西安电子科技大学
XIDIAN UNIVERSITY

基于视频图像的人脸替换技术研究

肖　清

目录

CONTENTS

01/研究意义

02/人脸图像预处理

03/人脸替换算法设计与实现

04/表情替换算法设计与实现

05/假脸检测方案及结果

06/总结与展望

研究意义

1.1研究意义

影视制作

利用人脸替换技术将演员的脸换成主角的脸，既确保了明星的人生安全，又不影响作品的效果。

虚拟体验

通过表情迁移（替换）技术将用户的面部表情呈现在VR世界中，将会增加虚拟现实体验的个性化。

活动营销

适用于广告代理、咨询、媒体、学校等企业和机构，在 H5 或其他传播形式中加入人脸融合的玩法来进行产品推广、活动宣传。

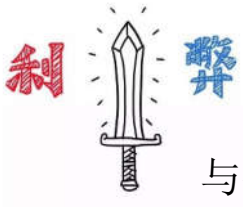
形象打造

适用于游戏、影视制作等公司根据游戏、影视形象做融合模板，打造形象周边。

大众娱乐

适用于用户的个性化形象定制，通过自拍或上传照片，可将用户人脸融合在不同形象中，打造独特的个人形象。

1.1研究意义



与任何变革性技术一样，人脸替换技术是一把双刃剑，在为人们日常生活提供乐趣的同时也带来了新的挑战。

虚假新闻　恶搞视频　金融诈骗

眼见不再为实，导致信任危机

损害公众名誉，侵犯个人隐私

扰乱社会治安，引起舆论混乱

因此，针对各种假脸视频图像的检测工作刻不容缓。

1.2研究现状

- 基于传统二维图像处理的人脸替换
  - 实现简单，无需复杂计算
  - 严重依赖于人脸数据库
  - 对夸张表情、姿态的替换效果较差
- 基于三维建模的人脸替换
  - 比较直观，能够克服大的姿态差异
  - 建模过程复杂，计算量大，效率低
  - 图像真实感差
- 基于深度学习的人脸替换
  - 生成的人脸图像真实自然
  - 自动化程度高
  - 模型具有通用性
  - 需要大量、丰富、完备的数据进行网络训练

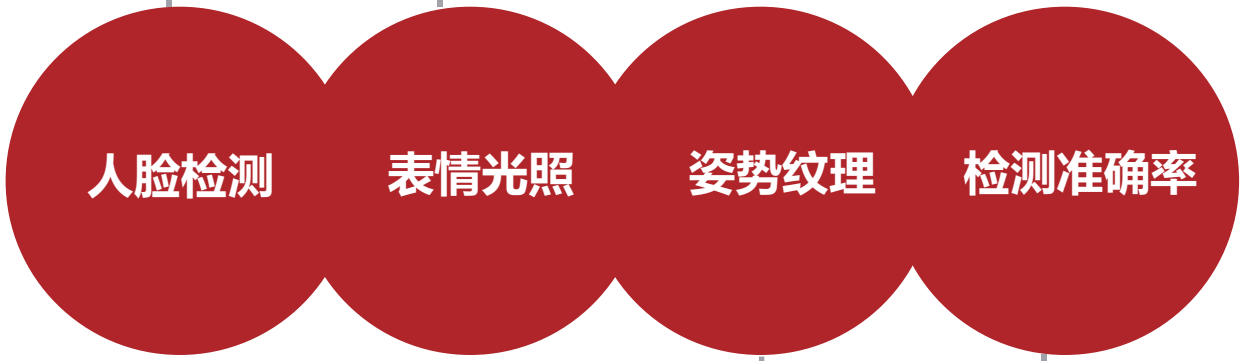
1.2研究现状

- 基于传统图像取证的检测技术：在像素级别构建深度伪造图像检测模型，但可通过在伪造图像中加噪声的方式绕过。
- 基于CNN架构定制化修改的检测技术：通过修改CNN架构和损失函数等方式监督学习实现检测，但容易受到对抗样本的攻击。
- 基于真伪图像特征差异比对的检测技术：通过分析提取真伪图像的差异化特征，进而训练分类器实现深度伪造图像的检测。
- 基于真伪图像频谱差异分析的检测技术：通过真伪图像频谱的差异化分析，实现深度伪造图像的检测。

1.3现存主要问题

人脸替换结果的好坏很大程度上取决于人脸检测器的精度和准确性。

替换结果中人脸表情、视线方向、光照等目标属性未完全忠于目标图像。



当源和目标人脸头部姿势与面部表情差异较大时，替换结果发生扭曲变形。提取纹理信息需要耗费大量的资源和时间。

假脸检测方法的准确率和检测速度均有待提高。

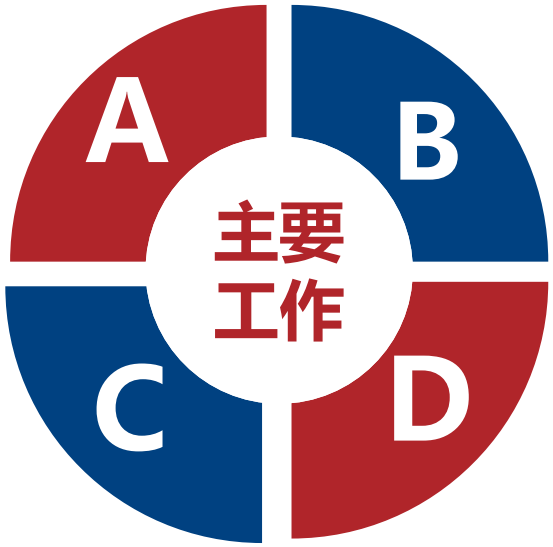
1.4主要工作

人脸图像预处理

对待替换的人脸图像进行预处理，其中包括人脸检测、定位以及特征提取。

人脸表情替换

利用图像动画网络框架，基于一阶运动模型实现实时人脸表情替换。



人脸替换

针对现存人脸替换存在的问题，设计算法准确全面提取目标图像属性，得到自然逼真的人脸替换视频图像。

假脸图像检测

通过学习大量真伪人脸图像训练网络模型，实现对深度伪造假脸图像的准确检测。

人脸图像预处理

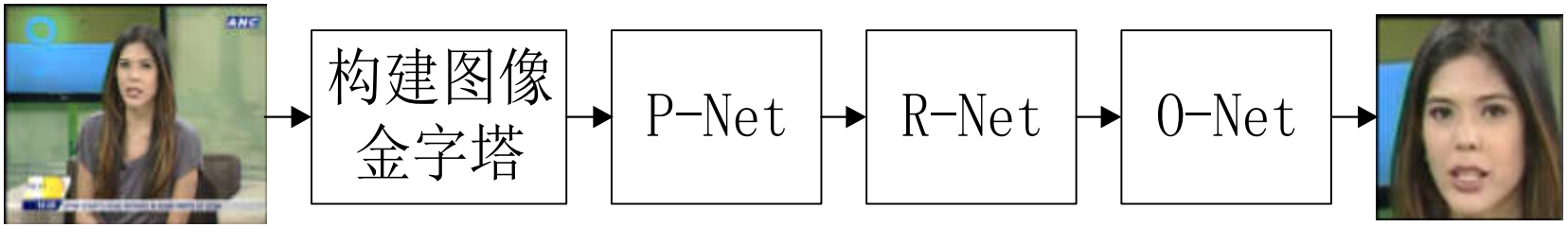
2.1人脸检测及定位

本文使用目前应用较广的多任务卷积神经网络(Multi Task Convolutional Neural Network, MTCNN）对人脸图像进行检测和定位。

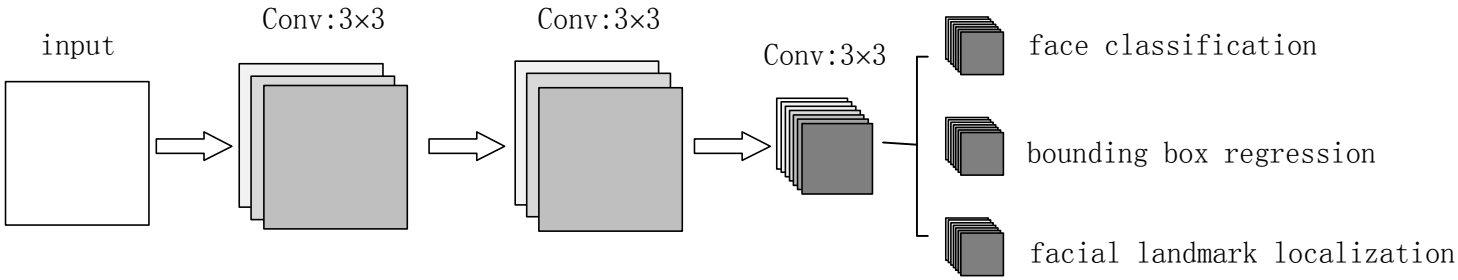
1.先使用小模型生成目标区域候选框

2.再使用更复杂的模型进行细分类和更高精度的区域框回归

3.递归执行第2步实现快速高效的人脸检测及定位



2.1人脸检测及定位



P-Net网络结构图

快速生成候选窗口的P-Net	全卷积网络，用于人脸区域建议。
高精度候选窗口过滤选择的R-Net	在P-Net基础上增加一个全连接层（128节点）。
生成边界框与定位关键点的O-Net	在R-Net基础上增加一个卷积层且全连接层节点数增加为256。

