GAMES Webinar 232期

语音驱动的艺术肖像说话视频生成

报告人：易冉
上海交通大学
2022.06.09

Animating Portrait Line Drawings from a Single Face Photo and a Speech Signal

Ran Yi, Zipeng Ye, Ruoyu Fan, Yezhi Shu, Yong-Jin Liu, Yu-Kun Lai, Paul L. Rosin

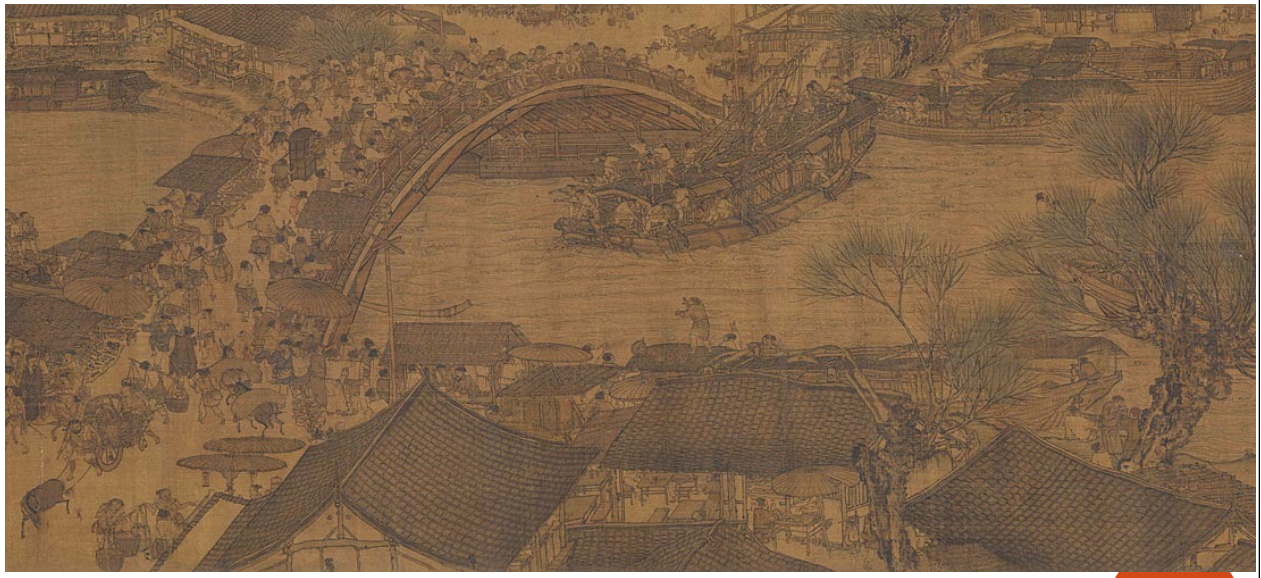
SIGGRAPH 2022 Conference Proceedings

目录

选题背景和意义

- 艺术创作
  - 人类对真实世界中感知的生活素材进行加工处理，进而表达情感的一种创造性的过程
  - 艺术作品来源于生活，亦丰富了人们的精神世界

梵高《星月夜》



张择端《清明上河图》

选题背景和意义

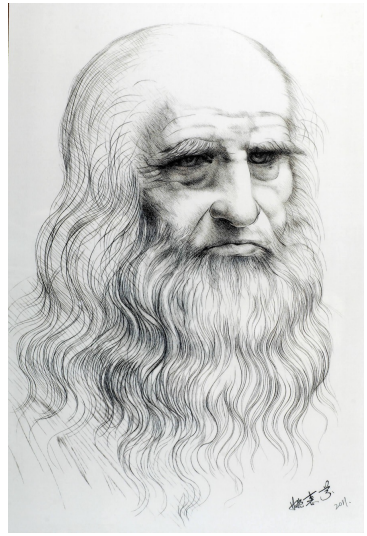
- 媒体艺术
  - 指结合计算机、数码技术等信息技术创作的艺术作品
  - 是科学与艺术的融合与交叉
  - 随着信息技术的发展，不断形成新的艺术形式：

信息技术	形成的新的艺术形式
便携式录像设备	录像艺术
电影特效技术	电影艺术
人机交互技术	交互艺术
人工智能技术	人工智能艺术
…	…

交互艺术 东京森大厦数字艺术馆《无界的世界》

选题背景和意义

- 肖像艺术和人脸图像的重要性



  - 肖像艺术是艺术领域的一个重要门类
    - 在摄影技术尚未成熟之前，肖像画是刻画和记载人物形象的主要方式

  - 人脸图像视频的分析与生成是计算机视觉和图形学中的一个重要课题
    - 人脸识别在安防监控等领域有重要应用
    - 人脸生成为数字媒体、虚拟现实等领域提供内容支撑

选题背景和意义

- 研究基于单张人脸照片的艺术动画制作：以单张人脸照片+语音信号作为输入，生成艺术肖像说话视频
- 在电影动画制作、虚拟现实、社交媒体等领域有广泛应用
- 与真实感的说话视频生成相比，具有更强的视觉效果，可以唤起人们不同的体验，实现新的交互和媒体应用

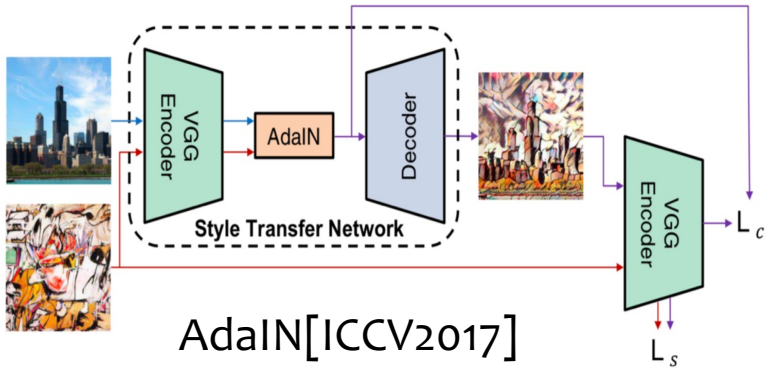
相关工作

1. 人脸图像艺术风格转换

- A. 基于卷积神经网络的艺术风格转换
- 目标风格由单张示例风格图片定义

1）基于图像优化的方法：

- Gatys, CNNMRF等
- 需要反复迭代，图像优化过程历时较长



AdaIN[ICCV2017]

