南京郵電大學
Nanjing University of Posts and Telecommunications

SadTalker：学习逼真的3D运动系数以实现风格化的音频驱动单图说话人脸动画

SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-DrivenSingle Image Talking Face Animation

Zhang W, Cun X, Wang X, et al.
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

汇报人：曹陈晨　　**指导老师**：徐鹤

汇报时间：2025年5月25日

目录

CONTENTS

Part.01

研究背景

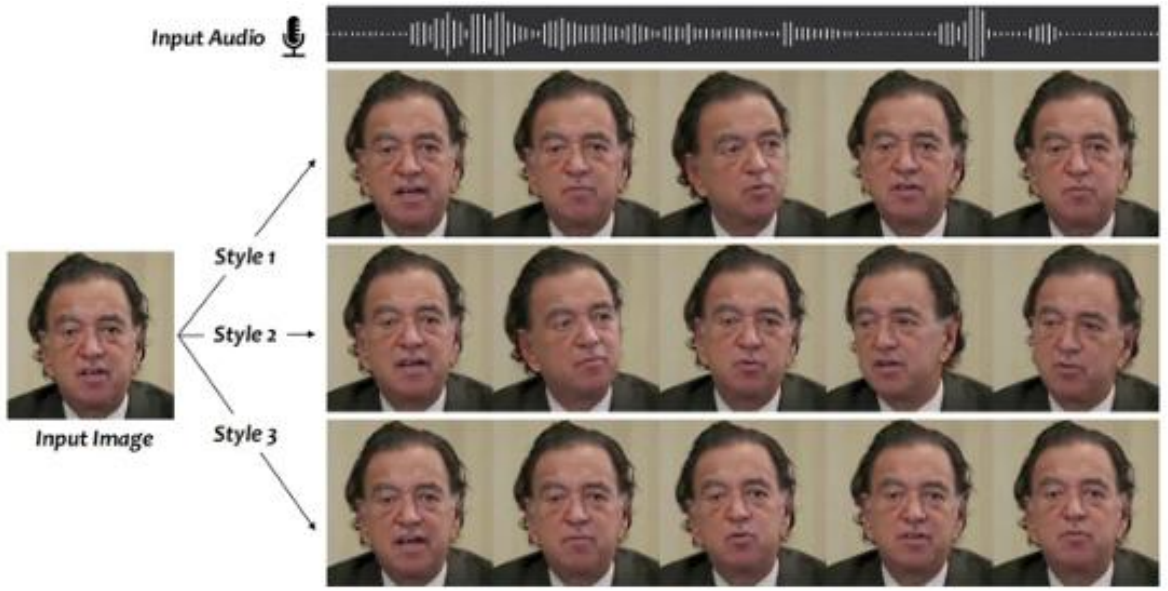
Research Background

1.1 研究背景

研究背景概述

单图像生成说话视频

利用一张静态人脸图像和一段对应的语音音频，生成一段图像中的人物说话的视频。
生成的面部动画旨在嘴唇动作与语音内容精准匹配，头部姿态、眼神变化、面部表情等也都要符合说话时的自然状态。



