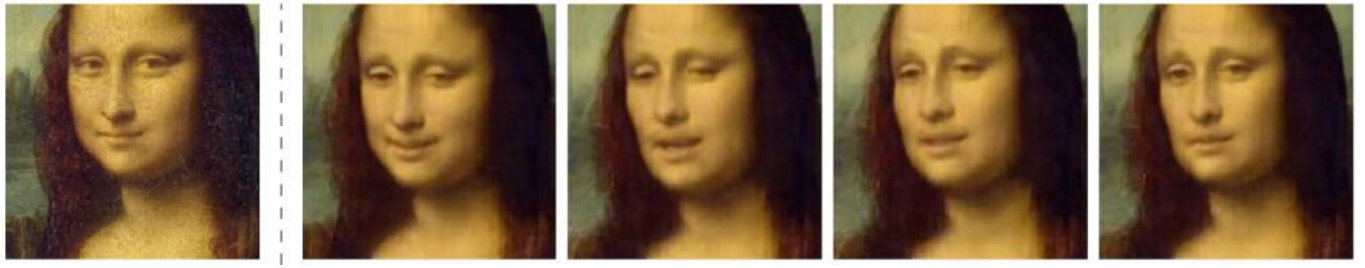
- 实验数据集
  - VoxCeleb1：1fps，256p
  - VoxCeleb2：25fps，10倍数据量
- 参数指标
  - FID：生成图像与真实图像特征向量距离
  - CSIM：编码之间余弦相似度
  - USER：真人成功分辨概率
    - 用户评价，2张真图，1张生成图
    - 极限为0.33
  - FF-无微调和$\mathcal{L}_{MCH}$，FT-正常流程

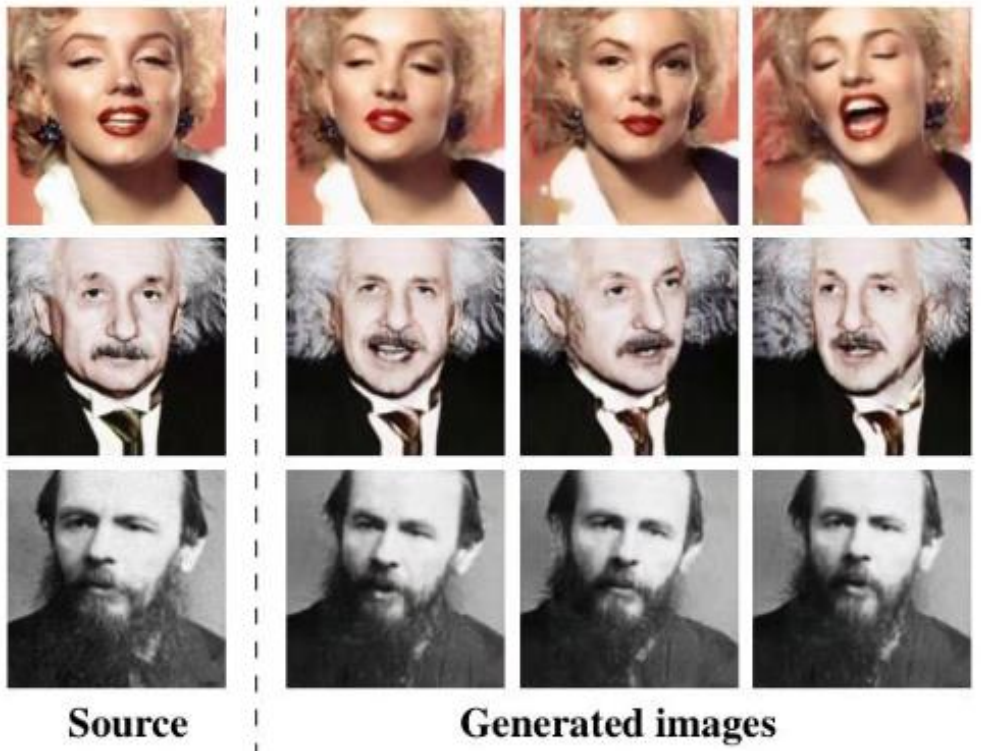
Method (T)	FID↓	SSIM↑	CSIM↑	USER↓
VoxCeleb1				
X2Face (1)	45.8	**0.68**	**0.16**	0.82
Pix2pixHD (1)	**42.7**	0.56	0.09	0.82
Ours (1)	43.0	0.67	0.15	**0.62**
X2Face (8)	51.5	**0.73**	**0.17**	0.83
Pix2pixHD (8)	**35.1**	0.64	0.12	0.79
Ours (8)	38.0	0.71	**0.17**	**0.62**
X2Face (32)	56.5	**0.75**	0.18	0.85
Pix2pixHD (32)	**24.0**	0.70	0.16	0.71
Ours (32)	29.5	0.74	**0.19**	**0.61**
VoxCeleb2				
Ours-FF (1)	**46.1**	0.61	**0.42**	**0.43**
Ours-FT (1)	48.5	**0.64**	0.35	0.46
Ours-FF (8)	42.2	0.64	**0.47**	0.40
Ours-FT (8)	**42.2**	**0.68**	0.42	**0.39**
Ours-FF (32)	40.4	0.65	**0.48**	0.38
Ours-FT (32)	**30.6**	**0.72**	0.45	**0.33**