2）基于模型优化的方法：

- 用前向神经网络代替图像优化的过程
- 单个模型只针对一个特定的风格的方法:[ECCV2016], [ICML2016]等
- 单个模型多个风格的方法：CIN, StyleBank等
- 单个模型可对任意风格转换的方法：AdaIN, WCT, AdaAttn等

- 缺点：从单张风格图像中定义风格困难、不精确

相关工作

1. 人脸图像艺术风格转换

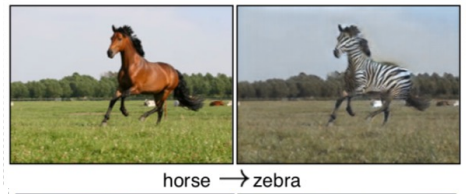
- B. 图像到图像转换(Image-to-Image Translation)
- 学习源域的图像到目标域的图像之间的映射

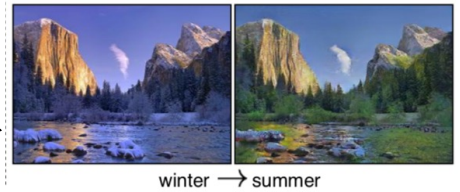
1）基于成对数据

- Pix2Pix, Pix2PixHD, BicycleGAN等

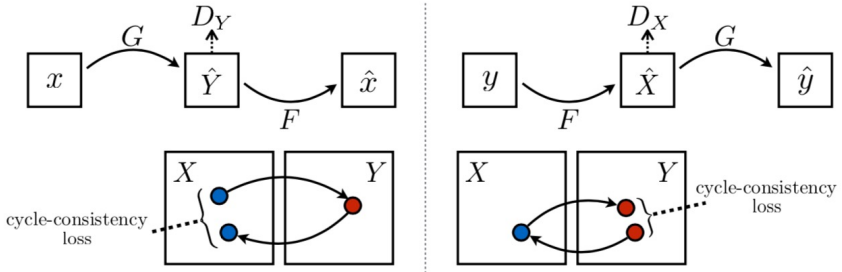
2）基于非成对数据 —— 循环一致性约束

- 双域之间转换：CycleGAN, DualGAN, UNIT, CouncilGAN等
- 多域之间转换：StarGAN, ComboGAN等
- 多模态转换：MUNIT, DRIT等





- 缺点：对人脸图像的面部特征保持差，在高度抽象的目标风格上失效



循环一致性约束

相关工作

1. 人脸图像艺术风格转换

- C. 针对人脸的图像到图像转换（艺术肖像域）
- 代表工作：
  - APDrawingGAN、 APDrawingGAN++
    - ——基于成对数据的层次化生成网络
  - Unpaired-Portrait-Drawing
    - ——基于非成对数据的非对称循环映射
  - QMUPD
    - ——肖像线条画评估指标引导的生成模型
- 缺点：只处理静态的肖像线条画生成。如应用于真实视频逐帧合成肖像画，结果可以清楚地观察到严重的不连续性。

相关工作

2. 语音驱动的说话人视频合成

MakeItTalk [TOG2020]

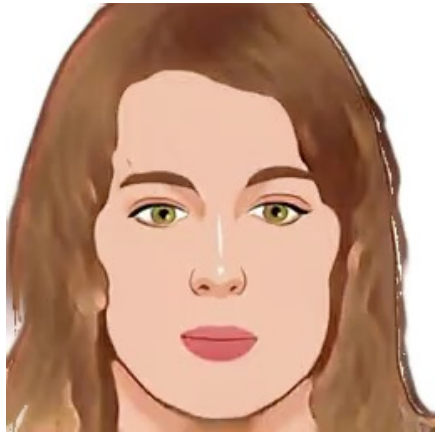
- 跨模态，从音频合成面部变化
- 真实感的说话人视频生成
- 代表工作：
- 针对特定人物的方法：
  - Synthesizing Obama, Neural Voice Puppetry, AudioDVP等
- 任意人物的通用方法：
  - YouSaidThat, ATVG, DAVS , Speech2Vid, MakeItTalk, Rhythmic Head, PC-AVS 等
- 输出结果和输入人像都属于同一个域，未进行域或风格的转换

语音驱动的艺术肖像说话视频生成

- 问题介绍
  - 研究语音驱动的艺术肖像视频的生成
  - 如何从单张人脸照片和一段语音信号生成艺术肖像说话视频（每一帧为艺术肖像画）
  - 要求与语音同步且保持人物身份

人脸照片

语音信号

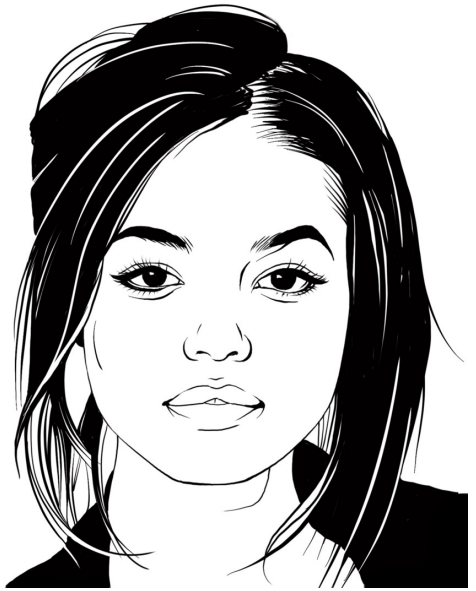


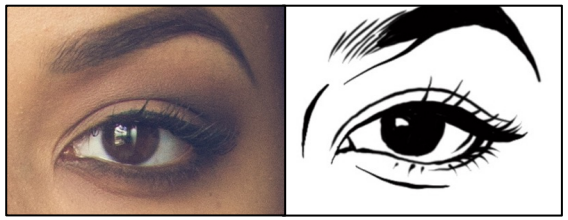
本方法合成的艺术肖像说话视频

语音驱动的艺术肖像说话视频生成

艺术肖像线条画的挑战

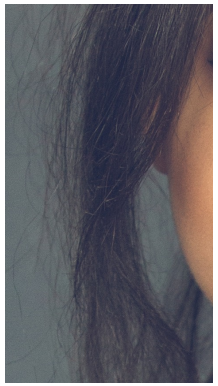
- 具有以下特点：
  1）高度抽象性：只由少数稀疏、连续的图形元素组成
  2）强语义性：包含面部特征
  3）非精确性：一些面部特征的轮廓没有完全精准定位
  4）概念性：艺术肖像画中包含概念性的线条





