

基于结构化潜码引导 NeRF 的语音驱动人脸重演

谢志峰^{1,2)}, 郑迦恒¹⁾, 王吉¹⁾, 梁佳佳¹⁾, 马利庄^{2,3)}

¹⁾(上海大学上海电影学院 上海 200072)

²⁾(上海电影特效工程技术研究中心 上海 200072)

³⁾(上海交通大学计算机科学与工程系 上海 200240)
(zhifeng_xie@shu.edu.cn)

摘要: 语音驱动的人脸重演的目标是生成与输入语音内容相匹配的高保真人脸面部动画。然而, 由于音频与视频模态之间存在鸿沟, 当前方法难以实现高质量的面部重演。针对现有方法保真度低、唇音同步效果差等问题, 提出一种基于结构化潜码引导隐式神经表示的语音驱动人脸重演方法, 以人脸点云序列作为中间表示, 将语音驱动人脸重演分解为跨模态映射和神经辐射场渲染两大任务分别解决。首先, 通过跨模态映射从音频预测人脸表情系数, 利用人脸三维重建技术获得人脸身份系数; 然后, 基于 3DMM 模型合成人脸点云动画序列; 接着, 使用顶点位置信息构建结构化隐式神经表示, 回归场景中每个采样点的密度和颜色值; 最后, 通过体绘制技术渲染人脸 RGB 帧, 并装配到原图像中。在多个时长为 3~5 min 的单人演讲视频上的可视化比较、量化评估、主观评估等实验结果表明, 文中所提方法在唇音同步效果与图像生成精度上优于 AD-NeRF 等方法, 能够实现高保真语音驱动人脸重演。

关键词: 音频驱动人脸重演; 隐式神经表示; 神经辐射场; 跨模态
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20026

Speech-Driven Facial Reenactment Based on Implicit Neural Representations with Structured Latent Codes

Xie Zhifeng^{1,2)}, Zheng Jiaheng¹⁾, Wang Ji¹⁾, Liang Jiajia¹⁾, and Ma Lizhuang^{2,3)}

¹⁾(Shanghai Film Academy, Shanghai University, Shanghai 200072)

²⁾(Shanghai Engineering Research Center of Motion Picture Special Effects, Shanghai 200072)

³⁾(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240)

Abstract: The goal of speech-driven facial reenactment aims to generate high-fidelity facial animation matching with the input speech content. However, existing methods can hardly achieve high-quality facial reenactment because of the gap between audio and video modals. In order to address the problems of existing methods such as low fidelity and poor lip sync effect, we propose a speech-driven facial reenactment method based on implicit neural representations with structured latent codes, which takes the point cloud sequence of human face as the intermediate representation, decomposing the speech-driven facial reenactment into two tasks: cross-modal mapping and neural radiance fields rendering. Firstly, we predict the facial expression coefficients through cross-modal mapping and get the facial identity coefficients by 3D face reconstruction; then, we synthesize face point cloud sequence based on 3DMM; next, we use the position of

收稿日期: 2022-10-24; 修回日期: 2023-07-11. 基金项目: 上海市自然科学基金(19ZR1419100); 上海市级科技重大专项(2021SHZDZX0102); 上海市科技创新行动计划人工智能科技支撑项目(21511101200); 上海市青年科技英才扬帆计划(22YF1420300). 谢志峰(1982—), 男, 博士, 副教授, 博士生导师, 主要研究方向为图像处理、深度学习; 郑迦恒(1998—), 男, 硕士研究生, 主要研究方向为深度学习、图像处理; 王吉(1999—), 女, 硕士研究生, 主要研究方向为深度学习、图像处理; 梁佳佳(1999—), 男, 硕士研究生, 主要研究方向为深度学习、图像处理; 马利庄(1963—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机图形学、图像处理.

vertices constructing the structured implicit neural representations and regress density and color for each sampling points; finally, we render RGB frames of human face through volume rendering techniques and assemble them into original image. Experiments results on multiple 3—5 min individual speech videos, including visual comparison, quantitative evaluation, and subjective assessment demonstrate that our method achieves better results than state-of-the-art methods such as AD-NeRF in terms of lip-sync accuracy and image generation precision, which can achieve high-fidelity speech-driven facial reenactment.

Key words: audio-driven facial reenactment; implicit neural representations; neural radiance field(NeRF); cross-modal

人脸重演^[1]任务旨在给定视频、音频等信号作为输入源, 驱动目标人脸将重演与输入源相匹配的面部动作, 同时保持人脸身份信息不变. 此类任务中, 最先发展的领域属视频驱动的人脸重演^[2], 经多年研究, 视频驱动的人脸重演技术如今已日趋完善. 然而, 以视频作为驱动源的人脸重演技术普遍存在缺乏对于遮挡的鲁棒性, 并且依赖于专业人士的表演. 相比视频序列, 语音信号获取成本更低, 使用更便捷, 还可以借助语音中的音节顿挫等信息个性化重演面部表情. 因此, 语音驱动的人脸重演技术^[3-9]近年来逐渐成为图形学与计算机视觉领域的研究热点, 可运用于虚拟主持人、影视后期编辑、视频会议等方面, 具有广泛的研究价值.