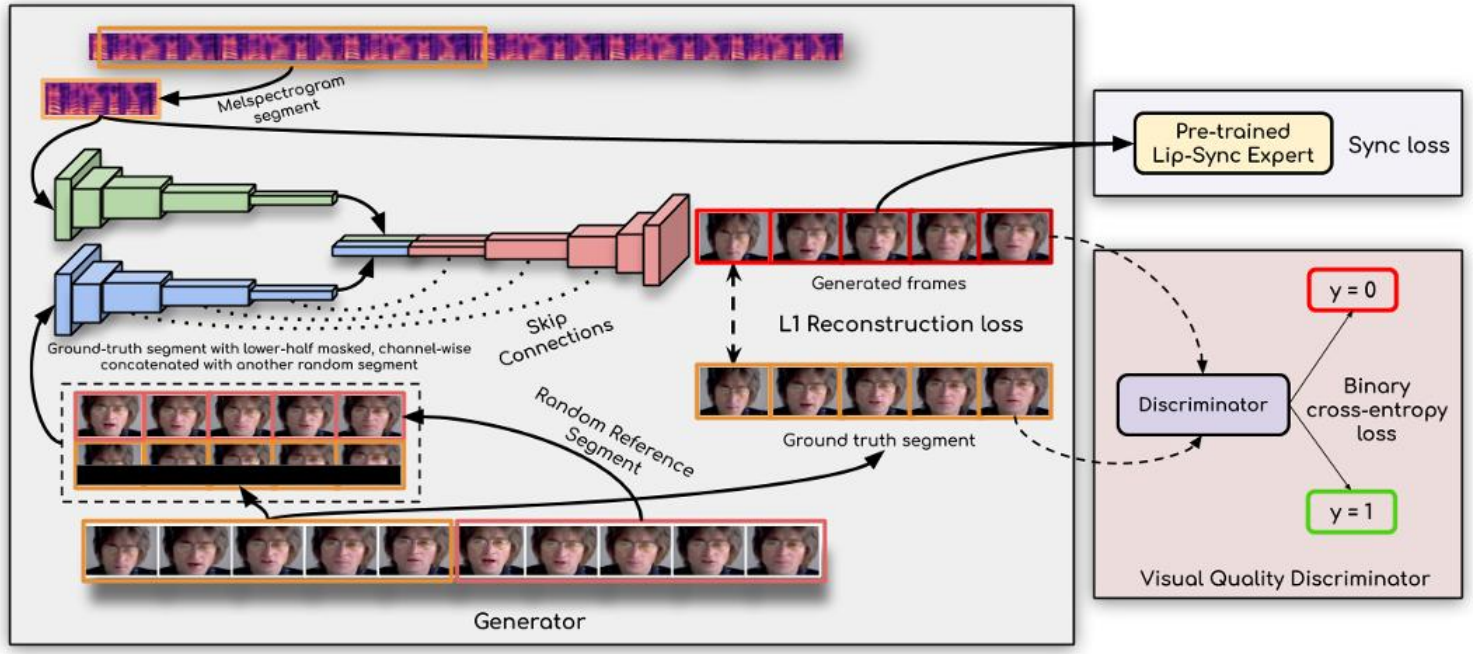
SadTalker的效果

1.1 研究背景

唇音同步

前期研究主要利用感知鉴别器生成准确的嘴唇运动，以wav2lip模型为例，它采用的是GAN的训练范式，一共有1个生成器和2个判别器。

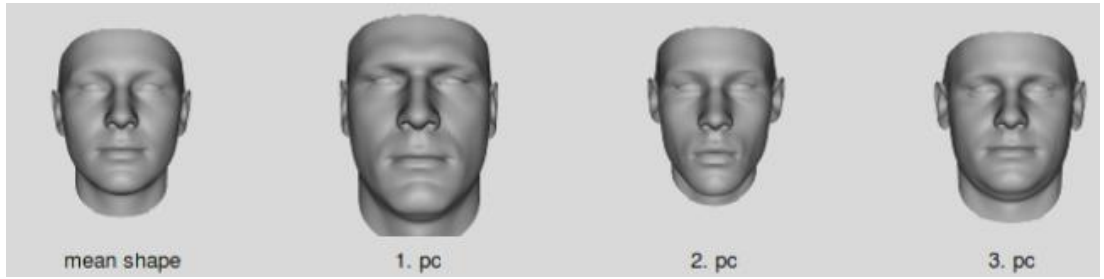
论文提出由于嘴唇区域在全脸占比小，导致常规模型前期主要学习重建其他信息，后期才处理嘴唇变形。因此需要一个额外判别器判断唇音同步。



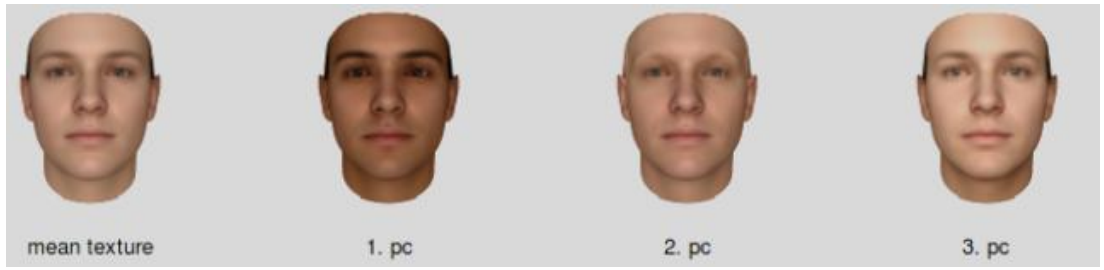
wav2lip模型架构

现有的判别器仅用单帧画面，缺乏时序上下文信息，且聚焦于生成图像伪影而非唇音同步。因此，基于 SyncNet 设计了预训练好的唇音同步判别器，训练后不再微调，防止模型权重被污染。

1.1 研究背景



形状向量Shape Vector：$S = (X_1, Y_1, Z_1, X_2, Y_2, Z_2, \ldots, Y_n, Z_n)$



纹理向量Texture Vector：$T = (R_1, G_1, B_1, R_2, G_2, B_2, \ldots, G_n, B_n)$

任意的人脸模型可以由数据集中的m个人脸模型进行加权组合获得：

$$\mathbf{S}_{mod} = \sum_{i=1}^{m} a_i \mathbf{S}_i, \quad \mathbf{T}_{mod} = \sum_{i=1}^{m} b_i \mathbf{T}_i, \quad \sum_{i=1}^{m} a_i = \sum_{i=1}^{m} b_i = 1$$

3D 可变形模型

3DMM 核心思想是人脸可由数据库中人脸正交基加权线性组合表示，求解模型即求基向量系数。左式降维分解得到：

$$S_{model} = \bar{S} + \sum_{i=1}^{m-1} \alpha_i s_i, T_{model} = \bar{T} + \sum_{i=1}^{m-1} \beta_i t_i$$

后续发展出了深度学习 3DMM 重建方法，包括：全监督方法，用模型直接回归系数，如 3DMM CNN；自监督方法，不依赖真实成对数据，以 MoFa 为代表；还有通过特殊特征编码提升重建效果的方法。