面向人类的伪造图像生成方法

实验效果 Few-Shot Adversarial Learning

- 实验效果
  - 目标样本数量越多，效果越好
  - 仅1张照片训练，同样效果极好





总结分析 Few-Shot Adversarial Learning

- 算法总结
  - 元学习训练阶段，构建嵌入器提取域特征，训练生成器和判别器参数
  - 应用阶段，通过元学习中的参数辅助初始化，微调生成器、判别器参数
  - P投影矩阵、$\mathcal{L}_{MCH}$嵌入器损失函数，确保模型可以快速收敛
  - 仅需少量样本进行快速在线学习完成域迁移，生成逼真伪造图像
- 算法优势
  - 泛化能力强，应用时无需重新训练，速度快、图像质量高，需要样本少
- 应用分析
  - 只需要一张照片，就能让他说任何话、做任何表情
- 能否做到检测?

越锋利的矛就会有越坚固的盾

视频深度伪造检测技术

背景介绍 Face Manipulation Detection

- 检测原理：伪造工具逐帧合成视频
  - 视频伪造工具不能增强帧之间的时间连贯性
  - 伪影、抖动、明暗不规则变化
- 伪影识别法(Artifact-Specific)
  - 连贯识别
    - CNN提取局部特征
    - RNN检测闪烁和抖动
    - LSTM检测面部区域光影
  - 取证识别：GAN留下的细微特征和样式
  - 生理识别：血液流动造成的面部像素变化
  - 融合识别、环境识别、行为识别、同步识别