然而, 构建语音信号到人脸视频的跨模态映射并非易事, 使用长短期记忆(long short term memory, LSTM)网络、卷积神经网络(convolutional neural network, CNN)等模型^[3,10]建模语音与视频间的映射关系会出现生成图像模糊、网络泛化性能弱等问题. 最近, 一些研究通过加入人脸关键点、人脸三维模型等介质作为重演管线的中间表示, 将语音驱动人脸重演过程分解为 2 部分, 分别构建音频到中间表示、中间表示到图像的映射, 显著地提升了重演效果. 但是此类优化方法在生成图像时受限于中间表示的表达精度, 在动态人脸眼睛、牙齿、皱纹等细节的生成和时间稳定性方面仍然存在问题.

为了生成高保真语音驱动人脸重演视频, 本文提出了一种新颖的重演方法. 给定输入音频, 首先使用基于 LSTM 的跨模态映射模块, 获取人脸点云动画序列; 然后, 构建基于结构化潜码引导的神经辐射场(neural radiance field, NeRF), 回归场景中采样点的密度和颜色值; 最后, 使用体绘制生成高质量重演视频. 通过在多个演讲视频上的实验表明, 本文方法在唇音同步效果和图像生成质量方面均优于对比方法, 能够实现高保真的语音

驱动人脸重演.

1 相关工作

1.1 音频驱动的人脸重演

语音驱动的人脸重演技术旨在生成与输入音频内容相匹配的高保真人脸面部动画. 2017 年, Chung 等^[3]提出一种基于编解码 CNN 与 LSTM 的方法学习音频与视频之间的映射关系. Suwajanakorn 等^[4]提出了一种特征映射方法, 通过网络学习音频特征到口型的映射, 使用基于主成分分析(principal component analysis, PCA)的模型合成嘴部纹理. 2020 年, Vougioukas 等^[5]提出了一种跨模态映射框架, 该方法基于时序生成对抗网络(temporal generative adversarial network, Temporal GAN)实现了端到端的语音驱动人脸表情重演. Chen 等^[6]以关键点作为中间表示, 使用基于 LSTM 的网络将音频转化到人脸关键点, 再通过关键点作为条件生成视频帧. 2020 年, Thies 等^[7]提出了 Neural Voice Puppetry, 以三维混合形状作为语音与图像的中间媒介, 能够实现带有个人谈话风格的嘴部重演. Yi 等^[8]以人脸模型作为重演中间表示, 通过基于生成对抗网络的模型细化人脸渲染结果, 实现了带有头部姿势的语音驱动人脸重演. 2021 年, Guo 等^[9]以 2 个 NeRF 直接构建由音频特征到重演视频的映射. 相较于其他深度学习方法, Guo 等^[9]方法能够生成更高精度的人脸动态细节, 但是也存在口型对应欠佳、接缝明显等问题. 上述方法通过引入中间表示, 音频驱动人脸重演方法的准确度和泛化性得到了提升. 其中, 三维人脸数据相较于二维人脸关键点, 能够反映更多的人脸空间结构与运动信息, 在带有姿势的人脸重演中可以取得更好的效果; 但是现有方法仍然存在生成结果带有伪影、细节模糊、帧间闪烁等问题. 本文使用三维点云序

列作为中间表示, 构建结构化潜码引导的 NeRF 生成重演图像, 能够进一步提升音频驱动人脸重演的唇音同步效果和图像生成质量.

1.2 基于 3DMM 的单目三维人脸重建

单目三维人脸重建技术^[11-12]旨在从单幅人脸图像中重建人脸的形状与外观, 其中, 基于三维可形变人脸模型^[13](3D morphable models, 3DMM)的技术被广泛使用. 基于 3DMM 的单目人脸三维重建方法通过解决受线性人脸统计模型约束的非线性优化问题表达目标人脸, 3DMM 通常包含人脸身份、表情、姿势等系数. Deep3DFace^[14]方法截取了 BFM(basel face model) 人脸模型^[15]的正脸部分作为基模型, 采用 FaceWarehouse^[16]构建的表情基, 通过 CNN 回归 3DMM 系数, 能够实现快速、准确的单目人脸三维重建, 并且对遮挡和大幅度运动具有鲁棒性. 本文使用基于 3DMM 的单目三维重建技术获取目标人脸身份和姿势系数, 结合 LSTM 模型预测的表情系数以合成人脸点云模型.