1.2 本文创新点

SadTalker模型用3D运动系数作为中间表示，通过音频驱动对单图片生成风格多样的说话人视频。

为了从音频中学习 3DMM 的真实运动系数，分别提出了 ExpNet（表情系数） 和PoseVAE（头部姿态）。

提出了一种新的语义解耦和3D感知的面部渲染方法，以生成逼真的说话人视频。

Part.02

思路与方法

Research Ideas And Methods

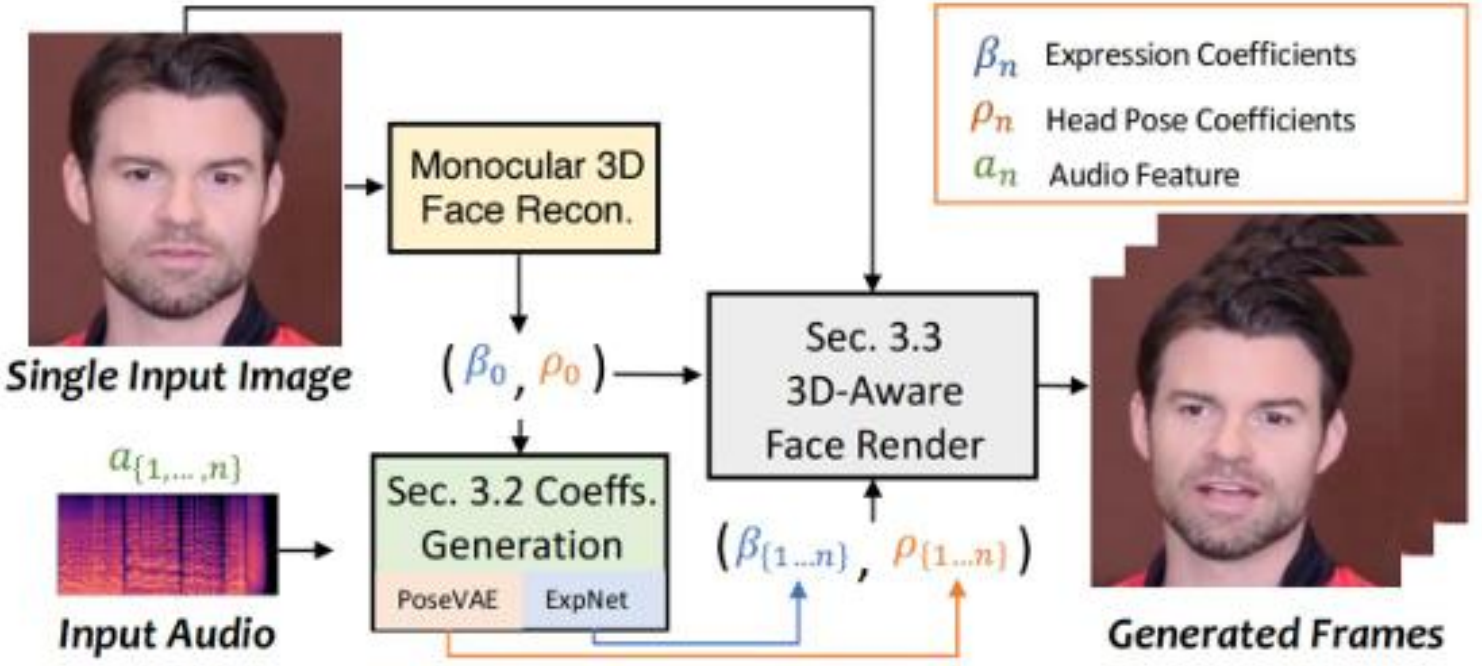
南京郵電大學
Nanjing University of Posts and Telecommunications

2.1 SadTalker总体框架

输入的图片通过单目 3D 人脸重建得到初始的表情系数β_0和头部姿态系数ρ_0，输入的音频经过处理后提取到音频特征$a_{\{1...n\}}$。

运动系数生成部分分为PoseVAE（生成头部姿态系数$\rho_{[1...n]}$）以及ExpNet（生成表情系数$\beta_{[1...n]}$）。

3D 感知人脸渲染部分对输入的初始的表情系数β_0、头部姿态系数ρ_0，以及生成的表情系数$\beta_{[1...n]}$、头部姿态系数$\rho_{[1...n]}$，结合音频特征等信息输入该模块，最终渲染生成说话人脸动画的一系列帧。



SadTalker 系统中音频驱动单图像生成说话人脸动画的流程图

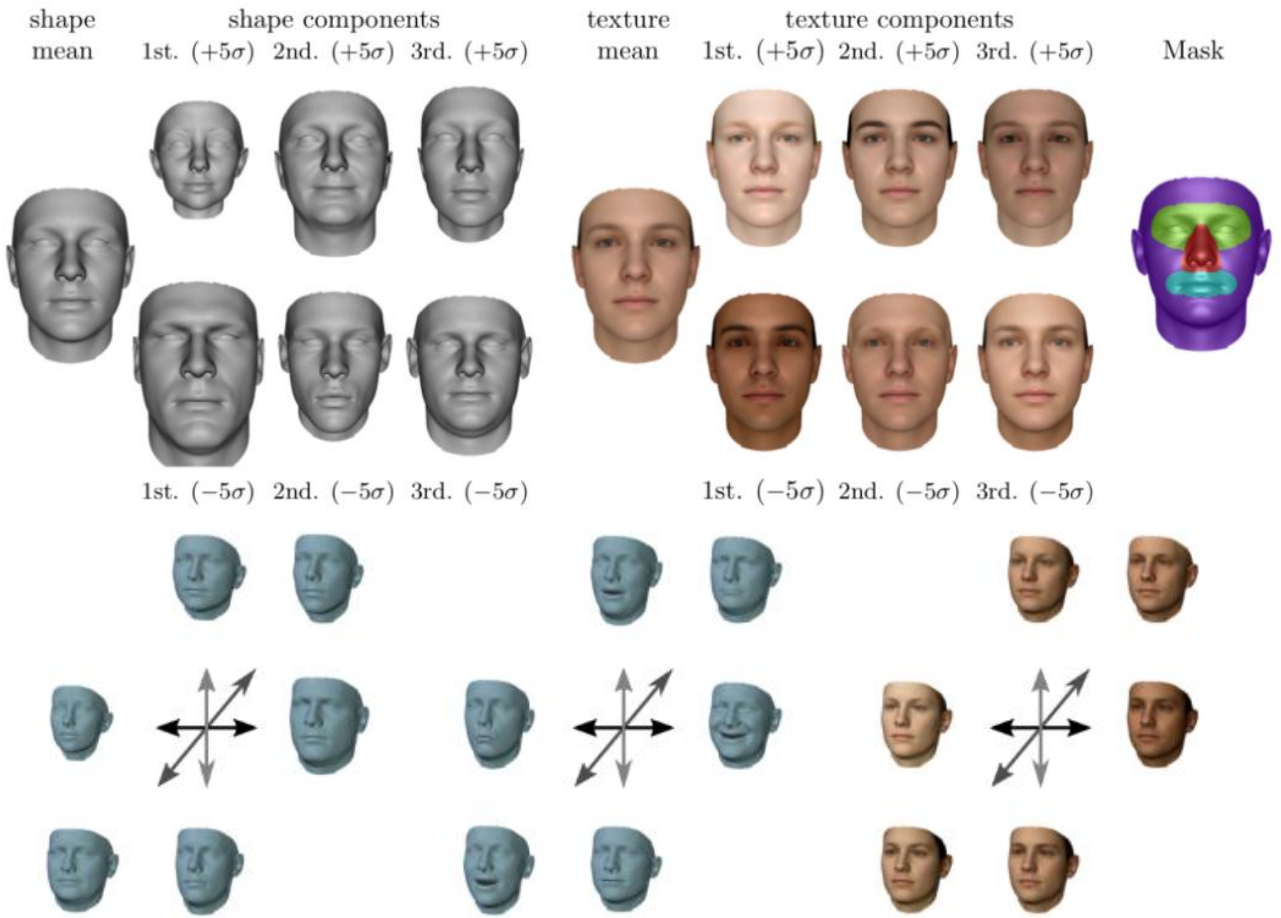
2.2 三维人脸模型基础

受到单图像深度三维重建方法的启发，将三维可变形模型的空间作为中间表示。在此空间中，3D 人脸形状可分解为：

$$S = \bar{S} + \alpha U_{id} + \beta U_{exp}$$

其中是$\bar{S}$是3D 人脸的平均形状，U_{id}和U_{exp}分别是基于身份和表情的可变形模型基；系数$\alpha \in \mathbb{R}^{80}$和$\beta \in \mathbb{R}^{64}$分别描述人物身份和表情。