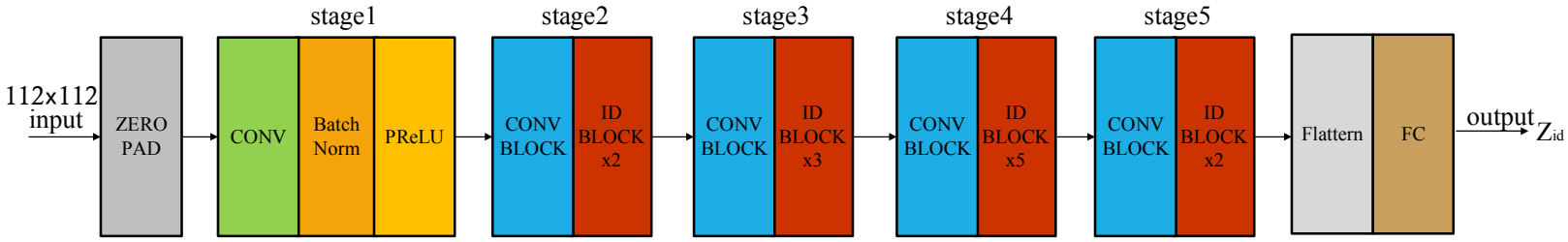
2.1人脸检测及定位

检测结果示例：

2.1人脸检测及定位

检测结果示例：

2.2提取人脸特征



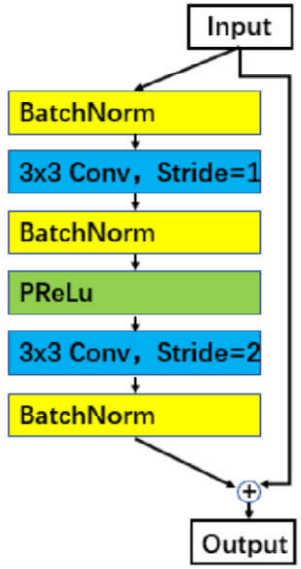
Arcface特征编码器结构图

基于Resnet-50模型进行调整：

将预训练模型中第一个7×7、步长为2的卷积层替换为3×3、步长为1的卷积层。
激活函数改为pReLU。

取消最大池化层和最后的平均池化层。

使用增强残差网络结构，使网络模型更适合于对人脸进行特征提取和识别。



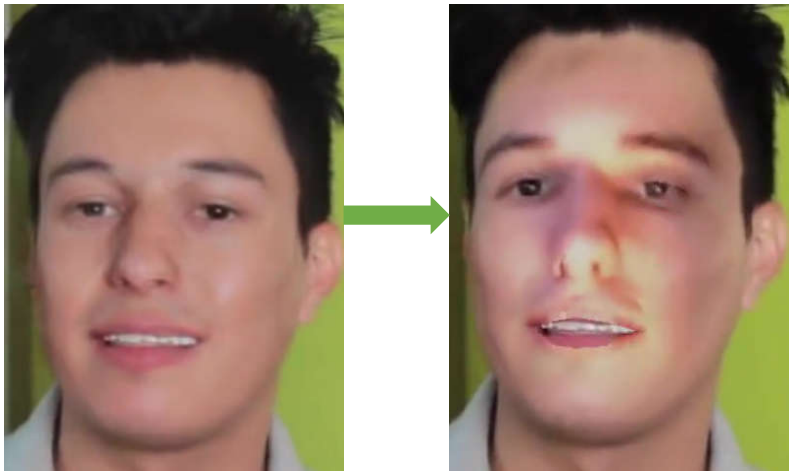
增强残差单元

人脸替换

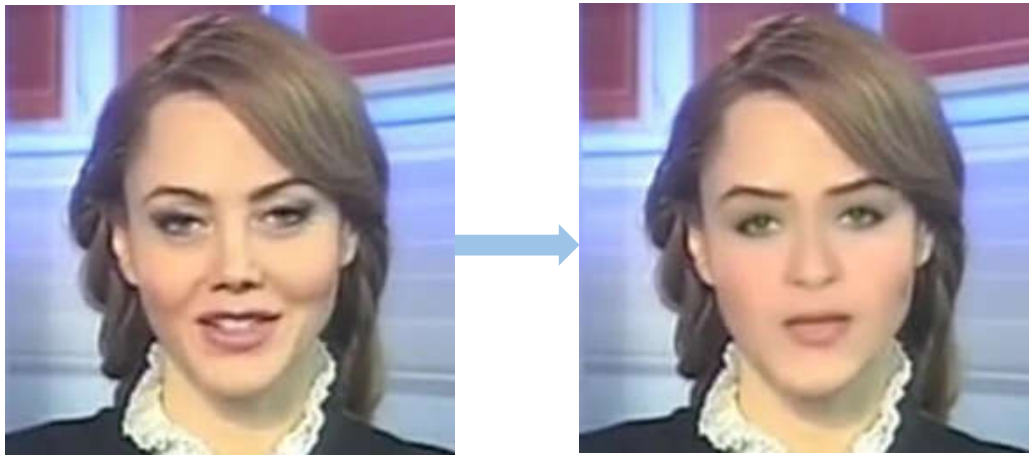
3.1多级属性编码器

目前许多方法生成的人脸替换图像常常忽略光照、背景等诸多目标图像属性（左图），而且生成人脸替换图像的表情、姿势与目标人脸的吻合度较低（右图）。因此本文基于U-Net网络结构设计了一种U型多级属性编码器，全面准确地提取目标图像属性。

目标人脸图像　　替换人脸图像



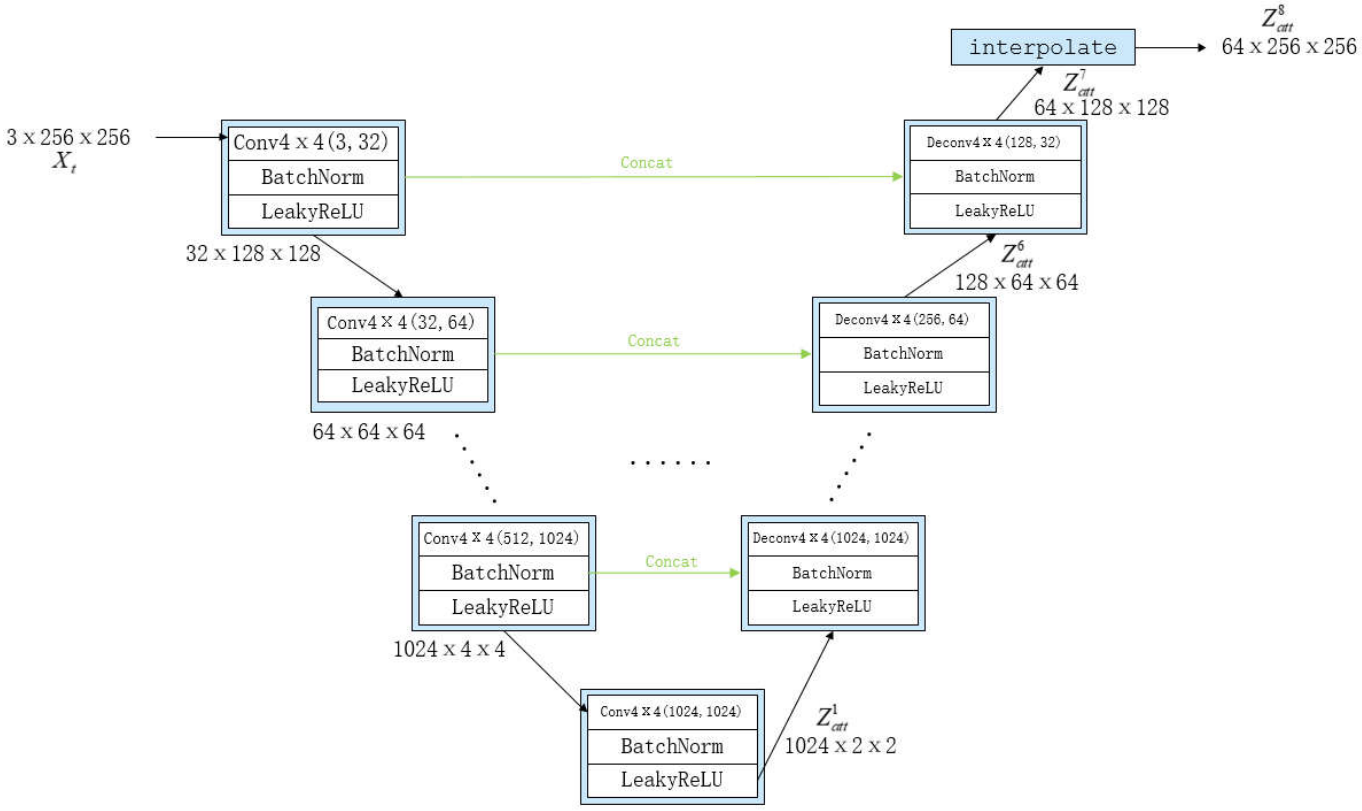
目标人脸图像　　替换人脸图像



3.1多级属性编码器

多级属性编码器结构：

- 左侧压缩路径由7个级联的卷积层和BN层组成
- 右侧扩展路径由6个反卷积层和BN层组成
- 中间通过拼接的方式，将压缩阶段获得的特征图同扩展阶段获得的特征图拼接起来
- 卷积层和反卷积层均使用4×4的卷积核，步长为2，填充为1。