1.3 NeRF 技术

2019 年, Sitzmann 等^[17]首次提出了一种场景表示网络, 通过对目标的形状和外观进行编码, 将场景表示为从世界坐标映射到场景特征表示的连续函数. 2020 年, Mildenhall 等^[18]提出了 NeRF, 通过使用隐式密度场和颜色场表示场景, 将对象形

状和外观的隐式表示转换为光线采样的结果, 实现了高质量的三维重建. 2021 年, Gafni 等^[19]通过 3DMM 模型跟踪人脸, 以面部表情系数作为条件驱动 NeRF, 可以基于 3~5 min 的输入视频创建目标人物动态 4D 化身. Peng 等^[20]提出了神经人体(neural body), 通过蒙皮多人线性模型(skinned multi-person linear, SMPL)锚定结构化潜码, 引导隐式神经表示回归密度与颜色, 能够进行动态人体三维重建和新视图合成. 本文使用 NeRF 技术, 通过体绘制生成人脸重演图像.

2 本文方法

本文方法如图 1 所示, 首先使用基于 3DMM 的人脸三维重建技术获取源对象的人脸身份系数 α 和姿势系数 ρ , 使用基于 LSTM 的跨模态的映射网络回归目标对象的人脸表情系数 β , 该网络从对齐的音频特征-表情系数数据中学习跨模态映射; 然后, 组合目标对象的表情系数和源对象的身份、姿势系数生成人脸点云模型作为中间表示; 接着, 基于人脸顶点位置锚定潜码, 构建结构化潜码引导的 NeRF, 回归每个采样点的密度和颜色值; 最后, 通过体绘制生成人脸 RGB 帧并装配回原始图像.

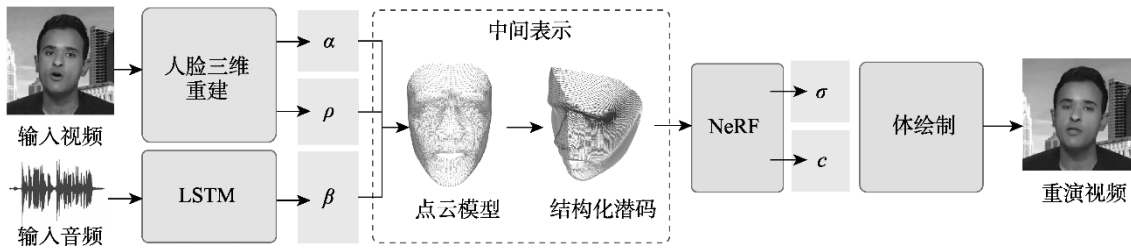


图 1 本文方法流程

2.1 人脸三维重建

本文使用 Deep3DFace^[14]作为人脸重建算法, 选取顶点数 35 709 的正面人脸模型作为基模型; 对于任一人脸图像 $\bar{I}$, 通过 CNN 拟合人脸的形状与外观, 该过程表示为 $\chi(\bar{I}) = (\alpha, \beta, \tau, o, \rho) \in \mathbb{R}^{257}$. 其中, α , β , τ , o 和 ρ 分别表示人脸的身份、表情、材质、光照和姿势的系数, 人脸身份系数 $\alpha \in \mathbb{R}^{80}$, 表情系数 $\beta \in \mathbb{R}^{64}$, 材质系数 $\tau \in \mathbb{R}^{80}$, 光照系数 $o \in \mathbb{R}^{27}$, 姿势系数 $\rho \in \mathbb{R}^6$. 对于任意的 3D 人脸, 其形状表示为 $S = \bar{S} + B_{id}\alpha + B_{exp}\beta$. 其中, $\bar{S}$ 表示人脸的平均形状; B_{id} 和 B_{exp} 分别是 3DMM 模型的

形状基底和表情基底. 在训练阶段, 本文记录目标人脸身份系数 α , 每帧姿势系数 ρ .

2.2 语音到顶点的映射

受 Yi 等^[8]的启发, 本文使用 LSTM 完成音频到表情系数的映射. 对于输入音频, 本文提取梅尔倒谱特征 $s = \{s^{(1)}, s^{(2)}, \dots, s^{(t)}\}$, 音频特征到表情系数的映射表示为

$$[\tilde{\beta}^{(t)}, h^{(t)}, c^{(t)}] = F_{LSTM}(E(s^{(t)}), h^{(t-1)}, c^{(t-1)}).$$

其中, E 是梅尔倒谱特征 $s^{(t)}$ 的编码器; $h^{(t-1)}$ 和 $c^{(t-1)}$ 分别是 LSTM 的隐藏层和细胞状态; $\tilde{\beta}^{(t)}$ 是第 t 帧的人脸表情系数的预测值. 本文 LSTM 模型

在 LRW 数据集^[21]上进行训练, 在测试阶段, 首先使用 500 帧视频对网络进行微调, 然后使用 LSTM 网络通过输入音频特征 $s = \{s^{(1)}, s^{(2)}, \dots, s^{(t)}\}$ 回归表情系数 $\tilde{\beta} = \{\tilde{\beta}^{(1)}, \tilde{\beta}^{(2)}, \dots, \tilde{\beta}^{(t)}\}$. 因此, 第 t 帧的人脸重演点云 $V^{(t)}$ 表示为 $V^{(t)} = \bar{S} + B_{id}\alpha + B_{exp}\tilde{\beta}^{(t)}$. 其中, $\bar{S}$ 表示 3DMM 预先构建的人脸模型的平均顶点.