为保留姿势变化系数，用$r \in SO(3)$和$t \in \mathbb{R}^3$表示头部旋转和平移。系统仅对头部姿势$\rho = [r, t]$和表情系数β进行建模，再通过之前引入的音频驱动系数，隐式调制这些运动参数以实现最终合成。



BFM2017数据集的平均人脸和三个线性系数

2.3.1 ExpNet（表情运动）

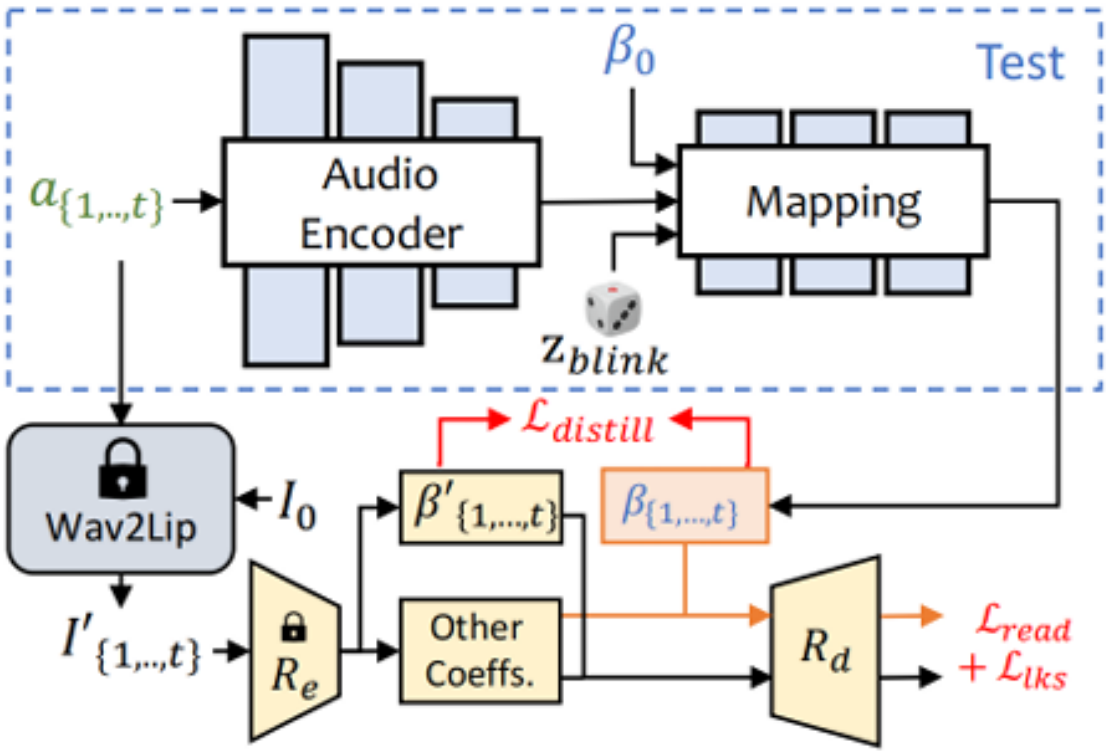
生成表情系数：

模型将对参考图像进行单目 3D 人脸重建获得的初始表情系数β_0，音频特征$a_{\{1...n\}}$处理得到的嵌入，以及眨眼控制信号z_{blink}拼接输入到映射网络中，解码得到每一帧的表情系数：

$$\beta_{\{1,...,t\}} = \phi_M\left(\phi_A\left(a_{\{1,...,t\}}\right), z_{blink}, \beta_0\right)$$

训练过程：

音频特征$a_{\{1...n\}}$和参考图片输入到Wav2Lip得到对应口型的视频帧$I'_{\{1,...,t\}}$，再将每一帧输入到3D人脸重建模型取其中的表情系数$\beta'_{\{1,...,t\}}$和Mapping输出的表情系数做蒸馏学习，其他系数构建面部标志点损失L_{lks}和唇读损失L_{read}.