人脸照片与肖像画对应图像块
非精确性

概念性线条



语音驱动的艺术肖像说话视频生成

问题介绍

- 任务：从一张人脸照片和一段语音信号生成与语音同步且保持人物身份的艺术肖像说话视频
- 目前没有针对此任务的解决方法
- 分解为子问题的解决方案：
  1. 真实感说话视频生成+逐帧生成肖像画：帧间不连续，稀疏图形元素的出现与消失
  2. 静态肖像画生成+基于变形的说话视频生成：所有帧均由一个肖像画变形得到，存在产生怪异感的变形，头部运动大时不自然

帧间不连续

怪异感的变形，不自然

语音驱动的艺术肖像说话视频生成

- 提出一个基于音频信号的跨模态的艺术肖像说话视频生成模型

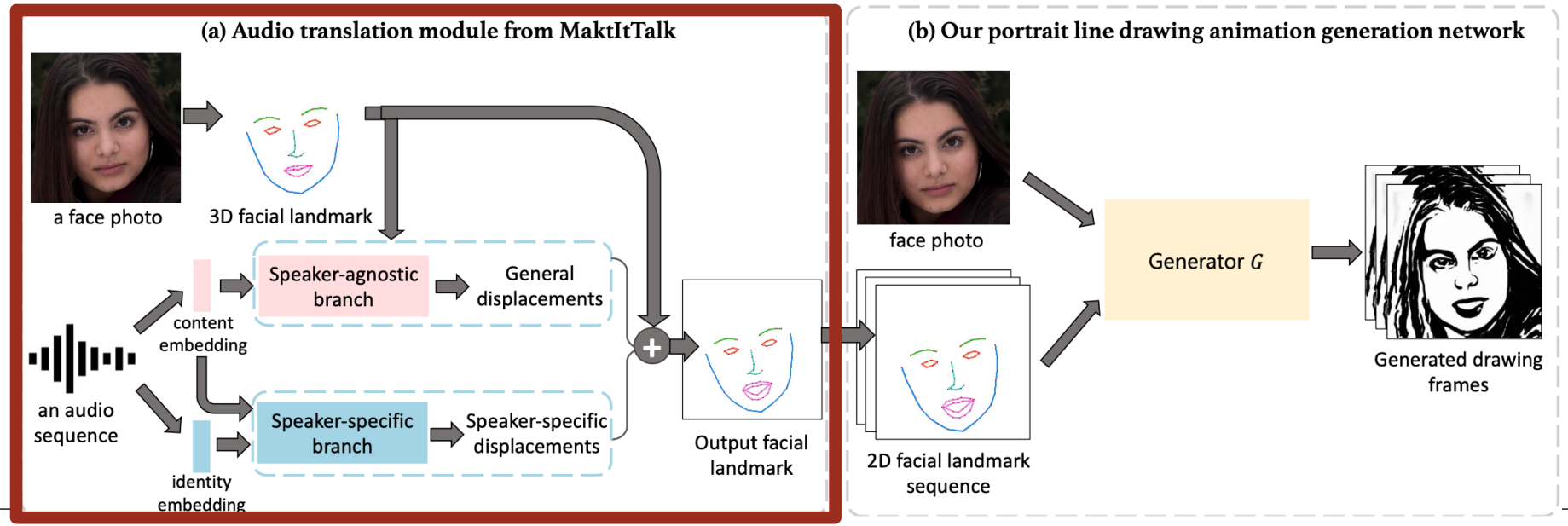
1）新的跨模态生成框架，特征空间扭曲

2）新的生成模型，同时实现几何变形和艺术风格转换，仅使用静态肖像画数据训练生成器

3）提出前景和背景分离处理的方案，以避免不适当的背景扭曲

语音驱动的艺术肖像说话视频生成

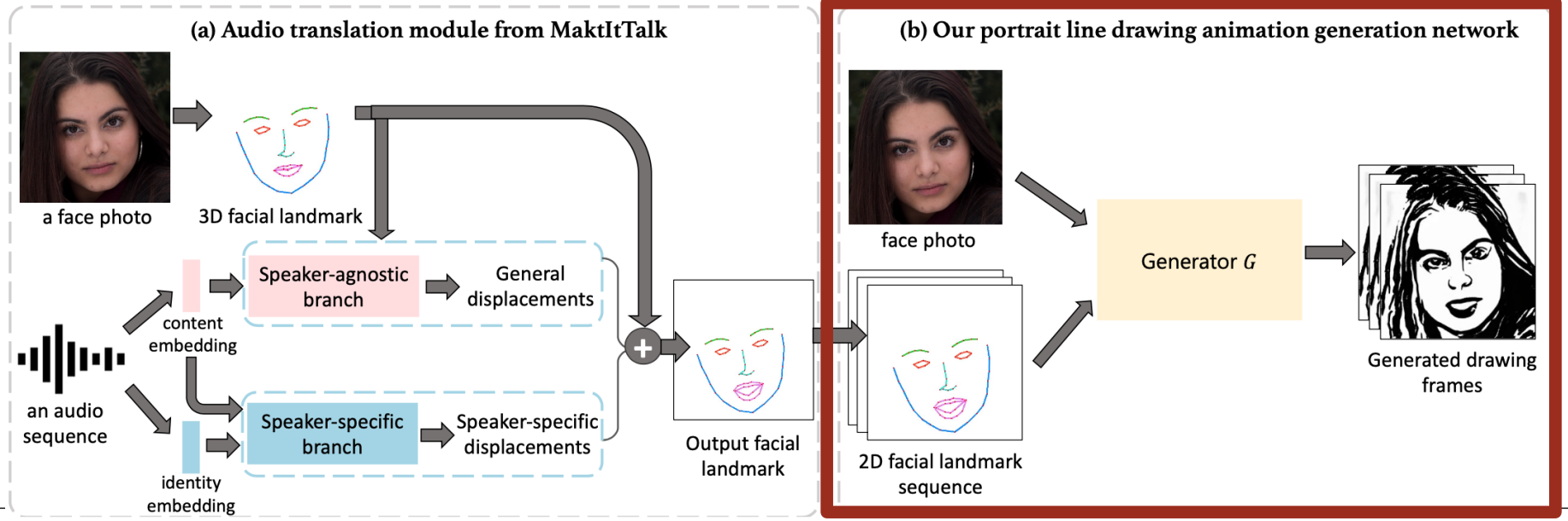
- 阶段1：语音到人脸特征点序列的转换
  - 基于MakeItTalk方法的语音网络
  - 学习语音特征到人脸特征点变化的映射
  - 双分支：身份无关分支学习通用的特征点变化，身份相关分支学习与身份相关的特征点变化