2.3 顶点到图像的渲染

2.3.1 结构化潜码引导的 NeRF

受 Neural Body^[20]的启发, 为使模型能够高质量地表达动态人脸的形状和外观, 本文使用结构化潜码引导的 NeRF 完成对采样点密度和颜色值的回归. 如图 2 所示, 对于世界空间下的人脸点云模型 V , 本文锚定潜码 $Z = \{Z_1, Z_2, \dots, Z_{35709}\}$, 每段潜码的维度 $d = 16$. 通过向潜码中加入三维位

置信息作为先验, 能够降低 NeRF 的训练难度; 同时, 将同组潜码映射到不同帧的 NeRF 中, 隐式场能够整合全局信息优化模型, 以精确地表达动态人脸细节.

由于 NeRF 需要对空间中每个点回归颜色和密度值, 因此为保证模型学习效果, 本文使用 SparseConvNet^[22]对潜码进行卷积. 首先计算输入顶点 V 的包围盒, 并按空间分辨率 $\bar{D} \times \bar{H} \times \bar{W}$ 将其分割为小体素, 非空体素内的潜码取该体素所包含的所有顶点潜码的平均值; 然后, 使用人脸姿势系数 ρ 将其变换到标准空间; 最后, 使用三维稀疏卷积处理输入体积, 将其转换为具有多尺度下采样尺寸的潜码体积. 在实验中, 本文设置小体素的边长为 5 mm, 连接三维稀疏卷积网络第 5, 9, 13, 17 层的输出构建潜码体积, 具体网络结构如图 3a 所示.

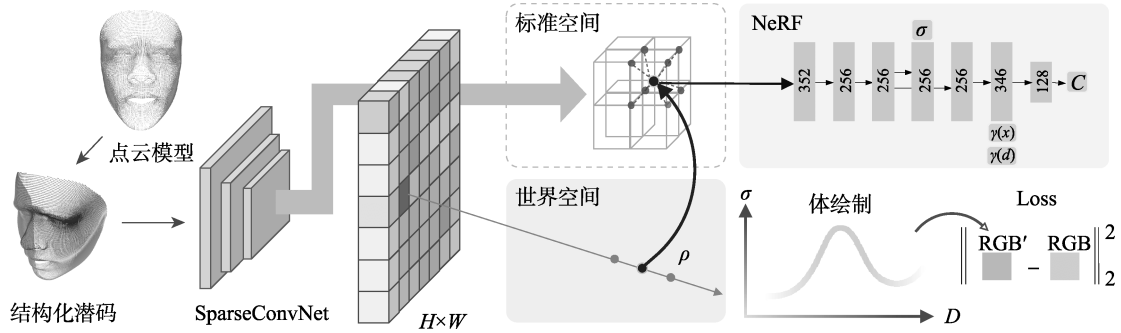


图 2 基于结构化潜码引导的 NeRF 模型

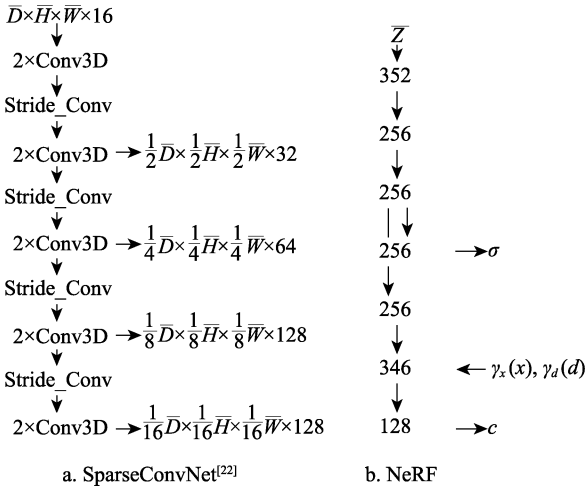


图 3 SparseConvNet 和 NeRF 网络结构图

对于 NeRF 在世界空间中的采样点 x , 首先利用头部姿势系数 ρ 将其与标准空间中的潜码体积对齐; 然后, 在潜码体积中通过三线性插值获取该采样点处的潜码 $\bar{Z}$, 该过程表示为

$\bar{Z} = \varphi(x, \rho, Z, V)$; 最后, 以潜码 $\bar{Z}$ 作为输入, 通过 NeRF 模型回归密度与颜色值. 本文将 t 时刻世界空间位置 x 处采样点的密度和颜色分别表示为

$$\sigma_t(x) = M_\sigma(\bar{Z}),$$

$$c_t(x) = M_c(\bar{Z}, \gamma_d(d), \gamma_x(x)).$$

其中, x 和 d 表示采样点位置和观察方向; M_σ 和 M_c 分别是 NeRF 回归密度和颜色的多层感知机, 网络结构如图 3b 所示. 为促进神经网络拟合包含高频变化的数据, 本文使用位置编码^[17]将输入 x 和 d 映射到高维空间, 即

$$\gamma(x) = (\sin(2^0 \pi x), \cos(2^0 \pi x), \dots, \sin(2^{l-1} \pi x), \cos(2^{l-1} \pi x)).$$

实验中, 本文设定对 x 编码的 $l = 10$, 对 d 编码的 $l = 4$.