(a) 双眼不一致性

(b) 反射缺失

(c) 牙齿细节

(d) 人脸拼接

(e) 光照估计和鼻子几何形状不精确

基于时域特征的检测

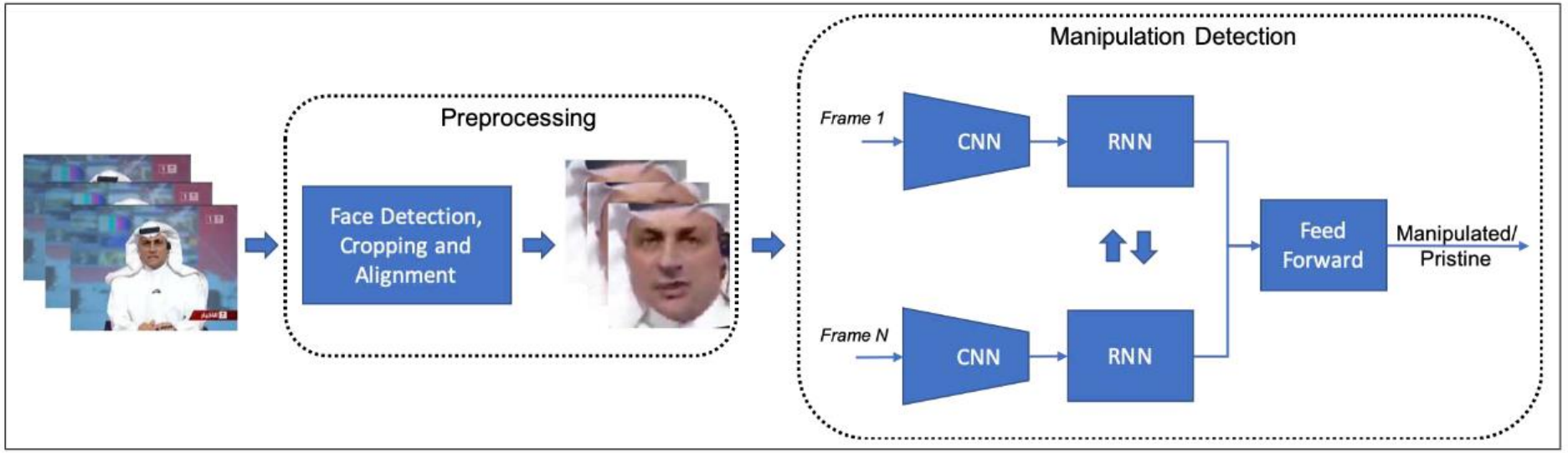
算法原理 Face Manipulation Detection

T	目标	检测目标视频是否经过伪造
I	输入	视频的某一帧或一段
P	处理	1、检测人脸信息，预处理生成蒙版、面部landmark标记 2、输入神经网络进行分析检测
O	输出	伪造/正常

P	问题	如何定位伪造视频时独有特征点并引入时序影响
C	条件	输入视频面部特征清晰，时序输入为相邻帧
D	难点	有效定位伪造特征点
L	水平	CVPR2019（计算机视觉顶会）

算法原理 Face Manipulation Detection

- 面部图像预处理
  - 检测帧是否包含面部信息，剪切、对齐面部图像
  - 生成面部landmark标记
- 循环卷积神经网络（RCNN）
  - RCNN = CNN + RNN
  - 既能像CNN一样提取局部特征，又能像RNN一样引入时序影响



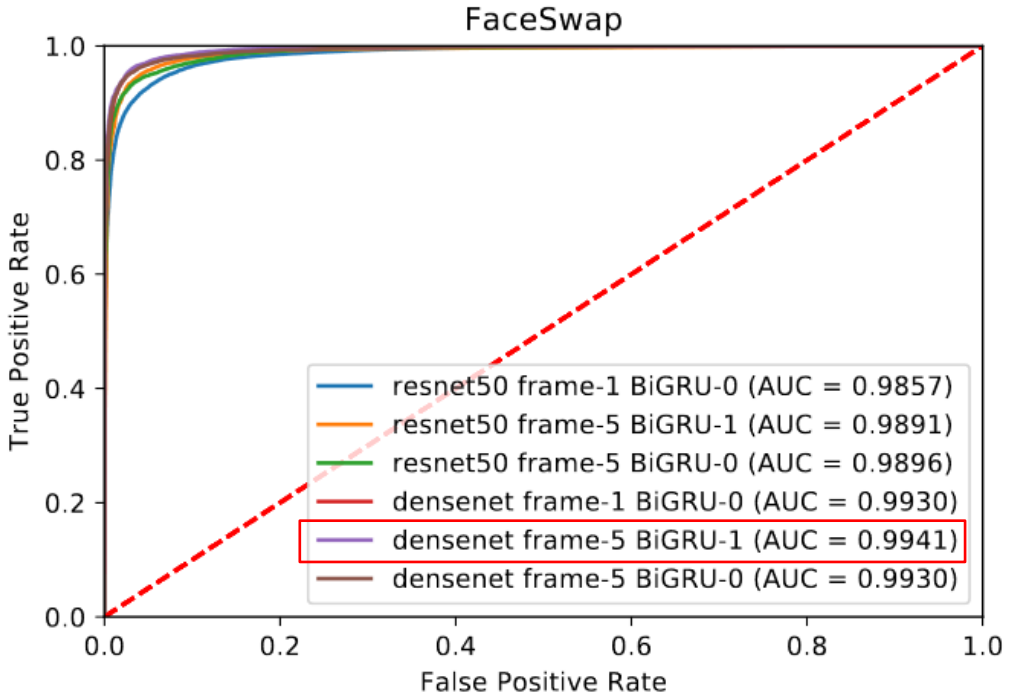
实验结果 Face Manipulation Detection

- 参数解释
  - Deepfake、Face2Face、FaceSwap：领域常用伪造算法
  - Frames：检测时输入帧数；FF++：伪造视频数据集&检测基准准确率
  - ResNet50、DenseNet：CNN；BiDir：RNN；Alignment：使用预处理
- 结论
  - 面部对齐等预处理方法提高性能、序列输入优于单帧输入、引入RNN提升准确率