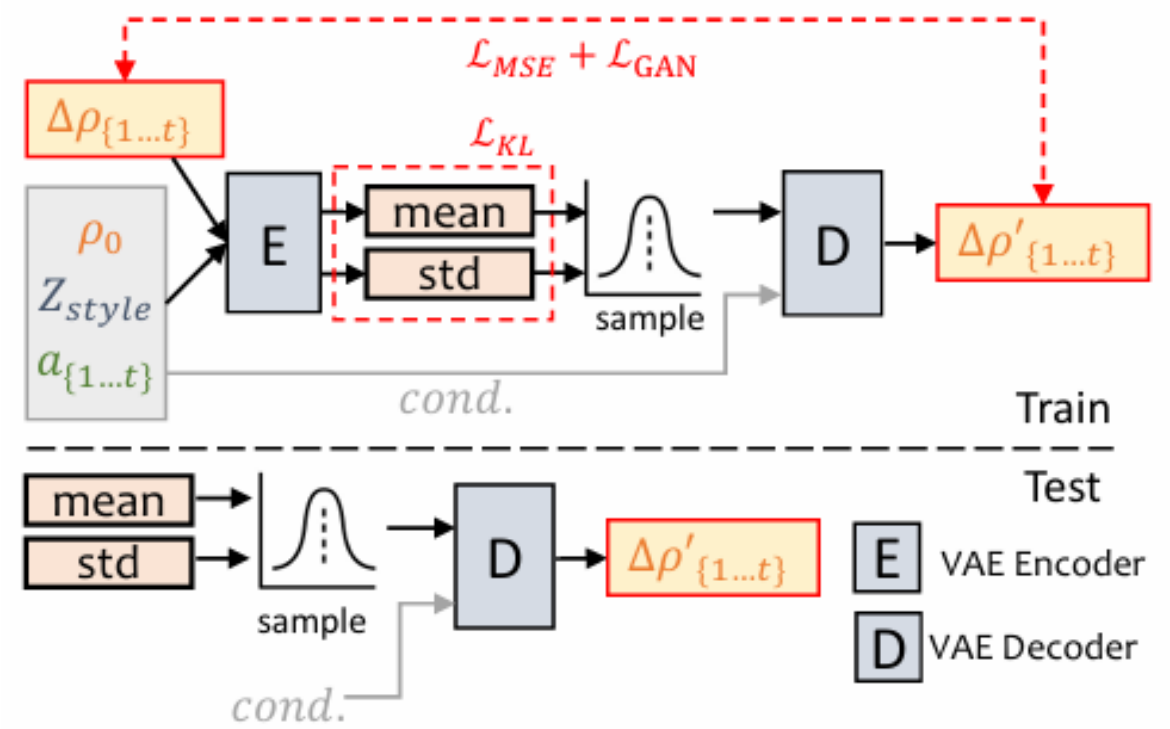
ExpNet 网络结构

2.3.2 PoseVAE（头部姿态）

参考条件VAE设计，头部姿态风格Z_{style}作为条件变量，连同初始头部姿态系数ρ_0和音频特征$a_{\{1...n\}}$处理得到的嵌入一同输入到VAE编码器中。

编码器将输入数据嵌入到一个高斯分布中，得到均值（mean）和标准差（std）；然后从分布中进行采样后输入到解码器。在解码器中，网络学习生成第一帧的条件姿态的残差。

损失函数由三个方面组成：
$\mathcal{L}_{KL}$用于衡量生成的头部姿态分布与期望分布之间的差异；$\mathcal{L}_{MSE}$和$\mathcal{L}_{GAN}$计算生成的头部姿态与真实头部姿态之间的平均误差。



PoseVAE网络结构

2.3.3 3D 感知面部渲染器

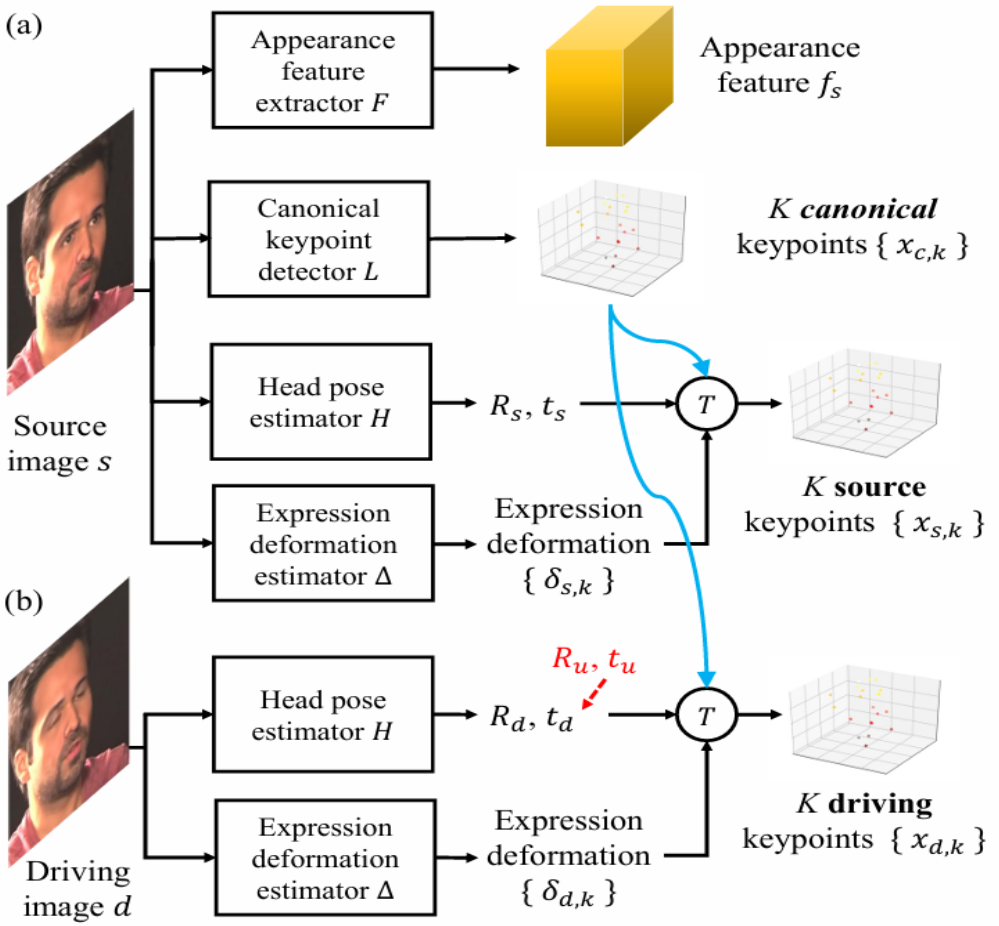
面部渲染器参考了face-vid2vid模型。该模型从视频中提取人脸隐式3D关键点作为驱动信号，具体流程如下：