2.3.2 体绘制

借助 NeRF 对场景密度和颜色值的回归, 本文使用体绘制方程^[23]计算尺寸为 $H \times W$ 的重演图像

的颜色值. 对于任一起点为 o , 方向为 d , 近远端分别为 m_n 和 m_f 的摄像机光线 $r(m) = o + md$, 其颜色值表示为 $C(r) = \int_{m_n}^{m_f} T(m) \sigma(r(m)) c(r(m), d) dm$. 其中, $\sigma(r(m))$ 和 $c(r(m), d)$ 分别表示采样点的密度和颜色值; $T(m) = \exp\left(-\int_{m_n}^m \sigma(r(s)) ds\right)$ 表示射线从 m_n 到 m_f 的累计透射率. 实验中, 本文在近远边界间对射线 $r(m)$ 取 $N_i = 64$ 个采样点 $\{x_i\}_{i=1}^{N_i}$, 实际颜色值表示为 $\hat{C}(r) = \sum_{i=1}^{N_i} T_i (1 - \exp(-\sigma_i(x_i) \delta_i)) c_i(x_i)$. 其中, $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_i(x_j) \delta_j\right)$; δ_i 表示采样点间的距离, 记 $\delta_i = \|x_{i+1} - x_i\|_2$.

2.3.3 损失函数

对于图像集 $\{I_t | t=1, 2, \dots, N_t\}$, 本文的优化目标是 minimized 生成图像与真实图像间的渲染误差, 即

$$\min_{Z, \theta} \sum_{t=1}^{N_t} L(I_t; Z, \theta).$$

其中, Z 表示输入潜码; θ 表示网络参数; 损失函数 L 定义为

$$L = \sum_{r \in R_{ray}} \|\hat{C}(r) - \bar{C}(r)\|_2^2.$$

其中, R_{ray} 表示摄像机光线的集合; $\bar{C}(r)$ 表示真实图像中的像素颜色值.

3 实验结果与分析

3.1 数据集构建

本文选取视频网站上 15 段演讲和播音视频构建训练数据集, 演讲者包含不同人种、肤色、年龄、性别; 每段视频仅包含一位演讲者, 画面大小裁剪到 450×450 像素, 总长度为 3~5 min. 为方便后续实验, 对于每段视频, 本文还分别截取 12 s, 40 s 和 120 s 片段构建测试数据.

3.2 实验环境及过程

对于数据集中的每段视频, 本文使用其中 40 s 片段训练结构化隐式神经表示模型; 实验在 Pytorch 框架下进行, 采用 Adam 优化器, 初始学习率设置为 5×10^{-4} , 在训练过程中使用指数衰减法进行优化; 每个 epoch 迭代 1000 幅图像, 共训练 150 个 epoch.

3.3 实验结果

本文分别对训练集中的 15 段视频进行了训练和测试, 部分可视化结果如图 4 所示, 可以看出,

本文方法泛化性能良好, 能够作用于不同肤色、年龄、性别和装扮的人物, 对于目标人物, 同样可以接受任意语音作为输入进行表情重演. 与受限于中间表示精度的生成式方法不同, 本文方法不依赖介质生成纹理和细节, 仅使用中间表示为 NeRF 提供三维先验, 然后通过隐式场对采样点回归密度和颜色, 使用体绘制生成 RGB 图像, 因此对于人脸眼睛、皱纹、牙齿等动态细节能够精确表达. 此外, 本文方法以中间表示连接音频和视频 2 种模态, 以两阶段生成的模式降低了网络学习跨模态映射的难度, 提升了口型重演的准确率. 最后, 本文方法使用了人脸三维点云作为中间表示, 打破了基于图像的方法受限于头部姿势运动的桎梏, 能够在重演时操控头部姿势. 如图 5 所示, 首先, 固定目标人物正对相机的嘟嘴表情(第 1 列), 然后, 匹配其点头(第 1 行)和摆头(第 2 行)头部姿势, 实现操控目标人物保持嘟嘴的重演表情点头和摆头.