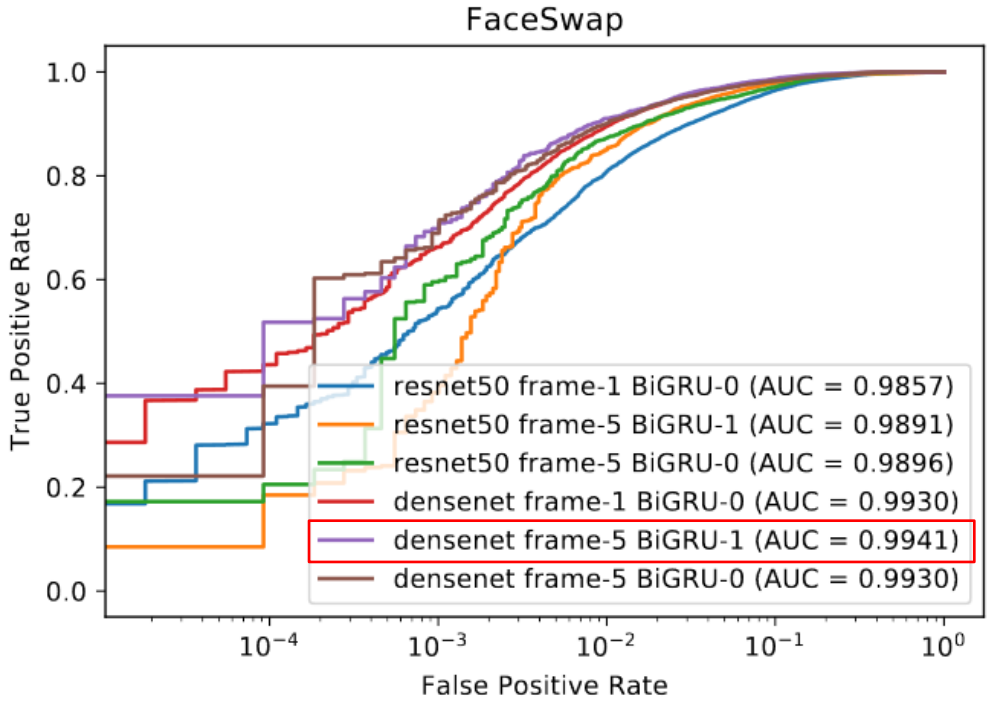
Manipulation	Frames	FF++ [34]	ResNet50	DenseNet	ResNet50 + Alignment	DenseNet + Alignment	ResNet50 + Alignment + BiDir	DenseNet + Alignment + BiDir
Deepfake	1	93.46	94.8	94.5	96.1	96.4	-	-
	5	-	94.6	94.7	96.0	96.7	94.9	**96.9**
Face2Face	1	89.8	90.25	90.65	89.31	87.18	-	-
	5	-	90.25	89.8	92.4	93.21	93.05	**94.35**
FaceSwap	1	92.72	91.34	91.04	93.85	96.1	-	-
	5	-	90.95	93.11	95.07	95.8	95.4	**96.3**

实验结果 Face Manipulation Detection

- ROC曲线
  - 在保证正确率的前提下，误报率在可接受范围内
  - DenseNet + BiDir + 预处理时，AUC = 0.9941



线性图



对数图

未来展望

- 视频伪造技术发展
  - 通过解耦和pix2pixHD网络组件改善面部质量和身份
  - 通过时间判别器和光流预测提升视频的流畅度和逼真度
  - 使用辅助网络以减轻边界伪影，将合成图像无缝融合
  - 在预训练的VGG人脸识别网络上使用感知损失
    - 已应用在流行的在线视频伪造工具中
- 视频伪造检测技术发展
  - 基于空间特征、时空融合特征、生物特征检测
  - 水印检测、区块链检测

矛和盾相互促进发展

参考文献

[1] Zakharov E, Shysheya A, Burkov E, et al. Few-shot adversarial learning of realistic neural talking head models[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9459-9468.

[2] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 694-711.

[3] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8798-8807.

[4] SABIR E, CHENG J X, JAISWAL A, et al. Recurrent convolutional strategies for face manipulation detection in videos [C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, Jun16 -20,2019 . Piscataway: IEEE, 2019:80 -87.

道德经

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。

谢谢！