语音驱动的艺术肖像说话视频生成

阶段2：艺术肖像说话视频生成模型

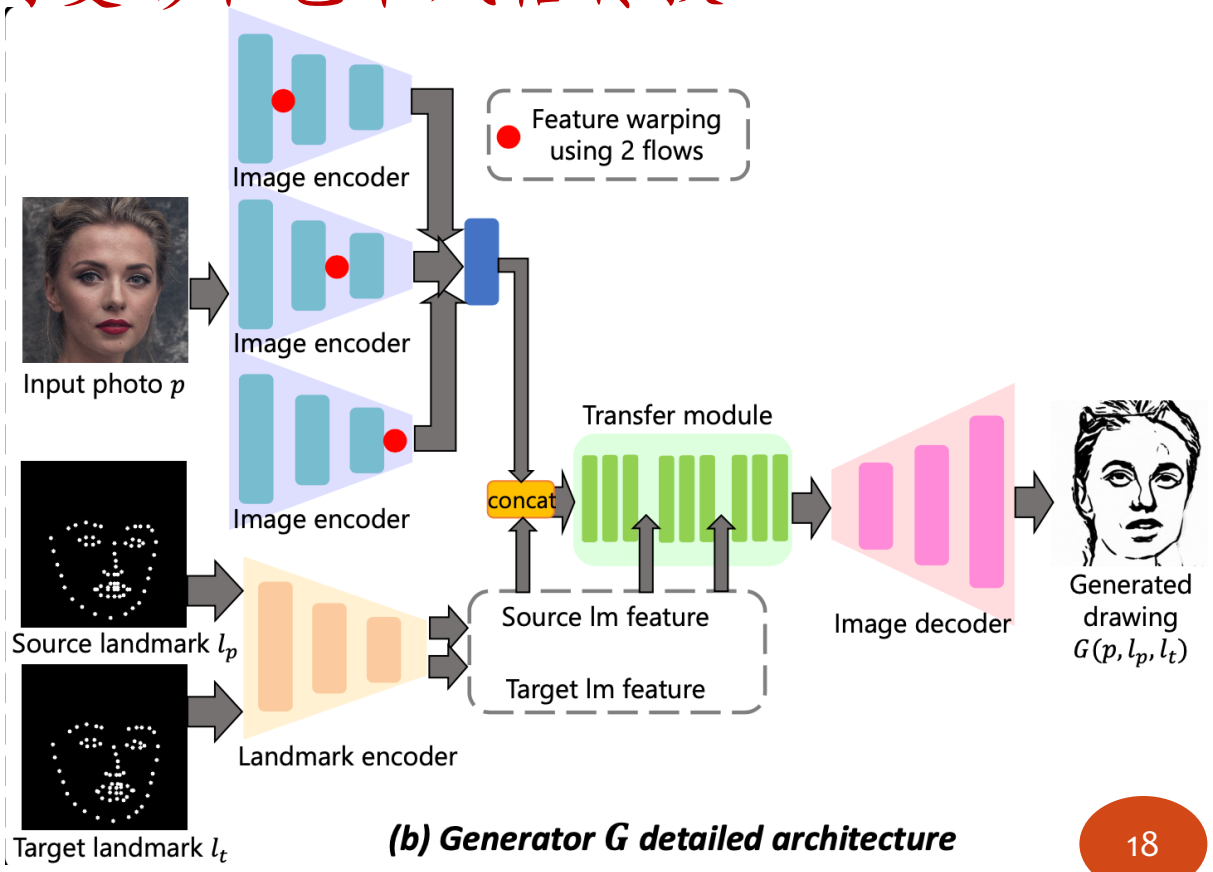
- 从人脸照片和目标特征点生成艺术肖像画，与目标特征点具有相同面部几何结构，与输入人脸照片具有相同的人物身份
- 难点1 同时进行面部几何变形和艺术风格转换
- 难点2 由于真实的艺术肖像视频数据难以获取，从互联网收集的非成对的静态人脸照片和艺术肖像画图像学习生成时间上连续的艺术肖像说话视频