多级属性编码器结构图

激活函数采用ReLU的变体LeakyReLU：

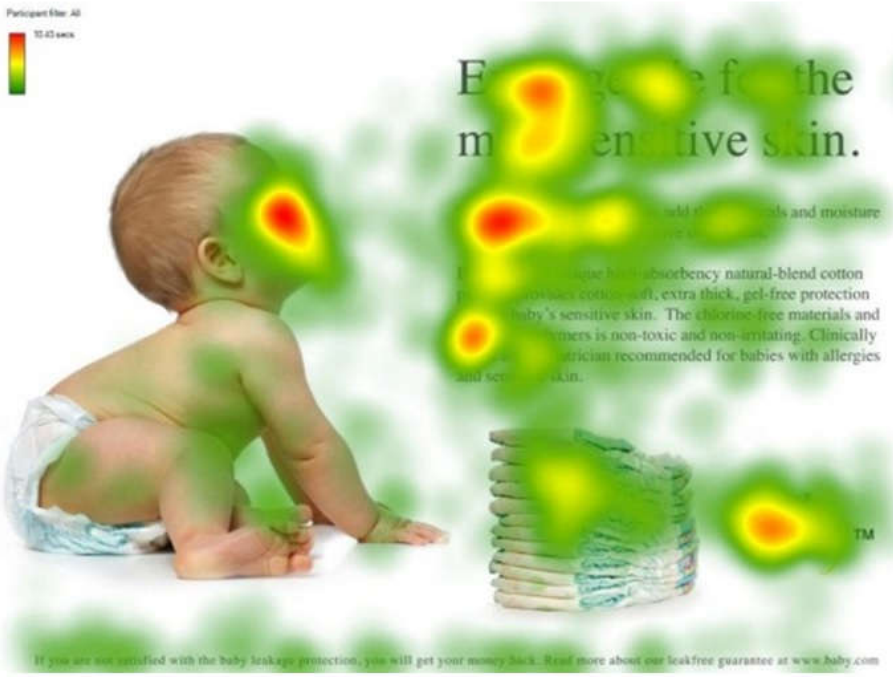
$$y=\begin{cases} x & if\ x\geq 0 \\ ax & if\ x<0\ (0<a<1) \end{cases}$$

最终得到多级目标属性 $Z_{att}=\{Z^1_{att}(X_t),Z^2_{att}(X_t),\dots\dots Z^8_{att}(X_t)\}$ 。

3.2结合注意力机制的生成器网络

视觉注意力：

- 人类特有的一种大脑信号处理机制，通过快速扫描整张图像，获得需要重点关注的目标区域，然后对该区域投入更多注意力资源，同时忽略周围的无用信息。



深度学习中的注意力机制：

- 通俗讲是一种记忆选择机制，能选择出对模型最重要的信息并加以学习，与人类的视觉注意力本质类似，核心目标也是从复杂众多的信息中选择出对当前任务更关键的信息。

注意力机制的基本原理：

- 网络中每层不同的通道或不同位置其特征的重要性各不相同，后面的层应该更注重其中有用的重要信息，抑制不重要的信息。

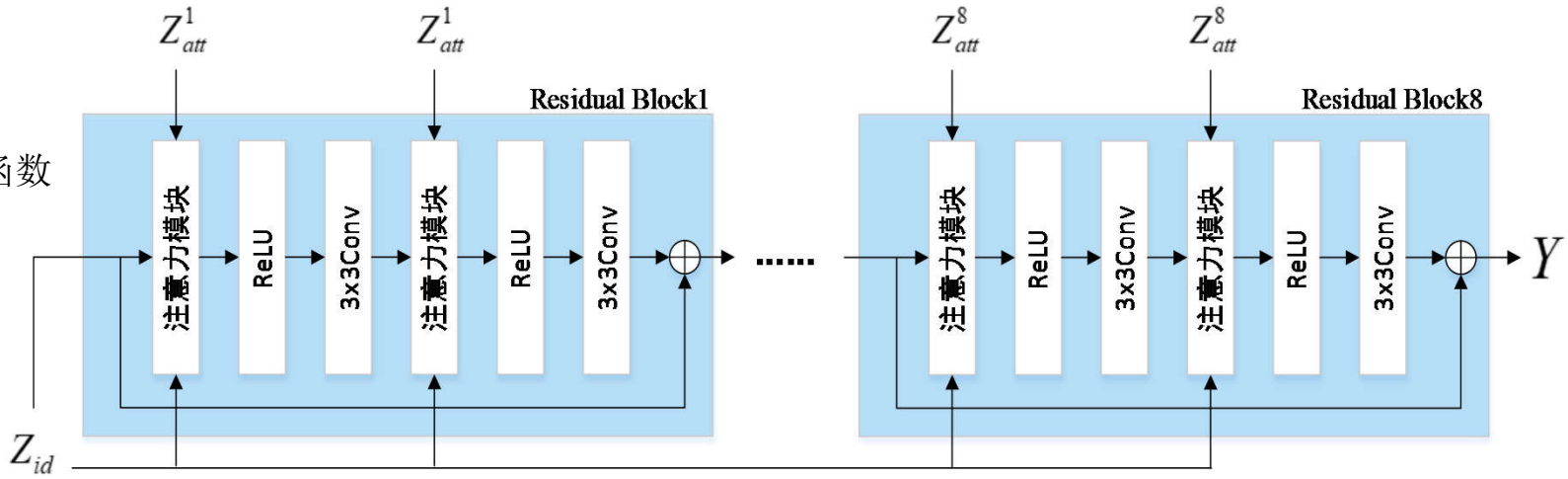
注意力机制的本质：

- 对输入的源图像中的每个Key分量分别计算出其对应Value的注意力权重系数，然后加权求和。

$$Attention(Query, Source) = \sum_{i=1}^{L_x} Similarity(Query, Key_i) * Value_i$$

3.2结合注意力机制的生成器网络

- 包含8个残差块
- 每个残差块由注意力模块、激活函数和卷积层三部分组成
- 激活函数采用ReLU函数
- 卷积核大小为3×3，步长为1



生成器网络结构图

损失函数

对抗损失函数：$L_{GAN} = -\log D(Y)$

目标属性保存损失：$L_{att} = \frac{1}{2}\sum_{k=1}^{n} \| Z_{att}^k(Y) - Z_{att}^k(X_t) \|_2^2$

相对ID距离损失：$L_{id} = \cos^{-1}\frac{Z_{id}(X_s)Z_{id}(Y)}{\| Z_{id}(X_s)\|_2 \| Z_{id}(Y)\|_2} + \frac{\cos^{-1}\frac{Z_{id}(X_s)Z_{id}(Y)}{\| Z_{id}(X_s)\|_2 \| Z_{id}(Y)\|_2}}{\cos^{-1}\frac{Z_{id}(X_t)Z_{id}(Y)}{\| Z_{id}(X_t)\|_2 \| Z_{id}(Y)\|_2}}$

$$L_G = \alpha L_{GAN} + \beta L_{att} + \gamma L_{id}$$

3.2结合注意力机制的生成器网络

1. 归一化第k层激活向量 h_i^k

$$\overline{h}_i^k = \frac{h_i^k - \mu^k}{\sigma^k}$$

2. 得到注意力模型权重 M^k

3. 将源人脸身份特征集成到激活向量中

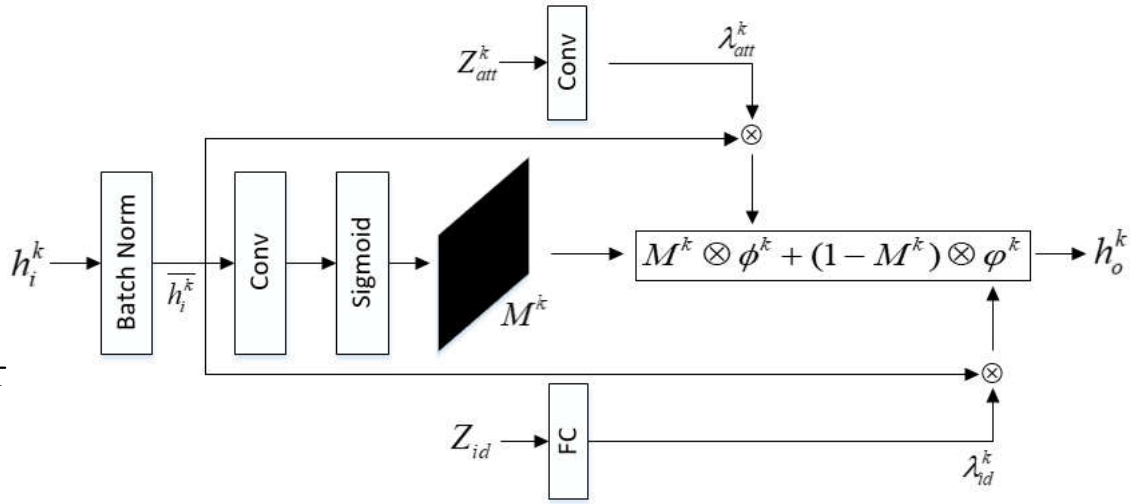
$$\varphi^k = \lambda_{id}^k \otimes \overline{h_i^k}$$