图 4 驱动不同目标对象重演



图 5 驱动不同目标对象重演

3.4 对比实验

为验证本文框架的有效性, 将其与已有的音频驱动人脸重演方法进行比较. 首先, 与基于图像

的方法进行对比, 如图 6 所示, Vougioukas 等^[5]方法存在生成图像结果模糊、身份信息改变等问题, Chen 等^[6]方法仅嘴部有明显运动。此外, 这 2 种方法仅适用于低分辨率的、裁剪到正脸的图像, 本文方法能够生成带有姿势的完整人脸, 并且在重演过程中身份特征保持稳定。

接着, 本文与能够生成完整面部的 Yi 等^[8]和 Guo 等^[9]对比, 结果如图 7 所示, 结果显示, Yi 等^[8]在眼睛、牙齿等细节生成上存在伪影; Guo 等^[9]存在身体衔接裂缝、口型对应错误等问题, 在图中由红框标出; 本文方法生成图像不存在伪影, 且语音能够与口型良好对应。

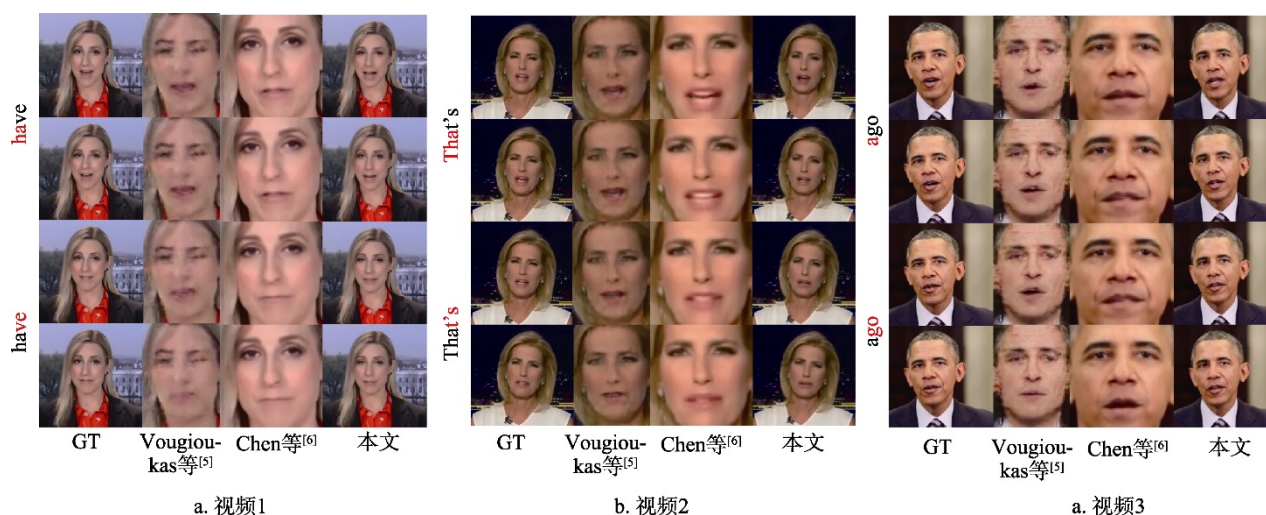


图 6 本文方法与基于图像的方法对比

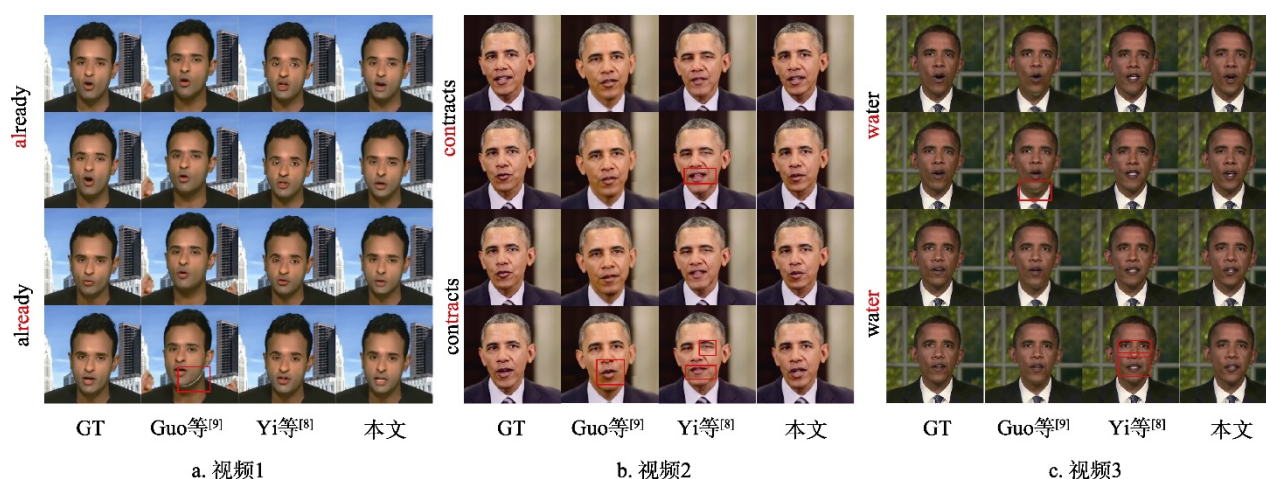


图 7 本文方法与完整面部生成方法对比

本文还在第 3.1 节所构建的数据集中与 Yi 等^[8]和 Guo 等^[9]进行了量化比较, 使用原始音频和头部姿势进行自重演计算 PSNR 与 SSIM 以评估图像生成质量, 使用 SyncNet 和 AU-MSE 指标测试唇音同步效果与口型匹配度。本文通过 SyncNet^[24]检测重演视频序列的置信度, 使用 OpenFace^[25]检测源音频与目标音频下半张脸相关面部动作单元(action units, AU), 并累计每项 AU 之间的 MSE 作为评估指标 AU-MSE。为保证模型泛化效果, 本文还从每段视频中截取 12 s 的片段作为测试集, 每种方法