原图像处理流程：

对于原图像，通过外观特征提取器获取外貌特征，结合关键点检测器、头部姿态估计器和表情变形估计器经过变换操作T得到K个源关键点。

驱动图像处理流程：

只需要通过头部姿态估计器和表情变形估计器获得特征参数，借助源图像的规范关键点经变换操作T得到K个驱动关键点。



Facevid2vid模型架构

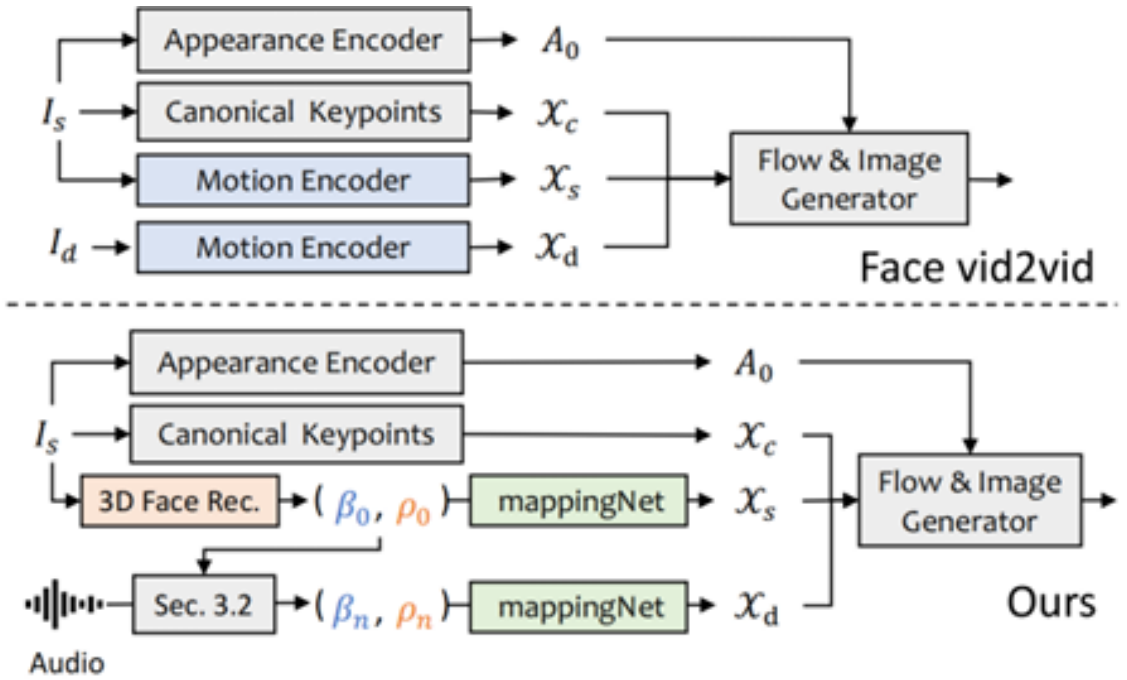
2.3.3 3D 感知面部渲染器

本模型通过3DMM系数作为中间表达，通过mappingNet 将其映射到无监督的 3D 关键点空间后进行面部生成。

为了人脸动作更平滑，mappingNet使用整个滑动时间窗口中的系数进行平滑处理，实验后只选择了表情和头的系数姿势驱动生成。最终将源图片、源3D关键点、驱动3D关键点输入Image Generator来合成每一帧的画面。

训练过程：

首先以自监督的方式训练face-vid2vid，冻结外观特征提取器、关键点检测器以及图像生成器后，在真实视频的 3DMM 系数上训练mappingNet。



本文的面部渲染器与facevid2vid的对比

Part.03

实验分析

Experiment and Analysis

南京郵電大學
Nanjing University of Posts and Telecommunications

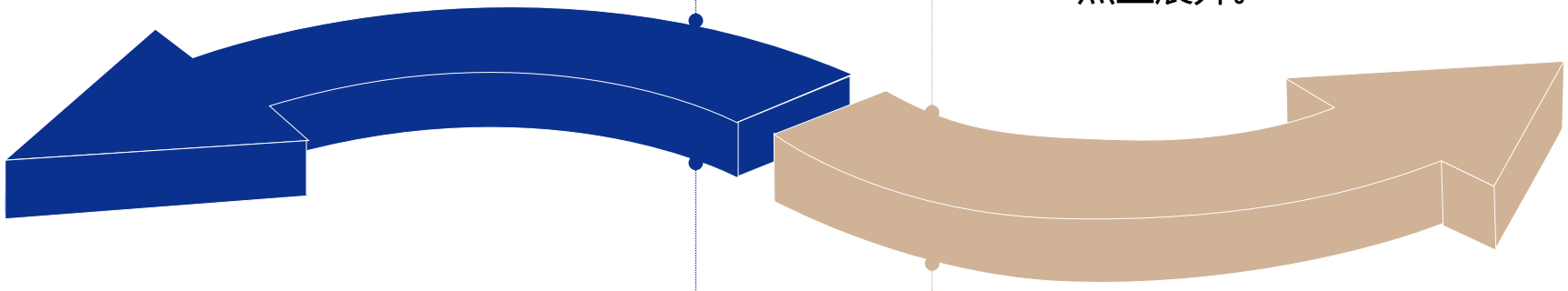
3.1 实验设置

数据集

选用 VoxCeleb 数据集，预处理后训练面部渲染器；HDTF数据集，作为测试集评估测试。

基线

选取HDTF数据集上最优的方法进行对比，评估直接在公共检查点上展开。



评估指标

通过 FID、CPBD 评估图像质量，CSIM 评估身份保持，LSE-D、LSE-C 评估唇形同步等

训练信息

ExpNet、PoseVAE 和 FaceRender用 Adam 优化器训练，3DMM 参数通过预训练的深度三维人脸重建方法提取。

3.2 结果分析

Method	Eye Blink	Lip Synchronization		Learned Head Motion		Video Quality		
		LSE-C↑	LSE-D↓	Diversity↑	Beat Align↑	FID↓	CPBD↑	CSIM↑
Real Video	N./A.	8.211	6.982	0.259	0.271	0.000	0.428	1.000
Wav2Lip* [28]	N./A.	10.221	5.535	N./A.	N./A.	21.725	0.368	0.849
PC-AVS** [49]	from ref.	9.053	6.355	N./A.	N./A.	69.127	0.206	0.683
MakeItTalk [50]	automatic	5.110	10.059	0.257	0.268	28.243	0.283	0.838
Audio2Head [37]	automatic	7.357	7.535	0.181	0.267	24.392	0.281	0.823
Wang *et al.* [38]	automatic	4.932	10.055	0.226	0.268	22.432	0.295	0.811
Ours	controllable	7.290	7.772	**0.278**	**0.293**	**22.057**	**0.335**	**0.843**