4. 将目标图像属性集成到激活向量中

$$\phi^k = \lambda_{att}^k \otimes \overline{h_i^k}$$

5. 将两个激活向量加权组合得到最终的输出

$$h_o^k = M^k \otimes \phi^k + \left(1 - M^k\right) \otimes \varphi^k$$



注意力模块结构图

3.3实验环境及流程

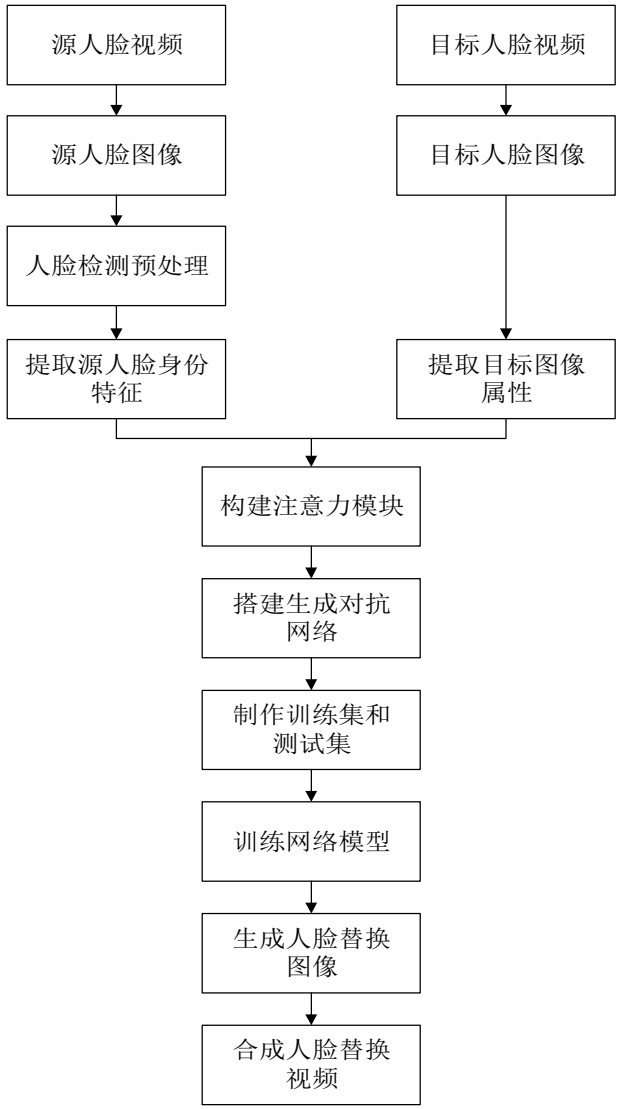
- 实验环境

CPU ： Intel(R) Core(TM) i7-9700八核处理器
GPU ： NVIDIA GeForce RTX 2080
运行内存：32G
开发平台：PyCharm
深度学习框架：PyTorch
编程语言：Python