语音驱动的艺术肖像说话视频生成

艺术肖像说话视频生成模型

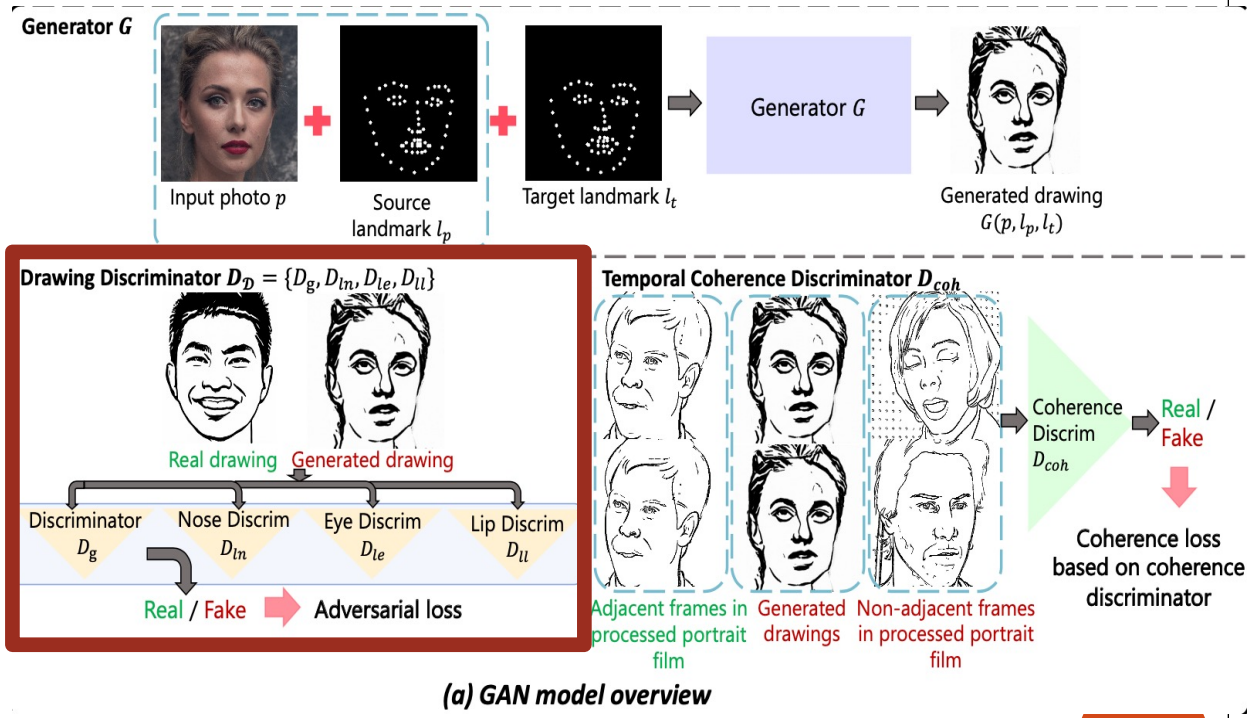
- 挑战一：同时进行面部几何变形和艺术风格转换
- 网络结构
  - 基于**特征扭曲**的生成器：由图像编码器、特征点编码器、特征扭曲模块、转换模块、图像解码器组成，**在特征空间进行变形——根据从特征点预测的Flow map对图像特征进行扭曲，多尺度特征扭曲**



语音驱动的艺术肖像说话视频生成

艺术肖像说话视频生成模型

- 挑战一：同时进行面部几何变形和艺术风格转换
- 网络结构
  - 基于特征扭曲的生成器：由图像编码器、特征点编码器、特征扭曲模块、转换模块、图像解码器组成，在特征空间进行变形
  - 肖像画鉴别器：区分生成的肖像画与真实的肖像画，1个全局鉴别器，3个局部鉴别器



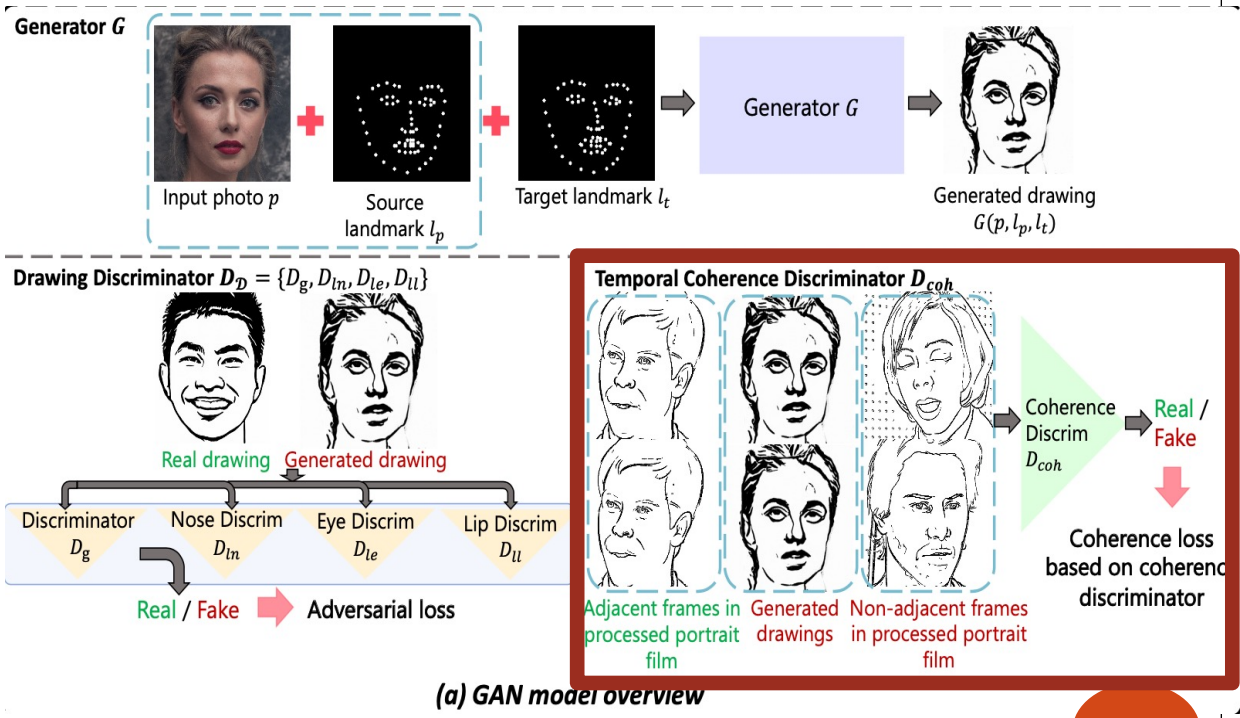
(a) GAN model overview

语音驱动的艺术肖像说话视频生成

艺术肖像说话视频生成模型

挑战二：真实的艺术肖像视频数据难以获取

- 网络结构
  - 帧间一致性鉴别器：判别两个肖像画是否在时间上是连续的。
  - 为获取正例，使用彩色肖像画电影，提取线条将其转换为线条画。处理后的相邻两帧作为近似的正例



(a) GAN model overview

语音驱动的艺术肖像说话视频生成

艺术肖像说话视频生成模型

- 挑战一：同时进行面部几何变形和艺术风格转换
- 损失函数：包括6个损失项

1. 对抗损失：衡量肖像画鉴别器的鉴别能力

$$L_{adv}(G, D_D) = \sum_{D \in D_D} \mathbb{E}_{d \in S(d)}[\log D(d)] + \sum_{D \in D_D} \mathbb{E}_{p \in S(p)}[\log(1 - D(G(p, l_p, l_t))]$$

2. 基于图像扭曲的内容损失：使用静态肖像画生成方法和图像扭曲的结果作为近似的Ground Truth

$$L_{content}(G) = \mathbb{E}_{p \in S(p)}[||W(d_s, l_p, l_t) - G(p, l_p, l_t)||_1]$$