SadTalker和HDTF数据集上最优方法的比较

Method	Lip Sync.	Motion Diversity	Video Sharpness	Overall Naturalness
Wav2Lip [28]	15.6%	3.1%	2.0%	2.8%
PC-AVS [49]	18.1%	9.6%	3.4%	9.1%
MakeItTalk [50]	5.6%	5.3%	5.7%	6.9%
Wang *et al.* [38]	12.5%	12.1%	16.3%	11.6%
Audio2Head [37]	9.5%	12.1%	9.7%	14.7%
Ours	**38.7%**	**57.9%**	**62.8%**	**54.8%**

模型的用户评估结果

3.2.1 SadTalker与最优方法比较

SadTalker在整体视频质量和头部姿态多样性方面优势明显，在嘴唇同步指标上与其他方法水平相当；同时考虑到得分与真实视频相似，因此更具有优势。

3.2.2 人工评估

测试者更加青睐SadTalker表明其更高的生成质量，同时部分认为SadTalker在嘴唇同步方面更优，则是因为人们更关注视频的整体质量。

3.2 结果分析

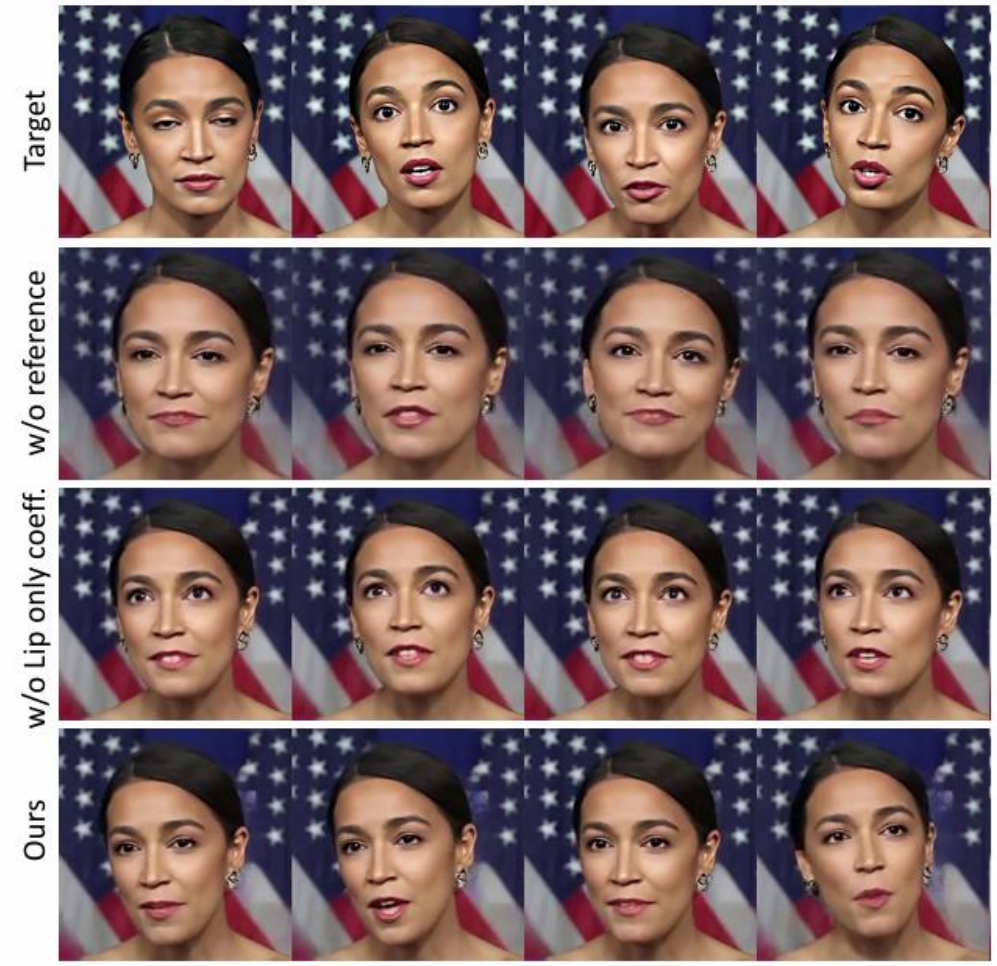
SadTalker及其他模型生成效果与目标图像的对比

3.2.3 与其他模型生成的图像对比

文章生成的视频在视觉质量上与原始目标视频极为相似，且实现了多样化头部姿态的生成。

相比之下，Wav2Lip 生成的半脸存在模糊现象；PC-AVS 和 Audio2Head 在身份保留方面表现欠佳，Audio2Head 甚至只能生成正面的说话头部；MakeItTalk 和 Audio2Head 由于采用 2D 变形技术，生成的面部视频出现扭曲。

3.2 结果分析



3.2.4 ExpNet 消融实验

Method	LSE-C ↑	LSE-D ↓
Speech2Gesture [10]	0.878	13.889
OursFull (Lip coeffs. + β_0 + $\mathcal{L}_{read}$)	**7.290**	**7.772**
w/o β_0 & $\mathcal{L}_{read}$	5.241	9.532
w/o $\mathcal{L}_{read}$	6.993	7.841
w/ real coeffs.	6.567	8.061