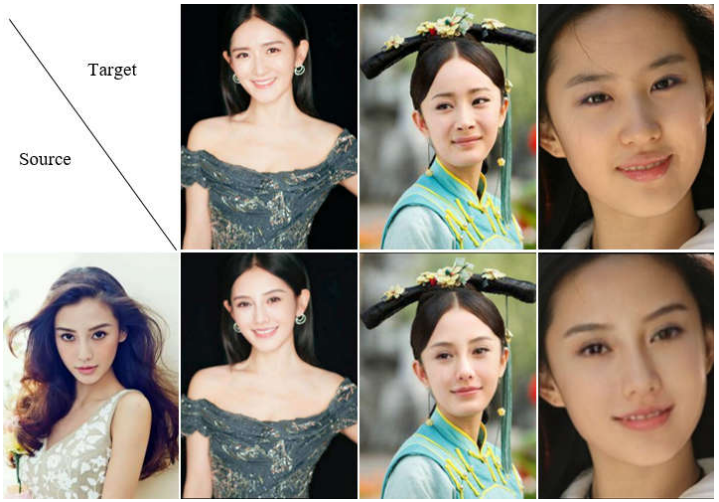
训练数据集：

CelebA-HQ数据集
FFHQ数据集
FaceForensic++数据集

训练50个epoch结果图



3.4人脸替换实验结果

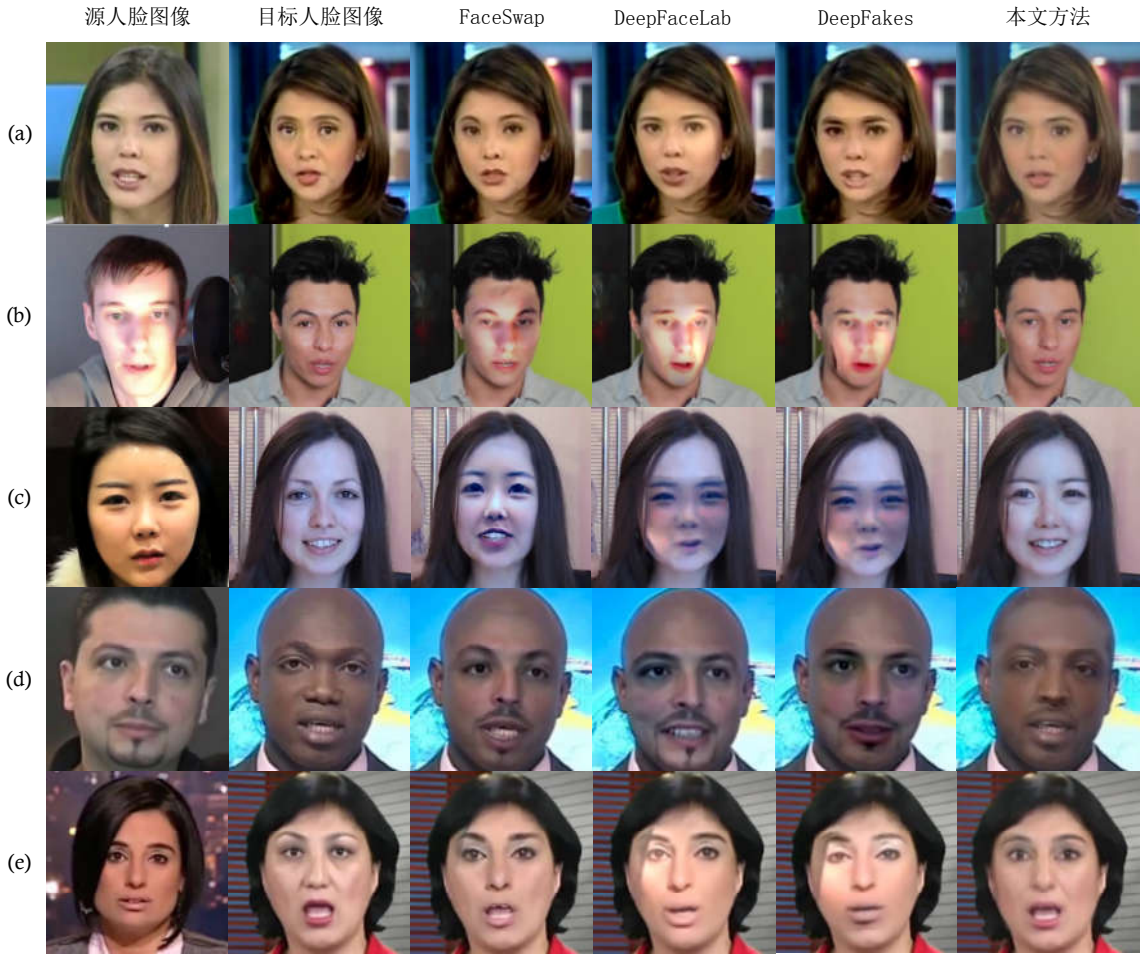


替换人脸面部表情、嘴角微笑程度、图像亮度以及分辨率等都与目标图像完全吻合，甚至连人物视线方向都高度一致。

克服了性别的局限性，能够将男性的面部替换到女性目标图像上，且未出现怪异或不适合的不良感受。

替换结果与源人脸高度相似，较好地集成了源人脸身份特征，在目标人脸面部有遮挡物时依然能够完成换脸，且没有产生明显缺陷。

3.4人脸替换实验结果



分析：

FaceSwap：替换结果中眼睛和嘴唇轮廓较为生硬不自然。

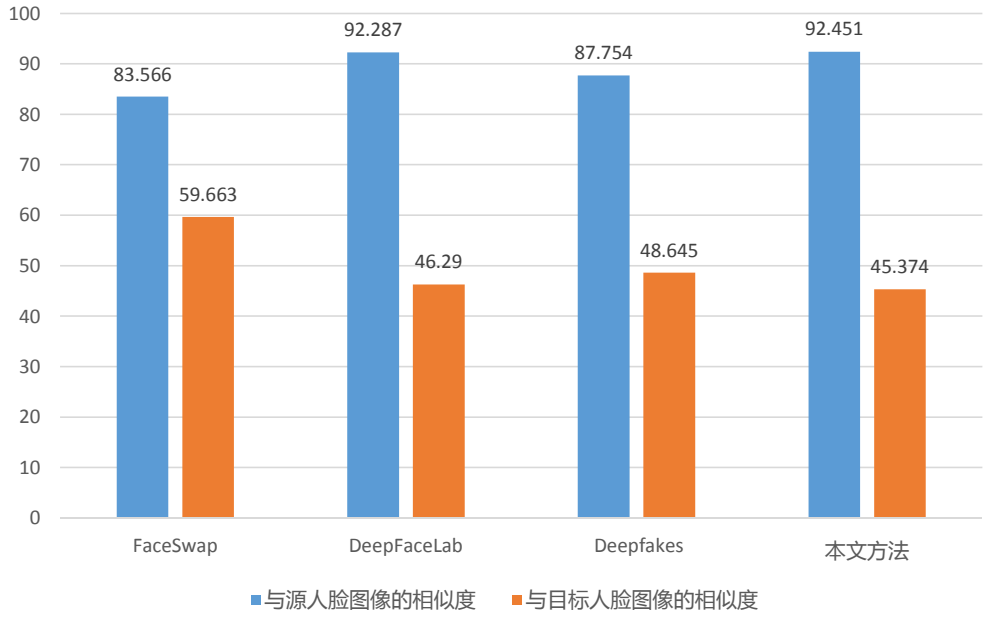
DeepFaceLab：先生成源面部区域图像，再嵌入到目标人脸中，会出现比较明显的嵌入痕迹。

DeepFakes：具有明显的嵌入痕迹，表情与目标人脸差异较大。

本文方法：人脸替换图像真实自然，表情到位且与目标图像高度一致，光线属性也完全忠于目标图像。

3.4人脸替换实验结果

客观评价



FaceSwap和DeepFakes方法的换脸结果准确率较低，而DeepFaceLab和本文方法与源人脸的相似度均高于92%，最大程度地保留了源人脸特征信息。

表1.1 人脸替换客观评价结果

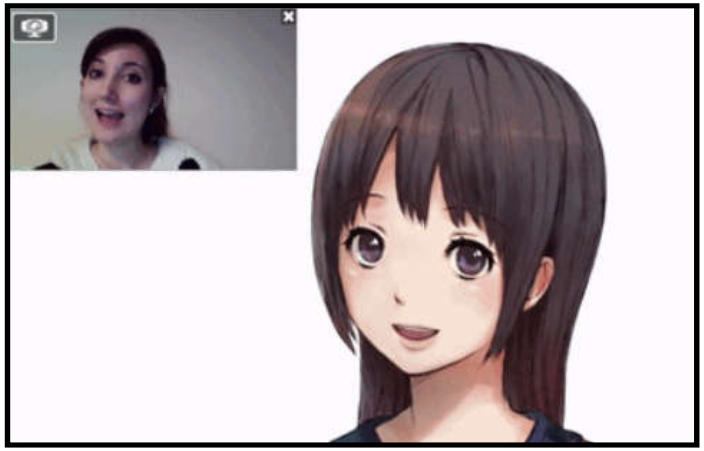
方法	结构相似性	头部姿势	面部表情
FaceSwap	0.7063	3.046	2.232
DeepFaceLab	0.7618	3.961	2.517
DeepFakes	0.7872	4.219	2.997
本文的方法	**0.8291**	**2.935**	**2.026**

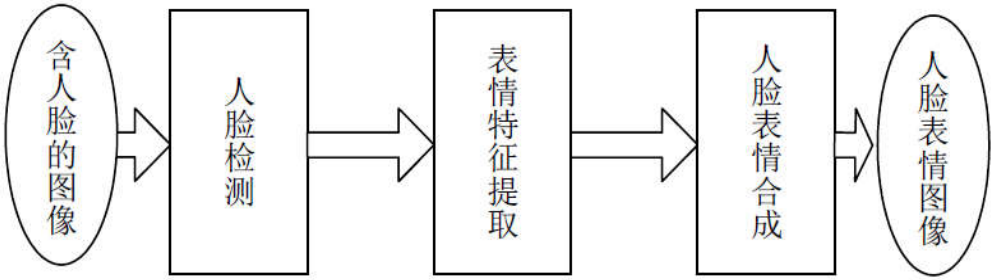
本文方法生成的人脸替换图像在结构相似性、头部姿势差异和人脸表情差异方面均具有明显优势。

表情替换

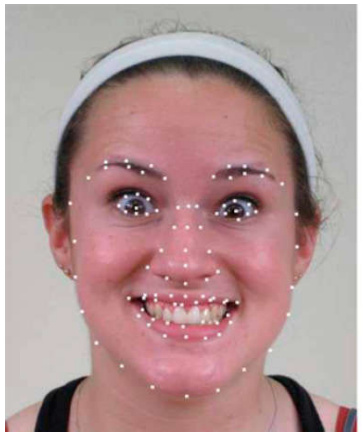
4.1表情替换介绍

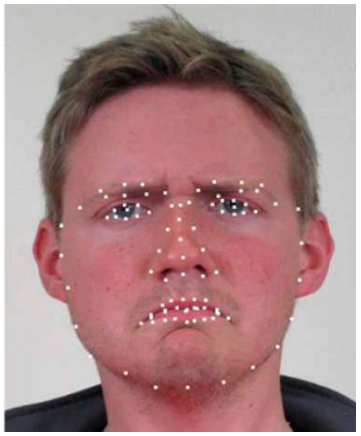
人脸表情替换，即表情迁移，是指通过先进的技术将源人脸视频图像中的表情替换为目标驱动视频或图像中的人脸表情。