语音驱动的艺术肖像说话视频生成

艺术肖像说话视频生成模型

- 挑战一：同时进行面部几何变形和艺术风格转换
- 损失函数：包括6个损失项

3. 几何损失：使用人脸特征点检测器，衡量生成肖像画中特征点与目标特征点距离 $L_{geom}(G) = \mathbb{E}_{p \in S(p)}[||x(l_t) - R_{land}(G(p, l_p, l_t))||_2]$

   引入嘴唇轮廓掩膜，提高嘴唇轮廓线质量

$$L_{geom\_lip}(G) = \mathbb{E}_{p \in S(p)}[||G(p, l_p, l_t) \cdot M_{lip\_line}||_1]$$

4. 身份保持损失：使用人脸识别网络提取的身份特征，衡量身份特征在转换前后相似性

$$L_{iden}(G) = \mathbb{E}_{p \in S(p)}[||R_{iden}(d_s) - R_{iden}(G(p, l_p, l_t))||_1]$$

语音驱动的艺术肖像说话视频生成

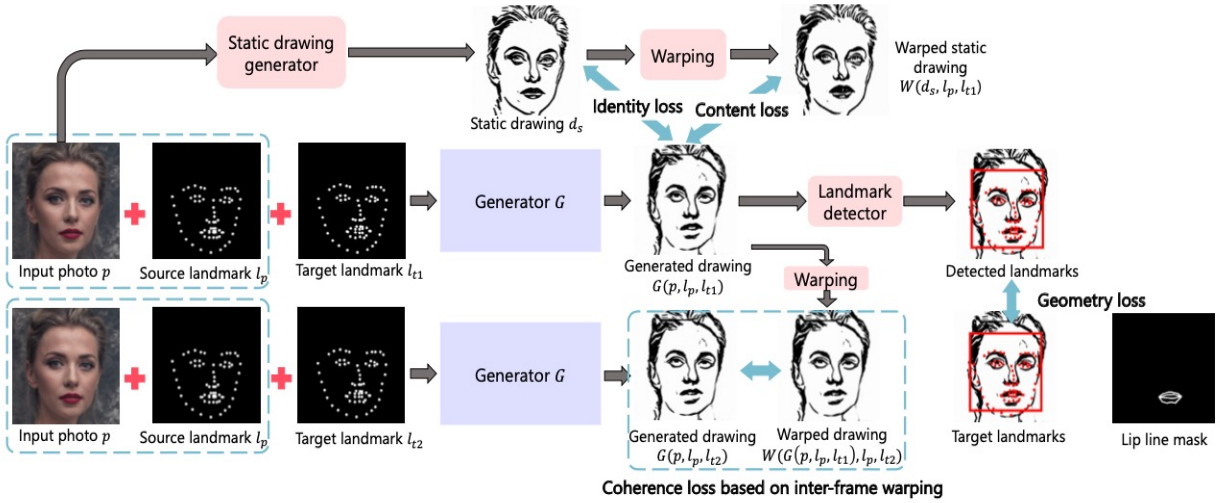
艺术肖像说话视频生成模型

- 挑战二：真实的艺术肖像视频数据难以获取
- 损失函数：包括6个损失项

5. 基于帧间图像扭曲的帧间一致性损失

- 人脸照片p分别与两组具有细微差别的目标特征点l_{t1}和l_{t2}作为输入，由此生成的两个肖像画$G(p, l_p, l_{t1})$和$G(p, l_p, l_{t2})$作为视频中连续两帧的模拟
- 两个肖像画经过图像扭曲后应近似

$$L_{coh1}(G) = \mathbb{E}_{p \in S(p)}[||W(G(p, l_p, l_{t1}), l_{t1}, l_{t2}) - G(p, l_p, l_{t2})||_1]$$



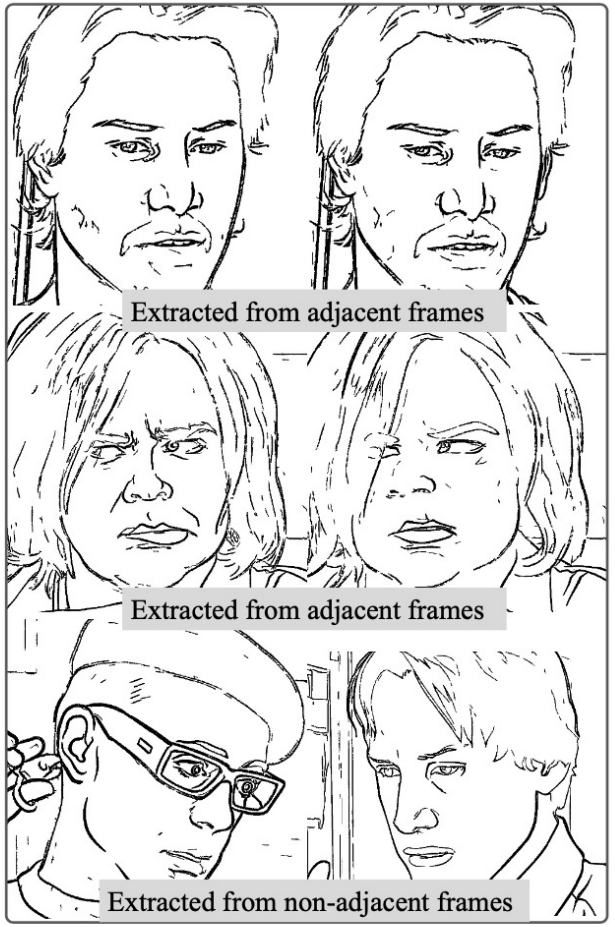
语音驱动的艺术肖像说话视频生成

艺术肖像说话视频生成模型

- 挑战二：真实的艺术肖像视频数据难以获取
- 损失函数：包括6个损失项

6. 基于帧间一致性鉴别器的帧间一致性损失
   - 1种真实样本：肖像电影提取线条后的相邻帧
   - 2种虚假样本：1）生成的两帧；
     2）彩色肖像电影提取线条后的不相邻帧

$$L_{coh2}(G, D_{coh}) = \mathbb{E}_{(d_1,d_2)\in S(adj)}[\log D_{coh}(d_1, d_2)]$$
$$+\mathbb{E}_{p\in S(p)}[\log(1 - D_{coh}(G(p, l_p, l_{t1}), G(p, l_p, l_{t2}))]$$
$$+\mathbb{E}_{(d_3,d_4)\in S(nadj)}[\log(1 - D_{coh}(d_3, d_4))]$$