单段视频的评估指标在 SyncNet 和 AU-MSE 上取所有测试音频的重演结果的平均值。实验中, 按照每种方法对训练素材时长的不同要求, 以不同的视频长短进行训练: Yi 等^[8]使用 12 s, 本文使用 40 s, Guo 等^[9]使用 120 s。表 1 展示了本文在图 7 中的 3 段视频上的量化比较结果, 每段视频片段中的 SyncNet 和 AU-MSE 指标取 3 段测试音频评估结果的均值。对于视频片段 1 和视频片段 3, 本文方法图像生成质量和唇音同步效果均优于其他 2 种完整面部生成方法; 在视频片段 2 中, 本文方法

AU-MSE 误差较高,但是在置信度和图像质量指标上超过另外 2 种方法,原因在于该视频人物说话方式多变,本文方法难以通过 40 s 的学习还原该人物说话风格。

表 1 对比完整面部生成方法量化结果

方法	视频片段 1			
	SyncNet	AU-MSE	PSNR/dB	SSIM
Guo 等 ^[9]	4.36	1.84	29.32	0.91
Yi 等 ^[8]	4.38	1.76	32.73	0.94
本文	4.40	0.89	36.80	0.97

方法	视频片段 2			
	SyncNet	AU-MSE	PSNR/dB	SSIM
Guo 等 ^[9]	3.33	1.86	23.81	0.86
Yi 等 ^[8]	3.50	1.97	27.28	0.89
本文	3.81	2.24	31.62	0.95

方法	视频片段 3			
	SyncNet	AU-MSE	PSNR/dB	SSIM
Guo 等 ^[9]	3.35	2.45	25.37	0.89
Yi 等 ^[8]	4.04	2.02	30.02	0.89
本文	4.51	1.80	33.88	0.96

注: 粗体表示最优值。

此外,本文也使用主观问卷评测的方式进行用户研究。挑选来自 Yi 等^[8], Guo 等^[9], Vougioukas 等^[5], Chen 等^[6]和本文方法的共 20 个重演结果(每种方法 4 个视频),随机排序后向 30 位观看者发放问卷,要求观看者从真实感(S_1)、流畅度(S_2)、唇音同步(S_3)、图像质量(S_4) 4 个维度对每个视频进行评分(1~5 分),评分结果如表 2 所示,结果显示,本文方法在真实感、流畅度、唇音同步、图像质量上均被认为取得了最佳效果。

表 2 主观问卷评测结果

方法	S_1	S_2	S_3	S_4
Vougioukas 等 ^[5]	1.83	1.47	2.00	1.17
Chen 等 ^[6]	2.13	3.07	3.40	3.17
Guo 等 ^[9]	3.80	3.63	3.67	4.20
Yi 等 ^[8]	3.57	3.37	3.60	2.97
本文	4.03	3.80	3.97	4.37

注: 粗体表示最优值。

3.5 消融实验

为了验证本文结构化隐式神经表示在音频重演框架中的有效性,本文将其替换为以表情系数驱动的 NeRF 进行消融实验。该实验在 40 s 单人演讲视频上进行,共迭代 150 个 epoch;实验结果可

视化如表 3 所示,最后一列中,每行图像的上下部分是对应的本文结果和消融实验结果图像的细节放大图。直接使用表情系数驱动的方法会导致在眼部和嘴部出现大量伪影和局部模糊。本文还对自重演测试进行了量化分析,结果如表 4 所示,本文方法在 PSNR, SSIM 和 AU-MSE 指标均上优于直接使用表情系数驱动人脸重演的方法;消融实验表明,本文方法将音频预测的表情系数转化为三维先验,利用同一组潜码整合全局数据,优化了 NeRF 学习动态人脸的效果。

表 3 本文消融实验结果可视化

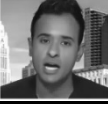
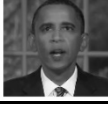
测试	GT	本文结果	消融实验结果	图像细节
自重演 音频测试				
				
泛化音频 测试				
				

表 4 自重演测试消融实验量化结果

自重演测试	PSNR/dB	SSIM	AU-MSE
结构化潜码驱动	34.15	0.94	0.64
表情系数驱动	33.12	0.92	0.75

注: 粗体表示最优值。

4 结 语

本文提出了一种基于结构化隐式神经表示的语音驱动人脸重演方法,以人脸点云序列作为中间介质,将语音驱动人脸重演任务分解为音频到顶点的映射和顶点到图像的渲染两大模块,降低了模型学习跨模态映射的难度;通过构建结构化潜码引导的 NeRF,以三维先验辅助 NeRF 学习人脸的形状和纹理,实现了高保真人脸渲染。

本文方法仍存在部分局限性:无法对人物眨眼进行个性化控制。当通过音频到表情映射获取的表情系数与目标人物表情系数在隐空间距离较远时,有时会出现异常眨眼情况,具体的失败结果如图 8 所示。图 8a 展示了本文方法正常眨眼的结

果, 图 8b 展示了异常眨眼的结果. 未来的工作中可以通过尝试对眨眼进行单独解耦的方式解决人物眨眼操控问题.