人脸表情替换系统





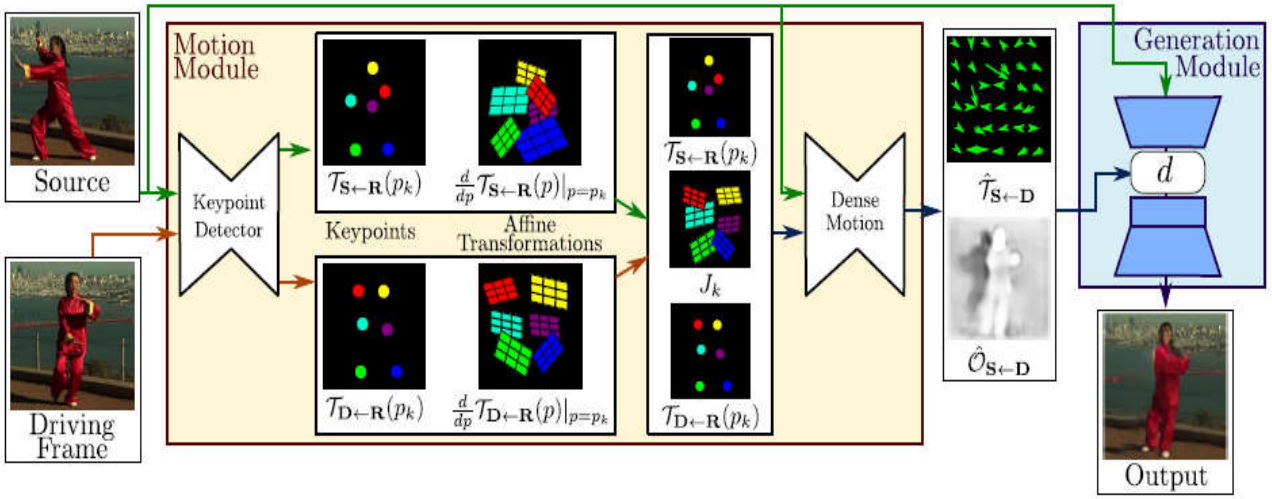
研究难点

人脸具有非常复杂的层次化纹理结构

同一个个体的面部表情千变万化

不同个体的面部特征和表情各不相同

4.2表情替换算法



图像动画的一阶运动模型

运动估计模块：估计目标驱动D到源图像S的密集运动场 $T_{S\leftarrow D}$ 并输出遮挡掩模 $\hat{O}_{S\leftarrow D}$

图像生成模块：利用生成器网络和运动估计模块的输出扭曲源图像并修复被遮挡的部分

4.2表情替换算法

分别定位源和目标人脸的面部关键点

↓

在局部范围内通过仿射变化对人脸表情建模

↓

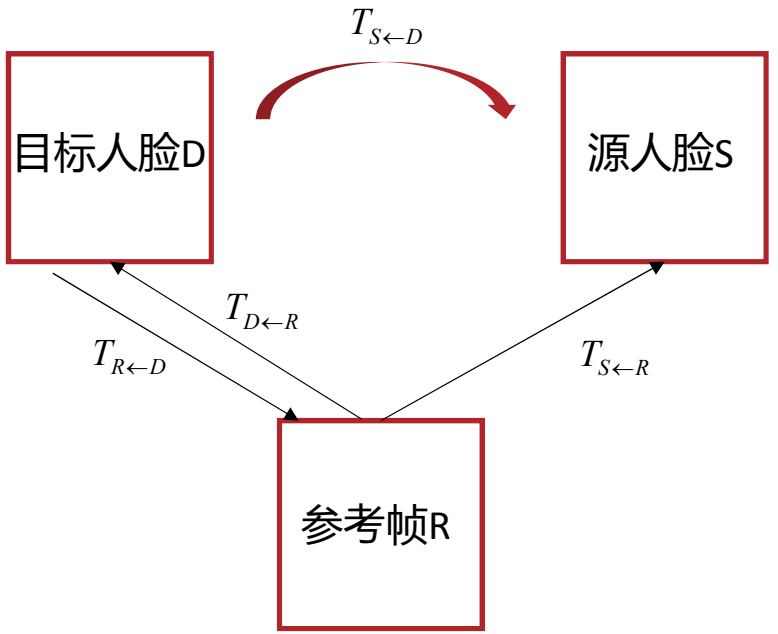
密集运动网络的输出经过卷积层得到遮挡掩模

↓

生成器根据目标D产生的遮挡对源S进行自动估计

↓

解码器解码得到人脸表情替换图像



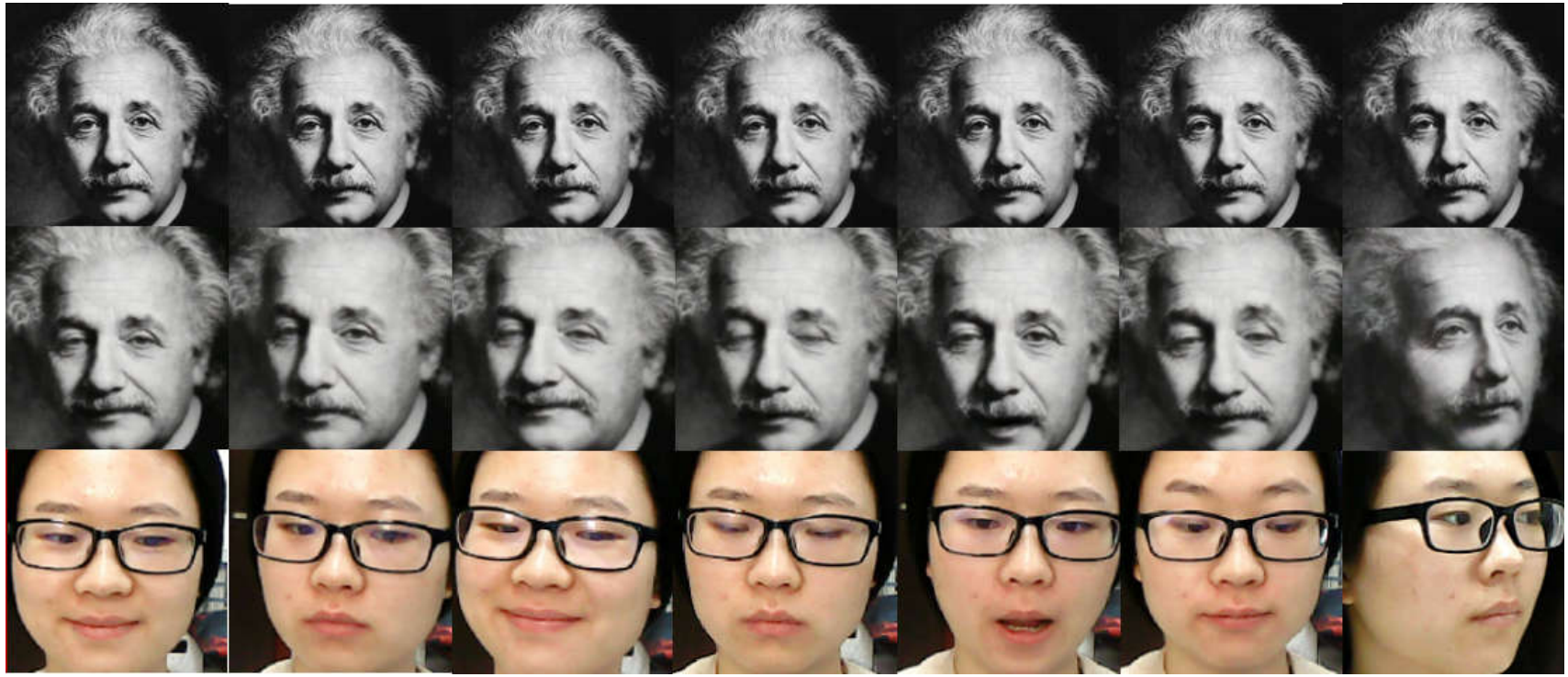
$$T_{S\leftarrow D} = T_{R\leftarrow D} \circ T_{S\leftarrow R} = T^{-1}{}_{D\leftarrow R} \circ T_{S\leftarrow R}$$