(b) Extracted line drawings from film frames

语音驱动的艺术肖像说话视频生成

- 艺术肖像说话视频生成模型
  - 挑战三：背景随前景发生扭曲
  - 提出一种前景和背景分离处理的方案，以避免不适当的背景扭曲
  - 基于ModNet的前景-背景分割
  - 前景基于提出的生成网络生成
  - 背景基于静态方法生成
  - 最后将结果进行融合

Ablation Study

Input photo (Flickr public domain) | *Our method* | *w.o. foreground-background separation*

Without the foreground-background separation scheme, the background (e.g. texts) is warped and generates worse results.

语音驱动的艺术肖像说话视频生成

- 训练数据

线条画风格

(a) Artistic portrait line drawings (b) Extracted line drawings from film frames

语音驱动的艺术肖像说话视频生成

- 训练数据

卡通画风格

语音驱动的艺术肖像说话视频生成

实验比较——基线方法介绍

- 基线方法1 真实感说话人视频生成+逐帧生成肖像画：ATVG+UPD，SDA+UPD

- 基线方法2 静态肖像画生成+基于变形的说话视频生成：UPD+Warp

语音驱动的艺术肖像说话视频生成

- 实验比较——基线方法1、2比较结果

1. Comparison with ATVG+UPD, SDA+UPD, UPD+Warp

Inputs: a face photo (Flickr public domain) + an audio clip collected separately

Inputs *ATVG+UPD* *SDA+UPD* *UPD+Warp* *Ours*

Our method generates natural and artistic results with inter-frame continuity, which both synchronize with the audio and capture identity.

语音驱动的艺术肖像说话视频生成

实验比较——基线方法介绍

- 基线方法3 MakeItTalk+静态肖像画生成逐帧生成肖像画：MakeItTalk+UPD，MakeItTalk+CycleGAN
- 基线方法4 基于关键帧的视频风格化方法：EbySynth (TOG19)

语音驱动的艺术肖像说话视频生成

实验比较——基线方法3比较结果

2. Comparison with MakeItTalk+UPD, MakeItTalk+CycleGAN

Input photo (Flickr public domain)

Landmark sequence

MakeItTalk +CycleGAN

facial features missing

MakeItTalk +UPD

flickering, less temporally coherent

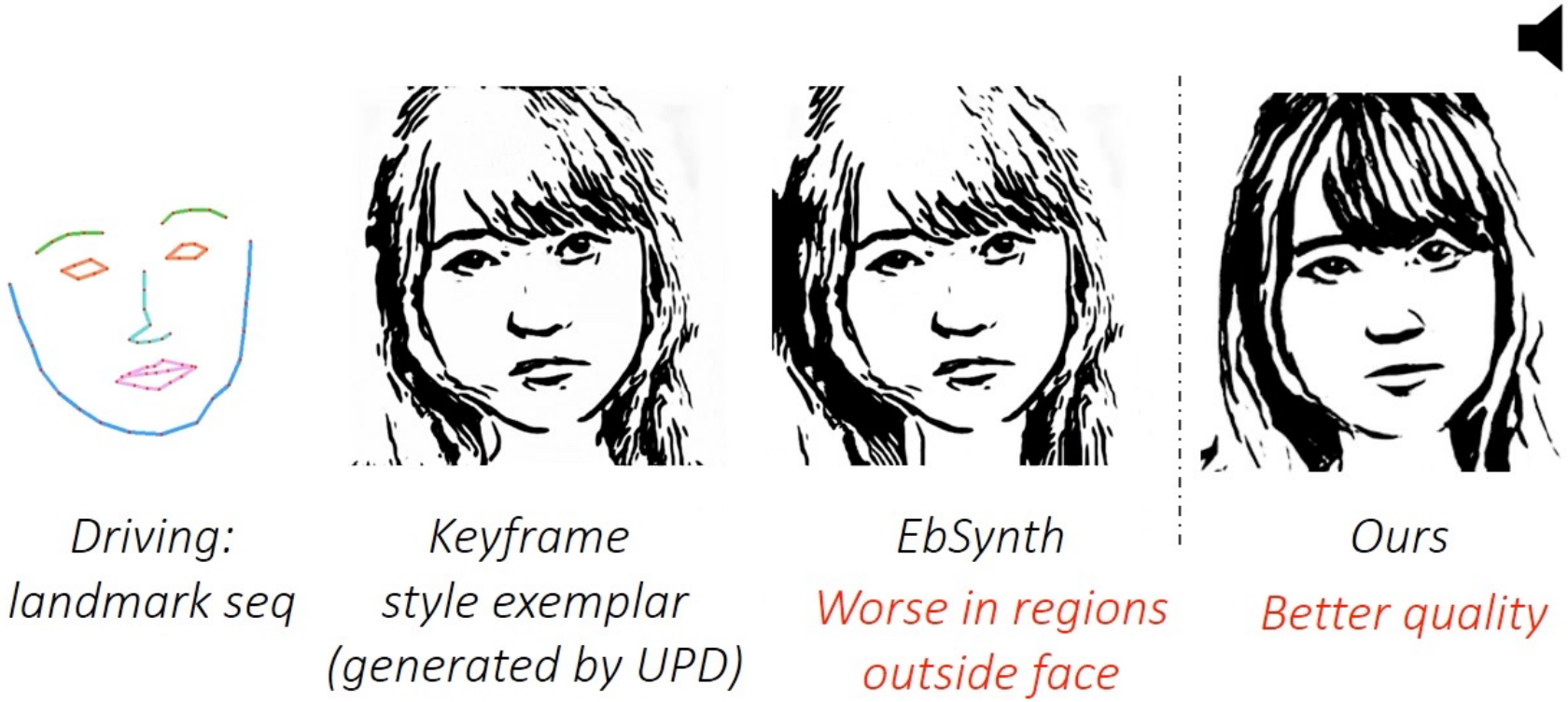
Ours

Better quality

语音驱动的艺术肖像说话视频生成

- 实验比较——基线方法4比较结果

3. Comparison with keyframe-based method EbSynth



语音驱动的艺术肖像说话视频生成

用户实验评估

- 57位用户参与评估
- 本方法在各个评估维度上表现最优

Methods	Naturalness	ID preserve	Lip sync	Take-all
ATVG+UPD	1.0%	0.6%	22.7%	1.3%
SDA+UPD	1.5%	0.8%	2.1%	1.1%
UPD+Warp	14.6%	23.7%	12.6%	14.2%
Ours	**82.9%**	**74.8%**	**62.6%**	**83.3%**