图 9 失败结果

参考文献(References):

- [1] Fei Jianwei, Xia Zhihua, Yu Peipeng, *et al.* Survey of face synthesis[J]. *Journal of Frontiers of Computer Science and Technology*, 2021, 15(11): 2025-2047(in Chinese)
(费建伟, 夏志华, 余佩鹏, 等. 人脸合成技术综述[J]. *计算机科学与探索*, 2021, 15(11): 2025-2047)
- [2] Thies J, Zollhöfer M, Stamminger M, *et al.* Face2Face: real-time face capture and reenactment of RGB videos[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2387-2395
- [3] Chung J S, Jamaludin A, Zisserman A. You said that? [OL]. [2022-10-24]. <https://arxiv.org/abs/1705.02966>
- [4] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing Obama: learning lip sync from audio[J]. *ACM Transactions on Graphics*, 2017, 36(4): Article No.95
- [5] Vougioukas K, Petridis S, Pantic M. Realistic speech-driven facial animation with GANs[J]. *International Journal of Computer Vision*, 2020, 128(5): 1398-1413
- [6] Chen L L, Maddox R K, Duan Z Y, *et al.* Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 7824-7833
- [7] Thies J, Elgharib M, Tewari A, *et al.* Neural voice puppetry: audio-driven facial reenactment[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 716-731
- [8] Yi R, Ye Z P, Zhang J Y, *et al.* Audio-driven talking face video generation with learning-based personalized head pose[OL]. [2022-10-24]. <https://arxiv.org/abs/2002.10137>
- [9] Guo Y D, Chen K Y, Liang S, *et al.* AD-NeRF: audio driven neural radiance fields for talking head synthesis[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 5764-5774
- [10] Xie Z F, Zhang W L, Sheng B, *et al.* BaGFN: broad attentive graph fusion network for high-order feature Interactions[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, 34(8): 4499-4513
- [11] Mao Aihua, Situ Hengge. Image-driven automatic 3D human face modeling and editing algorithm[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2019, 31(1): 17-25(in Chinese)
(毛爱华, 司徒亨哥. 图像驱动的三维人脸自动生成与编辑算法[J]. *计算机辅助设计与图形学学报*, 2019, 31(1): 17-25)
- [12] Wang Han, Xia Shihong. Construction of facial shape with details from single image[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2017, 29(7): 1256-1266(in Chinese)
(王涵, 夏时洪. 单张图片自动重建带几何细节的人脸形状[J]. *计算机辅助设计与图形学学报*, 2017, 29(7): 1256-1266)
- [13] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[C] //Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 1999: 187-194
- [14] Deng Y, Yang J L, Xu S C, *et al.* Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2019: 285-295
- [15] Paysan P, Knothe R, Amberg B, *et al.* A 3D face model for pose and illumination invariant face recognition[C] //Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance. Los Alamitos: IEEE Computer Society Press, 2009: 296-301
- [16] Cao C, Weng Y L, Zhou S, *et al.* FaceWarehouse: a 3D facial expression database for visual computing[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(3): 413- 425
- [17] Sitzmann V, Zollhöfer M, Wetzstein G. Scene representation networks: continuous 3D-structure-aware neural scene representations[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2019: Article No.101
- [18] Mildenhall B, Srinivasan P P, Tancik M, *et al.* NeRF: representing scenes as neural radiance fields for view synthesis[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 405-421
- [19] Gafni G, Thies J, Zollhöfer M, *et al.* Dynamic neural radiance fields for monocular 4D facial avatar reconstruction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8645-8654
- [20] Peng S D, Zhang Y Q, Xu Y H, *et al.* Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 9050-9059
- [21] Chung J S, Zisserman A. Lip reading in the wild[C] //Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 479-487

- ngs of the 13th Asian Conference on Computer Vision. Heidelberg: Springer, 2017: 87-103
- [22] Graham B, Engelcke M, van der Maaten L. 3D semantic segmentation with submanifold sparse convolutional networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 9224-9232
- [23] Kajiya J T, Von Herzen B P. Ray tracing volume densities[J]. ACM SIGGRAPH Computer Graphics, 1984, 18(3): 165-174
- [24] Chung J S, Zisserman A. Out of time: automated lip sync in the wild[C] //Proceedings of the Asian Conference on Computer Vision. Heidelberg: Springer, 2017: 251-263
- [25] Baltrusaitis T, Zadeh A, Lim Y C, *et al.* OpenFace 2.0: facial behavior analysis toolkit[C] //Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 59-66