4.3表情替换实验结果

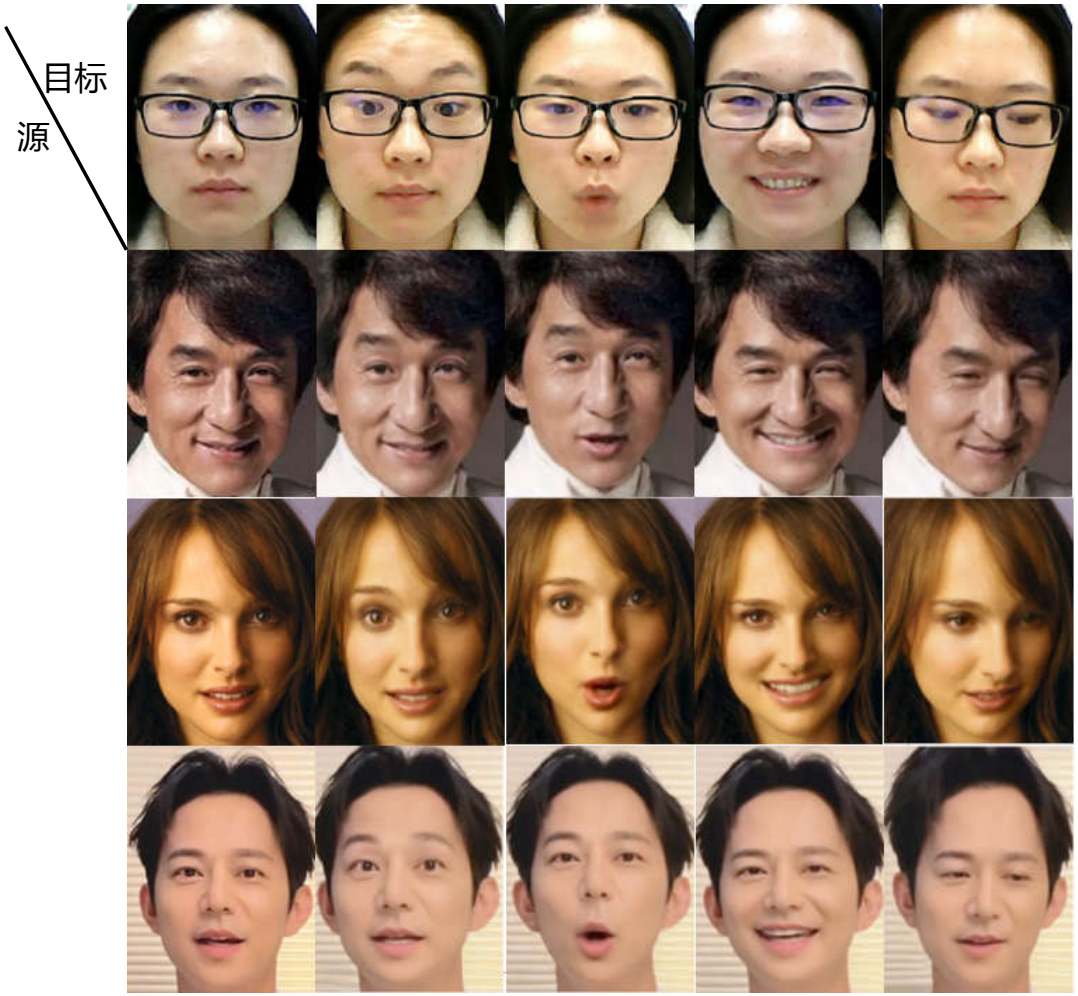
源人脸图像

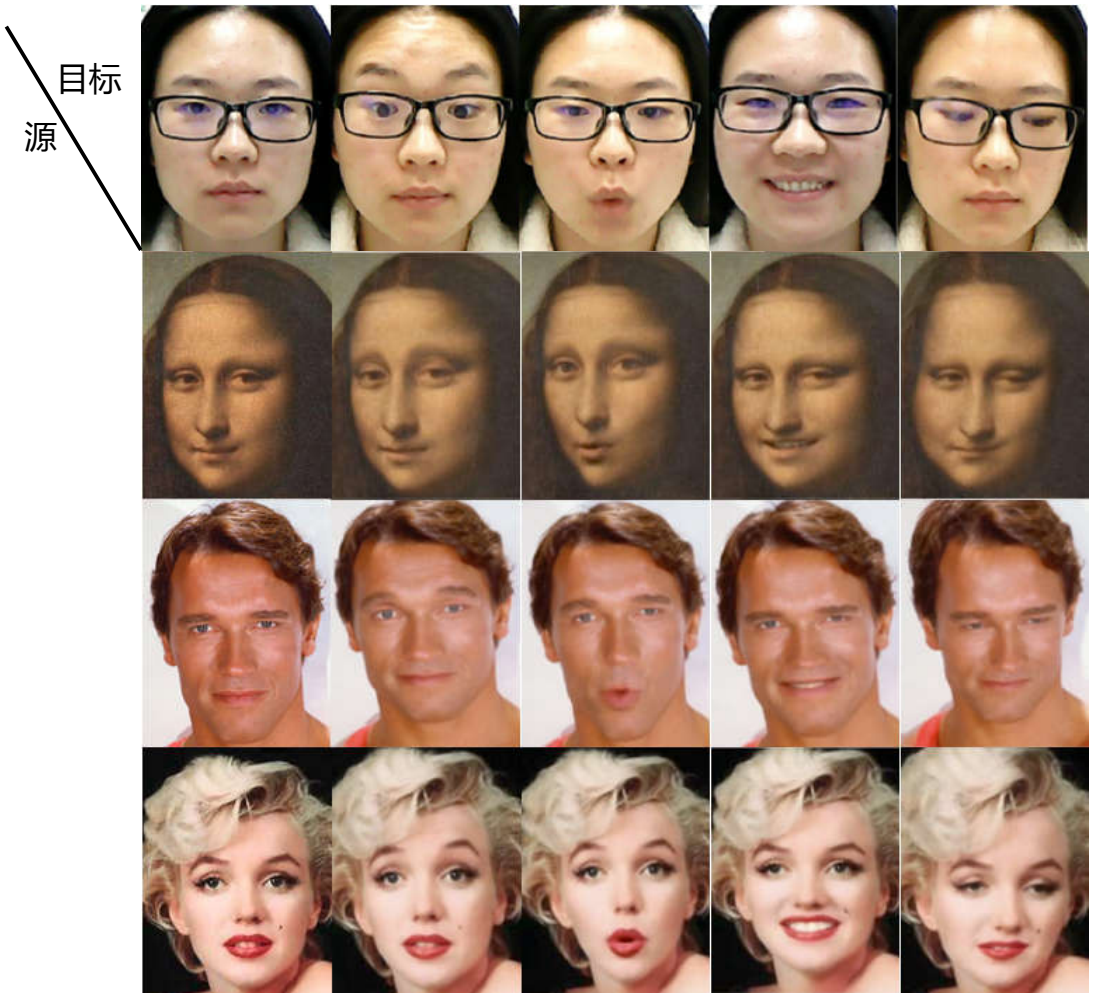
表情替换结果

目标人脸图像



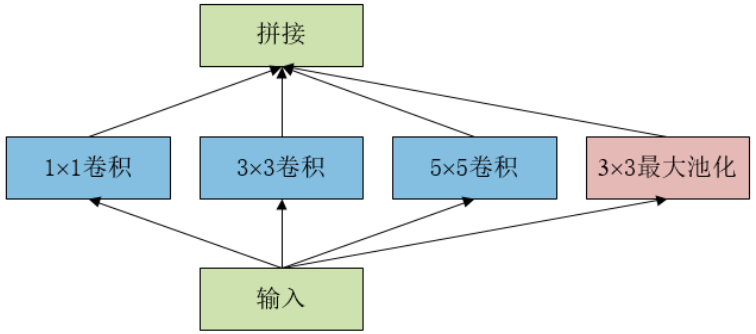
4.3表情替换实验结果



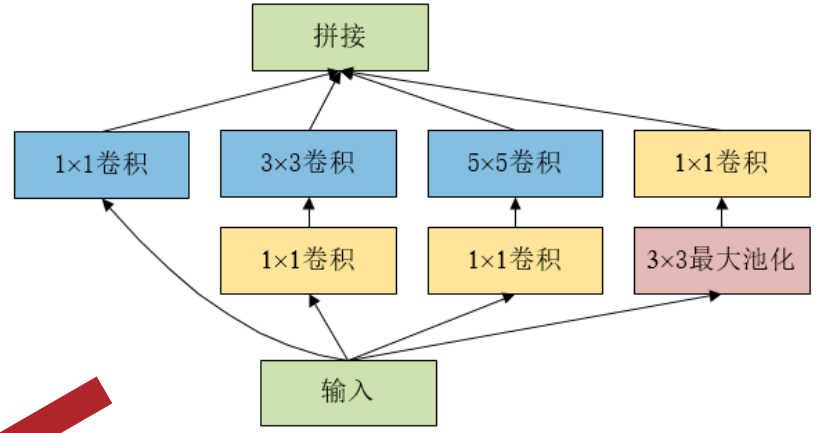


假脸检测

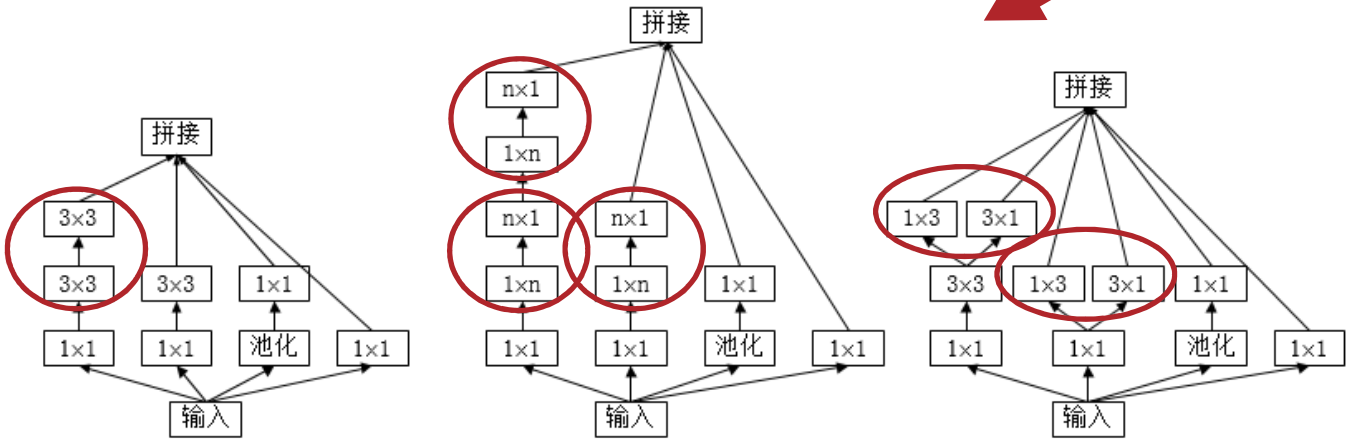
5.1深度可分离卷积



原始的Inception模块

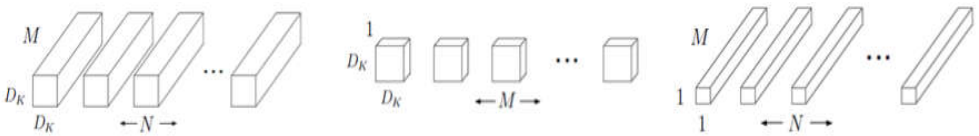


Inception v1模块



Inception v2模块

5.2假脸检测算法



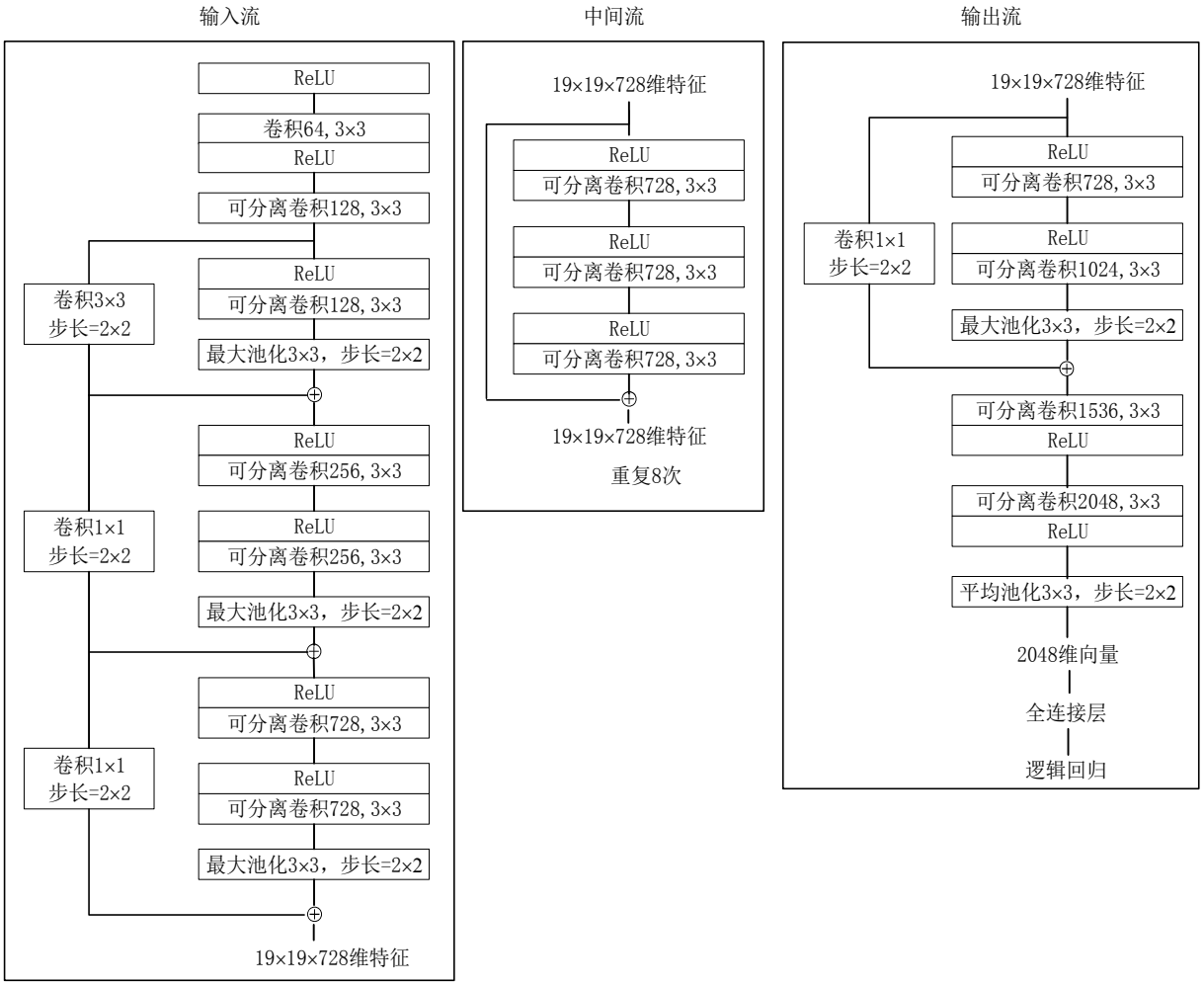
(a)标准卷积　(b)逐通道卷积　(c)通道融合卷积

深度可分离卷积

实验表明，深度可分离卷积结构在减少参数量的情况下增加模型的层数，既减少了存储空间，又增强了模型的表达能力。

Xception网络结构

主要是在Inception v3的基础上引入了深度可分离卷积，在基本不增加网络复杂度的前提下提高了模型的效果。每个小块之间采用残差连接。



Xception网络结构图

5.3数据集介绍

	Deepfake TIMIT 数据集	FaceForensics++ 数据集	FFW 数据集	DeepFake Detection 数据集
视频来源	相机拍摄	YouTube 网站	YouTube 网站	相机拍摄
分辨率	512×384	1280×720	1920×1080	1920×1080
真实视频数量	559	1000	0	363
伪造视频数量	640	4000	46	3000+
特点	视频中人脸区域占比大，背景较为单一	数据规模大，但合成质量最低	视频合成质量高，效果逼真，但缺乏真实视频样本	视频合成质量高，数据量大，但人数较少

真假脸视频

↓

分帧成真假脸图片

↓

按照6:2:2划分数据集并标注

↓

训练网络模型

↓

测试网络模型

	真脸数量/帧	假脸数量/帧
训练集	12079	35682
验证集	3905	12033
测试集	3821	11694

5.4假脸检测结果

库内测试结果

方法	样本数	准确率
FaceSwap	2365	98.05
Deepfakes	3509	**100**
Face2face	3455	96.91
NeuralTextures	2365	95.82
Youtube	3821	**100**