量化指标评估

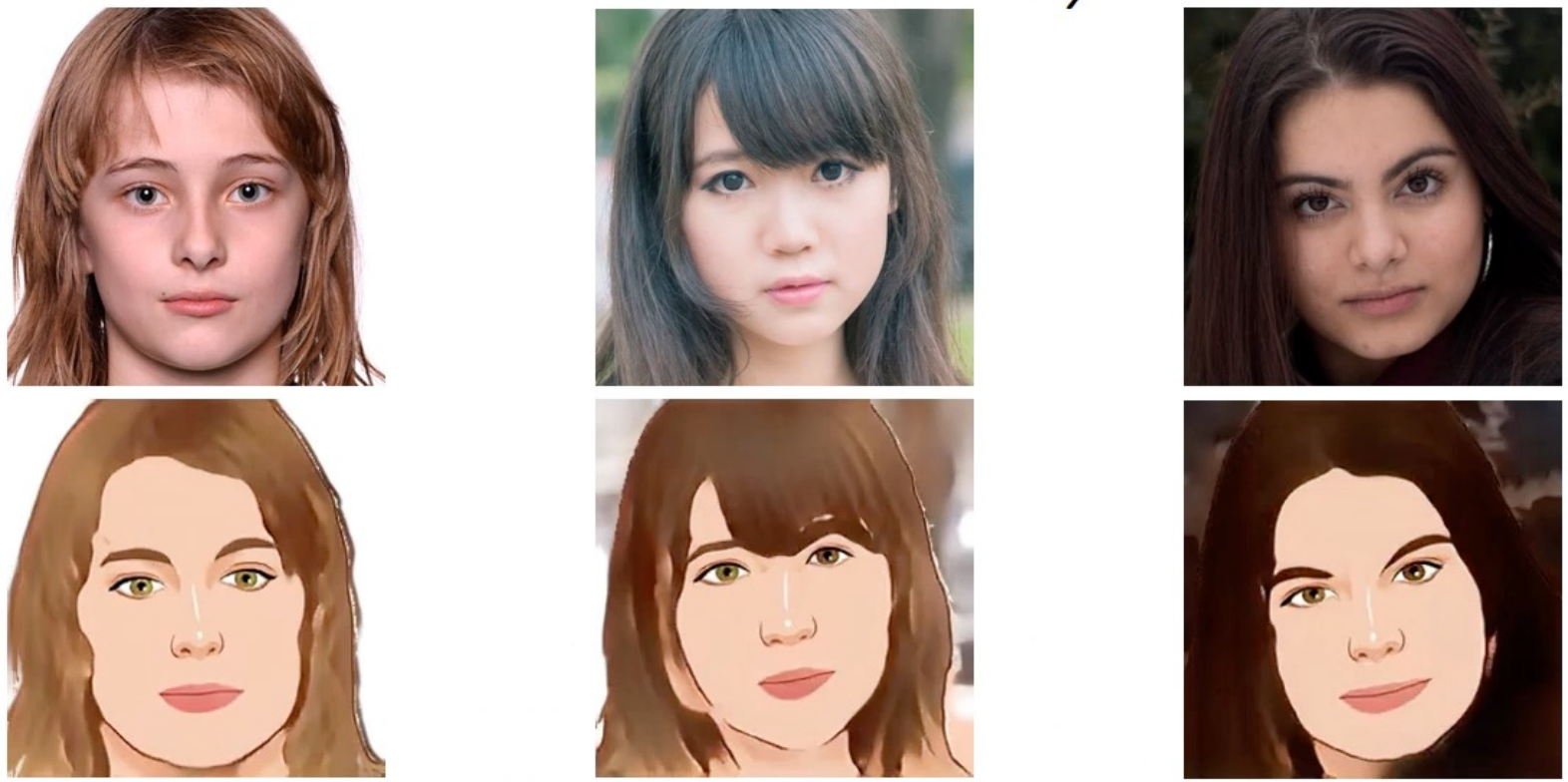
- FID距离衡量生成肖像画分布与真实分布距离(单帧质量)
- 帧间SSIM衡量帧间连续性
- LMD衡量语音-唇部运动同步性

Methods	FID ↓	SSIM ↑	LMD ↓
ATVG+UPD	162.1	0.923	2.74
SDA+UPD	185.2	0.913	2.76
MakeItTalk+UPD	143.0	0.951	**2.64**
UPD+Warp	151.9	**0.975**	**2.64**
Ours	**135.2**	0.974	**2.64**

语音驱动的艺术肖像说话视频生成

- 更多风格的生成结果（卡通、油画、水彩）

Extension to cartoon style



Our method can be easily extended to a cartoon style, and generate good results of talking portrait cartoons.

研究总结

- 研究基于单张照片和语音信号的艺术肖像说话视频生成，提出了一种基于特征空间扭曲的生成框架，首先从语音信号中预测面部特征点的运动，然后提出一个新的生成模型同时进行艺术风格转换和几何变形。
- 仅使用静态肖像画数据训练生成器，为了解决生成视频帧间不连续的问题，提出了两个新的帧间一致性优化项：(1) 基于帧间图像扭曲的损失项和 (2) 基于帧间一致性鉴别器的损失项。
- 在线条画、卡通、油画、水彩等风格上生成了高质量、具有表现力的艺术肖像说话视频，有助于提升动画创作效率。

Limitation

- 对于模糊或有夸张表情的输入，结果不佳
- 受到前景分割精度的影响

感谢各位老师同学！
请批评指正

上海交通大学 计算机系 易冉
数字媒体与计算机视觉实验室
https://dmcv.sjtu.edu.cn/
联系方式：ranyi@sjtu.edu.cn
个人主页：https://yiranran.github.io

实验室主页　个人主页