ExpNet 组件效果对比

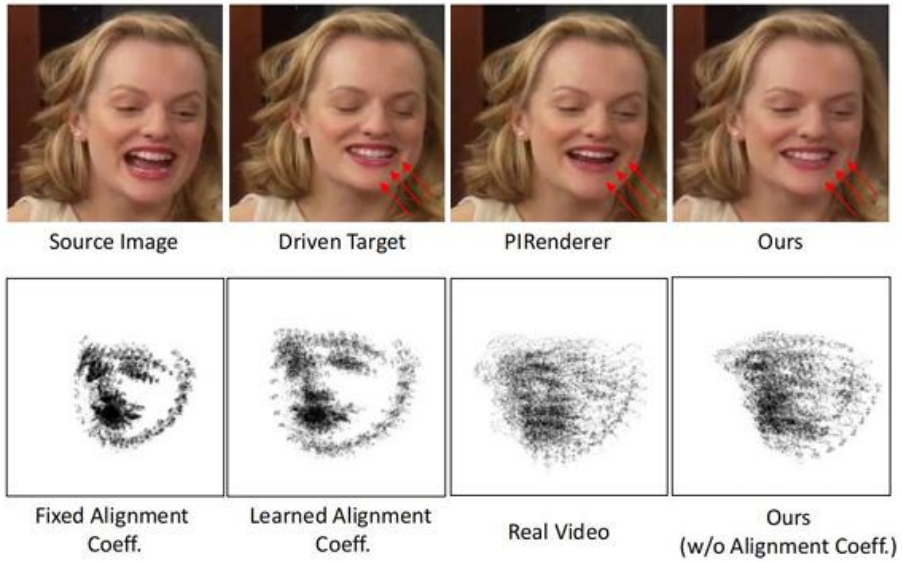
选择Speech2Gesture 为基线模型，对比发现将头部姿态和表情系数解耦训练更具有优势，同时网络添加的初始表情、嘴唇损失以及lip-only系数都能提高图片生成效果。

3.2 结果分析

3.2.5 PoseVAE及面部渲染消融实验

Method	Diversity↑	Beat Align↑
Speech2Gesture [10]	0.1574	0.274
OurFull (Single Fixed Style)	0.2735	0.287
w/o $\mathcal{L}_{gan}$	0.2500	0.271
w/o initial pose	0.2725	0.278
w/o audio	0.2566	0.274
w/o all conditions	0.2631	0.279
OursFull (Mixed Style)	**0.2778**	**0.293**

PoseVAE 消融实验结果



面部渲染的消融研究

从运动多样性和音频对齐方面评估，完整的SadTalker模型在这两项指标上都更优，各个条件对提升图像质量都由促进效果；此外，如果模型采用随机选择身份标签，在多样性性能上表现更为优异。

在重建质量上与 PIRenderer 相比，SadTalker 表情重建和唇同步更优；此外，使用固定或可学习对齐系数生成的头部姿势不佳，模型直接以头部姿势和表情作参数，效果更逼真。

Part.04

总结与思考

Summary and Reflection

南京郵電大學
Nanjing University of Posts and Telecommunications

4.1 研究成果总结

01 解耦系统架构

SadTalker利用 3DMM 的运动系数作为中间表示，通过 ExpNet 和 PoseVAE 分别生成表情和头部姿态运动系数，再借助 3D 感知人脸渲染技术合成最终视频。

02 网络结构高效

ExpNet 通过参考第一帧表情系数、使用lip-only系数作为目标减少了其他面部运动的干扰；PoseVAE 基于条件 VAE生成稳定且具有不同风格的头部运动。

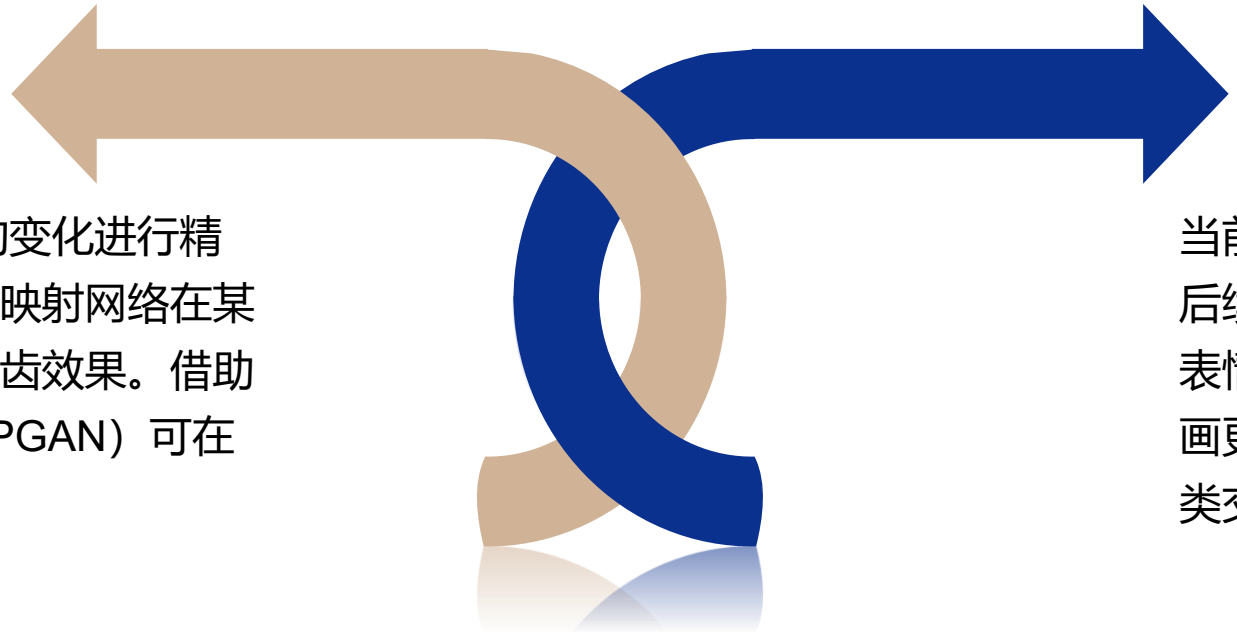
03 性能表现优异

在与其他先进方法的对比中，SadTalker在整体视频质量、头部姿态多样性上表现更优，在嘴唇同步指标上与其他方法相当，且生成的视频在视觉效果上更逼真。

4.2 未来工作展望

解决伪影现象

3DMM 无法对眼睛和牙齿的变化进行精确建模，导致面部渲染中的映射网络在某些情况下难以合成逼真的牙齿效果。借助盲人面部修复网络（如 GFPGAN）可在一定程度上改善这一问题。



面部表情的多样化

当前研究主要聚焦嘴唇运动和眨眼，后续可将情绪、注视方向等更多面部表情及身体动作纳入研究，使生成动画更具表现力与真实感，额外模拟人类交流时丰富的非语言信息。