Xception网络拥有较强的学习能力，在库内测试中的表现非常好，准确率均超过95%，其中对于Deepfakes方法生成的伪造视频和Youtube上获取的真实视频的检测准确率接近100%。

准确检测出的假脸图像

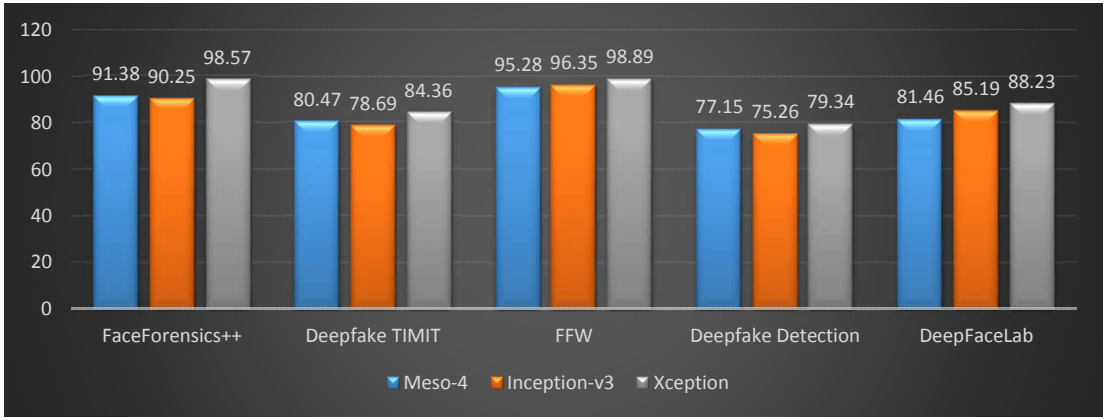
未检测出的假脸图像

5.4假脸检测结果

跨库测试结果

数据集	样本类别	样本数	检测准确率
Deepfake TIMIT数据集	Fake	2314	85.26
	Real	2158	82.37
FFW数据集	CGI	1420	100
	Face-Replacement	1348	100
	FakeApp	1342	94.44
	Part-of-face	1248	100
Deepfake Detection数据集	Fake	2936	79.69
	Real	1427	73.12
人脸身份替换	Fake	937	90.28
DeepFaceLab	Fake	3489	87.19
FFHQ数据集	Real	3000	86.73
Take by phone	Real	100	95

与库内测试相比，跨库测试得到的检测准确率有所降低，但从整体上看，该网络模型具有较好的泛化能力。



本文方法与假脸视频检测器Meso-4和Inception-v3在多个不同的数据集上进行测试比较，可以看出Xception表现最好，检测平均准确率约为92.51%，而Meso-4和Inception-v3表现相对比较接近，由此可以反映出Xception模块具有更强的特征提取能力。

总结展望

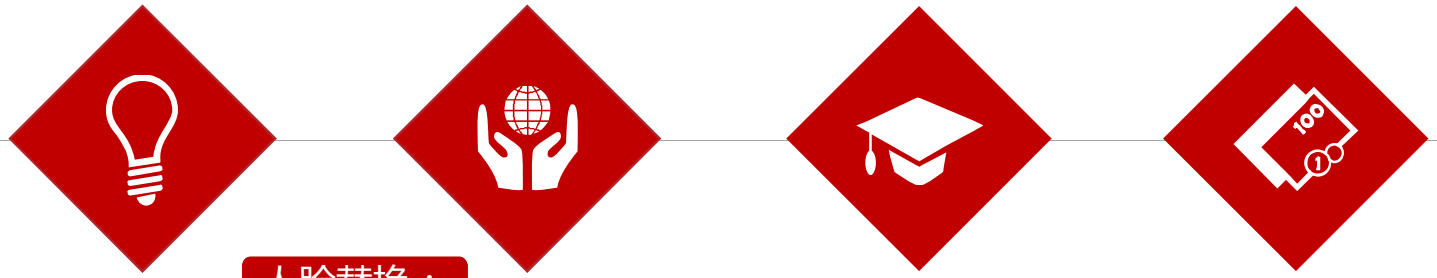
6.1研究总结

人脸图像预处理：

利用MTCNN网络对人脸图像进行检测及定位并提取面部特征。

表情替换：

利用图像动画网络框架，基于一阶运动模型，采用自监督的方式解耦身份外观和运动信息实现表情替换。



人脸替换：

设计了一种基于多级属性编码器和注意力机制的人脸替换算法，将注意力机制与生成对抗网络相结合，生成真实自然的人脸替换图像。

假脸检测：

提出了一种基于XceptionNet CNN架构的假脸检测器进行人脸反检测。

6.2工作展望

A 优化网络模型

本文提出的人脸替换算法生成的替换图像在脸型、肤色等方面与源人脸还有一定的差距，可以从优化网络模型角度克服生成脸型的不稳定性。

B 增加深度伪造音频

本文目前只针对视频中的人脸面部表情和简单的头部姿势进行替换，如果能把合成伪造的音频与口型等动作相结合，整段视频将会达到更加生动逼真的效果。

C 提高检测器鲁棒性

本文提出的假脸视频图像检测算法在大量训练数据集的基础上表现优异，但泛化能力还有较大的提升空间，应该深入研究鲁棒性更强的检测算法，使其对于无法大量获得或未知的假脸视频图像也能准确检测。

感谢聆听

THANK YOU